REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
26-06-2002	FINAL	05/01/99 - 04/30/02

4. TITLE AND SUBTITLE
EXPONENTIALLY UNSTABLE PLANTS WITH SATURATING ACTUATORS

5a. CONTRACT NUMBER
N00014-99-1-0670

5b. GRANT NUMBER
N00014-99-1-0670

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
ZONGLI LIN, PROJECT DIRECTOR

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF VIRGINIA, SCHOOL OF ENGINEERING AND APPLIED SCIENCE,
DEPT. OF ELECTRICAL & COMPUTER ENGINEERING, THORNTON HALL,
P.O. BOX 400258, CHARLOTTESVILLE, VA 22904-4258

8. PERFORMING ORGANIZATION REPORT NUMBER
GG10082

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
OFFICE OF NAVAL RESEARCH, BALLSTON CENTRE TOWER ONE,
800 NORTH QUINCY STREET, ARLINGTON, VA 22217-5660

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
NO DISTRIBUTION LIMITATIONS. NO CLASSIFIED OR RESTRICTED DATA.

20020701 073

13. SUPPLEMENTARY NOTES

14. ABSTRACT
EVERY PHYSICAL ACTUATOR IS SUBJECT TO SATURATION. FOR THIS REASON, THE ORIGINAL FORMULATIONS OF MANY FUNDAMENTAL CONTROL PROBLEMS, INCLUDING COTROLLABILITY AND TIME OPTIMAL CONTROL, ALL REFLECT THE CONSTRAINTS IMPLOSED BY ACTURATOR SATURATION. CONTROL PROBLEMS THAT INVOLVE HARD NONLINEARITIES SUCH AS ACTUATOR SATURATION,HOWEVER, TURNED OUT TO BE DIFFICULT TO DEAL WITH. AS A RESULT, EVEN THOUGH THERE HAVE BEEN CONTINUAL EFFORTS IN ADDRESSING ACTUATOR SATRUATION FOR A CHRONOLOGICAL BIBLIOGRAPHY ON THIS SUBJECT, ITS EFFECT HAS BEEN IGNORED IN MOST OF THE MODERN CONTROL LITERATURE. THIS REPORT SUMMARIEZES THE RESULTS OBTAINED UNDER THIS PROJECT. FOR CLARITY WE HAVE ITEMIZED EACH RESULT AS APPEARED IN THE BOOK OR JOURNAL PUBLICATIONS.

15. SUBJECT TERMS
PHYSICAL ACTUATOR SATURATION

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			ZONGLI LIN
U	U	U	UU		19b. TELEPHONE NUMBER *(Include area code)* 434-924-6342

Final Technical Report, June 2002

Project Title: Exponentially Unstable Plants with Saturating Actuators (N00014-99-1-0670)

To: Dr. Allen Moshfegh, Chief of Naval Research, ONR

From: Zongli Lin, ONR Young Investigator

Reproduced From Best Available Copy

Contents

1 Introduction and Organization 1

2 Main Results: Book and Journal Publications 3

3 Future Research Topics 10

4 Appendix: Copies of Publications 12

1 Introduction and Organization

Every physical actuator is subject to saturation. For this reason, the original formulations of many fundamental control problems, including controllability and time optimal control, all reflect the constraints imposed by actuator saturation. Control problems that involve hard nonlinearities such as actuator saturation, however, turned out to be difficult to deal with. As a result, even though there have been continual efforts in addressing actuator saturation for a chronological bibliography on this subject), its effect has been ignored in most of the modern control literature.

On the other hand, it has been well known that, when the actuator saturates, the performance of the closed-loop system designed without considering actuator saturation may seriously deteriorate. In the extreme case, even the stability may be lost. Actuator saturation is a particularly important consideration in reconfigurable control systems. Following a system failure, it is likely that the available control authority will be reduced, and that the reconfigurable control system will drive actuators into their saturated regime to stabilize the systems and to attempt to regain nominal performance. A well-known example of performance degradation (*e.g.*, large overshoot and large settling time) occurs when a linear compensator with integrators, say a PID compensator, is used in a closed-loop system. During the time when the actuator saturates, the error is continuously integrated even though the control is not what it should be, and hence, the states of the compensator attain values that lead to larger controls than the actuator limit. This phenomenon is called integrator windup. In the absence of integrators, a given reference setpoint might result in a different steady state response, causing the need to reset the reference to a value different from the desired setpoint. With integral control, the controllers automatically bring the output to the given reference setpoint and hence the integrator does the reset. For this reason, integrator windup is sometimes referred to as reset windup.

A practical approach to compensating this performance degradation due to actuator saturation is to add some problem-specific anti-windup schemes to deal with the adverse effects caused

by saturation. These schemes are typically introduced using *ad hoc* modifications and extensive simulations. The basic idea behind these schemes is to introduce additional feedbacks in such a way that the actuator stays properly within its limit. For example, several schemes have been proposed to solve the reset windup problem when integrators are present in the forward loop. Most of these schemes lead to improved performance but poorly understood stability properties. More recently, some researchers have attempted to provide more systematic and more general schemes to deal with the problem.

In this project, we have taken the approach of considering actuator saturation at the outset of control design. As seen in the recent literature, there has been a resurge of interest in this approach, possibly owing to its systematic nature. A fundamental issue is the identification of the class of linear systems that are (globally) asymptotically null controllable by bounded controls. A system is said to be globally asymptotically null controllable by bounded controls if, for a given bound on the controls, every state in its state space can be driven to the origin either in a finite time or asymptotically by a bounded control. In particular, it was established earlier that a linear stabilizable system having all its poles in the closed left-half plane is globally asymptotically null controllable. For this reason, a linear stabilizable system with all its poles in the closed left-half plane is commonly said to be asymptotically null controllable with bounded controls, or ANCBC, and most of the recent work has been focused on ANCBC systems. For such systems, various types of feedback laws have been proposed that work globally (on the entire state space) or semi-globally (on any a priori given arbitrarily large bounded set in the state space).

It is clear that a linear system having poles in the open right-half plane is not globally asymptotically null controllable with bounded controls. Any feedback laws designed for such a system would not work globally. Two natural questions to ask are:

- for a general, not necessarily ANCBC, linear system with saturating actuators, what is the asymptotically null controllable region, the set of all states that can be driven to the origin in a finite time or asymptotically by a bounded control?
- how to design feedback laws that work on the entire asymptotically null controllable region or a large portion of it?

This project answered these two questions, for both continuous-time and discrete-time systems. We started with explicit analytical descriptions of the asymptotically null controllable region. Once we have identified this region, we designed design feedback laws that achieve various closed-loop performance specifications on the entire asymptotically null controllable region or a large portion of it. These performances range from the basic control problem of stabilization to those beyond large domain of attraction such as transient properties, disturbance rejection and output regulation. The key to achieving stability and performance on the entire asymptotically null controllable region or on a large portion of it is to push actuators to work to their full capacity.

This final technical report is organized as follows. Section 2 summarizes the results obtained under this project. For clarity, we will itemize each result as appeared in the book or journal publications. Copies of these publications are included in the Appendix. Section 3 includes a few topics for future research.

We greatly appreciate the support of the ONR, in particular, that of Dr. Allen Moshfegh and the Young Investigator Program.

2 Main Results: Book and Journal Publications

The results obtained under this project have been reported in one book, one book chapter, 20 journal papers and many conference papers. Another 7 papers are currently under review for journal publication. In what follows, we will first list the book, which contains a systematic description of the majority of our results. The journal papers and the book chapter are then listed in the chronological order. The results originally reported in conference papers have either been published in journals or are currently under review for journal publication. For this reason, conference publications are not listed.

1. T. Hu and Z. Lin, *Control Systems with Actuator Saturation: Analysis and Design*, Birkhäuser, Boston, xvi, 392 p, 2001.

   Every physical actuator is subject to saturation. When the actuator saturates, the performance of the control system designed without considering actuator saturation will seriously deteriorate. Currently, there is a surge of interest in increasing the practical applicability of control theory by incorporating the effect of saturation into the design of a control system.

   Control Systems with Actuator Saturation: Analysis and Design examines the problem of actuator saturation in depth. The overall approach takes into account the saturation nonlinearities at the outset of the control design. In the case that a control law is designed a priori to meet either the performance or stability requirement, it analyzes the closed-loop system under actuator saturation systematically and redesigns the controller in such a way that the performance is retained while stability is improved. It also presents some related results on systems with state saturation or sensor saturation.

   Features and topics:

   - Results apply to general open loop systems, including exponentially unstable ones. Thus, they are widely applicable to practical systems.
   - Problem such as controllability, stabilizability, transience performance, disturbance rejection, and output regulation are treated in a systematic manner.
   - Analytic description of the null controllable region is first obtained. Various feedback laws are designed that work on the entire null controllable region or an arbitrarily large portion of it.
   - Analysis tools are developed for performance assessment of existing control systems under actuator saturation.
   - Examples are worked out in detail to demonstrate the usage of the results developed.

2. X. Bao, Z. Lin and E.D. Sontag, "Finite gain stabilization of discrete-time linear systems subject to actuator saturation," *Automatica*, Vol. 36, No. 2, pp. 269-277, February, 2000.

   It is shown that, for neutrally stable discrete-time linear systems subject to actuator saturation, finite gain l_p stabilization can be achieved by linear output feedback, for all $p \in (1, \infty]$. An explicit construction of the corresponding feedback laws is given. The feedback laws constructed also result in a closed-loop system that is globally asymptotically stable, and in an input-to-state estimate.

3. Z. Lin, A. Saberi, A. Stoorvogel and R. Mantri, "An improvement to the low gain design for discrete-time linear systems subject to input saturation – solutions of semi-global output

regulation problems," *International Journal of Robust and Nonlinear Control*, Vol. 10, No. 3, pp. 117-135, March, 2000.

A low gain design for linear discrete-time systems subject to input saturation was recently developed to solve both semi-global stabilization and semi-global output regulation problems. This paper proposes an improvement to the low gain design and determines controllers with the new design that achieve semi-global output regulation. The improvement is reflected in better utilization of available control capacity and consequently better closed-loop performance.

4. T. Hu and Z. Lin, "A complete stability analysis of planar linear systems under saturation," *IEEE Transactions on Circuits and Systems - Part I: Fundamental Theory and Applications*, Vol. 47, No. 4, pp. 498-512, April, 2000.

   A complete stability analysis is performed on a planar system of the form $\dot{x} = \text{sat}(Ax)$, where A is a Hurwitz matrix and sat is the saturation function. Necessary and sufficient conditions for the system to be globally asymptotically stable or to have a closed trajectory are explicitly given in terms of the entries of A. These conditions also indicate that the system always has a closed trajectory if it is not globally asymptotically stable.

5. Z. Lin and G. Tao, "Adaptive control of a weakly nonminimum phase linear system," *IEEE Transactions on Automatic Control*, Vol. 45, No. 4, pp. 824-829, April, 2000.

   For a weakly nonminimum phase linear system, we design an adaptive state feedback control law that causes the system output to track a desired trajectory to an arbitrarily high degree of precision. The key to this is the use of a low gain feedback design technique.

6. T. Hu and Z. Lin, "On enlarging the basin of attraction for linear systems under saturated linear feedback," *Systems & Control Letters*, Vol. 40, No. 1, pp. 59-69, May, 2000.

   We consider the problem of enlarging the basin of attraction for a linear system under saturated linear feedback. An LMI based approach to this problem is developed. For discrete-time system, this approach is enhanced by the lifting technique, which leads to further enlargement of the basin of attraction. The low convergence rate inherent with the large invariant set (hence, the large basin of attraction) is prevented by the construction of a sequence of invariant ellipsoids nested within the large one obtained.

7. X. Bao and Z. Lin, "On L_p input to state stabilizability of affine nonlinear systems subject to actuator saturation," *Journal of the Franklin Institute*, Vol. 337, pp. 691-712, September, 2000.

   The L_p input to state stabilizability of affine in control nonlinear systems subject to actuator saturation is examined. A few sets of conditions under which the system is (finite gain) L_p input to state stabilizable are identified and the stabilizing feedback laws are explicitly constructed.

8. T. Hu and Z. Lin, "Practical stabilization on the null controllable region of exponentially unstable linear systems subject to actuator saturation nonlinearities and disturbance," *International Journal of Robust and Nonlinear Control*, Vol. 11, No. 6, pp. 555-588, May, 2001.

   This paper investigates the problem of practical stabilization for linear systems subject to actuator saturation and input additive disturbance. Attention is restricted to systems with two anti-stable modes. For such a system, a family of linear feedback laws is constructed that achieves semi-global practical stabilization on the asymptotically null controllable

region. This is in the sense that, for any set $\mathcal{X}_0$ in the interior of the asymptotically null controllable region, any (arbitrarily small) set $\mathcal{X}_\infty$ containing the origin in its interior, and any (arbitrarily large) bound on the disturbance, there is a feedback law from the family such that any trajectory of the closed-loop system enters and remains in the set $\mathcal{X}_\infty$ in a finite time as long as it starts from the set $\mathcal{X}_0$. In proving the main results, the continuity and monotonicity of the domain of attraction for a class of second-order systems are revealed.

9. T. Hu and Z. Lin, "A complete stability analysis of planar discrete-time linear systems under saturation," *IEEE Transactions on Circuits and Systems - Part I: Fundamental Theory and Applications*, Vol. 48, No. 6, pp. 710-725, June, 2001.

   A complete stability analysis is performed on a planar discrete-time system of the form $x(k+1) = \text{sat}(Ax(k))$, where A is a Schur stable matrix and sat is the saturation function. Necessary and sufficient conditions for the system to be globally asymptotically stable are given. In the process of establishing these conditions, the behaviors of the trajectories are examined in detail.

10. T. Hu, Z. Lin and L. Qiu, "Stabilization of exponentially unstable linear systems with saturating actuators," *IEEE Transactions on Automatic Control*, Vol. 46, No. 6, pp. 973-979, June, 2001.

    We study the problem of stabilizing exponentially unstable linear systems with saturating actuators. The study begins with planar systems with both poles exponentially unstable. For such a system, we show that the boundary of the domain of attraction under a saturated stabilizing linear state feedback is the unique stable limit cycle of its time-reversed system. A saturated linear state feedback is designed that results in a closed-loop system having a domain of attraction that is arbitrarily close to the null controllable region. This design is then utilized to construct state feedback laws for higher order systems with two exponentially unstable poles.

11. T. Hu, Z. Lin and Y. Shamash, "Semi-global stabilization with guaranteed regional performance of linear systems subject to actuator saturation," *Systems & Control Letters*, Vol. 43, No. 3, pp. 203-210, July, 2001.

    For a linear system under a given saturated linear feedback, we propose feedback laws that achieve semi-global stabilization on the null controllable region while preserving the performance of the original feedback law in a fixed region. Here by semi-global stabilization on the null controllable region we mean the design of feedback laws that result in a domain of attraction that includes any *a priori* given compact subset of the null controllable region. Our design guarantees that the region on which the original performance is preserved would not shrink as the domain of attraction is enlarged by appropriately adjusting the feedback laws. Both continuous-time and discrete-time systems will be considered.

12. Z. Lin and T. Hu, "Semi-global stabilization of linear systems subject to output saturation," *Systems & Control Letters*, Vol. 43, No. 3, pp. 211-217, July, 2001.

    It is established that a SISO linear stabilizable and detectable system subject to output saturation can be semi-globally stabilized by linear output feedback if all its *invariant zeros* are in the closed left-half plane, no matter where the open loop poles are. This result complements a recent result that such systems can always be globally stabilized by discontinuous nonlinear feedback laws, and can be viewed as dual to a well-known result: a linear stabilizable and detectable system subject to input saturation can be semi-globally

stabilized by linear output feedback if all its *poles* are in the open left-half plane, no matter where the invariant zeros are.

13. T. Hu and Z. Lin, "Exact characterization of invariant ellipsoids for linear systems with saturating actuators," *IEEE Transactions on Automatic Control*, Vol. 47, No. 1, pp. 164-169, January 2002.

    We present a necessary and sufficient condition for an ellipsoid to be an invariant set of a linear system under a saturated linear feedback. The condition is given in terms of linear matrix inequalities and can be easily used for optimization based analysis and design.

14. T. Hu, Z. Lin and B.M. Chen, "Analysis and design for discrete-time linear systems subject to actuator saturation," *Systems & Control Letters*, Vol. 45, No. 2, pp. 97-112, February 2002.

    We present a method to estimate the domain of attraction for a discrete-time linear system under a saturated linear feedback. A simple condition is derived in terms of an auxiliary feedback matrix for determining if a given ellipsoid is contractive invariant. Moreover, the condition can be expressed as LMIs in terms of all the varying parameters and hence can easily be used for controller synthesis. The following surprising result is revealed for systems with single input: suppose that an ellipsoid is made invariant with a linear feedback, then it is invariant under the saturated linear feedback if and only if it can be made invariant with any saturated (nonlinear) feedback. Finally, the set invariance condition is extended to determine the invariant sets for systems with persistent disturbances. LMI based methods are developed for constructing feedback laws that achieve disturbance rejection with guaranteed stability requirements.

15. T. Hu, Z. Lin and B.M. Chen, "An analysis and design method for linear systems subject to actuator saturation and disturbance," *Automatica*, Vol. 38, No. 2, pp. 351-359, February 2002.

    We present a method for estimating the domain of attraction of the origin for a system under a saturated linear feedback. A simple condition is derived in terms of an auxiliary feedback matrix for determining if a given ellipsoid is contractive invariant. This condition is shown to be less conservative than the existing conditions which are based on the circle criterion or the vertex analysis. Moreover, the condition can be expressed as LMIs in terms of all the varying parameters and hence can easily be used for controller synthesis. This condition is then extended to determine the invariant sets for systems with persistent disturbances. LMI based methods are developed for constructing feedback laws that achieve disturbance rejection with guaranteed stability requirements. The effectiveness of the developed methods are illustrated with examples.

16. Y.-Y. Cao, Z. Lin and T. Hu, "Stability analysis of linear time-delay systems subject to input saturation," *IEEE Transactions on Circuits and Systems - Part I: Fundamental Theory and Applications*, Vol. 49, No. 2, pp. 233-240, February 2002.

    This paper is devoted to stability analysis of linear systems with state delay and input saturation. The domain of attraction resulting from an *a priori* designed state feedback law is analyzed using Lyapunov-Razumikhin and Lyapunov-Krasovskii functional approach. Both delay-independent and delay-dependent estimation of the domain of attraction are presented using the linear matrix inequality technique. The problem of designing linear state feedback laws such that the domain of attraction is enlarged is formulated and solved as an optimization problem with LMI constraints. Numerical examples are used to demonstrate the effectiveness of the proposed design technique.

17. Y.-Y. Cao, Z. Lin and Y. Shamash, "Set invariance analysis and gain-scheduling control for LPV systems subject to actuator saturation," *Systems & Control Letters*, Vol. 46, No. 2, pp. 137-151, May 2002.

    In this paper, a set invariance analysis and gain scheduling control design approach is proposed for the polytopic linear parameter-varying systems subject to actuator saturation. A set invariance condition is first established. By utilizing this set invariance condition, the design of a time-invariant state feedback law is formulated and solved as an optimization problem with LMI constraints. A gain-scheduling controller is then designed to further improve the closed-loop performance. Numerical examples are presented to demonstrate the effectiveness of the proposed analysis and design method.

18. T. Hu, A.N. Pitsillides and Z. Lin, "Null controllabilility and stabilization of linear systems subject to asymmetric actuator saturation," in *Actuator Saturation Control*, eds., V. Kapila and K.M. Grigoriadis, Dekker, pp. 47-75, 2002.

    This paper generalizes our recent results on the null controllable regions and the stabilizability of exponentially unstable linear systems subject to symmetric actuator saturation. The description of the null controllable region carries smoothly from the symmetric case to the asymmetric case. As to stabilization, we have to take a quite different approach since the development of our earlier results for planar anti-stable systems relies mainly on the symmetric property of the vector field and the trajectories. Specifically, in this paper, we construct a Lyapunov function from a closed trajectory to show that this closed trajectory forms the boundary of the domain of attraction for a planar anti-stable system under the control of a saturated linear feedback. If the linear feedback is designed by the LQR method, then there is a unique limit cycle which forms the boundary of the domain of attraction. We further show that if the gain is increased along the direction of the LQR feedback, then the domain of attraction can be made arbitrarily close to the null controllable region. This design is then utilized to construct state feedback laws for higher order systems with two exponentially unstable poles.

19. T. Hu and Z. Lin, "Output regulation of general discrete-time linear systems with saturating actuators," *International Journal of Robust and Nonlinear Control*, to appear.

    This paper studies the classical problem of output regulation for linear discrete-time systems subject to actuator saturation and extends the recent results on continuous-time systems to discrete-time systems. The asymptotically regulatable region, the set of all initial conditions of the plant and the exosystem for which the asymptotic output regulation is possible, is characterized in terms of the null controllable region of the anti-stable subsystem of the plant. Feedback laws are constructed that achieve regulation on the asymptotically regulatable region.

20. T. Hu and Z. Lin, "On semi-global stabilizability of anti-stable systems by saturated linear feedback," *IEEE Transactions on Automatic Control*, to appear.

    It was recently established that a second-order anti-stable linear system can be semi-globally stabilized on its null controllable region by saturated linear feedback and a higher order linear system with two or more anti-stable poles can be semi-globally stabilized on its null controllable region by more general bounded feedback laws. We will show in this paper that a system with three real-valued anti-stable poles cannot be semi-globally stabilized on its null controllable region by the simple saturated linear feedback.

21. T. Hu, Z. Lin and L. Qiu, "An explicit description of null controllable region of linear systems with saturating actuators, *Systems & Control Letters*, to appear.

We give simple exact descriptions of the null controllable regions for general linear systems with saturating actuators. The description is in terms of a set of extremal trajectories of the anti-stable subsystem. For lower order systems or systems with only real eigenvalues, this description is further simplified to result in explicit formulae for the boundaries of the null controllable regions.

22. T. Hu and Z. Lin, "On improving the performance with continuous feedback laws," *IEEE Transactions on Automatic Control*, to appear.

    We present controller design methods to smoothen the discontinuity resulting from a piecewise linear control (PLC) law which was proposed to improve the convergence performance for systems with input constraints. The continuous control laws designed in this paper are explicit functions of the state and are easily implementable. We also show that the convergence performance can be further improved by using a saturated high gain feedback law. The efficiency of the proposed methods is illustrated with the PUMA 560 robot model.

23. T. Hu and Z. Lin, "Compiste quadratic Lyapunov functions for constrained control systems," submitted for publication in *IEEE Transactions on Automatic Control.*

    A Lyapunov function based on a group of quadratic functions is introduced in this paper. We call this Lyapunov function a composite quadratic function. Some important properties of this Lyapunov function are revealed. We show that this function is continuously differentiable and its level set is the convex hull of a group of ellipsoids. These results are used to study the set invariance properties of linear systems with input and state constraints. We show that for a system under a given saturated linear feedback, the convex hull of a group of invariant ellipsoids is also invariant. If each ellipsoid in a group can be made invariant with a bounded control of the saturating actuator, then their convex hull can also be made invariant by the same actuator. For a group of ellipsoids, each invariant under a separate saturated linear feedback, we also present a method for constructing a nonlinear continuous feedback law which makes their convex hull invariant.

24. T. Hu and Z. Lin, "The equivalence of several set invariance conditions under saturation," submitted for publication in *IEEE Transactions on Automatic Control.*

    Several equivalent conditions or statements for set invariance were obtained for systems with one saturating actuator in a recent paper. In particular, it was shown that the invariance of an ellipsoid under a saturated linear feedback is equivalent to its controlled invariance and also to the existence of a feedback linear inside the ellipsoid that makes it invariant. In this paper, we attempt to extend the results to systems with multiple saturating actuators. Our analysis reveals that the equivalence holds conditionally for some pairs of the statements and does not hold for some other pairs.

25. T. Hu and Z. Lin, "On the necessity of a recent set invariance condition under actuator saturation," submitted for publication in *Systems & Control Letters.*

    A sufficient condition for an ellipsoid to be invariant was obtained recently and an LMI approach was developed to find the largest ellipsoid satisfying the condition. This condition was later shown to be necessary for the single input case. This paper is dedicated to the multi-input case. We will examine when this condition is also necessary for multi-input systems. Our investigation is based on studying the optimal solution to a related LMI problem. A criterion is presented to determine when the condition is necessary and when the largest invariant ellipsoid has been obtained by using the LMI method.

26. T. Hu, Z. Lin and Y. Shamash, "On maximizing the convergence rate for linear linear systems with input saturation," submitted for publication in *IEEE Transactions on Automatic Control.*

    In this paper, we consider the problem of maximizing the convergence rate inside a given level set for both continuous-time and discrete-time systems with input saturation. We also provide simple methods for finding the largest ellipsoid of a given shape that can be made invariant with a saturated control. For the continuous-time case, the maximal convergence rate is achieved by a bang-bang type control with a simple switching scheme. Sub-optimal convergence rate can be achieved with saturated high-gain linear feedback. We also study the problem of maximizing the convergence rate in the presence of disturbances. For the discrete-time case, the maximal convergence rate is achieved by a coupled saturated linear feedback.

27. T. Hu and Z. Lin, "Output regulation for general linear systems with saturating actuators," submitted for publication in *Automatica.*

    This paper studies the classical problem of output regulation for linear systems subject to actuator saturation. The asymptotically regulatable region, the set of all initial conditions of the plant and the exosystem for which output regulation is possible, is characterized in terms of the null controllable region of the anti-stable subsystem of the plant. Output regulation laws are constructed from a given stabilizing state feedback law. It is shown that a stabilizing feedback law that achieves a larger domain of attraction leads to a feedback law that achieves output regulation on a larger subset of the asymptotically regulatable region. A feedback law that achieves global stabilization on the asymptotically null controllable region leads to a feedback law that achieves output regulation on the entire asymptotically regulatable region.

28. Y. Xiao, Y.-Y. Cao and Z. Lin, "Robust filtering for discrete-time systems with saturation and its application to transmultiplexers," submitted for publication in *IEEE Transactions on Signal Processing.*

    This paper considers the problem of robust filtering for discrete-time linear systems subject to saturation. A generalized dynamic filter architecture is proposed and a filter design method is developed. Our approach incorporates the conventional linear H_2 and H_∞ filtering as well as a regional l_2 gain filtering feature developed specially for the saturation nonlinearity, and is applicable to the digital transmultiplexer systems for the purpose of separating filter bank design. It turns out that our filter design can be carried out by solving a constrained optimization problem with LMI constraints. Simulation shows that the resultant separating filters possess satisfactory reconstruction performance while working in the linear range, and less degraded reconstruction performance in the presence of saturation.

29. Y.-Y. Cao and Z. Lin, "Robust stability analysis and fuzzy-scheduling control for nonlinear systems subject to actuator saturation," submitted for publication in *IEEE Transactions on Fuzzy Systems.*

    Takagi-Sugeno (TS) fuzzy models can provide an effective representation of complex nonlinear systems in terms of fuzzy sets and fuzzy reasoning applied to a set of linear input-output submodels. In this paper, the TS fuzzy modeling approach is utilized to carry out the stability analysis and control design for nonlinear systems with actuator saturation. The TS fuzzy representation of a nonlinear system subject to actuator saturation is presented. In our TS fuzzy representation, the modeling error is also captured by

norm-bounded uncertainties. A set invariance condition for the system in the TS fuzzy representation is first established. Based on this set invariance condition, the problem of estimating the domain of attraction of a TS fuzzy system under a constant state feedback law is formulated and solved as an LMI optimization problem. By viewing the state feedback gain as an extra free parameter in the LMI optimization problem, we arrive at a method for designing state feedback gain that maximizes the domain of attraction. A fuzzy scheduling control design method is also introduced to further enlarge the domain of attraction. An inverted pendulum is used to show the effectiveness of the proposed fuzzy controller.

3 Future Research Topics

There are several research topics important to current research program. In what follows, we briefly describe three of these research topics.

1. **Intelligent Control of Nonlinear Systems with Actuator Saturation**

   Saturation is probably the most widely encountered and most dangerous nonlinearity in control systems. Actuator saturation is a particularly important consideration in reconfigurable control systems. Following a system failure, it is likely that the available control authority will be reduced, and that the reconfigurable control system will drive actuators into their saturated regime to stabilize the systems and to attempt to regain nominal performance. The destabilizing effects of actuator saturation have been cited as contributing factors in several mishaps involving high performance aircraft.[1] [2] [3] As a result, actuator saturation has been receiving increasing attention from research community. Most of these results however deal with linear systems that are at worst marginally unstable. Results that do deal with unstable systems are often conservative. However, the dynamics of many control systems such as flight control systems are nonlinear and their linearizations are strictly unstable.

   Under the support of this YIP award, we have been studying unstable linear systems with saturating actuators and have obtained several fundamental results.[4] In particular, we have for the first time obtained an explicit characterization of the null controllable region, the set of states that can be driven to the origin using bounded controls delivered by the saturating actuators. We have also constructed feedback laws that actually stabilize the system on an arbitrarily large portion of the null controllable region. These results and the experience gained in obtaining these results prepared us to tackle the more difficult problem of control design for nonlinear systems with actuator saturation.

   Our approach is inspired by the TS fuzzy models as originally proposed by Takagi and Sugeno.[5] The idea is to represent a nonlinear system subject to actuator saturation by a fuzzy blending of a group of (possibly strictly unstable) linear systems subject to actuator saturation. Such a fuzzy blended model takes the form of an linear parameter varying (LPV) systems with the parameter varying within a polytope $\mathcal{P}$. Corresponding to each

[1] J.M. Lenorovitz, "Gripen control problems resolved through in-flight, ground simulations," *Aviation Week Space Technol.*, pp. 74-75, June 18, 1990.

[2] M.A. Dornheim, "Report pinpoints factors leading to YF-22 crash," *Aviation Week Space Technol.*, pp. 53-54, Nov. 9, 1992.

[3] C.A. Shifrin, "Gripen likely to fly again soon," *Aviation Week Space Technol.*, pp. 72, Aug. 23, 1993.

[4] T. Hu and Z. Lin, *Control Systems with Actuator Saturation: Analysis and Design*, Birkäuser, Boston, 2001

[5] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Trans. Sys., Man, Cybern.*, Vol. 15, No. 1, PP. 116-132, 1985.

vertex of the polytope is a linear system subject to actuator saturation. A control law is then designed for each of these linear systems subject to actuator saturation. These "local" control laws are then fuzzy blended to form the overall controller for the original nonlinear systems.

Preliminary work as reported in our Publication 29 of Section 2 has indicated the power of combining the results developed in this project with the tools from intelligent control theory.

2. **Time Delay in Networked Systems**

Time delay is probably one of the most important issues to be considered in networked systems. Although there is a vast literature on control systems with time delays, the analysis and design methods that take into account practical constraints such as actuator saturation are very conservative. In our preliminary work, Publication 16 of Section 2, we have been able to demonstrate how the theory developed under this YIP project can be readily applied to drastically sharpen the analysis and design tools for systems with time delay.

Our work as reported in Publication 16 of Section 2 only deals with stability. The next step is to address closed-loop system performances beyond stability.

3. **Large Scale Systems with Limited Information**

Many large scale systems, such as a group of UCAVs working as a networked system, require the control action be designed and implemented locally, because of the lack of centralized information. Control design becomes more complicated when physical constraints such as time delay and imperfect actuators/sensors have to be taken into account. For example, in the design of local controllers, the limited information available to the subsystems are probably transmitted from other subsystems with some time delay. Many Sensors cannot provide precise information. Sensors might have finite precision when the signal to be sensed is small. Sensors might even only be able to pick up the sign of the signal to be sensed. Analysis and design tools for large scale systems with limited information is an important research topic.

The benefit of understanding the adverse effects of limited information and imperfect sensors and knowing ways of alleviating these effects is two fold. First, we will be able to fully utilize the available information and the precision of sensors without worrying about their adverse effects. On the other hand, by being able to fully utilize the limited information and limited sensor precision, we will be able to use the fewest and least sophisticated/expensive sensors to achieve the desired closed-loop system performance.

4 Appendix: Copies of Publications

Publication 1

Table of Contents[6]
Preface

[6]A copy of the book was earlier presented to Dr. Moshfegh.

CONTROL
ENGINEERING

Tingshu Hu
Zongli Lin

Control Systems with Actuator Saturation

Analysis and Design

Birkhäuser

Contents

Preface xiii

1 Introduction 1
- 1.1 Linear Systems with Actuator Saturation 1
- 1.2 Notation, Acronyms, and Terminology 3

2 Null Controllability – Continuous-Time Systems 11
- 2.1 Introduction 11
- 2.2 Preliminaries and Definitions 12
- 2.3 General Description of Null Controllable Region 15
- 2.4 Systems with Only Real Eigenvalues 21
- 2.5 Systems with Complex Eigenvalues 27
- 2.6 Some Remarks on the Description of $\mathcal{C}(T)$ 33
- 2.7 Asymptotically Null Controllable Region 34
- 2.8 Conclusions 35

3 Null Controllability – Discrete-Time Systems 37
- 3.1 Introduction 37
- 3.2 Preliminaries and Definitions 38
- 3.3 General Description of Null Controllable Region 41
- 3.4 Systems with Only Real Eigenvalues 44
- 3.5 Systems with Complex Eigenvalues 48
- 3.6 An Example 50
- 3.7 Asymptotically Null Controllable Region 51
- 3.8 Conclusions 53

4 Stabilization on Null Controllable Region – Continuous-Time Systems **55**
4.1 Introduction 55
4.2 Domain of Attraction – Planar System under Saturated Linear Feedback 57
4.3 Semi-Global Stabilization – Planar Systems 67
4.4 Semi-Global Stabilization – Higher Order Systems 74
4.5 Conclusions 83

5 Stabilization on Null Controllable Region – Discrete-Time Systems **85**
5.1 Introduction 85
5.2 Global Stabilization at Set of Equilibria – Planar Systems 86
5.3 Global Stabilization – Planar Systems 99
5.4 Semi-Global Stabilization – Planar Systems 105
5.5 Semi-Global Stabilization – Higher Order Systems 108
5.6 Conclusions 111

6 Practical Stabilization on Null Controllable Region **113**
6.1 Introduction 113
6.2 Problem Statement and Main Results 114
6.2.1 Problem Statement 114
6.2.2 Main Results: Semi-Global Practical Stabilization . 114
6.3 Proof of Main Results 115
6.3.1 Properties of the Trajectories of Second Order Linear Systems 115
6.3.2 Properties of the Domain of Attraction 119
6.3.3 Proof of Theorem 6.2.1: Second Order Systems . . . 127
6.3.4 Proof of Theorem 6.2.1: Higher Order Systems . . . 141
6.4 An Example 144
6.5 Conclusions 147
6.A Proof of Lemma 6.3.1 149
6.B Proof of Lemma 6.3.2 153

7 Estimation of the Domain of Attraction under Saturated Linear Feedback **157**
7.1 Introduction 157

7.2 A Measure of Set Size 159
7.3 Some Facts about Convex Hulls 160
7.4 Continuous-Time Systems under State Feedback 163
7.4.1 A Set Invariance Condition Based on Circle Criterion 164
7.4.2 An Improved Condition for Set Invariance 165
7.4.3 The Necessary and Sufficient Condition – Single Input Systems 167
7.4.4 Estimation of the Domain of Attraction 169
7.5 Discrete-Time Systems under State Feedback 173
7.5.1 Condition for Set Invariance 173
7.5.2 The Necessary and Sufficient Condition – Single Input Systems 177
7.5.3 Estimation of the Domain of Attraction 179
7.6 Extension to Output Feedback 180
7.7 Conclusions 181

8 On Enlarging the Domain of Attraction 183
8.1 Introduction 183
8.2 Continuous-Time Systems 183
8.3 Discrete-Time Systems 185
8.4 Conclusions 191

9 Semi-Global Stabilization with Guaranteed Regional Performance 195
9.1 Introduction 195
9.2 Expansion of the Domain of Attraction 197
9.3 Semi-Globalization – Discrete-Time Systems 199
9.4 Semi-Globalization – Continuous-Time Systems 205
9.5 An Example 207
9.6 Conclusions 208

10 Disturbance Rejection with Stability 211
10.1 Introduction 211
10.2 Continuous-Time Systems 213
10.2.1 Problem Statement 213
10.2.2 Condition for Set Invariance 214

10.2.3 Disturbance Rejection with Guaranteed Domain of Attraction 216
10.2.4 An Example 219
10.3 Discrete-Time Systems 221
10.3.1 Problem Statement 221
10.3.2 Condition for Set Invariance 223
10.3.3 Disturbance Rejection with Guaranteed Domain of Attraction 225
10.4 Conclusions 228

11 On Maximizing the Convergence Rate **229**
11.1 Introduction 229
11.2 Continuous-Time Systems 233
11.2.1 Maximal Convergence Control and Maximal Invariant Ellipsoid 233
11.2.2 Saturated High Gain Feedback 242
11.2.3 Overall Convergence Rate 247
11.2.4 Maximal Convergence Control in the Presence of Disturbances 255
11.3 Discrete-Time Systems 258
11.4 Conclusions 264

12 Output Regulation – Continuous-Time Systems **265**
12.1 Introduction 265
12.2 Preliminaries and Problem Statement 267
12.2.1 Review of Linear Output Regulation Theory 267
12.2.2 Output Regulation in the Presence of Actuator Saturation 270
12.3 The Regulatable Region 271
12.4 State Feedback Controllers 279
12.5 Error Feedback Controllers 290
12.6 An Example 297
12.7 Conclusions 301

13 Output Regulation – Discrete-Time Systems **305**
13.1 Introduction 305
13.2 Preliminaries and Problem Statement 306
13.2.1 Review of Linear Output Regulation Theory 306

13.2.2 Output Regulation in the Presence of Actuator Saturation 307
13.3 The Regulatable Region 309
13.4 State Feedback Controllers 315
13.5 Error Feedback Controllers 324
13.6 Conclusions 325

14 Linear Systems with Non-Actuator Saturation **327**
14.1 Introduction 327
14.2 Planar Linear Systems under State Saturation – Continuous-Time Systems 328
14.2.1 System Description and Problem Statement 328
14.2.2 Main Results on Global Asymptotic Stability 328
14.2.3 Outline of the Proof 330
14.3 Planar Linear Systems under State Saturation – Discrete-Time Systems 344
14.3.1 System Description and Problem Statement 344
14.3.2 Main Results on Global Asymptotic Stability 344
14.3.3 Outline of the Proof 347
14.4 Semi-Global Stabilization of Linear Systems Subject to Sensor Saturation 362
14.4.1 Introduction 362
14.4.2 Main Results 363
14.4.3 An Example 370
14.5 Conclusions 371

Bibliography **375**

Index **387**

Preface

Saturation nonlinearities are ubiquitous in engineering systems. In control systems, every physical actuator or sensor is subject to saturation owing to its maximum and minimum limits. A digital filter is subject to saturation if it is implemented in a finite word length format. Saturation nonlinearities are also purposely introduced into engineering systems such as control systems and neural network systems. Regardless of how saturation arises, the analysis and design of a system that contains saturation nonlinearities is an important problem. Not only is this problem theoretically challenging, but it is also practically imperative. This book intends to study control systems with actuator saturation in a systematic way. It will also present some related results on systems with state saturation or sensor saturation.

Roughly speaking, there are two strategies for dealing with actuator saturation. The first strategy is to neglect the saturation in the first stage of the control design process, and then to add some problem-specific schemes to deal with the adverse effects caused by saturation. These schemes, known as anti-windup schemes, are typically introduced using *ad hoc* modifications and extensive simulations. The basic idea behind these schemes is to introduce additional feedbacks in such a way that the actuator stays properly within its limits. Most of these schemes lead to improved performance but poorly understood stability properties.

The second strategy is more systematic. It takes into account the saturation nonlinearities at the outset of the control design. Or, in the case that a control law is designed a priori to meet either the performance or stability requirement, it analyzes the closed-loop system under actuator saturation systematically and redesigns the controller in such a way that

the performance is retained while stability is improved or the other way around. This is the approach we will take in this book. Such an approach to dealing with actuator saturation entails the characterization of the null controllable region, the set of all states that can be driven to the origin by the saturating actuators, and the design of feedback laws that are valid on the entire null controllable region or a large portion of it. More specifically, the results that are to be presented in this book are outlined as follows.

In Chapter 1, after a short introduction to linear systems with saturation nonlinearities, in particular, actuator saturation, we list some notation and acronyms that are used throughout the book. Some technical terms will also be defined here.

Chapters 2 and 3 give explicit descriptions of the null controllable region of a linear system with the bounded controls delivered by the saturating actuators. Chapter 2 deals with continuous-time systems. Chapter 3 deals with discrete-time systems.

Chapters 4 and 5 study the stabilizability at the origin of linear systems with saturating actuators. The main objective is to obtain a domain of attraction that is arbitrarily close to the null controllable region. We refer to such a stabilization problem as semi-global stabilization on the null controllable region. Chapter 4 deals with continuous-time systems. Chapter 5 deals with discrete-time systems.

Chapter 6 considers continuous-time linear systems that are subject to both actuator saturation and input-additive bounded disturbance. The objective is to construct feedback laws that will cause all trajectories starting from within any a priori specified (arbitrarily large) compact subset of the null controllable region to converge to another a priori specified (arbitrarily small) neighborhood of the origin. We refer to such a problem as semi-global practical stabilization on the null controllable region.

Chapter 7 looks at the problem of controlling a linear system with saturating actuators from a different angle. An LMI-based method is developed for estimating the domain of attraction of a linear system under an a priori designed saturated linear feedback law. This analysis method is then utilized in Chapter 8 to arrive at a method for designing linear state feedback laws that would result in the largest estimated domain of attraction. Each of these two chapters treats both continuous-time and discrete-time systems.

Chapter 9 develops a design method for arriving at simple nonlinear feedback laws that achieve semi-global stabilization on the null controllable region and, in the mean time, guarantee regional performance. Both continuous-time and discrete-time systems are considered.

Chapter 10 addresses the problem of controlling linear systems subject to both actuator saturation and disturbance. Unlike in Chapter 6, here the disturbance is not input additive and can enter the system from anywhere. Design problems that capture both large domains of attraction and strong disturbance rejection capability are formulated and solved. Both continuous-time and discrete-time systems are considered.

Chapter 11 examines the problem of maximizing the convergence rate inside a given ellipsoid for both continuous-time and discrete-time systems subject to actuator saturation. Simple methods are also proposed for determining the largest ellipsoid of a given shape that can be made invariant with a saturated control. For continuous-time systems, the maximal convergence rate is achieved by a bang-bang type control with a simple switching scheme. A sub-optimal convergence rate can be achieved with saturated high-gain linear feedback. For discrete-time systems, the maximal convergence rate is achieved by a coupled saturated linear feedback.

Chapters 12 and 13 formulate and solve the classical problem of output regulation for linear systems with saturating actuators. The problem is to design stabilizing feedback laws that, in the presence of disturbances, cause the plant output to track reference signals asymptotically. Both the reference signals and the disturbances are modeled by a reference system, called the exosystem. The asymptotically regulatable region, the set of all initial conditions of the plant and the exosystem for which the output regulation is possible, is characterized. Feedback laws that achieve output regulation on the asymptotically regulatable region are constructed. Chapter 12 deals with continuous-time systems. Chapter 13 deals with discrete-time systems.

Finally, Chapter 14 collects some results on the analysis and design of linear systems subject to saturation other than actuator saturation. This includes sensor saturation and state saturation.

The intended audience of this monograph includes practicing engineers and researchers in areas related to control engineering. An appropriate background for this monograph would be some first year graduate courses

in linear systems and multivariable control. Some background in nonlinear control systems would greatly facilitate the reading of the book.

In such an active current research area as actuator saturation, it is impossible to account for all the available results. Although we have tried our best to relate our work to the available research, we are still frustrated with our inability to do a better job in this regard and will strive to improve in future work.

We would like to thank some of our colleagues who, through collaboration on the topics of this book, motivated and helped us in many ways. They are Professor Ben M. Chen of National University of Singapore, Professor Daniel Miller of University of Waterloo, Professor Li Qiu of Hong Kong University of Science and Technology, and Professor Yacov Shamash of State University of New York at Stony Brook. We would also like to thank our colleague Dr. Yong-Yan Cao for a careful reading of the manuscript.

We are indebted to Professor William S. Levine, the series editor, for his enthusiasm and encouragement of our efforts in completing this book. We are also thankful to the staff at Birkhäuser, in particular, Ms. Louise Farkas, Ms. Shoshanna Grossman, and Ms. Lauren Schultz, for their excellent editorial assistance.

We are grateful to the United States Office of Naval Research's Young Investigator Program for supporting our research that leads to most of the results presented in this book. We are also grateful to the University of Virginia for an environment that allowed us to write this book.

Our special thanks go to our families, {Jianping, Sylvia, (T. H.)} and {Jian, Tony, Vivian, (Z. L.)}. Without their sacrifice, encouragement, and support, this book would not have been completed.

This monograph was typeset by the authors using LaTeX. All simulations and numerical computations were carried out in MATLAB.

Charlottesville, Virginia
November 2000

Tingshu Hu
Zongli Lin

Publication 2

PERGAMON

Automatica 36 (2000) 269-277

www.elsevier.com/locate/automatica

Brief Paper

Finite gain stabilization of discrete-time linear systems subject to actuator saturation☆

Xiangyu Bao[a], Zongli Lin[a,*], Eduardo D. Sontag[b]

[a]Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903, USA
[b]Department of Mathematics, Rutgers University, New Brunswick, NJ 08903, USA

Received 10 August 1998; received in final form 23 April 1999

Abstract

It is shown that, for neutrally stable discrete-time linear systems subject to actuator saturation, finite gain $l_p$sta bilization can be achieved by linear output feedback, for all $p\in(1,\infty]$. An explicit construction of the corresponding feedback laws is given. The feedback laws constructed also result in a closed-loop system that is globally asymptotically stable, and in an input-to-state estimate. © 1999 Elsevier Science Ltd. All rights reserved.

Keywords: Input saturation; Discrete-time linear systems; Finite gain stability; Lyapunov functions

1. Introduction

In this paper, we consider the problem of global stabilization of a discrete-time linear system subject to actuator saturation:

$$\mathscr{P}:\begin{cases}x^+ = Ax + B\sigma(u+u_1), & x\in\mathbb{R}^n,\ u\in\mathbb{R}^m,\\ y = Cx + u_2, & y\in\mathbb{R}^r\end{cases} \tag{1}$$

(we use the notation x^+ to indicate a forward shift, that is, for a function x and an integer t, $x^+(t)$ is $x(t+1)$), where $u_1\in\mathbb{R}^m$ is the actuator disturbance, $u_2\in\mathbb{R}^r$ is the sensor noise, and $\sigma:\mathbb{R}^m\to\mathbb{R}^m$ represents actuator saturation, i.e., $\sigma(s)=[\sigma_1(s_1)\ \sigma_2(s_2)\ \cdots\ \sigma_m(s_m)]$ with $\sigma_i(s_i)=\text{sign}(s_i)\min\{1,|s_i|\}$, and the pair (A,B) is stabilizable. The problem of global asymptotic stabilization (internal stabilization) of this system has recently been solved using nonlinear state feedback laws and under the condition that all the eigenvalues of A are inside or on the unit circle (Yang, Sontag & Sussmann, 1997), and for neutrally stable open-loop system using linear state feedback (Choi, 1999). Here, we are interested not only in closed-loop state space stability (internal stability), but also in stability with respect to both measurement and actuator noises. More specifically, we would like to construct a controller $\mathscr{C}$ so that the operator $(u_1,u_2)\mapsto(y_1,y_2)$ as defined by the following standard systems interconnection (see Fig. 1):

$$y_1 = \mathscr{P}(u_1 + y_2),$$

$$y_2 = \mathscr{C}(u_2 + y_1), \tag{2}$$

is well defined and finite gain stable.

We note that the disturbance u_1 we consider here is input additive and enters the system together with the control input u through the actuators. Simple examples show that the problem we consider does not always have a solution if the disturbance enters the system from outside the actuators.

The above problem was first studied for continuous-time systems. It was shown in Liu, Chitour and Sontag (1996) that, for neutrally stable open-loop systems, linear feedback laws can be used to achieve finite gain stability, with respect to every L_p-norm. For a neutrally stable system, all open loop poles are located in the closed left-half plane, with those on the $j\omega$-axis having Jordan blocks of size one. In the case that full state is available for feedback (i.e., $y_1 = x$ and $u_2 = 0$), it was shown in Lin, Saberi and Teel (1995) that if the external input signal is

☆This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Henk Nijmeijer under the direction of Editor T. Basar.

* Corresponding author. Tel.: + 1-804-924-6342; fax: + 1-804-924-8818.

E-mail address: zl5y@virginia.edu (Z. Lin)

0005-1098/00/$ - see front matter © 1999 Elsevier Science Ltd. All rights reserved.
PII: S0005-1098(99)00138-7

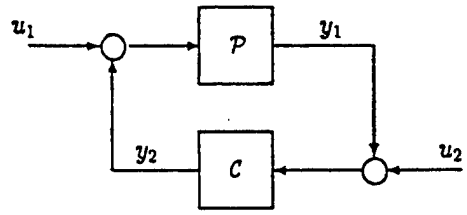


Fig. 1. Standard closed-loop connection.

uniformly bounded, then finite-gain L_p-stabilization and local asymptotic stabilization can always be achieved simultaneously by linear feedback, no matter where the poles of the open-loop system are. The uniform boundedness condition of Lin et al. (1995) was later removed by resorting to nonlinear feedback (Lin, 1997). Some other works related to the topic are Hou, Saberi and Lin (1997a), Chitour, Liu and Sontag (1995), Nguyen and Jabbari (1997), Saberi, Hou and Stoorvogel (1998), Suarez, Alvarez-Ramirez, Sznaier and Ibarra-Valdez (1997) and the references therein.

There are also several studies in the discrete-time setting, showing some of the continuous-time results carry over to discrete-time (for example, Hou, Saberi, Lin & Sannuti, 1997b; Yang et al., 1997) and some do not (for example, Hou et al., 1997a,b). In particular, Hou et al. (1997a,b) show that the results of Lin (1997) and Lin et al. (1995) on finite gain stabilization of continuous-time systems do not carry over to discrete-time systems. The objective of this paper is to show that the results of Liu et al. (1996), however, do carry over to discrete-time systems. More specifically, we show that, for neutrally stable discrete-time linear systems subject to actuator saturation, finite gain l_p stabilization can be achieved by linear output feedback for all $p\in(1,\infty]$. An explicit construction of the corresponding feedback laws is given. The feedback laws constructed also result in a closed-loop system that is globally asymptotically stable, and provide an input-to-state estimate. While many of the arguments used are conceptually similar to those used in the continuous-time case Liu et al. (1996), there are technical aspects that are very different and not totally obvious. For example, unlike in Liu et al. (1996), the feedback gain for the discrete-time case needs to be multiplied by a small factor, say κ, which causes the solution of a certain Lyapunov equation, and the subsequent estimation of the solution, to be dependent on κ (see Lemma 2). As another example, the difficulties in evaluating the difference of the non-quadratic Lyapunov function along the trajectories of the closed-loop system entail a careful estimation by Taylor expansion.

The remainder of the paper is organized as follows. Section 2 states the main results. Section 3 contains the proof of the results that were stated in Section 2. A brief concluding remark is given in Section 4.

2. Preliminary and problem statement

We first recall some notation. For a vector $X\in\mathbb{R}^r$, $|X|$ denotes the Euclidean norm of X, and for a matrix $X\in\mathbb{R}^{m\times n}$, the induced operator norm. For any $p\in[1,\infty)$, we write l_p^n for the set of all sequences $\{x(t)\}_{t=0}^{\infty}$, where $x\in\mathbb{R}^n$, such that $\sum_{t=0}^{\infty}|x(t)|^p<\infty$, and the l_p-norm of $x\in l_p^n$ is defined as $\|x\|_{l_p}=(\sum_{t=0}^{\infty}|x(t)|^p)^{1/p}$. We use l_∞^n to denote the set of all sequences $\{x(t)\}_{t=0}^{\infty}$, where $x\in\mathbb{R}^n$, such that $\sup_t|x(t)|<\infty$, and the l_∞-norm of $x\in l_\infty^n$ is defined as $\|x\|_{l_\infty}=\sup_t|x(t)|$.

The objective of this paper is to show the following result concerning the global asymptotic stabilization as well as l_p-stabilization of system $\mathcal{P}$, as given by (1), using linear output feedback.

Theorem 1. *Consider a system* (1). *Let A be neutrally stable, i.e., all the eigenvalues of A are inside or on the unit circle, with those on the unit circle having all Jordan blocks of size one. Also assume that (A,B) is stabilizable and (A,C) is detectable. Then, there exits a linear observer-based output feedback law of the form*

$$\hat{x}^+ = A\hat{x}+B\sigma(F\hat{x})-L(y-C\hat{x}),$$
$$u=F\hat{x} \tag{3}$$

which has the following properties:

1. *It is finite gain l_p-stable for all $p\in(1,\infty]$, i.e., there exists a $\gamma_p>0$ such that*

$$\|x\|_{l_p}\le\gamma_p[\|u_1\|_{l_p}+\|u_2\|_{l_p}],\quad \forall u_1\in l_p^m,\ u_2\in l_p^r$$
$$\text{and } x(0)=0,\ \hat{x}(0)=0. \tag{4}$$

2. *In the absence of actuator and sensor noises u_1 and u_2, the equilibrium $(x,\hat{x})=(0,0)$ is globally asymptotically stable.*

Remark 1. We will in fact actually obtain the following stronger ISS-like property (see Sontag, 1998 and references there in):

$$\|(x,\hat{x})\|_{l_p}\le\theta_p(|x(0)|+|\hat{x}(0)|)+\gamma_p[\|u_1\|_{l_p}+\|u_2\|_{l_p}], \tag{5}$$

where θ_p is a class-$\mathcal{K}$ function. Observe that the single estimate (5) encompasses both the gain estimate (4) and asymptotic stability. Obviously, (4) is the special case of (5) for zero initial states. On the other hand, when applied with arbitrary initial states but $u_1=u_2=0$, there follows that $(x,\hat{x})$ is in l_p, which implies, in particular, that $(x(t),\hat{x}(t))$ must converge to zero as $t\to\infty$ (global attraction) and that $|(x(t),\hat{x}(t))|$ is bounded by $\theta_p(|x(0)|+|\hat{x}(0)|)$ (stability).

3. Proof of Theorem 1

The proof of Theorem 1 will follow readily from the following proposition, which we establish first.

Proposition 1. *Let A be orthogonal (i.e., $A'A = I$),an d suppose that the pair (A, B) is controllable. Then,the system*

$$x^+ = Ax + B\sigma(-\kappa B'Ax + u), \quad x\in\mathbb{R}^n,\ u\in\mathbb{R}^m \tag{6}$$

is finite gain l_p-stable, $p\in(1,\infty]$, for sufficiently small $\kappa > 0$. Moreover, for each $p\in(1,\infty]$ there exist a real γ_p, a $\kappa^*\in(0,1]$, and a class-$\mathcal{K}$ function θ_p such that, for all $\kappa\in(0,\kappa^*]$,

$$\|x\|_{l_p} \le \gamma_p\|u\|_{l_p} + \theta_p(|x(0)|) \tag{7}$$

for all inputs $u\in l_p^m$ and all initial states $x(0)$.

To prove this proposition, we need to establish a few lemmas.

Lemma 1. *For any $p > l > 0$, there exist two scalars $M_1, M_2 > 0$ such that,for any two positive scalars ξ and ζ,*

$$\xi^{p-l}\zeta^l \le M_1\xi^p + M_2\zeta^p \tag{8}$$

and consequently, for any $n > 0$ and $\kappa > 0$,

$$\xi^{p-l}\zeta^l \le M_1\kappa^n\xi^p + \kappa^{n(l-p)/l}M_2\zeta^p. \tag{9}$$

Proof of Lemma 1. Let $h:\mathbb{R}^+\to\mathbb{R}^+$ be defined as $h(x) = x^{l/(p-l)}$, which is continuous and strictly increasing with $h(0) = 0$ and $h(\infty) = \infty$, and $k(x) = x^{(p-l)/l}$ be its pointwise inverse. Define

$$H(x) = \int_0^x h(v)\,dv = \frac{p-l}{p}x^{p/(p-l)} \tag{10}$$

and

$$K(x) = \int_0^x k(v)\,dv = \frac{l}{p}x^{p/l}. \tag{11}$$

Letting $a = \xi^{p-l}$ and $b = \zeta^l$, it follows from Young's inequality (Hardy, Littlewood & Polya, 1952), $ab \le H(a) + K(b)$ for all $a, b\in\mathbb{R}^+$, that

$$\xi^{p-l}\zeta^l \le \frac{p-l}{p}\xi^p + \frac{l}{p}\zeta^p = M_1\xi^p + M_2\zeta^p, \tag{12}$$

which also trivially implies (9). □

Lemma 2. *Let A and B be as given in Proposition* 1. *Then, for any $\kappa > 0$ such that $\kappa B'B < 2I$, $\bar{A}(\kappa) = A - \kappa BB'A$ is asymptotically stable. Moreover, let $P(\kappa)$ be the unique positive-definite solution to the Lyapunov equation,*

$$\bar{A}(\kappa)'P\bar{A}(\kappa) - P = -I. \tag{13}$$

Then, there exists a $\kappa^* > 0$ such that

$$\frac{\chi_1}{\kappa}I \le P(\kappa) \le \frac{\chi_2}{\kappa}I, \quad \forall\kappa\in(0,\kappa^*] \tag{14}$$

for some positive constants χ_1 and χ_2 independent of κ.

Proof of Lemma 2. The asymptotic stability of $\bar{A}$ follows from a simple Lyapunov/LaSalle argument (Choi, 1999). Let $\kappa_1^* > 0$ be such that $\kappa B'B < 2I$ for all $\kappa\in(0,\kappa_1^*]$. We recall that the solution to the Lyapunov equation (13) is given by

$$\begin{aligned} P(\kappa) &= \sum_{k=0}^{\infty}(\bar{A}^k(\kappa))'\bar{A}^k(\kappa) \\ &= \sum_{k=0}^{\infty}[(A-\kappa BB'A)']^k[(A-\kappa BB'A)]^k. \end{aligned} \tag{15}$$

Using the fact that $AA' = I$, we have

$$\begin{aligned} &(A-\kappa BB'A)'(A-\kappa BB'A) \\ &\quad = I - 2\kappa A'BB'A + \kappa^2A'BB'BB'A \\ &\quad = I - \kappa A'B(2I-\kappa B'B)B'A. \end{aligned} \tag{16}$$

Using now the fact that $\kappa B'B < 2I$ for $\kappa\in(0,\kappa_1^*]$, we know that there exists $\kappa_2^*\in(0,\kappa_1^*]$ such that

$$\tfrac{1}{2}I \le (A-\kappa BB'A)'(A-\kappa BB'A) \le I, \quad \forall\kappa\in(0,\kappa_2^*]. \tag{17}$$

Again using the fact that $A'A = I$, we verify in a straightforward way that

$$(A-\kappa BB'A)^n = A^n - \kappa C_{A,B}C'_{A,B}A^n + \kappa^2M_1(\kappa), \tag{18}$$

where $M_1(k)$ is a polynomial matrix in κ of order $n-2$, n being the order of the system (6), and

$$C_{A,B} = [B \quad AB \quad \cdots \quad A^{n-1}B]$$

is the controllability matrix of the pair (A, B) and is of full rank. It then follows that

$$\begin{aligned} &((A-\kappa BB'A)^n)'(A-\kappa BB'A)^n \\ &\quad = ((A^n)' - \kappa(A^n)'C_{A,B}C'_{A,B} + \kappa^2M_1'(\kappa)) \\ &\quad\quad (A^n - \kappa C_{A,B}C'_{A,B}A^n + \kappa^2M_1(\kappa)) \\ &\quad = I - 2\kappa(A^n)'C_{A,B}C'_{A,B}A^n + \kappa^2M_2(\kappa), \end{aligned} \tag{19}$$

where $M_2(\kappa)$ is a symmetric polynomial matrix in κ of order $2n-2$. Since $C_{A,B}$ is of full rank, and because A is nonsingular, there exists a $\kappa^*\in(0,\kappa_2^*]$ such that

$$\begin{aligned} 0 \le I - \kappa M_1^0 I &\le ((A-\kappa BB'A)^n)'(A-\kappa BB'A)^n \\ &\le I - \kappa M_2^0 I < I, \quad \forall\kappa\in(0,\kappa^*] \end{aligned} \tag{20}$$

for some constants $M_1^0, M_2^0 > 0$ independent of κ.

Using (17), (20) and the fact that $A'A = I$ in (15), we have that for all $\kappa \in (0, \kappa^*]$,

$$P(\kappa) \le \sum_{i=0}^{n-1} [(A-\kappa BB'A)^i]'(A-\kappa BB'A)^i \sum_{k=0}^{\infty} (1-\kappa M_2^0)^k I \le n\frac{1}{\kappa M_2^0} I = \frac{\chi_2}{\kappa} I \tag{21}$$

and

$$P(\kappa) \ge \sum_{i=0}^{n-1} [(A-\kappa BB'A)^i]'(A-\kappa BB'A)^i \sum_{k=0}^{\infty} (1-\kappa M_1^0)^k I \ge \left(\frac{1}{2}\right)^{n-1} n\frac{1}{\kappa M_1^0} I = \frac{\chi_1}{\kappa} I, \tag{22}$$

where $\chi_1 = n/2^{n-1}M_1^0$ and $\chi_2 = n/M_2^0$. □

Lemma 3. *Let $\bar{A}(\kappa)$ be as given in Proposition 1, $P(\kappa)$ as defined in Lemma 2, then for any $p \in (1, \infty)$, there exists a $\kappa^* > 0$ such that*

$$[x'\bar{A}'(\kappa)P(\kappa)\bar{A}(\kappa)x]^{p/2} - [x'P(\kappa)x]^{p/2} \le -\kappa^{(2-p)/2}\zeta|x|^p, \quad \kappa \in (0, \kappa^*], \tag{23}$$

where $\zeta > 0$ is some constant independent of κ.

Proof of Lemma 3. Inequality (23) holds trivially for $x = 0$. Hence, in what follows, we assume, without loss of generality, that $x \neq 0$.

For simplicity, we introduce from now the following notation:

$$\mu = x'\bar{A}'(\kappa)P(\kappa)\bar{A}(\kappa)x \tag{24}$$

(where x and κ will be clear from the context). By the definition of $P(\kappa)$, we have

$$\mu - x'P(\kappa)x = -x'x. \tag{25}$$

From Lemma 2, there exists a $\kappa_1^* > 0$ such that for all $\kappa \in (0, \kappa_1^*]$

$$\left|\frac{x'x}{x'P(\kappa)x}\right| \le \frac{4}{5}, \quad \forall x \neq 0. \tag{26}$$

With (25) and (26), we can continue the proof using Taylor expansion with remainder,

$$\begin{aligned}
&[x'\bar{A}'(\kappa)P(\kappa)\bar{A}(\kappa)x]^{p/2} - [x'P(\kappa)x]^{p/2} \\
&= [x'P(\kappa)x - x'x]^{p/2} - [x'P(\kappa)x]^{p/2} \\
&\le [x'P(\kappa)x]^{p/2}\left[1 - \frac{p}{2}\frac{x'x}{x'P(\kappa)x} + \delta\left(\frac{x'x}{x'P(\kappa)x}\right)^2\right] - [x'P(\kappa)x]^{p/2} \\
&= -\frac{p}{2}[x'P(\kappa)x]^{(p-2)/2}|x|^2 + \delta[x'P(\kappa)x]^{(p-4)/2}|x|^4, \\
&\kappa \in (0, \kappa_1^*],
\end{aligned} \tag{27}$$

where $\delta = \max_{|z| \le 4/5}\{\frac{p}{8}|(p-2)(1+z)^{p/2-2}|\}$ is a constant independent of κ.

Again by Lemma 2, there exists a $\kappa^* \in (0, \kappa_1^*]$ such that

$$[\mu]^{p/2} - [x'P(\kappa)x]^{p/2} \le -\kappa^{(2-p)/2}\zeta|x|^p, \quad \kappa \in (0, \kappa^*] \tag{28}$$

for some $\zeta > 0$ independent of κ. □

Lemma 4. *Let A and B be as given in Proposition 1. For any $l \in [1, \infty)$ and any $\kappa \in (0, 1]$,*

$$|\sigma(-\kappa B'Ax + u)|^l \le 2^{l-1}\kappa^l|B|^l|x|^l + 2^{l-1}|u|^l. \tag{29}$$

Proof of Lemma 4. Since σ is a standard saturation function and $|A| = 1$, for any $l \ge 1$, we have

$$\begin{aligned}
|\sigma(-\kappa B'Ax + u)|^l &\le (\kappa|B||x| + |u|)^l \\
&\le 2^{l-1}\kappa^l|B|^l|x|^l + 2^{l-1}|u|^l,
\end{aligned} \tag{30}$$

where the last inequality follows from Jensen's inequality applied to the convex function s^l:

$$(a+b)^l \le \tfrac{1}{2}(2a)^l + \tfrac{1}{2}(2b)^l, \quad \forall a, b \ge 0. \qquad \square$$

Lemma 5. *Let A and B be as given in Proposition 1. Pick any $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$, any number $\eta \ge 3$, and any non-negative real number l. Denote $\tilde{x} = -\kappa B'Ax + u$. Then, provided $|x| > \eta|B\sigma(\tilde{x})|$, we have*

$$|Ax + B\sigma(\tilde{x})|^l \le |x|^l + l|x|^{l-2}x'A'B\sigma(\tilde{x}) + M|x|^{l-2}|B\sigma(\tilde{x})|^2 \tag{31}$$

for some constant $M > 0$ which is independent of κ.

Proof of Lemma 5. We first note that, since $|x| > \eta|B\sigma(\tilde{x})| \ge 3|B\sigma(\tilde{x})|$,

$$\frac{|2x'A'B\sigma(\tilde{x}) + |B\sigma(\tilde{x})|^2|}{|x|^2} \le \frac{4}{5}. \tag{32}$$

Hence, using Taylor expansion with remainder, we have

$$\begin{aligned}
|Ax + B\sigma(\tilde{x})|^l &= [|x|^2 + 2x'A'B\sigma(\tilde{x}) + |B\sigma(\tilde{x})|^2]^{l/2} \\
&= |x|^l\left(1 + \frac{2x'A'B\sigma(\tilde{x}) + |B\sigma(\tilde{x})|^2}{|x|^2}\right)^{l/2} \\
&\le |x|^l\left[1 + \frac{l}{2}\frac{2x'A'B\sigma(\tilde{x}) + |B\sigma(\tilde{x})|^2}{|x|^2} + \delta\left(\frac{2x'A'B\sigma(\tilde{x}) + |B\sigma(\tilde{x})|^2}{|x|^2}\right)^2\right] \\
&\le |x|^l + l|x|^{l-2}x'A'B\sigma(\tilde{x}) + \frac{l}{2}|x|^{l-2}|B\sigma(\tilde{x})|^2 + \delta|x|^{l-4}((2+1/\eta)|x||B\sigma(\tilde{x})|)^2,
\end{aligned} \tag{33}$$

where $\delta = \max_{|z| \le 4/5}\{\frac{1}{8}l|(l-2)(1+z)^{l/2-2}|\}$ is a constant independent of κ.

So we can see that the inequality (31) holds for $M = \frac{1}{2} + \delta(2 + 1/\eta)^2$. □

We are now ready to prove Proposition 1.

Proof of Proposition 1. We separate the proof for $p \in (1, \infty)$ and for $p = \infty$.

Proof for $p \in (1, \infty)$: For clarity, let us repeat here the system equation (6)

$$x^+ = Ax + B\sigma(-\kappa B'Ax + u),\ x \in \mathbb{R}^n,\ u \in \mathbb{R}^m. \tag{34}$$

This may also be rewritten as

$$x^+ = \bar{A}(\kappa)x + B(-\tilde{x} + \sigma(\tilde{x}) + u),\quad x \in \mathbb{R}^n,\ u \in \mathbb{R}^m, \tag{35}$$

where $\bar{A}(\kappa) = A - \kappa BB'A$, $\tilde{x} = -\kappa B'Ax + u$.

For this system, define the function V_1 as

$$V_1(x) = (x'P(\kappa)x)^{p/2}, \tag{36}$$

where $P(\kappa)$ is as given in Lemma 2. We next evaluate the increments $V(x^+(t)) - V(x(t))$, which we denote as "ΔV_1" for short, along any given trajectory of (35). It is convenient to treat separately the cases $|x| > \eta|B\sigma(\tilde{x})|$ and $|x| \le \eta|B\sigma(\tilde{x})|$. Here $\eta \ge 3$ is a number to be specified soon.

Case 1: $|x| > \eta|B\sigma(\tilde{x})|$. Using the definition of V_1, we now give an upper bound on ΔV_1 along the trajectories of the system (35). To simplify the equations, we introduce the following notation:

$$v = 2\kappa x'A'P(\kappa)BB'Ax + 2x'A'P(\kappa)B\sigma(\tilde{x}) + \sigma'(\tilde{x})B'P(\kappa)B\sigma(\tilde{x}),$$

in addition to μ as defined in Eq. (24). Thus,

$$\begin{aligned}
\Delta V_1 &= V_1^+ - V_1 \\
&= [(x^+)'P(\kappa)x^+]^{p/2} - [x'P(\kappa)x]^{p/2} \\
&= [[\bar{A}(\kappa)x + \kappa BB'Ax + B\sigma(\tilde{x})]'P(\kappa)[\bar{A}(\kappa)x \\
&\quad + \kappa BB'Ax + B\sigma(\tilde{x})]]^{p/2} - [x'P(\kappa)x]^{p/2} \\
&= [\mu + 2x'A'P(\kappa)B(\kappa B'Ax + \sigma(\tilde{x})) \\
&\quad - \kappa^2 x'A'BB'P(\kappa)BB'Ax + \sigma'(\tilde{x})B'P(\kappa)B\sigma(\tilde{x})]^{p/2} \\
&\quad - [x'P(\kappa)x]^{p/2} \\
&\le [\mu + v]^{p/2} - [x'P(\kappa)x]^{p/2} \\
&= [\mu]^{p/2}\left[1 + \frac{v}{\mu}\right]^{p/2} - [x'P(\kappa)x]^{p/2}.
\end{aligned} \tag{37}$$

By Lemma 2, there exist a $\kappa_1^* > 0$ and $\eta \ge 3$ independent of κ, such that for all $|x| > \eta|B\sigma(\tilde{x})|$

$$\left|\frac{v}{\mu}\right| \le \frac{4}{5},\quad \kappa \in (0, \kappa_1^*]. \tag{38}$$

To see this, let $\kappa_0^* > 0$ be such that (14) of Lemma 2 and (17) in the proof of Lemma 2 both hold for all $\kappa \in (0, \kappa_0^*]$. Then, for all $\kappa \in (0, \kappa_0^*]$, we have

$$|v| \le \left[2\chi_2|B|^2 + \frac{2\chi_2}{\kappa\eta} + \frac{\chi_2}{\kappa\eta^2}\right]|x|^2 \tag{39}$$

and

$$|\mu| \ge \frac{\chi_1}{2\kappa}|x|^2, \tag{40}$$

from which it is clear that there exist κ_1^* and $\eta > 3$ such that (38) holds.

Next, we may use a Taylor expansion with remainder to continue the bounding of ΔV_1 as follows:

$$\Delta V_1 \le [\mu]^{p/2}\left[1 + \frac{p}{2}\frac{v}{\mu} + \delta\left[\frac{v}{\mu}\right]^2\right] - [x'P(\kappa)x]^{p/2}, \tag{41}$$

where $\delta = \max_{|z| \le 4/5}\{\frac{1}{8}p|(p-2)(1+z)^{p/2-2}|\}$ is a constant independent of κ.

By Lemmas 3 and 2, there exists $\kappa_2^* \in (0, \kappa_1^*]$ such that for any $\kappa \in (0, \kappa_2^*]$, we have

$$\begin{aligned}
\Delta V_1 &\le -\kappa^{(2-p)/2}\zeta|x|^p + \frac{p}{2}[\mu]^{(p-2)/2}[v] + \delta[\mu]^{(p-4)/2}[v]^2 \\
&\le -\kappa^{(2-p)/2}\zeta|x|^p + \psi_1\kappa^{(2-p)/2}|x|^{p-2} \\
&\quad \times [2|x||P(\kappa)B||\tilde{x} - \sigma(\tilde{x})| + 2|x||P(\kappa)B||u| \\
&\quad + |P(\kappa)||B\sigma(\tilde{x})|^2] + \psi_2\kappa^{(4-p)/2}|x|^{p-4} \\
&\quad \times [2|x|^2|\kappa P(\kappa)||B|^2 + 2|P(\kappa)||x||B\sigma(\tilde{x})| \\
&\quad + |P(\kappa)||B\sigma(\tilde{x})|^2]^2,
\end{aligned} \tag{42}$$

where $\zeta > 0$ is as defined in Lemma 3, and $\psi_1, \psi_2 > 0$ are some constants independent of κ.

Before continuing, we digress to observe that

$$|\tilde{x} - \sigma(\tilde{x})| \le \tilde{x}'\sigma(\tilde{x}). \tag{43}$$

Using (43), Lemmas 1 and 4, and the condition $|x| > \eta|B\sigma(\tilde{x})|$, we can show that there exists a $\kappa_3^* \in (0, \kappa_2^*]$ such that for all $\kappa \in (0, \kappa_3^*]$ the estimation of ΔV_1 can be now concluded as follows:

$$\begin{aligned}
\Delta V_1 \le\ & -\kappa^{(2-p)/2}\zeta|x|^p + 2\psi_1\kappa^{(2-p)/2}|x|^{p-1}|P(\kappa)B|\tilde{x}'\sigma(\tilde{x}) \\
& + M_{1a}\kappa^{(2-p)/2}\max\{\kappa, \kappa^{p-1}\}|x|^p + M_{2a}(\kappa)|u|^p,
\end{aligned} \tag{44}$$

where $M_{1a} > 0$, $M_{2a}(\kappa) > 0$ with M_{1a} independent of κ are defined in an obvious way. In deriving (44), we have also used the fact that $|x|^{p-2} < (|B\sigma(\tilde{x})|/\eta)^{p-2}$ for $p < 2$ and $B\sigma(\tilde{x}) \ne 0$.

Case 2: $|x| \le \eta|B\sigma(\tilde{x})|$. By using Lemmas 2–4, ΔV_1 along the trajectories of (35) is bounded as follows:

$$\begin{aligned}\Delta V_1 &= [(x^+)'P(\kappa)x^+]^{p/2} - [x'P(\kappa)x]^{p/2}\\ &\le |P(\kappa)|^{p/2}|Ax + B\sigma(\tilde{x})|^p - [x'P(\kappa)x]^{p/2} + [\mu]^{p/2}\\ &\le -\kappa^{(2-p)/2}\zeta|x|^p + \kappa^{-p/2}\chi_2^{p/2}(|x| + |B\sigma(\tilde{x})|)^p\\ &\le -\kappa^{(2-p)/2}\zeta|x|^p + (\eta+1)^p\kappa^{-p/2}\chi_2^{p/2}|B\sigma(\tilde{x})|^p\\ &\le -\kappa^{(2-p)/2}\zeta|x|^p + M_{1b}\kappa^{-(p-2)/2}\kappa^{p-1}|x|^p\\ &\quad + M_{2b}(\kappa)|u|^p, \quad \kappa\in(0,\kappa_3^*], \end{aligned} \tag{45}$$

where $\chi_2 > 0$ and $\zeta > 0$ are as defined in Lemmas 2 and 3, respectively, and $M_{1b} > 0$, $M_{2b}(\kappa) > 0$ are constants with M_{1b} being independent of κ.

Summarizing, we may combine Case 1 with Case 2, to obtain

$$\Delta V_1 \le \begin{cases} -\kappa^{(2-p)/2}\zeta|x|^p + 2\psi_1\kappa^{(2-p)/2}\ |P(\kappa)B||x|^{p-1}\tilde{x}'\sigma(\tilde{x})\\ \quad + M_1\kappa^{(2-p)/2}\max\{\kappa,\kappa^{p-1}\}\ |x|^p\\ \quad + M_2(\kappa)|u|^p, \quad \text{if } |x| > \eta|B\sigma(\tilde{x})|,\\ \kappa^{(2-p)/2}\zeta|x|^p + M_1\kappa^{(2-p)/2}\ \kappa^{p-1}|x|^p + M_2(\kappa)|u|^p\\ \text{if } |x| \le \eta|B\sigma(\tilde{x})|, \end{cases} \tag{46}$$

where

$$M_1 = \max\{M_{1a}, M_{1b}\}$$

and

$$M_2(\kappa) = \max\{M_{2a}(\kappa), M_{2b}(\kappa)\}.$$

For system (34), we next define another function

$$V_0(x) = |x|^{p+1}. \tag{47}$$

An estimation of its increments along the trajectories of (34) can also be carried out by separately considering each of the cases $|x| > \eta|B\sigma(\tilde{x})|$ and $|x| \le \eta|B\sigma(\tilde{x})|$.

Case 1: $|x| \le \eta|B\sigma(\tilde{x})|$. By Lemma 4, for any $\kappa\in(0,\kappa_3^*]$,

$$\begin{aligned}\Delta V_0 &= |Ax + B\sigma(\tilde{x})|^{p+1} - |x|^{p+1} \le |Ax + B\sigma(\tilde{x})|^{p+1}\\ &\le (|x| + |B\sigma(\tilde{x})|)^{p+1} \le ((\eta+1)|B\sigma(\tilde{x})|)^{p+1}\\ &\le \kappa N_{1a}|x|^p + N_{2a}|u|^p \end{aligned} \tag{48}$$

for some positive constants N_{1a} and N_{2a} independent of κ. In deriving (48), we have used the fact that both σ and κ are bounded.

Case 2: $|x| > \eta|B\sigma(\tilde{x})|$. By Lemmas 5, 4 and 1, there exists $\kappa_4^*\in(0,\kappa_3^*]$ such that for any $\kappa\in(0,\kappa_4^*]$,

$$\begin{aligned}\Delta V_0 &= |Ax + B\sigma(\tilde{x})|^{p+1} - |x|^{p+1},\\ &\le |x|^{p+1} + (p+1)|x|^{p-1}x'A'B\sigma(\tilde{x})\\ &\quad + N_{1b}|B\sigma(\tilde{x})|^2|x|^{p-1} - |x|^{p+1},\\ &\le -\frac{p+1}{\kappa}|x|^{p-1}\tilde{x}^T\sigma(\tilde{x}) + \kappa N_{1c}|x|^p + N_{2c}(\kappa)|u|^p, \end{aligned} \tag{49}$$

where $N_{1c}, N_{1b} > 0$, $N_{2c}(\kappa) > 0$ are constants, and N_{1b}, N_{1c} are independent of κ. In deriving (49), the first inequality by Lemma 5, the second inequality is the consequence of the fact that σ is bounded and Lemmas 4 and 1.

Combining Case 1 with Case 2, we have, for any $\kappa\in(0,\kappa_4^*]$,

$$\Delta V_0 \le \begin{cases} -\frac{p+1}{\kappa}|x|^{p-1}\tilde{x}'\sigma(\tilde{x})\\ \quad + \kappa N_1|x|^p + N_2(\kappa)|u|^p, & \text{if } |x| > \eta|B\sigma(\tilde{x})|,\\ \kappa N_1|x|^p + N_2(\kappa)|u|^p, & \text{if } |x| \le \eta|B\sigma(\tilde{x})|, \end{cases} \tag{50}$$

where

$$N_1 = \max\{N_{1a}, N_{1c}\}$$

and

$$N_2(\kappa) = \max\{N_{2a}, N_{2c}(\kappa)\}.$$

Finally, we define the following Lyapunov (or "storage") function:

$$V(x) = V_1(x) + \varpi V_0(x), \tag{51}$$

where

$$\varpi = \frac{2}{p+1}\kappa^{(4-p)/2}\psi_1|P(\kappa)B|.$$

It is straightforward to verify that there exists some $\kappa^*\in(0,\kappa_4^*]$ such that

$$\Delta V(x) \le -\kappa^{(2-p)/2}\alpha|x|^p + \beta(\kappa)|u|^p, \quad \forall\kappa\in(0,\kappa^*] \tag{52}$$

for some $\alpha\in(0,\zeta)$ and $\beta(\kappa) > 0$.

Now consider an arbitrary initial state $x(0)$ and control u, and the ensuing trajectory x. Summing both sides of (52) from $t = 0$ to ∞ and using the fact that V is non-negative, we conclude that

$$\kappa^{(2-p)/2}\alpha\|x\|_{l_p}^p \le \beta(\kappa)\|u\|_{l_p}^p + \theta_{p0}(|x(0)|), \tag{53}$$

where $\theta_{p0}(r) = \varpi r^{p+1} + (\chi_2 r^2/\kappa)^{p/2}$. This implies that

$$\|x\|_{l_p} \le \gamma_p\|u\|_{l_p} + \theta_p(|x(0)|), \tag{54}$$

where

$$\gamma_p = (\kappa^{(p-2)/2}\beta(\kappa)/\alpha)^{1/p},$$

and

$$\theta_p(r) = (\kappa^{(p-2)/2}\theta_{p0}(r)/\alpha)^{1/p}.$$

Proof for $p = \infty$. From (52) we get for $p = 2$,

$$\Delta V(x) \leq -\alpha|x|^2 + \beta(\kappa)\|u\|_{l_\infty}^2. \tag{55}$$

Hence, $\Delta V(x)$ is negative outside the ball of radius $(\beta(\kappa)/\alpha)^{1/2}\|u\|_{l_\infty}$ centered at the origin, from which it follows that, for any state $x(t)$ in the trajectory:

$$V(x(t)) \leq \left(\frac{\varpi\beta^{3/2}(\kappa)}{\alpha^{3/2}}\|u\|_{l_\infty} + \frac{\chi_2\beta(\kappa)}{\alpha\kappa}\right)\|u\|_{l_\infty}^2 + \theta_{\infty 0}(|x(0)|), \tag{56}$$

where $\theta_{\infty 0}(r) = (\chi_2/\kappa)r^2 + \varpi r^3$. If $\|u\|_{l_\infty} \leq 1$, we have

$$\frac{\chi_1}{\kappa}|x(t)|^2 \leq x(t)'P(\kappa)x(t) \leq V(x(t)) \tag{57}$$

and

$$V(x(t)) \leq \left(\frac{\varpi\beta^{3/2}(\kappa)}{\alpha^{3/2}} + \frac{\chi_2\beta(\kappa)}{\alpha\kappa}\right)\|u\|_{l_\infty}^2 + \theta_{\infty 0}(|x(0)|) \tag{58}$$

which implies the following estimate for the entire trajectory:

$$\|x\|_{l_\infty} \leq \left\{\frac{\kappa\varpi\beta^{3/2}(\kappa)}{\alpha^{3/2}\chi_1} + \frac{\chi_2\beta(\kappa)}{\alpha\chi_1}\right\}^{1/2}\|u\|_{l_\infty} + \theta_{\infty 1}(|x(0)|), \tag{59}$$

where $\theta_{\infty 1}(r) = (\kappa\theta_{\infty 0}(r)/\chi_1)^{1/2}$. If, instead, $\|u\|_{l_\infty} > 1$, we have

$$\varpi|x|^3 \leq V(x) \leq \left(\frac{\varpi\beta^{3/2}(\kappa)}{\alpha^{3/2}} + \frac{\chi_2\beta(\kappa)}{\alpha\kappa}\right)\|u\|_{l_\infty}^3 + \theta_{\infty 1}(|x(0)|), \tag{60}$$

from which we get that

$$\|x\|_{l_\infty} \leq \left(\frac{\beta^{3/2}(\kappa)}{\alpha^{3/2}\varpi} + \frac{\chi_2\beta(\kappa)}{\kappa\alpha\varpi}\right)^{1/3}\|u\|_{l_\infty} + \theta_{\infty 2}(|x(0)|), \tag{61}$$

where $\theta_{\infty 2}(r) = (\theta_{\infty 0}(r)/\varpi)^{1/3}$. Letting

$$\gamma_\infty = \max\left\{\left\{\frac{\kappa\varpi\beta^{3/2}(\kappa)}{\alpha^{3/2}\chi_1} + \frac{\chi_2\beta(\kappa)}{\alpha\chi_1}\right\}^{\frac{1}{2}}, \left\{\frac{\beta^{3/2}(\kappa)}{\alpha^{3/2}\varpi} + \frac{\chi_2\beta(\kappa)}{\alpha\varpi\kappa}\right\}^{1/3}\right\}$$

and $\theta_\infty = \max\{\theta_{\infty 1}, \theta_{\infty 2}\}$, we have, finally, the required conclusion:

$$\|x\|_{l_\infty} \leq \gamma_\infty\|u\|_{l_\infty} + \theta_\infty(|x(0)|) \tag{62}$$

for $p = \infty$ as well.

We are now ready to prove Theorem 1. □

Proof of Theorem 1. Without loss of generality, making a change of coordinates if required, we may assume that the system (1) has the following partitioned form:

$$\begin{aligned} x_1^+ &= A_1x_1 + B_1\sigma(u + u_1),\\ x_0^+ &= A_0x_0 + B_0\sigma(u + u_1),\\ y &= Cx + u_2. \end{aligned} \tag{63}$$

where A_1 is orthogonal and A_0 is asymptotically stable, and

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_0 \end{bmatrix}, \qquad B = \begin{bmatrix} B_1 \\ B_0 \end{bmatrix}.$$

We construct the output feedback law in the form of (3) with $F = [-\kappa B_1'A_1 \;\; 0]$, the matrix L being chosen such that $A + LC$ is asymptotically stable. Using this feedback, the closed-loop system is

$$\begin{cases} x_1^+ = A_1x_1 + B_1\sigma(-\kappa B_1'A_1\hat{x}_1 + u_1),\\ x_0^+ = A_0x_0 + B_0\sigma(-\kappa B_1'A_1\hat{x}_1 + u_1),\\ \hat{x}^+ = A\hat{x} + B\sigma(-\kappa B_1'A_1\hat{x}_1) - L(Cx - C\hat{x} + u_2). \end{cases} \tag{64}$$

Let $e = [e_1' \;\; e_0']'$, where $e_1 = x_1 - \hat{x}_1$ and $e_0 = x_0 - \hat{x}_0$. Here we have partitioned $\hat{x} = [\hat{x}_1' \;\; \hat{x}_0']'$ accordingly. In the new states (x, e), (64) can be written as follows,

$$\begin{aligned} x_1^+ &= A_1x_1 + B_1\sigma(-\kappa B'_1A_1x_1 + \kappa B'_1A_1e_1 + u_1),\\ x_0^+ &= A_0x_0 + B_0\sigma(-\kappa B'_1A_1x_1 + \kappa B'_1A_1e_1 + u_1),\\ e^+ &= (A + LC)e + B[\sigma(-\kappa B'_1A_1x_1 + \kappa B'_1A_1e_1 + u_1) - \sigma(-\kappa B'_1A_1x_1 + \kappa B'_1A_1e_1)] + Lu_2. \end{aligned} \tag{65}$$

Since σ is global Lipschitz with a Lipschitz constant 1,

$$|\sigma(-\kappa B_1'A_1x_1 + \kappa B_1'A_1e_1 + u_1) - \sigma(-\kappa B_1'A_1x_1 + \kappa B_1'A_1e_1)| \leq |u_1|. \tag{66}$$

Noting that $A + LC$ is asymptotically stable and viewing

$$\sigma(-\kappa B'_1A_1x_1 + \kappa B'_1A_1e_1 + u_1) - \sigma(-\kappa B'_1A_1x_1 + \kappa B'_1A_1e_1) + Lu_2$$

as an l_p input to the e-subsystem, we have that, for some constant $\gamma_{pe} > 0$,

$$\|e\|_{l_p} \leq \gamma_{pe}(\|u_1\|_{l_p} + \|u_2\|_{l_p} + |e(0)|). \tag{67}$$

Next, applying Proposition 1 to the x_1-subsystem, and viewing $\kappa B'_1A_1e_1 + u_1$ as an l_p input to this subsystem, we have,

$$\|x_1\|_{l_p} \leq \gamma_{p1}(\|u_1\|_{l_p} + \|u_2\|_{l_p} + |e(0)|) + \theta_{p1}(|x_1(0)|)$$

for some $\gamma_{p1} > 0$ and θ_{p1} of class $\mathcal{K}$.

On the other hand, viewing $\sigma(-\kappa B'_1 A_1 x_1 + \kappa B'_1 A_1 e_1 + u_1)$ as an l_p input to the x_0-subsystem, we have the estimate

$$\|x_0\|_{l_p} \le \gamma_{p0}(\|x_1\|_{l_p} + \|e\|_{l_p} + \|u_1\|_{l_p} + |x_0(0)|)$$

for some $\gamma_{p0} > 0$.

In conclusion, we have

$$\|x\|_{l_p} \le \|x_1\|_{l_p} + \|x_0\|_{l_p}$$
$$\le \gamma_p(\|u_1\|_{l_p} + \|u_2\|_{l_p}) + \varphi_p(|e(0)| + |x(0)|), \quad (68)$$

where $\gamma_p > 0$ is some constant and φ_p is a suitable class-$\mathcal{K}$ function. Together with (67), and changing back to the original coordinates, we also conclude that an estimate like the one in (5) holds. □

4. Conclusions

In this paper, we have established that a discrete-time, neutrally stable, stabilizable, and detectable linear system, when subject to actuator saturation, is finite gain l_p stabilizable by linear feedback, for any $p \in (1, \infty]$. A linear output feedback law which simultaneously achieves l_p stabilization and global asymptotic stabilization was constructed.

Acknowledgements

The work of Xiangyu Bao and Zongli Lin was supported in part by the US Office of Naval Research grant N00014-99-1-0670. The work of Eduardo D. Sontag was supported in part by US Air Force Grant F49620-97-1-0159.

References

Choi, J. (1999). On the stabilization of linear discrete-time systems subject to input saturation. *Systems & Control Letters, 36*, 241–244.

Hardy, G. H., Littlewood, J. E., & Polya, G. (1952). *Inequalities*. Cambridge: Cambridge University Press.

Hou, P., Saberi, A., & Lin, Z. (1997a). On l_p-stabilization of strictly unstable discrete-time linear systems with saturating actuators. *Proceedings of the CDC* (pp. 4510–4515).

Hou, P., Saberi, A., Lin, Z., & Sannuti, P. (1997b). Simultaneous external and internal stabilization for continuous and discrete-time critically unstable linear systems with saturating actuators. *Proceedings of* 1997 *ACC* (pp. 1292–1296).

Lin, Z. (1997). H_∞-almost disturbance decoupling with internal stability for linear systems subject to input saturation. *IEEE Transactions on Automatic Control, 42*, 992–995.

Lin, Z., Saberi, A., & Teel, A. (1995). Simultaneously L_p — stabilization and internal stabilization of linear system subject to input saturation — state feedback case. *Systems & Control Letters, 25*, 219–226.

Liu, W., Chitour, Y., & Sontag, E. (1996). On finite gain stabilizability of linear systems subject to input saturation. *SIAM J. Control and Optimization, 34*, 1190–1219.

Chitour, Y., Liu, W., & Sontag, E. (1995). On the continuity and incremental-gain properties of certain saturated linear feedback loops. *International Journal of Robust and Nonlinear Control, 5*, 413–440.

Nguyen, T., & Jabbari, F. (1997). Output feedback controllers for disturbance attenuation with bounded control. *Proceedings of the 36th CDC* (pp. 177–182).

Saberi, A., Hou, P., & Stoorvogel, A. (1998). On simultaneous global external and global internal stabilization of critically unstable linear systems with saturating actuators. *Proceedings of 1998 ACC* (pp. 1463–1467).

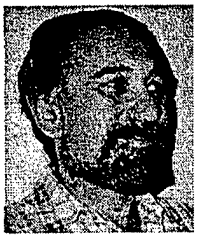
Sontag, E. D. (1998). Comments on integral variants of ISS. *Systems & Control Letters, 34*, 93–100.

Suarez, R., Alvarez-Ramirez, J., Sznaier, M., & Ibarra-Valdez, C. (1997). L_2-disturbance attenuation for linear systems with bounded controls: an ARE-Based Approach. In S. Tarbouriech, and G. Garcia, *Control of uncertain systems with bounded inputs* (pp. 25–38). Lecture notes in control and information sciences, vol. 227. Berlin: Springer.

Yang, Y., Sontag, E. D., & Sussmann, H. J. (1997). Global stabilization of linear discrete-time systems with bounded feedback. *Systems & Control Letters, 30*, 273–281.

Xiangyu Bao received the B.S. degree in automatic control from Tsinghua University, Beijing, China, and the M.S. degree in control engineering from Beijing Institute of Control Engineering, Chinese Academy of Space Technology, Beijing, China, in 1993 and 1996, respectively. She is currently working toward her Ph.D. degree in electrical engineering at the University of Virginia, Charlottesville. Her research interests include nonlinear control theory and control of systems with saturating actuators.

Zongli Lin was born in Fuqing, Fujian, China on February 24, 1964. He received his B.S. degree in mathematics and computer science from Amoy University, Xiamen, China, in 1983, his Master of Engineering degree in automatic control from Chinese Academy of Space Technology, Beijing, China, in 1989, and his Ph.D. degree in electrical and computer engineering from Washington State University, Pullman, Washington, in May 1994. From July 1983 to July 1986, Dr. Lin worked as a control engineer at Chinese Academy of Space Technology. In January 1994, he joined the Department of Applied Mathematics and Statistics, State University of New York at Stony Brook as a visiting assistant professor, where he became an assistant professor in September 1994. Since July 1997, he has been an assistant professor in electrical engineering at University of Virginia. His current research interests include nonlinear control, robust control, and control of systems with saturating actuators. In these areas he has published several papers. He is also the author of the recent book, *Low Gain Feedback* (Springer-Verlag, London, 1998). Dr. Lin currently serves as an associate editor on the Conference Editorial Board of the IEEE Control Systems Society. He is the recipient of an ONR Young Investigator Award.



Eduardo Sontag's major current research interests lie in several areas of control theory and biologically-inspired mathematics (such as neural networks). He received his Licenciado degree from the Mathematics Department at the University of Buenos Aires in 1972, and his Ph.D. (Mathematics) under Rudolf E. Kalman at the Center for Mathematical Systems Theory, University of Florida, in 1976. Since 1977, Dr. Sontag has been with the Department of Mathematics at Rutgers, The State University of New Jersey, where he is currently Professor II of Mathematics as well as a Member of the Graduate Faculties of the Department of Computer Science and of the Department of Electrical and Computer Engineering. He is also the director of SYCON, the Rutgers Center for Systems and Control. Sontag has authored over two hundred and fifty journal and conference papers and book chapters in the above areas, as well as the books Topics in Artificial Intelligence (in Spanish, Buenos Aires: Prolam, 1972), Polynomial Response Maps (Berlin: Springer, 1979), and Mathematical Control Theory: Deterministic Finite Dimensional Systems (Texts in Applied Mathematics, Volume 6, Second Edition, New York: Springer, 1998). Among recent or upcoming major presentations are: a 75-min lecture at the AMS Short Course on Nonlinear Control (AMS Winter Meeting, San Antonio, Jan. 1999), a course on stabilization problems for nonlinear systems at the NATO Advanced Study Institute "Nonlinear Analysis, Differential Equations, and Control" (Montreal, August 1998), a keynote address to the British Applied Mathematics Colloquium'99 (Bath, April 1999), an hour talk at the Symposium on the Mathematical Theory of Networks and Systems (Padova, July 1998), a course on learning theory as part of the 1997 Cambridge University's Isaac Newton Institute Neural Networks and Machine Learning Programme, a 45-minute invited lecture at the 1994 International Congress of Mathematicians, a plenary talk at the 1995 SIAM Control Conference, a course on neurocontrol as part of the 1993 European Control Conference, an expository survey talk at the SPIE's 1997 International Symposium on Aerospace/Defense Sensing and Controls Conference, an hour talk at the AMS Summer Research Institute "Differential Geometry and Control" (Boulder, July 1997), two plenary talks at the 1995 Neural Information Processing Systems Conference, and plenaries at the 1992 IFAC Conference on Nonlinear Control and the 1993 Jerusalem Conference on Control Theory and Applications. Sontag is an Associate Editor for various journals, including: IEEE Transactions in Automatic Control, Control-Theory and Advanced Technology, SMAI-COCOV, Dynamics and Control, Journal of Computer and Systems Sciences, Neurocomputing, and Neural Computing Surveys (Board of Advisors), and a former Associate Editor for Neural Networks and for Systems and Control Letters. In addition, he is a co-founder and co-Managing Editor of the Springer journal MCSS (Mathematics of Control, Signals, and Systems). Sontag is an IEEE Fellow, and has been Program Director and Vice-Chair of the Activity Group in Control and Systems Theory of SIAM. He has been a member of several committees at SIAM and the AMS, and is a former Chair of the Committee on Human Rights of Mathematicians of the latter. He is listed in "Who's Who in Frontier Science and Technology" and in "American Men and Women of Science".

Publication 3

INTERNATIONAL JOURNAL OF ROBUST AND NONLINEAR CONTROL
Int. J. Robust Nonlinear Control 2000; 10:117–135

An improvement to the low gain design for discrete-time linear systems in the presence of actuator saturation nonlinearity

Zongli Lin[1], Ali Saberi[2], Anton A. Stoorvogel[3,*,†] and Ravi Mantri[2]

[1]*Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903-2442, U.S.A.*
[2]*School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164-2752, U.S.A.*
[3]*Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, Netherlands*

SUMMARY

A low-gain design for linear discrete-time systems subject to input saturation was recently developed to solve both semi-global stabilization and semi-global output regulation problems. This paper proposes an improvement to the low-gain design and determines controllers with the new design that achieve semi-global output regulation. The improvement is reflected in better utilization of available control capacity and consequently better closed-loop performance. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: actuator saturation; input saturation; output regulation; low-gain feedback

1. INTRODUCTION

Recently, there has been a surge of interest in the study of linear systems subject to input saturation due to a wide recognition of the inherent constraints on the control actuator. Although most of the results in this study pertain to the problem of global stabilization (see, for example, References [1–5]) and semi-global stabilization (see, for example, Reference [6–12]), some attempts have also been made in the solution of output regulation problems for continuous-time systems. Roughly speaking, this problem is one of controlling a linear system subject to input saturation in order to have its output track (or reject) a family of reference (or disturbance) signals

* Correspondence to: Anton A. Stoorvogel, Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, Netherlands.
† E-mail: A.A.Stoorvogel@tue.nl.

Contract/grant sponsor: US Office of Naval Research Young Investigator Program; contract/grant number: N00014-99-1-0670

Received 25 March 1999
Revised 28 August 1999
Accepted 23 September 1999

CCC 1049-8923/2000/030117–19$17.50
Copyright © 2000 John Wiley & Sons, Ltd.

generated by some external system, usually called the exosystem. The control laws that solve the output regulation problems are referred to as regulators. A global output regulation problem results if tracking and disturbance rejection is required to occur for all initial conditions of the closed-loop system, while a semi-global output regulation problem results if tracking and disturbance rejection is only required to occur when the initial conditions of the closed-loop system are inside an *a priori* given (arbitrarily large) bounded set of the state space. The global output regulation problem and semi-global output regulation problem for linear continuous-time systems subject to input saturation were recently studied in References [13–15]. More recently, for the discrete-time systems, the semi-global output regulation problem was also studied in detail in Reference [16]. As was concluded for both the continuous-time case [15] and the discrete-time case [16], although the global output regulation is appealing by definition, the semi-global output regulation is achievable for a much larger class of systems and allows for linear feedbacks.

This paper represents a continued effort of [16] on the study of semi-global output regulation problem for linear discrete-time systems subject to input saturation. In Reference [16] a set of solvability conditions are given and linear feedback laws which solve the semi-global output regulation problem are constructed. These feedback laws are constructed in such a way that for any *a priori* given bounded set of initial conditions the control signals will not saturate after a finite time, which, for large sets of initial conditions, entails very low feedback gains. As a result, because of linearity, whenever the state is close to the origin, the control input will be far away from its maximum allowable value and thus the closed-loop system will be operating far from its full capacity. To avoid such a situation from happening, in this paper, we construct new feedback laws, which also solve the semi-global output regulation problems under the same solvability conditions. These new feedback laws are constructed based on a new design technique to be developed in the paper. The new design reflects fuller utilization of the available control capacity and consequently better closed-loop performance. We should point out that the fundamental significance of the new design technique introduced here is due to the fact that it is of a semi-global nature. A design conceptually similar to ours was introduced in the interesting paper [17] for local stabilization.

The remainder of this paper is organized as follows. In Section 2 we recall the problem formulation and the solvability conditions from [16]. The new design for state feedback output regulators is given in Section 3, while its error feedback counterpart is given in Section 4. In Section 5, we show that the generalized semi-global output regulation problems we formulated in References [15, 16], which allow an external driving signal to the exosystem, can also be solved by regulators based on the new design. Finally, we draw a brief conclusion in Section 6.

We will mostly use standard notation in this paper. However, we have denoted the shift operator by a superscript $^+$, i.e. $x^+(k) = x(k+1)$. Moreover, for notational brevity, we will suppress the time index k. For a vector $q = [q_1, q_2, \cdots, q_n]'$ we define

$$|q|_\infty = \max_i |q_i|$$

On the other hand, for a vector-valued function w and $K \geqslant 0$ we define

$$\|w\|_\infty = \sup_k |w(k)|_\infty, \qquad \|w\|_{\infty,K} = \sup_{k \geqslant K} |w(k)|_\infty$$

Finally, $\|\cdot\|$ denotes the standard Euclidean norm.

Copyright © 2000 John Wiley & Sons, Ltd.

2. PRELIMINARIES AND PROBLEM STATEMENT

We consider the output regulation problems when the plant inputs are subject to saturation. More specifically, we consider a multivariable system with inputs that are subject to saturation together with an exosystem that generates disturbance and reference signals as described by the following system:

$$\begin{aligned} x^+ &= Ax + B\sigma(u) + Pw \\ w^+ &= Sw \\ e &= Cx + Qw \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^n$, $w \in \mathbb{R}^s$, $u \in \mathbb{R}^m$, $e \in \mathbb{R}^p$, and σ is a vector-valued saturation function defined as

$$\sigma(s) = [\sigma_1(s_1), \sigma_2(s_2), \ldots, \sigma_m(s_m)]' \tag{2}$$

with

$$\sigma_i(s_i) = \begin{cases} s_i & \text{if } |s_i| \leqslant 1 \\ -1 & \text{if } s_i < -1 \\ 1 & \text{if } s_i > 1 \end{cases} \tag{3}$$

Also, without loss of generality, we assume that matrix B is of full rank.

Remark

Unlike in the semi-global stabilization problem (see e.g. [15]), each component of the saturation function σ in the semi-global output regulation problem needs to be exactly known at a neighbourhood of the origin for possible disturbance rejection. Without loss of generality, we have assumed that this neighbourhood has a radius of 1 and, to ensure the existence of linear feedback regulators, we have also assumed that σ_i is linear in this neighbourhood.

Following [18] for linear systems in the absence of input saturation, the semi-global linear feedback output regulation problem and semi-global linear observer-based error feedback output regulation problem for linear systems subject to input saturation were first formulated in References [15, 16] as follows.

Problem 2.1

Consider system (1) and a compact set $\mathcal{W}_0 \subset \mathbb{R}^s$. *The semi-global state feedback regulator problem* is defined as follows:

For any *a priori* given (arbitrarily large) bounded set $\mathcal{X}_0 \subset \mathcal{R}^n$, find, if possible, a static feedback law $u = \alpha(x, w)$ such that

(i) The equilibrium $x = 0$ of

$$x^+ = Ax + B\sigma(\alpha(x, 0)) \tag{4}$$

is asymptotically stable with $\mathcal{X}_0$ contained in its basin of attraction;

(ii) For all $x(0) \in \mathcal{X}_0$ and $w(0) \in \mathcal{W}_0$, the solution of the closed-loop system satisfies

$$\lim_{k \to \infty} e(k) = 0 \tag{5}$$

Copyright © 2000 John Wiley & Sons, Ltd.

Problem 2.2

Consider system (1) and a compact set $\mathscr{W}_0 \subset \mathbb{R}^s$. *The semi-global linear observer-based error feedback regulator problem* is defined as follows:

For any *a priori* given (arbitrarily large) bounded sets $\mathscr{X}_0 \subset \mathbb{R}^n$ and $\mathscr{Z}_0 \subset \mathbb{R}^{n+s}$, find, if possible, an error feedback law of the form

$$\begin{cases} \begin{pmatrix} \hat{x}^+ \\ \hat{w}^+ \end{pmatrix} = \begin{pmatrix} A & P \\ 0 & S \end{pmatrix} \begin{pmatrix} \hat{x} \\ \hat{w} \end{pmatrix} + \begin{pmatrix} B \\ 0 \end{pmatrix} \sigma(u) + \begin{pmatrix} L_A \\ L_S \end{pmatrix} \left[e - (C \quad Q) \begin{pmatrix} \hat{x} \\ \hat{w} \end{pmatrix} \right] \\ u = \alpha(\hat{x}, \hat{w}) \end{cases} \tag{6}$$

such that

(i) The equilibrium $(x, \hat{x}, \hat{w}) = (0, 0, 0)$ of

$$\begin{cases} x^+ = Ax + B\sigma(\alpha(\hat{x}, \hat{w})) \\ \begin{pmatrix} \hat{x}^+ \\ \hat{w}^+ \end{pmatrix} = \begin{pmatrix} A & P \\ 0 & S \end{pmatrix} \begin{pmatrix} \hat{x} \\ \hat{w} \end{pmatrix} + \begin{pmatrix} B \\ 0 \end{pmatrix} \sigma(\alpha(\hat{x}, \hat{w})) + \begin{pmatrix} L_A \\ L_S \end{pmatrix} (C \quad Q) \begin{pmatrix} x - \hat{x} \\ -\hat{w} \end{pmatrix} \end{cases} \tag{7}$$

is asymptotically stable with $\mathscr{X}_0 \times \mathscr{Z}_0$ contained in its basin of attraction.

(ii) For all $(x(0), \hat{x}(0), \hat{w}(0)) \in \mathscr{X}_0 \times \mathscr{Z}_0$ and $w(0) \in \mathscr{W}_0$, the solution of the closed-loop system satisfies

$$\lim_{k \to \infty} e(k) = 0. \tag{8}$$

The solvability conditions for these problems were also established in Reference [16] and are recalled as follows.

Theorem 2.3

Consider system (1) and the given compact set $\mathscr{W}_0 \subset \mathbb{R}^s$. The semi-global state feedback regulator problem is solvable if

(i) (A, B) is stabilizable and A has all eigenvalues inside or on the unit circle;

(ii) There exist matrices Π and Γ such that:

(a) They solve the following linear matrix equations:

$$\Pi S = A\Pi + B\Gamma + P$$

$$0 = C\Pi + Q \tag{9}$$

(b) There exist a $\delta > 0$ and a $K \geqslant 0$ such that $\|\Gamma w\|_{\infty, K} \leqslant 1 - \delta$ for all w with $w(0) \in \mathscr{W}_0$.

Theorem 2.4

Consider system (1) and the given compact set $\mathscr{W}_0 \subset \mathbb{R}^s$. The semi-global linear observer-based error feedback regulator problem is solvable if

Copyright © 2000 John Wiley & Sons, Ltd.

(i) (A, B) is stabilizable and A has all eigenvalues inside or on the unit circle; Moreover, the pair

$$\left[(C \quad Q), \begin{pmatrix} A & P \\ 0 & S \end{pmatrix}\right]$$

is detectable;

(ii) There exist matrices Π and Γ such that:

(a) They solve the following linear matrix equations:

$$\begin{aligned} \Pi S &= A\Pi + B\Gamma + P \\ 0 &= C\Pi + Q \end{aligned} \tag{10}$$

(b) There exists a $\delta > 0$ and a $K \geqslant 0$ such that $\|\Gamma w\|_{\infty,K} \leqslant 1 - \delta$ for all w with $w(0) \in \mathcal{W}_0$.

Linear low-gain feedback-based regulators that solve the above problems were also explicitly constructed in Reference [16]. The necessity of these solvability conditions was also discussed in Reference [16].

As explained in the introduction, this low-gain-based design results in under-utilization of the available control capacity. The goal of this paper is to provide a new design methodology which incorporates significant improvement to the low-gain design technique as developed in References [8, 10] and leads to fuller utilization of the available control capacity and hence better closed-loop performance. We will also show that this new design methodology is also applicable to the so-called generalized output regulation problems we formulated earlier in References [15, 16].

3. AN IMPROVED DESIGN FOR THE STATE FEEDBACK REGULATOR

In this section we construct a family of nonlinear state feedback laws, parameterized in ε and ρ, and then show that such a family of state feedback laws solves the semi-global output regulation problem. Significant improvement on the closed-loop performance over the earlier design [16] is then shown by an example.

This family of nonlinear feedback laws reduces to a linear one for the single input case ($m = 1$), that is, the function $\alpha(x, w)$ as in Problems 2.1 and 2.2 reduces to the form of $Fx + Gw$.

The new state feedback regulator design is carried out in the following two steps.

Step 1 (Solution of an algebraic Riccati equation):

We start by choosing a continuous function H: $(0, 1) \to \mathbb{R}^{n \times n}$ such that $H(\varepsilon)$ is positive definite for each $\varepsilon \in (0, 1]$ and

$$\lim_{\varepsilon \to 0} H(\varepsilon) = 0 \tag{11}$$

While a simple choice of $H(\varepsilon)$ is $H(\varepsilon) = \varepsilon I$, the choice of $H(\varepsilon)$ does significantly affect the closed-loop performance. We leave the issue of judicious choice of $H(\varepsilon)$ for future investigation.

$$X = A'XA + H(\varepsilon) - A'XB(B'XB + I)^{-1}B'XA, \quad \varepsilon \in (0, 1] \tag{12}$$

Copyright © 2000 John Wiley & Sons, Ltd.

We now recall the following lemmas regarding the properties of the Riccati equation (12) from Lin *et al.* [10].

Lemma 3.1

Assume that (A, B) is stabilizable and all the eigenvalues of A are located inside or on the unit circle. Then, for any $\varepsilon \in (0, 1]$, there exists a unique matrix $X(\varepsilon) > 0$ which solves the algebraic Riccati equation (12) and is such that $A - B(B'X(\varepsilon)B + I)^{-1}B'X(\varepsilon)A$ is Schur stable. Moreover,

$$\lim_{\varepsilon \to 0} X(\varepsilon) = 0 \tag{13}$$

Lemma 3.2

Assume that (A, B) is stabilizable and all the eigenvalues of A are located inside or on the unit circle. Then, there exists an $\varepsilon^* \in (0, 1]$ such that, for $\varepsilon \in (0, \varepsilon^*]$,

$$\|X^{1/2}(\varepsilon)AX^{-1/2}(\varepsilon)\| \leqslant \sqrt{2} \tag{14}$$

Step 2 (Composition of state feedback laws):

$$u = -[F(\varepsilon) + \rho\kappa(x, w)K(\varepsilon)](x - \Pi w) + \Gamma w, \quad \rho \in [0, 2] \tag{15}$$

where $F(\varepsilon) = (B'X(\varepsilon)B + I)^{-1}B'X(\varepsilon)A$, $K(\varepsilon) = (B'X(\varepsilon)B)^{-1}B'X(\varepsilon)A_c$, $A_c = A - BF(\varepsilon)$, with $X(\varepsilon)$ being the solution of the Riccati equation (12), and κ: $\mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}_+$ is defined as

$$\kappa(x, w) = \max_{z \in [0, 1]} \{z : |-[F(\varepsilon) + \rho zK(\varepsilon)](x - \Pi w) + \Gamma w|_\infty \leqslant 1\} \tag{16}$$

If, in the above maximization, there exists no z for which the inequality is satisfied then z is chosen equal to 0. Moreover, for the case of single input (i.e. $m = 1$), we can choose

$$\kappa(x, w) \equiv 1 \tag{17}$$

Remark

We note that when $\rho = 0$ the new state feedback laws as given in (15) reduce to the low-gain-based linear state feedback laws as given in Reference [16] and is ε referred to as a low-gain parameter. Moreover, it will become clear shortly that the choice of $K(\varepsilon)$ guarantees that, for any $\rho \in [0, 2]$, the family of state feedback laws (15) also solves the semi-global linear state feedback output regulation problem. A non-zero value of ρ represents fuller utilization of the actuator capacity. While the use of the freedom in the choice of ρ needs to be further explored for achieving control objectives beyond semi-global output regulation, we will demonstrate its contribution to the improvement of the closed-loop transient performance in the output regulation problems.

We then have the following results.

Theorem 3.3

Consider system (1) and the given compact set $\mathcal{W}_0 \subset \mathbb{R}^s$. Assume the sufficient conditions of Theorem 2.3 are satisfied. Then the family of state feedback laws as given in (15) solves the semi-global state feedback regulator problem.

Copyright © 2000 John Wiley & Sons, Ltd.

More specifically, for any a priori given (arbitrarily large) bounded set $\mathscr{X}_0 \subset \mathbb{R}^n$, there exists an $\varepsilon^* \in (0, 1]$ such that for each $\varepsilon \in (0, \varepsilon^*]$ and for each $\rho \in [0, 2]$,

(i) The equilibrium $x = 0$ of

$$x^+ = Ax + B\sigma(-(F(\varepsilon) + \rho\kappa(x, 0)\, K(\varepsilon))\, x) \tag{18}$$

is locally exponentially stable with $\mathscr{X}_0$ contained in its basin of attraction.

(ii) For all $x(0) \in \mathscr{X}_0$ and $w(0) \in \mathscr{W}_0$, the solution of the closed-loop system satisfies

$$\lim_{k \to \infty} e(k) = 0 \tag{19}$$

Proof. We prove this theorem by showing that for each given bounded set $\mathscr{X}_0$, there exists an $\varepsilon^* \in (0, 1]$ such that for all $\varepsilon \in (0, \varepsilon^*)$ and all $\rho \in [0, 2]$, both items (i) and (ii) of the theorem hold.

We first show that there exists an $\varepsilon_1^* \in (0, 1]$ such that for each $\varepsilon \in (0, \varepsilon_1^*)$ and each $\rho \in [0, 2]$, (i) holds. To this end, rewrite (18) as

$$x^+ = A_c x + B[\sigma(-(F(\varepsilon) + \rho\kappa(x, 0)\, K(\varepsilon))x) + F(\varepsilon)x] \tag{20}$$

Consider the Lyapunov function

$$V_1(x) = x'X(\varepsilon)x \tag{21}$$

and let $c_1 > 0$ be such that

$$c_1 \geqslant \sup_{x \in \mathscr{X}_0,\, \varepsilon \in (0,1)} x'X(\varepsilon)x \tag{22}$$

Such a c_1 exists since $\lim_{\varepsilon \to 0} X(\varepsilon) = 0$ by Lemma 3.1 and $\mathscr{X}_0$ is bounded. We note here that such a c_1 guarantees that $\mathscr{X}_0 \subset L_{V_1}(c_1)$, $\forall \varepsilon \in (0, 1]$, where the level set $L_{V_1}(c_1)$ is defined as $L_{V_1}(c_1) = \{x \in \mathbb{R}^n\colon V_1(x) \leqslant c_1\}$. Let ε_1^* be such that for all $\varepsilon \in (0, \varepsilon_1^*]$, $x \in L_{V_1}(c_1)$ implies that $|F(\varepsilon)\, x|_\infty \leqslant 1$. Such an ε_1^* exists because of Lemma 3.2 and the fact that $\lim_{\varepsilon \to 0} X(\varepsilon) = 0$. Note also that from (12) it follows

$$A_c' X(\varepsilon) A_c - X(\varepsilon) = -H(\varepsilon) - F'(\varepsilon) F(\varepsilon) \tag{23}$$

By the definition of κ, (16), for all $x \in L_{V_1}(c_1)$, the saturation functions in the closed-loop system (20) operate in their respective linear regions, and hence the control input remains unsaturated. The evaluation of the deference of V_1 along the trajectories of this linear closed-loop system shows that, for $x \in L_{V_1}(c_1)$,

$$\begin{aligned} V_1(x^+) - V_1(x) &= -x'H(\varepsilon)\, x - x'F'(\varepsilon)\, F(\varepsilon)x \\ &\quad - \rho\kappa(x, 0)\, [2 - \rho\kappa(x, 0)]x'A_c'XB(B'XB)^{-1}B'XA_c x \\ &\leqslant -x'H(\varepsilon)\, x \end{aligned} \tag{24}$$

Now consider the choice of $\kappa(x, w) \equiv 1$ for the single input case, the evaluation of the difference of V_1 along the trajectories of (18) inside the set $L_{V_1}(c_1)$ gives

$$V_1(x^+) - V_1(x) = -x'H(\varepsilon)x - x'F'(\varepsilon)\, F(\varepsilon)x + \phi_1(\gamma_1)$$

Copyright © 2000 John Wiley & Sons, Ltd.

where $\phi_1(\gamma_1) = 2x'A_c'X(\varepsilon)\,B\gamma_1 + \gamma_1 B'X(\varepsilon)\,B\gamma_1$ and

$$\gamma_1 = \sigma(-(F(\varepsilon) + \rho(B'X(\varepsilon)B)^{-1}B'X(\varepsilon)\,A_c)\,x) + F(\varepsilon)x$$

Denoting $\mu_1 = -F(\varepsilon)\,x$ and $v_1 = -K(\varepsilon)\,x$, $\phi_1(\gamma_1)$ can be written as

$$\begin{aligned}\phi_1(\gamma_1) &= -2v_1(B'X(\varepsilon)B)\,[\sigma(\mu_1 + \rho v_1) - \mu_1] \\ &\quad + [\sigma(\mu_1 + \rho v_1) - \mu_1](B'X(\varepsilon)B)\,[\sigma(\mu_1 + \rho v_1) - \mu_1] \\ &= [\sigma(\mu_1 + \rho v_1) - (\mu_1 + 2v_1)](B'X(\varepsilon)B)\,[\sigma(\mu_1 + \rho v_1) - \mu_1]\end{aligned}$$

Noting that $\rho \in [0, 2]$, the definition of σ and for all $x \in L_{V_1}(c_1)$, $|\mu_1| \leqslant 1$, we have

$$|\mu_1 + \rho v_1| \leqslant 1 \Rightarrow \phi_1(\gamma_1) = -\rho(2-\rho)(B'X(\varepsilon)B)v_1^2 \leqslant 0$$

$$\mu_1 + \rho v_1 > 1 \Rightarrow v_1 > 0, \sigma(\mu_1 + \rho v_1) - (\mu_1 + \rho v_1) < 0$$

$$\Rightarrow \phi_1(\gamma_1) \leqslant -(2-\rho)\,(B'X(\varepsilon)B)v_1[1 - \mu_1] \leqslant 0$$

and

$$\mu_1 + \rho v_1 < -1 \Rightarrow v_1 < 0, \sigma(\mu_1 + \rho v_1) - (\mu_1 + \rho v_1) > 0$$

$$\Rightarrow \phi_1(\gamma_1) \leqslant -(2-\rho)\,(B'X(\varepsilon)B)v_1[-1 - \mu_1] \leqslant 0$$

We conclude that for all $x \in L_{V_1}(c_1)$, $\phi_1(\gamma_1) \leqslant 0$ and hence $V_1(x^+) - V_1(x) \leqslant -x'H(\varepsilon)x$.

So far, we have shown that both for the multiple input and single input case,

$$V_1(x^+) - V_1(x) \leqslant -x'H(\varepsilon)x, \quad \forall x \in L_{V_1}(c_1) \tag{25}$$

which implies that the closed-loop system (18) is locally exponentially stable with $\mathscr{X}_0$ contained in its basin of attraction. We note here that the choice of ρ determines the decay rate of $V_1(x^+) - V_1(x)$ and hence the freedom in choosing ρ can be utilized to ensure fast convergence.

Next, we show that there exists an $\varepsilon_2^* \in (0, 1]$ such that for each $\varepsilon \in (0, \varepsilon_2^*]$, item (ii) of the theorem holds.

To this end, let us introduce an invertible, triangular co-ordinate change $\xi = x - \Pi w$. Using condition (ii) (a) (see Theorem 2.3), we have

$$\begin{aligned}\xi^+ &= x^+ - \Pi w^+ \\ &= Ax + B\sigma(u) + Pw - \Pi S w \\ &= A\xi + B[\sigma(u) - \Gamma w]\end{aligned} \tag{26}$$

With the family of state feedback laws given by (15), the closed-loop system can be written as

$$\begin{aligned}\xi^+ &= A\xi + B[\sigma(\Gamma w - (F(\varepsilon) + \rho\kappa(\xi + \Pi w, w)\,K(\varepsilon)\xi) - \Gamma w] \\ &= A_c\xi + B[\sigma(\Gamma w - (F(\varepsilon) + \rho\kappa(\xi + \Pi w, w)\,K(\varepsilon))\xi) - \Gamma w + F(\varepsilon)\xi]\end{aligned} \tag{27}$$

By Condition (ii) (b) (see Theorem 2.3), $\|\Gamma w\|_{\infty,K} < 1 - \delta$. Moreover, for any $x(0) \in \mathscr{X}_0$ and any $w(0) \in \mathscr{W}_0$, $\xi(K)$ belongs to a bounded set, say $\mathscr{U}_K$, independent of ε since $\mathscr{X}_0$ and $\mathscr{W}_0$ are both bounded and $\xi(K)$ is determined by a linear difference equation with bounded inputs $\sigma(\cdot)$ and Γw.

Copyright © 2000 John Wiley & Sons, Ltd.

We then pick a Lyapunov function

$$V_2(\xi) = \xi' X(\varepsilon)\xi \tag{28}$$

and let $c_2 > 0$ be such that

$$c_2 \geqslant \sup_{\xi \in \mathscr{U}_K,\, \varepsilon \in (0,1]} \xi' X(\varepsilon)\xi \tag{29}$$

Such a c_2 exists since $X(\varepsilon)$ and $\mathscr{U}_K$ are bounded. Let $\varepsilon_2^* \in (0, 1]$ be such that $\xi \in L_{V_2}(c_2)$ implies that $|F(\varepsilon)\xi| \leqslant \delta$ where

$$L_{V_2}(c_2) = \{\xi \in \mathbb{R}^n | V_2(\xi) < c_2\}$$

The existence of such an ε_2^* is again due to Lemma 3.2 and the fact that $\lim_{\varepsilon \to 0} X(\varepsilon) = 0$.

By the definition of function κ, (16), for all $x \in L_{V_1}(c_1)$ and $k \geqslant K$, the saturation functions in the closed-loop system (27) operate in their respective linear regions, and hence the closed-loop system reduces to

$$\xi^+ = A_c\xi + \rho\kappa(\xi + \Pi w, w)\, B(B'X(\varepsilon)B)^{-1}B'X(\varepsilon)A_c\xi \tag{30}$$

The evaluation of the deference of V_1 along the trajectories of this linear closed-loop system shows that, for $\xi \in L_{V_2}(c_2)$,

$$\begin{aligned} V_2(\xi^+) - V_2(\xi) &= -\xi' H(\varepsilon)\xi - \xi' F'(\varepsilon)F(\varepsilon)\xi \\ &\quad - [2 - \rho\kappa(\xi + \Pi w, w)]\xi' A_c' X B(B'XB)^{-1}B'XA_c\xi \\ &\leqslant -\xi' H(\varepsilon)\xi \end{aligned} \tag{31}$$

Now consider the choice $\kappa(x, \omega) \equiv 1$ for the single input case, the evaluation of the difference of V, $k \geqslant K$, inside the set $L_{V_2}(c_2)$, using (23), now shows that for all $\xi \in L_{V_2}(c_2)$,

$$V_2(\xi^+) - V_2(\xi) = -\xi' H(\varepsilon)\xi - \xi' F'(\varepsilon)F(\varepsilon)\xi + \phi_2(\gamma_2) \tag{32}$$

where $\phi_2(\gamma_2) = 2\xi' A_c' X(\varepsilon)\, B\gamma_2 + \gamma_2 B'X(\varepsilon)\, B\gamma_2$ and

$$\gamma_2 = \sigma(\Gamma w - (F(\varepsilon) + \rho(B'X(\varepsilon)B)^{-1}B'X(\varepsilon)A_c)\xi) - \Gamma w + F(\varepsilon)\xi$$

Denoting $\theta_2 = \Gamma w$, $\mu_2 = -F(\varepsilon)\xi$ and $\nu_2 = -K(\varepsilon)\xi$, $\phi_2(\gamma_2)$ can be written as

$$\begin{aligned} \phi_2(\gamma_2) &= -2\nu_2(B'X(\varepsilon)B)[\sigma(\theta_2 + \mu_2 + \rho\nu_2) - \theta_2 - \mu_2] \\ &\quad + [\sigma(\theta_2 + \mu_2 + \rho\nu_2) - \theta_2 - \mu_2](B'X(\varepsilon)B)[\sigma(\theta_2 + \mu_2 + \rho\nu_2) - \theta_2 - \mu_2] \\ &= [\sigma(\theta_2 + \mu_2 + \rho\nu_2) - (\theta_2 + \mu_2 + 2\nu_2)](B'X(\varepsilon)B)[\sigma(\theta_2 + \mu_2 + \rho\nu_2) - \theta_2 - \mu_2] \end{aligned}$$

Noting that $\rho \in [0, 2]$, the definition of σ and for all $\xi \in L_{V_2}(c_2)$, $|\theta_2 + \mu_2| \leqslant 1$, we have

$$\begin{aligned} &|\theta_2 + \mu_2 + \rho\nu_2| \leqslant 1 \Rightarrow \phi_2(\gamma_2) = -\rho(2 - \rho)(B'X(\varepsilon)B)\nu_2^2 \leqslant 0 \\ &\theta_2 + \mu_2 + \rho\nu_2 > 1 \Rightarrow \nu_2 > 0, \sigma(\theta_2 + \mu_2 + \rho\nu_2) - (\theta_2 + \mu_2 + \rho\nu_2) < 0 \\ &\qquad \Rightarrow \phi_2(\gamma_2) \leqslant -(2 - \rho)(B'X(\varepsilon)B)\nu_2[1 - (\theta_2 + \mu_2)] \leqslant 0 \end{aligned}$$

Copyright © 2000 John Wiley & Sons, Ltd.

and

$$\theta_2 + \mu_2 + \rho v_2 < -1 \Rightarrow v_2 < 0, \sigma(\theta_2 + \mu_2 + \rho v_2) - (\theta_2 + \mu_2 + \rho v_2) > 0$$

$$\Rightarrow \phi_2(\gamma_2) \leqslant -(2-\rho)(B'X(\varepsilon)B)v_2[-1-(\theta_2+\mu_2)] \leqslant 0$$

We conclude that for all $\xi \in L_{V_2}(c_2)$, $\phi_2(\gamma_2) \leqslant 0$ and hence $V_2(\xi^+) - V_2(\xi) \leqslant -\xi' H(\varepsilon)\xi$. This shows that any trajectory of the closed-loop system (27) starting from $\{\xi = x - \Pi w: x \in \mathscr{X}_0, w \in \mathscr{W}_0\}$ remains inside the set $L_{V_2}(c_2)$ and approaches the equilibrium $\xi = 0$ as $k \to \infty$, which implies that $e(k) = C\xi(k) \to 0$ as $k \to \infty$.

Finally, setting $\varepsilon^* = \min\{\varepsilon_1^*, \varepsilon_2^*\}$, we conclude our proof of Theorem 2.3. □

Remark

As is clear from the above derivation, the choice of $\kappa(x, w)$ as in (16) prevents the control input from saturating the actuators while increasing the utilization of their capacities. While the avoidance of actuator saturation is essential in establishing (24) due to multiinput coupling, the choice of $\kappa(x, w) \equiv 1$ for single input case allows the control input to saturate the actuators and thus further increase the utilization of their capacities.

We now illustrate the state feedback design by an example.

Example 3.4

Consider the following system:

$$\begin{cases} x^+ = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} x + \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \sigma(u) + \begin{pmatrix} -1 & 2 \\ 0 & 0 \\ -2 & 1 \end{pmatrix} w \\ w^+ = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} w \\ e = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \end{pmatrix} x + \begin{pmatrix} -1 & 0 \\ 1 & -1 \end{pmatrix} w \end{cases} \tag{33}$$

with $w(0) \in \mathscr{W}_0$ where $\mathscr{W}_0 = \{w \in R^2: \|w\| < 0.9\}$. It is straightforward to show that, the solvability conditions for the semi-global linear state feedback output regulation problem are satisfied. More specifically, the matrices,

$$\Pi = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \Gamma = (1 \quad 0) \tag{34}$$

solve the linear matrix equations (10). Also, $\delta = 0.1$, since $\|\Gamma w\|_\infty \leqslant 0.9$ for all $w(0) \in \mathscr{W}_0$. Let the set $\mathscr{X}_0$ be given by $\mathscr{X}_0 = \{x \in \mathbb{R}^4: \|x\| \leqslant 10\}$.

Then, following the proof of Theorem 3.3, a suitable choice of ε^* is 6.628×10^{-6}. For $\varepsilon = \varepsilon^*$ and $\rho = 1$, the feedback law (15) is given by

$$u = (0.3037 \quad 0.1523 \quad -0.2403)x + (0.9366 \quad -0.1523)w$$

Copyright © 2000 John Wiley & Sons, Ltd.

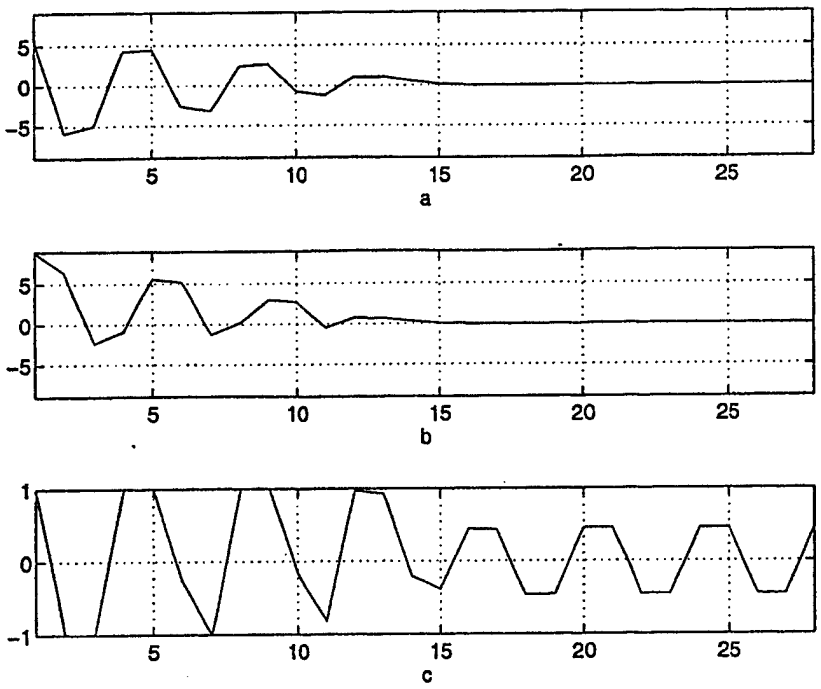


Figure 1. $\varepsilon = 6.628 \times 10^{-6}$, $\rho = 1$: (a) e_1; (b) e_2; (c) $\sigma(u)$.

For the initial conditions $x(0) = (6, 6, -2)'$, $w(0) = (0.45, -0.45)'$, Figure (1) shows the control action and the closed-loop performance of the regulator for $\rho = 1$. Figure (2) shows the control action and the closed-loop performance of the regulator design in Reference [16], i.e. for $\rho = 0$.

4. AN IMPROVED DESIGN FOR THE ERROR FEEDBACK REGULATOR

This section provides an improved design for the error feedback regulators. Significant improvement on the closed-loop performance over the earlier design [16] is again shown by an example. The strategy taken in the new design is to implement the new state feedback regulators as constructed in the previous section with the state of a fast linear observer. The observer is chosen to be of a deadbeat-type. Arbitrary fast observers can also be used, however, the use of a deadbeat-type observer simplifies our proof drastically since the states of the system will be exactly the same as those of the observer in a finite time. More specifically, the new linear observer-based error feedback regulator takes the following form:

$$\hat{x}^+ = A\hat{x} + B\sigma(u) + P\hat{w} + L_A e - L_A(C\hat{x} + Q\hat{w})$$

$$\hat{w}^+ = S\hat{w} + L_S e - L_S(C\hat{x} + Q\hat{w})$$

$$u = -(F(\varepsilon) + \rho\kappa(\hat{x}, \hat{w})K(\varepsilon))\hat{x} + ((F(\varepsilon) + \rho\kappa(x, w)K(\varepsilon))\Pi + \Gamma)\hat{w} \tag{35}$$

Copyright © 2000 John Wiley & Sons, Ltd.

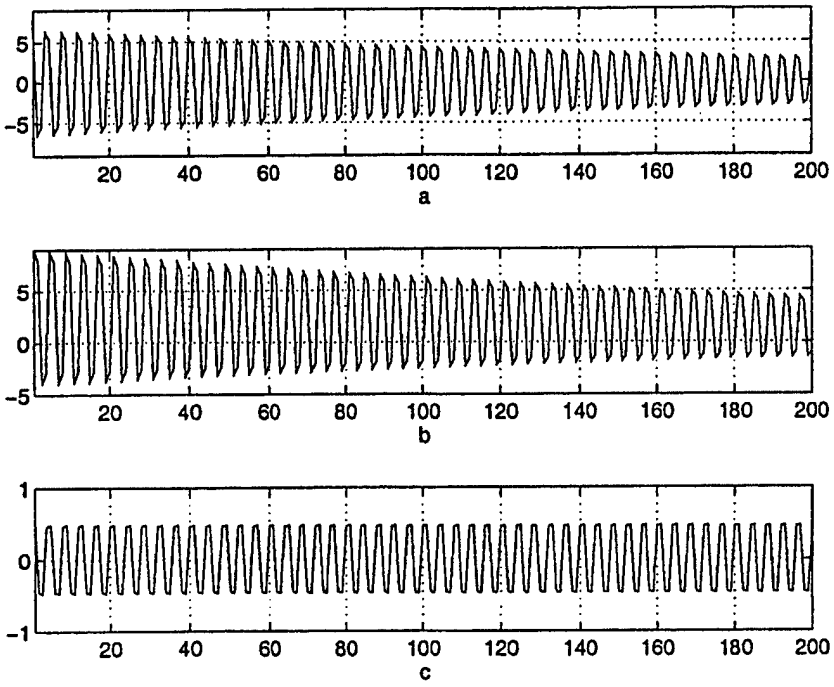


Figure 2. $\varepsilon = 6.628 \times 10^{-6}, \rho = 0$. (a) e_1; (b) e_2; (c) $\sigma(u)$.

where $F(\varepsilon) = (B'X(\varepsilon)B + I)^{-1}B'X(\varepsilon)A$, $K(\varepsilon) = (B'X(\varepsilon)B)^{-1}B'X(\varepsilon)A_c$with $X(\varepsilon)$ being the solution to the Riccati equation (12), and the function κ is as defined by (16) [or (17)]]. The matrices L_A and L_S are chosen such that all eigenvalues of the following matrix:

$$\bar{A} = \begin{pmatrix} A - L_A C & P - L_A Q \\ -L_S C & S - L_S Q \end{pmatrix} \tag{36}$$

are at the origin. Here we, of course, assumed that the pair

$$\left[(C \quad Q), \begin{pmatrix} A & P \\ 0 & S \end{pmatrix} \right] \tag{37}$$

is observable.

We have the following results.

Theorem 4.1

Consider system (1) and the given compact set $\mathscr{W}_0 \subset \mathbb{R}^s$. Assume the sufficient conditions of theorem 2.4 are satisfied. In addition, assume that the pair (37) is observable. Then the error feedback laws (35) solves the semi-global linear observer-based error feedback regulator problem.

More specifically, for any *a priori* given (arbitrarily large) set $\mathscr{X}_0 \subset \mathbb{R}^n$ and $\mathscr{Z}_0 \subset \mathbb{R}^{n+s}$, there exists an $\varepsilon^* \in (0, 1]$, such that for each $\rho \in [0, 2]$,

Copyright © 2000 John Wiley & Sons, Ltd.

(i) The equilibrium $(x, \hat{x}, \hat{w}) = (0, 0, 0)$ of

$$x^+ = Ax + B\sigma(u)$$

$$\begin{pmatrix} \hat{x}^+ \\ \hat{w}^+ \end{pmatrix} = \begin{pmatrix} A & P \\ 0 & S \end{pmatrix} \begin{pmatrix} \hat{x} \\ \hat{w} \end{pmatrix} + \begin{pmatrix} B \\ 0 \end{pmatrix} \sigma(u) + \begin{pmatrix} L_A \\ L_S \end{pmatrix} (C \quad Q) \begin{pmatrix} x - \hat{x} \\ -\hat{w} \end{pmatrix} \tag{38}$$

$$u = -(F(\varepsilon) + \rho\kappa(\hat{x}, \hat{w})K(\varepsilon))\hat{x} + ((F(\varepsilon) + \rho\kappa(\hat{x}, \hat{w})K(\varepsilon))\Pi + \Gamma)\hat{w})$$

is asymptotically stable with $\mathscr{X}_0 \times \mathscr{Z}_0$ contained in its basin of attraction.

(ii) For all $(x(0), \hat{x}(0), \hat{w}(0)) \in \mathscr{X}_0 \times \mathscr{Z}_0$ and $w(0) \in \mathscr{W}_0$, the solution of the closed-loop system satisfies

$$\lim_{k \to \infty} e(k) = 0. \tag{39}$$

Proof. With the family of feedback laws as given by (35), the closed-loop system consisting of system (1) and the dynamic error feedback laws (35) can be written as (we do not write the equation $w^+ = Sw$ explicitly but it is, of course, always there):

$$\begin{aligned} x^+ &= Ax + B\sigma(u) + Pw \\ \hat{x}^+ &= A\hat{x} + B\sigma(u) + P\hat{w} + L_A C(x - \hat{x}) + L_A Q(w - \hat{w}) \\ \hat{w}^+ &= S\hat{w} + L_S C(x - \hat{x}) + L_S Q(w - \hat{w}) \\ u &= -(F(\varepsilon) + \rho\kappa(\hat{x}, \hat{w})K(\varepsilon))\hat{x} + ((F(\varepsilon) + \rho\kappa(\hat{x}, \hat{w})K(\varepsilon))\Pi + \Gamma)\hat{w} \end{aligned} \tag{40}$$

We then adopt the invertible change of state variable,

$$\tilde{x} = x - \hat{x}, \quad \tilde{w} = w - \hat{w} \tag{41}$$

and rewrite the closed-loop system (40) as

$$\begin{aligned} x^+ &= Ax + B\sigma(\Gamma\hat{w} - (F(\varepsilon) + \rho\kappa(x - \tilde{x}, w - \tilde{w})K(\varepsilon))(\hat{x} - \Pi\hat{w})) + Pw \\ \tilde{x}^+ &= (A - L_A C)\tilde{x} + (P - L_A Q)\tilde{w} \\ \tilde{w}^+ &= -L_S C\tilde{x} + (S - L_S Q)\tilde{w} \end{aligned} \tag{42}$$

Since all eigenvalues of $\bar{A}$ are at the origin, it is easy to verify that for time $k \geqslant n + s$, $\tilde{x}(k) \equiv 0$ and $\tilde{w}(k) \equiv 0$. As a result, for $k \geqslant n + s$, $\hat{x}(k) \equiv x(k)$ and $\hat{w}(k) \equiv w(k)$. On the other hand, for all $x(0) \in \mathscr{X}_0$, $(\hat{x}(0), \hat{w}(0)) \in \mathscr{Z}_0$ and $w(0) \in \mathscr{W}_0$, $x(n + s)$ belongs to a bounded set. Hence, the rest of the proof becomes the same as the proof of Theorem 2.3. □

Example 4.2

We consider the same plant and the exosystem as in Example 3.4. However, this time, the state x and w are not available for feedback, which forces us to use error feedback regulators. Let the

Copyright © 2000 John Wiley & Sons, Ltd.

sets $\mathscr{W}_0 = \{w \in \mathbb{R}^2: \|w\| < 0.9\}$ and $\mathscr{X}_0 = \{x \in R^4: \|x\| < 10\}$. Let the set $\mathscr{Z}_0$ be given by $\mathscr{Z}_0 = \{z \in \mathbb{R}^6: \|z\| \leqslant 2\}$. Following the proof of Theorem 4.1, a suitable choice of ε^* is 6.628×10^{-6}. It can be verified that the matrix $\bar{A}$ as defined in (36) has all its eigenvalues located at the origin if we choose

$$L_A = \begin{pmatrix} 0.3748 & -0.8750 \\ 1.3752 & 0.1250 \\ -0.6252 & -0.8750 \end{pmatrix} \quad L_S = \begin{pmatrix} 0.25 & -0.25 \\ -0.3752 & -0.125 \end{pmatrix}$$

For $\varepsilon = \varepsilon^*$ and $\rho = 1$, the feedback laws (35) are given by

$$\begin{aligned} \hat{x}^+ &= A\hat{x} + B\sigma(u) + P\hat{w} + L_A C(x - \hat{x}) + L_A Q(w - \hat{w}) \\ \hat{w}^+ &= S\hat{w} + L_S C(x - \hat{x}) + L_S Q(w - \hat{w}) \\ u &= -(-0.3037 \quad -0.1523 \quad 0.2403)\hat{x} \\ &\quad + (0.9366 \quad -0.1523)\hat{w} \end{aligned} \tag{43}$$

For the initial conditions $x(0) = (-6, 2, 6)'$, $w(0) = (0.45, -0.45)'$, $\hat{x}(0) = (0, 0, 0, 0)'$, $\hat{w}(0) = (0, 0)'$, Figure (3) shows the control action and the closed-loop performance for the

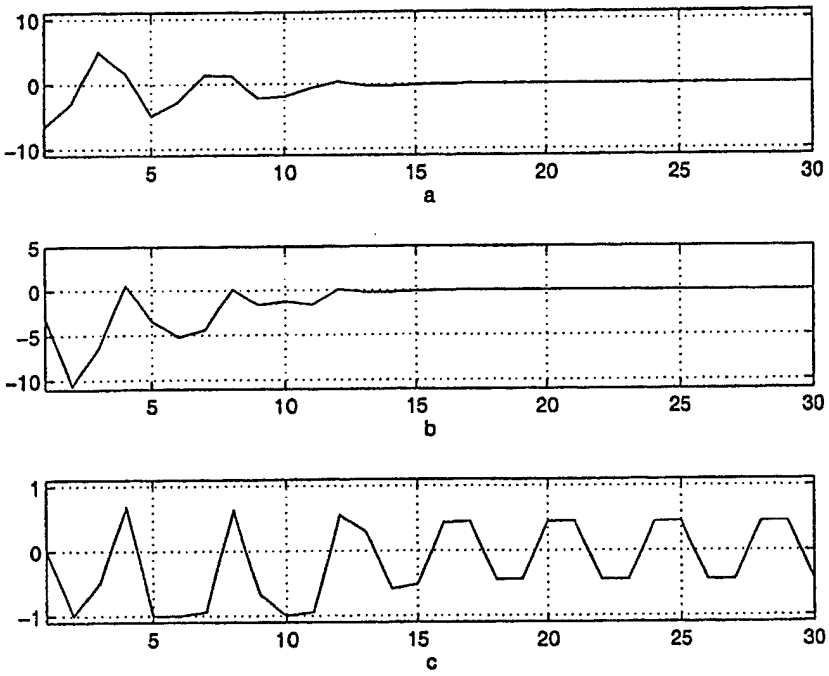


Figure 3. $\varepsilon = 6.628 \times 10^{-6}$, $\rho = 1$. (a) e_1; (b) e_2; (c) $\sigma(u)$.

Copyright © 2000 John Wiley & Sons, Ltd.

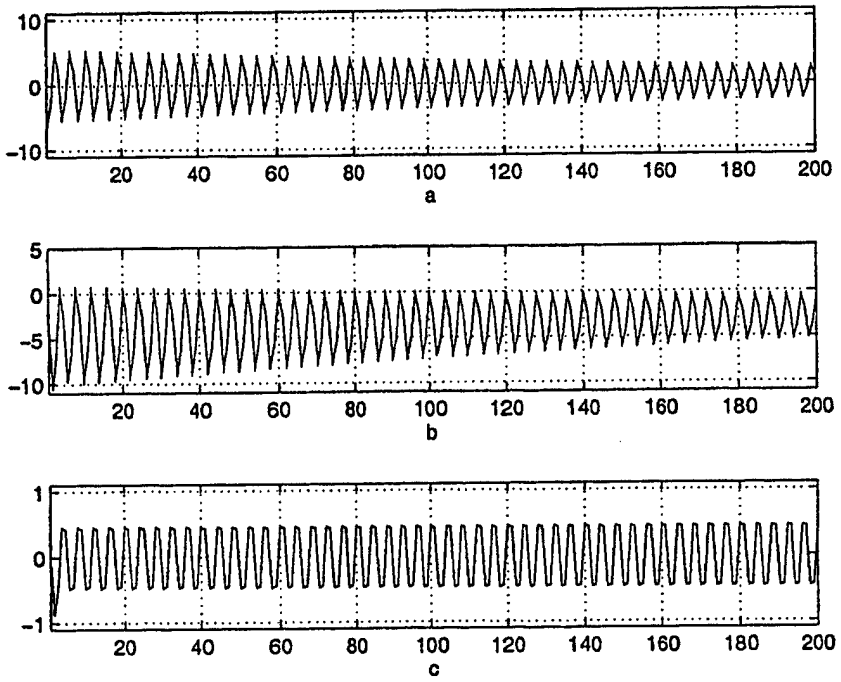


Figure 4. $\varepsilon = 6.628 \times 10^{-6}$, $\rho = 0$. (a) e_1; (b) e_2; (c) $\sigma(u)$.

dynamic error feedback regulator for $\rho = 1$. Figure (4) shows the control action and the closed-loop performance of the dynamic error feedback regulator design in Reference [16], i.e. for $\rho = 0$.

5. GENERALIZED SEMI-GLOBAL OUTPUT REGULATION PROBLEMS

In an effort to broaden the class of disturbance and reference signals, in References [15, 16], we formulated the generalized semi-global feedback regulator problems, in which an external driving signal to the exosystem is included. More specifically, we consider a multivariable system with inputs that are subject to saturation together with an exosystem that generates disturbance and reference signals as described by the following system:

$$\begin{aligned} x^+ &= Ax + B\sigma(u) + Pw \\ w^+ &= Sw + r \\ e &= Cx + Qw \end{aligned} \tag{44}$$

where $x \in \mathbb{R}^n$, $w \in \mathbb{R}^s$, $u \in \mathbb{R}^m$, $e \in \mathbb{R}^p$, $r \in L_\infty$ is an external signal to the exosystem, and σ is a vector-valued saturation function as defined in Section 2.

The generalized semi-global state feedback regulation problem and the generalized semi-global linear observer-based error feedback regulation problem are formulated as follows.

Copyright © 2000 John Wiley & Sons, Ltd.

Problem 5.1

Consider system (44), two compact sets $\mathscr{W}_0 \subset \mathbb{R}^s$ and $\mathscr{R} \subset L_\infty$. *The generalized semi-global state feedback regulator problem* is defined as follows:

For any *a priori* given (arbitrarily large) bounded set $\mathscr{X}_0 \subset \mathbb{R}^n$, find, if possible, a static feedback law $u = \alpha(x, w, r)$, such that

(i) The equilibrium $x = 0$ of

$$x^+ = Ax + B\sigma(\sigma(\alpha(x, 0, 0)) \tag{45}$$

is locally aymptotically stable with χ_0 contained in its basin of attraction;

(ii) For all $x(0) \in \chi_0$, $w(0) \in \mathscr{W}_0$ and $r \in \mathscr{R}$, the solution of the closed-loop system satisfies

$$\lim_{k\to\infty} e(k) = 0. \tag{46}$$

Problem 5.2

Consider system (44) and two compact sets $\mathscr{W}_0 \subset \mathbb{R}^s$ and $\mathscr{R} \subset L_\infty$. *The generalized semi-global linear observer-based error feedback regulator problem* is defined as follows.

For any *a priori* given (arbitrarily large) bounded sets $\mathscr{X}_0 \subset \mathbb{R}^n$ and $\mathscr{Z}_0 \subset \mathbb{R}^{n+s}$,fi nd, if possible, a linear observer-based error feedback law of the form:

$$\begin{cases} \begin{pmatrix} \hat{x}^+ \\ \hat{w}^+ \end{pmatrix} = \begin{pmatrix} A & P \\ 0 & S \end{pmatrix}\begin{pmatrix} \hat{x} \\ \hat{w} \end{pmatrix} + \begin{pmatrix} B \\ 0 \end{pmatrix}\sigma(u) + \begin{pmatrix} 0 \\ I \end{pmatrix} r + \begin{pmatrix} L_A \\ L_S \end{pmatrix}\left[e - (C \quad Q)\begin{pmatrix} \hat{x} \\ \hat{w} \end{pmatrix}\right] \\ u = \alpha(\hat{x}, \hat{w}, r) \end{cases} \tag{47}$$

such that

(i) The equilibrium $(x, \hat{x}, \hat{w}) = (0, 0, 0)$ of

$$\begin{cases} x^+ = Ax + B\sigma(\alpha(\hat{x}, \hat{w}, 0)) \\ \begin{pmatrix} \hat{x}^+ \\ \hat{w}^+ \end{pmatrix} = \begin{pmatrix} A & P \\ 0 & S \end{pmatrix}\begin{pmatrix} \hat{x} \\ \hat{w} \end{pmatrix} + \begin{pmatrix} B \\ 0 \end{pmatrix}\sigma(\alpha(\hat{x}, \hat{w}, 0)) + \begin{pmatrix} L_A \\ L_S \end{pmatrix}(C \quad Q)\begin{pmatrix} x - \hat{x} \\ -\hat{w} \end{pmatrix}\end{cases} \tag{48}$$

is locally asymptotically stable with $\mathscr{X}_0 \times \mathscr{Z}_0$ contained in its basin of attraction;

(ii) For all $x(0), \hat{x}(0), \hat{w}(0) \in \mathscr{X}_0 \times \mathscr{Z}_0$, $w(0) \in \mathscr{W}_0$, and all $r \in \mathscr{R}$, the solution of the closed-loop system satisfies

$$\lim_{k\to\infty} e(k) = 0. \tag{49}$$

In this section we show that the improved design technique as developed in the previous two sections can also be used to construct the regulators to solve the generalized semi-global output regulation problems. We summarize these results in the following two theorems.

Theorem 5.3

Consider system (44) and given compact sets $\mathscr{W}_0 \subset \mathbb{R}^s$ and $\mathscr{R} \subset L_\infty$. If

(i) (A, B) is stabilizable and A has all its eigenvalues on or inside the unit circle.

(ii) There exist matrices Π and Γ such that:

Copyright © 2000 John Wiley & Sons, Ltd.

(a) They solve the following linear matrix equations:

$$\begin{aligned} \Pi S &= A\Pi + B\Gamma + P \\ 0 &= C\Pi + Q \end{aligned} \tag{50}$$

(b) For each $r \in \mathscr{R}$, there exists a function $\tilde{r} \in L_\infty$ such that $\Pi r = B\tilde{r}$.
(c) There exists a $\delta > 0$ and a $K \geqslant 0$ such that $\|\Gamma w + \tilde{r}\|_{\infty,K} \leqslant 1 - \delta$ for all w with $w(0) \in \mathscr{W}_0$ and all $r \in \mathscr{R}$.

Then the following family of state feedback laws solves the generalized semi-global state feedback output regulation problem,

$$u = -[F(\varepsilon) + \rho\kappa(x, w)K(\varepsilon)]x + [(F(\varepsilon) + \rho\kappa(x, w)K(\varepsilon))\Pi + \Gamma]w + \tilde{r}, \quad \rho \in [0, 2] \tag{51}$$

where, $F(\varepsilon) = (B'X(\varepsilon)B + I)^{-1}B'X(\varepsilon)A, K(\varepsilon) = (B'X(\varepsilon)B)^{-1}B'X(\varepsilon)A, A_c = A - BF(\varepsilon)$, with $X(\varepsilon)$ being the solution of the Riccati equation (12) and the function κ is as defined by (16).

More specifically, for any *a priori* given (arbitrarily large) bounded set $\mathscr{X} \in \mathbb{R}^n$, there exists an $\varepsilon \in (0, 1]$ such that for each $\varepsilon \in (0, \varepsilon^*)$ and for each $\rho \in [0, 2]$,

(i) The equilibrium $x = 0$ of

$$x^+ = Ax + B\sigma(-F(\varepsilon)x - \rho\kappa(x, 0)K(\varepsilon)x)$$

is locally exponentially stable with $\mathscr{X}_0$ contained in its basin of attraction.
(ii) For all $x(0) \in \mathscr{X}_0$, $w(0) \in \mathscr{W}_0$, and $r \in \mathscr{R}$, the solutions of the closed-loop system satisfies

$$\lim_{k\to\infty} e(k) = 0$$

Proof. The proof is similar, *mutatis mutandis* to that of Theorem 3.3, except that (27) takes the following slightly different form:

$$\xi^+ = A_c\xi + B[\sigma\Gamma w + \tilde{r} - (F(\varepsilon) + \rho\kappa(x, w)K(\varepsilon)\xi) - \Gamma w + F(\varepsilon)\xi - \tilde{r}] \qquad \square$$

Theorem 5.4
Consider system (44) and the given compact sets $\mathscr{W}_0 \subset \mathbb{R}^s$ and $\mathscr{R} \subset L_\infty$. If

(i) (A, B) is stabilizable and A has all its eigenvalues inside or on the unit circle. Moreover, the pair

$$\left[(C \quad Q), \begin{pmatrix} A & P \\ 0 & S \end{pmatrix}\right]$$

is observable;
(ii) There exist matrices Π and Γ such that:

(a) They solve the following linear matrix equations:

$$\begin{aligned} \Pi S &= A\Pi + B\Gamma + P \\ 0 &= C\Pi + Q \end{aligned} \tag{52}$$

(b) For each $r \in \mathscr{R}$, there exists a function $\tilde{r} \in L_\infty$ such that $\Pi r = B\tilde{r}$ for all $k \geqslant 0$.
(c) There exist a $\delta > 0$ and a $K \geqslant 0$ such that $\|\Gamma w + \tilde{r}\|_{\infty,K} \leqslant 1 - \delta$ for all w with $w(0) \in \mathscr{W}_0$ and all $r \in \mathscr{R}$.

Copyright © 2000 John Wiley & Sons, Ltd.

Then, the following family of linear observer-based error feedback laws solves the generalized semi-global linear observer-based error feedback output regulation problem.

$$\begin{aligned}\hat{x}^+ &= A\hat{x} + B\sigma(u) + P\hat{w} + L_A e - L_A(C\hat{x} + Q\hat{w})\\ \hat{w}^+ &= S\hat{w} + L_S e - L_S(C\hat{x} + Q\hat{w}) + r\\ u &= -(F(\varepsilon) + \rho\kappa(\hat{x}, \hat{w})K(\varepsilon))\hat{x} + ((F(\varepsilon) + \rho\kappa(\hat{x}, \hat{w})K(\varepsilon))\Pi + \Gamma)\hat{w} + \tilde{r}\end{aligned} \tag{53}$$

with $X(\varepsilon)$ being the solution of the Riccati equation (12), and the function κ is as defined by (16). The matrices L_A and L_S are chosen such that all eigenvalues of the following matrix:

$$\bar{A} = \begin{pmatrix} A - L_A C & P - L_A Q \\ -L_S C & S - L_S Q \end{pmatrix} \tag{54}$$

are at the origin.

More specifically, for any *a priori* given (arbitrarily large) bounded set $\mathscr{X}_0 \in \mathbb{R}^n$ and $\mathscr{Z}_0 \in \mathbb{R}^{n+s}$, there exists an $\varepsilon^* \in (0, 1]$ such that for each $\varepsilon \in (0, \varepsilon^*)$ and for each $\rho \in [0, 2]$,

(i) The equilibrium $(x, \hat{x}, \hat{w}) = (0, 0, 0)$ of

$$\begin{aligned} x^+ &= Ax + B\sigma(u)\\ \begin{pmatrix}\hat{x}^+\\ \hat{w}^+\end{pmatrix} &= \begin{pmatrix} A & P\\ 0 & S\end{pmatrix}\begin{pmatrix}\hat{x}\\ \hat{w}\end{pmatrix} + \begin{pmatrix}B\\ 0\end{pmatrix}\sigma(u) + \begin{pmatrix}0\\ I\end{pmatrix}r + \begin{pmatrix}L_A\\ L_S\end{pmatrix}\begin{pmatrix}C & Q\end{pmatrix}\begin{pmatrix}x - \hat{x}\\ -\hat{w}\end{pmatrix}\\ u &= -(F(\varepsilon) + \rho\kappa(\hat{x}, \hat{w})K(\varepsilon))\hat{x} + ((F(\varepsilon) + \rho\kappa(\hat{x}, \hat{w})K(\varepsilon))\Pi + \Gamma)\hat{w}\end{aligned} \tag{55}$$

is asymptotically stable with $\mathscr{X}_0 \times \mathscr{Z}_0$ contained in its basin of attraction.

(ii) For all $(x(0), \hat{x}(0), \hat{w}(0)) \in \mathscr{X}_0 \times \mathscr{Z}_0$, $w(0) \in \mathscr{W}_0$, and all $r \in \mathscr{R}$, the solution of the closed-loop system satisfies

$$\lim_{k\to\infty} e(k) = 0 \tag{56}$$

Proof. The proof is similar to that of Theorem 4.1. □

6. CONCLUSIONS

New regulators were constructed to solve the semi-global output regulation problems for linear discrete-time systems subject to input saturation. These regulators make better use of the available control capacity to achieve better closed-loop system performance.

ACKNOWLEDGEMNET

Work supported in part by the US Office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

Copyright © 2000 John Wiley & Sons, Ltd.

REFERENCES

1. Fuller AT. In-the-large stability of relay and saturating control systems with linear controller. International Journal of Control 1969; **10**(4):457–480.
2. Sontag ED, Sussmann HJ. Nonlinear output feedback design for linear systems with saturating controls. In *Proceedings of the 29th CDC*, Honolulu, 1990, 3414–3416.
3. Sussmann HJ, Sontag ED, Yang Y. A general result on the stabilization of linear systems using bounded controls. *IEEE Transaction on Automatic Control* 1994; **39**(12):2411–2425.
4. Sussmann HJ, Yang Y. On the stabilizability of multiple integrators by means of bounded feedback controls. In *Proceedings of the 30th CDC*, Brighton, U.K., 1991; 70–72.
5. Teel AR. Global stabilization and restricted tracking for multiple integrators with bounded controls. *System & Control Letters* 1992; **18**(3):165–171.
6. Lin Z. Global and semi-global control problems for linear systems subject to input saturation and minimum-phase input–output linearizable systems, *Ph.D. Thesis*, Washington State University, May 1994.
7. Lin Z, Saberi A. Semi-global exponential stabilization of linear systems subject to "input saturation" via linear feedbacks. *Systems and Control Letters* 1993; **21**(3):225–239.
8. Lin Z, Saberi A. Semi-global exponential stabilization of linear discrete-time systems subject to 'input saturation' via linear feedbacks. *Systems and Control Letters* 1995; **24**:125–132.
9. Lin Z, Saberi A. A semi-global low-and-high gain design technique for linear systems with input saturation – stabilization and disturbance rejection. *International Journal of Robust and Nonlinear Control* 1995; **5**:381–398.
10. Lin Z, Saberi A, Stoorvogel A. Semi-global stabilization of linear discrete-time systems subject to input saturation via linear feedback — an ARE-based approach. *IEEE Transactions on Automatic Control* 1996; **41**(8):1203–1207.
11. Saberi A, Lin Z, Teel A. Control of linear systems with saturating actuators. *IEEE Transactions on Automatic Control* 1996; **41**(3):368–378.
12. Teel AR. Semi-global stabilization of linear null-controllable systems with input nonlinearities. *IEEE Transactions on Automatic Control* 1995; **40**(1):96–100.
13. Teel AR. Feedback stabilization: nonlinear solutions to inherently nonlinear problems. *Ph.D. Thesis*, Electronics Research Laboratory, College of Engineering, University of California, 1992.
14. Lin Z, Mantri R, Saberi A. Semi-global output regulation for linear systems subject to input saturation – a low-and-high gain design. *Control Theory and Advanced Technology* 1995; **4**(5):2209–2232.
15. Lin Z, Stoorvogel AA, Saberi A. Output regulation for linear systems subject to input saturation. *Automatica* 1996; **32**(1):29–47.
16. Mantri R, Saberi A, Lin Z, Stoorvogel A. Output regulation for linear discrete-time systems subject to input saturation. *International Journal of Robust and Nonlinear Control* 1997; **7**(11):1003–1021.
17. Gutman P-O, Hagander P. A new design of constrained controllers for linear systems. *IEEE Transactions on Automatic Control* 1985; **30**(1):22–33.
18. Francis BA. The linear multivariable regulator problem. *SIAM Journal on Control and Optimisation* 1997; **15**(3):486–505.

Copyright © 2000 John Wiley & Sons, Ltd.

Publication 4

A Complete Stability Analysis of Planar Linear Systems Under Saturation

Tingshu Hu, *Student Member, IEEE,* and Zongli Lin, *Senior Member, IEEE*

***Abstract*—A complete stability analysis is performed on a planar system of the form $\dot{x} = \sigma(Ax)$ where A is a Hurwitz matrix and σ is the saturation function. Necessary and sufficient conditions for the system to be globally asymptotically stable (GAS) or to have a closed trajectory are explicitly given in terms of the entries of A. These conditions also indicate that the system always has a closed trajectory if it is not GAS.**

***Index Terms*—Closed trajectories, neural networks, saturation, stability.**

I. Introduction

DYNAMICAL systems with saturation nonlinearities arise frequently in neural networks, analog circuits, and control systems (see, for example, [2], [4], [5], and [8] and the references therein). In this paper, we consider the systems of the following form:

$$\dot{x} = \sigma(Ax), \qquad x \in R^n \tag{1}$$

where σ: $R^n \to R^n$ is the standard saturation function. With a slight abuse of notation, we use the same symbol to denote both the vector saturation function and the scalar saturation function, i.e., if $v \in R^n$, then $\sigma(v) = [\sigma(v_1), \sigma(v_2), \cdots, \sigma(v_n)]^T$ and

$$\sigma(v_i) = \begin{cases} -1, & \text{if } v_i < -1 \\ v_i, & \text{if } -1 \le v_i \le 1, \\ 1 & \text{if } v_i > 1. \end{cases} \tag{2}$$

Systems of the form (1) and their discrete counterparts mainly arise in neural networks and in digital filters.

As with any dynamical system, stability of these systems is of primary concern and has been heavily studied in the literature for a long period of time (see, for example, [1], [6], [7], [8], and [10] and the references therein). As seen in the literature, the stability analysis of such systems are highly nontrivial. Even for the planar case, only sufficient conditions for global asymptotic stability are available [1], [8], [10]. In this paper, we present a complete analysis of the planar system of the form (1). In particular, necessary and sufficient conditions for the system to be globally asymptotically stable (GAS) or to have a limit circle are explicitly given in terms of the entries of the matrix A. We will also describe a surprising, but appealing, phenomenon that even with an unstable matrix A it is still possible for the system to have a bounded global attractor.

We would like to point out that the necessary and sufficient conditions for planar linear systems operating on the unit square to be GAS were recently identified in [8] and [9]. The class of linear systems operating on the unit square can be put in a form similar to (1) with σ being a state-dependent function that takes zero value instead of ± 1 as the saturation function does, whenever the state is to leave the unit square. By forcing the state within the unit square, the dynamical behavior is completely different. For example, the closed trajectory would not exist [9].

We will begin searching for the necessary and sufficient condition for the system to be GAS by drawing a general picture of the vector field in Section III. Some constants are captured to characterize the vector field. In Section IV, we show that it is these constants, rather than the stability of the A matrix, that determine the global boundedness of the trajectories. An interesting example is presented to show that even if A is unstable, the system can still have a bounded global attractor.

The condition for the existence of a bounded global attractor as given in Section IV, along with the stability of the matrix A, guarantees the system to be GAS. This is shown in Section V. Now that all the trajectories are bounded, the only problem to be solved in Section V is the nonexistence of a closed trajectory. This problem turns out to be quite complicated due to the partition of the vector field by the saturation. In the central unit square $\sigma(x) = x$, and a trajectory in this region follows that of a linear system. Off the central square, the sequence of the intersections of a trajectory with a straight line is governed by a first-order linear time invariant discrete-time system. The real complexity arises when a trajectory traverses between the central square and other regions. We will approach this problem through evolving models with

$$A = \begin{bmatrix} -1, & a_{12} \\ -ka_{12}a_{22}, & a_{22} \end{bmatrix}, \qquad a_{12} > 1,\ k \ge 1.$$

In the primary model $a_{22} = 1$ and $k = 1$. In the secondary model $a_{22} = 1$ and $k \ge 1$. In the third-level model $a_{22} \in (0, 1)$ and $k \ge 1$. The trajectories of the secondary model are very appealing. Inside a certain ellipse all the trajectories are closed and outside this ellipse all the trajectories converge to this ellipse. We will establish our main results by comparing the trajectories of the general model with those of a secondary model, which in turn are characterized by comparing with the primary model.

Manuscript received April 14, 1999; revised September 15, 1999. This work was supported in part by the U.S. Office of Naval Research Young Investigator Program under Grant N00014-99-1-0670. This paper was recommended by Associate Editor V. Pérez Villar.

The authors are with the Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903 USA (e-mail: th7f@virginia.edu; zl5y@virginia.edu).

Publisher Item Identifier S 1057-7122(00)02918-4.

1057–7122/00$10.00 © 2000 IEEE

II. Main Results

Consider the following system

$$\dot{x} = \sigma(Ax), \qquad x \in \boldsymbol{R}^2 \tag{3}$$

where σ: $\boldsymbol{R}^2 \to \boldsymbol{R}^2$ is the saturation function, i.e., if $v \in \boldsymbol{R}^2$, then $\sigma(v) = \begin{bmatrix} \sigma(v_1) \\ \sigma(v_2) \end{bmatrix}$ and σ is as defined by (2). For this system, we observe that its dynamics are left unchanged under the state transformation $z = Tx$ if T is of the form $T = PS$ where P is a permutation matrix and $S = \text{diag}(\pm 1, \pm 1)$.

We assume through out this paper that $\det(A) \neq 0$. This implies that A is nonsingular and the system has a unique equilibrium point at the origin. Following the idea of [1], let $z = Ax$. The system (3) is then transformed into the following form:

$$\dot{z} = A\dot{x} = A\sigma(Ax) = A\sigma(z).$$

We see that the dynamics of the system (3) and hence its stability properties are equivalent to those of the system

$$\dot{x} = A\sigma(x). \tag{4}$$

We will focus on (4) in this paper.

Given an initial state x_0, denote the trajectory of the system (4) that passes through x_0 at $t = 0$ as $\psi(t, x_0)$. Mainly, we consider the positive trajectory $\psi(t, x_0)$, $t \geq 0$. However, occasionally we use $\psi(-t, x_0)$, $t \geq 0$ for the purpose of comparison.

Definition 2.1: The system (4) is said to be stable at its equilibrium $x_e = 0$ if, for any $\epsilon > 0$, there exists a $\delta > 0$ such that $\|\psi(t, x_0)\| \leq \epsilon$ for all $t \geq 0$ and $\|x_0\| \leq \delta$. It is said to be GAS if $x_e = 0$ is a stable equilibrium and $\lim_{t \to \infty} \psi(t, x_0) = 0$ for all $x_0 \in \boldsymbol{R}^2$. Also, it is said to be locally asymptotically stable if it is stable and $\lim_{t \to \infty} \psi(t, x_0) = 0$ for $x_0 \in U_0$, a neighborhood of $x_e = 0$.

Obviously, $x_e = 0$ is a locally asymptotically stable equilibrium if and only if A is Hurwitz. In this case, at least one of its diagonal elements must be negative. Without loss of generality, we assume throughout the remaining part of this paper that

$$A = \begin{bmatrix} -a_{11}, & a_{12} \\ -a_{21}, & a_{22} \end{bmatrix}, \qquad a_{11} > 0,\ a_{21} \geq 0. \tag{5}$$

Otherwise, we can use $T = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ as the state transformation matrix to make $a_{11} > 0$ or use $T = \left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$ to make $a_{21} \geq 0$.

Our main result in this paper, presented in the following theorem, gives a complete description of the stability properties of the system (4) with A given in (5). As explained above, any Hurwitz A can be transformed into the form of (5).

Theorom 2.1: The system (4) is GAS if and only if A is Hurwitz and one of the following conditions is satisfied:

a) $a_{22} < 0$;

b) $a_{22} \geq 0$ and $a_{11}a_{21} \geq a_{12}a_{22}$.

On the other hand, if none of a) and b) is satisfied, the system will have diverging trajectories and there will be a closed trajectory.

In proving this main result, we will also obtain conditions under which all the trajectories of the system (4) are bounded.

Remark 2.1: We recall a recent sufficient condition for global asymptotic stability of the system (4) from [1]. The results of [1], tailored to the special form of A in (5), is summarized as follows. the system (4) is GAS if A is Hurwitz and one of the following conditions is satisfied.

a) $a_{22} < 0$.

b) $a_{22} \geq 0$ and $a_{11} > a_{12}$.

The fact that $a_{22} \geq 0$ and A is Hurwitz imply that $a_{12} > 0$.

In view of Remark 2.1, we only need to consider the case where $a_{22} \geq 0$ and $a_{11} \leq a_{12}$. In this case, the four parameters a_{11}, a_{12}, a_{21}, a_{22} are all nonnegative.

III. The Vector Field

In this section, we present a general picture of the following vector field:

$$\begin{aligned} \dot{x}_1 &= -a_{11}\sigma(x_1) + a_{12}\sigma(x_2) =: f_1(x), \\ \dot{x}_2 &= -a_{21}\sigma(x_1) + a_{22}\sigma(x_2) =: f_2(x) \end{aligned} \tag{6}$$

where $0 < a_{11} \leq a_{12}, a_{21}, a_{22} \geq 0$ and $\det(A) \neq 0$. Denote the slope of the trajectory at x as

$$\eta(x) := \frac{f_2(x)}{f_1(x)}.$$

The vector field of (6) is partitioned into nine regions, according to the saturation function, by two vertical lines $x_1 = \pm 1$ and two horizontal lines $x_2 = \pm 1$ (see Fig. 1). In the central unit square, $\dot{x} = Ax$.

In the region $U := \{x: |x_1| \leq 1,\ x_2 \geq 1\}$

$$\begin{aligned} \dot{x}_1 &= -a_{11}x_1 + a_{12} \\ \dot{x}_2 &= -a_{21}x_1 + a_{22}. \end{aligned}$$

Since $a_{11} \leq a_{12}$, $\dot{x}_1 \geq 0$ in this region and the trajectories go rightward. Also note that $\dot{x}$ is independent of x_2, so for all the points on a vertical line $x_1 = c$, $|c| \leq 1$ in this region $\dot{x}$ is the same. Because of this, if $x_0 \in U$ and $\psi(t, x_0) \in U$ for all $t \in [0, t_1]$, then with $\Delta > 0$

$$\psi\left(t, x_0 + \begin{bmatrix} 0 \\ \Delta \end{bmatrix}\right) = \psi(t, x_0) + \begin{bmatrix} 0 \\ \Delta \end{bmatrix} \qquad \forall t \in [0, t_1]. \tag{7}$$

We call (7) the vertical shifting property in the region U. Specifically, let $x_0 = \left[\begin{smallmatrix} -1 \\ x_{02} \end{smallmatrix}\right]$, $x_{02} \geq 1$ be a point on the line $x_1 = -1$, then

$$\begin{aligned} x_1(t) &= e^{-a_{11}t}(-1) + \frac{a_{12}}{a_{11}}\left(1 - e^{-a_{11}t}\right) \\ &= -\left(1 + \frac{a_{12}}{a_{11}}\right)e^{-a_{11}t} + \frac{a_{12}}{a_{11}} \end{aligned} \tag{8}$$

$$\begin{aligned} x_2(t) &= x_{02} + \int_0^t \left(-a_{21}x_1(\tau) + a_{22}\right) d\tau \\ &= x_{02} + \frac{a_{21}(a_{12} + a_{11})}{a_{11}^2}\left(1 - e^{-a_{11}t}\right) - \frac{\det A}{a_{11}}t. \end{aligned} \tag{9}$$

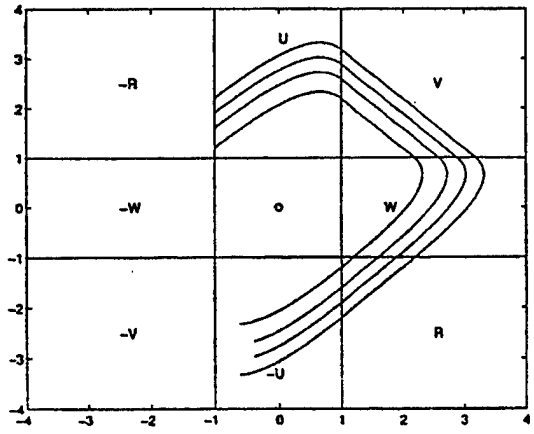


Fig. 1. The partion of the vector field.

Suppose $x(t)$ remains in the region U before it intersects with the line $x_1 = 1$ at $t = T$ with $x(T) = \begin{bmatrix} 1 \\ x_{T2} \end{bmatrix}$, then by solving (8) with $x_1(T) = 1$ we get

$$T = \frac{1}{a_{11}} \log \frac{a_{12} + a_{11}}{a_{12} - a_{11}}$$

and from (9) with $t = T$ we have

$$x_{T2} - x_{02} = \frac{2a_{21}}{a_{11}} - \frac{\det A}{a_{11}^2} \log \frac{a_{12} + a_{11}}{a_{12} - a_{11}}.$$

In the above derivation, we have assumed that $a_{11} < a_{12}$. As Proposition 3.1 will show, $a_{11} = a_{12}$ automatically ensures the global asymptotic stability of the system (4) if A is Hurwitz.

As expected, the increment of x_2 from $t = 0$ to $t = T$ is independent of x_{02}. We denote this constant as

$$h_2 := x_{T2} - x_{02} = \frac{2a_{21}}{a_{11}} - \frac{\det A}{a_{11}^2} \log \frac{a_{12} + a_{11}}{a_{12} - a_{11}}. \quad (10)$$

In the region $V := \{x:\ x_1 \geq 1, x_2 \geq 1\}$

$$\dot{x} = \begin{bmatrix} -a_{11} + a_{12} \\ -a_{21} + a_{22} \end{bmatrix}$$

is a constant. So the slope of the trajectories is a constant. We denote this constant slope $\eta(x)$ as

$$\alpha := \frac{-a_{21} + a_{22}}{-a_{11} + a_{12}}. \quad (11)$$

In the region $W := \{x:\ x_1 \geq 1, |x_2| \leq 1\}$

$$\dot{x}_1 = -a_{11} + a_{12}x_2$$
$$\dot{x}_2 = -a_{21} + a_{22}x_2.$$

In contrast to the region U, $\dot{x}$ is independent of x_1. If $x_0 \in W$ and $\psi(t, x_0) \in W$ for all $t \in [0, t_1]$, then with $\Delta > 0$ we have

$$\psi\left(t, x_0 + \begin{bmatrix} \Delta \\ 0 \end{bmatrix}\right) = \psi(t, x_0) + \begin{bmatrix} \Delta \\ 0 \end{bmatrix} \quad \forall t \in [0, t_1]. \quad (12)$$

We call (12) the horizontal shifting property in the region W. As Proposition 3.1 will show, if $a_{21} \leq a_{22}$, the system (4) will not be GAS. Now for the case that $a_{21} > a_{22}$, $\dot{x}_2 < 0$ and $\dot{x}$ points downward in this region. In this case, if a trajectory starts at a point $x_0 = \begin{bmatrix} x_{01} \\ 1 \end{bmatrix}$, $x_{01} \geq 1$ on the line $x_2 = 1$ and crosses the line $x_2 = -1$ at a point $x(T) = \begin{bmatrix} x_{T1} \\ -1 \end{bmatrix}$, $x_{T1} \geq 1$, then $x_{T1} - x_{01}$ is a constant. We denote this constant as h_1. It can be verified, as with the the constant h_2, that

$$h_1 = x_{T1} - x_{01} = \begin{cases} -\dfrac{2a_{12}}{a_{22}} + \dfrac{\det A}{a_{22}^2} \log \dfrac{a_{21} + a_{22}}{a_{21} - a_{22}}, & \text{if } a_{22} > 0 \\ -\dfrac{2a_{11}}{a_{21}}, & \text{if } a_{22} = 0. \end{cases} \quad (13)$$

In the region $R := \{x:\ x_1 \geq 1, x_2 \leq -1\}$

$$\dot{x} = \begin{bmatrix} -a_{11} - a_{12} \\ -a_{21} - a_{22} \end{bmatrix}.$$

We denote the constant slope $\eta(x)$ in this region as

$$\beta := \frac{a_{21} + a_{22}}{a_{11} + a_{12}}. \quad (14)$$

The remaining four regions are symmetric to U, V, W, and R. We denote them as $-U$, $-V$, $-W$, $-R$.

For a general second-order nonlinear system that has a unique equilibrium point at the origin, its GAS can be proven if we can show that all its trajectories are bounded and there exists no closed trajectory. Here we have some criteria to determine the existence of closed trajectories for the system (4).

Lemma 3.1:

a) Let Q be a closed bounded region that does not contain the origin. If no trajectory leaves Q or no trajectory enters Q, then there will be a closed trajectory within Q.

b) Let Q be a simply connected region. If $(\partial f_1/\partial x_1) + (\partial f_2/\partial x_2)$ is not identically zero and does not change sign in Q, there will be no closed trajectory in Q. (Note that for f_1 and f_2 as defined by (6), $(\partial f_1/\partial x_1) + (\partial f_2/\partial x_2)$ exists inside each region of the partition of the state space Fig. 1.)

Lemma 3.1 a) is a simple application of the Poincaré–Bendixon Theorem to the system (4) and its time reversed system $\dot{x} = -A\sigma(x)$. In addition, b) follows from the Bendixon Theorem. It can also be easily obtained from Green's Theorem. This theorem will be frequently applied in this paper.

If the system (4) has a closed trajectory, say, Γ, then Γ must enclose the origin (by the index theory) since the origin is the unique equilibrium point. And Γ must be symmetric to the origin. Since the vector field is symmetric to the origin, $-\Gamma$ is also a closed trajectory. If Γ is not symmetric, then $\Gamma \neq -\Gamma$ and the two different trajectories will have intersections. This is impossible since no trajectories can intersect.

We next digress to address two special cases, $a_{11} = a_{12}$ and $a_{21} \leq a_{22}$.

Proposition 3.1: Assume A is Hurwitz.

a) If $a_{11} = a_{12}$, then the system (6) is GAS.

b) If $a_{21} \leq a_{22}$ (which implies that $a_{11} \neq a_{12}$), then the system has diverging trajectories and also has a closed trajectory.

Proof:

a) We see that a) is an extension of b) in Remark 2.1 and can be proven by a method similar to [1]. First we claim that

the vertical strip $H = \{x: |x_1| \leq 1\}$ is an invariant set and a global attractor. Since A is Hurwitz (det $A > 0$) and $a_{11} = a_{12}$, we must have $a_{21} > a_{22}$.

On the line $x_1 = 1$, $\dot{x}_1 \leq 0$ and on $x_1 = -1$, $\dot{x}_1 \geq 0$. So no trajectory in H points out of it and hence it is an invariant set. In the region V, $\dot{x}_1 = 0$, $\dot{x}_2 = -a_{21} + a_{22} < 0$, so all the trajectories in this region will enter W. In W, $\dot{x}_1 \leq 0$, $\dot{x}_2 < 0$, so all the trajectories will enter the central square or the region R. In R, $\dot{x}_1 = -a_{11} - a_{12} < 0$, $\dot{x}_2 = -a_{21} - a_{22} < 0$, so all the trajectories will enter the region $-U$. Similar arguments apply to the regions $-V$, $-W$, $-R$. This shows that all the trajectories outside of the strip H will enter it. Hence, it is a global attractor.

Next we show that all the trajectories in H are bounded. Let $p_1 = \begin{bmatrix} -1 \\ h \end{bmatrix}$, $h \geq 1$ be a point on the line $x_1 = -1$. Then $\eta(p_1) = \beta > 0$. In the region U, $\eta(x)$ depends only on x_1 and it can be easily verified that $\eta(x)$ is a decreasing function of x_1 and $\dot{x}_1 > 0$ for $x_1 \in [-1, 1)$. So if we draw a straight line E with slope β at p_1, then no trajectory in U will cross E upward. Symmetrically, no trajectory will cross $-E$ downward. This shows that the parallelogram enclosed by E, $-E$ and $x_1 = \pm 1$, denoted as P, is also an invariant set and $\psi(t, x_0) \in P$ for all $t \geq 0$ as long as $x_0 \in P$. Since for every $x_0 \in U$ there exists such a parallelogram that encloses x_0, it follows that all the trajectories are bounded.

In H, $(\partial f_1/\partial x_1) + (\partial f_2/\partial x_2) < 0$ ($= -a_{11}$ in U, $-U$ or $-a_{11} + a_{22}$ in the central square) so by Lemma 3.1b), there exists no closed trajectory in H. Since H is a global attractor, all the trajectories will enter it and then converge to the origin. Thus, the system is GAS.

b) From $0 \leq a_{21} \leq a_{22}$ we have $a_{22} > 0$, otherwise $\det(A) = 0$. We also have $\alpha \geq 0$, $\|\dot{x}\| > 0$, and $\angle \dot{x} \in [0, \pi/2]$ in the region V. So every trajectory starting from within this region will diverge along a straight line with slope α and is unbounded. Let Q_1 be the polygon with vertices 1, 2, $\cdots$, 8 (see Fig. 2, where at point 2 $x = \begin{bmatrix} 1 \\ a_{21}/a_{22} \end{bmatrix}$ and the line from 3 to 4 has slope β). From 2 to 3 $\dot{x}_2 = 0$ and $\dot{x}_1 > 0$. From 3 to 4, $\angle \dot{x} \geq \tan^{-1} \beta - \pi$, so the trajectories direct outward from Q_1. It is also easy to see that on other parts of the boundary of Q_1, all the trajectories remain on it or direct outward from it. Since A is Hurwitz, there exists a Lyapunov level set Q_0 in the central square such that all the trajectories inside Q_0 will stay inside and converge to the origin. Let $Q = Q_1 \setminus Q_0$, then no trajectory will enter Q, so by Lemma 3.1, there is a closed trajectory in Q. □

Now that the two special cases are cleared, we now turn to the remaining case where $a_{11} < a_{12}$ and $a_{21} > a_{22}$. For this case, all the trajectories go clockwise, see Fig. 1 for some typical trajectories. Here we summarize the properties of the trajectories as follows.

In the region U, $\dot{x}_1 > 0$ and the trajectories go rightward. If $x_1 < a_{22}/a_{21}$, $\dot{x}_2 > 0$ and if $x_1 > a_{22}/a_{21}$, $\dot{x}_2 < 0$. On the line $x_1 = a_{22}/a_{21}$, the trajectories turn from upward to downward.

In the region W, $\dot{x}_2 < 0$ and the trajectories go downward. On the line $x_2 = a_{11}/a_{12}$, the trajectories turn from rightward to leftward.

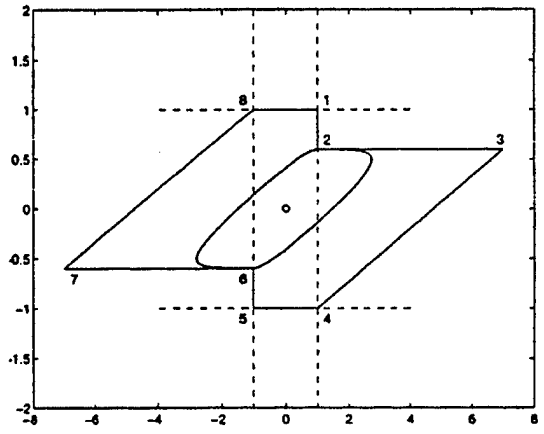


Fig. 2. Illustration for the proof of Proposition 3.1.

In the region V, the trajectories are straight lines that go downward-rightward; In the region R, the trajectories go downward-leftward.

In the central square, on the line $x_2 = (a_{11}/a_{12})x_1$, $\dot{x}_1 = 0$ and on the line $x_2 = (a_{22}/a_{21})x_1$, $\dot{x}_2 = 0$.

Finally, in this case, h_1, h_2, α and β are all well defined.

IV. Conditions for the Global Boundedness of the Trajectories

In this section, we consider the system

$$\dot{x} = A\sigma(x) = \begin{bmatrix} -a_{11} & a_{12} \\ -a_{21} & a_{22} \end{bmatrix} \sigma(x)$$
$$a_{11}, a_{21}, a_{12} > 0, \quad a_{22} \geq 0. \tag{15}$$

Assume that $a_{11} < a_{12}$ and $a_{21} > a_{22}$ (this implies $\det(A) \neq 0$.) We do not assume that A is Hurwitz in this section since the critical case where A has a pair of pure imaginary eigenvalues will be useful to our study. It turns out that the system can have a bounded global attractor, even if A is unstable. The global boundedness depends on β/α, h_1 and h_2, rather than the stability of A.

Proposition 4.1: Assume $a_{11} < a_{12}$ and $a_{21} > a_{22}$. The system (15) has a bounded global attractor if and only if one of the following conditions is satisfied.

a) $a_{11}a_{21} > a_{12}a_{22}$.

b) $a_{11}a_{21} = a_{12}a_{22}$ and $\beta h_1 + h_2 < 0$.

If $a_{11}a_{21} = a_{12}a_{22}$ and $\beta h_1 + h_2 = 0$, then outside certain region, all the trajectories are closed. If $a_{11}a_{21} < a_{12}a_{22}$ (or $a_{11}a_{21} = a_{12}a_{22}$ and $\beta h_1 + h_2 > 0$), there will be unbounded trajectories and if, in addition, A is Hurwitz, there exists a closed trajectory.

Proof: Under the assumption that $a_{11} < a_{12}$ and $a_{21} > a_{22}$, we have $\alpha < 0$, $\beta > 0$, $|h_1|$, $|h_2| < \infty$.

Let

$$p_1 = \begin{bmatrix} 1 \\ u_k + 1 \end{bmatrix}, \quad u_k \geq \max\left(0, \frac{\alpha}{\beta}(\beta h_1 + h_2), \alpha h_1\right)$$

be a point on the line $x_1 = 1$. See the point labeled 1 in Fig. 3. Let the trajectory starting from p_1 be $\psi(t, p_1)$. We will show later that $\psi(t, p_1)$ will go through regions V, W, R, and $-U$ consecutively (not fall into the central square before leaving

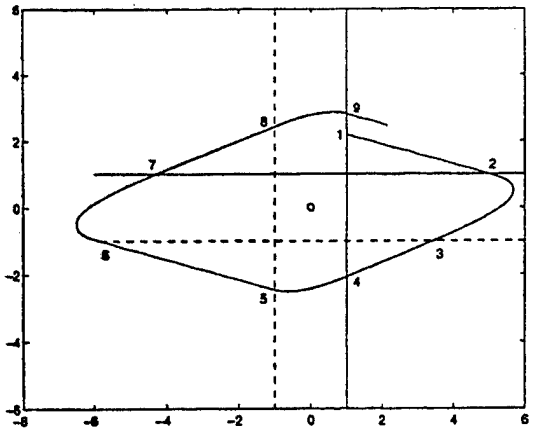


Fig. 3. Illustration for the Proof of Proposition 4.1.

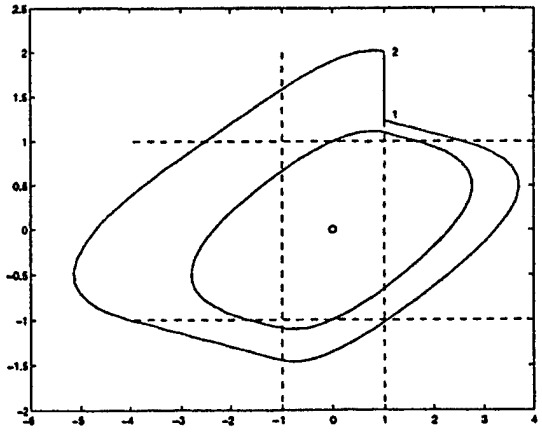


Fig. 4. Illustration for the Proof of Proposition 4.1: Case 1.

$-U$). Let the intersections of $\psi(t, p_1)$ with the lines $x_2 = 1$, $x_2 = -1$, $x_1 = 1$, $x_1 = -1$ be

$$\begin{bmatrix} 1+v_k \\ 1 \end{bmatrix}, \begin{bmatrix} 1+w_k \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1-r_k \end{bmatrix} \text{ and } \begin{bmatrix} -1 \\ -1-u_{k+1} \end{bmatrix}$$

which correspond to the points 2, 3, 4, 5 in Fig. 3. Then

$$\begin{aligned} v_k &= -\frac{1}{\alpha}u_k \\ w_k &= v_k + h_1 = -\frac{1}{\alpha}u_k + h_1 \\ r_k &= \beta w_k = -\frac{\beta}{\alpha}u_k + \beta h_1 \end{aligned}$$

and $u_{k+1} = r_k + h_2$, i.e.,

$$u_{k+1} = -\frac{\beta}{\alpha}u_k + \beta h_1 + h_2. \tag{16}$$

The requirement that the trajectory does not enter the central square is equivalent to $v_k, w_k, r_k, u_{k+1} \geq 0$. This can be guaranteed by $u_k \geq \max\left(0, (\alpha/\beta)(\beta h_1 + h_2), \alpha h_1\right)$. If we also have $u_{k+1} \geq \max\left(0, (\alpha/\beta)(\beta h_1 + h_2), \alpha h_1\right)$, then we can continue with the above process symmetrically to get an intersection with the line $x_1 = 1$, $\begin{bmatrix} 1 \\ 1+u_{k+2} \end{bmatrix}$ (point 9 in Fig. 3) where

$$u_{k+2} = -\frac{\beta}{\alpha}u_{k+1} + \beta h_1 + h_2$$

and so on. Equation (16) defines a first-order linear time invariant discrete-time system.

Case 1 ($a_{11}a_{21} < a_{12}a_{22}$): This inequality is equivalent to $-(\beta/\alpha) > 1$. So in this case, the discrete-time system (16) is unstable. If $u_k > \max\left(0, (\alpha/(\alpha+\beta))(\beta h_1 + h_2), (\alpha/\beta)(\beta h_1 + h_2), \alpha h_1\right)$, then $u_{k+1} > u_k$, $u_{k+2} > u_{k+1}, \cdots$ will be an exponentially increasing sequence and the trajectory starting from $\begin{bmatrix} 1 \\ u_k+1 \end{bmatrix}$ will be unbounded.

Let $p_1 = \begin{bmatrix} 1 \\ 1+u_1 \end{bmatrix}$ where $u_1 > \max\left(0, (\alpha/(\alpha+\beta))(\beta h_1 + h_2), (\alpha/\beta)(\beta h_1 + h_2), \alpha h_1\right)$ (see point 1 in Fig. 4). Then by the foregoing argument, $\psi(t, p_1)$ will return to the line $x_1 = 1$ at a point above p_1 (see 2 in Fig. 4). By connecting 1 and 2, we get a closed curve. Let the region enclosed by this closed curve be Q_1. From 1 to 2, $\dot{x}$ is a constant and $\angle\dot{x} \in (-(\pi/2), 0)$ since $\alpha < 0$. So $\dot{x}$ directs outward from Q_1. If A is Hurwitz, there will be a Lyapunov level set Q_0 in the central square which is invariant. Let $Q = Q_1 \setminus Q_0$, then no trajectory will enter Q and, by Lemma 3.1, there is a closed trajectory in Q as illustrated in Fig. 4.

Case 2 ($a_{11}a_{21} > a_{12}a_{22}$): This inequality is equivalent to $-(\beta/\alpha) < 1$. Since $-(\beta/\alpha) > 0$, in this case the discrete-time system (16) is stable. Let u^* be chosen such that

$$\begin{aligned} &\min\left(-\frac{\beta}{\alpha}u^* + \beta h_1 + h_2,\ u^*\right) \\ &\quad > \max\left(0, \frac{\alpha}{\beta}(\beta h_1 + h_2), \alpha h_1\right) \end{aligned} \tag{17}$$

and

$$\left(\left(\frac{\beta}{\alpha}\right)^2 - 1\right)u^* + \left(1 - \frac{\beta}{\alpha}\right)(\beta h_1 + h_2) < -1 \tag{18}$$

then with $u_k \geq u^*$, the trajectory $\psi(t, \begin{bmatrix} 1 \\ u_k+1 \end{bmatrix})$ does not fall into the central square before it returns to the line $x_1 = 1$ (This is guaranteed by (17). Moreover, because of (18), we have

$$\begin{aligned} u_{k+2} - u_k &= \left(\left(\frac{\beta}{\alpha}\right)^2 - 1\right)u_k + \left(1 - \frac{\beta}{\alpha}\right)(\beta h_1 + h_2) \\ &\leq \left(\left(\frac{\beta}{\alpha}\right)^2 - 1\right)u^* + \left(1 - \frac{\beta}{\alpha}\right)(\beta h_1 + h_2) \\ &< -1. \end{aligned} \tag{19}$$

Let $p_1 = \begin{bmatrix} 1 \\ 1+u^* \end{bmatrix}$ (see point 1 in Fig. 5). Then by the foregoing argument, $\psi(t, p_1)$ will return to the line $x_1 = 1$ at a point 2 between p_1 and $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. By connecting 1 and 2, we get a closed curve. Denote the region enclosed by this closed curve as Q_1. Since on the line between 1 and 2 $\dot{x}$ directs inward of Q_1. Thus, $\psi(t, x_0)$ will stay in Q_1 as long as $x_0 \in Q_1$. Therefore, Q_1 is an invariant set.

Let x_0 be any point outside of Q_1, then $\psi(t, x_0)$ goes clockwise and will intersect with the line $x_1 = 1$ above 2, say at $p = \begin{bmatrix} 1 \\ 1+u_0 \end{bmatrix}$. If $u_0 < u^*$, i.e., p is between 1 and 2, then $\psi(t, x_0)$ will enter Q_1 afterward and stay there. If $u_0 \geq u^*$, then by (19), we have $u_2 < u_0 - 1$, $u_4 < u_2 - 1, \cdots$ until $u_k < u^*$ for some finite k. This implies $\psi(t_1, x_0) \in Q_1$ for some $t_1 > 0$. Therefore, Q_1 is a global attractor.

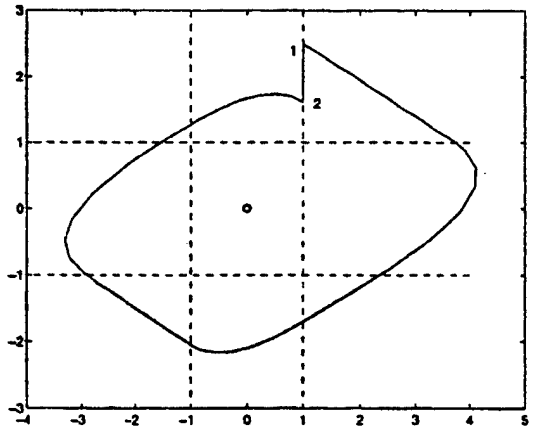


Fig. 5. Illustration for the Proof of Proposition 4.1: Case 2.

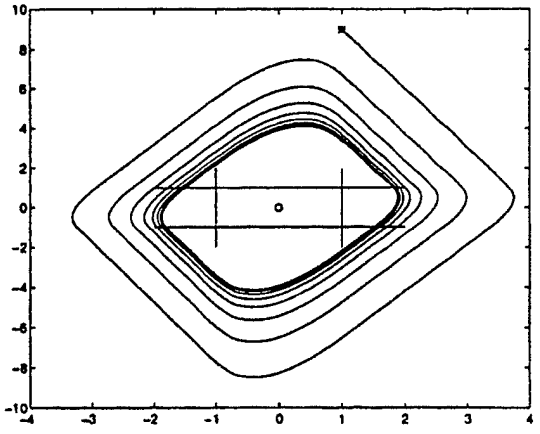


Fig. 6. A global attractor.

Case 3 $a_{11}a_{21} = a_{12}a_{22}$: In this case, $-(\beta/\alpha) = 1$ and

$$u_{k+2} - u_k = 2(\beta h_1 + h_2).$$

Suppose $\beta h_1 + h_2 < 0$, the sequence u_k, u_{k+2}, $\cdots$ will decrease steadily before the trajectory touches the central square, similar to Case 2, there exists a global attractor.

Suppose $\beta h_1 + h_2 > 0$, the sequence will increase steadily and the trajectory will go unbounded. Also, similarly to Case 2, there exists a closed trajectory.

Suppose $\beta h_1 + h_2 = 0$, then if $\geq \max(0, (\alpha/\beta)(\beta h_1 + h_2), \alpha h_1)$, we will have $u_k = u_{k+1} = \cdots = u_{k+N}$ for all N. So $\begin{bmatrix} 1 \\ 1+u_k \end{bmatrix}$ is on a closed trajectory. Let Q_1 be the region enclosed by the closed trajectory passing through $\begin{bmatrix} 1 \\ 1+u_k \end{bmatrix}$, then all the trajectories outside of Q_1 are closed. □.

To demonstrate Proposition 4.1, consider the system with $A = \begin{bmatrix} -1 & 2 \\ -5 & 2 \end{bmatrix}$. Clearly, A is exponentially unstable, but $a_{11}a_{21} > a_{12}a_{22}$. So the system has a global attractor (see Fig. 6).

An interesting case is that $a_{11} = a_{22}$ and $a_{21} = a_{12}$. In this case, A has a pair of pure imaginary eigenvalues. For the linear system $\dot{x} = Ax$ every point in the plane is on a closed trajectory. This is also true for the saturated system

$$\dot{x} = A\sigma(x) = \begin{bmatrix} -a_{11} & a_{12} \\ -a_{12} & a_{11} \end{bmatrix} \sigma(x), \quad a_{12} > a_{11} > 0. \quad (20)$$

Denote the trajectory of (20) as $\psi_1(t, x_0)$.

Proposition 4.2: All the trajectories of (20) are closed. Each trajectory is symmetric with respect to the line $x_1 = x_2$ and the line $x_1 = -x_2$.

Proof: For this system, $a_{11}a_{21} = a_{12}a_{22}$ and it can also be verified that $\beta h_1 + h_2 = 0$. By Proposition 4.1, $\psi_1(t, x_0)$ is bounded for every x_0. On the other hand, since A has a pair of pure imaginary eigenvalues, there are closed trajectories in any neighborhood of the origin. Thus, any x_0 is outside of a closed trajectory. Therefore, $\psi_1(t, x_0)$ will be a closed curve or go to a closed curve. Since $\psi_1(t, x_0)$ goes clockwise, it will intersect the line $x_1 = x_2$ somewhere, say, at $\begin{bmatrix} r \\ r \end{bmatrix}$. So, for simplicity, we can assume that $x_0 = \begin{bmatrix} r \\ r \end{bmatrix}$ for some $r > 0$. To show $\psi_1(t, x_0)$ is a closed trajectory and is symmetric to the line $x_1 = x_2$, it suffices to show that $\psi_1(t, x_0) = J\psi_1(-t, x_0)$ where $J = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

Consider the time-reversed system of (20),

$$\dot{z} = -A\sigma(z). \quad (21)$$

Denote its trajectory as $\phi(t, x_0)$. Then $\psi_1(-t, x_0) = \phi(t, x_0)$. From (21), we have

$$J\dot{z} = -JA\sigma(z) = -JAJ\sigma(Jz) = A\sigma(Jz)$$

thus, $J\phi(t, x_0) = \psi_1(t, Jx_0)$. Since $Jx_0 = x_0$, it follows that $\psi_1(t, x_0) = J\psi_1(-t, x_0)$.

To show that the trajectory is symmetric with respect to $x_1 = -x_2$, we write (20) as

$$\begin{bmatrix} \dot{x}_1 \\ -\dot{x}_2 \end{bmatrix} = \begin{bmatrix} -a_{11} & -a_{12} \\ a_{12} & a_{11} \end{bmatrix} \begin{bmatrix} \sigma(x_1) \\ \sigma(-x_2) \end{bmatrix} = A^T \begin{bmatrix} \sigma(x_1) \\ \sigma(-x_2) \end{bmatrix}.$$

We also have $-JA^TJ = A^T$. Following the same procedure as above by considering the state $\begin{bmatrix} x_1 \\ -x_2 \end{bmatrix}$ instead, we can show that all the trajectories are symmetric with respect to the line $x_1 = -x_2$. □

The system (20) is not GAS but is useful for us to develop the condition for global asymptotic stability. We will establish the main result of the paper by comparing the trajectory of a general system with that of (20).

V. Proof of Theorem 2.1

In view of Remark 2.1 and Proposition 3.1, we only need to consider the following system:

$$\dot{x} = A\sigma(x) = \begin{bmatrix} -a_{11} & a_{12} \\ -a_{21} & a_{22} \end{bmatrix} \sigma(x)$$
$$a_{11}, a_{12}, a_{21} > 0, a_{22} \geq 0 \quad (22)$$

with $a_{12} > a_{11}$, $a_{21} > a_{22}$.

Proposition 5.1: Assume that $a_{12} > a_{11}$ and A is Hurwitz, the system (22) is GAS if and only if $a_{11}a_{21} \geq a_{12}a_{22}$.

This proposition can be established as follows. First note that $a_{12} > a_{11}$ and $a_{11}a_{21} \geq a_{12}a_{22}$ imply $a_{21} > a_{22}$. If $a_{11}a_{21} = a_{12}a_{22}$ and A is Hurwitz, then it can be verified that $\beta h_1 + h_2 < 0$. Hence, this proposition shows that the stability of A

along with the global boundedness condition in Proposition 4.1 guarantees the system to be GAS.

If $a_{11}a_{21} < a_{12}a_{22}$, then by Propositions 3.1 and 4.1, the system is not GAS whether $a_{21} > a_{22}$ or not. So the necessity of the condition is obvious. What remains is to show the sufficiency of the condition. Now that the global boundedness of the trajectories is guaranteed, the only thing needed to be shown is that the system has no closed trajectory.

Since all the trajectories are kept unchanged when the vector field is multiplied by a positive constant, we assume that $a_{11} = 1$ in the sequel for simplicity. Now we have

$$A = \begin{bmatrix} -1 & a_{12} \\ -a_{21} & a_{22} \end{bmatrix}.$$

We first deal with the case where $a_{22} = 0$.

Lemma 5.1: Assume $a_{12} > a_{11}$ and A is Hurwitz. If $a_{22} = 0$, then (22) is GAS.

Proof: See the Appendix. □

In what follows, we consider the case that $a_{22} > 0$. Let $k = a_{21}/a_{12}a_{22}$, then we can assume that A takes the form

$$A = \begin{bmatrix} -1 & a_{12} \\ -ka_{12}a_{22} & a_{22} \end{bmatrix}, \quad k > 0,\ a_{22} > 0. \tag{23}$$

The assumption in Proposition 5.1 that $a_{11} < a_{12}$ and A is Hurwitz translates to

$$a_{12} > 1, \qquad a_{22} < 1, \qquad ka_{12}^2 > 1$$

and the condition $a_{11}a_{21} \geq a_{12}a_{22}$ is equivalent to

$$k \geq 1.$$

Therefore, we can establish Proposition 5.1 by showing that the system

$$\dot{x} = A\sigma(x) = \begin{bmatrix} -1 & a_{12} \\ -ka_{12}a_{22} & a_{22} \end{bmatrix} \sigma(x)$$
$$a_{12} > 1,\ \ k \geq 1,\ \ 0 < a_{22} < 1 \tag{24}$$

is GAS. The proof will be carried out by evolving A from the simplest form where $a_{22} = 1$, $k = 1$ to the case $a_{22} = 1$, $k \geq 1$ and finally to the general case $0 < a_{22} < 1$, $k \geq 1$. When $a_{22} = 1$, the system is surely not GAS because A is not Hurwitz, but the trajectories in this case will be used as a reference to show the convergence of the trajectories when a_{22} is decreased.

To proceed, we need a technical lemma. Recall (13)

$$h_1 = \begin{cases} -\dfrac{2a_{12}}{a_{22}} + \dfrac{\det A}{a_{22}^2} \log \dfrac{a_{21} + a_{22}}{a_{21} - a_{22}}, & \text{if } a_{22} > 0 \\ -\dfrac{2a_{11}}{a_{21}}, & \text{if } a_{22} = 0. \end{cases}$$

Now we have $a_{22} > 0$ and $a_{21} = ka_{12}a_{22}$, so

$$h_1 = \frac{1}{a_{22}} \left(-2a_{12} + (ka_{12}^2 - 1) \log \frac{ka_{12} + 1}{ka_{12} - 1} \right).$$

Lemma 5.2: If $a_{22} > 0$, $a_{12} > 1$ and $k \geq 1$, then $h_1 < 0$.

Proof: See the Appendix. □

Now we consider the case where $a_{22} = 1$ and $k \geq 1$,

$$\dot{x} = \begin{bmatrix} -1 & a_{12} \\ -ka_{12} & 1 \end{bmatrix} \sigma(x), \quad a_{12} > 1,\ \ k \geq 1. \tag{25}$$

Given an initial point x_0, denote the trajectory of (25) as $\psi_2(t, x_0)$ and as a comparison, denote the trajectory of

$$\dot{x} = \begin{bmatrix} -1 & a_{12} \\ -a_{12} & 1 \end{bmatrix} \sigma(x) \tag{26}$$

as $\psi_1(t, x_0)$. Then $\psi_1(t, x_0)$ is closed for every x_0 by Proposition 4.2.

In the following, we present three lemmas about the intersections of $\psi_2(t, x_0)$ with some straight lines.

Lemma 5.3: Assume $k > 1$.

a) Let $x_0 = \begin{bmatrix} x_{01} \\ 1 \end{bmatrix}$. If $x_{01} \in (1/a_{12}, 1]$, then $\psi_2(t, x_0)$ (see the dashed curve in Fig. 7) may intersect with the line $x_1 = 1$. Let the first intersection be $x_c = \begin{bmatrix} 1 \\ x_{c2} \end{bmatrix}$, then $x_{c2} < x_{01}$, i.e,

$$\left\| x_0 - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\| < \left\| x_c - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\|.$$

If $x_{01} \in [1/ka_{12}, 1/a_{12}]$, $\psi_2(t, x_0)$ will go downward-rightward at first, but will not intersect with the line $x_1 = 1$ before it turns leftward (see the dotted curve in Fig. 7).

b) Let $x_0 = \begin{bmatrix} 1 \\ x_{02} \end{bmatrix}$, $x_{02} < 1/a_{12}$. Then $\psi_2(t, x_0)$ goes downward-leftward (see the dash-dotted curve in Fig. 7). Let the first intersection of $\psi_2(t, x_0)$ with the line $x_2 = -1$ be $x_c = \begin{bmatrix} x_{c1} \\ -1 \end{bmatrix}$, then $x_{02} > -x_{c1}$, i.e.,

$$\left\| x_0 - \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\| > \left\| x_c - \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\|.$$

As a comparison, two $\psi_1(t, x_0)$ are also shown in Fig. 7 (see the solid curves). In Fig. 7, x_0's are marked with *.

Proof: See the Appendix. □

Lemma 5.4: Given $\gamma \geq 0$, let $x_0 = \begin{bmatrix} 1+\gamma \\ s_1 \end{bmatrix}$, $s_1 \in (1/a_{12}, 1]$ be a point on the line $x_1 = 1 + \gamma$. Then $\psi_2(t, x_0)$ will go downward-rightward at first, then turn leftward and return to the line $x_1 = 1 + \gamma$. Let the intersection be $x_c = \begin{bmatrix} 1+\gamma \\ s_2 \end{bmatrix}$, then $s_1 + s_2 > 0$.

Proof: See the Appendix. □

Lemma 5.5: Let $x_0 = \begin{bmatrix} x_{01} \\ 1 \end{bmatrix}$, $x_{01} > 1/ka_{12}$ be a point on the line $x_2 = 1$. Then $\psi_2(t, x_0)$ goes downward-rightward at first and turns leftward. Suppose $\psi_2(t, x_0)$ has an intersection with the line $x_2 = -1$ at $x_c = \begin{bmatrix} x_{c1} \\ -1 \end{bmatrix}$, then $x_{c1} < x_{01}$ (see Fig. 8).

Proof: See the Appendix

The following two lemmas give a complete characterization of the trajectories of the system (25).

Lemma 5.6: Assume $k > 1$. Let $x^* = \begin{bmatrix} 1/ka_{12} \\ 1 \end{bmatrix}$. Then, $\psi_2(t, x^*)$ is a closed curve that lies within the central square. Denote the region enclosed by $\psi_2(t, x^*)$ as S_0, then every point inside S_0 is on a closed trajectory. And outside S_0, any trajectory will converge to $\psi_2(t, x^*)$ (see Fig. 9).

Proof: See the Appendix. □

Lemma 5.7: Assume $k > 1$. Let $x_0 = \begin{bmatrix} x_{01} \\ 1 \end{bmatrix}$, $x_{01} < 1/ka_{12}$ be a point on the line $x_2 = 1$, then $\psi_2(t, x_0)$ goes upward at first and will return to the line $x_2 = 1$. Suppose $\psi_2(t, x_0)$ intersects

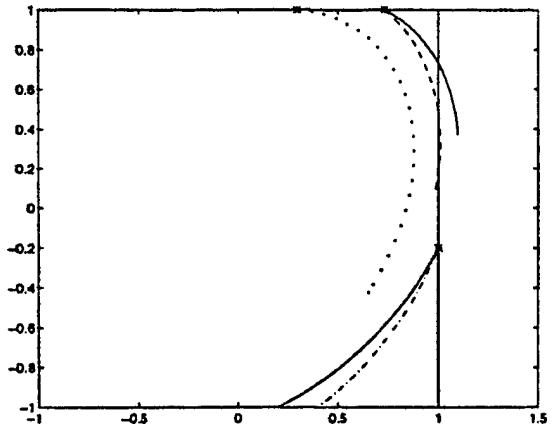


Fig. 7. Illustration for Lemma 5.3.

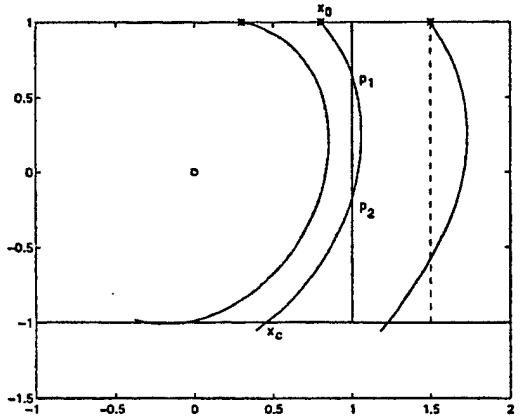


Fig. 8. Illustration for Lemma 5.5.

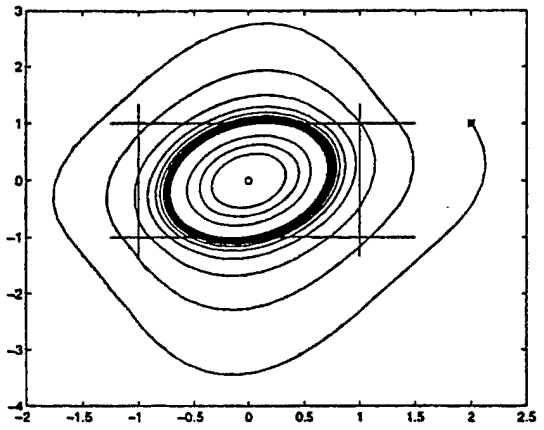


Fig. 9. Illustration for Lemma 5.6.

with the line $x_2 = -1$. Let $x_c = \begin{bmatrix} x_{c1} \\ -1 \end{bmatrix}$ be the first intersection, then $-(1/ka_{12}) < x_{c1} < -x_{01}$.

Proof: See the Appendix. □

Lemmas 5.6 and 5.7 give us a clear picture of the trajectories of (25), where $a_{22} = 1$, $k > 1$. Lemma 5.7 shows that if x_0 is outside of S_0, a trajectory $\psi_2(t, x_0)$ will move closer and closer to S_0 as it reaches the lines $x_2 = \pm 1$. Next we will show that as a_{22} is decreased, a trajectory $\psi(t, x_0)$ of (24) will move even closer to S_0, as compared with $\psi_2(t, x_0)$. This will lead to our final result about the global asymptotic stability of the system (24) and hence the proof of Proposition 5.1. Rewrite (24) as follows:

$$\begin{aligned} \dot{x}_1 &= -\sigma(x_1) + a_{12}\sigma(x_2) \\ \dot{x}_2 &= a_{22}(-ka_{12}\sigma(x_1) + \sigma(x_2)) \end{aligned} \tag{27}$$

where $a_{12} > 1$, $k \geq 1$ and $0 < a_{22} < 1$. We will consider the perturbation of the trajectories as a_{22} is varied, so denote the trajectory of (27) as $\psi(t, x_0, a_{22})$ and the slope of a trajectory at x be $\eta(x, a_{22})$. As compared with (25), $\dot{x}_1$ is the same but $\dot{x}_2$ is multiplied with a scalar a_{22}. Because of this, the trajectories of (27) exhibit some interesting properties.

Fact 5.1:

a) Let $x_0 = \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix}$, $x_{02} \geq 1$ be a point above the line $x_2 = 1$, then for all $a_{22} > 0$

$$[0 \quad 1]\psi(t, x_0, a_{22}) - x_{02} = a_{22}([0 \quad 1]\psi_2(t, x_0) - x_{02})$$

$$[1 \quad 0]\psi(t, x_0, a_{22}) = [1 \quad 0]\psi_2(t, x_0)$$

as long as $\psi(t, x_0, a_{22})$ stays above the line $x_2 = 1$.

b) Let $x_0 = \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix}$, $x_{01} \geq 1$ be a point to the right of the line $x_1 = 1$, then for all $a_{22} > 0$

$$[1 \quad 0]\psi(t, x_0, a_{22}) - x_{01} = \frac{1}{a_{22}}([1 \quad 0]\psi_2(a_{22}t, x_0) - x_{01})$$

$$[0 \quad 1]\psi(t, x_0, a_{22}) = [0 \quad 1]\psi_2(a_{22}t, x_0)$$

as long as $\psi(t, x_0, a_{22})$ stays to the right of $x_1 = 1$.

See Fig. 10 for an illustration, where the solid curves are $\psi_2(t, x_0)$ and $\psi_2(a_{22}t, x_0)$, and the dashed curves are $\psi(t, x_0, a_{22})$, $a_{22} < 1$.

Fact 5.1a) implies that $\psi(t, x_0, a_{22})$ and $\psi_2(t, x_0)$ are on the same vertical line but the distance from $\psi(t, x_0, a_{22})$ to the line $x_2 = x_{02}$ is a_{22} times that from $\psi_2(t, x_0)$ to $x_2 = x_{02}$. In particular, $\psi_2(t, x_0)$ and $\psi(t, x_0, a_{22})$ return to the line $x_2 = x_{02}$ at the same time and the same point. This simply follows from the fact that $\dot{x}_2$ of (27) is a_{22} times that of (25) and that $\dot{x}$ is independent of x_2 above the line $x_2 = 1$. It can also be directly verified from the expression of $\psi(t, x_0, a_{22})$ and $\psi_2(t, x_0)$.

Fact 5.1b) implies that $\psi(t, x_0, a_{22})$ and $\psi_2(a_{22}t, x_0)$ are on the same horizontal line but the distance from $\psi(t, x_0, a_{22})$ to the line $x_1 = x_{01}$ is $1/a_{22}$ times that from $\psi_2(a_{22}t, x_0)$ to $x_1 = x_{01}$. In particular, $\psi_2(a_{22}t, x_0)$ and $\psi(t, x_0, a_{22})$ return to the line $x_1 = x_{01}$ at the same time and the same point. This also follows from the fact that $\dot{x}_2$ is scaled by a_{22}. If we scale the vector field to the right of the line $x_1 = 1$ by $1/a_{22}$, then $\dot{x}_2$ is the same as that of (25) but $\dot{x}_1$ is amplified by $1/a_{22}$. Note that the scaling of the vector field results in the time scaling of $\psi_2(a_{22}t, x_0)$.

With Fact 5.1, we are ready to present a final lemma that leads to the proof of Proposition 5.1.

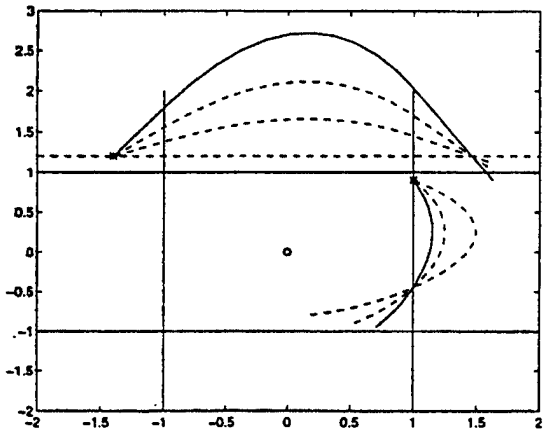


Fig. 10. Illustration for Fact 5.1.

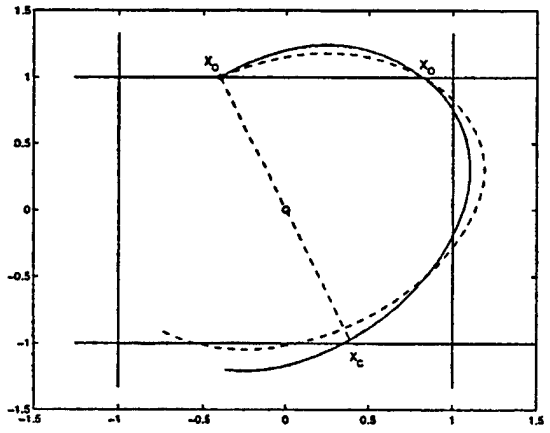


Fig. 11. Illustration for the Proof of Proposition 5.1.

Lemma 5.8: Let $x_0 = \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix}$, $x_{02} \in (0, 1]$, $x_{01} \geq x_{02}/ka_{12}$. Suppose $\psi(t, x_0, a_{22})$ intersects with the line $x_2 = -x_{02}$ at a point $x_c = \begin{bmatrix} x_{01} \\ -x_{02} \end{bmatrix}$ and $x_{c1} < x_{01}$, then if $\delta \in (0, a_{22})$ is sufficiently small, $\psi(t, x_0, a_{22} - \delta)$ will intersect with $\psi(t, x_0, a_{22})$ at a point to the right of x_0. If it also intersects with the line $x_2 = -x_{02}$, the intersection will be to the left of x_c.

For an illustration of Lemma 5.8, see Figs. 16–18 where the solid curves are $\psi(t, x_0, a_{22})$ and the dashed curves are $\psi(t, x_0, a_{22} - \delta)$.

Proof: See the Appendix. □

Proof of Proposition 5.1: The necessity of the condition simply follows from Propositions 3.1 and 4.1. With Lemma 5.1, it remains to be shown that the system (27) or (24) is GAS. We will first show that any point on the line $x_2 = 1$ is not on a closed trajectory. We can restrict our attention to the points to the left of $\begin{bmatrix} 1/ka_{12} \\ 1 \end{bmatrix}$, since for the points to its right, they can be traced back to the left as the trajectories go rightward above the line $x_2 = 1$. Let $x_0 = \begin{bmatrix} x_{01} \\ 1 \end{bmatrix}$, $x_{01} < 1/ka_{12}$, then $\psi_2(t, x_0)$ of the system (25) (see the solid curve in Fig. 11) will return to the line $x_2 = 1$ at a point x_0'. From Fact 5.1, $\psi(t, x_0, a_{22})$ will also return to x_0' for all $a_{22} > 0$ (see the dashed curve in Fig. 11).

We have shown in Lemma 5.7 that for any x_0 to the left of $\begin{bmatrix} 1/ka_{12} \\ 1 \end{bmatrix}$, if $\psi_2(t, x_0)$ reaches the line $x_2 = -1$ at some point

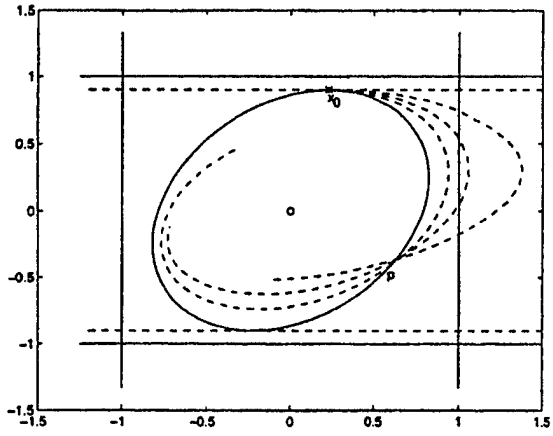


Fig. 12. Illustration for the Proof of Proposition 5.1.

$x_c = \begin{bmatrix} x_{c1} \\ -1 \end{bmatrix}$, then $x_{c1} < -x_{01}$, i.e., x_c is to the left of $-x_0$. By Lemma 5.5, x_c is also to the left of x_0'. From Lemma 5.8, we know that as a_{22} is decreased from 1, the intersection of $\psi(t, x_0', a_{22})$ and $x_2 = -1$ will move leftward and, hence, remain to the left of $-x_0$ and x_0'. Note that x_0' is on $\psi(t, x_0, a_{22})$, so $\psi(t, x_0', a_{22})$ overlaps with $\psi(t, x_0, a_{22})$. Therefore, x_0 is not on a closed trajectory (note that a closed trajectory must be symmetric).

Next we exclude the possibility of the existence of a closed trajectory that does not intersect with $x_2 = 1$. Suppose there is one, then it must intersect with the line $x_2 = ka_{12}x_1$ at some point, say $x_0 = \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix}$, $x_{02} < 1$, then at x_0, $\dot{x}_2 = 0$ and $x_0 \in S_0$. When $a_{22} = 1$, $\psi(t, x_0, a_{22}) = \psi_2(t, x_0)$ is an ellipse that touches the lines $x_2 = \pm x_{02}$ (see the solid curve in Fig. 12). By Lemma 5.8, as a_{22} is decreased to $a_{22} - \delta$, $\psi(t, x_0, a_{22} - \delta)$ will intersect with $\psi(t, x_0, a_{22})$ at a point to the right of x_0 and $-x_0$. After that, it will stay above $\psi(t, x_0, a_{22})$ and above the line $x_2 = -x_{02}$ (see the dashed curves in Fig. 12). Thus no closed trajectory can be formed.

If $k > 1$, then $a_{11}a_{21} > a_{12}a_{22}$. If $k = 1$, then $a_{11}a_{21} = a_{12}a_{22}$ and it can be verified that $\beta = a_{22}$, $\beta h_1 + h_2 = a_{22}(1 - a_{22})h_1 < 0$ (assume $a_{11} = 1$). So by Proposition 4.1, all the trajectories of (27) are bounded, and they must converge to the origin.

Proof of Theorem 2.1: Combining Remark 2.1, Propositions 3.1, 4.1, and 5.1, we can obtain the necessary and sufficient condition for the system (4) to be GAS. Condition b) in Theorem 2.1 is a simple combination of Condition b) in Remark 2.1 and Propositions 3.1, 4.1, and 5.1. This simplification is justified as follows.

Since A is Hurwitz and $a_{22} \geq 0$, we must have $a_{12} > 0$.

If $a_{12} \leq a_{11}$, Remark 2.1 b) and Proposition 3.1 say the system is GAS. If $a_{12} > a_{11}$, because $a_{11}a_{21} \geq a_{12}a_{22}$, the system is also GAS by Proposition 5.1.

Conversely, suppose $a_{22} \geq 0$ but $a_{11}a_{21} < a_{12}a_{22}$, we have

$$\frac{a_{12}}{a_{11}} > \frac{a_{21}}{a_{22}}.$$

Since A is Hurwitz

$$\frac{a_{12}}{a_{11}} > \frac{a_{22}}{a_{21}}.$$

Therefore

$$\frac{a_{12}}{a_{11}} > \max\left(\frac{a_{21}}{a_{22}}, \frac{a_{22}}{a_{21}}\right) \geq 1$$

i.e., $a_{12} > a_{11}$. Hence by Propositions 3.1 and 4.1, the system is not GAS whether $a_{21} > a_{22}$ or not. And in both cases, the system has unbounded trajectories and there is also a closed trajectory. □

VI. Conclusions

We gave a complete stability analysis of a planar linear system under saturation. The analysis involves intricate investigation on the vector field and the intersections of the trajectories with the lines $x_1 = \pm 1$ and $x_2 = \pm 1$. Our main result provides a necessary and sufficient condition for such a system to be GAS.

Appendix
Proof of Lemmas

Proof of Lemma 5.1

Under the condition, we have $a_{21} > a_{22}$, $a_{11}a_{21} > a_{12}a_{22}$. By Proposition 4.1, the system has a bounded global attractor. We need to show that there exists no closed trajectory. Suppose, on the contrary, that there is such a one. Denote the region enclosed by the closed trajectory as Q, then by Green's Theorem

$$\iint_Q \left(\frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2}\right) dx_1\, dx_2 = 0.$$

Since a closed trajectory must enclose the origin, the area of the intersection of Q and the central square is nonzero. Thus, the left-side integral is strictly smaller than zero. Note that $(\partial f_1/\partial x_1) + (\partial f_2/\partial x_2) = -1$ in the central square and nonpositive in other parts of the plane. This is a contradiction. □

Proof of Lemma 5.2

We only need to show that the lemma is true when $a_{22} = 1$. First, let $k = 1$, then

$$h_1(a_{12}) = -2a_{12} + \left(a_{12}^2 - 1\right) \log \frac{a_{12}+1}{a_{12}-1}. \tag{28}$$

As $a_{12} \to 1$, $h_1 \to -2$ and as $a_{12} \to \infty$, $h_1 \to 0$. Suppose there is an extremum a_{12}^* in the interval $(1, \infty)$, then at this extremum $dh_1/da_{12} = 0$. From routine computation, this implies

$$\log \frac{a_{12}^*+1}{a_{12}^*-1} = \frac{2}{a_{12}^*}.$$

Put this into the formula (28), we get the only possible extremum value

$$h_1(a_{12}^*) = -\frac{2}{a_{12}^*} < 0.$$

Since $h_1 \leq 0$ at the two end points of the interval $(1, \infty)$, we must have $h_1 < 0$ on the whole interval.

Next we show that for any fixed $a_{12} > 1$, $h_1 < 0$ for all $k \geq 1$. Here we have

$$h_1(k) = -2a_{12} + (ka_{12}^2 - 1) \log \frac{ka_{12}+1}{ka_{12}-1}.$$

We have just shown that for any given $a_{12} > 1$, $h_1(1) < 0$. As $k \to \infty$, we also have $h_1 \to 0$. Suppose there is an extremum k^* between $(1, \infty)$, then $dh_1/dk = 0$. This implies

$$\log \frac{k^*a_{12}+1}{k^*a_{12}-1} = \frac{2\left(k^*a_{12}^2 - 1\right)}{a_{12}\left(k^*a_{12}-1\right)\left(k^*a_{12}+1\right)}.$$

Put this into the function $h_1(k)$, we get

$$h_1(k^*) = \frac{2a_{12}^2 + 2 - 4k^*a_{12}^2}{a_{12}\left(k^*a_{12}-1\right)\left(k^*a_{12}+1\right)} < 0$$

(note that $a_{12} > 1$, $k \geq 1$). It follows that $h_1(k) < 0$ for all $k \geq 1$ and $a_{12} > 1$. □

Proof of Lemma 5.3

a) This can be shown by comparing $\psi_2(t, x_0)$ with $\psi_1(t, x_0)$. Since $\psi_1(t, x_0)$ is symmetric with respect to the line $x_1 = x_2$, it will intersect with $x_1 = 1$ at $\begin{bmatrix}1\\x_{01}\end{bmatrix}$ for any $x_{01} \in [1/a_{12}, 1]$. Since at the same point x, if $x_1 > 0$, then $\dot{x}_2$ of (25) is smaller (more negative) than that of (26) and $\dot{x}_1$ of the two is the same, so $\psi_2(t, x_0)$ is below $\psi_1(t, x_0)$. Hence, the first intersection of $\psi_2(t, x_0)$ with $x_1 = 1$, if there is one, must be below that of $\psi_1(t, x_0)$ with $x_1 = 1$, which is $\begin{bmatrix}1\\x_{01}\end{bmatrix}$. This shows that $\left\|x_0 - \begin{bmatrix}1\\1\end{bmatrix}\right\| < \left\|x_c - \begin{bmatrix}1\\1\end{bmatrix}\right\|$.

If $x_{01} \in [1/ka_{12}, 1/a_{12}]$, $\psi_2(t, x_0)$ will go downward-rightward at first and when it reaches the line $x_1 = 1/a_{12}$, it is below the point $\begin{bmatrix}1/a_{12}\\1\end{bmatrix}$. Since $\psi_1\left(t, \begin{bmatrix}1/a_{12}\\1\end{bmatrix}\right)$ does not go beyond the line $x_1 = 1$ (at the intersection $\dot{x}_1 = 0$), $\psi_2(t, x_0)$ will not intersect with the line before it turns leftward.

b) If $x_{02} < 1/a_{12}$, then $\psi_2(t, x_0)$ goes downward-leftward. Suppose $\psi_2(t, x_0)$ intersects with $x_2 = -1$ at x_c. Let the region enclosed by $\psi_2(t, x_0)$ and the two lines $x_1 = 1$, $x_2 = -1$ be S, then by Green's Theorem

$$\begin{aligned}&\oint_{\partial S} f_2(x)\, dx_1 - f_1(x)\, dx_2 \\ &= \iint_S \left(\frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2}\right) dx_1\, dx_2 \\ &= 0\end{aligned}$$

where ∂S denotes the boundary of S that goes clockwise. Note that along the trajectory $\psi_2(t, x_0)$, the integral on the boundary is zero, so we have

$$-\int_{x_{02}}^{-1} (-1 + a_{12}x_2)\, dx_2 + \int_1^{x_{c1}} (-ka_{12}x_1 - 1)\, dx_1 = 0.$$

Let

$$c = \left\| x_c - \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\| = 1 - x_{c1}$$
$$e = \left\| x_0 - \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\| = x_{02} + 1$$

then from the above equation

$$\frac{ka_{12}}{2} c^2 - (1 + ka_{12})\, c = \frac{a_{12}}{2} e^2 - (1 + a_{12})\, e. \qquad (29)$$

Since $\psi_2(t, x_0)$ intersects with $x_2 = -1$ downward, so at x_c we must have $\dot{x}_2 \le 0$. This implies x_c must be to the right of $\begin{bmatrix} -(1/ka_{12}) \\ -1 \end{bmatrix}$, i.e., $c \le 1 + 1/ka_{12}$. Similarly, $e \le 1 + 1/a_{12}$. It can be shown by the manipulation of quadratic functions that for any $e \le 1 + 1/a_{12}$ there is a unique $c \le 1 + 1/ka_{12}$ satisfying (29) and $c < e$ due to $k > 1$ and $a_{12} > 1$. □

Proof of Lemma 5.4

Since $s_1 \in (1/a_{12}, 1]$, at x_0, $\dot{x}_1 > 0$, and $\dot{x}_2 < 0$. Thus, $\psi_2(t, x_0)$ goes downward-rightward. In the region W, above the line $x_2 = 1/a_{12}$, $\dot{x}_1 > 0$ and below the line $\dot{x}_1 < 0$, so $\psi_2(t, x_0)$ turns downward-leftward on meeting this line.

By the horizontal shifting property (12) of the trajectories in the region W, it suffices to show that the lemma is true for one $\gamma > 0$. Choose γ sufficiently large such that when $\psi_2(t, x_0)$ reaches the line $x_2 = -s_1$ at $\begin{bmatrix} 1+\gamma+\Delta \\ -s_1 \end{bmatrix}$, it is still inside W, i.e., $\gamma + \Delta \ge 0$. Obviously, the quantity Δ is independent of γ like h_1. If $\Delta < 0$, then $\begin{bmatrix} 1+\gamma+\Delta \\ -s_1 \end{bmatrix}$ is to the left of the line $x_1 = 1 + \gamma$. This implies $\psi_2(t, x_0)$ must have intersected the line $x_1 = 1 + \gamma$ at $\begin{bmatrix} 1+\gamma \\ s_2 \end{bmatrix}$ before it reaches $\begin{bmatrix} 1+\gamma+\Delta \\ -s_1 \end{bmatrix}$. Note that $\psi_2(t, x_0)$ goes downward-leftward below the line $x_2 = 1/a_{12}$. Thus $s_2 > -s_1$, i.e., $s_1 + s_2 > 0$.

What remains to be shown is that $\Delta < 0$ is indeed the case. Trivial calculation shows that

$$\Delta = -2a_{12}s_1 + \left(ka_{12}^2 - 1\right) \log \frac{ka_{12} + s_1}{ka_{12} - s_1}.$$

When $s_1 = 1$ we get $\Delta = h_1$. Let $\bar{a}_{12} = a_{12}s_1$, $\bar{k} = k/s_1^2$, then

$$\Delta = -2\bar{a}_{12} + \left(\bar{k}\bar{a}_{12}^2 - 1\right) \log \frac{\bar{k}\bar{a}_{12} + 1}{\bar{k}\bar{a}_{12} - 1}.$$

This is similar to h_1. Since $s_1 \in (1/a_{12}, 1]$, we have $\bar{a}_{12} > 1$, $\bar{k} \ge 1$, so by Lemma 5.2, $\Delta < 0$. □

Proof of Lemma 5.5

When $x_{01} \ge 1$ we must have $x_{c1} < x_{01}$, otherwise we would get $h_1 \ge 0$. However, we know that $h_1 < 0$ by Lemma 5.2. What remains to be shown is the case where $x_{01} \in (1/ka_{12}, 1)$. Since x_{c1} depends continuously on x_{01}, it suffices to show that $x_{c1} \ne x_{01}$ for any $x_{01} \in (1/ka_{12}, 1)$. We prove this by contradiction.

Assume that $x_{c1} = x_{01}$ for some $x_{01} \in (1/ka_{12}, 1)$. Then the line x_0 to x_c is vertical.

Case 1: $\psi_2(t, x_0)$ does not intersect with the line $x_1 = 1$ before it reaches $x_2 = -1$. Applying Green's Theorem to the region enclosed by $\psi_2(t, x_0)$ and the vertical line x_0 to x_c. Since $\partial f_1/\partial x_1 + \partial f_2/\partial x_2 = 0$ in the region, we have

$$\oint f_2(x)\, dx_1 - f_1(x)\, dx_2 = -\int_{-1}^{1} (-x_{01} + a_{12}x_2)\, dx_2 = 0.$$

This leads to $x_{01} = 0$, which contradicts the condition that $x_{01} > 1/ka_{12}$.

Case 2: $\psi_2(t, x_0)$ intersects with the line $x_1 = 1$ before it reaches $x_2 = -1$. Let the intersections be $p_1 = \begin{bmatrix} 1 \\ s_1 \end{bmatrix}$, $p_2 = \begin{bmatrix} 1 \\ s_2 \end{bmatrix}$, see Fig. 8. Again applying Green's Theorem to the region enclosed by the line x_c to x_0, the line p_1 to p_2 and the trajectory $\psi_2(t, x_0)$, we get

$$\int_{-1}^{1} (-x_{01} + a_{12}x_2)\, dx_2 + \int_{s_1}^{s_2} (-1 + a_{12}x_2)\, dx_2 = 0.$$

This leads to

$$2x_{01} = (s_1 - s_2)\left(1 - \frac{a_{12}}{2}(s_1 + s_2)\right). \qquad (30)$$

By Lemma 5.3, $s_1 \le x_{01}$, $s_2 \ge -x_{01}$ (= is taken when $k = 1$), so $s_1 - s_2 \le 2x_{01}$. By Lemma 5.4, $s_1 + s_2 > 0$, so $(1 - (a_{12}/2)(s_1 + s_2)) < 1$. This contradicts (30).

Combining the two cases, we must have $x_{c1} \ne x_{01}$ for any $x_{01} \in (1/ka_{12}, 1)$. Also, by continuity, $x_{01} > x_{c1}$. □

Proof of Lemma 5.6

At x^*, $\dot{x}_2 = 0$, $\dot{x}_1 > 0$, so $\psi_2(t, x^*)$ goes rightward. By Lemma 5.3, $\psi_2(t, x^*)$ will not intersect with the line $x_1 = 1$ before it turns leftward. Since A has a pair of pure imaginary eigenvalues, the trajectory will touch the line $x_2 = -1$ at $-x^*$. And by symmetry, it will return to x^* thus form a closed curve. Note that because at $\pm x^*$, $\dot{x}_2 = 0$, so $\psi_2(t, x^*)$ has only one intersection with each of the lines $x_2 = 1$ and $x_2 = -1$. It follows that $\psi_2(t, x^*)$ is inside the central square.

If $x_0 \in S_0$, then $\psi_2(t, x_0)$ will stay within S_0 since the trajectories will not intersect with $\psi_2(t, x^*)$. Thus $\psi_2(t, x_0)$ is in the linear region and will be a closed trajectory.

Since $a_{11}a_{21} = ka_{12} > a_{12} = a_{12}a_{22}$, $a_{12} > a_{11}$ and $a_{21} > a_{22}$, Condition a) in Proposition 4.1 is satisfied, thus, every trajectory $\psi_2(t, x_0)$ of (25) will enter a bounded attractor and hence is bounded. To prove the remaining part of the lemma, it suffices to show that there is no closed trajectory outside S_0. We prove this by contradiction.

Suppose there is a closed trajectory outside of S_0, say Γ, then Γ goes clockwise and must have two intersections with the line

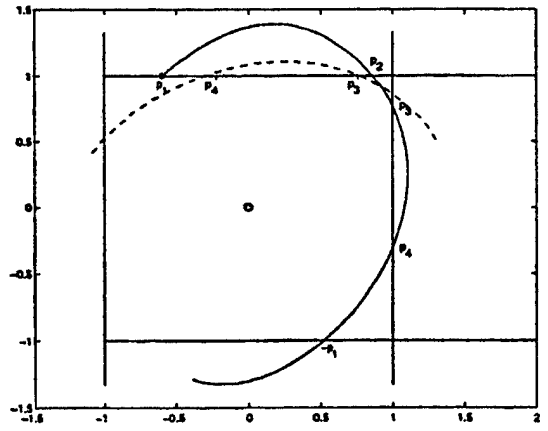


Fig. 13. Illustration for the Proof of Lemma 5.6: Case 1.

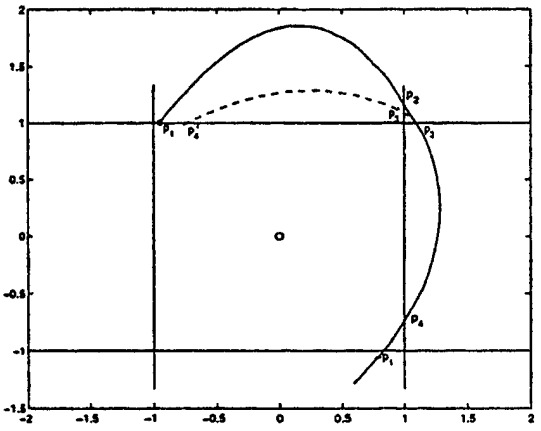


Fig. 14. Ilustration for the Proof of Lemma 5.6: Case 2.

$x_2 = 1$. Denote the region enclosed by Γ as S, then by Green's Theorem, we have

$$\iint_S \left(\frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2}\right) dx_1\, dx_2 = 0.$$

Denote the area of a region X as $\mathcal{A}(X)$, then

$$\mathcal{A}(S \cap U) - \mathcal{A}(S \cap W) = 0. \tag{31}$$

Note that in the central square and in the region V, R, $\partial f_1/\partial x_1 + \partial f_2/\partial x_2 = 0$, in U, $\partial f_1/\partial x_1 + \partial f_2/\partial x_2 = -1$ and in W, $\partial f_1/\partial x_1 + \partial f_2/\partial x_2 = 1$. Also note that S must be symmetric with respect to the origin. Equation (31) implies that Γ must also intersect with the line $x_1 = 1$ and the area of the part of S in the region U and that in the region W must be equal. We will show that this is impossible.

Case 1: The intersections are all on the boundary of the central square. See Fig. 13, where p_1, p_2 are the intersections with $x_2 = 1$, p_3, p_4 are the intersections with $x_1 = 1$. By symmetry, Γ should intersect with $x_2 = -1$ at $-p_1$. The contradiction will be $\mathcal{A}(S \cap U) > \mathcal{A}(S \cap W)$.

Denote

$$p_1 = \begin{bmatrix} s_1 \\ 1 \end{bmatrix}, \quad p_2 = \begin{bmatrix} s_2 \\ 1 \end{bmatrix}, \quad p_3 = \begin{bmatrix} 1 \\ s_3 \end{bmatrix}, \quad p_4 = \begin{bmatrix} 1 \\ s_4 \end{bmatrix}.$$

Then by Lemma 5.3, $s_3 < s_2$, $s_4 > -(-s_1)$ and hence

$$s_1 < s_4 < s_3 < s_2. \tag{32}$$

Get a symmetric projection of $\Gamma_{p_3 \to p_4}$ with respect to the line $x_1 = x_2$ on the region U and denote it as $\Gamma'_{p_3 \to p_4}$ (see the dashed curve in Fig. 13). The corresponding intersections with the line $x_2 = 1$ are $p'_3 = \left[\begin{smallmatrix} s_3 \\ 1 \end{smallmatrix}\right]$, $p'_4 = \left[\begin{smallmatrix} s_4 \\ 1 \end{smallmatrix}\right]$. From (32), p'_3 and p'_4 are between p_1 and p_2.

At a point x $in \Gamma_{p_1 \to p_2}$, the slope of Γ is

$$\eta_1(x_1) = \frac{-ka_{12}x_1 + 1}{-x_1 + a_{12}}.$$

By symmetry (exchanging x_1 and x_2 and taking the inverse of the slope), at a point $x \in \Gamma'_{p_3 \to p_4}$ the slope of Γ' is

$$\eta_2(x_1) = \frac{-1 + a_{12}x_1}{-ka_{12} + x_1}.$$

Clearly, $\mathcal{A}(S \cap W)$ equals the area of the region enclosed by $\Gamma'_{p_3 \to p_4}$ and the line $x_2 = 1$. For the two areas $\mathcal{A}(S \cap U)$ and $\mathcal{A}(S \cap W)$ to be equal, $\Gamma_{p_1 \to p_2}$ and $\Gamma'_{p_3 \to p_4}$ must have two intersections, say $x_l = \left[\begin{smallmatrix} x_{l1} \\ x_{l2} \end{smallmatrix}\right]$ and $x_r = \left[\begin{smallmatrix} x_{r1} \\ x_{r2} \end{smallmatrix}\right]$, with $x_{l1} < x_{r1}$. At the left intersection x_l, $\Gamma'_{p_3 \to p_4}$ crosses $\Gamma_{p_1 \to p_2}$ upward and at the right intersection x_r, $\Gamma'_{p_3 \to p_4}$ crosses $\Gamma_{p_1 \to p_2}$ downward. This implies

$$\eta_1(x_{l1}) < \eta_2(x_{l1}), \qquad \eta_1(x_{r1}) > \eta_2(x_{r1}). \tag{33}$$

Let

$$\begin{aligned} \eta_{12}(x_1) &= \eta_1(x_1) - \eta_2(x_1) \\ &= \frac{-a_{12}(k-1)(x_1^2 - (k+1)a_{12}x_1 + 1)}{(-x_1 + a_{12})(-ka_{21} + x_1)} \end{aligned}$$

then $\eta_{12}(-1) > 0$, $\eta_{12}(1) < 0$ and from (33), we have

$$\eta_{12}(x_{l1}) < 0, \quad \eta_{12}(x_{r1}) > 0.$$

The function changes sign three times, so $\eta_{12}(x_1)$ has at least three zeros between -1 and -1. Obviously there are only two zeros in this interval, hence, $\Gamma'_{p_3 \to p_4}$ and $\Gamma_{p_1 \to p_2}$ cannot have two intersections. Consequently, $\mathcal{A}(S \cap U) > \mathcal{A}(S \cap W)$. A contradiction.

Case 2: Γ intersects with the lines $x_1 = 1, x_2 = 1$ as in Fig. 14. From p_2 to p_3, the slope of the straight line is $\alpha = (-ka_{12} + 1)/(-1 + a_{12}) < -1$, so $\left\| p_2 - \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right] \right\| > \left\| p_3 - \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right] \right\|$. From Lemma 5.3, $\left\| p_4 - \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right] \right\| > \left\| -p_1 - \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right] \right\| = \left\| p_1 - \left[\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right] \right\|$, note that Γ is symmetric and must intersect with $x_2 = -1$ at $-p_1$. Hence, if we get a symmetric projection of $\Gamma_{p_3 \to p_4}$ on the region U (see the dashed curve in Fig. 14), then p'_4 is to the right of p_1 and p'_3 is below p_2. Suppose $\mathcal{A}(S \cap U) = \mathcal{A}(S \cap W)$, $\Gamma_{p_1 \to p_2}$ will

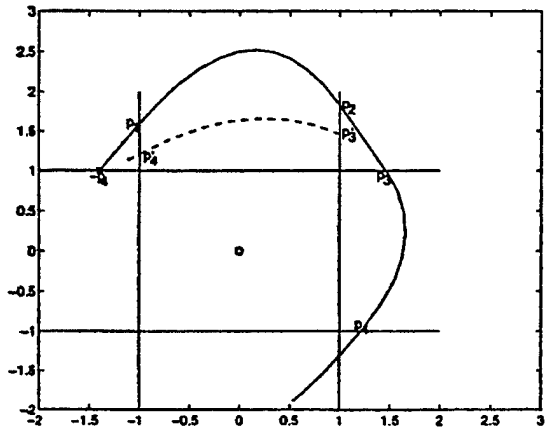


Fig. 15. Illustration for the Proof of Lemma 5.6: Case 3.

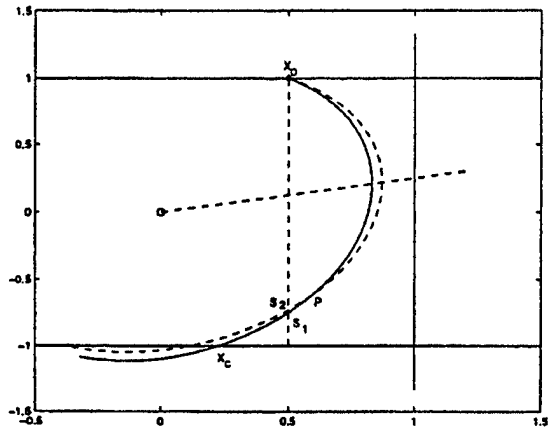


Fig. 16. Illustration for the Proof of Lemma 5.8: Case 1.

intersect with $\Gamma'_{p_3 \to p_4}$ twice. This is impossible as was shown in Case 1.

Case 3: The four intersections are as shown in Fig. 15.

Similarly, since $\alpha < -1$, $\beta > 1$, we have $\left\|p_2 - \begin{bmatrix}1\\1\end{bmatrix}\right\| > \left\|p_3 - \begin{bmatrix}1\\1\end{bmatrix}\right\|$ and $\left\|p_4 - \begin{bmatrix}1\\-1\end{bmatrix}\right\| < \left\|-p_1 - \begin{bmatrix}1\\-1\end{bmatrix}\right\| = \left\|p_1 - \begin{bmatrix}-1\\1\end{bmatrix}\right\|$. Also get a symmetric projection of $\Gamma_{p_3 \to p_4}$ on the region U (see the dashed curve in Fig. 15), then p'_4 is below p_1 and p'_3 is below p_2. Similar to Case 1, we can show that $\mathcal{A}(S \cap U) > \mathcal{A}(S \cap W)$. A contradiction.

Because $h_1 < 0$, there is no such case where Γ only encloses $\begin{bmatrix}1\\-1\end{bmatrix}$ and $\begin{bmatrix}-1\\1\end{bmatrix}$ but not $\begin{bmatrix}1\\1\end{bmatrix}$ and $\begin{bmatrix}-1\\-1\end{bmatrix}$.

Combining the above three cases, we see that there is no closed trajectory that intersects with any of the lines $x_1 = \pm 1$ and $x_2 = \pm 1$ twice. So there is no closed trajectory outside of S_0 and if a trajectory starts outside of S_0, it will converge to the boundary of S_0. □

Proof of Lemma 5.7

Since x_0 is outside of S_0 and is to the left of x^*, so $\psi_2(t, x_0)$ goes upward-rightward at first. After crossing the line $x_1 = 1/ka_{12}$, it goes downward and returns to the line $x_2 = 1$. Since $\psi_2(t, x_0)$ goes clockwise, at the first intersection with the line $x_2 = -1$ it crosses the line downward, so $\dot{x}_2 < 0$ at x_c and $x_{c1} > -1/ka_{12}$. Let t_m be the time when $\psi_2(t, x_0)$ intersects with $x_2 = -1$, i.e., $\psi_2(t_m, x_0) = x_c$. Suppose $x_{c1} = -x_{01}$, then $x_c = -x_0$ and $\{\psi_2(t, x_0), t \in [0, 2t_m]\}$ is a closed curve. This is impossible by Lemma 5.6. Now suppose $x_{c1} > -x_{01}$, then x_c is to the right of $-x_0$. Let the region enclosed by $\{\psi_2(t, x_0), t \in [0, t_m]\}$, $\{\psi_2(t, -x_0), t \in [0, t_m]\}$, the line x_c to $-x_0$ and the line $-x_c$ to x_0 be S, then on the line from x_c to $-x_0$, $\dot{x}_2 < 0$ and $\dot{x}$ points outward from S. Similarly, on the line from $-x_c$ to x_0, $\dot{x}$ also points outward from S. Thus, no trajectory outside of S will enter it. This contradicts with Lemma 5.6 since S_0 is in the interior of S.

Therefore, we must have $-(1/ka_{12}) < x_{c1} < -x_{01}$. □

Proof of Lemma 5.8

Without loss of generality, assume $x_{02} = 1$. When $x_{02} \in (0, 1)$, the proof can be carried out similarly. There are three cases.

Case 1: $x_{01} < 1$ and $\psi(t, x_0, a_{22})$ does not intersect with the line $x_1 = 1$ (see Fig. 16).

Because $x_{01} \geq 1/ka_{12}$, both $\psi(t, x_0, a_{22})$ and $\psi(t, x_0, a_{22} - \delta)$ go downward-rightward until reaching the line $x_2 = x_1/a_{12}$ (see the dashed line passing through the origin). (On this line, $\dot{x}_1 = 0$). After that, the trajectories turn leftward.

At $t = 0$, the slopes of both trajectories are negative and $\eta(x_0, a_{22}) < \eta(x_0, a_{22} - \delta)$, so $\psi(t, x_0, a_{22} - \delta)$ will go to the right of $\psi(t, x_0, a_{22})$ at the begining. On the part of $\psi(t, x_0, a_{22})$ that is above the line $x_2 = x_1/a_{12}$, $\eta(x, a_{22}) < \eta(x, a_{22} - \delta)$ and on the part that is below the line $\eta(x, a_{22}) > \eta(x, a_{22} - \delta)$. So $\psi(t, x_0, a_{22} - \delta)$ can only cross $\psi(t, x_0, a_{22})$ leftward below the line. We will show that the crossing point p (see Fig. 16) is to the right of x_0 and x_c. After the crossing, $\psi(t, x_0, a_{22} - \delta)$ will stay to the left of $\psi(t, x_0, a_{22})$ until meeting the line $x_2 = -1$. This leads to the desired result.

Let $s_1 = \begin{bmatrix}x_{01}\\e_1\end{bmatrix}$, $s_2 = \begin{bmatrix}x_{01}\\e_2\end{bmatrix}$ be the intersections of $\psi(t, x_0, a_{22})$ and $\psi(t, x_0, a_{22} - \delta)$ with the vertical line $x_1 = x_{01}$, respectively. Since x_c is to the left of x_0, so s_1 is above the line $x_2 = -1$. Assume on the contrary that there is no intersection of $\psi(t, x_0, a_{22} - \delta)$ with $\psi(t, x_0, a_{22})$ that is to the right of the line $x_1 = x_{01}$, then s_1 must be above s_2, i.e., $e_1 > e_2$ and $\psi(t, x_0, a_{22} - \delta)$ is to the right of $\psi(t, x_0, a_{22})$ before meeting the line $x_1 = x_{01}$. Denote the area of the region enclosed by $\psi(t, x_0, a_{22})$ and the line from x_0 to s_1 as $\mathcal{A}$ and the area of the region enclosed by $\psi(t, x_0, a_{22} - \delta)$ with the line from x_0 to s_2 as $\mathcal{B}$, then $\mathcal{A} < \mathcal{B}$. Applying Green's theorem to the vector fields corresponding to a_{22} and $a_{22} - \delta$, we have

$$-\int_{e_1}^{1} f_1(x)\, dx_2 = \mathcal{A}(a_{22} - 1) \tag{34}$$

$$-\int_{e_2}^{1} f_1(x)\, dx_2 = \mathcal{B}(a_{22} - \delta - 1). \tag{35}$$

Note that $f_1(x) = -x_1 + a_{12}x_2$ is the same for both the vector fields. Subtracting (35) from (34), we obtain

$$\int_{e_2}^{e_1} f_1(x) dx_2 = (\mathcal{B} - \mathcal{A})(1 - a_{22}) + \mathcal{B}\delta > 0.$$

We know that $f_1(x) < 0$ from s_1 to s_2 since the trajectories go leftward. By assumption, $e_1 > e_2$, so $\int_{e_2}^{e_1} f_1(x) dx_2 < 0$. A contradiction. Therefore, we must have $e_1 < e_2$ and

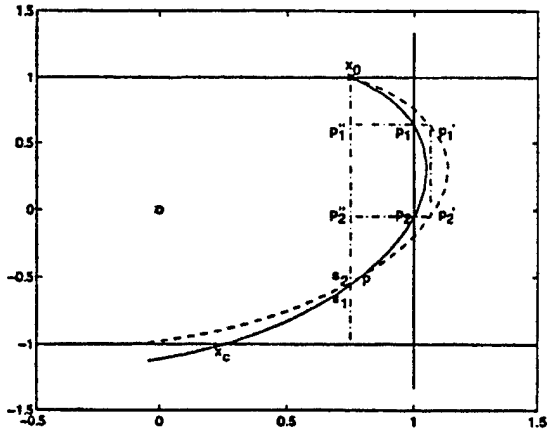


Fig. 17. Illustration for the Proof of Lemma 5.8: Case 2.

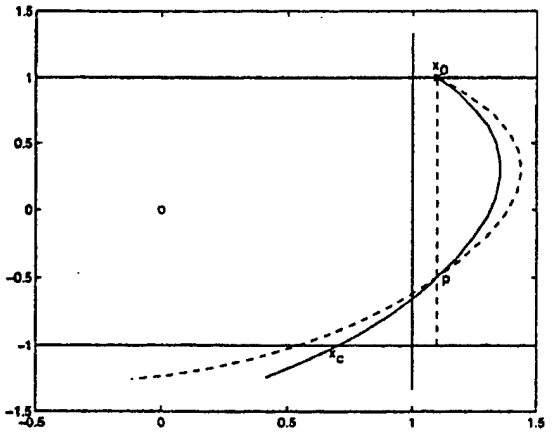


Fig. 18. Illustration for the Proof of Lemma 5.8: Case 3.

$\psi(t, x_0, a_{22} - \delta)$ intersects $\psi(t, x_0, a_{22})$ at a point p to the right of x_0.

Case 2: $x_{01} < 1$ and $\psi(t, x_0, a_{22})$ intersects with $x_1 = 1$, see Fig. 17.

Let p_1 be the first intersection of $\psi(t, x_0, a_{22})$ with $x_1 = 1$, p_2 be the second one.

Let the horizontal distance from $\psi(t, x_0, a_{22})$ to $\psi(t, x_0, a_{22} - \delta)$ at p_1 be Δ_1 and that at p_2 be Δ_2. Then $p_1 + \begin{bmatrix}\Delta_1\\0\end{bmatrix} =: p_1'$ is on $\psi(t, x_0, a_{22} - \delta)$. By the horizontal shifting property (12) of the trajectories, $\psi(t, p_1', a_{22})$ will intersect the line $x_1 = 1 + \Delta_1$ at $p_2 + \begin{bmatrix}\Delta_1\\0\end{bmatrix} =: p_2'$. From Fact 1, $\psi(t, p_1', a_{22} - \delta)$ also returns to the line $x_1 = 1 + \Delta_1$ at p_2'. Because p_1' is on $\psi(t, x_0, a_{22} - \delta)$, $\psi(t, p_1', a_{22} - \delta)$ overlaps with $\psi(t, x_0, a_{22} - \delta)$. It follows that $\Delta_2 = \Delta_1$.

Let s_1, s_2 and p be defined similarly to Case 1, we will also show that s_2 is above s_1 by contradiction. First, we need an upper bound for Δ_1.

Let $v = \begin{bmatrix}v_1\\v_2\end{bmatrix}$ be a point on $\psi(t, x_0, a_{22})$ and $\begin{bmatrix}v_1+\Delta\\v_2\end{bmatrix}$ be a point on $\psi(t, x_0, a_{22} - \delta)$. Define

$$g(v_1, v_2) := \frac{(-ka_{12}v_1 + v_2)}{-v_1 + a_{12}v_2}$$

then the slope of $\psi(t, x_0, a_{22})$ at v is $a_{22}g(v_1, v_2) =: \eta_1$ and the slope of $\psi(t, x_0, a_{22} - \delta)$ at $\begin{bmatrix}v_1+\Delta\\v_2\end{bmatrix}$ is $(a_{22} - \delta)g(v_1 + \Delta, v_2) =: \eta_2$. It easily can be verified that for $v_2 > 0$, $g(v_1, v_2)$ is a decreasing function of v_1, so

$$g(v_1, v_2) > g(v_1 + \Delta, v_2), \qquad \forall \Delta > 0, v_2 > 0. \tag{36}$$

We can view Δ as a function of v_1. Routine analysis shows that

$$\frac{d\Delta}{dv_1} = \frac{\eta_1 - \eta_2}{\eta_2}.$$

Note that the part of $\psi(t, x_0, a_{22})$ from x_0 to p_1 is above the line $x_2 = 0$. It follows from (36) that

$$\begin{aligned}\eta_1 - \eta_2 &= a_{22}g(v_1, v_2) - (a_{22} - \delta)g(v_1 + \Delta, v_2)\\ &= a_{22}(g(v_1, v_2) - g(v_1 + \Delta, v_2)) + \delta g(v_1 + \Delta, v_2)\\ &> \delta g(v_1 + \Delta, v_2).\end{aligned}$$

Since $g(v_1 + \Delta, v_2) < 0$, we have

$$\frac{d\Delta}{dv_1} < \frac{\delta}{a_{22} - \delta}.$$

Therefore

$$\Delta_1 = \int_{x_{01}}^{1} \frac{d\Delta}{dv_1}\, dv_1 < \frac{\delta}{a_{22} - \delta}(1 - x_{01}). \tag{37}$$

Let p_1'' be the intersection of the extension of the line from p_1 to p_1' with $x_1 = x_{01}$ and p_2'' be that of the line from p_2 to p_2' (see Fig. 17). Denote $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4$ as the areas of the regions enclosed by straight lines and $\psi(t, x_0, a_{22})$ corresponding to the sets of vertices $\{x_0, p_1'', p_1\}$, $\{p_2'', p_2, s_1\}$, $\{p_1'', p_1, p_2, p_2''\}$, $\{p_1, p_2\}$, respectively. Denote $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_4, \mathcal{B}_5$ as the areas of the regions enclosed by straight lines and $\psi(t, x_0, a_{22} - \delta)$ corresponding to the sets of vertices $\{x_0, p_1'', p_1'\}$, $\{p_2'', p_2', s_2\}$, $\{p_1', p_2'\}$, $\{p_1, p_1', p_2', p_2\}$, respectively. Then by Green's theorem, we have

$$\begin{aligned}-\int_{e_1} f_1(x)\, dx_2 &= (-1 + a_{22})(\mathcal{A}_1 + \mathcal{A}_2) + (-1 + a_{22})\mathcal{A}_3\\ &\quad + a_{22}\mathcal{A}_4\end{aligned} \tag{38}$$

$$\begin{aligned}-\int_{e_2}^{1} f_1(x)\, dx_2 &= (-1 + a_{22} - \delta)(\mathcal{B}_1 + \mathcal{B}_2)\\ &\quad + (-1 + a_{22} - \delta)\mathcal{A}_3 + (a_{22} - \delta)\mathcal{B}_4\\ &\quad + (a_{22} - \delta)\mathcal{B}_5.\end{aligned} \tag{39}$$

Note that there are small triangle areas in $\mathcal{B}_1$ and $\mathcal{B}_2$ that are in the region W. They are of the order δ^2. Since δ is arbitrarily small, we can treat them as a region in the central square for simplicity.

It follows from Fact 5.1 b) and the horizontal shifting property in the region W that $a_{22}\mathcal{A}_4 = (a_{22} - \delta)\mathcal{B}_4$. From (37), we have

$$\begin{aligned}&(-1 + a_{22})\mathcal{A}_3 - (-1 + a_{22} - \delta)\mathcal{A}_3 - (a_{22} - \delta)\mathcal{B}_5\\ &\quad = \delta\mathcal{A}_3 - (a_{22} - \delta)\mathcal{B}_5\\ &\quad = \delta\|p_1 - p_2\|(1 - x_{01}) - (a_{22} - \delta)\|p_1 - p_2\|\Delta_1 > 0.\end{aligned}$$

By assumption, s_1 is above s_2, so $\mathcal{A}_1 + \mathcal{A}_2 < \mathcal{B}_1 + \mathcal{B}_2$. Subtracting (39) from (38) we get

$$\int_{e_2}^{e_1} f_1(x)dx_2 > 0.$$

A contradiction with $e_1 > e_2$ and $f_1(x) < 0$.

Case 3: $x_{01} \geq 1$ (see Fig. 18).

In this case, $\psi(t, x_0, a_{22})$ goes downward-rightward, then turns downward-leftward and returns to the line $x_1 = x_{01}$ at a point, say, $p = \begin{bmatrix} x_{01} \\ x_{p2} \end{bmatrix}$. Because $h_1 < 0$, p is above the line $x_2 = -1$. By Fact 5.1, $\psi(t, x_0, a_{22} - \delta)$ will also return to the line $x_1 = x_{01}$ at the same point p. After that, $\psi(t, x_0, a_{22} - \delta)$ remains to the left of $\psi(t, x_0, a_{22})$ untill it meets the line $x_2 = -1$ and the desired result follows.

REFERENCES

[1] F. Albertini and D. D'Aless, "Asymptotic stability of continuous-time systems with saturation nonlinearities," *Syst. Contr. Lett.*, vol. 29, pp. 175–180, Nov. 1996.

[2] D. S. Bernstein and A. N. Michel, "A chronological bibliography on saturating actuators," *Int. J. Robust Nonlinear Contr.*, vol. 5, pp. 375–380, Aug. 1995.

[3] L. Hou and A. N. Michel, "Asymptotic stability of systems with saturation constraints," *IEEE Trans. Automat. Contr.*, vol. 43, pp. 1148–1154, Aug. 1998.

[4] L. Jin, P. N. Nikiforuk, and M. M. Gupta, "Absolute stability conditions for discrete-time recurrent neural networks," *IEEE Trans. Neural Networks*, vol. 5, pp. 954–964, Nov. 1994.

[5] Z. Lin, *Low Gain Feedback*, ser. Lecture Notes in Control and Information Sciences. London, U.K.: Springer-Verlag, 1998.

[6] D. Liu and A. N. Michel, "Asymptotic stability of systems operating on a closed hypercube," *Syst. Contr. Lett.*, vol. 19, pp. 281–285, Oct. 1992.

[7] ——, "Sparsely interconnected neural networks for associative memories with applications to cellular neural networks," *IEEE Trans. Circuits Syst.*, vol. 41, pp. 295–307, Apr. 1994.

[8] ——, *Dynamical Systems with Saturation Nonlinearities*, ser. Lecture Notes in Control and Information Sciences. London, U.K.: Springer, 1994.

[9] R. Mantri, A. Saberi, and V. Venkatasubramanian, "Stability analysis of continuous time planar systems with state saturation nonlinearity," *IEEE Trans. Circuits Syst. I*, vol. 45, pp. 989–993, Sept. 1998.

[10] J. H. F. Ritzerfeld, "A condition for the overflow stability of second-order digital filters that is satisfied by all scaled state-space structures using saturation," *IEEE Trans. Circuits Syst.*, vol. 36, pp. 1049–1057, Aug. 1989.

Tingshu Hu (S'99) was born in Sichuan, China, in 1966. She received the B.S. and M.S. degrees in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 1985 and 1988 respectively.

She is currently working toward a Ph.D. degree at the Department of Electrical Engineering, University of Virginia, Charlottesville. Her research interests include systems with saturation nonlinearities and robust control theory.

Zongli Lin (S'89–M'90–SM'98) was born in Fuqing, Fujian, China, on February 24, 1964. He received the B.S. degree in mathematics and computer science from Amoy University, Xiamen, China, in 1983, the M.Eng. degree in automatic control from the Chinese Academy of Space Technology, Beijing, China, in 1989, and the Ph.D. degree in electrical and computer engineering from Washington State University, Pullman, WA, in May 1994.

From July 1983 to July 1986, He worked as a Control Engineer at the Chinese Academy of Space Technology. In January 1994 he joined the Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, as a Visiting Assistant Professor. He is now an Assistant Professor in Electrical Engineering at the University of Virginia, Charlottesville. His current research interests include nonlinear control, robust control, and control of systems with saturating actuators. In these areas he has published several papers. He is also the Author of the recent book, *Low Gain Feedback* (London, U.K.: Springer-Verlag, 1998).

Dr. Lin currently serves as an Associate Editor on the Conference Editorial Board of the IEEE Control Systems Society. He is the Recipient of an ONR Young Investigator Award.

Publication 5

ACKNOWLEDGMENT

The authors are grateful to the reviewers for their helpful comments and suggestions.

REFERENCES

[1] I. B. Rhodes, "A parallel decomposition for Kalman filters," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 322–324, 1990.
[2] R. A. Singer and R. G. Sea, "Increasing the computational efficiency of discrete Kalman filters," *IEEE Trans. Automat. Contr.*, vol. 16, pp. 254–257, 1971.
[3] E. Tse and M. Athans, "Optimal minimal-order observer-estimators for discrete linear time-varying systems," *IEEE Trans. Automat. Contr.*, vol. 15, pp. 416–426, 1970.
[4] C. T. Leondes and L. M. Novak, "Optimal minimal-order observers for discrete-time systems—A unified theory," *Automatica*, vol. 8, pp. 379–387, 1972.
[5] E. Fogel and Y. F. Huang, "Reduced-order optimal state estimator for linear systems with partially noise corrupted measurement," *IEEE Trans. Automat. Contr.*, vol. 25, pp. 994–996, 1980.
[6] Y. Halevi, "The optimal reduced-order estimator for systems with singular measurement noise," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 777–781, 1989.
[7] B. Friedland, "Treatment of bias in recursive filtering," *IEEE Trans. Automat. Contr.*, vol. 14, pp. 359–367, 1969.
[8] A. T. Alouani, P. Xia, T. R. Rice, and W. D. Blair, "On the optimality of two-stage state estimation in the presence of random bias," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 1279–1282, 1993.
[9] C.-S. Hsieh and F.-C. Chen, "Optimal solution of the two-stage Kalman estimator," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 194–199, 1999.
[10] H. A. P. Blom and Y. Bar-Shalom, "The interacting multiple model algorithm for systems with Markovian switching coefficients," *IEEE Trans. Automat. Contr.*, vol. 23, pp. 780–783, 1988.
[11] G. A. Watson and W. D. Blair, "Interacting acceleration compensation algorithm for tracking maneuvering targets," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 31, pp. 1152–1159, 1995.
[12] W. D. Blair and G. A. Watson, "Interacting multiple bias model algorithm with application to tracking maneuvering targets," in *Proc. IEEE 31th Conf. Decision Contr.*, Tucson, AZ, Dec. 1992, pp. 3790–3795.

Adaptive Control of a Weakly Nonminimum Phase Linear System

Zongli Lin and Gang Tao

Abstract—**For a weakly nonminimum phase linear system, we design an adaptive state feedback control law that causes the system output to track a desired trajectory to an arbitrarily high degree of precision. The key to this is the use of a low gain feedback design technique.**

Index Terms—**Adaptive control, low gain feedback, nonminimum phase, tracking.**

I. INTRODUCTION

It is well known in both classical and modern control that the system minimum phase property facilitates control designs [1], [3]–[5], [7], [9], [10]. However, many practical systems are nonminimum phase [3]. The nonminimum phase property may prevent some desired control objectives from being achieved. One of such objectives is that the system output tracks a desired trajectory. Since the unstable system zeros cannot be cancelled by standard state or output feedback controllers, special designs are usually needed for output tracking purpose. Controllers based on the internal model principal [1], [4], [5] are commonly used for nonminimum phase systems in order to achieve output tracking of reference signals at the internal model frequencies.

In this paper we explore a different way of treating nonminimum phase systems and identify conditions under which certain design objective can be met for nonminimum phase systems. In particular, by utilizing the recent development of low gain feedback design techniques [8], we show how adaptive state feedback control laws can be constructed to cause the output of a weakly nonminimum phase linear system (a system whose invariant zeros are in the closed left-half of the complex plane) to track a given reference output trajectory to an arbitrarily high degree of precision. The reference output trajectory can be any smooth signals that do not contain the frequency components of the jw axis invariant zeros.

The rest of the paper is organized as follows. In Section II, we formulate the problem of designing adaptive state feedback controllers for weakly nonminimum phase linear systems to ensure desired tracking properties. In Section III, we will show that, with the use of a low gain feedback design, our adaptive controller, when applied to a weakly nonminimum phase linear system, ensures the closed-loop signal boundedness and an output tracking error whose steady-state trajectory can be made arbitrarily small. We will also present an example with simulation results to illustrate our low gain adaptive controller and its desired tracking performance. A brief concluding remark is made in Section 4.

II. PROBLEM STATEMENT

Consider the following linear system:

$$\begin{cases} \dot{x}_0 = A_0 x_0 + B_0 x_1, \quad x_0 \in \mathbf{R}^{n_0} \\ \dot{x}_1 = x_2, \quad x_1 \in \mathbf{R} \\ \quad \vdots \\ \dot{x}_{r-1} = x_r \\ \dot{x}_r = E_0 x_0 + a_1 x_1 + \cdots + a_r x_r + bu, \quad b > 0 \\ y = x_1 \end{cases} \tag{1}$$

where $x = [x_0', x_1, x_2, \ldots, x_r]' \in \mathbf{R}^{n_0+r}$ is the state vector, $u \in \mathbf{R}$ is the control input, $y \in \mathbf{R}$ is the system output, and b and a_i's are unknown system parameters. We have assumed that $b \neq 0$, and without loss of generality, further assumed that $b \geq b_0 > 0$ for some known b_0.

We note here that the dynamics of x_0 is the zero dynamics of the system and the eigenvalues of A_0 are the invariant zeros of the system. Also note that this system has a relative degree of r. We further make the following assumptions on the system.

Assumption 1: The system (1) is of weakly nonminimum phase, i.e., all the eigenvalues of A_0 lie in the closed left-half plane.

Assumption 2: The system (1) is stabilizable, i.e., the pair (A_0, B_0) is stabilizable.

Our objective is to construct an adaptive state feedback control law that causes the system output $y(t)$ to track a desired output trajectory y_d to an arbitrarily high degree of precision without knowing the values of the system parameters b and a_i's. Unlike in the adaptive control of minimum phase systems in which A_0 is stable, here A_0 is unstable and needs to be stabilized through B_0. For this reason, we require the knowledge of A_0 and B_0. We also make the following assumption on the desired trajectory y_d.

Manuscript received February 2, 1999; revised May 20, 1999. Recommended by Associate Editor, B. M. Chen. This work was supported in part by the U.S. Office of Naval Research Young Investigator Program under Grant N00014-99-1-0670 and the National Science Foundation under Grant CS-9619363.

The authors are with the Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903-2442 USA.

Publisher Item Identifier S 0018-9286(00)01087-4.

0018–9286/00$10.00 © 2000 IEEE

Assumption 3: The desired output trajectory y_d is a $\mathcal{C}^r$ function of time t with all $y_d^{(i)}$, $i = 1, 2, \ldots, r$, bounded, and all its frequency components are away from the $j\omega$-axis invariant zeros of the system.

More precisely, we will solve the following control problem under Assumptions 1, 2 and 3.

Problem 1: Consider the system (1). For any *a priori* given (arbitrarily small) number $\eta > 0$, design an adaptive state feedback control law,

$$\begin{cases} u = \phi\left(\alpha_0, \alpha_1, \ldots, \alpha_r, x_0, x_1, \ldots, x_r, y_d, \dot{y}_d, \ldots, y_d^{(r)}\right) \\ \dot{\alpha}_i = \psi\left(x_0, x_1, \ldots, x_r, y_d, \dot{y}_d, \ldots, y_d^{(r)}\right), \quad i = 0, 1, \ldots, r \end{cases} \tag{2}$$

such that all signals of the closed-loop system are bounded, and:

a) for $y_d \equiv 0$, $\lim_{t\to\infty} y(t) = 0$;

b) for any desired trajectory y_d that satisfies Assumption 3

$$|e_{ss}(t)| \leq \eta \tag{3}$$

where $e = y - y_d$, and $e_{ss}(t)$ stands for the steady state trajectory of $e(t)$.

III. Main Results

In this section, we first explicitly design an adaptive state feedback law of the form (2) and then show that it indeed solves Problem 1. Our design is carried out in the following three steps.

Step 1: Low Gain Feedback. Consider the pair (A_0, B_0). By Assumptions 1 and 2, the pair (A_0, B_0) is stabilizable, and all the eigenvalues of A_0 are in the closed left-half plane. The low gain feedback design can be described in the following three sub-steps.

Step 1.1. Find nonsingular transformation matrices T_S such that the pair (A_0, B_0) is transformed into the following control canonical form:

$$T_S^{-1} A_0 T_S = \begin{bmatrix} A_0^0 & 0 \\ 0 & A_0^- \end{bmatrix}, \quad T_S^{-1} B_0 = \begin{bmatrix} B_0^0 \\ 0 \end{bmatrix} \tag{4}$$

where, $A_0^- \in \mathbf{R}^{n_0^- \times n_0^-}$ contains all the open left-half plane eigenvalues of A_0, and all eigenvalues of $A_0^0 \in \mathbf{R}^{n_0^0 \times n_0^0}$ are on the $j\omega$ axis and hence (A_0^0, B_0^0) is controllable as given by,

$$A_0^0 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_{n_0^0}^0 & -a_{n_0^0-1}^0 & -a_{n_0^0-2}^0 & \cdots & -a_1^0 \end{bmatrix}, \quad B_0^0 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{5}$$

We note here that the existence of such a T_S is due to the fact that the set of eigenvalues of A_0^0 are disjoint from that of A_0^-. An efficient algorithm for the contruction of T_S can be found in [2].

Step 1.2. For the (A_0^0, B_0^0), let $F_0^0(\varepsilon) \in \mathbf{R}^{1\times n_0^0}$ be the state feedback gain such that

$$\lambda\left(A_0^0 + B_0^0 F_0^0(\varepsilon)\right) = -\varepsilon + \lambda\left(A_0^0\right) \in \mathbf{C}^-, \quad \varepsilon > 0. \tag{6}$$

Note that $F_0^0(\varepsilon)$ is unique. Here ε is a design parameter whose value is to be determined by the desired degree of tracking precision η.

Step 1.3. Form a feedback gain for the pair (A_0, B_0) as follows:

$$F_0(\varepsilon) = \begin{bmatrix} F_0^0(\varepsilon) & 0 \end{bmatrix} T_S^{-1}. \tag{7}$$

It is clear that $F_0^0(\varepsilon)$ tends to zero as ε tends to zero. For this reason, the feedback gain $F_0^0(\varepsilon)$ is referred to as low gain feedback. The properties of such a low gain design were examined in detail in [8]. In what follows, we establish another property that is pertinent to the development of the current paper.

Lemma 1: Consider the triple $(A_0^0, B_0^0, F_0^0(\varepsilon))$ as given by (5) and (6). Then, there exists an $\varepsilon^* > 0$ such that for all $\varepsilon \in (0, \varepsilon^*]$,

$$\left|F_0^0(\varepsilon)\left(j\omega I - A_0^0 - B_0^0 F_0^0(\varepsilon)\right)^{-1}\right| \leq \gamma\varepsilon \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left|\frac{1}{(j\omega - j\omega_i + \varepsilon)^j}\right| \tag{8}$$

where $j\omega_i$'s are the distinct eigenvalues of A_0^0, n_i is the algebraic multiplicity of $j\omega_i$, and γ is a positive constant independent of ε.

Proof: See Appendix. □

Step 2: Output Renaming. Define a new output as

$$\tilde{y} = x_1 - F_0(\varepsilon) x_0 - y_d \tag{9}$$

where y_d is the desired trajectory for the original output y to track. With this new output, the system (1) can be rewritten as (10), shown at the bottom of the page, where

$$\begin{cases} \dot{x}_0 = (A_0 + B_0 F_0(\varepsilon)) x_0 + B_0 \tilde{y} + B_0 y_d, \quad x_0 \in \mathbf{R}^{n_0} \\ \dot{\tilde{x}}_1 = \tilde{x}_2 \\ \quad \vdots \\ \dot{\tilde{x}}_{r-1} = \tilde{x}_r \\ \dot{\tilde{x}}_r = \bar{a}_0(\varepsilon) x_0 + \bar{a}_1(\varepsilon)[\tilde{x}_1 + y_d] + \cdots + \bar{a}_r(\varepsilon)\left[\tilde{x}_r + y_d^{(r-1)}\right] + bu - y_d^{(r)} \\ \tilde{y} = \tilde{x}_1 \end{cases} \tag{10}$$

$$\begin{cases} \tilde{x}_1 = x_1 - F_0(\varepsilon) x_0 - y_d \\ \tilde{x}_2 = x_2 - F_0(\varepsilon) A_0 x_0 - F_0(\varepsilon) B_0 x_1 - \dot{y}_d \\ \tilde{x}_3 = x_3 - F_0(\varepsilon) A_0^2 x_0 - F_0(\varepsilon) A_0 B_0 x_1 - F_0(\varepsilon) B_0 x_2 - \ddot{y}_d \\ \qquad\qquad \vdots \\ \tilde{x}_r = x_r - F_0(\varepsilon) A_0^{r-1} x_0 - F_0(\varepsilon) A^{r-2} B_0 x_1 - \cdots - F_0(\varepsilon) B_0 x_{r-1} - y_d^{(r-1)} \end{cases} \tag{11}$$

we have defined the new state variables as shown in (11), also shown at the bottom of the previous page, and $\bar{a}_i(\varepsilon)$'s are the unknown parameters, defined as functions of the unknown parameters a_i's and the design parameter ε in an obvious way.

Step 3: Adaptive feedback controller design. Let us first consider the $\tilde{x} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_r]'$ subsystem

$$\begin{cases} \dot{\tilde{x}}_1 = \tilde{x}_2 \\ \quad \vdots \\ \dot{\tilde{x}}_{r-1} = \tilde{x}_r \\ \dot{\tilde{x}}_r = \bar{a}_0(\varepsilon)x_0 + \bar{a}_1(\varepsilon)z_1 + \cdots + \bar{a}_r(\varepsilon)z_r + bu - y_d^{(r)} \\ \tilde{y} = \tilde{x}_1 \end{cases} \tag{12}$$

where for $i = 1$ to r

$$z_i = \tilde{x}_i + y_d^{(i-1)}. \tag{13}$$

Our task here is to stabilize this system. Since $\bar{a}_i(\varepsilon)$'s depend on the unknown parameters a_i's, we choose to design an adaptive stabilizer of form:

$$\begin{cases} u = \frac{1}{\hat{b}} v \\ v = y_d^{(r)} \quad -\hat{a}_0 x_0 - \hat{a}_1 z_1 - \cdots - \hat{a}_r z_r \\ \qquad\quad -k_1\tilde{x}_1 - k_2\tilde{x}_2 - \cdots - k_r\tilde{x}_r \end{cases} \tag{14}$$

where k_i's are positive constants such that

$$s^r + k_r s^{r-1} + \cdots + k_2 s + k_1 \tag{15}$$

is a Hurwitz polynomial whose roots are at desired locations, and $\hat{b}$ and $\hat{a}_i$'s are design parameters, which are estimates of b and $\bar{a}_i(\varepsilon)$'s, respectively, and are to be tuned adaptively.

To determine the adaptation laws for tuning the parameter estimates $\hat{b}$ and $\hat{a}_i$'s, we consider the closed-loop system under the control law (14), shown in (16), at the bottom of the page, which can be written in the following compact form:

$$\begin{cases} \dot{x}_0 = (A_0 + B_0 F_0(\varepsilon))x_0 + B_0\tilde{y} + B_0 y_d \\ \dot{\tilde{x}} = A\tilde{x} + B(\bar{a} - \hat{a})'z + B(b - \hat{b})u \end{cases} \tag{17}$$

where $\tilde{x} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_r]'$, and

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -k_1 & -k_2 & -k_3 & \cdots & -k_r \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{18}$$

$$\bar{a} = \begin{bmatrix} \bar{a}_0(\varepsilon) \\ \bar{a}_1(\varepsilon) \\ \vdots \\ \bar{a}_r(\varepsilon) \end{bmatrix}, \quad \hat{a} = \begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \vdots \\ \hat{a}_r \end{bmatrix}, \quad z = \begin{bmatrix} x_0 \\ z_1 \\ \vdots \\ z_r \end{bmatrix}. \tag{19}$$

The error equation (17) then motivates us to choose the adaptation laws for $\hat{b}$ and $\hat{a}(t)$ as,

$$\dot{\hat{b}} = \begin{cases} 0, & \text{if } \hat{b} = b_0 \text{ and } \tilde{x}'PBu < 0 \\ \gamma\tilde{x}'PBu, & \text{otherwise} \end{cases}, \quad \hat{b}(0) \geq b_0 \tag{20}$$

and

$$\dot{\hat{a}} = \Gamma z\tilde{x}'PB \tag{21}$$

where $\gamma > 0$ and $\Gamma = \Gamma' > 0$ are adaptation gains and $P = P' > 0$ is the unique positive definite solution to the Lyapunov equation

$$A'P + PA = -Q \tag{22}$$

for any chosen $Q = Q' > 0$. The existence of such a P is due to the fact that A is asymptotically stable by the choice of k_i's.

The following theorem shows that our adaptive state feedback design indeed meets the stated control objective.

Theorem 1 (Solution to **Problem 1***):* Under Assumptions 1–3, the adaptive feedback control law comprising (14), (20) and (21) ensures that all signals of the closed-loop system are bounded, and the closed-loop system has the following two desired properties,

a) for $y_d \equiv 0$, $\lim_{t\to\infty} x(t) = 0$;
b) for any y_d that satisfies Assumption 3 and any $\eta > 0$, there exists an $\varepsilon^* > 0$ such that for all $\varepsilon \in (0, \varepsilon^*]$

$$|e_{ss}(t)| \leq \eta. \tag{23}$$

$$\begin{cases} \dot{x}_0 = (A_0 + B_0 F_0(\varepsilon))x_0 + B_0\tilde{y} + B_0 y_d \\ \dot{\tilde{x}}_1 = \tilde{x}_2 \\ \quad \vdots \\ \dot{\tilde{x}}_{r-1} = \tilde{x}_r \\ \dot{\tilde{x}}_r = -k_1\tilde{x}_1 - k_2\tilde{x}_2 - \cdots - k_r\tilde{x}_r \\ \qquad +(\bar{a}_0(\varepsilon) - \hat{a}_0)x_0 + (\bar{a}_1(\varepsilon) - \hat{a}_1)z_1 + \cdots + (\bar{a}_r(\varepsilon) - \hat{a}_r)z_r + (b - \hat{b})u \\ \tilde{y} = \tilde{x}_1 \end{cases} \tag{16}$$

Proof: We first establish that the solution to the closed-loop system equations under the control law (14)–(21) exists for all times. To this end, left us consider the Lyapunov function candidate,

$$W = x_0' P_0(\varepsilon) x_0 + \beta \tilde{x}' P \tilde{x} + \beta(\bar{a} - \hat{a})' \Gamma^{-1}(\bar{a} - \hat{a}) + \frac{\beta}{\gamma}(b - \hat{b})(b - \hat{b}) \tag{24}$$

where $P_0(\varepsilon) > 0, \varepsilon > 0$ is the unique solution to the Lyapunov equation

$$(A_0 + B_0 F_0(\varepsilon))' P_0 + P_0(A_0 + B_0 F_0(\varepsilon)) = -I \tag{25}$$

and

$$\beta = 2\lambda_{\min}^{-1}(Q)\|P_0(\varepsilon)B_0\|^2. \tag{26}$$

The existence of such a $P(\varepsilon)$ is guaranteed by the fact that $A_0 + B_0 F_0(\varepsilon)$ is asymptotically stable.

The evaluation of the derivative of W along the trajectories of the closed-loop system, using (20) and (21), then yields

$$\begin{aligned} \dot{W} &\le -x_0' x_0 + 2x_0' P_0(\varepsilon) B_0 \tilde{y} + 2x_0' P_0(\varepsilon) B_0 y_d - \beta \tilde{x}' Q \tilde{x} \\ &\quad + 2\beta \tilde{x}' P B(\bar{a} - \hat{a})' z + 2\beta \tilde{x}' P B(b - \hat{b}) u \\ &\quad - 2\beta(\bar{a} - \hat{a})\Gamma^{-1}\dot{\hat{a}} - \frac{2\beta}{\gamma}(b - \hat{b})\dot{\hat{b}} \\ &\le -\frac{1}{2}x_0' x_0 + 2x_0' P_0 B_0 y_d \\ &\le 2\|P_0(\varepsilon)B_0\| y_d^2 \end{aligned} \tag{27}$$

which, together with the fact that y_d is bounded, shows that the solution indeed exists for all times.

To continue our proof, we consider the dynamics of $\tilde{x}$, $\hat{a}$ and $\hat{b}$, to which x_0 is viewed as an external signal. Choose a Lyapunov function candidate as

$$V = \tilde{x}' P \tilde{x} + (\bar{a} - \hat{a})' \Gamma^{-1}(\bar{a} - \hat{a}) + \frac{1}{\gamma}(b - \hat{b})'(b - \hat{b}). \tag{28}$$

Then the derivative of V along the trajectories of (17) and (14) is given by

$$\begin{aligned} \dot{V} &= -\tilde{x}' Q \tilde{x} + 2\tilde{x}' P B(\bar{a} - \hat{a})' z + 2\tilde{x}' P B(b - \hat{b}) u \\ &\quad - 2(\bar{a} - \hat{a})\Gamma^{-1}\dot{\hat{a}} - \frac{2}{\gamma}(b - \hat{b})\dot{\hat{b}} \\ &\le -\tilde{x}' Q \tilde{x} \end{aligned} \tag{29}$$

from which, we conclude that $\tilde{x} \in L_2 \cap L_\infty$ and $\hat{a}, \hat{b} \in L_\infty$. Using the facts that $y_d \in L_\infty$ and $\tilde{y} = \tilde{x}_1 \in L_\infty$, we see from the first equation of (10) that $x_0 \in L_\infty$, which, together with the fact that $y_d^{(i-1)} \in L_\infty$ for $i = 1, 2, \ldots, r$, implies that $z \in L_\infty$. From the definition of u in (14), we see that $u \in L_\infty$. Hence we have shown that all signals of the closed-loop system are bounded.

We now proceed to show the desired properties (a) and (b). It follows from (17) that $\dot{\tilde{x}} \in L_\infty$, which, together with the fact that $\tilde{x} \in L_2$, implies that $\tilde{x}(t) \to 0$ as $t \to \infty$ [11].

In the case that $y_d \equiv 0$, it follows from the first equation of (10) that $\lim_{t\to\infty} x_0(t) = 0$. Using this in (11), we have that $\lim_{t\to\infty} x_i(t) = 0$, for all $i = 1$ to r. This establishes Property (a). To establish Property (b), we consider the first equation of the closed-loop system (10)

$$\dot{x}_0 = (A_0 + B_0 F_0(\varepsilon))x_0 + B_0 \tilde{y} + B_0 y_d. \tag{30}$$

Since $\lim_{t\to\infty} \tilde{y}(t) = \lim_{t\to\infty} \tilde{x}_1(t) = 0$, the steady state trajectory of $\tilde{x}_0$ is all due to the input y_d. Assumption 3 and Lemma 1 then imply that, for any given $\eta > 0$, there exists an $\varepsilon^* > 0$ such that

$$\left|\left(F_0^0(\varepsilon)x_0\right)_{ss}\right| \le \eta, \quad \forall \varepsilon \in (0, \varepsilon^*]. \tag{31}$$

Finally, from the expression

$$e(t) = y - y_d = x_1 - y_d = \tilde{y} + F_0^0(\varepsilon)x_0 \tag{32}$$

and the fact that $\lim_{t\to\infty} \tilde{y}(t) = 0$, we have

$$|e_{ss}(t)| \le \eta, \quad \forall \varepsilon \in (0, \varepsilon^*]. \tag{33}$$

This completes the proof of the theorem. □

In what follows, we use an example to demonstrate the performance of our adaptive low gain state feedback controller.

Example: Consider the following linear: system,

$$\begin{cases} \dot{x}_0 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x_0 + \begin{bmatrix} 0 \\ 1 \end{bmatrix} y, \quad x_0 = [x_{01}, x_{02}]' \\ \dot{x}_1 = x_2 \\ \dot{x}_2 = a_1 x_1 + a_2 x_2 + bu \\ y = x_1 \end{cases} \tag{34}$$

where a_1, a_2 and $b \ge 1$ are unknown parameters. Clearly this system is in the form of (1) and it is weakly nonminimum phase and has a relative degree of 2.

Following the design procedure and choosing

$$k_1 = 1, \quad k_2 = 2, \quad \Gamma = 2I, \quad \gamma = 2, \quad Q = I \tag{35}$$

we construct the family of adaptive control laws parameterized in ε, shown at the bottom of the page, with the adaptation laws given by

$$\begin{cases} \dot{\hat{a}} = \begin{bmatrix} x_0 \\ \varepsilon^2 x_{01} + 2\varepsilon x_{02} + x_1 \\ \varepsilon^2 x_{02} + 2\varepsilon x_1 + x_2 \end{bmatrix} w, \\ \dot{\hat{b}} = \begin{cases} 0 & \text{if } \hat{b} = b_0 \text{ and } wu < 0 \\ wu & \text{otherwise} \end{cases}, \quad \hat{b}(0) \ge 1 \\ w = [\varepsilon^2 x_{01} + (2\varepsilon - \varepsilon^2)x_{02} + (1 - 2\varepsilon)x_1 - x_2 - y_d + \dot{y}_d]u. \end{cases} \tag{37}$$

Figs. 1 and 2 show some simulation results. In these simulations, the initial conditions are chosen as random numbers, $a_1 = 1$, $a_2 = 2$, $b = 2$, the desired input is $y_d = \sin t$, and $\varepsilon = 0.1, 0.01$. It is clear from these simulations that the adaptive control law (36) does achieve the desired properties of the closed-loop system.

IV. Conclusion

In this paper, we have utilized the recent development in low gain feedback design techniques in the design of adaptive state feedback control laws. The resulting adaptive control laws are capable of causing the output of a weakly nonminimum phase linear system to track a desired trajectory to an arbitrarily high degree of accuracy. The extension

$$\begin{cases} u = \frac{v}{\hat{b}} \\ v = y_d + 2\dot{y}_d + \ddot{y}_d - \hat{a}_0 x_0 - [(\hat{a}_1 + 1)\varepsilon^2 \quad 2\hat{a}_1\varepsilon + 2\hat{a}_2\varepsilon^2 + 2\varepsilon + 2\varepsilon^2]x_0 \\ \qquad -(\hat{a}_1 + 2\varepsilon\hat{a}_2 + 1 + 4\varepsilon)x_1 - (\hat{a}_2 + 2)x_2 \end{cases} \tag{36}$$

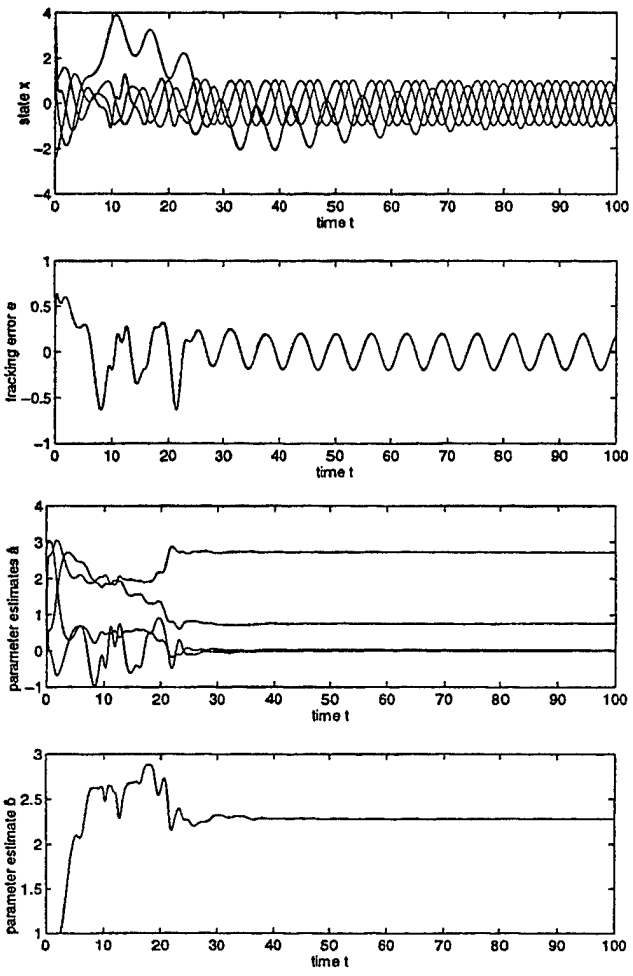


Fig. 1. Closed-loop performance for $\varepsilon = 0.1$.

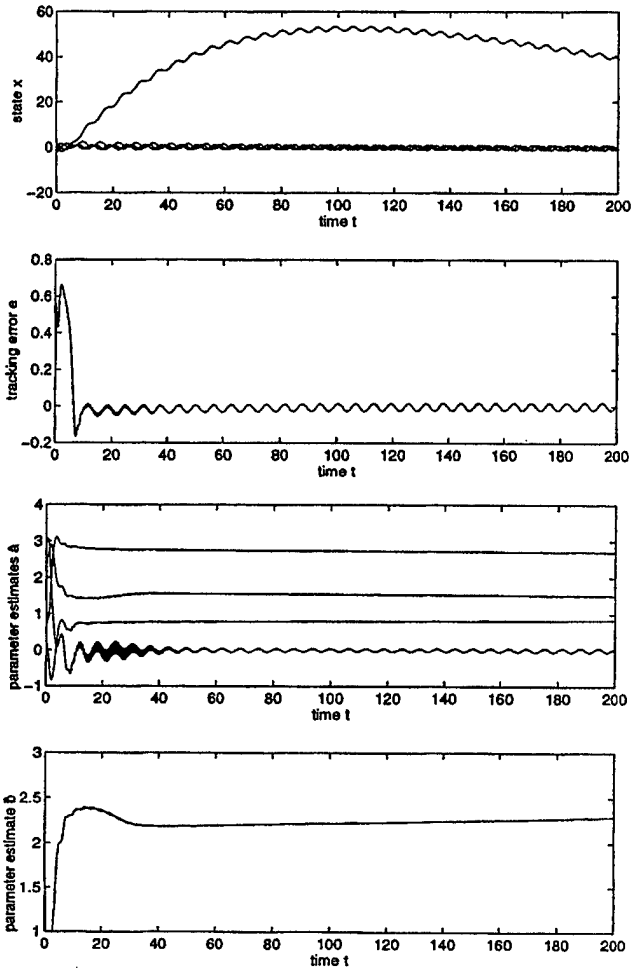


Fig. 2. Closed-loop performance for $\varepsilon = 0.01$.

of this design approach to strictly nonminimum phase systems and to the output feedback case is currently under investigation.

Appendix
Proof of Lemma 1

Let

$$\det(sI - A_0^0) = \prod_{i=1}^{p}(s-\lambda_i)^{n_i}, \quad n_1 + n_1 + \cdots + n_p = n_0^0 \quad \text{(A.1)}$$

where $\lambda_i = j\omega_i$ and $\lambda_i \neq \lambda_j$, $i \neq j$. Then, for each $i = 1$ to p, the n_i generalized eigenvectors of A_0^0 are given by the following ([6]):

$$p_1^i = \begin{bmatrix} 1 \\ \lambda_i \\ \lambda_i^2 \\ \vdots \\ \lambda_i^{n_0^0-2} \\ \lambda_i^{n_0^0-1} \end{bmatrix}, \quad p_2^i = \begin{bmatrix} 0 \\ 1 \\ 2\lambda_i \\ 3\lambda_i^2 \\ \vdots \\ (n_0^0-1)\lambda_i^{n_0^0-2} \end{bmatrix}, \ldots,$$

$$p_{n_i}^i = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ C_{n_0^0-2}^{n_i-1}\lambda_i^{n_0^0-n_i-1} \\ C_{n_0^0-1}^{n_i-1}\lambda_i^{n_0^0-n_i} \end{bmatrix}.$$

Similarly, for each $i = 1$ to p, the n_i generalized eigenvectors of $A + BF(\varepsilon)$ are given by

$$q_1^i = \begin{bmatrix} 1 \\ \lambda_{\varepsilon i} \\ \lambda_{\varepsilon i}^2 \\ \vdots \\ \lambda_{\varepsilon i}^{n_0^0-2} \\ \lambda_{\varepsilon i}^{n_0^0-1} \end{bmatrix}, \quad q_2^i = \begin{bmatrix} 0 \\ 1 \\ 2\lambda_{\varepsilon i} \\ 3\lambda_{\varepsilon i}^2 \\ \vdots \\ (n_0^0-1)\lambda_{\varepsilon i}^{n_0^0-2} \end{bmatrix}, \ldots,$$

$$q_{n_i}^i = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ C_{n_0^0-2}^{n_i-1}\lambda_{\varepsilon i}^{n_0^0-n_i-1} \\ C_{n_0^0-1}^{n_i-1}\lambda_{\varepsilon i}^{n_0^0-n_i} \end{bmatrix}$$

where $\lambda_{\varepsilon i} = -\varepsilon + \lambda_i$ and $C_{n_0^0}^i$ is defined as

$$C_{n_0^0}^i = \begin{cases} \frac{n_0^0!}{(n_0^0-i)!i!} & 0 \le i \le n_0^0 \\ 0 & \text{otherwise} \end{cases}.$$

We next form the following two nonsingular transformation matrices

$$P = [p_1^1, p_2^1, \ldots, p_{n_1}^1, \ldots, p_1^p, p_2^p, \ldots, p_{n_p}^p]$$

$$\begin{aligned}\left|F_0^0(\varepsilon)\left(sI-A_0^0-B_0^0F_0^0(\varepsilon)\right)^{-1}\right| &= \left|\mathcal{L}\left[F_0^0(\varepsilon)e^{\left(A_0^0+B_0^0F_0^0(\varepsilon)\right)t}\right]\right| = \left|\mathcal{L}\left[F_0^0(\varepsilon)Q(\varepsilon)e^{Q^{-1}(\varepsilon)\left(A_0^0+B_0^0F_0^0(\varepsilon)\right)Q(\varepsilon)t}Q^{-1}(\varepsilon)\right]\right| \\ &\le (|P^{-1}|+1)\sum_{i=1}^{p}\left|F_0^0(\varepsilon)\left[q_1^i,q_2^i,\ldots,q_{n_i}^i\right]\mathcal{L}[e^{J_{\varepsilon i}t}]\right| \\ &= \delta\varepsilon(|P^{-1}|+1)\sum_{i=1}^{p}\left|\mathcal{L}\begin{bmatrix} e^{\lambda_{\varepsilon i}t} & te^{\lambda_{\varepsilon i}t} & t^2e^{\lambda_{\varepsilon i}t}/2! & \cdots & t^{n_i-1}e^{\lambda_{\varepsilon i}t}/(n_i-1)! \\ 0 & e^{\lambda_{\varepsilon i}t} & te^{\lambda_{\varepsilon i}t} & \cdots & t^{n_i-2}e^{\lambda_{\varepsilon i}t}/(n_i-2)! \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & e^{\lambda_{\varepsilon i}t}\end{bmatrix}\right| \\ &= \delta\varepsilon(|P^{-1}|+1)\sum_{i=1}^{p}\sum_{j=1}^{n_i}(n_i-j+1)\left|\frac{1}{(s-j\omega_i+\varepsilon)^j}\right| \end{aligned} \tag{A.7}$$

which is independent of ε, and

$$Q(\varepsilon)=\left[q_1^1,q_2^1,\ldots,q_{n_1}^1,\ldots,q_1^p,q_2^p,\ldots,q_{n_p}^p\right].$$

It then follows that

$$P^{-1}A_0^0P=\begin{bmatrix}J_1 & 0 & \cdots & 0\\ 0 & J_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & J_p\end{bmatrix},\quad J_i=\begin{bmatrix}\lambda_i & 1 & 0 & \cdots & 0\\ 0 & \lambda_i & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \cdots & \vdots\\ 0 & 0 & 0 & \cdots & \lambda_i\end{bmatrix} \tag{A.2}$$

and

$$Q^{-1}(\varepsilon)\left(A_0^0+B_0^0F_0^0(\varepsilon)\right)Q(\varepsilon)=\begin{bmatrix}J_{\varepsilon 1} & 0 & \cdots & 0\\ 0 & J_{\varepsilon 2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & J_{\varepsilon p}\end{bmatrix}$$

$$J_{\varepsilon i}=\begin{bmatrix}\lambda_{\varepsilon i} & 1 & 0 & \cdots & 0\\ 0 & \lambda_{\varepsilon i} & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \cdots & \vdots\\ 0 & 0 & 0 & \cdots & \lambda_{\varepsilon i}\end{bmatrix}. \tag{A.3}$$

It was shown in [8] that

$$\left|F_0^0(\varepsilon)q_j^i\right|\le\delta_j^i\varepsilon^{n_i-j+1} \tag{A.4}$$

where δ_j^i is some positive constant independent of ε.

We also note that both $Q(\varepsilon)$ and $Q^{-1}(\varepsilon)$ are continuous functions of ε satisfying $Q(0)=P$ and $Q^{-1}(0)=P^{-1}$. Now from the continuity of norm functions it follows that there exists an $\varepsilon^*\in(0,1]$, such that for all $\varepsilon\in(0,\varepsilon^*]$

$$|Q(\varepsilon)|\le|P|+1,\quad |Q^{-1}(\varepsilon)|\le|P^{-1}|+1. \tag{A.5}$$

We are now ready to show (8). Using the the following fact about the Laplace transform

$$e^{At}=\mathcal{L}^{-1}[(sI-A)^{-1}] \tag{A.6}$$

on the matrix $A_0^0+B_0^0F_0^0(\varepsilon)$, we have (A.7), shown at the top of the page, where $\delta=\max_{i=1}^{p}\max_{j=1}^{n_i}\delta_j^i$.

Hence, we have

$$\left|F_0^0(\varepsilon)\left(j\omega I-A_0^0-B_0^0F_0^0(\varepsilon)\right)^{-1}\right|\le\gamma\varepsilon\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left|\frac{1}{(j\omega-j\omega_i+\varepsilon)^j}\right| \tag{A.8}$$

here $\gamma=\delta(|P^{-1}|+1)n_0^0$. This completes the proof of Lemma 1. □

REFERENCES

[1] K. J. Astrom and B. Wittenmark, *Adaptive Control*, 2nd ed. Reading, MA: Addison-Wesley, 1995.

[2] B. M. Chen, "A simple algorithm for the stable/unstable decomposition of a linear discrete-time system," *International Journal of Control*, vol. 61, no. 1, pp. 255–260, 1995.

[3] G. F. Franklin, J. D. Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems*: Addison Wesley, 1991.

[4] G. C. Goodwin and K. S. Sin, *Adaptive Filtering Prediction and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[5] P. A. Ioannou and J. Sun, *Robust Adaptive Control*: Prentice Hall, 1996.

[6] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1980.

[7] M. Krstić, I. Kanellakopoulos, and P. V. Kokotović, *Nonlinear and Adaptive Control Design*, New York: John Wiley & Sons, 1995.

[8] Z. Lin, *Low Gain Feedback*, London: Springer, 1999, vol. 240, Lecture Notes in Control and Information Sciences.

[9] K. S. Narendra and A. M. Annaswamy, *Stable Adaptive Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[10] S. Sastry and M. Bodson, *Adaptive Control: Stability, Convergence, and Robustness*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[11] G. Tao, "A simple alternative to the Barbalat lemma," *IEEE Transactions on Automatic Control*, vol. 42, no. 5, p. 698, 1997.

Publication 6

ELSEVIER

Systems & Control Letters 40 (2000) 59–69

SYSTEMS & CONTROL LETTERS

www.elsevier.com/locate/sysconle

On enlarging the basin of attraction for linear systems under saturated linear feedback ☆

Tingshu Hu *, Zongli Lin

Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903, USA

Received 9 May 1999; received in revised form 9 September 1999; accepted 18 December 1999

Abstract

We consider the problem of enlarging the basin of attraction for a linear system under saturated linear feedback. An LMI-based approach to this problem is developed. For discrete-time system, this approach is enhanced by the lifting technique, which leads to further enlargement of the basin of attraction. The low convergence rate inherent with the large invariant set (hence, the large basin of attraction) is prevented by the construction of a sequence of invariant ellipsoids nested within the large one obtained. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Basin of attraction; Invariant set; Saturation; Lifting technique

1. Introduction

The notion of invariant set has played an important role in the analysis and design of dynamical systems (see, e.g., [2,3,7,15] and the references therein). For a stable linear system, a simple and popular type of invariant set is the level set $\Omega(P) = \{x: x^{\mathrm{T}}Px \leqslant 1\}$, associated with the Lyapunov function $V(x) = x^{\mathrm{T}}Px$. Another type of well-studied invariant set is polyhedra (see, e.g., [1,7]).

For linear systems under saturated stabilizing linear state feedback, both the problem of estimating the basin of attraction (the largest invariant set) for a specific feedback gain matrix and that of searching for an appropriate feedback gain matrix to result in a large basin of attraction are of paramount importance and have attracted a great deal of attention from the control research community. Although these problems are still far from being completely solved, recent literature shows that they have been examined extensively from various aspects (see, e.g., [1,6,9,11] and the recent survey paper [2]). In particular, in [9], we investigated continuous-time linear systems under saturated stabilizing linear feedback. We showed that, if the system is of second order and has both open-loop poles in the open right half-plane, the boundary of the basin of attraction is the unique unstable limit cycle of the closed-loop system and can be easily obtained from its time-reversed system. Moreover, a family of gain matrices can be designed to obtain a basin of attraction that is arbitrarily close to the null controllable region, the largest possible basin of attraction under any bounded

☆ This work was supported in part by the US Office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

* Corresponding author.

E-mail address: th7f@virginia.edu (T. Hu)

0167-6911/00/$ - see front matter © 2000 Elsevier Science B.V. All rights reserved.

PII: S0167-6911(00)00002-5

controls. Though complete for the second-order continuous-time systems, these results cannot be extended in an obvious way to either general higher-order systems or discrete-time systems, which do not always have time-reversed systems.

For general linear systems under saturated linear stabilizing feedback, usually only an estimate of the basin of attraction can be obtained. A simple way of estimating the basin of attraction is the largest linear invariant ellipsoid associated with a quadratic Lyapunov function. By linear invariant ellipsoid, we mean an invariant ellipsoid that is completely within the linear region of the saturation function. This estimate, though often conservative, can be improved by an appropriate choice of the Lyapunov function and the feedback gain matrix. For example, in the case that the open-loop system is not exponentially unstable, the linear invariant ellipsoid can be made large enough to cover any a priori given (arbitrarily large) bounded set [11].

The objective of this paper is to present a systematic approach to the design of the feedback laws that result in large basin of attraction for general linear systems, both in continuous- and in discrete-time. More specifically, we will present an LMI-based approach to maximizing the linear invariant ellipsoid. Given a reference set X_R, the maximization is in the sense that the linear invariant ellipsoid contains the set αX_R with α being maximized. In the case that the open-loop system is not exponentially unstable, our design results in linear invariant ellipsoid that includes any a priori given (arbitrarily large) bounded set as a subset. This is the so-called semi-global stabilization [11–13]. For discrete-time systems, we show that this approach can be enhanced by the lifting technique, which leads to further enlargement of the basin of attraction. Finally, the low convergence rate inherent with the large basin of attraction can be prevented by constructing a sequence of invariant ellipsoids nested within the large one obtained and optimizing the convergence rate of the piecewise linear controller of Wredenhagen and Belanger [15].

The remainder of this paper is organized as follows. In Section 2, we present an LMI approach to the maximization of the linear invariant ellipsoid for both continuous- and discrete-time systems. In Section 3, we show how the lifting technique can be used to further enlarge the basin of attraction for discrete-time systems. In Section 4, we show how the closed-loop system convergence rate can be increased by switching the feedback gains between nested sequence of linear invariant ellipsoids. Two examples are included in Section 5 to demonstrate the effectiveness of the proposed design techniques. Concluding remarks are made in Section 6.

2. Maximizing the linear invariant ellipsoid

Consider the system

$$x(k+1) = Ax(k) + B\sigma(u(k)), \quad x \in \mathbb{R}^n, \ u \in \mathbb{R}^m, \tag{1}$$

where (A,B) is stabilizable. In this paper, we use $\sigma(\cdot)$ to denote a standard saturation function of appropriate dimensions. For example, in the above system, $\sigma: \mathbb{R}^m \to \mathbb{R}^m$, and $\sigma(u) = [\sigma(u_1), \sigma(u_2), \ldots, \sigma(u_m)]^T$, where $\sigma(u_i) = \mathrm{sgn}(u_i)\min\{1, |u_i|\}$. The closed-loop system under the feedback law $u = Fx$ is given by

$$x(k+1) = Ax(k) + B\sigma(Fx(k)). \tag{2}$$

Let f_i be the ith row of F. Denote the linear region of system (2) as

$$L(F) := \{x \in \mathbb{R}^n : |f_i x| \leqslant 1, i = 1, 2, \ldots, m\}.$$

Let F be such that $A + BF$ has all its eigenvalues inside the unit circle, then there exists $P > 0$ such that

$$(A + BF)^T P (A + BF) - P < 0.$$

Denote the Lyapunov level set as $\Omega(P) := \{x \in \mathbb{R}^n : x^T P x \leqslant 1\}$. If $\Omega(P) \subset L(F)$, then $\Omega(P)$ is an invariant set and we call it a linear invariant ellipsoid.

Our design objective is to choose F and P such that $\Omega(P) \subset L(F)$ is maximized in some sense. In the literature, e.g., [4], the largeness of a set is usually measured by its volume. Here we will take its shape

into consideration and will maximize $\Omega(P)$ with respect to some reference set. To make the problem well formulated, we introduce a reference set X_R. Let X_R be a bounded convex set, denote

$$\alpha X_R = \{\alpha x : x \in X_R\}.$$

The linear invariant ellipsoid $\Omega(P)$ is said to be maximized over F and P if α is maximized such that $\alpha X_R \subset \Omega(P) \subset L(F)$. In practice, the shape of X_R can be determined by the partial knowledge of the initial states. It can also be chosen according to the shape of the null controllable region, as identified in [9,8]. In this paper, we will consider two types of X_R:

- The polygon: $X_R = \mathrm{co}\{x_1, x_2, \ldots, x_l\}$ is the convex hull of a given set of states $x_1, x_2, \ldots, x_l \in \mathbb{R}^n$;
- The ellipsoid: $X_R = \{x \in \mathbb{R}^n : x^T R x \leqslant 1\}$, $R > 0$.

Given system (1) and X_R, the optimization problem can be described as follows:

$$\begin{aligned} &\sup_{P>0,F} && \alpha \\ &\text{s.t.} && \text{(a) } \alpha X_R \subset \Omega(P), \\ &&& \text{(b) } \Omega(P) \subset L(F), \\ &&& \text{(c) } (A+BF)^T P (A+BF) - P < 0. \end{aligned} \tag{3}$$

We also define the supremum of α as α^*.

If X_R is a polygon, then constraint (a) is equivalent to

$$\alpha^2 x_i^T P x_i \leqslant 1, \quad i = 1, 2, \ldots, l. \tag{4}$$

If X_R is an ellipsoid, then constraint (a) is equivalent to

$$\alpha^2 P \leqslant R. \tag{5}$$

On the other hand, constraint (b) is equivalent to

$$\min\{x^T P x : f_i x = 1\} \geqslant 1, \quad i = 1, 2, \ldots, m. \tag{6}$$

To see this, note that $\Omega(P) \subset L(F)$ if and only if all the hyperplanes $f_i x = \pm 1$, $i = 1, 2, \ldots, m$, lie completely outside of $\Omega(P) = \{x \in \mathbb{R}^n : x^T P x \leqslant 1\}$, i.e., at each point x on the hyperplanes $f_i x = \pm 1$, we have $x^T P x \geqslant 1$.

The left-hand side of (6) is a convex optimization problem and has a unique minimum. By using the Lagrange multiplier method, we obtain

$$\min\{x^T P x : f_i x = 1\} = (f_i P^{-1} f_i^T)^{-1}.$$

Consequently, constraint (b) is equivalent to

$$f_i P^{-1} f_i^T \leqslant 1, \quad i = 1, 2, \ldots, m. \tag{7}$$

Thus, if X_R is a polygon, then (3) can be rewritten as follows:

$$\begin{aligned} &\sup_{P>0,F} && \alpha \\ &\text{s.t.} && \text{(a) } \alpha^2 x_i^T P x_i \leqslant 1, \quad i = 1, 2, \ldots, l, \\ &&& \text{(b) } f_i P^{-1} f_i^T \leqslant 1, \quad i = 1, 2, \ldots, m, \\ &&& \text{(c) } (A+BF)^T P (A+BF) - P < 0. \end{aligned} \tag{8}$$

If X_R is an ellipsoid, we just need to replace (a) with (5).

Constraints (a)–(c) are nonlinear and convex. The standard tool to transform such constraints into LMI is Schur complements: Suppose $Q > 0$, then the LMI

$$\begin{bmatrix} R & S \\ S^T & Q \end{bmatrix} \geqslant 0$$

if and only if $R-SQ^{-1}S^{\mathrm{T}}\geqslant 0$. Let $\gamma=1/\alpha^2$, $Q=P^{-1}$, $Y=FP^{-1}$, then we can transform (8) into the following LMI problem:

$$\begin{aligned}
&\inf_{Q,Y} \quad \gamma\\
&\text{s.t.} \quad \text{(a)} \begin{bmatrix} \gamma & x_i^{\mathrm{T}} \\ x_i & Q \end{bmatrix} \geqslant 0, \quad i=1,2,\ldots,l,\\
&\qquad \text{(b)} \begin{bmatrix} 1 & y_i \\ y_i^{\mathrm{T}} & Q \end{bmatrix} \geqslant 0, \quad i=1,2,\ldots,m,\\
&\qquad \text{(c)} \begin{bmatrix} Q & QA^{\mathrm{T}}+Y^{\mathrm{T}}B^{\mathrm{T}} \\ AQ+BY & Q \end{bmatrix} > 0,
\end{aligned} \tag{9}$$

where we have used y_i to denote the ith row of Y. For the case where X_{R} is an ellipsoid, we can simply replace (a) in (9) with $\gamma Q\geqslant R^{\vdash}$. We will denote the infimum of γ in the above optimization problem as $\gamma^*=1/(\alpha^*)^2$.

Remark 1. When $\gamma=\gamma^*$, there may not exist Q and Y that satisfy (a)–(c) in (9). In this case, we can choose $\gamma=\gamma^*+\varepsilon$ with ε arbitrarily small and solve for feasible solutions satisfying the constraints. For example, suppose A has all its eigenvalues on or inside the unit circle, then $\gamma^*=0$ [11] and no $Q>0$ satisfies (a) or (5). By taking γ arbitrarily small, we can make the set $\alpha X_{\mathrm{R}}\subset\Omega(P)\subset L(F)$ arbitrarily large, i.e., semi-global stabilization [12,13] can be achieved.

Remark 2. The above optimization method can be easily adapted to the continuous-time system by replacing (c) in (8) with $(A+BF)^{\mathrm{T}}P+P(A+BF)<0$ and (c) in (9) with $QA^{\mathrm{T}}+AQ+Y^{\mathrm{T}}B^{\mathrm{T}}+BY<0$.

3. Further enlargement of basin of attraction via lifting technique

The lifting technique has been used to improve the robust performance of discrete-time systems in [10] and to design semi-global stabilizing controller in [5]. Here we will show that it can also be efficiently used to enlarge the basin of attraction. Let $N\geqslant 1$ be a positive integer. Denoting

$$\bar{A}=A^N, \quad \bar{B}=[A^{N-1}B\ A^{N-2}B\ \cdots B]$$

and

$$\bar{x}(k)=x(kN), \quad \bar{u}(k)=\begin{bmatrix} u(kN) \\ u(kN+1) \\ \vdots \\ u(kN+N-1) \end{bmatrix},$$

we obtain the lifted N-step system

$$\bar{x}(k+1)=\bar{A}\bar{x}(k)+(\bar{B}\sigma\ \bar{u}(k)), \quad \bar{x}\in\mathbb{R}^n,\ \bar{u}\in\mathbb{R}^{Nm}. \tag{10}$$

Let $\bar{u}(k)=\bar{F}\bar{x}(k)$, $\bar{F}\in\mathbb{R}^{Nm\times n}$ be a stabilizing feedback. The closed-loop system is

$$\bar{x}(k+1)=\bar{A}\bar{x}(k)+(\bar{B}\sigma\ \bar{F}\bar{x}(k)). \tag{11}$$

Similar to the one-step case, the problem of maximizing the linear invariant ellipsoid can be described as

$$\begin{aligned}
&\sup_{P>0,\bar{F}} \quad \alpha\\
&\text{s.t.} \quad \text{(a) } \alpha X_{\mathrm{R}}\subset\Omega(P),\\
&\qquad \text{(b) } \Omega(P)\subset L(\bar{F}) \quad (\text{or } \bar{f}_i P^{-1}\bar{f}_i^{\mathrm{T}}\leqslant 1,\ i=1,2,\ldots,Nm),\\
&\qquad \text{(c) } (\bar{A}+\bar{B}\bar{F})^{\mathrm{T}}P(\bar{A}+\bar{B}\bar{F})-P<0
\end{aligned} \tag{12}$$

which can be solved by the LMI approach proposed in the previous section. Denoting the supremum of α as $\alpha^*(N)$, we have the following theorem that justifies the use of lifting technique.

Theorem 1. *For any integers* $p, N \geqslant 1$, $\alpha^*(p) \leqslant \alpha^*(pN)$.

Proof. *Case* 1: $p = 1$. Denote the set of feasible (α, P) satisfying constraints (a)–(c) as

$$\Phi(N) = \{(\alpha, P): \exists \bar{F} \text{ s.t. (a), (b) and (c) are satisfied}\}.$$

It suffices to show that $\Phi(1) \subset \Phi(N)$.

Suppose that $(\alpha, P) \in \Phi(1)$, then there exists an $F \in \mathbb{R}^{m\times n}$ such that

$$f_i P^{-1} f_i^{\mathrm{T}} \leqslant 1, \quad i = 1, 2, \ldots, m \tag{13}$$

and

$$(A + BF)^{\mathrm{T}} P (A + BF) - P < 0 \tag{14}$$

which is equivalent to

$$\begin{bmatrix} P & (A+BF)^{\mathrm{T}} \\ A+BF & P^{-1} \end{bmatrix} > 0$$

and to

$$(A + BF) P^{-1} (A + BF)^{\mathrm{T}} - P^{-1} < 0. \tag{15}$$

Let

$$\bar{F} = \begin{bmatrix} F \\ F(A+BF) \\ \vdots \\ F(A+BF)^{N-1} \end{bmatrix},$$

then

$$\bar{A} + \bar{B}\bar{F} = A^N + A^{N-1}BF + A^{N-2}BF(A+BF) + \cdots + BF(A+BF)^{N-1} = (A+BF)^N.$$

It then follows from (14) that

$$(\bar{A} + \bar{B}\bar{F})^{\mathrm{T}} P (\bar{A} + \bar{B}\bar{F}) = ((A+BF)^{\mathrm{T}})^N P (A+BF)^N < ((A+BF)^{\mathrm{T}})^{N-1} P (A+BF)^{N-1} < \cdots < P$$

which shows that P and $\bar{F}$ satisfy constraint (c).

Since $\bar{f}_j = f_i (A+BF)^q$ for some $i \leqslant m$, $q \leqslant N-1$, we have

$$\bar{f}_j P^{-1} \bar{f}_j^{\mathrm{T}} = f_i (A+BF)^q P^{-1} ((A+BF)^{\mathrm{T}})^q f_i^{\mathrm{T}}.$$

It follows from (13) and (15) that

$$\bar{f}_j P^{-1} \bar{f}_j^{\mathrm{T}} \leqslant f_i (A+BF)^{q-1} P^{-1} ((A+BF)^{\mathrm{T}})^{q-1} f_i^{\mathrm{T}} \leqslant \cdots \leqslant f_i P^{-1} f_i^{\mathrm{T}} \leqslant 1$$

which shows that P and $\bar{F}$ also satisfy constraint (b). Hence $(\alpha, P) \in \Phi(N)$.

Case 2: $p > 1$. Let

$$\hat{A} = A^p, \quad \hat{B} = [A^{p-1}B \ A^{p-2}B \ \cdots B]$$

and

$$\bar{A} = A^{pN}, \quad \bar{B} = [A^{pN-1}B \ A^{pN-2}B \ \cdots B],$$

then

$$\bar{A} = \hat{A}^N, \quad \bar{B} = [\hat{A}^{N-1}\hat{B} \ \hat{A}^{N-2}\hat{B} \cdots \hat{B}].$$

Suppose we first lift system (1) with step p to get $\hat{x}(k)=x(kp)$,

$$\hat{x}(k+1)=\hat{A}\hat{x}(k)+\hat{B}\sigma(\hat{u}(k)),$$

then lift the above system with step N to get $\bar{x}(k)=\hat{x}(kN)=x(kNp)$,

$$\bar{x}(k+1)=\bar{A}\bar{x}(k)+(\bar{B}\sigma\ \bar{u}(k)).$$

Applying the result in Case 1, we immediately have $\alpha^*(p)\leqslant\alpha^*(pN)$. □

Remark 3. The equality $\alpha^*(p)=\alpha^*(pN)$ with $N>1$ can occur in some special cases. For example, let $A=a>1$, $B=1$, and $X_R=[-1,1]$. It can be verified that $\alpha^*(N)=1/(a-1)$ for all $N\geqslant 1$.

From the above theorem, we see that

$$\alpha^*(1)\leqslant\alpha^*(2)\leqslant\alpha^*(4)\leqslant\alpha^*(8)\cdots,$$

$$\alpha^*(1)\leqslant\alpha^*(3)\leqslant\alpha^*(6)\leqslant\alpha^*(12)\cdots.$$

But $\alpha^*(N_1)\leqslant\alpha^*(N_2)$ does not necessarily hold for all $N_1<N_2$. It should also be noted that, because of lifting, the resulting $\Omega(P)$ is not necessarily invariant for the original system at each step (see Fig. 2).

4. Performance improvement

Inherent with the achieved large basin of attraction is however the low convergence rate. To improve the convergence performance, we can use the idea of piecewise linear control [15] to design a set of nested ellipsoids $\Omega(P_M)\subset\Omega(P_{M-1})\subset\cdots\subset\Omega(P_1)\subset\Omega(P_0)$, such that when the state enters an inner ellipsoid, the controller is switched to another feedback which makes this ellipsoid invariant with an increased convergence rate. Here we would like to explore the possibility of further increasing the overall convergence rate by maximizing the convergence rate in each of the nested ellipsoids. The nested invariant sets can be simply chosen by setting

$$P_i=\beta_i P_0,\quad 1<\beta_1<\beta_2<\cdots<\beta_M.$$

The convergence rate inside $\Omega(P)$ under a feedback $u=Fx$ can be measured by a positive number $c<1$ such that

$$(A+BF)^{\mathrm{T}}P(A+BF)-cP\leqslant 0. \tag{16}$$

We note that such F and c always exist for any $P=P_i$ since P_0, F_0 and $c=1$ satisfy (16). Smaller c indicates faster convergence rate. Now let $P=\beta P_0$ be fixed, we need to design F such that c is minimized. The problem can be stated as follows. For a given β,

$$\begin{aligned}&\min_{F}\quad c\\ &\text{s.t.}\quad \text{(a) } f_i P_0^{-1} f_i^{\mathrm{T}}\leqslant\beta,\quad i=1,2,\ldots,m,\quad (\Omega(\beta P_0)\subset L(F)),\\ &\qquad\text{(b) } (A+BF)^{\mathrm{T}}P_0(A+BF)-cP_0\leqslant 0.\end{aligned} \tag{17}$$

We denote the minimum of c as $c^*(\beta)$. For the lifted N-step controller design, we can replace A,B and F with $\bar{A},\bar{B}$ and $\bar{F}$, respectively. As in the previous sections, this optimization problem can also be put into the LMI framework. We note here that other performance criteria can also be formulated into a similar optimization problem (see, e.g., [14]).

Proposition 1. *$c^*(\beta)$ is decreased as β is increased. If B has full row rank, then there exists a β_0 such that $c^*(\beta)=0$ for all $\beta>\beta_0$.*

Table 1
The increase of $\alpha^*(N)$

N	1	2	4	8	16	32
$\alpha^*(N)$	1.0650	1.0930	1.1896	1.4017	1.5164	1.5426

Proof. Constraint (b) in (17) is equivalent to

$$P_0^{-(1/2)}(A+BF)^{\mathrm{T}}P_0(A+BF)P_0^{-(1/2)} \leqslant cI.$$

Hence

$$\begin{aligned} c^*(\beta) = \min \quad & \lambda_{\max}[P_0^{-(1/2)}(A+BF)^{\mathrm{T}}P_0(A+BF)P_0^{-(1/2)}] \\ \text{s.t.} \quad & f_i P_0^{-1} f_i^{\mathrm{T}} \leqslant \beta, \quad i = 1,2,\ldots,m. \end{aligned}$$

As β is increased, the constraint $f_i P_0^{-1} f_i^{\mathrm{T}} \leqslant \beta$ becomes less restrictive, hence $c^*(\beta)$ will decrease.

If B has full row rank, then there exists F_1 such that $A + BF_1 = 0$. Let the ith row of F_1 be f_{1i}. Let $\beta_0 = \max\{f_{1i}P_0^{-1}f_{1i}^{\mathrm{T}} : i = 1,2,\ldots,m\}$, then for all $\beta > \beta_0$, $c^*(\beta) = 0$. □

Usually, for system (1), B does not have full row rank. For the lifted system (10), if (A,B) is controllable, then $\bar{B}$ will have full row rank when $N \geqslant n$. We will see in the examples that the lifting design method is efficient not only in enlarging the linear invariant ellipsoid, but also in increasing the convergence rate.

Now, let $1 < \beta_1 < \beta_2 < \cdots < \beta_M$ be a sequence of numbers. Denote the optimal solution of (17) corresponding to β_i as c_i^* and F_i^*. A switching feedback law can be designed as

$$u(k) = \begin{cases} F_0 x(k) & \text{if } x(k) \in \Omega(P_0)\backslash\Omega(\beta_1 P_0), \\ F_1^* x(k) & \text{if } x(k) \in \Omega(\beta_1 P_0)\backslash\Omega(\beta_2 P_0), \\ \quad\vdots & \\ F_M^* x(k) & \text{if } x(k) \in \Omega(\beta_M P_0). \end{cases}$$

In the set $\Omega(\beta_i P_0) \setminus \Omega(\beta_{i+1}P_0)$, the convergence rate is c_i^*. As the state enters the inner set $\Omega(\beta_{i+1}P_0)\backslash \Omega(\beta_{i+2}P_0)$, the convergence rate is increased to c_{i+1}^*.

5. Examples

Example 1. Consider a second-order system in the form of (1) with

$$A = \begin{bmatrix} 0.9510 & 0.5408 \\ -0.2704 & 1.7622, \end{bmatrix}, \quad B = \begin{bmatrix} 0.0980 \\ 0.5408 \end{bmatrix}.$$

A has two unstable eigenvalues $\{1.2214, 1.4918\}$. The reference set $X_{\mathrm{R}} = \{x \in \mathbb{R}^2 : x^{\mathrm{T}}Rx \leqslant 1\}$, where

$$R = \begin{bmatrix} 1.2862 & -1.0310 \\ -1.0310 & 4.7138 \end{bmatrix},$$

is chosen according to the shape of the null controllable region. Table 1 shows the computational result for $\alpha^*(N), N = 1,2,4,8,16,32$.

Fig. 1 shows the effectiveness of the lifting design. The innermost curve is the boundary of $\alpha^*(1)X_{\mathrm{R}}$. For $N = 2,4,8,16,32$, the set $\alpha^*(N)X_{\mathrm{R}}$ grows bigger. The outermost curve is the boundary of the null controllable region obtained by the method proposed in [8].

We see that the increase from $\alpha^*(16)$ to $\alpha^*(32)$ is small so we take $N = 16$ as the lifting step. Now we design for the 16-step lifted system a set of nested invariant ellipsoids to accelerate the convergence rate. The optimal P_0 corresponding to $\alpha^*(16)$ is

$$P_0 = \begin{bmatrix} 0.5593 & -0.4483 \\ -0.4483 & 2.0497 \end{bmatrix} = 0.4348R = \frac{1}{(\alpha^*(16))^2}R.$$

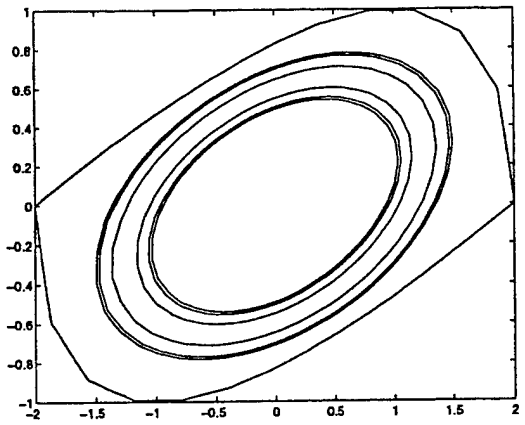


Fig. 1. The sets $\alpha^*(N)X_R$.

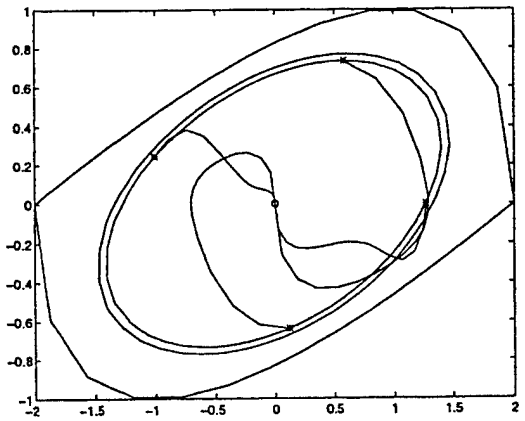


Fig. 2. The deadbeat control.

So $\Omega(P_0) = \alpha^*(16)X_R$. The optimal feedback is

$$\bar{F}_0^T = \begin{bmatrix} 0.3504 & 0.4636 & 0.6129 & 0.7324 & 0.7279 & 0.6374 & 0.5467 & 0.4777 \\ -1.4294 & -1.3917 & -1.2360 & -0.8490 & -0.2872 & 0.1679 & 0.4481 & 0.6167 \end{bmatrix}$$

$$\begin{bmatrix} 0.4279 & 0.3918 & 0.3653 & 0.3454 & 0.3302 & 0.3185 & 0.3094 & 0.3021 \\ 0.7225 & 0.7924 & 0.8406 & 0.8750 & 0.9003 & 0.9193 & 0.9337 & 0.9447 \end{bmatrix}.$$

The eigenvalues of $\bar{A} + \bar{B}\bar{F}$ are $0.2758 \pm j0.8814$, which indicates a low convergence rate.

We take $\beta = 1.04, 1.08, 1.1$, and get the corresponding $c^*(\beta)$ as $0.2650, 0.005, 0$. This shows that the convergence rate is accelerated. The fact that $c^*(1.1) = 0$ implies that all the states in $\Omega(1.1P_0)$ can be steered to the origin in 16 steps (counted for the original unlifted system) by a linear feedback controller. The deadbeat feedback matrix is

$$\bar{F}_0^T = \begin{bmatrix} 0.3115 & 0.4671 & 0.6665 & 0.7789 & 0.7236 & 0.6235 & 0.5433 & 0.4864 \\ -1.4990 & -1.4655 & -1.2419 & -0.6839 & -0.0779 & 0.3118 & 0.5342 & 0.6662 \end{bmatrix}$$

$$\begin{bmatrix} 0.4461 & 0.4170 & 0.3953 & 0.3788 & 0.3659 & 0.3555 & 0.3470 & 0.3398 \\ 0.7494 & 0.8047 & 0.8428 & 0.8697 & 0.8887 & 0.9020 & 0.9110 & 0.9164 \end{bmatrix}.$$

Fig. 2 illustrates this design result, where the innermost ellipsoid is $\Omega(1.1P_0)$ and the larger ellipsoid is $\Omega(P_0) = \alpha^*(16)X_R$. The outermost curve is the boundary of the null controllable region. The initial states

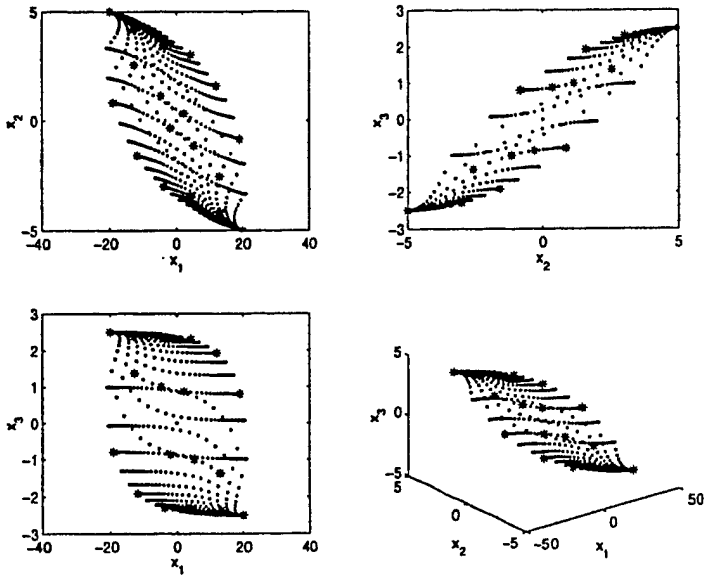


Fig. 3. The vertices of X_R.

Table 2
The increase of $\alpha^*(N)$

N	1	2	4	8	16	32
$\alpha^*(N)$	0.4274	0.4382	0.4593	0.4868	0.5564	0.6041

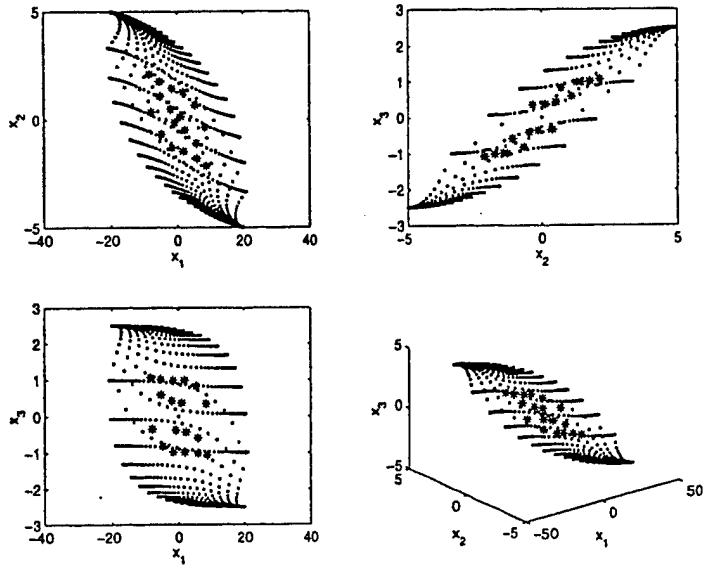


Fig. 4. $\alpha^*(1)X_R$an d the null controllable region.

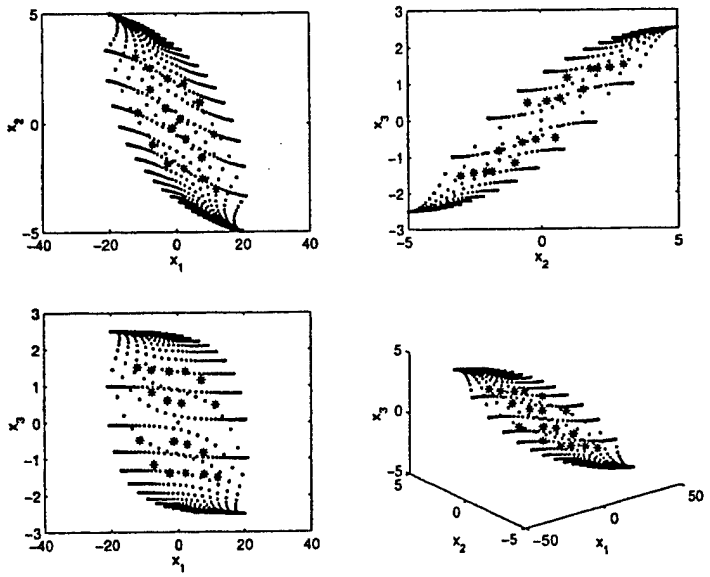


Fig. 5. $\alpha^*(32)X_R$an d the null controllable region.

on the boundary of $\Omega(1.1P_0)$ are marked with "*". They are all driven to the origin by the linear feedback control in 16 steps. Fig. 2 also shows some trajectories of the unlifted system under this 16-step control law.

Example 2. Consider a third-order system in the form of (1) with

$$A = \begin{bmatrix} 1.1972 & 1.0775 & 0 \\ 0 & 1.1972 & 0 \\ 0 & 0 & 1.4333 \end{bmatrix}, \quad B = \begin{bmatrix} 1.4431 \\ 0.9861 \\ 1.0833 \end{bmatrix}.$$

All of the eigenvalues of A are unstable. For the purpose of comparison, we choose 18 points on the boundary of the null controllable region as the vertices of X_R (see Fig. 3), where the vertices of X_R are marked with '*' and the vertices of the null controllable region are marked with '.'. Table 2 shows the computational result for $\alpha^*(N)$, $N = 1,2,4,8,16,32$.

We also see that $\alpha^*(N)$ increases significantly as N is increased. (See Fig. 4 for the vertices of $\alpha^*(1)X_R$ and Fig. 5 for the vertices of $\alpha^*(32)X_R$, both in comparison with the null controllable region.)

6. Conclusions

We have proposed an LMI-based approach to the maximization of the linear invariant ellipsoid for linear systems under saturated linear feedback. The proposed approach applies to both continuous- and discrete-time systems. For discrete-time systems, we also showed that the lifting technique can be used to further enlarge the basin of attraction. Finally, the low convergence rate inherent with the large basin of attraction is increased by switching feedback laws between a sequence of nested invariant ellipsoids. Two examples are worked out to demonstrate the effectiveness of the proposed design techniques.

References

[1] G. Bitsoris, On the positive invariance of polyhedral sets for discrete-time systems, Systems Control Lett. 11 (1988) 243–248.
[2] F. Blanchini, Feedback control for LTI systems with state and control bounds in the presence of disturbances, IEEE Trans. Automat. Control 35 (1990) 1231–1243.

[3] F. Blanchini, Set invariance in control – a survey, Automatica 35 (1999) 1747–1767.
[4] S. Boyd, L. El Ghaoui, E. Feron, V. Balakrishnan, Linear Matrix Inequalities in Systems and Control Theory, SIAM Studies in Appl. Mathematics, Philadelphia, 1994.
[5] J. Collado, R. Lozano, A. Ailon, Semi-global stabilization of discrete-time systems with bounded inputs using periodic controller, Systems Control Lett. 36 (1999) 267–275.
[6] E.G. Gilbert, K.T. Tan, Linear systems with state and control constraints: the theory and application of maximal output admissible sets, IEEE Trans. Automat. Control 36 (1991) 1008–1020.
[7] P.O. Gutman, M. Cwikel, Admissible sets and feedback control for discrete-time linear dynamical systems with bounded control and dynamics, IEEE Trans. Automat Control 31 (1986) 373–376.
[8] T. Hu, D.E. Miller, L. Qiu, Controllable regions of LTI discrete-time systems with input saturation, Proceedings of 1998 CDC, 1998, pp. 371–376.
[9] T. Hu, Z. Lin, L. Qiu, The controllability and stabilizability of exponentially unstable linear systems with saturating actuators, submitted for publication.
[10] P.P. Khargonekar, K. Poolla, A. Tannenbaum, Robust control of linear time invariant plants using periodic compensation, IEEE Trans. Automat Control 30 (1985) 1088–1096.
[11] Z. Lin, Low Gain Feedback, Lecture Notes in Control and Information Sciences, Vol. 240, Springer, London, 1998.
[12] Z. Lin, A. Saberi, Semi-global exponential stabilization of linear systems subject to 'input saturation' via linear feedbacks, Systems Control Lett. 21 (1993) 225–239.
[13] Z. Lin, A. Saberi, Semi-global exponential stabilization of linear discrete-time systems subject to input saturation via linear feedbacks, Systems Control Lett. 24 (1995) 125–132.
[14] M. Sznaier, A set induced norm approach to the robust control of constrained systems, SIAM J. Control Optim. 31 (3) (1993) 733–746.
[15] G.F. Wredenhagen, P.R. Belanger, Piecewise-linear LQ control for systems with input constraints, Automatica 30 (1994) 403–416.

Publication 7

PERGAMON Journal of the Franklin Institute 337 (2000) 691–712

Journal *of* The Franklin Institute

www.elsevier.nl/locate/jfranklin

On $L_p$input to state stabilizability of affine in control, nonlinear systems subject to actuator saturation☆

Xiangyu Bao, Zongli Lin*

Department of Electrical Engineering, University of Virginia, Thornton Hall, Room E313, Charlottesville, VA 22903, USA

Received 30 August 1999; received in revised form 3 May 2000

Abstract

The L_p input to state stabilizability of affine in control nonlinear systems subject to actuator saturation is examined. A few sets of conditions under which the system is (finite gain) L_p input to state stabilizable are identified and the stabilizing feedback laws are explicitly constructed. © 2000 The Franklin Institute. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Nonlinear systems; Actuator saturation; Input output stabilization; Lyapunov analysis

1. Introduction

Recently, there has been a surge of interest in control systems with saturating actuators. This surge of interest has led to many fundamental results (see, for example, [1,2] and the references therein). Most of these results however pertain to linear plants. In this paper, we continue some previous efforts on the stabilization of nonlinear systems with saturating actuators. More specifically, we consider the following affine in control nonlinear system subject to actuator saturation:

$$\dot{x} = f(x) + g(x)\sigma(u), \quad x \in \mathbb{R}^n, \quad u \in \mathbb{R}, \tag{1}$$

☆ Work supported in part by the US office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

*Corresponding author. Tel.: +1-804-924-6342; fax: +1-804-924-8818.

E-mail address: zl5y@virginia.edu (Z. Lin).

0016-0032/00/$20.00 © 2000 The Franklin Institute. Published by Elsevier Science Ltd. All rights reserved.

PII: S0016-0032(00)00046-6

where $f: \mathbb{R}^n \to \mathbb{R}^n$ and $g: \mathbb{R}^n \to \mathbb{R}^n$ are smooth functions with $f(0) = 0$, $\sigma: \mathbb{R} \to \mathbb{R}$ is the standard saturation function defined as follows:

$$\sigma(u) = \begin{cases} 1 & \text{if } u > 1, \\ u & \text{if } |u| \leqslant 1, \\ -1 & \text{if } u < -1. \end{cases} \tag{2}$$

Under the assumption that the free dynamics

$$\dot{x} = f(x) \tag{3}$$

is Lyapunov stable, system (1) has been well studied in the context of global asymptotic stabilization, both with and without actuator saturation. For the case that $f(x) = Ax$ and $g(x) = Bx + N$, system (1) reduces to a bilinear system and its asymptotic stabilizability via smooth state or bounded state feedback was investigated by Gauthier and Kupka [3]. In particular, Gauthier and Kupka identified a set of sufficient conditions for global asymptotic stabilization using the Jurdjevic and Quinn technique [4]. More recently, Lin [5–7] established conditions under which the general nonlinear system (1) is globally asymptotically stabilizable. Under these conditions, bounded feedback laws were constructed to achieve global asymptotic stabilization.

The objective of this paper is to examine the problem of external stabilization of system (1). To define this problem, we recall the definitions of L_p^n space and L_p norm.

Definition 1. Suppose $x \in \mathbb{R}^n$.

1. For any $p \in [1, \infty)$, the signal space L_p^n is defined as

$$L_p^n = \left\{ x(t): \mathbb{R}^+ \to \mathbb{R}^n \;\middle|\; \int_{t=0}^{\infty} |x(t)|^p \, dt < \infty \right\}, \tag{4}$$

where $|\cdot|$ represents Euclidean norm. The signal space L_∞^n is defined as

$$L_\infty^n = \left\{ x(t): \mathbb{R}^+ \to \mathbb{R}^n \;\middle|\; \sup_t |x(t)| < \infty \right\}. \tag{5}$$

2. For any $p \in [1, \infty)$, the L_p norm of any signal $x \in L_p^n$ is defined as

$$\|x\|_{L_p} = \left[\int_{t=0}^{\infty} |x(t)|^p \, dt \right]^{1/p}. \tag{6}$$

The L_∞ norm of any signal $x \in L_\infty^n$ is defined as

$$\|x\|_{L_\infty} = \sup_t |x(t)|. \tag{7}$$

In the event that $n = 1$, we denote L_p^1 simply as L_p.

We define some notions of stabilizability to be discussed in this paper as follows.

Definition 2 (*L_p input to state stabilizability*). System (1) is said to be L_p input to state stabilizable by state feedback, for some $p \in [1, \infty]$, if there exists a state feedback law $u = \phi(x)$ with $\phi(0) = 0$ such that the closed-loop system

$$\dot{x} = f(x) + g(x)\sigma(\phi(x) + v), \quad x(0) = 0 \tag{8}$$

has the following property:

$$x \in L_p^n, \quad \forall v \in L_p, \tag{9}$$

where v is the external input such as a disturbance. Furthermore, the system is said to be finite gain L_p input to state stabilizable if for the closed-loop system (8),

$$\|x\|_{L_p} \leqslant \gamma \|v\|_{L_p}, \ \forall v \in L_p, \tag{10}$$

for some $\gamma \geqslant 0$.

Definition 3 (*Semi-global stabilizability*). System (1) is said to be semi-globally stabilizable if, for any bounded (arbitrarily large) set $\mathcal{X}_0 \in \mathbb{R}^n$, there exists a feedback law $u = \phi_{\mathcal{X}_0}(x)$ such that the equilibrium $x = 0$ of the closed-loop system

$$\dot{x} = f(x) + g(x)\sigma(\phi_{\mathcal{X}_0}(x)) \tag{11}$$

is locally asymptotically stable with $\mathcal{X}_0$ contained in its domain of attraction.

Several recent papers have investigated this problem for linear systems subject to actuator saturation, a special case of (1),

$$\dot{x} = Ax + B\sigma(u). \tag{12}$$

These investigations have led to a rather clear understanding of the problem. More specifically, it was shown by Liu et al., [8] that system (12) is finite gain L_p input to state stabilizable via linear feedback if (A, B) stabilizable and A is Lyapunov stable, that is, all the eigenvalues of A are in the closed left-half plane with those on the $j\omega$-axis having Jordan blocks of size 1. More recently, it was shown in [9] that the system (12) is finite gain L_2 input to state stabilizable via *nonlinear* feedback without any conditions on the open-loop system matrix A. For a more detailed account of these results, see [10].

Despite the recent understanding of finite gain L_p input to state stabilization of the linear system subject to actuator saturation (12), the issues related to its nonlinear counterpart (1) have not been carefully examined. This paper makes an attempt to study the problem of input to state stabilizability of system (1) with stable free dynamics. We will identify a few sets of sufficient conditions under which the system is L_p input to state stabilizable or finite gain L_p input to state stabilizable. Under these conditions, feedback laws are also explicitly constructed.

The organization of this paper is as follows. Section 2 recalls some results from [7] concerning global asymptotic stabilizability of system (1). These results serve as our motivating references for some of our main results to be given in Section 3. A brief concluding remark is made in Section 4.

2. Preliminaries and motivation of the problem

To begin with, we first introduce some notations. Throughout this paper, let $f(x)$ and $g(x)$ be smooth vector fields in $\mathbb{R}^n$, and $V : \mathbb{R}^n \to \mathbb{R}_+$ be a smooth function. Then Lie derivative $L_g V$ is defined by

$$L_g V(x) = \frac{\partial V(x)}{\partial x} g(x) \tag{13}$$

and Lie bracket $[f, g]$, which is again a smooth vector field, is defined by

$$[f, g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g. \tag{14}$$

The following notation for higher-order Lie brackets will also be used:

$$ad_f^0 g = g, \quad ad_f^1 g = [f, g], \ \ldots, \ ad_f^{k+1} g = [f, ad_f^k g], \quad k = 0, 1, \ldots. \tag{15}$$

Now we consider system (1) with its free dynamics satisfying the following assumption.

Assumption 1. For the free dynamics (3), there exists a C^r $(r \geqslant 1)$ function $V : \mathbb{R}^n \to \mathbb{R}_+$, which is positive definite and proper on $\mathbb{R}^n$, such that

$$L_f V(x) \leqslant 0, \quad \forall x \in \mathbb{R}^n. \tag{16}$$

That is, the free dynamics (3) is globally stable in the sense of Lyapunov.

With the vector fields f and g, we define the distribution

$$D = \text{span}\,\{ad_f^k g : 0 \leqslant k \leqslant n-1\} \tag{17}$$

and two sets associated with D:

$$\Omega = \{x \in \mathbb{R}^n : L_f^k V = 0, \ k = 1, \ldots, r\} \tag{18}$$

and

$$S = \{x \in \mathbb{R}^n : L_f^k L_\tau V(x) = 0, \ \forall \tau \in D, k = 0, \ldots, r-1\}. \tag{19}$$

The following theorem concerning global asymptotic stabilization of (1) is recalled from [7].

Theorem 1. *Under Assumption* 1, *system* (1) *is globally asymptotically stable if* $\Omega \cap S = \{0\}$. *Moreover, a stabilizing feedback law is given by*

$$u = -L_g V(x). \tag{20}$$

In the case of a linear system (12), the conditions of Theorem 1 reduces to the facts that all eigenvalues of A are in the closed left-half plane with those on the $j\omega$-axis having Jordan blocks of size 1 and that the pair (A, B) is stabilizable. In this case,

without loss of generality, we can assume that the pair (A, B) is in the following form:

$$A = \begin{bmatrix} A_0 & 0 \\ 0 & A_- \end{bmatrix}, \quad B = \begin{bmatrix} B_0 \\ B_- \end{bmatrix}, \tag{21}$$

where all eigenvalues of $A_- \in \mathbb{R}^{n_- \times n_-}$ have negative real parts and $A_0 \in \mathbb{R}^{n_0 \times n_0}$ is skew symmetric, i.e., $A_0 + A_0^{\mathrm{T}} = 0$. Moreover, the pair (A_0, B_0) are controllable. Clearly, subsystem

$$\dot{x}_0 = A_0 x_0 + B_0 \sigma(u) \tag{22}$$

satisfies the conditions of Theorem 1 with $V = x_0^{\mathrm{T}} x_0, \Omega = \mathbb{R}^{n_0}$ and $S = \{0\}$. Noticing the fact that the subsystem

$$\dot{x}_- = A_- x_- + B_- \sigma(u) \tag{23}$$

is an asymptotically stable linear system, we conclude from Theorem 1 that a globally asymptotically stabilizing feedback law is given by

$$u = -L_{B_0} V(x_0) = -2B_0^{\mathrm{T}} x_0. \tag{24}$$

It was also shown in [8] that the feedback law (24) also achieves finite gain $L_p$inpu t to state stabilization for (12), a linear system subject to actuator saturation. The implication of these discussions is the following: for a linear system subject to actuator saturation, the conditions of Theorem 1 ensure both global asymptotic stability and finite gain L_p input to state stabilizability. The following proposition and example show that, for general affine in control nonlinear system (1), the conditions of Theorem 1 do not always imply L_p input to state stabilizability. This motivates further investigation into the problem of L_p input to state stabilizability for general affine in control nonlinear system (1).

Lemma 1. *Let* g: $\mathbb{R}^n \to \mathbb{R}^n$, $n > 1$, *be continuous. If* $\{x : x^{\mathrm{T}} P g(x) = 0\} = \{0\}$ *for some* $P > 0$, *then* $g(0) = 0$.

Proof. Without loss of generality, assume $P = I$. If $P \neq I$, this condition could also be met by redefining x, P and g. We prove this lemma by contradiction. Assume that $g(0) \neq 0$. Then, there is at least one element of $g(0)$ which is nonzero. Without loss of generality, assume that $g_1(0) > 0$, where $g_1(x)$ is the first element of vector $g(x)$. By continuity of $g(x)$ at $x = 0$, there exists a $\delta > 0$ and a neighborhood $\mathcal{N}(0, r)$, $r > 0$, of the origin, such that $g_1(x) \geqslant \delta$, $\forall x \in \mathcal{N}(0, r)$. Clearly, $x_1 = [r/2, 0, 0, \ldots, 0]^{\mathrm{T}} \in \mathcal{N}(0, r)$ and hence $x_1^{\mathrm{T}} g(x_1) \geqslant \delta r/2 > 0$. Similarly, $x_2 = -x_1 \in \mathcal{N}(0, r)$ and hence $x_2^{\mathrm{T}} g(x_2) \leqslant -\delta r/2 < 0$. By continuity of $g(x)$ and the fact that $n > 1$, there exists an $x \neq 0$ such that $x^{\mathrm{T}} g(x) = 0$. □

Proposition 1. *Consider system* (1) *with* $n > 1$. *Assume that the conditions of Theorem* 1 *is satisfied with* $V = x^{\mathrm{T}} P x$, *where* $P > 0$. *Also assume that* $S \subset \{x \in \mathbb{R}^n : L_g V(x) = 2x^{\mathrm{T}} P g(x) = 0\} = \{0\}$. *Then, the feedback law* $u = -L_g V(x)$ *also achieves finite gain* L_p *input to state stabilization.*

Proof. Lemma 1 shows that $g(0) = 0$. Hence, the closed-loop system is trivially finite gain L_p input to state stable. □

Example 1. Consider the following system in the form of (1):

$$\dot{x} = e^x \sigma(u), \quad x \in \mathbb{R}. \tag{25}$$

The conditions of Theorem 1 hold with $V = x^2/2$ and $S = \{0\}$. Hence, the feedback law $u = -xe^x$ achieves global asymptotic stabilization. However, the closed-loop system

$$\dot{x} = e^x \sigma(-xe^x + v), \tag{26}$$

where v is the actuator disturbance, is not finite gain L_∞ stable. To see this, let $v(t) = -2 \in L_\infty$. Since $|xe^x| \leqslant e^{-1}$, $\forall x \leqslant 0$, the closed-loop system behaves as

$$\dot{x} = -e^x, \quad x(0) = 0, \tag{27}$$

Hence, $x(t) = -\ln(t+1) \notin L_\infty$.

3. Main results

Consider system (1). We will present a few sets of sufficient conditions for system (1) to be L_p input to state stabilizable or finite gain L_p input to state stabilizable.

To present these sufficient conditions, we need some preliminary results. First, the following lemma gives conditions under which the linearized system $(f'(0), g(0))$ is controllable, where $f'(0)$ is the first derivative of $f(x)$ with respect to x at $x = 0$.

Lemma 2. *If*

$$\text{Rank } D|_{x=0} = \text{Rank span } \{ad_f^k g : 0 \leqslant k \leqslant n-1\}|_{x=0} = n, \tag{28}$$

then the pair $(f'(0), g(0))$ *is controllable.*

Proof. It follows from a straightforward rank check of the controllability matrix of $(f'(0), g(0))$. □

We recall from [6] the following controllability-like condition:

$$\dim D = \dim \text{span } \{ad_f^k g : 0 \leqslant k \leqslant n-1\} = n, \quad \forall x \in \mathbb{R}^n \setminus \{0\}. \tag{29}$$

In [6], it was shown that affine in control nonlinear system is asymptotically stabilizable when the free dynamics of the system is stable and this controllability-like condition is satisfied. However, the following example shows that this controllability-like condition does not imply controllability of the linearized system $(f'(0), g(0))$.

Example 2. Consider system (1) with

$$f(x) = Ax = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} x, \quad g(x) = \begin{bmatrix} x^{\mathrm{T}}x \\ 0 \\ 1 \end{bmatrix}. \tag{30}$$

For this pair of $(f(x), g(x))$, we have,

$$\begin{aligned} \dim D &= \dim \operatorname{span}\left\{ g, \frac{\partial g}{\partial x}Ax - Ag, \frac{\partial((\partial g/\partial x)Ax - Ag)}{\partial x}Ax - A\left(\frac{\partial g}{\partial x}Ax - Ag\right) \right\} \\ &= \dim \operatorname{span}\left\{ \begin{bmatrix} x^{\mathrm{T}}x \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 2x^{\mathrm{T}} \\ 0 \\ 0 \end{bmatrix} Ax - Ag, -\frac{\partial Ag}{\partial x}Ax + A^2 g \right\} \\ &= \dim \operatorname{span}\{ g, -Ag, A^2 g \} \\ &= \dim \operatorname{span}\left\{ \begin{bmatrix} x^{\mathrm{T}}x \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} \right\} = 3, \quad \forall x \in \mathbb{R}^3 \setminus \{0\}. \end{aligned} \tag{31}$$

However, $(A, g(0))$ is not controllable.

The next lemma characterizes some conditions under which the system

$$\dot{x} = f(x) + g(x)u \tag{32}$$

is globally asymptotically stable and locally exponentially stable.

Lemma 3. *Consider system* (32). *Assume that the following conditions are satisfied:*

(a) *Assumption* 1 *holds with a* V *that also satisfies*

$$\frac{\partial V}{\partial x} = k(x)x^{\mathrm{T}} \tag{33}$$

for some $k(x) \geqslant 0$ *with* $k(0) > 0$;

(b) $\dim D = \dim \operatorname{span}\{ad_f^k g : 0 \leqslant k \leqslant n-1\} = n, \quad \forall x \in \mathbb{R}^n$;

(c) $f'(0)$ *is skew symmetric.*

Then, the closed-loop system under the state feedback $u = -L_g V = -k(x)g(x)^{\mathrm{T}}x$ *is globally asymptotically stable and locally exponentially stable.*

Proof of Lemma 3. Conditions (a) and (b) implies global asymptotic stability of the closed-loop system [7]. By Lemma 2, Condition (b) implies the controllability of the linearized system $(f'(0), g(0))$. Local exponential stability of the closed-loop system then follows from the exponential stability of its linearized system

$$\dot{x} = f'(0)x - k(0)g(0)g(0)^{\mathrm{T}}x, \tag{34}$$

which in turn is due to Conditions (b), (c) and $k(0) > 0$. □

Remark 1. Note here that condition (33) is not as restrictive as it appears to be. For example, for the first-order system, if V is $C^r(r > 1)$ function, it is satisfied as long as $\mathrm{d}V(x)/\mathrm{d}x|_{x=0} = 0$. For higher-order systems, it can be satisfied if $V(x) = V_1(|x|)$, for some function V_1, and $\mathrm{d}V(x)/\mathrm{d}x|_{x=0} = 0$.

In what follows, we recall a converse Lyapunov theorem [11].

Lemma 4. *Consider the autonomous system*

$$\dot{x} = f(x). \tag{35}$$

If the equilibrium $x = 0$ is globally asymptotically stable and locally exponentially stable, then for any $p \in (1, \infty)$ there exists a function $V(x)$ such that

$$a|x|^p \leqslant V(x) \leqslant \alpha(|x|)|x|^p, \tag{36}$$

$$\frac{\partial V}{\partial x} f(x) \leqslant -c|x|^p, \tag{37}$$

$$\left|\frac{\partial V}{\partial x}\right| \leqslant \beta(|x|)|x|^{p-1}, \tag{38}$$

where a and c are some positive scalars, and $\alpha(\cdot)$ and $\beta(\cdot)$ are positive monotonically nondecreasing continuous function on $[0, \infty)$. Moreover, in the absence of global asymptotic stability, (37), (38) *will hold within the domain of attraction.*

Finally, we recall from [12] the following lemma.

Lemma 5. *Let ξ and ζ be two positive scalars. For any $p > l > 0$, there exists two scalars $M_1, M_2 > 0$ such that*

$$\xi^{p-l}\zeta^l \leqslant M_1\xi^p + M_2\zeta^p \tag{39}$$

and consequently, for any $n > 0$ and $\kappa > 0$,

$$\xi^{p-l}\zeta^l \leqslant M_1\kappa^n\xi^p + \kappa^{n(l-p)/l}M_2\zeta^p. \tag{40}$$

We are now ready to present our first result on L_p input to state stabilizability of the system (1).

Theorem 2. *Consider system* 1. *Assume that the following conditions are satisfied:*

1. *Assumption* 1 *is satisfied with a V that also satisfies*

$$\frac{\partial}{\partial x}V(x) = k(x)x^{\mathrm{T}}, \tag{41}$$

where $k(x) \geqslant 0$, $k(0) > 0$.

2. $f'(0)$ *is skew-symmetric and* $\dim D = \dim \text{span}\{ad_f^k g : 0 \leqslant k \leqslant n-1\} = n$, $\forall x \in \mathbb{R}^n$.
3. *There exists a* $\hat{g} : \mathbb{R}^n \to \mathbb{R}_+$ *with* $|g(x)| \leqslant \hat{g}(x), \forall x \in \mathbb{R}^n$, *such that for some positive real number* θ,

$$\beta(|x|)\hat{g}(x) \leqslant \theta(1 + |x|^{\delta}), \quad \delta \in [0, 1], \tag{42}$$

where $\beta(\cdot)$ *is as defined in Lemma* 4 *for system* $\dot{x} = f(x) - g(x)L_g V$, *and if* $n > 1$,

$$\frac{\partial[\hat{g}(x)k(x)x_i]}{\partial x_j} = \frac{\partial[\hat{g}(x)k(x)x_j]}{\partial x_i}, \quad i \neq j, \ i,j = 1, \ldots, n. \tag{43}$$

Then, if $\delta \in [0, 1)$, *system* (1) *is* L_p *input to state stabilizable for any* $p \in (1, \infty]$, *and if* $\delta = 1$, *system* (1) *is* L_p *input to state stabilizable for any* $p \in (1, \infty)$. *Under the above conditions, the feedback law* $u = -L_g V$ *achieves global asymptotic stability and* L_p *input to state stability simultaneously.*

Remark 2. The reason for introducing $\hat{g}(x)$ is to increase the possibility of satisfying both (42) and (43). Suppose for some system, $g(x) = [2x_1 \ x_2]^{\mathrm{T}}$, $k(x) = 1$ and (42) is satisfied with $\hat{g}(x) = |g(x)|$. But (43) cannot be satisfied with this choice of $\hat{g}(x)$. If, however, we let $\hat{g}(x) = 2 \cdot \sqrt{x_1^2 + x_2^2}$ and suppose (42) is also true for the new $\hat{g}(x)$, it is easy to see that $|g(x)| \leqslant \hat{g}(x)$ and (43) is also satisfied.

Proof. Under the state feedback $u = -L_g V$, the closed-loop system, in the presence of the external input v, is written as follows:

$$\begin{aligned} \dot{x} &= f(x) + g(x)\sigma(-L_g V + v) \\ &= f(x) - g(x)L_g V + g(x)L_g V + g(x)\sigma(-L_g V + v) \\ &= f(x) - g(x)L_g V - g(x)(\tilde{x} - \sigma(\tilde{x})) + g(x)v, \end{aligned} \tag{44}$$

where $\tilde{x} = -L_g V + v$.

By Lemma 3, Conditions 1 and 2 imply that the closed-loop system (44) in absence of saturation and disturbance v, which is

$$\dot{x} = f(x) - g(x)L_g V \tag{45}$$

is globally asymptotically stable and locally exponentially stable.

Lemma 4 then shows that, for any $p \in (1, \infty)$, there exists a function $V_1(x)$ such that,

$$a|x|^p \leqslant V_1(x) \leqslant \alpha(|x|)|x|^p, \tag{46}$$

$$\frac{\partial V_1}{\partial x}(f(x) - g(x)L_g V) \leqslant -c|x|^p, \tag{47}$$

$$\left|\frac{\partial V_1}{\partial x}\right| \leqslant \beta(|x|)|x|^{p-1}, \tag{48}$$

where a and c are some positive scalars, and $\alpha(\cdot)$ and $\beta(\cdot)$ are positive monotonically nondecreasing continuous function on $[0, \infty)$.

Hence, the derivative of V_1 along the trajectories of (44) is bounded by

$$\dot{V}_1 \leqslant -c|x|^p + \beta(|x|)|x|^{p-1}|g(x)|\tilde{x}\sigma(\tilde{x}) + \beta(|x|)|x|^{p-1}|g(x)||v|$$

$$\leqslant -c|x|^p + \beta(|x|)|x|^{p-1}\hat{g}(x)\tilde{x}\sigma(\tilde{x}) + \beta(|x|)|x|^{p-1}\hat{g}(x)|v|. \tag{49}$$

For the case that $n = 1$, define

$$V_2(x) = \int_0^x \beta(|z|)|z|^{p-1}\hat{g}(z)k(z)z\,\mathrm{d}z. \tag{50}$$

For the case that $n > 1$, define

$$V_2(x) = \int_0^x \sum_{i=1}^{n} \beta(|z|)|z|^{p-1}\hat{g}(z)k(z)z_i\,\mathrm{d}z_i. \tag{51}$$

Note that V_2 is well defined due to (43). The partial derivative of V_2 is then given by

$$\frac{\partial}{\partial x}V_2(x) = \beta(|x|)|x|^{p-1}\hat{g}(x)k(x)x^{\mathrm{T}} \tag{52}$$

and $V_2(x)$ is nonnegative since $k(x) \geqslant 0$ for any $x \in \mathbb{R}^n$.

Now the derivative of V_2 along the trajectories of (44) is bounded as follows:

$$\dot{V}_2 \leqslant \beta(|x|)|x|^{p-1}\hat{g}(x)k(x)x^{\mathrm{T}}g(x)\sigma(\tilde{x})$$

$$\leqslant -\beta(|x|)|x|^{p-1}\hat{g}(x)\tilde{x}\sigma(\tilde{x}) + \beta(|x|)|x|^{p-1}\hat{g}(x)|v|. \tag{53}$$

Choosing $V_3 = V_1 + V_2$ and using (42) and Lemma 5, we have

$$\dot{V}_3 \leqslant -c|x|^p + 2\beta(|x|)|x|^{p-1}\hat{g}(x)|v| \leqslant -c|x|^p + 2\theta(1 + |x|^{\delta})|x|^{p-1}|v|$$

$$\leqslant -\frac{c}{2}|x|^p + \bar{\theta}(1 + |x|^{\delta p})|v|^p \tag{54}$$

for some $\bar{\theta} > 0$.

We now first establish the theorem for $p \in (1, \infty)$.

By inequality (46) and the definition of V_2, function V_3 has a lower bound,

$$V_3(x) \geqslant a|x|^p, \quad \forall x \in \mathbb{R}^n. \tag{55}$$

where a is as defined in (46).

Hence, inequality (54) can be continued as,

$$\dot{V}_3 \leqslant \tilde{\theta}(1 + V_3^{\delta})|v|^p \leqslant 2\tilde{\theta}(1 + V_3)^{\delta}|v|^p \tag{56}$$

for some $\tilde{\theta} > 0$.

Inequality (56) further implies

$$\frac{\mathrm{d}V_3}{(1 + V_3)^{\delta}} \leqslant 2\tilde{\theta}|v|^p\,\mathrm{d}t. \tag{57}$$

Integrating both sides of (57) from 0 to t and noting that $V_3(x(0)) = V_3(0) = 0$, we obtain,

(i) if $\delta \in [0, 1)$,

$$(1 + V_3(t))^{1-\delta} \leqslant 2\tilde{\theta}\|v\|_{L_p} < \infty, \quad \forall v \in L_p. \tag{58}$$

(ii) if $\delta = 1$,

$$\ln(1 + V_3(t)) \leqslant 2\tilde{\theta}\|v\|_{L_p} < \infty, \quad \forall v \in L_p. \tag{59}$$

Combining these two cases, we conclude that for any $v \in L_p$, V_3 is bounded by some positive real number depending on $\|v\|_{L_p}$, and hence $|x|$ is bounded by some positive real number depending on $\|v\|_{L_p}$.

Hence, by continuity of function $\beta(\cdot)$ and $g(\cdot)$, derivative of V_3 along the trajectories of (44) is bounded by

$$\dot{V}_3 \leqslant -M_1(\|v\|_{L_p})|x|^p + M_2(\|v\|_{L_p})|v|^p, \tag{60}$$

where $M_1(\|v\|_{L_p})$ and $M_2(\|v\|_{L_p})$ are positive real numbers depending on $\|v\|_{L_p}$.

Again integrating both sides of (60) from $t = 0$ to ∞ and noting that $V_3(x(0)) = V_3(0) = 0$ and $V_3(x(t)) \geqslant 0$ for any t, we have,

$$\int_0^\infty |x|^p \, dt < \infty, \quad \forall v \in L_p. \tag{61}$$

This concludes the proof of the Theorem for $p \in (1, \infty)$.

Next, we consider the case that $p = \infty$. To this end, we start with (54) for $p = 2$. More specifically, we have

$$\dot{V}_3 \leqslant -c|x|^2 + 2\theta|x||v| + 2\theta|x|^{1+\delta}|v|. \tag{62}$$

By Lemma 6, inequality (62) could be continued as follows:

$$\begin{aligned} \dot{V}_3 &\leqslant -\tilde{c}|x|^2 + \theta_1(|v|^2 + |v|^{2/(1-\delta)}) \\ &\leqslant -\tilde{c}|x|^2 + \theta_1(\|v\|_{L_\infty}^2 + \|v\|_{L_\infty}^{2/(1-\delta)}), \quad \forall v \in L_\infty, \end{aligned} \tag{63}$$

for some $\tilde{c}$, $\theta_1 > 0$.

Hence $\dot{V}_3$ is negative outside the ball with radius $\sqrt{(\theta_1/\tilde{c})(\|v\|_{L_\infty}^2 + \|v\|_{L_\infty}^{2/(1-\delta)})}$. Let V_M be the maximum of V_3 inside this ball, then $V_3(x(t)) \leqslant V_M$ for any $t \geqslant 0$. Inequality (46) now implies that

$$|x|^2 \leqslant \frac{1}{a} V_1(x) \leqslant \frac{1}{a}(V_1(x) + V_2(x)) = \frac{1}{a} V_3(x) \leqslant \frac{1}{a} V_M. \tag{64}$$

This completes the proof. □

We next establish some other sets of sufficient conditions for (finite gain) L_p input to state stability of (1). These sets of conditions are weaker than those of Theorem 2 and some even allow for stronger stabilization result. However, the stabilizing feedback laws under these conditions only achieve semi-global or local asymptotic

stability instead of global asymptotic stability. For simplicity in the presentation, we will restrict ourselves to the L_2 case.

We first establish the following preliminary result.

Lemma 6. *Consider system* (1) *under the feedback law*

$$u = -L_g V - \frac{L_g V}{1-(\kappa V)^2}, \tag{65}$$

where V is defined in Assumption 1 *and $\kappa > 0$ is a design parameter. Then, $L_V^0(1/\kappa) = \{x : V < 1/\kappa\}$ is an invariant set, that is, any trajectory that starts from inside this set will remain inside.*

Proof. The closed-loop system is given by

$$\dot{x} = f(x) + g(x)\sigma\left(-L_g V - \frac{L_g V}{1-(\kappa V)^2} + v\right), \quad v \in L_2, \tag{66}$$

where v is actuator disturbance. Inside the set $L_V^0(1/\kappa)$, the derivative of V along the trajectories of (66) is bounded by

$$\dot{V} \leqslant L_g V \sigma\left(-L_g V\left(1+\frac{1}{1-(\kappa V)^2}\right) + v\right). \tag{67}$$

If $\left|\dfrac{L_g V}{1-(\kappa V)^2}\right| \geqslant |v|$, then $\dot{V} \leqslant 0$.

If $\left|\dfrac{L_g V}{1-(\kappa V)^2}\right| < |v|$, then

$$\begin{aligned}\dot{V} &\leqslant L_g V\left[\sigma\left(-L_g V\left(1+\frac{1}{1-(\kappa V)^2}\right)+v\right) - \sigma\left(-L_g V\left(1+\frac{1}{1-(\kappa V)^2}\right)\right)\right.\\ &\quad \left.+\sigma\left(-L_g V\left(1+\frac{1}{1-(\kappa V)^2}\right)\right)\right]\\ &\leqslant |L_g V||v| \leqslant (1-(\kappa V)^2)|v|^2.\end{aligned} \tag{68}$$

Combining the two cases, we get

$$\dot{V} \leqslant (1-(\kappa V)^2)|v|^2. \tag{69}$$

Integrating both sides from 0 to t, we have

$$\frac{1}{2}\ln\frac{1+\kappa V(t)}{1-\kappa V(t)} - \frac{1}{2}\ln\frac{1+\kappa V(x_0)}{1-\kappa V(x_0)} \leqslant \kappa \int_0^t |v|^2 < \infty, \tag{70}$$

which implies $V(t) < 1/\kappa$, $\forall t$, for any $v \in L_2$.

For later use, we note that,

$$V(t) \leqslant \frac{1}{\kappa} \frac{e^{2\kappa\|v\|_{L_2}^2} - 1}{e^{2\kappa\|v\|_{L_2}^2} + 1} \quad \text{for} \quad x_0 = 0 \quad \text{and} \quad \forall v \in L_2. \tag{71}$$

□

We are now ready to present our second set of conditions for L_2 input to state stabilizability.

Theorem 3. *Assume system* (1) *satisfies the following conditions*:

1. *Assumption* 1 *is satisfied with a* V *such that*,

$$\frac{\partial V}{\partial x} = k(x)x^{\mathrm{T}}, \tag{72}$$

where $k(x) \geqslant 0$ *and any* $x \in \mathbb{R}^n$ *and* $k(0) > 0$.
2. $f'(0)$ *is skew-symmetric and* $\dim D = \dim \text{span}\{ad_f^k g : 0 \leqslant k \leqslant n-1\} = n$, $\forall x \in \mathbb{R}^n$;
3. *If* $n > 1$, *there exists a* $\kappa^* > 0$ *such that for all* $x \in L_V^0(1/\kappa)$, $\forall \kappa \in (0, \kappa^*]$,

$$\frac{\partial k(x)(1 + 1/(1 - (\kappa V(x))^2))x_i}{\partial x_j} = \frac{\partial k(x)(1 + 1/(1 - (\kappa V(x))^2))x_j}{\partial x_i}, \tag{73}$$

Then system (1) *is* L_2 *input to state stabilizable. Moreover, there exists a family of feedback laws that achieves semi-global asymptotic stabilization and* L_2 *input to state stabilization simultaneously.*

Proof. We prove this theorem by explicit construction of feedback laws. The family of feedback laws we construct is parameterized in $\kappa > 0$ and is given by

$$u = -L_g V - \frac{L_g V}{1 - (\kappa V)^2}. \tag{74}$$

By Lemma 6, level set $L_V^0(1/\kappa)$ is invariant with $v = 0$, which means any trajectory starting from inside the level set $L_V^0(1/\kappa)$ will stay inside for ever. Hence, the equilibrium $x = 0$ of the closed-loop system

$$\dot{x} = f(x) + g(x)\sigma\left(-L_g V - \frac{L_g V}{1 - (\kappa V)^2}\right) \tag{75}$$

is asymptotically stable and $L_V^0(1/\kappa)$ is the domain of attraction [7]. Since V is a proper function on $\mathbb{R}^n$, by decreasing the value of κ, we can include any bounded set $\mathcal{X}_0$ of the state space in $L_V^0(1/\kappa)$. This shows that the family of feedback laws (74) indeed achieves semi-global stabilization.

In what follows we show that the same family of feedback laws also achieves L_2 input to state stabilization. To this end, we consider the closed-loop system in the

presence of external input v,

$$\begin{aligned}\dot{x} &= f(x) + g(x)\sigma\left(-\frac{L_gV}{1-(\kappa V)^2} - L_gV + v\right)\\ &= f(x) - g(x)L_gV\left(1+\frac{1}{1-(\kappa V)^2}\right) + g(x)L_gV\left(1+\frac{1}{1-(\kappa V)^2}\right)\\ &\quad + g(x)\sigma\left(-\frac{L_gV}{1-(\kappa V)^2} - L_gV + v\right)\\ &= f(x) - g(x)L_gV\left(1+\frac{1}{1-(\kappa V)^2}\right) - g(x)[\tilde{x} - \sigma(\tilde{x})] + g(x)v, \end{aligned} \tag{76}$$

where v is the actuator disturbance and $\tilde{x} = -((1/(1-(\kappa V)^2)) + 1)L_gV + v$.

By Lemma 6, $L_V^0(1/\kappa)$ is an invariant set for any $v \in L_2$. Moreover, from (71), for any given $v \in L_2$,

$$x \in L_V\left(\frac{1}{\kappa}\frac{e^{2\kappa\|v\|_{L_2}^2} - 1}{e^{2\kappa\|v\|_{L_2}^2} + 1}\right) = \left\{x \in \mathbb{R}^n : V \leqslant \frac{1}{\kappa}\frac{e^{2\kappa\|v\|_{L_2}^2} - 1}{e^{2\kappa\|v\|_{L_2}^2} + 1}\right\}. \tag{77}$$

By Lemma 3, Conditions 1 and 2 imply that the equilibrium $x = 0$ of system

$$\dot{x} = f(x) + g(x)\left(-L_gV - \frac{L_gV}{1-(\kappa V)^2}\right) \tag{78}$$

is locally exponentially stable and $L_V^0(1/\kappa)$ is its domain of attraction. By Lemma 4, there exists V_1 such that, for $x \in L_V^0(1/\kappa)$,

$$\frac{\partial V_1}{\partial x}\left[f(x) + g(x)\left(-L_gV - \frac{L_gV}{1-(\kappa V)^2}\right)\right] \leqslant -c|x|^2, \tag{79}$$

$$\left|\frac{\partial V_1}{\partial x}\right| \leqslant \beta(|x|)|x|, \tag{80}$$

where $c > 0$ and β: $\mathbb{R}_+ \to \mathbb{R}_+$ is some monotonically nodecreasing continuous function.

Let $\mu, b > 0$ be such that

$$|g(x)| \leqslant \mu, \quad \beta(x) \leqslant b, \quad \forall x \in L_V\left(\frac{1}{\kappa}\frac{e^{2\kappa\|v\|_{L_2}^2} - 1}{e^{2\kappa\|v\|_{L_2}^2} + 1}\right). \tag{81}$$

Hence, the derivative of V_1 along the trajectories of (76) is bounded by

$$\begin{aligned}\dot{V}_1 &\leqslant -c|x|^2 + \beta(x)|x||g(x)|\tilde{x}\sigma(\tilde{x}) + \beta(x)|x||g(x)||v|\\ &\leqslant -c|x|^2 + b\mu|x|\tilde{x}\sigma(\tilde{x}) + b\mu|x||v|. \end{aligned} \tag{82}$$

For case $n = 1$, define

$$V_2 = \int_0^x |z|k(z)z\left(1 + \frac{1}{1 - (\kappa V(z)^2)}\right) dz. \tag{83}$$

When $n > 1$, define

$$V_2(x) = \int_0^x \sum_{i=1}^{n} |z|k(z)\left(1 + \frac{1}{1 - (\kappa V(z))^2}\right) z_i \, dz_i. \tag{84}$$

Note that this $V_2(x)$ is well defined due to Condition 3 of the theorem. The derivative of V_2 along the closed-loop system is bounded as follows:

$$\dot{V}_2 \leqslant |x|k(x)\left(1 + \frac{1}{1 - (\kappa V(|x|))^2}\right) x^{\mathrm{T}} g\sigma(\tilde{x}) = -|x|\tilde{x}\sigma(\tilde{x}) + |x||v|. \tag{85}$$

Now define $V_3 = V_1 + b\mu V_2$,

$$\dot{V}_3 \leqslant -c|x|^2 + 2b\mu|x||v| \leqslant -\frac{3c}{4}|x|^2 + \frac{4b^2\mu^2}{c}|v|^2. \tag{86}$$

Integrating both sides from $t = 0$ to ∞, and noting that $V_3(x(0)) = V_3(0) = 0$ and $V_3(x(t)) \geqslant 0$ for all t, we have

$$\|x\|_{L_2} \leqslant \gamma \|v\|_{L_2} \tag{87}$$

with $\gamma = 4b\mu/\sqrt{3}c$. □

Now we give a simple example to illustrate the result given in Theorem 3.

Example 3. Consider system

$$\dot{x} = \begin{bmatrix} -x_1^3 + x_2 \\ -x_1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 + x_1^2 + x_2^2 \end{bmatrix} \sigma(u). \tag{88}$$

It is easy to verify that Conditions 1–3 in Theorem 3 are satisfied with a function $V = x_1^2 + x_2^2$ for any $\kappa > 0$. Then the family of feedback laws $u = -L_g V - (L_g V/(1 - (\kappa V)^2))$ achieves both semi-global asymptotic stabilization and L_2 input to state stabilization simultaneously. Fig. 1 shows the simulation results for $x(0) = (2, 1)$ and a disturbance $v = [1(t - 5) - 1(t - 10)]$. The value of κ chosen is 0.1.

Our next set of condition exploits some special structures of $g(x)$. Here we will see the L_2 gain from the disturbance to the state is finite and could be made arbitrarily small by appropriate choice of control law.

Theorem 4. *System* (1) *is finite gain* L_2 *input to state stabilizable, if*

1. $(f'(0), g(0))$ *stabilizable; and*
2. $g(x)$ *is in the form of* $g(x) = \alpha(x)B$, *where* $\alpha(x)$ *is a scalar function and* $B \in \mathbb{R}^n$, *or* $g(x)$ *satisfies* $g'(0) = 0$.

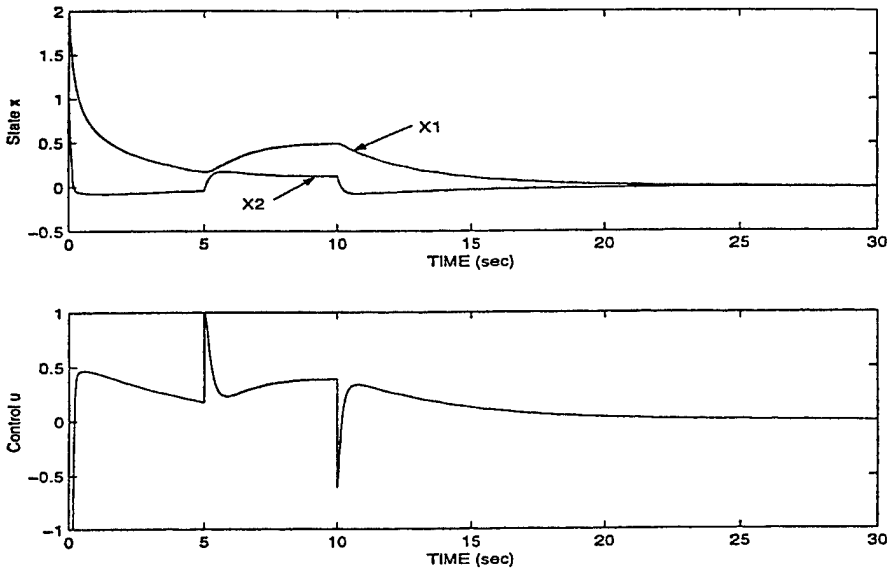


Fig. 1. Simulation results for Example 3 with $\kappa = 0.1$.

Moreover, under these conditions, there exists a feedback law that simultaneously achieves local asymptotic stability and finite gain L_2 input to state stability with an arbitrarily small gain.

Proof. Design the state feedback law as follows:

$$u = -g(0)^{\mathrm{T}} Px - \frac{\rho}{1 - (\kappa x^{\mathrm{T}} Px)^2} g(0)^{\mathrm{T}} Px, \tag{89}$$

where $\rho > 0$ is a design parameter and P is the unique positive-definite solution to

$$Pf'(0) + (f'(0))^{\mathrm{T}} P - Pg(0)g(0)^{\mathrm{T}} P = -I. \tag{90}$$

We note that such a P exists because $(f'(0), g(0))$ is stabilizable.

Under this feedback law, the closed-loop system takes the following form:

$$\dot{x} = f(x) + g(x)\sigma\left(-g(0)^{\mathrm{T}} Px - \frac{\rho}{1 - (\kappa x^{\mathrm{T}} Px)^2} g(0)^{\mathrm{T}} Px + v\right). \tag{91}$$

Choose Lyapunov function candidate $V = x^{\mathrm{T}} Px$, the derivative of V along the closed-loop system (91) is given by

$$\dot{V} = -|x|^2 + x^{\mathrm{T}}P[x^{\mathrm{T}}f_1''(\xi_1)x, \ldots, x^{\mathrm{T}}f_i''(\xi_i)x, \ldots, x^{\mathrm{T}}f_n''(\xi_n)x]^{\mathrm{T}} + x^{\mathrm{T}}Pg(0)g(0)^{\mathrm{T}}Px + 2x^{\mathrm{T}}Pg(x)\sigma\left(-g(0)^{\mathrm{T}}Px - \frac{\rho}{1-(\kappa x^{\mathrm{T}}Px)^2}g(0)^{\mathrm{T}}Px + v\right), \tag{92}$$

where

$$f_i''(\xi_i) = \begin{bmatrix} \dfrac{\partial^2 f_i(\xi_i)}{\partial^2 x_1} & \dfrac{\partial^2 f_i(\xi_i)}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f_i(\xi_i)}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f_i(\xi_i)}{\partial x_1 \partial x_2} & \dfrac{\partial^2 f_i(\xi_i)}{\partial^2 x_2} & \cdots & \dfrac{\partial^2 f_i(\xi_i)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f_i(\xi_i)}{\partial x_1 \partial x_n} & \dfrac{\partial^2 f_i(\xi_i)}{\partial x_2 \partial x_n} & \cdots & \dfrac{\partial^2 f_i(\xi_i)}{\partial^2 x_n} \end{bmatrix} \tag{93}$$

and $\xi_i = [\xi_{i1}, \ldots, \xi_{ij}, \ldots, \xi_{in}]^{\mathrm{T}}$ with $\xi_{ij} \in [0, x_j]$, $j = 1, \ldots, n$.

In the case that $g(x) = \alpha(x)B$, (92) is continued as

$$\begin{aligned} \dot{V} &= -|x|^2 + x^{\mathrm{T}}P[x^{\mathrm{T}}f_1''(\xi_1)x, \ldots, x^{\mathrm{T}}f_i''(\xi_i)x, \ldots, x^{\mathrm{T}}f_n''(\xi_n)x]^{\mathrm{T}} \\ &\quad + x^{\mathrm{T}}Pg(0)g(0)^{\mathrm{T}}Px + 2x^{\mathrm{T}}P(\alpha(0)B + \alpha'(\zeta)xB) \\ &\quad \times \sigma\left(-g(0)^{\mathrm{T}}Px - \frac{\rho}{1-(\kappa x^{\mathrm{T}}Px)^2}g(0)^{\mathrm{T}}Px + v\right) \\ &\leqslant -|x|^2 + F(\xi)|x|^3 + x^{\mathrm{T}}Pg(0)g(0)^{\mathrm{T}}Px + 2x^{\mathrm{T}}P(\alpha(0)B + \alpha'(\zeta)xB) \\ &\quad \times \sigma\left(-g(0)^{\mathrm{T}}Px - \frac{\rho}{1-(\kappa x^{\mathrm{T}}Px)^2}g(0)^{\mathrm{T}}Px + v\right), \end{aligned} \tag{94}$$

where $\zeta \in \mathbb{R}^n$ with $\zeta_i \in [0, x_i]$, $i = 1, \ldots, n$ and $F(\cdot)$ is a continuous scalar function.

Let $\kappa > 0$ be such that

$$x \in L_V^0(1/\kappa) \Rightarrow F(\xi)|x|^3 \leqslant |x|^2/2, \ |\alpha'(\zeta)x| \leqslant |\alpha(0)|/2 \quad \text{and} \quad |g(0)^{\mathrm{T}}Px| \leqslant 1. \tag{95}$$

Fix κ. We now consider (94) inside $L_V^0(1/\kappa)$, which, by Lemma 6, is an invariant set if $v \in L_2$.

If $\left|\dfrac{\rho}{1-(\kappa x^{\mathrm{T}}Px)^2}g(0)^{\mathrm{T}}Px\right| \geqslant |v|$, then $\dot{V} < 0$.

If $\left|\dfrac{\rho}{1-(\kappa x^{\mathrm{T}}Px)^2}g(0)^{\mathrm{T}}Px\right| < |v|$, then,

$$\dot{V} \leqslant -\tfrac{1}{2}|x|^2 - 2x^{\mathrm{T}}P[\alpha(0)B + \alpha'(\zeta)xB]$$
$$\times \left[-\sigma\left(-g(0)^{\mathrm{T}}Px - \frac{\rho}{1-(\kappa x^{\mathrm{T}}Px)^2}g(0)^{\mathrm{T}}Px + v\right) - g(0)^{\mathrm{T}}Px\right]$$
$$\leqslant -\frac{1}{2}|x|^2 + \frac{4(1-(\kappa x^{\mathrm{T}}Px)^2)}{\rho}|v|^2$$
$$\leqslant -\frac{1}{2}|x|^2 + \frac{4}{\rho}|v|^2. \tag{96}$$

Integrating both sides from $t=0$ to ∞ and noting that $V(x(0)) = V(0) = 0$ and $V(x(t)) \geqslant 0$, we conclude,

$$\|x\|_{L_2} \leqslant \frac{\sqrt{8}}{\rho}\|v\|_{L_2}, \quad \forall v \in L_2. \tag{97}$$

Moreover, we note that $\sqrt{8/\rho} \to 0$ as $\rho \to \infty$.

In the case when $g'(0) = 0$, (92) is bounded by

$$\dot{V} = -|x|^2 + x^{\mathrm{T}}P[x^{\mathrm{T}}f_1''(\xi_1)x, \ldots, x^{\mathrm{T}}f_i''(\xi_i)x, \ldots, x^{\mathrm{T}}f_n''(\xi_n)x]^{\mathrm{T}}$$
$$+ x^{\mathrm{T}}Pg(0)g(0)^{\mathrm{T}}Px + 2x^{\mathrm{T}}P(g(0)$$
$$+ [x^{\mathrm{T}}g_1''(\zeta_1)x/2, \ldots, x^{\mathrm{T}}g_i''(\zeta_i)x/2, \ldots, x^{\mathrm{T}}g_n''(\zeta_n)x/2]^{\mathrm{T}})$$
$$\times \sigma\left(-g(0)^{\mathrm{T}}Px - \frac{\rho}{1-(\kappa x^{\mathrm{T}}Px)^2}g(0)^{\mathrm{T}}Px + v\right)$$
$$\leqslant -|x|^2 + G(\xi,\zeta)|x|^3 + x^{\mathrm{T}}Pg(0)g(0)^{\mathrm{T}}Px$$
$$+ 2x^{\mathrm{T}}Pg(0)\sigma\left(-g(0)^{\mathrm{T}}Px - \frac{\rho}{1-(\kappa x^{\mathrm{T}}Px)^2}g(0)^{\mathrm{T}}Px + v\right), \tag{98}$$

where $G(\cdot)$ is some continuous scalar function. Let $\kappa > 0$ be some sufficiently large number such that

$$x \in L_V^0(1/\kappa) \Rightarrow G(\xi,\zeta)|x|^3 \leqslant |x|^2/2 \quad \text{and} \quad |g(0)^{\mathrm{T}}Px| \leqslant 1. \tag{99}$$

Fix κ. Now consider (98) inside $L_V^0(1/\kappa)$, which, by Lemma 6, is an invariant set in $v \in L_2$. First rewrite (98) as

$$\dot{V} \leqslant -\frac{1}{2}|x|^2 + x^{\mathrm{T}}Pg(0)g(0)^{\mathrm{T}}Px$$
$$+ 2x^{\mathrm{T}}Pg(0)\sigma\left(-g(0)^{\mathrm{T}}Px - \frac{\rho}{1-(\kappa x^{\mathrm{T}}Px)^2}g(0)^{\mathrm{T}}Px + v\right).$$

$$\text{If } \left|\frac{\rho}{1-(\kappa x^{\mathrm{T}}Px)^2}g(0)^{\mathrm{T}}Px\right| \geqslant |v|, \text{ then } \dot{V}<0.$$

$$\text{If } \left|\frac{\rho}{1-(\kappa x^{\mathrm{T}}Px)^2}g(0)^{\mathrm{T}}Px\right| < |v|, \text{ then}$$

$$\begin{aligned}\dot{V} &\leqslant -\frac{1}{2}|x|^2 - 2x^{\mathrm{T}}Pg(0)\\ &\quad\times\left[-\sigma\left(-g(0)^{\mathrm{T}}Px - \frac{\rho}{1-(\kappa x^{\mathrm{T}}Px)^2}g(0)^{\mathrm{T}}Px + v\right) - g(0)^{\mathrm{T}}Px\right]\\ &\leqslant -\frac{1}{2}|x|^2 + \frac{4(1-(\kappa x^{\mathrm{T}}Px)^2)}{\rho}|v|^2\\ &\leqslant -\frac{1}{2}|x|^2 + \frac{4}{\rho}|v|^2. \end{aligned} \tag{100}$$

Integrating both sides from $t=0$ to ∞ and noting that $V(x(0))=V(0)=0$ and $V(x(t))\geqslant 0$, we conclude

$$\|x\|_{L_2} \leqslant \frac{\sqrt{8}}{\rho}\|v\|_{L_2}, \quad \forall v \in L_2. \tag{101}$$

Moreover, we note that $\sqrt{8/\rho}\to 0$ as $\rho\to\infty$. □

Now, we give an example to illustrate the result in Theorem 4.

Example 4. First, we consider system

$$\dot{x} = \begin{bmatrix} x_1^3 + x_2 \\ x_1 + x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ (1+x_2^2)\cos x_1 \end{bmatrix}\sigma(u). \tag{102}$$

It is easy to see that Conditions 1 and 2 are both satisfied and

$$g(x) = \alpha(x)B = (1+x_2^2)\cos x_1 \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

From the proof of Theorem 4, we obtain a feedback law of

$$\begin{aligned} u = &-[1+\rho/(1-\kappa^2(2.6955x_1^2 + 4.8284x_1x_2 + 3.6131x_2^2)^2)]\\ &\times(2.4142x_1 + 3.6131x_2). \end{aligned} \tag{103}$$

The simulation result is shown in Figs. 2 and 3, where the disturbance $v = [1(t-5) - 1(t-10)]$ and the initial condition $(x_1(0), x_2(0)) = (0.5, 0.1)$. From the simulation we can see the effect of the disturbance v on the state x is indeed reduced by increasing the value of ρ.

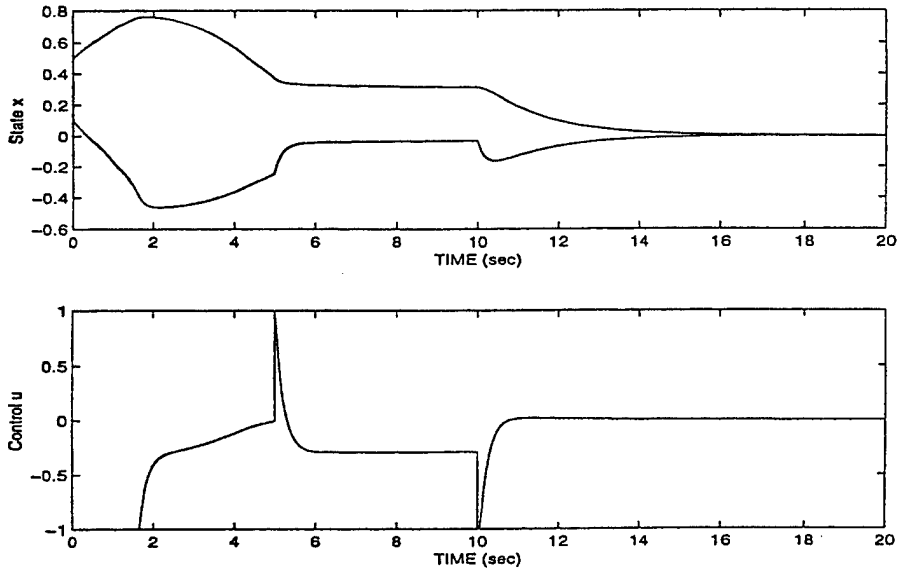


Fig. 2. Simulation results for system (102) with $\kappa = 0.8$, $\rho = 1$.

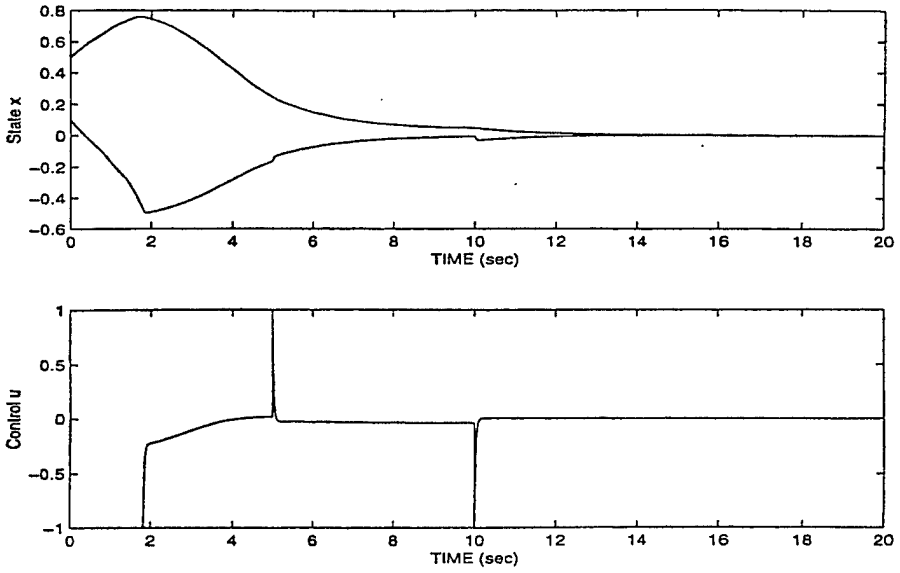


Fig. 3. Simulation results for system (102) with $\kappa = 0.8$, $\rho = 10$.

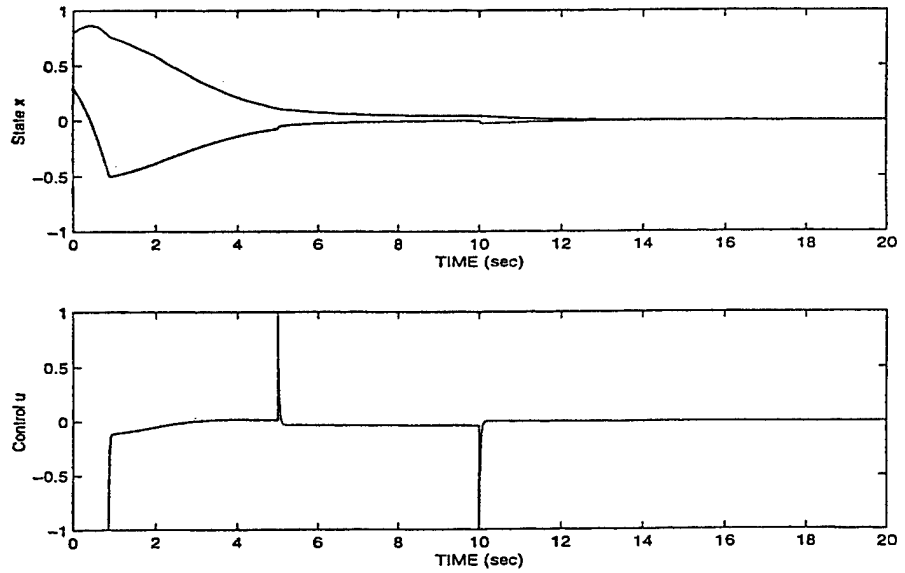


Fig. 4. Simulation results for system (104) with $\kappa = 0.2$, $\rho = 10$.

Next, we consider system

$$\dot{x} = \begin{bmatrix} x_1^3 + x_2 \\ x_1 + x_2 \end{bmatrix} + \begin{bmatrix} x_1^3 \\ 1 + x_1^2 + x_2^7 \end{bmatrix} \sigma(u). \tag{104}$$

It is easy to see that Conditions 1 and 2 are both satisfied and $g'(0) = 0$. From the proof of Theorem 4, we have the feedback law in the form of

$$u = -\left[1 + \rho / \left(1 - \kappa^2 (2.6955x_1^2 + 4.8284x_1x_2 + 3.6131x_2^2)^2\right)\right] \times (2.4142x_1 + 3.6131x_2). \tag{105}$$

The simulation result is shown in Fig. 4, where the disturbance $v = [1(t-5) - 1(t-10)]$ and the initial condition $(x_1(0), x_2(0)) = (0.8, 0.3)$.

4. Conclusions

In this paper we have studied the external stabilizability of affine in control nonlinear systems with saturating actuators. A few sets of conditions under which the system is L_p input to state stabilizable or finite gain L_p input to state stabilizable are given. Under those conditions, the stabilizing feedback laws are explicitly constructed. Our future work would concentrate in further relaxation of these conditions and establishing their counterparts for discrete-time systems.

Acknowledgements

The authors would like to thank C. Qian and W. Lin for detailed comments on the first draft of the paper.

References

[1] D.S. Bernstein, A.N. Michel, A chronological bibliography on saturating actuators, Int. J. Robust Nonlinear Control 5 (1995) 375–380.
[2] Z. Lin, Low Gain Feedback, Springer, Berlin, 1998.
[3] J.P. Gauthier, I. Kupka, Controllability and stability, J. Differential Equations 28 (1978) 381–389.
[4] V. Jurdjevic, J.P. Quinn, Controllability and stability, J. Differential Equations 28 (1978) 381–389.
[5] W. Lin, Input saturation and global stabilization by output feedback for affine systems, IEEE Trans. Automat. Control 40 (1995) 776–782.
[6] W. Lin, Feedback stabilization of general nonlinear control systems: a passive system approach, System Control Lett. 25 (1995) 41–52.
[7] W. Lin, Global asymptotic stabilization of general nonlinear systems with stable free dynamics via passivity and bounded feedback, Automatica 32 (1996) 916–924.
[8] W. Liu, Y. Chitour, E. Sontag, On finite gain stabilizability of linear systems subject to input saturation, SIAM J. Control Optim. 34 (4) (1996) 1190–1219.
[9] Z. Lin, H_∞-almost disturbance decoupling with internal stability for linear systems subject to input saturation, IEEE Trans. Automat. Control 42 (1997) 992–995.
[10] P. Hou, A. Saberi, Z. Lin, P. Sannuti, Simultaneous external and internal stabilization for continuous and discrete-time critically unstable linear systems with saturating actuators, Automatica 34 (1998) 1547–1557.
[11] W. Lin, C. Qian, Semi-global robust stabilization of nonlinear systems by partial state and output feedback. Proceedings of 1998 IEEE CDC, 1998, pp. 3105–3110.
[12] X. Bao, Z. Lin, E. Sontag, Finite gain stabilization of discrete-time linear systems subject to actuator saturation, Automatica 36 (2) (2000) 269–277.

Publication 8

INTERNATIONAL JOURNAL OF ROBUST AND NONLINEAR CONTROL
Int. J. Robust Nonlinear Control 2001; **11**:555–588 (DOI: 10.1002/rnc.565)

Practical stabilization of exponentially unstable linear systems subject to actuator saturation nonlinearity and disturbance

Tingshu Hu*,† and Zongli Lin

Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903, U.S.A.

SUMMARY

This paper investigates the problem of practical stabilization for linear systems subject to actuator saturation and input additive disturbance. Attention is restricted to systems with two anti-stable modes. For such a system, a family of linear feedback laws is constructed that achieves semi-global practical stabilization on the asymptotically null controllable region. This is in the sense that, for any set χ_0 in the interior of the asymptotically null controllable region, any (arbitrarily small) set χ_∞ containing the origin in its interior, and any (arbitrarily large) bound on the disturbance, there is a feedback law from the family such that any trajectory of the closed-loop system enters and remains in the set χ_∞ in a finite time as long as it starts from the set χ_0. In proving the main results, the continuity and monotonicity of the domain of attraction for a class of second-order systems are revealed. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: nonlinearities; semi-global stabilization; disturbance rejection; actuator saturation; limit cycle; high gain feedback

1. INTRODUCTION

We consider the problem of controlling an exponentially unstable linear system with saturating actuators. This control problem involves issues ranging from such basic ones as controllability and stabilizability to closed-loop performances beyond stabilization. In regard to controllability, the issue is the characterization of the null controllable region (or the asymptotically null controllable region), the set of all initial states that can be driven to the origin by the bounded

*Correspondence to: Tingshu Hu, Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903, U.S.A.
†E-mail: th7f@virginia.edu

Contract/grant sponsor: US office of Naval Research Young Investigator Program; contract/grant number: N00014-99-1-0670

Copyright © 2001 John Wiley & Sons, Ltd.

Received 14 September 1999
Accepted 4 September 2000

input provided by the saturating actuators in a finite time (or asymptotically). On the other hand, the issue of stabilizability is the determination of the existence of feedback laws that stabilize the system within the asymptotically null controllable region and the actual construction of these feedback laws.

It turns out that these seemingly simple issues are actually quite difficult to address for general linear systems. As a result, they have been systematically studied only for linear systems that are not exponentially unstable (all open-loop poles are in the closed left-hand-plane). In particular, it is now well known [1–3] that if a linear system has all its open-loop poles in the closed left-half-plane and is stabilizable in the usual linear system sense, then, when subject to actuator saturation, its asymptotically null controllable region is the entire state space. For this reason, such a linear system is usually referred to as asymptotically null controllable with bounded controls (ANCBC).

In regard to stabilizability, it is shown in Reference [4] that a linear system subject to actuator saturation can be globally asymptotically stabilized by nonlinear feedback if and only if it is ANCBC. A nested feedback design technique for designing nonlinear globally asymptotically stabilizing feedback laws was proposed in References [5–7]. Alternative solutions to the global stabilization problem consisting of scheduling a parameter in an algebraic Riccati equation according to the size of the state vector were later proposed in References [8–10]. The question of whether or not a general linear ANCBC system subject to actuator saturation can be globally asymptotically stabilized by linear feedback was answered in References [11, 12], where it was shown that a chain of integrators of length greater than 2 cannot be globally asymptotically stabilized by saturated linear feedback.

The notion of semi-global asymptotic stabilization (on the asymptotically null controllable region) for linear systems subject to actuator saturation was introduced in References [13, 14]. The semi-global framework for stabilization requires feedback laws that yield a closed-loop system which has an asymptotically stable equilibrium whose domain of attraction includes an *a priori* given (arbitrarily large) bounded subset of the asymptotically null controllable region. In References [13, 14], it was shown that, for linear ANCBC systems subject to actuator saturation, one can achieve semi-global asymptotic stabilization by using linear feedback laws.

In an effort to address closed-loop performances beyond large domain of attraction, [15] formulates and solves the problem of practical semi-global stabilization for ANCBC systems with saturating actuators. In particular, low-and-high gain feedback laws are constructed that not only achieve semi-global stabilization in the presence of input additive uncertainties but also have the ability to reject bounded input additive disturbance.

Despite the numerous results on linear ANCBC systems, the counterparts of the above-mentioned results for exponentially unstable linear systems are less understood. Recently, we made an attempt to systematically study issues related to the null controllable regions (or asymptotically null controllability regions) and the stabilizability for exponentially unstable linear systems subject to actuator saturation and gave a rather clear understanding of these issues [16]. Specifically, we gave a simple exact description of the null controllable region for a general anti-stable linear system in terms of a set of extremal trajectories of its time-reversed system. For a linear planar anti-stable system under a saturated linear stabilizing feedback law, we established that the boundary of the domain of attraction is the unique stable limit circle of its time-reversed system. Furthermore, we constructed feedback laws that semi-globally asymptotically stabilize any system with two anti-stable modes on its asymptotically null controllable region. This is in the sense that, for any *a priori* given set in the interior of the asymptotically null controllable

Copyright © 2001 John Wiley & Sons, Ltd.

region, there exists a saturated linear feedback law that yields a closed-loop system which has an asymptotically stable equilibrium whose domain of attraction includes the given set.

The goal of this paper is to design feedback laws that, not only achieve semi-global stabilization on the asymptotically null controllable region, but also has the ability to reject bounded disturbance to an arbitrary level of accuracy. Our attention will be restricted to systems that have two anti-stable modes. Our problem formulation is motivated by its counterpart for ANCBC systems [15].

This paper is organized as follows. Section 2 formulates the problem and summarizes the main results. Sections 3 and 4 establish some fundamental properties of the behaviours of planar systems. These properties lead to the proof of the main results in Sections 5 and 6 . Section 7 uses an aircraft model to demonstrate the results obtained in this paper. Section 8 contains a brief concluding remark.

For a set X, we use ∂X, $\bar{X}$ and $\text{int}(X)$ to denote its boundary, closure and interior, respectively. For a measurable function, $w:[0, \infty) \to \mathbf{R}$, $\|w\|_\infty$ is its L_∞-norm. For a vector v, we use $(v)_i$ to denote its ith co-ordinate. For two bounded subsets X_1, X_2 of $\mathbf{R}^n$, their Hausdorff distance is defined as

$$d(X_1, X_2) := \max\{\vec{d}(X_1, X_2), \vec{d}(X_2, X_1)\}$$

where

$$\vec{d}(X_1, X_2) = \sup_{x_1 \in X_1} \inf_{x_2 \in X_2} \|x_1 - x_2\|$$

Here the vector norm used is arbitrary.

2. PROBLEM STATEMENT AND THE MAIN RESULTS

2.1. *Problem statement*

Consider an open-loop system subject to both actuator saturation and disturbance,

$$\dot{x} = Ax + b\,\text{sat}(u + w) \tag{1}$$

where $x \in \mathbf{R}^n$ is the state, $u \in \mathbf{R}$ is the control input, $w \in \mathbf{R}$ is the disturbance and $\text{sat}(s) = \text{sign}(s) \min\{1, |s|\}$ is the standard saturation function. Assume that (A, b) is stabilizable. We consider the following set of disturbances:

$$\mathscr{W} := \{w:[0, \infty) \to \mathbf{R},\ w \text{ is measurable and } \|w\|_\infty \leqslant D\},$$

where D is a known constant.

In addressing the practical stabilization problem, we need to describe the largest possible region in the state space that can be stabilized. For this purpose, we introduce the notions of null controllability and asymptotic null controllability.

Definition 1

Consider system (1) in the absence of the disturbance w. A state x_0 is said to be null controllable if there exist a $T \in [0, \infty)$ and a measurable control u such that the state trajectory $x(t)$ satisfies

Copyright © 2001 John Wiley & Sons, Ltd.

$x(0) = x_0$ and $x(T) = 0$. The set of all null controllable states is called the null controllable region of the system and is denoted by $\mathscr{C}$.

Definition 2

Consider system (1) in the absence of the disturbance w. A state x_0 is said to be asymptotically null controllable if there exists a measurable control u such that the state trajectory $x(t)$ satisfies $x(0) = x_0$ and $\lim_{t\to\infty} x(t) = 0$. The set of all asymptotically null controllable states is called the asymptotic null controllable region of the system and is denoted by $\mathscr{C}_a$.

In this paper, the matrix A (or the corresponding linear system) is said to be anti-stable if all of its eigenvalues are in the open right-half-plane and semi-stable if all of its eigenvalues are in the closed left-half-plane.

Proposition 1

Assume that (A, b) is stabilizable.

(a) if A is semi-stable, then $\mathscr{C}_a = \mathbf{R}^n$.
(b) If A is anti-stable, then $\mathscr{C}_a = \mathscr{C}$ is a bounded convex open set containing the origin.
(c) If

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$$

with $A_1 \in \mathbf{R}^{n_1 \times n_1}$ anti-stable and $A_2 \in \mathbf{R}^{n_2 \times n_2}$ semi-stable, and b is partitioned as

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

accordingly, then $\mathscr{C}_a = \mathscr{C}_1 \times \mathbf{R}^{n_2}$ where $\mathscr{C}_1$ is the null controllable region of the anti-stable system $\dot{x}_1 = A_1 x_1 + b_1 \mathrm{sat}(u)$.

Note that if (A, b) is controllable, then $\mathscr{C}_a = \mathscr{C}$.

Proposition 1 follows from a similar result on the null controllable region in Reference [16] by further partitioning A_2 and b_2 as

$$A_2 = \begin{bmatrix} A_{20} & 0 \\ 0 & A_{2-} \end{bmatrix}, \quad b_2 = \begin{bmatrix} b_{20} \\ b_{2-} \end{bmatrix}$$

where A_{20} has all its eigenvalues on the imaginary axis and A_{2-} is Hurwitz. Let the state be partitioned accordingly as $x = [x_1^{\mathrm{T}}\ x_{20}^{\mathrm{T}}\ x_{2-}^{\mathrm{T}}]^{\mathrm{T}}$, with $x_{20} \in \mathbf{R}^{n_{20}}$, $x_{2-} \in \mathbf{R}^{n_{2-}}$. Then

$$\left(\begin{bmatrix} A_1 & 0 \\ 0 & A_{20} \end{bmatrix}, \begin{bmatrix} b_1 \\ b_{20} \end{bmatrix} \right)$$

is controllable and the null controllable region corresponding to the state $[x_1^{\mathrm{T}}\ x_{20}^{\mathrm{T}}]^{\mathrm{T}}$ is $\mathscr{C}_1 \times \mathbf{R}^{n_{20}}$ by Reference [16]. After $[x_1^{\mathrm{T}}\ x_{20}^{\mathrm{T}}]^{\mathrm{T}}$ is steered to the origin, the control can be removed and the state x_{2-} will approach the origin asymptotically.

Our objective is to design a family of feedback laws such that given any (arbitrarily large) set χ_0 in the interior of $\mathscr{C}_a$ and any (arbitrarily small) set χ_∞ containing the origin in its interior, there is a feedback law from this family such that any trajectory of the closed-loop system that starts from

Copyright © 2001 John Wiley & Sons, Ltd.

χ_0 will enter χ_∞ in a finite time and remain there. A complete treatment of this problem was provided in Reference [15] for ANCBC systems. For such a system, a linear feedback can be designed so that the domain of attraction of a small neighbourhood of the origin includes any prescribed bounded set and the disturbance is rejected to an arbitrary level of accuracy. It should be noted that Reference [15] allows for multi-input and more general saturation functions but has the limitation that A has no exponentially unstable eigenvalues, i.e. A is semi-stable. Many earlier papers on control with saturating actuators also have this limitation. The main reason is that if A has exponentially unstable eigenvalues, the largest possible region that can be asymptotically stabilized, i.e. the null controllable region, was unknown.

To achieve our control objectives for exponentially unstable systems, we must know how to describe $\mathscr{C}_a$. In Reference [16], we gave some simple exact descriptions of $\mathscr{C}_a$, and constructed a family of switching saturated linear controllers for a system with two exponentially unstable modes that semi-globally stabilizes the system on $\mathscr{C}_a$. For easy reference, we give a brief review of the results in Reference [16] in the following subsection.

2.2. *Background*

Consider the system

$$\dot{x} = Ax + b\,\mathrm{sat}(u), \tag{2}$$

If A is anti-stable, then $\mathscr{C}_a = \mathscr{C}$ is a bounded convex open set. It was shown in Reference [16] that $\partial\mathscr{C}$ is composed of a set of extremal trajectories of the time reversed system of (2).

The second main result in Reference [16] is about the stability analysis of the following closed-loop system:

$$\dot{x} = Ax + b\,\mathrm{sat}(fx), \quad x \in \mathbf{R}^2 \tag{3}$$

where $A \in \mathbf{R}^{2\times 2}$ is anti-stable and $A + bf$ is Hurwitz. The time-reversed system of (3) is

$$\dot{z} = -Az - b\,\mathrm{sat}(fz) \tag{4}$$

Denote the state transition map of (3) by $\phi:(t, x_0) \mapsto x(t)$ and that of (4) by $\psi:(t, z_0) \mapsto z(t)$. Then the domain of attraction of the equilibrium $x_e = 0$ for (3) is defined by

$$\mathscr{S} := \left\{ x_0 \in \mathbf{R}^2 : \lim_{t\to\infty} \phi(t, x_0) = 0 \right\}$$

Proposition 2

$\mathscr{S}$ is convex and symmetric. $\partial\mathscr{S}$ is the unique limit cycle of systems (3) and (4), and has two intersections with each of the lines $fx = 1$ and $fx = -1$. Furthermore, $\cdot\mathscr{S}$ is the positive limit set of $\psi(\cdot, z_0)$ for all $z_0 \neq 0$.

It was also shown that $\mathscr{S}$ can be made arbitrarily close to $\mathscr{C}$ by suitably choosing f. Since A is anti-stable and (A, b) is controllable, the following Riccati equation

$$A'P + PA - Pbb'P = 0 \tag{5}$$

has a unique positive-definite solution $P > 0$. Let $f_0 = -b'P$. Then the origin is a stable equilibrium of the system

$$\dot{x} = Ax + b\,\mathrm{sat}(kf_0x), \quad x \in \mathbf{R}^2 \tag{6}$$

for all $k > 0.5$. Let $\mathscr{S}(k)$ be the domain of attraction of the equilibrium $x_e = 0$ for (6).

Copyright © 2001 John Wiley & Sons, Ltd.

Proposition 3
$\lim_{k\to\infty} d(\mathscr{S}(k), \mathscr{C}) = 0.$

Hence, the domain of attraction can be made to include any compact subset of $\mathscr{C}$ by simply increasing the feedback gain. We say that the system is semi-globally stabilized (on its null controllable region) by the family of feedbacks $u = \text{sat}(kf_0x)$, $k > 0.5$. This result was then extended to construct a family of switching saturated linear feedback laws that semi-globally stabilizes a higher-order system with two anti-stable modes.

2.3. *Main results of this paper*

Given any (arbitrarily small) set that contains the origin in its interior, we will show that its domain of attraction can be made to include any compact subset of $\mathscr{C}_a$ in the presence of disturbances bounded by an (arbitrarily large) given number. More specifically, we will establish the following result on semi-global practical stabilization on the asymptotically null controllable region for system (1).

Theorem 1
Consider system (1) with A having two exponentially unstable eigenvalues. Given any set $\chi_0 \subset \text{int}(\mathscr{C}_a)$, any set χ_∞ such that $0 \in \text{int}(\chi_\infty)$, and any positive number D, there is a feedback law $u = F(x)$ such that any trajectory of the closed-loop system enters and remains in the set χ_∞ in a finite time as long as it starts from the set χ_0.

To prove Theorem 1, we need to establish some properties of planar linear systems, both in the absence and in the presence of actuator saturation.

3. PROPERTIES OF THE TRAJECTORIES OF SECOND-ORDER LINEAR SYSTEMS

We first consider the second-order anti-stable system

$$\dot{x} = Ax = \begin{bmatrix} 0 & -a_1 \\ 1 & a_2 \end{bmatrix} x, \quad a_1, a_2 > 0 \tag{7}$$

We will examine its trajectories with respect to a horizontal line $kfx = 1$ where $f = [0\ 1]$, $k > 0$. On this line, $x_2 = 1/k$ and if $x_1 > -a_2/k$, then $\dot{x}_2 > 0$, i.e. the vector $\dot{x}$ points upward; if $x_1 < -a_2/k$, then $\dot{x}_2 < 0$, i.e. the vector $\dot{x}$ points downward. Above the line, $\dot{x}_1 < 0$, hence the trajectories all go leftward. Denote.

$$a_m = \begin{cases} -\frac{\lambda_1}{k} & \text{if } A \text{ has real eigenvalues } \lambda_1 \geqslant \lambda_2 > 0 \\ \infty & \text{if } A \text{ has a pair of complex eigenvalues} \end{cases}$$

Then we have

Lemma 1
Let $x_{11} \geqslant -a_2/k$ and

$$p = \begin{bmatrix} x_{11} \\ \frac{1}{k} \end{bmatrix}$$

Copyright © 2001 John Wiley & Sons, Ltd.

be a point on the line $kfx = 1$. The trajectory $x(t) = e^{At}p, t \geqslant 0$ will return to this line if and only if $x_{11} < a_m$. Let T be the first time when it returns and

$$p' = \begin{bmatrix} y_{11} \\ \frac{1}{k} \end{bmatrix}$$

be the corresponding intersection, i.e. $p' = e^{AT}p$. This defines two functions: $x_{11} \to y_{11}$ and $x_{11} \to T$. Then for all $x_{11} \in (-a_2/k, a_m)$,

$$\frac{dy_{11}}{dx_{11}} < -1, \quad \frac{d^2y_{11}}{dx_{11}^2} < 0, \quad \frac{dT}{dx_{11}} > 0 \tag{8}$$

Proof. See Appendix A. □

It may be easier to interpret Lemma 1 by writing (8) as

$$\frac{d(-y_{11})}{dx_{11}} > 1, \quad \frac{d^2(-y_{11})}{dx_{11}^2} > 0$$

An illustration of Lemma 1 is given in Figure 1, where p_1, p_2, p_3 are three points on $kfx = 1$,

$$p_i = \begin{bmatrix} x_{11}^i \\ \frac{1}{k} \end{bmatrix}, \quad x_{11}^i \in [-\tfrac{a_2}{k}, a_m), \quad i = 1, 2, 3,$$

and p_1', p_2' and p_3' are the first intersections of the trajectories that start from p_1, p_2 and p_3. Then

$$\frac{\|p_3' - p_2'\|}{\|p_3 - p_2\|} > \frac{\|p_2' - p_1'\|}{\|p_2 - p_1\|} > 1 \tag{9}$$

It follows that

$$\frac{\|p_2' - p_1'\|}{\|p_3' - p_2'\|} < \frac{\|p_2 - p_1\|}{\|p_3 - p_2\|} \Rightarrow \frac{1 + \frac{\|p_2' - p_1'\|}{\|p_3' - p_2'\|}}{1 + \frac{\|p_2 - p_1\|}{\|p_3 - p_2\|}} < 1$$

Hence

$$\frac{\|p_3' - p_1'\|}{\|p_3 - p_1\|} = \frac{\|p_3' - p_2'\| + \|p_2' - p_1'\|}{\|p_3 - p_2\| + \|p_2 - p_1\|} = \frac{\|p_3' - p_2'\|}{\|p_3 - p_2\|} \frac{1 + \frac{\|p_2' - p_1'\|}{\|p_3' - p_2'\|}}{1 + \frac{\|p_2 - p_1\|}{\|p_3 - p_2\|}} < \frac{\|p_3' - p_2'\|}{\|p_3 - p_2\|} \tag{10}$$

Also from (9)

$$\frac{\|p_3' - p_2'\|}{\|p_2' - p_1'\|} > \frac{\|p_3 - p_2\|}{\|p_2 - p_1\|} \Rightarrow \frac{1 + \frac{\|p_3' - p_2'\|}{\|p_2' - p_1'\|}}{1 + \frac{\|p_3 - p_2\|}{\|p_2 - p_1\|}} > 1$$

Hence

$$\frac{\|p_3' - p_1'\|}{\|p_3 - p_1\|} = \frac{\|p_3' - p_2'\| + \|p_2' - p_1'\|}{\|p_3 - p_2\| + \|p_2 - p_1\|} = \frac{\|p_2' - p_1'\|}{\|p_2 - p_1\|} \frac{1 + \frac{\|p_3' - p_2'\|}{\|p_2' - p_1'\|}}{1 + \frac{\|p_3 - p_2\|}{\|p_2 - p_1\|}} > \frac{\|p_2' - p_1'\|}{\|p_2 - p_1\|} \tag{11}$$

Combining (10) and (11), we obtain

$$\frac{\|p_3' - p_2'\|}{\|p_3 - p_2\|} > \frac{\|p_3' - p_1'\|}{\|p_3 - p_1\|} > \frac{\|p_2' - p_1'\|}{\|p_2 - p_1\|} > 1 \tag{12}$$

Copyright © 2001 John Wiley & Sons, Ltd.

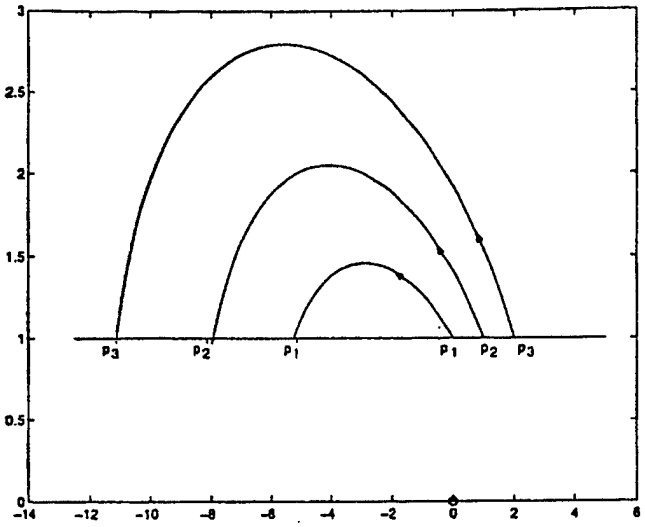


Figure 1. Illustration of Lemma 1.

We next consider a second-order stable linear system,

$$\dot{x} = Ax = \begin{bmatrix} 0 & -a_1 \\ 1 & -a_2 \end{bmatrix} x, \quad a_1, a_2 > 0 \tag{13}$$

We will study the trajectories of (13) with respect to two horizontal lines $kfx = 1$ and $kfx = -1$ where $f = [0\ 1]$, $k > 0$. On the line $kfx = -1$, if $x_1 < -a_2/k$, the vector $\dot{x}$ points downward; if $x_1 > -a_2/k$, the vector $\dot{x}$ points upward.
Let

$$p_0 = \begin{bmatrix} -\frac{a_2}{k} \\ -\frac{1}{k} \end{bmatrix}$$

be a point on $kfx = -1$. There is a point p_0' on $kfx = 1$ and $T_d > 0$ such that $e^{AT_d} p_0' = p_0$, $|kfe^{At} p_0'| \leqslant 1$, $\forall t \in [0, T_d]$ (see Figure 2). Denote the first coordinate of p_0' as x_m, i.e.

$$p_0' = \begin{bmatrix} x_m \\ \frac{1}{k} \end{bmatrix}$$

Let

$$p' = \begin{bmatrix} x_{11} \\ \frac{1}{k} \end{bmatrix}, \ x_{11} \in (-\infty, x_m]$$

be a point on $kfx = 1$, then there is a unique

$$p = \begin{bmatrix} y_{11} \\ -\frac{1}{k} \end{bmatrix}$$

Copyright © 2001 John Wiley & Sons, Ltd.

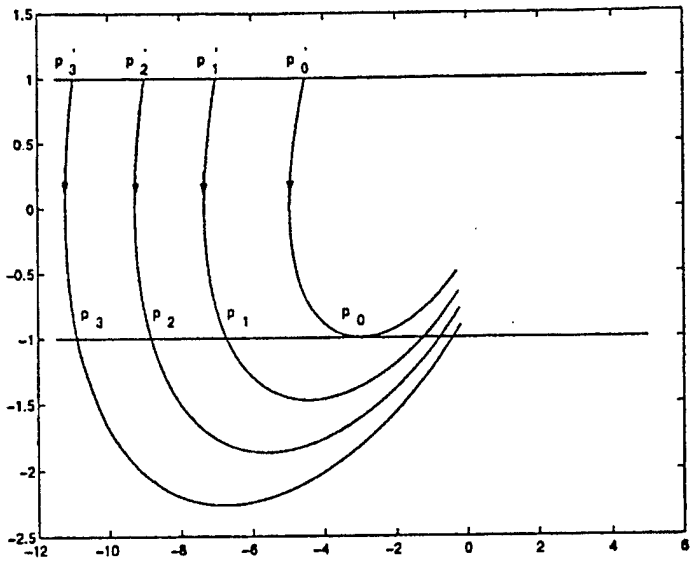


Figure 2. Illustration of Lemma 2.

on $kfx = -1$, where $y_{11} \in (-\infty, -a_2/k]$ and $T \in (0, T_d]$ such that

$$p = e^{AT}p', \qquad |kfe^{At}p'| \leqslant 1, \quad \forall t \in [0, T] \tag{14}$$

This defines two functions $x_{11} \to y_{11}$, and $x_{11} \to T$.

Lemma 2

For all $x_{11} \in (-\infty, x_m)$, we have $x_{11} < y_{11}$ and

$$\frac{dy_{11}}{dx_{11}} > 1, \qquad \frac{d^2y_{11}}{dx_{11}^2} > 0, \quad \frac{dT}{dx_{11}} > 0$$

Proof. See Appendix B. □

This lemma is illustrated with Figure 2, where p'_1, p'_2, p'_3 are three points on $kfx = 1$ and p_1, p_2, p_3 are the three first intersections of $kfx = -1$ with the three trajectories starting from p'_1, p'_2, p'_3, respectively. Then

$$\frac{\|p_1 - p_2\|}{\|p'_1 - p'_2\|} > \frac{\|p_1 - p_3\|}{\|p'_1 - p'_3\|} > \frac{\|p_2 - p_3\|}{\|p'_2 - p'_3\|} > 1$$

4. PROPERTIES OF THE DOMAIN OF ATTRACTION

Consider the closed-loop system

$$\dot{x} = Ax + b\,\mathrm{sat}(kfx), \quad x \in \mathbf{R}^2 \tag{15}$$

Copyright © 2001 John Wiley & Sons, Ltd.

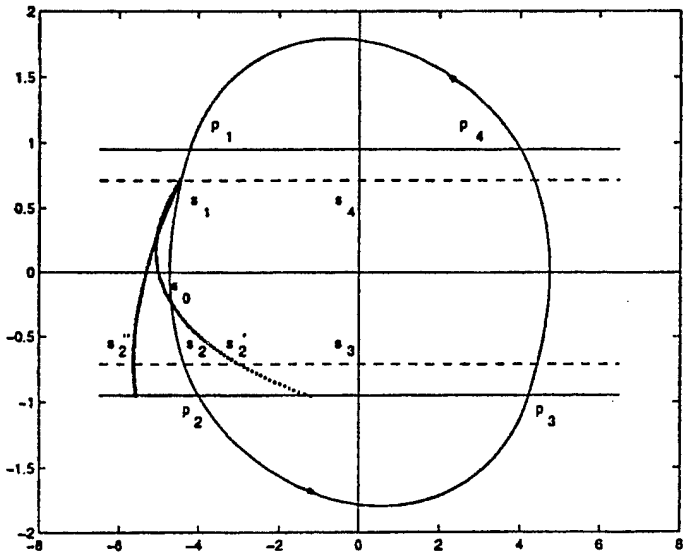


Figure 3. Illustration for the proof of Proposition 4.

where

$$A = \begin{bmatrix} 0 & -a_1 \\ 1 & a_2 \end{bmatrix}, \quad b = \begin{bmatrix} -b_1 \\ -b_2 \end{bmatrix},$$

$a_1, a_2, b_2 > 0$, $b_1 \geqslant 0$, and $f = [0\ 1]$. If $k > a_2/b_2$, then $A + kbf$ is Hurwitz and the origin is the unique equilibrium point of (15) and it is stable. Denote the domain of attraction of the origin as $\mathscr{S}(k)$, then by Proposition 2, $\partial\mathscr{S}(k)$ is the unique limit cycle of (15). We will further show that the domain of attraction $\mathscr{S}(k)$ increases as k is increased.

Consider $k_0 > a_2/b_2$. Denote the increment of k as δ_k. Proposition 2 says that $\partial\mathscr{S}(k_0)$ is symmetric with respect to the origin and has two intersections with each of the lines $k_0 fx = 1$ and $k_0 fx = -1$. In Figure 3, the closed curve is $\partial\mathscr{S}(k_0)$ and $p_1, p_2, p_3\ (= -p_1), p_4 (= -p_2)$ are the four intersections. Since at p_2, the trajectory goes downward, i.e. $\dot{x}_2 < 0$, so $(p_2)_1 < (a_2 - k_0 b_2)/k_0 < 0$. From Lemma 2, we have $(p_1)_1 < (p_2)_1 < 0$. Hence both p_1 and p_2 are on the left half plane. Define

$$\Delta(k_0) = -\frac{(p_2)_1}{b_2} k + \frac{a_2}{b_2} - k_0$$

Then $\Delta(k_0) > 0$ due to the fact that the trajectory goes downward at p_2.

Proposition 4

Suppose $k_0 > a_2/b_2$. Then for all $\delta_k \in (0, \Delta(k_0))$, $\mathscr{S}(k_0) \subset \mathscr{S}(k_0 + \delta_k)$.

Proof. Since $\delta_k > 0$, the two lines $(k_0 + \delta_k) fx = \pm 1$ lie in between $k_0 fx = \pm 1$. It follows that the vector field above $k_0 fx = 1$ and that below $k_0 fx = -1$ are the same for

$$\dot{x} = Ax + b\,\mathrm{sat}(k_0 fx) \tag{16}$$

Copyright © 2001 John Wiley & Sons, Ltd.

and

$$\dot{x} = Ax + b\,\mathrm{sat}((k_0 + \delta_k)fx) \tag{17}$$

So, if a trajectory of (17) starts at p_4 (or p_2), it will go along $\partial\mathscr{S}(k_0)$ to p_1 (or p_3).

Claim

If a trajectory of (17) starts at a point on $\partial\mathscr{S}(k_0)$ between p_1 and p_2 and intersects the line $k_0fx = -1$, then the intersection must be inside $\mathscr{S}(k_0)$.

It follows from the claim that any trajectory of (17) that starts from $\partial\mathscr{S}(k_0)$ will stay inside of $\overline{\mathscr{S}(k_0)}$ when it returns to the lines $k_0fx = \pm 1$. So it is bounded and hence belongs to $\overline{\mathscr{S}(k_0 + \delta_k)}$. Note that any trajectory outside of $\overline{\mathscr{S}(k_0 + \delta_k)}$ will diverge because the system has a unique limit cycle. Since the two sets are convex and open, we will have $\mathscr{S}(k_0) \subset \mathscr{S}(k_0 + \delta_k)$.

It remains to prove the claim.

Since $\mathscr{S}(k_0)$ is convex, $\angle(A + bk_0f)x$ from p_1 to p_2 along $\partial\mathscr{S}(k_0)$ is increasing. Let s_0 be the intersection of $\partial S(k_0)$ with the abscissa. Then at s_0, $\angle(A + bk_0f)x = -\pi/2$; from p_1 to s_0, $\angle(A + bk_0f)x \in (-\pi, -\pi/2)$; and from s_0 to p_2, $\angle(A + bk_0f)x \in (-\pi/2, 0)$. Now consider a point x along $\partial\mathscr{S}(k_0)$ between p_1 and p_2,

(1) If x is between p_1 and s_0, then $k_0fx \leqslant \mathrm{sat}((k_0 + \delta_k)fx)$. If $\angle(A + bk_0f)x < \angle b$, then $\dot{x}$ of (17) directs inward of $\partial\mathscr{S}(k_0)$ and if $\angle(A + bk_0f)x) > \angle b$, then $\dot{x}$ of (17) directs outward of $\partial\mathscr{S}(k_0)$. Since $\angle(A + bk_0f)x$ is increasing, the vector $\dot{x}$ may direct outward of $\partial\mathscr{S}(k_0)$ for the whole segment or for a lower part of the segment.

(2) If x is between s_0 and p_2, then $k_0fx \geqslant \mathrm{sat}((k_0 + \delta_k)fx)$. Since $\angle b \in (-\pi, -\pi/2)$, we have

$$\angle(A + bk_0f)x) \leqslant \angle(Ax + b\,\mathrm{sat}((k_0 + \delta_k)fx))$$

i.e. the vector $\dot{x}$ of (17) directs inward of $\partial\mathscr{S}$ (k_0).

Let

$$s_1 = \begin{bmatrix} x_{11} \\ h \end{bmatrix}, \quad h > 0$$

be a point on $\partial\mathscr{S}(k_0)$ between p_1 and s_0 such that $\dot{x}$ of (17) at s_1 directs outward of $\partial\mathscr{S}(k_0)$.

Let

$$s_2 = \begin{bmatrix} y_{11} \\ -h \end{bmatrix}$$

be the intersection of $\partial\mathscr{S}(k_0)$ with $x_2 = -h$. Then by (1) the trajectory of (17) starting at s_1 will remain outside of $\partial\mathscr{S}(k_0)$ above the abscissa. We need to show that when the trajectory reaches the line $x_2 = -h$ at s_2', it must be inside $\partial\mathscr{S}(k_0)$.

Let

$$s_3 = \begin{bmatrix} 0 \\ -h \end{bmatrix}, \qquad s_4 = \begin{bmatrix} 0 \\ h \end{bmatrix}$$

(see Figure 3). Denote the region enclosed by $s_1s_2s_3s_4s_1$ as G_0, where the part s_1s_2 is on $\partial\mathscr{S}(k_0)$ and the other parts are straight lines. Since this region lies between $k_0fx = \pm 1$, the vector field of

Copyright © 2001 John Wiley & Sons, Ltd.

(16) on this region is

$$\dot{x}_1 = -(a_1 + k_0 b_1)x_2 =: f_1(x)$$
$$\dot{x}_2 = x_1 + (a_2 - k_0 b_2)x_2 =: f_2(x)$$

Applying Green's Theorem to system (16) on G_0, we get

$$\oint_{\partial G_0} f_2\,\mathrm{d}x_1 - f_1\,\mathrm{d}x_2 = -\iint_{G_0}\left(\frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2}\right)\mathrm{d}x_1\,\mathrm{d}x_2 \tag{18}$$

Note that the left-hand side integral from s_1 and s_2 and that from s_3 to s_4 are zero. Denote the area of G_0 as Q_0, then from (18), we have

$$\tfrac{1}{2}x_{11}^2 + (a_2 - k_0 b_2)hx_{11} - \tfrac{1}{2}y_{11}^2 + (a_2 - k_0 b_2)hy_{11} = -(a_2 - k_0 b_2)Q_0 \tag{19}$$

Clearly $Q_0 > -h(x_{11} + y_{11})$ by the convexity of $\mathscr{S}(k_0)$ and the region G_0.

On the other hand, we consider a trajectory of (17) starting at s_1 but cross the line $x_2 = -h$ at

$$s_2' = \begin{bmatrix} y_{11} + \delta_{y11} \\ -h \end{bmatrix}$$

Firstly, we assume that s_1 lies between $(k_0 + \delta_k)fx = \pm 1$. Apply Green's Theorem to (17) on the region enclosed by $s_1 s_2' s_3 s_4 s_1$, where the part $s_1 s_2'$ is on a trajectory of (17). Denote the area of the region as $Q_0 + \delta_Q$. Similarly,

$$\tfrac{1}{2}x_{11}^2 + (a_0 - k_0 b_2 - \delta_k b_2)hx_{11} - \tfrac{1}{2}(y_{11} + \delta_{y11})^2 + (a_2 - k_0 b_2 - \delta_k b_2)h(y_{11} + \delta_{y11})$$
$$= -(a_2 - k_0 b_2 - \delta_k b_2)(Q_0 + \delta_Q) \tag{20}$$

Subtracting (19) from (20), we obtain

$$-[y_{11} - (a_2 - k_0 b_2 - \delta_k b_2)h]\delta_{y11} = (k_0 b_2 - a_2)\delta_Q + \delta_k b_2(Q_0 + hx_{11} + hy_{11}) + \tfrac{1}{2}\delta_{y11}^2 + \delta_k b_2 \delta_Q \tag{21}$$

Note that $Q_0 + hx_{11} + hy_{11} > 0$ and $k_0 b_2 - a_2 > 0$.

From the definition of $\Delta(k_0)$, we have

$$(p_2)_1 - (a_2 - k_0 b_2 - \delta_k b_2)\frac{1}{k_0} < 0$$

for all $\delta_k \in [0, \Delta(k_0))$. Since $y_{11} < (p_2)_1$, $h < 1/k_0$ and $-(a_2 - k_0 b_2 - \delta_k b_2) > 0$, it follows that

$$y_{11} - (a_2 - k_0 b_2 - \delta_k b_2)h < 0, \quad \forall\, \delta_k \in [0, \Delta(k_0))$$

Now, suppose that $\delta_k \in [0, \Delta(k_0))$. If $\delta_{y11} < 0$, then s_2' is outside of $\partial\mathscr{S}(k_0)$ and we must have $\delta_Q > 0$. In this case the left-hand side of (21) is negative and the right-hand side is positive. A contradiction. This shows that δ_{y11} must be positive and s_2' must be inside $\partial\mathscr{S}(k_0)$. By (2), the vector $\dot{x}$ of system (17) directs inward of $\partial\mathscr{S}(k_0)$ from s_2 to p_2, we know that when the trajectory reaches $k_0 fx = -1$, it must be to the right of p_2, i.e. still inside $\partial\mathscr{S}(k_0)$.

Now suppose s_1 lies between $(k_0 + \delta_k)fx = 1$ and $k_0 fx = 1$. Then by applying Green's Theorem, we get exactly the same equation as (21), although we need to partition the region enclosed by $s_1 s_2' s_3 s_4 s_1$ into 3 parts. And similar argument applies. Thus we conclude that for all $\delta_k \in [0, \Delta(k_0))$, $\mathscr{S}(k_0) \subset \mathscr{S}(k_0 + \delta_k)$. □

Copyright © 2001 John Wiley & Sons, Ltd.

Proposition 5
Consider

$$\dot{x} = Ax + b\,\mathrm{sat}(fx), \quad x \in \mathbf{R}^2 \tag{22}$$

where $A \in \mathbf{R}^{2\times 2}$, $b \in \mathbf{R}^{2\times 1}$ are constant matrices, A is anti-stable and $f \in \mathbf{R}^{1\times 2}$ is a variable. Denote the domain of attraction of the origin for (22) as $\mathscr{S}(f)$. Then, at any f such that $A + bf$ is Hurwitz and has distinct eigenvalues, $\overline{\mathscr{S}(f)}$ is continuous.

Proof. We only need to show that $\partial\mathscr{S}(f)$ is continuous. Recall from Proposition 2 that $\partial\mathscr{S}(f)$ is a closed trajectory and has four intersections with $fx = \pm 1$. Since the vector $\dot{x} = Ax + b\,\mathrm{sat}(fx)$ is continuous in f at each x, it suffices to show that one of the intersections is continuous in f. Actually, we can show that the intersections are also differentiable in f. For simplicity and for direct use of Lemmas 1 and 2, we apply a state-space transformation, $\hat{x} = V(f)x$, to system (22), such that

$$V(f)AV^{-1}(f) = \begin{bmatrix} 0 & -a_1 \\ 1 & a_2 \end{bmatrix} =: \hat{A}, \qquad V(f)b = \begin{bmatrix} b_1(f) \\ b_2(f) \end{bmatrix} =: \hat{b}(f), \qquad fV^{-1}(f) = [0\ 1] =: \hat{f} \tag{23}$$

Such a transformation always exists. To see this, assume that A is already in this form. Since A is anti-stable and $A + bf$ is stable, (f, A) must be observable. So

$$V(f) = \begin{bmatrix} fA - a_2 f \\ f \end{bmatrix}$$

in non-singular and it can be verified that this $V(f)$ is the desired transformation matrix. Moreover, $V(f)$, $V^{-1}(f)$, $b_1(f)$ are all analytic in f. Now consider the transformed system

$$\dot{\hat{x}} = \hat{A}\hat{x} + \hat{b}(f)\,\mathrm{sat}(\hat{f}\hat{x}) \tag{24}$$

Note that $\hat{A}$ and $\hat{f}$ are both independent of f. Under the state transformation, $\mathscr{S}(f)$ is transformed into $\hat{\mathscr{S}}(f) = \{V(f)x\colon x \in \mathscr{S}(f)\}$, the domain of attraction for (24) and $\partial\hat{\mathscr{S}}(f)$ is its unique limit cycle. Let

$$p_1 = \begin{bmatrix} \hat{x}_{11} \\ 1 \end{bmatrix}$$

be a point on $\hat{f}\hat{x} = 1$ such that a trajectory starting at p_1 will go above the line and return to the line (for the first time) at

$$p_1' = \begin{bmatrix} \hat{y}_{11} \\ 1 \end{bmatrix}$$

Let T_1 be the time for the trajectory to go from p_1 to p_1', then

$$e^{\hat{A}T_1}(p_1 + \hat{A}^{-1}\hat{b}(f)) = (p_1' + \hat{A}^{-1}\hat{b}(f))$$

or equivalently,

$$e^{\hat{A}T_1}\begin{bmatrix} \hat{x}_{11} + (\hat{A}^{-1}\hat{b}(f))_1 \\ 1 + (\hat{A}^{-1}\hat{b}(f))_2 \\ 1 \end{bmatrix} = \begin{bmatrix} \hat{y}_{11} + (\hat{A}^{-1}\hat{b}(f))_1 \\ 1 + (\hat{A}^{-1}\hat{b}(f))_2 \\ 1 \end{bmatrix}$$

Copyright © 2001 John Wiley & Sons, Ltd.

where $(\cdot)_i$, $i = 1, 2$, denotes the ith coordinate of a vector. It can be verified from the stability of $\hat{A} + \hat{b}(f)\hat{f}$ that $1 + (\hat{A}^{-1}\hat{b}(f))_2 > 0$. So Lemma 1 applies here with a changing of variables. We can write $\hat{y}_{11} = \hat{y}_{11}(f, \hat{x}_{11})$. By Lemma 1, $\hat{y}_{11}$ is continuously differentiable in $\hat{x}_{11}$. It is easy to see that $\hat{y}_{11}$ is also continuously differentiable in f.

Suppose that the trajectory continuous from p_1' and intersects the line $\hat{f}\hat{x} = -1$ at a non-zero angle. Let

$$p_1'' = \begin{bmatrix} \hat{z}_{11} \\ -1 \end{bmatrix}$$

be the first intersection of the trajectory with $\hat{f}\hat{x} = -1$. Note that between $\hat{f}\hat{x} = 1$ and $\hat{f}\hat{x} = -1$, the vector field of (24) is

$$\dot{\hat{x}} = (\hat{A} + \hat{b}(f)\hat{f})\hat{x} = \begin{bmatrix} 0 & -a_1 + b_1(f) \\ 1 & a_2 + b_2(f) \end{bmatrix} \hat{x}$$

and that $\hat{A} + \hat{b}(f)\hat{f}$ is Hurwitz, so Lemma 2 applies and we know that $\hat{z}_{11}$ is continuously differentiable in $\hat{y}_{11}$. To see that $\hat{z}_{11}$ is also continuously differentiable in f, recall we have assumed that $A + bf$ has distinct eigenvalues, so the eigenvalues are analytic in f. From (48) in the proof of Lemma 2, we see that T is continuously differentiable in λ_1, λ_2 and hence in f for $T < T_d$. Thus $\hat{z}_{11}$ is also continuously differentiable in f. (Here $\hat{z}_{11}$ corresponds to y_{11} in (B2) and $\hat{y}_{11}$ to x_{11} in (B1).) In summary, we can write

$$\hat{z}_{11} = \hat{z}_{11}(f, \hat{x}_{11})$$

where $\hat{z}_{11}$ is continuously differentiable in f and $\hat{x}_{11}$. Now suppose

$$p_1 = \begin{bmatrix} \hat{x}_{11} \\ 1 \end{bmatrix}$$

is a point in the limit cycle $\partial\hat{\mathcal{P}}(f)$, then we must have $\hat{z}_{11} = -\hat{x}_{11}$, i.e.,

$$\hat{z}_{11}(f, \hat{x}_{11}) + \hat{x}_{11} = 0 \tag{25}$$

due to the symmetry of $\partial\hat{\mathcal{P}}(f)$. We write $g(f, \hat{x}_{11}) = \hat{z}_{11}(f, \hat{x}_{11}) + \hat{x}_{11} = 0$.

By the uniqueness of the limit cycle, $\hat{x}_{11}$ is uniquely determined by f. By Lemmas 1 and 2, we know $\partial\hat{z}_{11}/\partial\hat{x}_{11} = (\partial\hat{z}_{11}/\partial\hat{y}_{11})\partial\hat{y}_{11}/\partial\hat{x}_{11} < -1$, so $\partial g/\partial\hat{x}_{11} \neq 0$ and by the implicit function theorem, $\hat{x}_{11}$ is differentiable in f. Recall that

$$p_1 = \begin{bmatrix} \hat{x}_{11} \\ 1 \end{bmatrix}$$

is a point in the vector field of (24). The corresponding intersection in the original system (22) is

$$V^{-1}(f)\begin{bmatrix} \hat{x}_{11} \\ 1 \end{bmatrix}$$

Clearly, it is also differentiable in f. □

Combining Propositions 4 and 5, we have

Copyright © 2001 John Wiley & Sons, Ltd.

Corollary 1

Consider system (15) with A, b and f in the specified form. Given k_1 and k_2, $k_2 > k_1 > a_2/b_2$. Suppose that $A + kbf$ has distinct eigenvalues for all $k \in [k_1, k_2]$. Then $\mathscr{S}(k) \subset \mathscr{S}(k + \delta_k)$ for all $k \in [k_1, k_2]$, $\delta_k \in [0, k_2 - k]$.

Proof. By proposition 5, $\partial\mathscr{S}(k)$ is continuous in k for all $k \in [k_1, k_2]$. So $(p_2)_1$ and hence the function $\Delta(k)$ are also continuous in k. If follows that $\min\{\Delta(k): k \in [k_1, k_2]\} > 0$. By applying Proposition 4, we have the corollary. □

It can be seen that there exists a $k_0 > 0$ such that $A + kbf$ has distinct eigenvalues for all $k > k_0$. Thus by Corollary 1, $\mathscr{S}(k)$ will be continuous and monotonically increasing for all $k > k_0$.

5. PROOF OF THEOREM 1: THE SECOND-ORDER CASE

We will prove the theorem by explicit construction of a family of feedback laws that solve the problem. To this end, let us first establish some preliminary results for a general system (1), not necessarily second order or anti-stable. Let $P(\varepsilon)$ be the positive definite solution of the Riccati equation.

$$A'P + PA - Pbb'P + \varepsilon I = 0 \tag{26}$$

It is known that $P(\varepsilon)$ is continuous for $\varepsilon \geqslant 0$. Let $f(\varepsilon) = -b'P(\varepsilon)$. With $u = kf(\varepsilon)x$, we have the closed-loop system

$$\dot{x} = Ax + b\,\mathrm{sat}(kf(\varepsilon)x + w) \tag{27}$$

Clearly, $A + kbf(\varepsilon)$ is Hurwitz for all $k \geqslant 0.5$. For $x(0) = x_0$, $w \in \mathscr{W}$, denote the state trajectory of (27) as $\psi(t, x_0, w)$.

Lemma 3

Consider system (27). Let $\varepsilon > 0$ be given. Let $c_\infty = \sigma_{\max}(P(\varepsilon))D^2/\varepsilon(2k - 1)$, $c_0 = 4/b'P(\varepsilon)b$. Suppose k is sufficiently large such that $c_\infty < c_0$. Denote

$$\mathscr{S}_p(\varepsilon) := \{x: x'P(\varepsilon)x \leqslant c_0\}$$

$$\mathscr{S}_\infty(\varepsilon, k) := \{x: x'P(\varepsilon)x \leqslant c_\infty\}$$

Then, $\mathscr{S}_p(\varepsilon)$ and $\mathscr{S}_\infty(\varepsilon, k)$ are invariant sets, and, for any $w \in \mathscr{W}$, $x_0 \in \mathscr{S}_p(\varepsilon)$, $\psi(t, x_0, w)$ will enter $\mathscr{S}_\infty(\varepsilon, k)$ in a finite time and remain there.

Proof. Let $V(x) = x'P(\varepsilon)x$. It suffices to show that for all $x \in \mathscr{S}_p(\varepsilon)\backslash\mathscr{S}_\infty(\varepsilon, k)$ and for all $|w| \leqslant D$, $\dot{V} < 0$. In the following, we simply write $P(\varepsilon)$ as P and $f(\varepsilon)$ as f, since in this lemma, ε is fixed. Note that

$$\dot{V} = x'(A'P + PA)x + 2x'Pb\,\mathrm{sat}(kfx + w)$$

We will consider the case where $x'Pb \geqslant 0$. The case where $x'Pb \leqslant 0$ is similar.

Copyright © 2001 John Wiley & Sons, Ltd.

If $kfx + w \leqslant -1$, then

$$\begin{aligned}\dot{V} &= x'(A'P + PA)x - 2x'Pb \\ &= x'Pbb'Px - 2x'Pb - \varepsilon x'x \\ &= x'Pb(x'Pb - 2) - \varepsilon x'x\end{aligned}$$

Since $x'Px \leqslant c_0 = 4/b'Pb$, we have $b'Px \leqslant \|b'P^{1/2}\| \, \|P^{1/2}x\| \leqslant 2$, and hence $\dot{V} < 0$.

If $kfx + w > -1$, then $\text{sat}(kfx + w) \leqslant kfx + w$, and,

$$\begin{aligned}\dot{V} &\leqslant x'(A'P + PA)x + 2x'Pb(kfx + w) \\ &= -(2k-1)x'Pbb'Px - \varepsilon x'x + 2x'Pbw \\ &= -\left(\sqrt{2k-1}\,x'Pb - \frac{w}{\sqrt{2k-1}}\right)^2 + \frac{w^2}{2k-1} - \varepsilon x'x\end{aligned}$$

Since $x'Px > c_\infty = \sigma_{\max}(P)D^2/\varepsilon(2k-1)$, we have $x'x > D^2/\varepsilon(2k-1)$. It follows that $\dot{V} < 0$. □

It is clear from Lemma 3 that as k goes to infinity, $\mathscr{S}_\infty(\varepsilon, k)$ converges to the origin. In particular, there exists a k such that $\mathscr{S}_\infty(\varepsilon, k) \subset \chi_\infty$.

For any ANCBC system, as $\varepsilon \to 0$, $P(\varepsilon) \to 0$, and $c_0 \to \infty$. Thus $\mathscr{S}_p(\varepsilon)$ can be made arbitrarily large; and with a fixed ε, we can increase k to make c_∞ arbitrarily small. So the proof of Theorem 1 would have been completed here. However, for exponentially unstable systems, $\mathscr{S}_p(\varepsilon)$ is a quite small subset of $\mathscr{C}_a$ as $\varepsilon \to 0$ [16] and hence considerable work needs to be carried out before completing the proof.

Define the domain of attraction of the origin in the absence of disturbance as

$$\mathscr{S}(\varepsilon, k) := \left\{x_0 : \lim_{t \to \infty} \psi(t, x_0, 0) = 0\right\}$$

and in the presence of disturbance, define the domain of attraction of the set $\mathscr{S}_\infty(\varepsilon, k)$ as

$$\mathscr{S}_D(\varepsilon, k) := \left\{x_0 : \lim_{t \to \infty} d(\psi(t, x_0, w), \mathscr{S}_\infty(\varepsilon, k)) = 0, \forall w \in \mathscr{W}\right\}$$

where $d(\psi(t, x_0, w), \mathscr{S}_\infty(\varepsilon, k))$ is the distance between the point $\psi(t, x_0, w)$ and the set $\mathscr{S}_\infty(\varepsilon, k)$. Our objective is to choose ε and k such that $\chi_0 \subset \mathscr{S}_D(\varepsilon, k)$ and $\mathscr{S}_\infty(\varepsilon, k) \subset \chi_\infty$.

Clearly $\mathscr{S}_p(\varepsilon) \subset \mathscr{S}_D(\varepsilon, k) \subset \mathscr{S}(\varepsilon, k)$. By using the Lyapunov function $V(x) = x'P(\varepsilon)x$, we can only determine a subset $\mathscr{S}_p(\varepsilon)$ of $\mathscr{S}_D(\varepsilon, k)$. As ε decreases, $P(\varepsilon)$ decreases. It was shown in Reference [17] that if $\varepsilon_1 < \varepsilon_2$, then $\mathscr{S}_p(\varepsilon_2) \subset \mathscr{S}_p(\varepsilon_1)$. So by decreasing ε, we can enlarge $\mathscr{S}_p(\varepsilon)$. However, since $\lim_{\varepsilon \to 0} \mathscr{S}_p(\varepsilon)$ can be much smaller than $\mathscr{C}_a$, we are unable to prove that $\mathscr{S}_D(\varepsilon, k)$ is close to $\mathscr{C}_a$ by simply enlarging $\mathscr{S}_p(\varepsilon)$ as was done in Reference [15]. For this reason, we will resort to the detailed investigation on the vector field of (27) in the presence of the disturbance.

We now continue with the proof of the theorem and focus on the second order systems. Also assume that A is anti-stable. In this case $\mathscr{C}_a = \mathscr{C}$.

We will prove the theorem by showing that, given any $\chi_0 \subset \text{int}(\mathscr{C})$, any (arbitrarily small) χ_∞ such that $0 \in \text{int}(\chi_\infty)$, and any $D > 0$, there exist an $\varepsilon > 0$ and a $k \geqslant 0.5$ such that $\chi_0 \subset \mathscr{S}_D(\varepsilon, k)$ and $\mathscr{S}_\infty(\varepsilon, k) \subset \chi_\infty$.

Copyright © 2001 John Wiley & Sons, Ltd.

Proposition 3 applies to the case where $\varepsilon = 0$. It means that $\lim_{k\to\infty} d(\mathscr{S}(0,k),\mathscr{C}) = 0$. But when $\varepsilon = 0$, it is impossible to achieve disturbance rejection by increasing the value of k even if there is no saturation. We can first let $\varepsilon = 0$, choose k_0 sufficiently large so that $A + k_0 bf(\varepsilon)$ has distinct eigenvalues and $\chi_0 \subset \text{int}(\mathscr{S}(0,k_0))$. Then by the continuity of the domain of attraction stated in Proposition 5 and the continuity of the solution of the Ricatti equation, we can fix this k_0 and choose ε sufficiently small so that $\chi_0 \subset \text{int}(\mathscr{S}(\varepsilon,k_0))$. By Corollary 1, we know that $\mathscr{S}(\varepsilon,k)$ is non-decreasing, so $\chi_0 \subset \text{int}(\mathscr{S}(\varepsilon,k))$ for all $k \geqslant k_0$. What remains to be shown is that for any given positive number D and a fixed ε, we can choose k sufficiently large so that $d(\mathscr{S}_D(\varepsilon,k), \mathscr{S}(\varepsilon,k))$ is arbitrarily small. Then we will have $\chi_0 \subset \mathscr{S}_D(\varepsilon,k)$ for some k.

Now, let us fix an ε such that $\chi_0 \subset \text{int}(\mathscr{S}(\varepsilon,k))$, $\forall k \geqslant k_0$. Since ε is fixed, we can assume that a state transformation $\hat{x} = Vx$ like (23) is performed so that

$$\hat{f} = -b'P(\varepsilon)V^{-1} = [0 \;\; 1] \tag{28}$$

$$\hat{A} = VAV^{-1} = \begin{bmatrix} 0 & -a_1 \\ 1 & a_2 \end{bmatrix}, \quad \hat{b} = Vb = \begin{bmatrix} -b_1 \\ -b_2 \end{bmatrix}, \qquad a_1, a_2, b_1, b_2 > 0 \tag{29}$$

where $a_1, a_2 > 0$ is from the anti-stability of A and $b_1, b_2 > 0$ follows from the fact that an LQ controller has infinite gain margin and $\varepsilon \neq 0$. ($b_1 = 0$ iff $\varepsilon = 0$). Under this state transformation, the sets $\mathscr{S}_p(\varepsilon)$, $\mathscr{S}_D(\varepsilon,k)$, $\mathscr{S}(\varepsilon,k)$, $\mathscr{S}_\infty(\varepsilon,k)$, $\mathscr{C}$, χ_0 and χ_∞ are transformed, respectively, into $\hat{\mathscr{S}}_p(\varepsilon)$, $\hat{\mathscr{S}}_D(\varepsilon,k)$, $\hat{\mathscr{S}}(\varepsilon,k)$, $\hat{\mathscr{S}}_\infty(\varepsilon,k)$, $\hat{\mathscr{C}}$, $\hat{\chi}_0$ and $\hat{\chi}_\infty$, all defined in an obvious way. For example, $\hat{\mathscr{C}} = \{Vx : x \in \mathscr{C}\}$. Let $\hat{P}(\varepsilon) = (V^{-1})'P(\varepsilon)V^{-1}$. Since ε is now fixed, we denote $\hat{P}(\varepsilon)$, $\hat{\mathscr{S}}_p(\varepsilon)$, $\hat{\mathscr{S}}_D(\varepsilon,k)$, $\hat{\mathscr{S}}(\varepsilon,k)$, and $\hat{\mathscr{S}}_\infty(\varepsilon,k)$, as $\hat{P}$, $\hat{\mathscr{S}}_p$, $\hat{\mathscr{S}}_D(k)$, $\hat{\mathscr{S}}(k)$ and $\hat{\mathscr{S}}_\infty(k)$, respectively.

Now we consider

$$\dot{\hat{x}} = \hat{A}\hat{x} + \hat{b}\,\text{sat}(k\hat{f}\hat{x} + w) \tag{30}$$

This standard form fits very well into Corollary 1, so we can be sure that $\hat{\mathscr{S}}(k)$ increases as k is increased. It follows that

$$\hat{\mathscr{S}}(k_0) \subset \hat{\mathscr{S}}(k), \quad \forall k > k_0$$

To satisfy the design requirement, it is necessary that no point in $\hat{\chi}_0 \backslash \hat{\chi}_\infty$ can be made stationary with any $|w| \leqslant D$. Let us first exclude this possibility by appropriate choice of k.

For a constant w, there are three candidate equilibrium points, $\hat{x}_e^+ = -\hat{A}^{-1}\hat{b}$, $\hat{x}_e^- = \hat{A}^{-1}\hat{b}$ and $\hat{x}_e^w = -(\hat{A} + k\hat{b}\hat{f})^{-1}\hat{b}w$, corresponding to $\text{sat}(k\hat{f}\hat{x} + w) = 1$, $\text{sat}(k\hat{f}\hat{x} + w) = -1$ and $\text{sat}(k\hat{f}\hat{x} + w) = k\hat{f}\hat{x} + w$, respectively. For each of them to be an actual equilibrium point, we must have

$$k\hat{f}\hat{x}_e^+ + w \geqslant 1,\; k\hat{f}\hat{x}_e^- + w \leqslant -1 \text{ or } |k\hat{f}\hat{x}_e^w + w| \leqslant 1$$

respectively.

Here we have

$$\hat{x}_e^+ = \frac{1}{a_1}\begin{bmatrix} a_2b_1 + a_1b_2 \\ -b_1 \end{bmatrix}, \qquad \hat{x}_e^- = -\hat{x}_e^+, \qquad \hat{x}_e^w = \frac{1}{a_1 + b_1k}\begin{bmatrix} a_2b_1 + a_1b_2 \\ -b_1 \end{bmatrix} w$$

If $\hat{A}$ has no complex eigenvalues, then $\hat{x}_e^+, \hat{x}_e^- \in \partial\hat{\mathscr{C}}$ [16], so $\hat{x}_e^+, \hat{x}_e^- \notin \hat{\chi}_0$ for any $\hat{\chi}_0 \subset \text{int}(\hat{\mathscr{C}})$. But if $\hat{A}$ has a pair of complex eigenvalues, $\hat{x}_e^+, \hat{x}_e^- \in \text{int}(\hat{\mathscr{C}})$ and will be in $\hat{\chi}_0$ if $\hat{\chi}_0$ is close enough to $\hat{\mathscr{C}}$. So, it is desirable that $\hat{x}_e^+$ and $\hat{x}_e^-$ cannot be made stationary by any $|w| \leqslant D$. This requires

$$k\hat{f}\hat{x}_e^+ + w < 1, \qquad k\hat{f}\hat{x}_e^- + w > -1, \quad \forall |w| \leqslant D$$

Copyright © 2001 John Wiley & Sons, Ltd.

which is equivalent to $k(b_1/a_1) + w > -1, \forall |w| \leqslant D$. If $D \leqslant 1$, this is satisfied for all k; if $D > 1$, we need to choose k such that

$$k > \frac{a_1}{b_1}(D-1)$$

Note that this will be impossible if $b_1 = 0$, which corresponds to the case where $\varepsilon = 0$. This is one reason that ε should be non-zero.

Finally, as $k \to \infty$, $\hat{x}_e^w \to 0$ for all $|w| \leqslant D$. So k can be chosen large enough such that $\hat{x}_e^w \notin \hat{\chi}_0 \setminus \hat{\chi}_\infty$.

In summary, from the above analysis, we will restrict ourselves to k such that

$$k > \frac{a_1}{b_1}(D-1), \qquad \frac{D}{a_1 + b_1 k}\begin{bmatrix} a_2 b_1 + a_1 b_2 \\ -b_1 \end{bmatrix} \in \chi_\infty \tag{31}$$

To study the vector field of (30), we rewrite it as

$$\dot{\hat{x}}_1 = -a_1 \hat{x}_2 - b_1 \operatorname{sat}(k\hat{f}\hat{x} + w)$$

$$\dot{\hat{x}}_2 = \hat{x}_1 + a_2 \hat{x}_2 - b_2 \operatorname{sat}(k\hat{f}\hat{x} + w)$$

The vector field is much complicated by the presence of the disturbance. However, it still exhibits some properties which we will make use in our construction of the desired controller:

- Above the line $k\hat{f}\hat{x} = D + 1$, $k\hat{f}\hat{x} + w \geqslant 1$ for all $|w| \leqslant D$, so $\operatorname{sat}(k\hat{f}\hat{x} + w) = 1$, i.e. the vector $\dot{\hat{x}}$ is independent of w and is affine in $\hat{x}$. Similarly, below $k\hat{f}\hat{x} = -(D+1)$, $\operatorname{sat}(k\hat{f}\hat{x} + w) = -1$.
- In the ellipsoid $\hat{\mathscr{P}}_p$, we have shown that all the trajectories will converge to $\hat{\mathscr{P}}_\infty(k)$, which can be made arbitrarily small by increasing the value of k.

Suppose that k is sufficiently large such that the boundary of $\hat{\mathscr{P}}_p$ intersects with the lines $k\hat{f}\hat{x} = \pm(D+1)$. Denote the region between $k\hat{f}\hat{x} = (D+1)$ and $k\hat{f}\hat{x} = -(D+1)$, and to the left of $\hat{\mathscr{P}}_p$ as $Q(k)$, see the shaded region in Figure 4. Let

$$\hat{x}_m(k) = -\max\{\hat{x}_1 : \hat{x} \in Q(k)\}$$

If k is sufficiently large, then $Q(k)$ lies entirely in the left-half-plane, so $\hat{x}_m(k) > 0$. Choose K such that

$$-x_m(K) + a_2 \frac{D+1}{K} < 0, \qquad \frac{-x_m(K) + a_2(D+1)/K}{-a_1(D+1)/K} > \frac{b_2}{b_1} \tag{32}$$

(Note that $x_m(k)$ increases as k is increased.) Then the vector field in $Q(k)$ has the following property:

Lemma 4

Suppose $k > K$. Then for all $\hat{x} \in Q(k)$, $|w| \leqslant D$,

$$\tan^{-1}\left(\frac{b_2}{b_1}\right) - \pi < \angle\left(\hat{A}\begin{bmatrix} -x_m(k) \\ \frac{D+1}{k} \end{bmatrix} + b\right) \leqslant \angle\dot{\hat{x}} < \tan^{-1}\left(\frac{b_2}{b_1}\right) \tag{33}$$

This implies that for any straight line E with slope b_2/b_1, if $\hat{x} \in E \cap Q(k)$, then the vector $\dot{\hat{x}}$ points to the right of E for all $|w| \leqslant D$.

Copyright © 2001 John Wiley & Sons, Ltd.

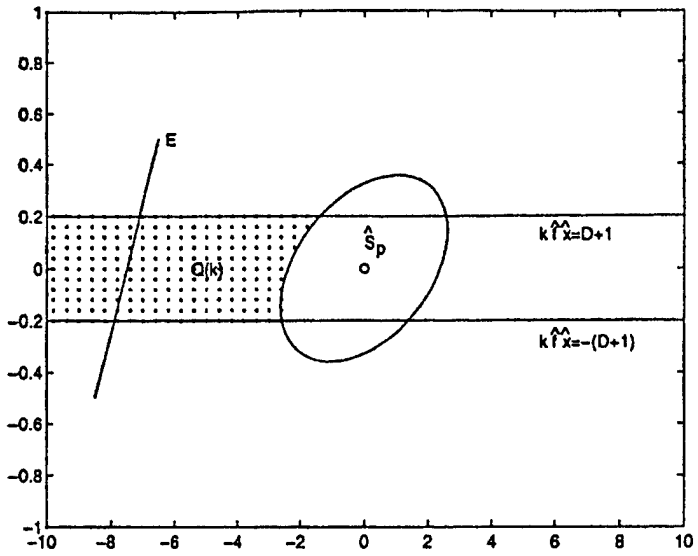


Figure 4. The vector field of system (30).

Proof. Between the lines $k\hat{f}\hat{x} = D + 1$ and $k\hat{f}\hat{x} = -(D + 1)$, $\text{sat}(k\hat{f}\hat{x} + w)$ takes values in $[-1, 1]$ and hence,

$$\dot{\hat{x}} \in \left\{ \begin{bmatrix} -a_1\hat{x}_2 \\ \hat{x}_1 + a_2\hat{x}_2 \end{bmatrix} + \lambda \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} : \lambda \in [-1, 1] \right\} \tag{34}$$

For $\hat{x} \in Q(k)$, if

$$\tan^{-1}\left(\frac{b_2}{b_1}\right) - \pi < \angle \hat{A}\hat{x} < \tan^{-1}\left(\frac{b_2}{b_1}\right)$$

then

$$\tan^{-1}\left(\frac{b_2}{b_1}\right) - \pi < \angle(\hat{A}\hat{x} + \lambda\hat{b}) < \tan^{-1}\left(\frac{b_2}{b_1}\right), \quad \forall \lambda \in [-1, 1] \tag{35}$$

Since $x_m(k)$ is increasing, we see from (32) that for all $k > K$,

$$-x_m(k) + a_2\frac{D+1}{k} < 0, \quad \frac{-x_m(k) + a_2\frac{D+1}{k}}{-a_1\frac{D+1}{k}} > \frac{b_2}{b_1}.$$

It follows that

$$\tan^{-1}\left(\frac{b_2}{b_1}\right) - \pi < \angle\left(\hat{A}\begin{bmatrix} -x_m(k) \\ \frac{D+1}{k} \end{bmatrix}\right) = \angle\begin{bmatrix} -a_1\frac{D+1}{k} \\ -x_m(k) + a_2\frac{D+1}{k} \end{bmatrix} < -\frac{\pi}{2} < \tan^{-1}\left(\frac{b_2}{b_1}\right)$$

For all $\hat{x} \in Q(k)$, we have $\hat{x}_1 \leqslant -x_m(k)$, $|\hat{x}_2| \leqslant (D + 1)/k$. So

$$\angle\left(\hat{A}\begin{bmatrix} -x_m(k) \\ \frac{D+1}{k} \end{bmatrix}\right) \leqslant \angle\hat{A}\hat{x} \leqslant 0 \Rightarrow \tan^{-1}\left(\frac{b_2}{b_1}\right) - \pi < \angle\hat{A}\hat{x} < \tan^{-1}\left(\frac{b_2}{b_1}\right)$$

Copyright © 2001 John Wiley & Sons, Ltd.

Hence by (35),

$$\tan^{-1}\left(\frac{b_2}{b_1}\right) - \pi < \angle \hat{\dot{x}} < \tan^{-1}\left(\frac{b_2}{b_1}\right)$$

for all $\hat{x} \in Q(k)$ and $|w| \leqslant D$. It can be further verified that

$$\min\{\angle \hat{A}\hat{x} + \lambda \hat{b} : \hat{x} \in Q(k), \lambda \in [-1, 1]\} \geqslant \angle\left(\hat{A}\begin{bmatrix} -x_m(k) \\ \frac{D+1}{k} \end{bmatrix} + b\right) > \tan^{-1}\left(\frac{b_2}{b_1}\right) - \pi$$

so (33) follows. □

This lemma means that any trajectory of (30) starting from inside of $Q(k)$ and to the right of E will remain to the right of E before it leaves $Q(k)$.

Based on Lemma 4, we can construct an invariant set $\hat{\mathscr{S}}_I(k) \subset \hat{\mathscr{S}}(k)$ and show that it is also a subset of $\hat{\mathscr{S}}_D(k)$. Moreover, it can be made arbitrarily close to $\hat{\mathscr{S}}(k)$.

Lemma 5

(a) if $k > K$ satisfies (31) and

$$\left(b_2 - \frac{a_2(D+1)}{k}\right) > \frac{b_1(D+1)}{kb_2} \tag{36}$$

then there exist unique p_1, $p_2 \in \hat{\mathscr{S}}(k)$ on the line $k\hat{f}\hat{x} = D + 1$ such that the trajectory of (30) starting at p_1 goes upward, returns to the line at p_2 and the line from p_2 to $-p_1$ has slope b_2/b_1 (see Figure 5, where the outer closed curve is $\partial\hat{\mathscr{S}}(k)$).

(b) Denote the region enclosed by the trajectories from $\pm p_1$ to $\pm p_2$, and the straight lines from $\pm p_2$ to $\mp p_1$ as $\hat{\mathscr{S}}_I(k)$. (In Figure 5, the region enclosed by the inner closed curve.) Then

$$\lim_{k \to \infty} d(\hat{\mathscr{S}}_I(k), \hat{\mathscr{S}}(k)) = 0$$

(c) $\hat{\mathscr{S}}_I(k)$ is an invariant set and $\hat{\mathscr{S}}_I(k) \subset \hat{\mathscr{S}}_D(k)$, i.e., it is inside the domain of attraction of $\hat{\mathscr{S}}_\infty(k)$.

Proof. Recall that $\partial\hat{\mathscr{S}}(k)$ is a closed trajectory of (30) with $w \equiv 0$. Denote the intersections of $\partial\hat{\mathscr{S}}(k)$ with $k\hat{f}\hat{x} = D + 1$ as s_1 and s_2 (see Figure 5). Let

$$p_0 = \begin{bmatrix} b_2 - \frac{a_2(D+1)}{k} \\ \frac{D+1}{k} \end{bmatrix}$$

then $\hat{\dot{x}}_2 = 0$ at p_0 and to the left (right) of p_0, $\hat{\dot{x}}_2 < 0 (> 0)$. Let p_1 be a point on $k\hat{f}\hat{x} = D + 1$ between p_0 and s_1, then a trajectory starting at p_1 goes upward and will return to $k\hat{f}\hat{x} = D + 1$ at some p_2 between p_0 and s_2. p_2 is uniquely determined by p_1. We then draw a straight line from p_2 with slope b_2/b_1. Let the intersection of the line with $k\hat{f}\hat{x} = -(D + 1)$ be p_3. Clearly, p_2 and p_3 depends on p_1 continuously. And the quantity

$$r(p_1) := \frac{(p_3 - (-s_1))_1}{(s_1 - p_1)_1}$$

Copyright © 2001 John Wiley & Sons, Ltd.

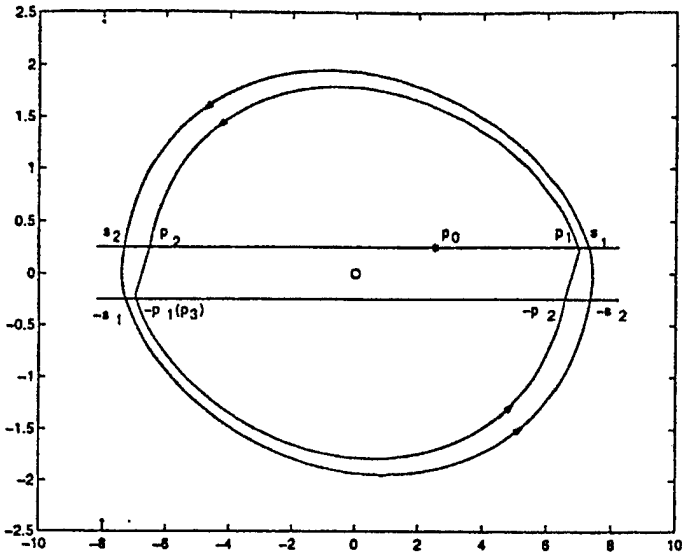


Figure 5. Illustration for Lemma 5.

also depends on p_1 continuously. If $p_1 = s_1$, then $p_2 = s_2$. Note that the trajectories above the line $k\hat{f}\hat{x} = D + 1$ are independent of w and hence are the same with those with $w = 0$. Since s_2 and $-s_1$ are on a trajectory of (30) with $w = 0$, so by Lemma 4, $-s_1$ must be to the right of the straight line with slope b_2/b_1 that passes s_2. This shows $-s_1$ is to the right of p_3 (with $p_1 = s_1$) and hence $\lim_{p_1 \to s_1} r(p_1) = -\infty$. If $p_1 = p_0$, then $p_2 = p_0$ and

$$p_3 = \begin{bmatrix} b_2 - \frac{a_2(D+1)}{k} - \frac{2(D+1)b_1}{kb_2} \\ -\frac{D+1}{k} \end{bmatrix}$$

So

$$r(p_1 = p_0) = \frac{(s_1)_1 + b_2 - \frac{a_2(D+1)}{k} - \frac{2(D+1)b_1}{kb_2}}{(s_1)_1 - (b_2 - \frac{a_2(D+1)}{k})}$$

And by condition (36), $r(p_1 = p_0) > 1$. Since $\lim_{p_1 \to s_1} r(p_1) = -\infty$, by the continuity of $r(p_1)$, there exists a p_1 between s_1 and p_0 such that $r(p_1) = 1$, i.e. $p_3 = -p_1$ and hence the line from p_2 to $-p_1$ has slope b_2/b_1. This shows the existence of (p_1, p_2) in (a). Suppose on the contrary that such pair (p_1, p_2) is not unique and there exists (p_1', p_2') with the same property, say, p_1' to the left of p_1 and p_2' to the right of p_2, by Lemma 1, $\|p_2 - p_2'\| > \|p_1 - p_1'\|$. But $\|(-p_1) - (-p_1')\| = \|p_2 - p_2'\|$ since the line from p_2 to $-p_1$ and that from p_2' to $-p_1'$ have the same slope. This is a contradiction.

(b) We see that $\dot{\hat{x}}_2 = 0$ at p_0, so by applying Lemma 1 with a shifting of the origin,

$$\frac{\|p_2 - s_2\|}{\|p_1 - s_1\|} > \frac{\|p_0 - s_2\|}{\|p_0 - s_1\|} > 1$$

Copyright © 2001 John Wiley & Sons, Ltd.

(refer to (12)). As $k \to \infty$, $s_2 + s_1 \to 0$, and

$$p_0 \to \begin{bmatrix} b_2 \\ 0 \end{bmatrix}$$

Since s_1 and s_2 are restricted to the null controllable region $\hat{\mathscr{C}}$, there exist some $K_1 > 0$, $\gamma > 0$, such that for all $k > K_1$,

$$\frac{\|p_0 - s_2\|}{\|p_0 - s_1\|} \geqslant 1 + \gamma \Rightarrow \frac{\|p_2 - s_2\|}{\|p_1 - s_1\|} > 1 + \gamma \tag{37}$$

From Figure 5, we see that

$$\|p_2 - s_2\| = \|p_1 - s_1\| + (-s_1 - s_2)_1 + (p_2 - (-p_1))_1$$

As $k \to \infty$, $2(D+1)/k \to 0$, so $(p_2 - (-p_1))_1 \to 0$. Since $s_1 + s_2 \to 0$, we have

$$\|p_2 - s_2\| - \|p_1 - s_1\| \to 0$$

From (37), $\|p_2 - s_2\| - \|p_1 - s_1\| > \gamma \|p_1 - s_1\|$. So we must have $\|p_1 - s_1\| \to 0$ and hence $\|p_2 - s_2\| \to 0$. Therefore, $\lim_{k\to\infty} d(\hat{\mathscr{S}}_I(k), \hat{\mathscr{S}}(k)) = 0$.

(c) First we show that $\hat{\mathscr{S}}_I(k)$ is an invariant set. Note that $\partial\hat{\mathscr{S}}_I(k)$ from p_1 to p_2 and that from $-p_1$ to $-p_2$ are trajectories of (30) under any $|w| \leqslant D$. At any point on the line from p_2 to $-p_1$, Lemma 4 says that $\dot{\hat{x}}$ directs to the right side of the line, i.e. no trajectory can cross the line from p_2 to $-p_1$ leftward, symmetrically, no trajectory can cross the line from $-p_2$ to p_1 rightward. These show that no trajectory can cross $\partial\hat{\mathscr{S}}_I(k)$ outward, thus $\hat{\mathscr{S}}_I k)$ is an invariant set. Since $\hat{\mathscr{S}}_p$ is also an invariant set and any trajectory that starts from inside of it will converge to $\hat{\mathscr{S}}_\infty(k)$, it suffices to show that any trajectory that starts from inside of $\hat{\mathscr{S}}_I(k)$ will enter $\hat{\mathscr{S}}_p$. We will do this by contradiction.

Suppose that there exist an $\hat{x}_0 \in \hat{\mathscr{S}}_I(k) \backslash \hat{\mathscr{S}}_p$ and a $w \in \mathscr{W}$ such that $\psi(t, \hat{x}_0, w) \in \hat{\mathscr{S}}_I(k) \backslash \hat{\mathscr{S}}_p$ for all $t > 0$, then there must be a point $\hat{x}^* \in \hat{\mathscr{S}}_I(k) \backslash \hat{\mathscr{S}}_p$ either

(1) $\lim_{t\to\infty} \psi(t, \hat{x}_0, w) = \hat{x}^*$; or
(2) there exists a sequence $t_1, t_2, \ldots, t_i, \ldots$ such that $\lim_{i\to\infty} \psi(t_i, \hat{x}_0, w) = \hat{x}^*$ and there is an $\varepsilon > 0$ such that for any $T > 0$, there exists $t > T$ satisfying $\|\psi(t, \hat{x}_0, w) - \hat{x}^*\| > \varepsilon$.

Item (1) implies that $\hat{x}^*$ can be made stationary by some $w \in \mathscr{W}$. This is impossible as we have shown that k has been chosen such that all the stationary points are inside $\hat{\mathscr{S}}_\infty(k)$. Item (2) implies that there is a closed trajectory with length greater than 2ε that passes through $\hat{x}^*$. There are two possibilities here: the closed trajectory encloses $\hat{\mathscr{S}}_p$ or it does not enclose $\hat{\mathscr{S}}_p$. We will show that none of the cases is possible.

Suppose that there is a closed trajectory that encloses $\hat{\mathscr{S}}_p$. Let q_1, q_2, q_3, q_4 be the four intersections of the closed trajectory with $k\hat{f}\hat{x} = \pm(D+1)$ as shown in Figure 6. By Lemma 1

$$\|p_2 - q_2\| > \|p_1 - q_1\|, \|q_4 - (-p_2)\| > \|q_3 - (-p_1)\|$$

and by Lemma 4,

$$\|q_3 - (-p_1)\| \geqslant \|p_2 - q_2\|, \quad \|p_1 - q_1\| \geqslant \|q_4 - (-p_2)\|$$

So we have

$$\|p_2 - q_2\| > \|p_1 - q_1\| \geqslant \|q_4 - (-p_2)\| > \|q_3 - (-p_1)\| \geqslant \|p_2 - q_2\|$$

Copyright © 2001 John Wiley & Sons, Ltd.

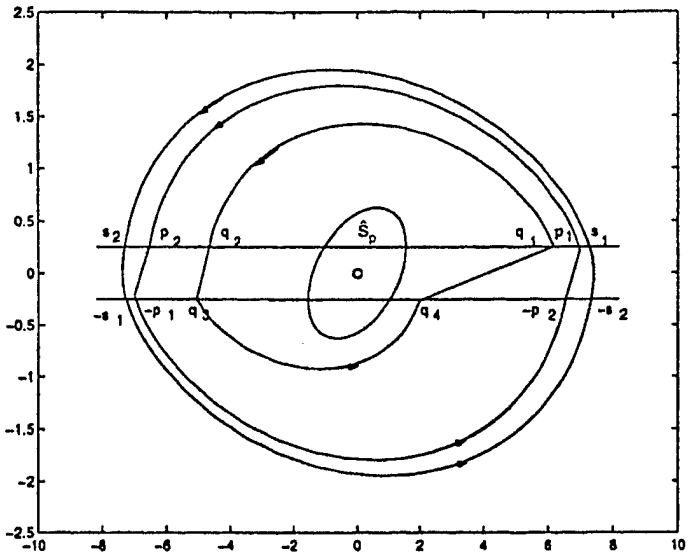


Figure 6. Illustration of the proof.

A contradiction. Therefore, there exists no closed trajectory that encloses $\hat{\mathscr{S}}_p$. We next exclude the other possibility.

Clearly, there can be no closed trajectory that is completely above $k\hat{f}\hat{x} = D + 1$ or below $k\hat{f}\hat{x} = -(D + 1)$. So if there is a closed trajectory, it must intersect $k\hat{f}\hat{x} = D + 1$ or $k\hat{f}\hat{x} = -(\bar{D} + 1)$ to the left (or to the right) of $\hat{\mathscr{S}}_p$ at least twice, or lies completely within $Q(k)$. We assume it is to the left of $\hat{\mathscr{S}}_p$. Since $k > K$ satisfies (36), so $x_m(k) > 0$ and $b_2 - a_2(D + 1)/k > 0$. Hence for all points on the line $k\hat{f}\hat{x} = D + 1$ to the left of $\hat{\mathscr{S}}_p$, $\dot{\hat{x}}_2 < 0$, so no closed trajectory lying between $\hat{\mathscr{S}}_I(k)$ and $\hat{\mathscr{S}}_p$ will cross this piece of straight line twice. On the line $k\hat{f}\hat{x} = -(D + 1)$ to the left of $\hat{\mathscr{S}}_p$, $\dot{\hat{x}}_1 > 0$. Since no trajectory in $Q(k)$ will cross a line that is parallel to the line from p_2 to $-p_1$ leftward, there will be no closed trajectory crossing the line $k\hat{f}\hat{x} = -(D + 1)$ to the left of $\hat{\mathscr{S}}_p$ twice. In view of Lemma 4, there exists no closed trajectory completely inside $Q(k)$. These show that there exist no closed trajectory that does not enclose $\hat{\mathscr{S}}_p$ either.

In conclusion, for every $\hat{x}_0 \in \hat{\mathscr{S}}_I(k)$, there must be a $T < \infty$ such that $\psi(T, \hat{x}_0, w) \in \hat{\mathscr{S}}_p$. And since $\hat{\mathscr{S}}_p$ is in the domain of attraction of $\hat{\mathscr{S}}_\infty(k)$, it follows that $\hat{x}_0 \in \hat{\mathscr{S}}_D(k)$ and hence $\hat{\mathscr{S}}_I(k) \subset \hat{\mathscr{S}}_D(k)$. □

The proof of Theorem 1 can be completed by invoking Lemmas 3 and 5. For clarity, we organize it as follows, including a constructive method to choose the parameters ε and k.

Proof of Theorem 1. Given $\chi_0 \subset \text{int}(\mathscr{C})$, χ_∞ such that $0 \in \text{int}\,\chi_\infty$ and $D > 0$, we need to choose ε and k such that $\chi_0 \subset \mathscr{S}_D(\varepsilon, k)$ and $\mathscr{S}_\infty(\varepsilon, k) \subset \chi_\infty$.

Step 1: Let $\varepsilon = 0$ and find k_0 such that $\chi_0 \subset \text{int}(\mathscr{S}(0, k_0))$. This is guaranteed by Proposition 3. Increase k_0, if necessary, such that $A + k_0 bf(\varepsilon)$ has distinct eigenvalues.

Step 2: Find $\varepsilon > 0$ such that $\chi_0 \subset \text{int}(\mathscr{S}(\varepsilon, k_0))$. This is guaranteed by Proposition 5 that $\mathscr{S}(\varepsilon, k_0)$ is continuous in $f(\varepsilon)$ and $f(\varepsilon)$ is continuous in ε.

Copyright © 2001 John Wiley & Sons, Ltd.

Step 3: Fix ε and perform state transformation $\hat{x} = Vx$ such that $(\hat{f}, \hat{A}, \hat{b})$ is in the form of (28) and (29). Also perform this transformation to the sets χ_0, χ_∞ to get $\hat{\chi}_0$, $\hat{\chi}_\infty$. We do not need to transform $\mathscr{S}(\varepsilon, k_0)$ to $\hat{\mathscr{S}}(k_0)$ but should remember that $\hat{\chi}_0 \subset \text{int}(\hat{\mathscr{S}}(k_0))$.

Step 4: Find $k > K$ satisfying (31) and (36) such that $\hat{\chi}_0 \subset \hat{\mathscr{S}}_I(k)$. Since $\hat{\chi}_0 \subset \text{int}(\hat{\mathscr{S}}(k_0))$, we have $\hat{\chi}_0 \subset \text{int}(\hat{\mathscr{S}}(k))$ for all $k > k_0$. So by Lemma 5, $\hat{\chi}_0 \subset \hat{\mathscr{S}}_I(k) \subset \hat{\mathscr{S}}_D(k)$ for some $k > 0$. It follows that $\chi_0 \subset \mathscr{S}_D(\varepsilon, k)$.

Step 5: Increase k, if necessary, so that $\mathscr{S}_\infty(\varepsilon, k) \subset \chi_\infty$. This is possible due to Lemma 3. □

6. PROOF OF THEOREM 1: HIGHER-ORDER SYSTEMS

As with the stabilization problem in Reference [16], where the disturbance is absent, the main idea in this section is first to bring those exponentially unstable states to a 'safe set' by using partial state feedback, then to switch to a full state feedback that steers all the states to a neighbourhood of the origin. The first step control is justified in the last section and the second step control is guaranteed by the property of the solution of the Riccati equation and Lemma 3, which allow the states that are not exponentially unstable to grow freely.

Without loss of generality, assume that the matrix pair (A, b) in system (1) is in the form of

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

where $A_1 \in \mathbf{R}^{2\times 2}$ is anti-stable and $A_2 \in \mathbf{R}^n$ is semi-stable. Assume that (A, b) is stabilizable. Denote the null controllable region of the subsystem

$$\dot{x}_1 = A_1 x_1 + b_1 \text{sat}(u)$$

as $\mathscr{C}_1$. Then the asymptotically null controllable region of (1) is $\mathscr{C}_a = \mathscr{C}_1 \times \mathbf{R}^n$. Given any $\gamma_1 \in (0, 1)$, and $\gamma_2 > 0$, denote

$$\Omega_1(\gamma_1) := \{\gamma_1 x_1 \in \mathbf{R}^2 : x_1 \in \bar{\mathscr{C}}_1\}, \qquad \Omega_2(\gamma_2) := \{x_2 \in \mathbf{R}^n : \|x_2\| \leqslant \gamma_2\} \tag{38}$$

For any compact subset χ_0 of $\mathscr{C}_a = \mathscr{C}_1 \times \mathbf{R}^n$, there exist γ_1 and γ_2 such that $\chi_0 \subset \Omega_1(\gamma_1) \times \Omega_2(\gamma_2)$. For this reason, we assume, without loss of generality, that $\chi_0 = \Omega_1(\gamma_1) \times \Omega_2(\gamma_2)$.

For $\varepsilon > 0$, let

$$P(\varepsilon) = \begin{bmatrix} P_1(\varepsilon) & P_2(\varepsilon) \\ P_2'(\varepsilon) & P_3(\varepsilon) \end{bmatrix} \in \mathbf{R}^{(2+n)\times(2+n)}$$

be the unique positive definite solution to the ARE

$$A'P + PA - Pbb'P + \varepsilon I = 0 \tag{39}$$

Clearly, as $\varepsilon \downarrow 0$, $P(\varepsilon)$ decreases. Hence $\lim_{\varepsilon \to 0} P(\varepsilon)$ exists.

Let P_1 be the unique positive definite solution to the ARE

$$A_1'P_1 + P_1A_1 - P_1b_1b_1'P_1 = 0$$

Then by the continuity property of the solution of the Riccati equation [18],

$$\lim_{\varepsilon \to 0} P(\varepsilon) = \begin{bmatrix} P_1 & 0 \\ 0 & 0 \end{bmatrix}$$

Copyright © 2001 John Wiley & Sons, Ltd.

Let $f(\varepsilon) := -b'P(\varepsilon)$. Let us first study the following closed-loop system

$$\dot{x} = Ax + b\,\mathrm{sat}(kf(\varepsilon)x + w) \tag{40}$$

Recall from Lemma 3, the invariant set $\mathscr{S}_p(\varepsilon)$ is a domain of attraction of the set $\mathscr{S}_\infty(\varepsilon, k)$.

Lemma 6

Denote

$$r_1(\varepsilon) := \frac{1}{\|P_1^{1/2}(\varepsilon)\|\|b'P^{1/2}(\varepsilon)\|}$$

$$r_2(\varepsilon) := \frac{-\|P_2(\varepsilon)\| + \sqrt{\|P_2(\varepsilon)\|^2 + 3\|P_1(\varepsilon)\|\,\|P_3(\varepsilon)\|}}{\|P_3(\varepsilon)\|} r_1(\varepsilon)$$

Then

$$D_1(\varepsilon) := \{x \in \mathbf{R}^{2+n} : \|x_1\| \leqslant r_1(\varepsilon),\ \|x_2\| \leqslant r_2(\varepsilon)\} \subset \mathscr{S}_p(\varepsilon)$$

Moreover, $\lim_{\varepsilon \to 0} r_2(\varepsilon) = \infty$, and $r_1(\varepsilon)$ increases with an upper bound as ε tends to zero.

Proof. Similar to the proof of Lemma 4.4.1 in Reference [16]. □

Proof of Theorem 1. Denote $\chi_{10} = \Omega(\gamma_1)$, then $\chi_{10} \subset \mathrm{int}(\mathscr{C}_1)$. Given $\varepsilon_0 > 0$, let $\chi_{1\infty} = \{x_1 \in \mathbf{R}^2 : \|x_1\| \leqslant r_1(\varepsilon_0)\}$. By the result of the second-order case, there exists a controller $u = f_1 x_1$ such that any trajectory of

$$\dot{x}_1 = A_1 x_1 + b_1\,\mathrm{sat}(f_1 x_1 + w) \tag{41}$$

that starts from within χ_{10} will converge to $\chi_{1\infty}$ at a finite time and stay there. Denote the trajectory of (41) that starts at x_{10} as $\psi_1(t, x_{10}, w)$ and define

$$T_M := \max_{x_{10} \in \partial\chi_{10}, w \in \mathscr{W}} \min\{t > 0 : \psi_1(t, x_{10}, w) \in \chi_{1\infty}\}$$

(An upper bound on T_M can be obtained by estimating the largest possible length of a trajectory $\psi_1(t, x_{10}, w)$, $x_{10} \in \chi_{10}$ before it enters $\chi_{1\infty}$ from Lemma 1 and (33), and the minimal $\|\dot{x}_1\|$ outside of $\chi_{1\infty}$. To apply (33), we can construct a region similar to $Q(k)$ by using $\chi_{1\infty}$ instead of $\mathscr{S}_p$.) Let

$$\gamma = \max_{t \in [0, T_M]} \|e^{A_2 t}\|\,\gamma_2 + \int_0^{T_M} \|e^{A_2(T_M - \tau)} b_2\|\,d\tau \tag{42}$$

then by Lemma 6, there exists an $\varepsilon < \varepsilon_0$ such that $r_1(\varepsilon) \geqslant r_1(\varepsilon_0)$, $r_2(\varepsilon) \geqslant \gamma$ and

$$D_1(\varepsilon) = \{x \in \mathbf{R}^{2+n} : \|x_1\| \leqslant r_1(\varepsilon),\ \|x_2\| \leqslant r_2(\varepsilon)\} \subset \mathscr{S}_p(\varepsilon)$$

lies in the domain of attraction of $\mathscr{S}_\infty(\varepsilon, k)$.

Choose k such that $\mathscr{S}_\infty(\varepsilon, k) \subset \chi_\infty$, and let the combined controller be

$$u(t) = \begin{cases} f_1 x_1(t), & x \notin \mathscr{S}_p(\varepsilon) \\ kf(\varepsilon)x(t), & x \in \mathscr{S}_p(\varepsilon) \end{cases} \tag{43}$$

and consider an initial state of the closed-loop system of (1) with (43), $x_0 \in \Omega_1(\gamma_1) \times \Omega_2(\gamma_2)$. If $x_0 \in \mathscr{S}_p(\varepsilon)$, then $x(t)$ will enter $\mathscr{S}_\infty(\varepsilon, k) \subset \chi_\infty$. If $x_0 \notin \mathscr{S}_p(\varepsilon)$, we conclude that $x(t)$ will enter $\mathscr{S}_p(\varepsilon)$ at some $T_1 \leqslant T_M$ under the control $u = f_1 x_1$. Observe that under this control, $x_1(t)$ goes along

Copyright © 2001 John Wiley & Sons, Ltd.

a trajectory of (41). If there is no switch, $x_1(t)$ will hit the ball $\chi_{1\infty}$ at $T_1 \leqslant T_M$ and at this instant $\|x_2(T_1)\| \leqslant \gamma \leqslant r_2(\varepsilon)$, so $x(T_1) \in D_1(\varepsilon)$. Thus we see that if there is no switch, $x(t)$ will be in $D_1(\varepsilon)$ at T_1. Since $D_1(\varepsilon) \subset \mathscr{S}_p(\varepsilon)$, $x(t)$ must have entered $\mathscr{S}_p(\varepsilon)$ at some earlier time $T \leqslant T_1 \leqslant T_M$. So we have the conclusion. With the switching control applied, once $x(t)$ enters the invariant set $\mathscr{S}_p(\varepsilon)$, it will converge to $\mathscr{S}_\infty(\varepsilon, k)$ and remain there. □

7. EXAMPLE

In this section, we will use an aircraft model to demonstrate the results obtained in this paper. Consider the longitudinal dynamics of the TRANS3 aircraft under certain flight condition [19],

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \dot{z}_4 \end{bmatrix} = \begin{bmatrix} 0 & 14.3877 & 0 & -31.5311 \\ -0.0012 & -0.4217 & 1.0000 & -0.0284 \\ 0.0002 & -0.3816 & -0.4658 & 0 \\ 0 & 0 & 1.0000 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} + \begin{bmatrix} 4.526 \\ -0.0337 \\ -1.4566 \\ 0 \end{bmatrix} v$$

The states z_1, z_2, z_3 and z_4 are the velocity, the angle of attack, the pitch rate and the Euler angle rotation of aircraft about the inertial y-axis, respectively. The control v is the elevator input, which is bounded by 10°, or 0.1745 rad. With a state transformation of the form $x = Tz$ and the input normalization such that the control is bounded by 1, we obtain

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \text{sat}(u + w)$$

where

$$A_1 = \begin{bmatrix} 0.0212 & 0.1670 \\ -0.1670 & 0.0212 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -0.4650 & 0.6247 \\ -0.6247 & -0.4650 \end{bmatrix}$$

and

$$b_1 = \begin{bmatrix} 8.2856 \\ -2.4303 \end{bmatrix}, \quad b_2 = \begin{bmatrix} 0.7584 \\ -1.8562 \end{bmatrix}$$

The system has two stable modes -0.4650 ± 0.6247i and two anti-stable ones, 0.0212 ± 0.1670i. Suppose that w is bounded by $|w| \leqslant D = 2$.

For the anti-stable x_1-subsystem, we take $\gamma_1 = 0.9$. With the technique in Section 5, we obtain a feedback $u = f_1 x_1$, where $f_1 = [-0.4335 \;\; 0.2952]$, such that $\Omega_1(\gamma_1)$ (as defined in (38)) is inside some invariant set $\mathscr{S}_I$. Moreover, for all initial $x_{10} \in \mathscr{S}_I$, under the control $u = f_1 x_1$, $x_1(t)$ will enter a ball $\chi_1 = \{x_1 \in \mathbf{R}^2 : \|x_1\|_2 \leqslant 29.8501\}$. In Figure 7, the outermost dotted closed curve is the boundary of the null controllable region $\partial\mathscr{C}_1$, the inner dash-dotted closed curve is $\partial\mathscr{S}_I$, the dashed closed curve is $\partial\Omega_1(\gamma_1)$, and the innermost solid closed curve is $\partial\chi_1$.

The x_2-subsystem is exponentially stable. Under the saturated control, it can be shown that for any initial value $x_{20} \in \mathbf{R}^2$, there exists a $T > 0$ such that $x_2(t)$ will enter a bounded ball at time T and remain there. The bounded ball is computed as

$$\chi_2 = \{x_2 \in \mathbf{R}^2 : \|x_2\|_2 \leqslant 4\}.$$

Copyright © 2001 John Wiley & Sons, Ltd.

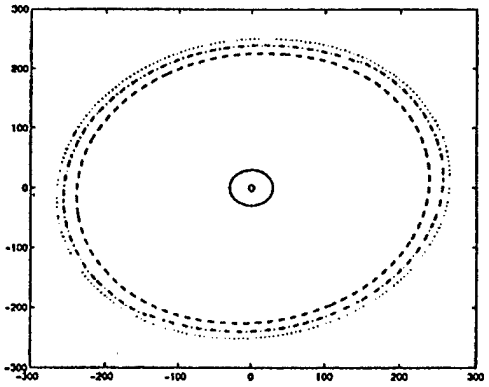


Figure 7. Design partial feedback $u = f_1 x_1$ such that $\Omega_1(\gamma_1) \subset \mathscr{S}_I$.

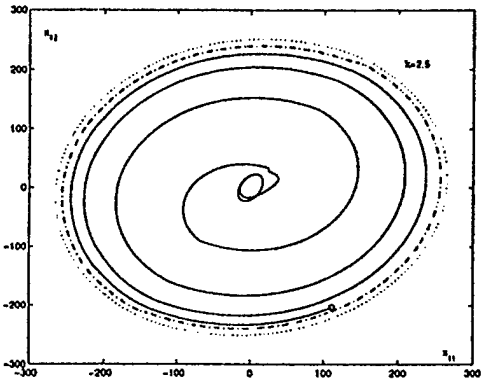


Figure 8. A trajectory of x_1 with $\varepsilon = 0.03$, $k = 2.5$.

We see that, for any $(x_{10}, x_{20}) \in \mathscr{S}_I \times \mathbf{R}^2$, under the partial feedback control $u = f_1 x_1$, the state (x_1, x_2) will enter the set $\chi_1 \times \chi_2$ in a finite time and remain there. The next step is to design a full state feedback to make the set $\chi_1 \times \chi_2$ inside the domain of attraction of an arbitrarily small set.

Choose $\varepsilon = 0.03$, we get

$$P(\varepsilon) = 0.001 \times \begin{bmatrix} 0.9671 & 0.0005 & -0.0686 & 0.0375 \\ 0.0005 & 0.9664 & 0.0345 & -0.0410 \\ -0.0686 & 0.0345 & 4.1915 & -0.7462 \\ 0.0375 & -0.0410 & -0.7462 & 11.3408 \end{bmatrix}$$

$f(\varepsilon) = 0.001 \times [-0.0729 \;\; 1.408 \;\; -36.4271 \;\; 33.6402]$, and $\mathscr{S}_p(\varepsilon) = \{x \in \mathbf{R}^4 : x'P(\varepsilon)x \leqslant 10.3561\}$. It can be verified that $\chi_1 \times \chi_2 \subset \mathscr{S}_p(\varepsilon)$. This implies that under the control $u = f_1 x_1$, the state will enter $\mathscr{S}_p(\varepsilon)$ at a finite time. If k is sufficiently large, then under the control $u = kf(\varepsilon)x$, $\mathscr{S}_p(\varepsilon)$ will be an invariant set. In this case, the switching controller (43) is well defined.

The final step is to choose k sufficiently large such that the state will converge to an arbitrarily small subset. We illustrate this point by simulation results for different values of k. In the

Copyright © 2001 John Wiley & Sons, Ltd.

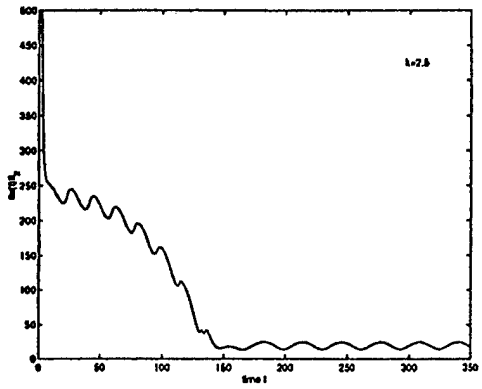


Figure 9. Time response of $\|x(t)\|_2$ with $\varepsilon = 0.03$, $k = 2.5$.

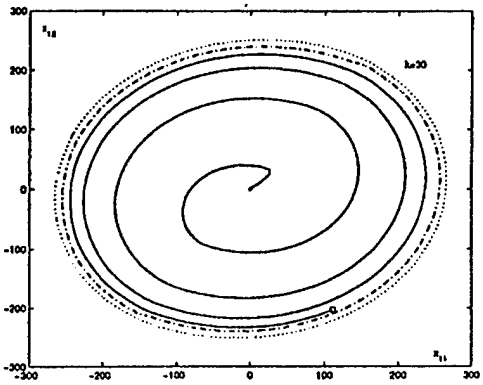


Figure 10. A trajectory of x_1 with $\varepsilon = 0.03$, $k = 30$.

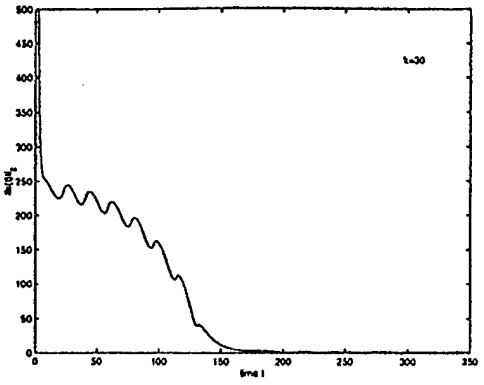


Figure 11. Time response of $\|x(t)\|_2$ with $\varepsilon = 0.03$, $k = 30$.

Copyright © 2001 John Wiley & Sons, Ltd.

simulation, we choose $w(t) = 2\sin(0.1t)$ and x_{10} to be a point very close to the boundary of $\mathscr{S}_I$, see the point market with 'o' in Figures 8 and 10. We also set $x_{20} = [1000\ 1000]^T$, which is very far away from the origin. When $k = 2.5$, the disturbance is not satisfactorily rejected (see Figure 8 for a trajectory of x_1 and Figure 9 for the time response of $\|x(t)\|_2$). When $k = 30$, the disturbance is rejected to a much higher level of accuracy (see Figures 10 and 11).

8. CONCLUSIONS

For linear exponentially unstable systems subject to actuator saturation and input additive disturbance, we have solved the problem of semi-global practical stabilization. We have assumed that the open-loop system has only two anti-stable modes and our results generalized the existing results on systems that do not have any exponentially unstable poles. Our analysis relies heavily on limit cycle theory and vector fields analysis of the exponentially unstable subsystem. It is not expected that our results can be further extended in a direct way to systems with more than two exponentially unstable open-loop poles.

APPENDIX A: PROOF OF LEMMA 1

Since at the intersection p', the trajectory goes downward, so $y_{11} < -a_2/k$. Using the fact that $fp = fp' = 1/k$ and $p' = e^{AT}p$, we have

$$[0\ k]e^{AT}\begin{bmatrix} x_{11} \\ \frac{1}{k} \end{bmatrix} = 1 \tag{A1}$$

$$[0\ k]e^{-AT}\begin{bmatrix} y_{11} \\ \frac{1}{k} \end{bmatrix} = 1 \tag{A2}$$

From (A1) and (A2), x_{11} and y_{11} can be expressed as functions of T. In other words, x_{11} and y_{11} are related to each other through the parameter T. Since the domain of valid x_{11} can be finite or infinite depending on the location of the eigenvalues of A, it is necessary to break the proof for different cases. We will see later that the relation among x_{11}, y_{11} and T are quite different for different cases.

Case 1:

$$A = \begin{bmatrix} 0 & -\lambda_1\lambda_2 \\ 1 & \lambda_1 + \lambda_2 \end{bmatrix}$$

has two different real eigenvalues $\lambda_1, \lambda_2 > 0$. Assume that $\lambda_1 > \lambda_2$.
Let

$$V = \begin{bmatrix} -\lambda_2 & -\lambda_1 \\ 1 & 1 \end{bmatrix}$$

then

$$e^{AT} = V\begin{bmatrix} e^{\lambda_1 T} & 0 \\ 0 & e^{\lambda_2 T} \end{bmatrix}V^{-1}$$

Copyright © 2001 John Wiley & Sons, Ltd.

From (A1) and (A2) we have

$$x_{11}(T) = \frac{1}{k}\frac{\lambda_1 - \lambda_2 + \lambda_2 e^{\lambda_2 T} - \lambda_1 e^{\lambda_1 T}}{e^{\lambda_1 T} - e^{\lambda_2 T}} \tag{A3}$$

$$y_{11}(T) = \frac{1}{k}\frac{\lambda_1 - \lambda_2 + \lambda_2 e^{-\lambda_2 T} - \lambda_1 e^{-\lambda_1 T}}{e^{-\lambda_1 T} - e^{-\lambda_2 T}} \tag{A4}$$

Due to the uniqueness of the trajectory, T is also uniquely determined by x_{11}. So, $x_{11} \leftrightarrow T$, $x_{11} \leftrightarrow y_{11}$, $y_{11} \leftrightarrow T$ are all one to one maps. From the above two equations, we know that $x_{11}(T)$ and $y_{11}(T)$ are analytic on $(0, \infty)$. It can be verified from (A3) that

$$\lim_{T \to 0} x_{11} = -\frac{\lambda_1 + \lambda_2}{k} = -\frac{a_2}{k}, \quad \lim_{T \to \infty} x_{11} = -\frac{\lambda_1}{k} = a_m$$

so we know the valid domain of x_{11} is $(-a_2/k, a_m)$. It can also be verified that $dx_{11}/dT > 0$, or $dT/dx_{11} > 0$.

Denote $g(T) := -dy_{11}/dx_{11}$, then

$$g(T) = \frac{\lambda_1 - \lambda_2 + \lambda_2 e^{\lambda_1 T} - \lambda_1 e^{\lambda_2 T}}{\lambda_1 - \lambda_2 + \lambda_2 e^{-\lambda_1 T} - \lambda_1 e^{-\lambda_2 T}}$$

It can be verified that $\lim_{T \to 0} g(T) = 1$ and

$$\frac{dg}{dT} = \frac{\lambda_1 \lambda_2 (e^{\lambda_1 T} - e^{\lambda_2 T})}{(\lambda_1 - \lambda_2 + \lambda_2 e^{-\lambda_1 T} - \lambda_1 e^{-\lambda_2 T})^2} h(T)$$

where $h(T) = (\lambda_1 - \lambda_2)(1 - e^{-(\lambda_1 + \lambda_2)T}) + (\lambda_1 + \lambda_2)(e^{-\lambda_1 T} - e^{-\lambda_2 T})$. Since $h(0) = 0$ and

$$\frac{dh}{dT} = (\lambda_1 + \lambda_2) e^{-(\lambda_1 + \lambda_2)} (\lambda_1 - \lambda_2 + \lambda_2 e^{\lambda_1 T} - \lambda_1 e^{\lambda_2 T}) > 0$$

for all $T > 0$, we have $h(T) > 0$, hence $dg/dT > 0$. This shows $g(T) > 1$ for all $T > 0$, i.e. $dy_{11}/dx_{11} < -1$. Since $dg/dT = dg(T)/dx_{11} \cdot dx_{11}/dT = -d^2 y_{11}/dx_{11}^2 \cdot dx_{11}/dT$, and $dg/dT > 0$, $dx_{11}/dT > 0$, it follows that

$$d^2 y_{11}/dx_{11}^2 < 0$$

Case 2:

$$A = \begin{bmatrix} 0 & -\lambda^2 \\ 1 & 2\lambda \end{bmatrix}$$

has two identical real eigenvalues $\lambda > 0$.
Let

$$V = \begin{bmatrix} -\lambda & 1 \\ 1 & 0 \end{bmatrix}$$

then

$$e^{AT} = V \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} V^{-1} e^{\lambda T}$$

In this case

$$x_{11}(T) = -\frac{1}{kT}(1 + \lambda T - e^{-\lambda T})$$

$$y_{11}(T) = \frac{1}{kT}(1 - \lambda T - e^{\lambda T})$$

Copyright © 2001 John Wiley & Sons, Ltd.

Similar to Case 1, it can be shown that

$$\lim_{T\to 0} x_{11} = -\frac{2\lambda}{k} = -\frac{a_2}{k}, \qquad \lim_{T\to\infty} x_{11} = -\frac{\lambda}{k} = a_m$$

so the valid domain of x_{11} is $(-a_2/k, a_m)$. It can also be verified that $\mathrm{d}x_{11}/\mathrm{d}T > 0$. Denote $g(T) := -\mathrm{d}y_{11}/\mathrm{d}x_{11}$, then

$$g(T) = \frac{1 - \mathrm{e}^{\lambda T} + \lambda T \mathrm{e}^{\lambda T}}{1 - \mathrm{e}^{-\lambda T} - \lambda T \mathrm{e}^{-\lambda T}}, \qquad \lim_{T\to\infty} g(T) = 1$$

and

$$\frac{\mathrm{d}g}{\mathrm{d}T} = \frac{\lambda^2 T}{(1 - \mathrm{e}^{-\lambda T} - \lambda T \mathrm{e}^{-\lambda T})^2} h(T)$$

where $h(T) = \mathrm{e}^{\lambda T} - \mathrm{e}^{-\lambda T} - 2\lambda T$. It can be shown that $h(T) > 0$, hence $\mathrm{d}g/\mathrm{d}T > 0$. The remaining part is similar to Case 1.

Case 3:

$$A = \begin{bmatrix} 0 & -(\alpha^2 + \beta^2) \\ 1 & 2\alpha \end{bmatrix}$$

has two complex eigenvalues $\alpha \pm \mathrm{j}\beta$, $\alpha, \beta > 0$.

Let

$$V = \begin{bmatrix} \beta & -\alpha \\ 0 & 1 \end{bmatrix}$$

then

$$\mathrm{e}^{AT} = V \begin{bmatrix} \cos\beta T & -\sin\beta T \\ \sin\beta T & \cos\beta T \end{bmatrix} V^{-1} \mathrm{e}^{\alpha T}$$

From (A1) and (A2) we have,

$$x_{11}(T) = \frac{1}{k \sin\beta T}(-\beta\cos\beta T - \alpha\sin\beta T + \beta \mathrm{e}^{-\alpha T})$$

$$y_{11}(T) = \frac{1}{k \sin\beta T}(\beta\cos\beta T - \alpha\sin\beta T - \beta \mathrm{e}^{\alpha T})$$

The valid domain of T is $(0, \pi/\beta)$, this can be obtained directly from the vector field and also from the above equations. Notice that

$$\lim_{T\to 0} x_{11}(T) = -\frac{2\alpha}{k} = -\frac{a_2}{k}, \qquad \lim_{T\to\pi/\beta} x_{11}(T) = \infty$$

So we have $a_m = \infty$ in this case.

Define $g(T)$ similarly as in Case 1, we have

$$g(T) = \frac{\beta + (\alpha\sin\beta T - \beta\cos\beta T)\mathrm{e}^{\alpha T}}{\beta - (\alpha\sin\beta T + \beta\cos\beta T)\mathrm{e}^{-\alpha T}}, \quad \lim_{T\to 0} g(T) = 1$$

and

$$\frac{\mathrm{d}g}{\mathrm{d}T} = \frac{(\alpha^2 + \beta^2)\sin\beta T}{(\beta - (\alpha\sin\beta T + \beta\cos\beta T)\mathrm{e}^{-\alpha T})^2} h(T)$$

Copyright © 2001 John Wiley & Sons, Ltd.

where $h(T) = \beta e^{\alpha T} - \beta e^{-\alpha T} - 2\alpha \sin \beta T$. It can be verified that $h(0) = 0$, $dh/dT > 0$, thus $dg/dT > 0$ for all $T \in (0, \pi/\beta)$. The remaining part is similar to case 1. □

APPENDIX B: PROOF OF LEMMA 2

Similar to the proof of Lemma 1, from (14), we can express x_{11} and y_{11} as functions of T, $x_{11}(T)$ and $y_{11}(T)$. Clearly these functions are analytic. Denote

$$g(T) := \frac{dy_{11}(T)/dT}{dx_{11}(T)/dT}$$

It suffices to show that $dx_{11}/dT > 0$, $g(T) > 1$, and $dg/dT > 0$. We need to break the proof into three different cases.

Case 1:

$$A = \begin{bmatrix} 0 & -\lambda_1\lambda_2 \\ 1 & -(\lambda_1 + \lambda_2) \end{bmatrix}$$

has two different real eigenvalues $-\lambda_1$, $-\lambda_2 < 0$. Assume that $\lambda_1 > \lambda_2$.
Let

$$V = \begin{bmatrix} \lambda_2 & \lambda_1 \\ 1 & 1 \end{bmatrix}$$

then

$$e^{AT} = V \begin{bmatrix} e^{-\lambda_1 T} & 0 \\ 0 & e^{-\lambda_2 T} \end{bmatrix} V^{-1}$$

From (14) and the fact that $kfp' = 1$, $kfp = -1$, we have

$$x_{11}(T) = \frac{1}{k} \frac{\lambda_2 - \lambda_1 + \lambda_2 e^{-\lambda_2 T} - \lambda_1 e^{-\lambda_1 T}}{e^{-\lambda_2 T} - e^{-\lambda_1 T}} \tag{B1}$$

$$y_{11}(T) = \frac{1}{k} \frac{\lambda_2 - \lambda_1 + \lambda_2 e^{\lambda_2 T} - \lambda_1 e^{\lambda_1 T}}{e^{\lambda_1 T} - e^{\lambda_2 T}} \tag{B2}$$

and

$$g(T) = \frac{\lambda_2 - \lambda_1 + \lambda_2 e^{-\lambda_1 T} - \lambda_1 e^{-\lambda_2 T}}{\lambda_2 - \lambda_1 + \lambda_2 e^{\lambda_1 T} - \lambda_1 e^{\lambda_2 T}}, \quad T \in (0, T_d)$$

By the definition of T_d, $y_{11}(T_d) = a_2/k = -(\lambda_1 + \lambda_2)/k$. It can be shown that as $T \to T_d$, $g(T) \to \infty$. Since $g(0) = 1$ and

$$\frac{dg}{dT} = \frac{2\lambda_1\lambda_2}{(\lambda_2 - \lambda_1 + \lambda_2 e^{\lambda_1 T} - \lambda_1 e^{\lambda_2 T})^2} \times \{(\lambda_1 + \lambda_2)[\mathrm{ch}(\lambda_1 T - \lambda_2 T) - 1] + (\lambda_2 - \lambda_1)[\mathrm{ch}(\lambda_2 T) - \mathrm{ch}(\lambda_1 T)]\} > 0$$

where $\mathrm{ch}(a) = (e^a + e^{-a})/2 \geqslant 1$ is monotonously increasing, we have that $g(T) > 1$ for all $T \in (0, T_d)$.

It can also be verified that $dx_{11}/dT > 0$. The remaining proof is similar to Appendix A.

Case 2:

$$A = \begin{bmatrix} 0 & -\lambda^2 \\ 1 & -2\lambda \end{bmatrix}$$

Copyright © 2001 John Wiley & Sons, Ltd.

has two identical real eigenvalues.
Let

$$V = \begin{bmatrix} \lambda & 1 \\ 1 & 0 \end{bmatrix}$$

then

$$e^{AT} = V \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} V^{-1} e^{-\lambda T}$$

In this case,

$$x_{11}(T) = -\frac{1}{kT}(1 - \lambda T + e^{\lambda T})$$

$$y_{11}(T) = -\frac{1}{kT}(1 + \lambda T + e^{-\lambda T})$$

and

$$g(T) = \frac{1 + \lambda T e^{-\lambda T} + e^{-\lambda T}}{1 - \lambda T e^{\lambda T} + e^{\lambda T}}$$

Since $g(0) = 1$ and

$$\frac{dg}{dT} = \frac{\lambda^2 T(2\lambda T + e^{\lambda T} - e^{-\lambda T})}{(1 - \lambda T e^{\lambda T} + e^{\lambda T})^2} > 0$$

we have $g(T) > 1$ for all $T \in (0, T_d)$. It can be verified that $dx_{11}/dT > 0$.

Case 3:

$$A = \begin{bmatrix} 0 & -(\alpha^2 + \beta^2) \\ 1 & -2\alpha \end{bmatrix}$$

has two complex eigenvalues $-\alpha \pm j\beta$, $\alpha, \beta > 0$.
Let

$$V = \begin{bmatrix} \beta & \alpha \\ 0 & 1 \end{bmatrix}$$

then

$$e^{AT} = V \begin{bmatrix} \cos \beta T & -\sin \beta T \\ \sin \beta T & \cos \beta T \end{bmatrix} V^{-1} e^{-\alpha T}$$

In this case,

$$x_{11}(T) = -\frac{1}{k \sin \beta T}(\beta \cos \beta T - \alpha \sin \beta T + \beta e^{\alpha T})$$

$$y_{11}(T) = -\frac{1}{k \sin \beta T}(\beta \cos \beta T + \alpha \sin \beta T + \beta e^{-\alpha T})$$

and

$$g(T) = \frac{\beta + (\beta \cos \beta T + \alpha \sin \beta T)e^{-\alpha T}}{\beta + (\beta \cos \beta T - \alpha \sin \beta T)e^{\alpha T}}, \quad T \in (0, T_d)$$

Copyright © 2001 John Wiley & Sons, Ltd.

Since $g(0) = 1$ and

$$\frac{dg}{dT} = \frac{\alpha^2 + \beta^2}{(\beta + (\beta \cos \beta T - \alpha \sin \beta T) e^{\alpha T})^2} [2\alpha \sin^2 \beta T + \beta(e^{\alpha T} - e^{-\alpha T}) \sin \beta T] > 0$$

we have $g(T) > 1$ for all $T \in (0, T_d)$. It can also be verified that $dx_{11}/dT > 0$.

For all the above three cases, Since $g(T) > 1$, i.e. $dy_{11}/dT > dx_{11}/dT$ for all T and $\lim_{T \to 0} x_{11}(T)/y_{11}(T) = 1$, we finally have $y_{11} > x_{11}$.

REFERENCES

1. Macki J, Strauss M. *Introduction to Optimal Control.* Springer: Berlin, 1982.
2. Schmitendorf WE, Barmish BR. Null controllability of linear systems with constrained controls. *SIAM Journal on Control and Optimization* 1980; **18**:327–345.
3. Sontag ED. An algebraic approach to bounded controllability of linear systems. *International Journal of Control* 1984; **39**:181–188.
4. Sontag ED, Sussmann HJ. Nonlinear output feedback design for linear systems with saturating controls. *Proceedings of the 29th IEEE Conference on Decision and Control*, 1990; 3414–3416.
5. Teel AR. Global stabilization and restricted tracking for multiple integrators with bounded controls. *System and Control Letters* 1992; **18**:165–171.
6. Sussmann HJ, Sontag ED, Yang Y. A general result on the stabilization of linear systems using bounded controls. *IEEE Transactions on Automatic Control.* 1994; **39**:2411–2425.
7. Teel AR. A nonlinear small gain theorem for the analysis of control systems with saturation. *IEEE Transactions on Automatic Control* 1996; **41**:1256–1270.
8. Megretski A. L_2 BIBO output feedback stabilization with saturated control. *Proceedings of the 13th IFAC World Congress*, vol. D, 1996; pp. 435–440.
9. Suarez R, Alvarez-Ramirez J, Solis-Daun J. Linear systems with bounded inputs: global stabilization with eigenvalue placement. *International Journal of Robust and Nonlinear Control* 1997; **7**:835–845.
10. Teel AR. Linear systems with input nonlinearities: global stabilization by scheduling a family of H_∞-type controllers. *International Journal of Robust and Nonlinear Control* 1995; **5**:399–441.
11. Fuller AT. In-the-large stability of relay and saturating control systems with linear controller. *International Journal of Control* 1969; **10**:457–480.
12. Sussmann HJ, Yang Y. On the stabilizability of multiple integrators by means of bounded feedback controls. *Proceedings of the 30th IEEE Conference on Decision and Control*, 1991; 70–72.
13. Lin Z, Saberi A. Semi-global exponential stabilization of linear systems subject to 'input saturation' via linear feedbacks. *Systems and Control Letters* 1993; **21**:225–239.
14. Lin Z. *Low Gain Feedback*, Lecture Notes in Control and Information Sciences, vol. 240. Springer: London, 1998.
15. Saberi A, Lin Z, Teel AR. Control of linear systems with saturating actuators. *IEEE Transactions Automatic Control* 1996; **41**:368–378.
16. Hu T, Lin Z. Control Systems with Actuator Saturation: Analysis and Design, Birkhäuser, Boston, 2001.
17. Wredenhagen GF, Belanger PR. Piecewise-linear LQ control for systems with input constraints. *Automatica* 1994; **30**:403–416.
18. Willems JC. Least squares stationary optimal control and the algebraic Riccati equations. *IEEE Transactions on Automatic Control* 1971; **16**:621–634.
19. Junkins J, Valasek J, Ward D. *Report on ONR UCAV Modeling Effort*, Department of Aerospace Engineering, Texas A&M University, 1999.

Copyright © 2001 John Wiley & Sons, Ltd.

Publication 9

A Complete Stability Analysis of Planar Discrete-Time Linear Systems Under Saturation

Tingshu Hu and Zongli Lin, *Senior Member, IEEE*

***Abstract*—A complete stability analysis is performed on a planar discrete-time system of the form $x(k+1) = \text{sat}(Ax(k))$, where A is a Schur stable matrix and sat is the saturation function. Necessary and sufficient conditions for the system to be globally asymptotically stable are given. In the process of establishing these conditions, the behaviors of the trajectories are examined in detail.**

***Index Terms*—Limit trajectories, neural networks, saturation, stability.**

I. Introduction

DYNAMICAL systems with saturation nonlinearities arise frequently in neural networks, analogue circuits and control systems (see, for example, [9], [5], [2], [6] and the references therein). In this paper, we consider systems of the following form:

$$x(k+1) = \text{sat}(Ax(k)), \qquad x \in \mathbf{R}^n \tag{1}$$

where sat $\mathbf{R}^n \to \mathbf{R}^n$ is the standard saturation function. With a slight abuse of notation, we use the same symbol to denote both the vector saturation function and the scalar saturation function, i.e., if $v \in \mathbf{R}^n$, then sat $(v) = [\text{sat}(v_1), \text{sat}(v_2), \ldots, \text{sat}(v_n)]^T$ and

$$\text{sat}(v_i) = \begin{cases} -1, & \text{if } v_i < -1 \\ v_i, & \text{if } -1 \le v_i \le 1 \\ 1, & \text{if } v_i > 1. \end{cases} \tag{2}$$

Systems of the form (1) and their continuous-time counterparts mainly arise in neural networks and in digital filters.

As with any dynamical system, stability of these systems is of primary concern and has been heavily studied in the literature for a long period of time (see, for example, [1], [7]–[11] and the references therein). As seen in the literature, the stability analysis of such systems are highly nontrivial even for the planar case. For the continuous-time counterpart of (1), only until recently have the necessary and sufficient conditions for global asymptotic stability (GAS) been established for the planar case [4]. For the planar discrete-time system of the form (1), to the best of our knowledge, no necessary and sufficient conditions have been known, although various sufficient conditions are available [9], [11]. This paper attempts to carry out a complete stability analysis of planar systems of the form (1). In particular, necessary and sufficient conditions for the system to be GAS will be identified. In the process of establishing these conditions, the behaviors of the trajectories are examined in detail.

This work is motivated by our recent result [4] on the planar continuous-time system

$$\dot{x} = \text{sat}(Ax), \qquad x \in \mathbf{R}^2. \tag{3}$$

However, the two systems (1) and (3) behave quite differently even though they have a similar description. First of all, (3) operates on the entire plane while (1) operates only on the unit square. The trajectories of (3) do not intersect each other but the connected trajectory of (1) [by connecting $x(k)$ and $x(k+1)$] can intersect itself. The limit trajectories of (3) must be periodic but a limit trajectory of (1) need not be. Finally, it is known that in the stability analysis for nonlinear systems, many more tools are available for continuous-time systems than for discrete-time systems.

We will start our investigation of the planar system (1) by characterizing some general properties of its limit trajectories. An important feature is that a nontrivial limit trajectory can only intersect two opposite pair of boundaries of the unit square and it cannot have intersections with both of the neighboring boundaries. This result turns our attention to a simpler system which has only one saturated state

$$x(k+1) = \begin{bmatrix} a_{11}x_1(k) + a_{12}x_2(k) \\ \text{sat}\,(a_{21}x_1(k) + a_{22}x_2(k)) \end{bmatrix}. \tag{4}$$

For this simpler system, we will establish a relation between the present intersection of a trajectory with the lines $x_2 = \pm 1$ and the next one in terms of a set of points on the line $x_2 = 1$. The relation is discontinuous but piecewise linear. The set of points are the places where the discontinuity occurs. Some attractive properties about these points and the relation between the next intersection and the present one are revealed. These properties help us to establish the condition for the system (4) to be GAS and to characterize an interval on the line $x_2 = 1$ from which the trajectories of (4) will converge to the origin. This in turn leads to our final result on the necessary and sufficient conditions for the GAS of a planar system of the form (1).

This paper is organized as follows. In Section II, we give the necessary and sufficient conditions for the GAS of the planar system in the form of (1). An example is also given to help interpret these conditions. These conditions are established in Sections III–V. In the process of establishing these conditions, the intricate properties of the system trajectories are also revealed. In particular, Section III reveals some general properties of the

Manuscript received July 6, 2000; revised December 23, 2000. This work was supported in part by the US Office of Naval Research Young Investigator Program under Grant N00014-99-1-0670. This paper was recommended by Associate Editor P. K. Rajan.

The authors are with the Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903 USA.

Publisher Item Identifier S 1057-7122(01)04290-8.

1057–7122/01$10.00 © 2001 IEEE

possible limit trajectories of the system which help us to exclude the existence of limit trajectories under the condition of ... main theorem and focus our attention to the simpler system ...th one saturated state. Section IV investigates system (4) and gives a necessary and sufficient condition for the system to be GAS. Section V proves the main result of this paper. Finally, a brief concluding remark is made in Section VI.

II. Main Results

Consider the following system:

$$x(k+1) = \text{sat}(Ax(k)), \qquad x \in \mathbf{R}^2 \tag{5}$$

where $A = \left[\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right]$ and sat: $\mathbf{R}^2 \to \mathbf{R}^2$ is the saturation function, i.e., if $v \in \mathbf{R}^2$, then $\text{sat}(v) = \left[\begin{smallmatrix} \text{sat}(v_1) \\ \text{sat}(v_2) \end{smallmatrix}\right]$ and $\text{sat}(\cdot)$ is as defined by (2).

Given an initial state $x(0) = x_0$, denote the trajectory of the system (5) that passes through x_0 at $k = 0$ as $\psi(k, x_0)$. In this paper, we only consider the positive trajectories. Hence, throughout the paper, $k \geq 0$.

Definition 2.1: The system (5) is said to be stable at its equilibrium $x_e = 0$ if, for any $\epsilon > 0$, there exists a $\delta > 0$ such that, $\|\psi(k, x_0)\| \leq \epsilon$, for all $k \geq 0$ and $\|x_0\| \leq \delta$. It is said to be globally asymptotically stable (GAS) if $x_e = 0$ is a stable equilibrium and $\lim_{k\to\infty} \psi(k, x_0) = 0$ for all $x_0 \in \mathbf{R}^2$. Also, it is said to be locally asymptotically stable if it is stable and $\lim_{k\to\infty} \psi(k, x_0) = 0$ for all x_0 in a neighborhood U_0 of $= 0$.

The system is GAS only if it is locally asymptotically stable, which is equivalent to that A has eigenvalues inside the unit circle. In this case, A is said to be Schur stable, or simply stable. In this paper, we assume that A is stable. Denote the closed unit square as $\mathbf{S}$ and its boundary as $\partial\mathbf{S}$. It is easy to see that no matter where $x(0)$ is, we always have $x(1) \in \mathbf{S}$. Hence, the global asymptotic stability is equivalent to $\lim_{k\to\infty} \psi(k, x_0) = 0$ for all $x_0 \in \mathbf{S}$. The main result of this paper is presented as follows:

Theorem 2.1: The system (5) is globally asymptotically stable if and only if A is stable and there exists no $x_0 \in \partial\mathbf{S}$ and $N > 0$ such that $\psi(N, x_0) = \pm x_0$ and $\psi(k, x_0) = A^k x_0 \in \mathbf{S}$ for all $k < N$.

If $\psi(k, x_0) = A^k x_0 \in \mathbf{S}$ for all $k < N$, then $\psi(N, x_0) = \text{sat}(A^N x_0)$. Hence, this theorem can be interpreted as follows. Assume that A is stable, then the system (5) is GAS if and only if none of the following statements are true.

1) There exist $N \geq 1$ and $d_1, d_2 \geq 0$ such that

$$A^N \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \pm \begin{bmatrix} 1+d_1 \\ 1+d_2 \end{bmatrix} \quad \text{and} \quad A^k \begin{bmatrix} 1 \\ 1 \end{bmatrix} \in \mathbf{S} \qquad \forall k < N.$$

2) There exist $N \geq 1$ and $d_1, d_2 \geq 0$ such that

$$A^N \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \pm \begin{bmatrix} -1-d_1 \\ 1+d_2 \end{bmatrix}$$

and

$$A^k \begin{bmatrix} -1 \\ 1 \end{bmatrix} \in \mathbf{S} \qquad \forall k < N.$$

3) There exist $N \geq 1$, $d_2 > 0$ and $x_1 \in (-1, 1)$ such that

$$A^N \begin{bmatrix} x_1 \\ 1 \end{bmatrix} = \pm \begin{bmatrix} x_1 \\ 1+d_2 \end{bmatrix}$$

and

$$A^k \begin{bmatrix} x_1 \\ 1 \end{bmatrix} \in \mathbf{S} \qquad \forall k < N.$$

4) There exist $N \geq 1$, $d_1 > 0$ and $x_2 \in (-1, 1)$ such that

$$A^N \begin{bmatrix} 1 \\ x_2 \end{bmatrix} = \pm \begin{bmatrix} 1+d_1 \\ x_2 \end{bmatrix}$$

and

$$A^k \begin{bmatrix} 1 \\ x_2 \end{bmatrix} \in \mathbf{S} \qquad \forall k < N.$$

Each of the above conditions implies that there is a simple periodic trajectory that starts at some x_0 with period N or $2N$. The trajectory stays inside $\mathbf{S}$ as that of the corresponding linear system for the first $N-1$ steps, and when the linear trajectory goes out of $\mathbf{S}$ at step N, the saturation function makes $\psi(N, x_0) = \text{sat}(A^N x_0)$ return exactly at x_0 or $-x_0$. These conditions can be verified. Since A is stable, there exists an integer N_0 such that $A^k x_0 \in \mathbf{S}$ for all $k > N_0$ and all $x_0 \in \partial\mathbf{S}$. Hence, it suffices to check the four conditions only for $N < N_0$.

Conditions 1) and 2) are very easy to check. As to 3) or 4), for each N, at most two $x_1's$ can be solved from

$$A^N \begin{bmatrix} x_1 \\ 1 \end{bmatrix} = \pm \begin{bmatrix} x_1 \\ 1+d_2 \end{bmatrix}.$$

To see this, denote the elements of A^N as $(A^N)_{ij}$, $i, j = 1, 2$. Then from 3), we have

$$(A^N)_{11} x_1 + (A^N)_{12} = \pm x_1. \tag{6}$$

If $(A^N)_{11} \neq \pm 1$, then there are two x_1's that satisfy (6). If $(A^N)_{11} = \pm 1$, we must have $(A^N)_{12} \neq 0$. Otherwise A^N would have an eigenvalue ± 1, which is impossible since A is stable. In this case, (6) has only one solution. It remains to check if $d_2 > 0$ and

$$A^k \begin{bmatrix} x_1 \\ 1 \end{bmatrix} \in \mathbf{S} \qquad \forall k < N.$$

In the process of proving Theorem 2.1, we will develop a more efficient method to check the conditions.

Example 2.1: Consider (5) with

$$A = \begin{bmatrix} 1.5840 & -1.3990 \\ 3.9702 & -2.9038 \end{bmatrix}.$$

The following results are presented with accuracy up to four decimal digits. There are two points on $\partial\mathbf{S}$ that satisfy condition 3), one with

$$x_1 = \frac{(A^3)_{12}}{1 - (A^3)_{11}} = 0.7308$$

and the other with

$$x_1 = \frac{(A^2)_{12}}{-1 - (A^2)_{11}} = 0.9208.$$

But there are four periodic trajectories as listed

1) $\begin{bmatrix}0.7308\\1.0000\end{bmatrix}$ $\begin{bmatrix}-0.2414\\-0.0023\end{bmatrix}$ $\begin{bmatrix}-0.3791\\-0.9516\end{bmatrix}$ $\begin{bmatrix}0.7308\\1.0000\end{bmatrix}$

2) $\begin{bmatrix}0.9028\\1.0000\end{bmatrix}$ $\begin{bmatrix}0.0310\\0.6804\end{bmatrix}$ $\begin{bmatrix}-0.9028\\-1.0000\end{bmatrix}$ $\begin{bmatrix}-0.0310\\-0.6804\end{bmatrix}$ $\begin{bmatrix}0.9028\\1.0000\end{bmatrix}$

3) $\begin{bmatrix}0.7424\\1.0000\end{bmatrix}$ $\begin{bmatrix}-0.2230\\0.0438\end{bmatrix}$ $\begin{bmatrix}-0.4145\\-1.0000\end{bmatrix}$ $\begin{bmatrix}0.7424\\1.0000\end{bmatrix}$

4) $\begin{bmatrix}1.0000\\1.0000\end{bmatrix}$ $\begin{bmatrix}0.1850\\1.0000\end{bmatrix}$ $\begin{bmatrix}-1.0000\\-1.0000\end{bmatrix}$ $\begin{bmatrix}-0.1850\\-1.0000\end{bmatrix}$ $\begin{bmatrix}1.0000\\1.0000\end{bmatrix}$.

In the third periodic trajectory, the first coordinate $x_1 = 0.7424$ of the initial state is computed from

$$x_1 = \frac{a_{12} - a_{11}(A^2)_{12}}{a_{11}(A^2)_{11} - 1} = 0.7424.$$

It should be noted that 4) is the only stable periodic trajectory.

As we can see from the example, there are other kinds of periodic trajectories than what are inferred by the conditions 1)–4), e.g., trajectories 3) and 4). There may also be trajectories that neither are periodic nor converge to the origin. We will prove in the subsequent sections that if none of the conditions 1)–4) is true, then there exist no nonconvergent trajectory of any kind.

III. Limit Trajectories

To prove that (5) is GAS, we need to show that the only limit point of any trajectory is the origin. It is known that A being stable alone is not sufficient to guarantee the GAS of the system. Actually, it is well-known [9] that the system may have stationary points other than the origin; there may be periodic trajectories and even trajectories that neither are periodic, nor converge to a stationary point. In this section, we are going to characterize some general properties of the nonconvergent trajectories. These properties will facilitate us to exclude the existence of such nonconvergent trajectories under the condition of Theorem 2.1.

Since every trajectory is bounded by the unit square, there exists a set of points such that the trajectory will go arbitrarily close to them infinitely many times.

Definition 3.1: For a given $x_0 \in \mathbf{R}^2$, a point $x^* \in \mathbf{R}^2$ is called a (positive) limit point of the trajectory $\psi(k, x_0)$ if there exists a subsequence of $\psi(k, x_0)$, $\psi(k_i, x_0)$, $i = 1, 2, \ldots$, such that $\lim_{i\to\infty} \psi(k_i, x_0) = x^*$. The set of all such limit points is called the limit set of the trajectory. We denote this limit set as $\Gamma(x_0)$.

Since the function $\text{sat}(Ax)$ is continuous in x, if a trajectory $\psi(k, x_0)$ returns arbitrarily close to $x \in \Gamma(x_0)$, it will also return arbitrarily close to $\text{sat}(Ax)$. We state this property in the following lemma.

Lemma 3.1: If $y_0 \in \Gamma(x_0)$, then $\psi(k, y_0) \in \Gamma(x_0)$ for all $k \geq 0$. Given any $\epsilon > 0$ (arbitrarily small) and any integer $N > 0$ (arbitrarily large), there exists an integer $K_0 > 0$ such that

$$|\psi(k + K_0, x_0) - \psi(k, y_0)|_\infty < \epsilon \qquad \forall k \leq N.$$

Because of Lemma 3.1, for $y_0 \in \Gamma(x_0)$, $\psi(k, y_0)$ is called a limit trajectory of $\psi(k, x_0)$. It is periodic if and only if $\Gamma(x_0)$ has finite number of elements.

The following notation is defined for simplicity. Denote

$$L_h = \left\{\begin{bmatrix}x_1\\1\end{bmatrix} : x_1 \in (-1, 1)\right\}$$

$$L_v = \left\{\begin{bmatrix}1\\x_2\end{bmatrix} : x_2 \in (-1, 1)\right\}.$$

We see that L_h and $-L_h$ are the two horizontal sides of S, and L_v and $-L_v$ are the two vertical sides of S. Notice that they do not include the four vertices of the unit square. Also, denote $v_1 = \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$, $v_2 = \left[\begin{smallmatrix}-1\\1\end{smallmatrix}\right]$ as the two upper vertices of the square.

Let y_0 be a limit point of some trajectory and for simplicity, let $y_k = \psi(k, y_0)$. Denote $Y = \{\pm y_k : k \geq 0\}$ and $AY = \{\pm Ay_k : k \geq 0\}$. Clearly, Y must have an intersection with the boundary of the unit square. If $Y \cap L_h$ is not empty, define

$$\gamma_1 = \inf\left\{x_1 : \begin{bmatrix}x_1\\1\end{bmatrix} \in Y \cap (L_h \cup \{v_1, v_2\})\right\}$$

and

$$\gamma_2 = \sup\left\{x_1 : \begin{bmatrix}x_1\\1\end{bmatrix} \in Y \cap (L_h \cup \{v_1, v_2\})\right\}.$$

If $Y \cap L_v$ is not empty, define

$$\gamma_3 = \sup\left\{x_2 : \begin{bmatrix}1\\x_2\end{bmatrix} \in Y \cap (L_v \cup \{v_1, -v_2\})\right\}$$

and

$$\gamma_4 = \inf\left\{x_2 : \begin{bmatrix}1\\x_2\end{bmatrix} \in Y \cap (L_v \cup \{v_1, -v_2\})\right\}.$$

The following proposition shows that a limit trajectory can only intersect one opposite pair of the sides of the unit square, not both of the neighboring pair. This result will reduce our problem to a much simpler one.

Proposition 3.1: Let y_0 be a limit point of some trajectory.

1) If $y_0 \in L_h$, then $\psi(k, y_0)$ will not touch L_v or $-L_v$ for all $k \geq 0$. Moreover, $\psi(k, y_0)$ will stay inside the strip
$$\left\{\begin{bmatrix}x_1\\x_2\end{bmatrix} : |x_1| \leq \max\{|\gamma_1|, |\gamma_2|\}\right\}.$$

2) If $y_0 \in L_v$, then $\psi(k, y_0)$ will not touch L_h or $-L_h$ and will stay inside the strip
$$\left\{\begin{bmatrix}x_1\\x_2\end{bmatrix} : |x_2| \leq \max\{|\gamma_3|, |\gamma_4|\}\right\}.$$

3) The set Y cannot include both v_1 and v_2.

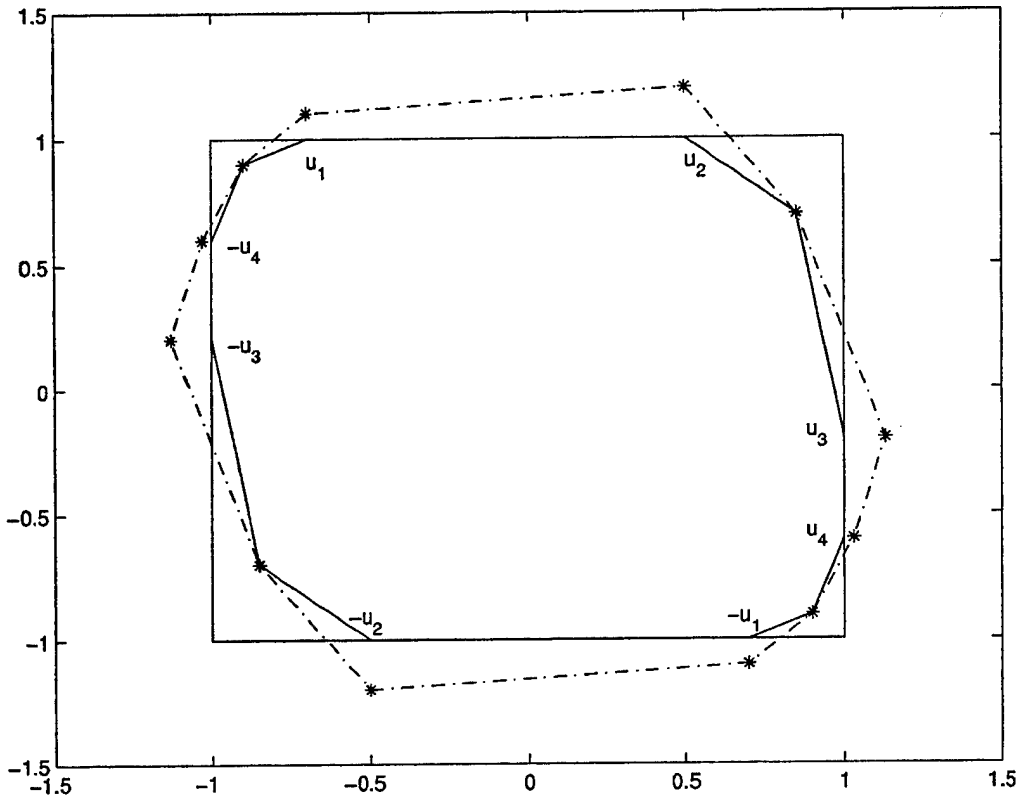


Fig. 1. Illustration for the proof of Proposition 3.1.

Proof: The proof is built up on a simple geometric fact. Let X be a set in $\mathbf{R}^2$ and let AX be the image of X under the linear map $x \rightarrow Ax$. Then, the area of AX equals to the area of X times $|\det(A)|$.

1)We first assume that Y contains a finite number of elements, i.e., $\psi(N, y_0) = y_0$ for some N. Suppose on the contrary that the trajectory will touch L_v or $-L_v$ at some step. The main idea of the proof is to show that the area of the convex hull of AY is no less than that of Y, which contradicts the fact that $|\det(A)| < 1$.

Since Y contains points on both L_h and L_v, γ_i, $i = 1, 2, 3, 4$, are all defined.

If y_k is in the interior of the unit square, then $y_k = Ay_{k-1}$; if $y_k \in L_h$, then $y_k = \text{sat}(Ay_{k-1})$ and

$$Ay_{k-1} = \begin{bmatrix} x_1 \\ 1+d \end{bmatrix}$$

for some $|x_1| < 1$ and $d \geq 0$ [note that $y_0 = y_N = \text{sat}(Ay_{N-1})$]; if $y_k \in L_v$, then

$$Ay_{k-1} = \begin{bmatrix} 1+d \\ x_2 \end{bmatrix}$$

for some $|x_2| < 1$ and $d \geq 0$. If $y_k = v_1$ (or v_2), then $y_k = \text{sat}(Ay_{k-1})$ and

$$Ay_{k-1} = \begin{bmatrix} 1+d_1 \\ 1+d_2 \end{bmatrix} \quad \left(\text{or } \begin{bmatrix} -1-d_1 \\ 1+d_2 \end{bmatrix}\right)$$

for some $d_1, d_2 \geq 0$. Hence, AY contains all the elements of Y which are in the interior of S, and for those y_k on the boundary of S, if $y_k \in L_h$, there is a point in AY that is just above y_k (on the same vertical line) and if $y_k \in L_v$, then there is a point in AY that is just to the right of y_k (on the same horizontal line).

Denote the areas of the convex hulls of Y and AY as $\mathcal{A}(Y)$ and $\mathcal{A}(AY)$, respectively. Also, let

$$u_1 = \begin{bmatrix} \gamma_1 \\ 1 \end{bmatrix} \quad u_2 = \begin{bmatrix} \gamma_2 \\ 1 \end{bmatrix}$$
$$u_3 = \begin{bmatrix} 1 \\ \gamma_3 \end{bmatrix} \quad u_4 = \begin{bmatrix} 1 \\ \gamma_4 \end{bmatrix}$$

as shown in Fig. 1. In the figure, the points marked with "$*$" belong to AY, the polygon with dash–dotted boundary is the convex hull of AY and the polygon with vertices $\pm u_i$, $i = 1, 2, 3, 4$, and some points in the interior of S is the convex hull of Y. Since there is at least one point in Y that is to the left of u_1, one to the right of u_2, one above u_3 and one below u_4, the convex hull of Y is a subset of the convex hull of AY. (This may not be true if u_1 is the leftmost point in Y, or if u_2 is the rightmost). It follows that $\mathcal{A}(AY) \geq \mathcal{A}(Y)$. This is a contradiction since $\mathcal{A}(AY) = |\det(A)|\mathcal{A}(Y)$ and $|\det(A)| < 1$.

If, on the contrary, Y has a point outside of the strip

$$\left\{ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} : |x_1| < \max\{|\gamma_1|, |\gamma_2|\} \right\}$$

then, there will be a point in Y that is to the left of u_1 (or on the same vertical line with u_1), and a point to the right of u_2 (or on the same horizontal line with u_2). In this case, we also have $\mathcal{A}(AY) \geq \mathcal{A}(Y)$, which is a contradiction.

Now we extend the result to the case that Y has infinite many elements. Also suppose on the contrary that the trajectory will touch L_v, $-L_v$ or go outside of the strip at some step. By

Lemma 3.1, for any $\varepsilon > 0$ and any integer $N \geq 1$, there exists a $K_0 > 0$ such that

$$|\psi(k + K_0, x_0) - \psi(k, y_0)|_\infty < \varepsilon$$

for all $k \leq N$ and in particular

$$|\psi(K_0, x_0) - y_0|_\infty < \varepsilon.$$

So, the trajectory $\psi(k + K_0, x_0)$, $k \geq 0$, will also touch (or almost touch) L_v, $-L_v$, or go outside of the strip. Since y_0 is a limit point of $\psi(k + K_0, x_0)$, there exists a $K_1 > 0$ such that

$$|\psi(K_1 + K_0, x_0) - y_0|_\infty < \varepsilon.$$

Define

$$Z(\varepsilon) = \{\psi(k + K_0, x_0): 0 \leq k \leq K_1\}$$

and

$$AZ(\varepsilon) = \{A\psi(k + K_0, x_0): 0 \leq k \leq K_1\}.$$

Using similar arguments as in the finite element case, we can show that

$$|\det(A)| = \frac{\mathcal{A}(AZ(\varepsilon))}{\mathcal{A}(Z(\varepsilon))} \geq 1 - O(\varepsilon).$$

Letting $\varepsilon \to 0$, we obtain $|\det(A)| \geq 1$, which is a contradiction.

2) Similar to 1).

3) If, on the contrary, Y contains both v_1 and v_2, then the convex hull of Y is $\mathbf{S}$. Also, AY contains a point

$$Ay_j = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad x_1 \leq -1 \quad x_2 \geq 1$$

and a point

$$Ay_k = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad x_1 \geq 1 \quad x_2 \geq 1$$

hence, the convex hull of AY contains $\mathbf{S}$. This also leads to $\mathcal{A}(AY) \geq \mathcal{A}(Y)$, a contradiction. □

IV. Systems with One Saturated State

Now, we are clear from Proposition 3.1 that if there is any limit trajectory, it can intersect only one opposite pair of the sides of the unit square, either $(L_h, -L_h)$, or $(L_v, -L_v)$, not both of them. So, we only need to investigate the possibility that a limit trajectory only intersects $\pm L_h$. The other possibility that it only intersects $\pm L_v$ is similar. For this reason, we consider the following system:

$$x(k+1) = \begin{bmatrix} a_{11}x_1(k) + a_{12}x_2(k) \\ \text{sat}\,(a_{21}x_1(k) + a_{22}x_2(k)) \end{bmatrix} := \text{sat}_2\,(Ax(k)). \tag{7}$$

Assume that $A = \left[\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right]$ is stable. If $a_{21} = 0$ or $a_{12} = 0$, it is easy to see that both systems (5) and (7) are GAS and none of the conditions 1)–4) following Theorem 2.1 can be true. So we assume in the following that a_{21}, $a_{12} \neq 0$.

The terms *GAS, limit point* and *limit trajectory* for (5) are extended to (7) in a natural way.

For a given initial state $x(0) = x_0$, denote the trajectory of the system (7) as $\psi_2(k, x_0)$. Denote the line $x_2 = 1$ as L_h^e, the line $x_2 = -1$ as $-L_h^e$ and the region between these two lines (including $\pm L_h^e$) as $\mathbf{S}^e$. We will show later that (7) has nontrivial limit trajectory in $\mathbf{S}$ if and only if (5) has nontrivial limit trajectory that intersects $\pm L_h$. In the sequel, when we say "limit trajectory," we mean a limit trajectory other than the trivial one at the origin.

In this section, we study the GAS of the system (7) and will also determine a subset in L_h^e which is free of limit points. Our investigation will be based on the study of the linear system

$$x(k+1) = Ax(k). \tag{8}$$

For a stable continuous-time linear planar system, if a trajectory stays in $\mathbf{S}^e$ for a whole cycle [$\angle x(t)$ increases or decreases by 2π], then $x(t)$ will be in $\mathbf{S}^e$ for all $t > 0$. But, for the discrete-time linear planar system (8), a trajectory might go out of $\mathbf{S}^e$ after staying within $\mathbf{S}^e$ for several cycles. In the continuous-time case, the trajectories never intersect but in the discrete-time case, the connected trajectory [by connecting $x(k)$ and $x(k+1)$] can intersect itself. These facts make the analysis much more complicated than the continuous-time system as discussed in [1], [3], [4] and [10].

A simple one or two point periodic trajectory can be formed if $A\left[\begin{smallmatrix} x_1 \\ 1 \end{smallmatrix}\right] = \pm\left[\begin{smallmatrix} x_1 \\ 1+d \end{smallmatrix}\right]$ for some $d > 0$. An N or $2N$ point periodic trajectory will be formed if $A^N\left[\begin{smallmatrix} x_1 \\ 1 \end{smallmatrix}\right] = \pm\left[\begin{smallmatrix} x_1 \\ 1+d \end{smallmatrix}\right]$, $d > 0$ and $A^k\left[\begin{smallmatrix} x_1 \\ 1 \end{smallmatrix}\right] \in \mathbf{S}^e$, for all $k < N$.

Proposition 4.1: The system (7) is GAS if and only if A is stable and the following statement is not true for any $x_1 \in \mathbf{R}$: There exist an integer $N > 0$ and a real number $d > 0$ such that

$$A^N \begin{bmatrix} x_1 \\ 1 \end{bmatrix} = \pm \begin{bmatrix} x_1 \\ 1+d \end{bmatrix}$$

and

$$A^k \begin{bmatrix} x_1 \\ 1 \end{bmatrix} \in \mathbf{S}^e \qquad \forall k < N. \tag{9}$$

Let $\alpha_s = \min\{|x_1|: x_1 \text{ satisfies (9)}\}$, then no limit trajectory can exist completely within the strip

$$\left\{ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} : |x_1| < \alpha_s \right\}.$$

Remark: If (9) is true for some x_1, then there will be a stationary point or periodic trajectory such as $x_0, Ax_0, \ldots, A^{N-1}x_0$, $\text{sat}_2(A^N x_0) = x_0, Ax_0, \ldots$. There may also exist other kind of limit trajectories. Proposition 4.1 says that if there is no simple periodic trajectory as inferred by (9), there will be no limit trajectory of any kind (except the one at the origin).

To prove Proposition 4.1, we need to establish the relation between the next intersection of a trajectory with $\pm L_h^e$ and the present one.

For $x_0 \in L_h^e$, suppose that $\psi_2(k, x_0)$ will intersect $\pm L_h^e$ at $k = k_i, i = 1, 2, \ldots$, with $0 < k_1 < k_2 < \ldots$. Since the trajectory can be switched to $-\psi_2(k, x_0)$ at any k without changing

its convergence property, we assume for simplicity that all the intersections $\psi_2(k_i, x_0)$ are in L_h^e (If not so, just multiply it with -1). Denote

$$x_0^1 = \psi_2(k_1, x_0) \quad x_0^2 = \psi_2(k_2, x_0) \quad \ldots.$$

We call x_0, x_0^1 and x_0^2 the first, the second and the third intersections, respectively. We also call x_0 and x_0^1 the present and the next intersections.

Clearly, x_0^1 is uniquely determined by x_0. We also see that the relation $x_0 \to x_0^1$ is a map from L_h^e to itself. To study the GAS of the system (7), it suffices to characterize the relation between x_0 and x_0^1. Through this relation, we can show that if (9) is not true for any x_1, then for every $x_0 \in L_h^e$, the intersections x_0^1, x_0^2, ... will move closer and closer toward an interval, and all the trajectories starting from this interval will not touch the lines $\pm L_h^e$ and will converge to the origin.

Let $x_0 \in L_h^e$. The next intersection of $\psi_2(k, x_0)$ with L_h^e occurs at step k_1 if

$$\left|[0 \quad 1]A^{k_1}x_0\right| \geq 1$$

and

$$\left|[0 \quad 1]A^{k}x_0\right| < 1 \qquad \forall k < k_1.$$

The next intersection is $x_0^1 = \psi_2(k_1, x_0) = \text{sat}_2(A^{k_1}x_0)$ [or $-\text{sat}_2(A^{k_1}x_0)$]. Since for different $x_0 \in L_h^e$, the number of steps for the trajectories to return to $\pm L_h^e$, i.e., the number k_1 as defined above, is different, we see that the relation between x_0 and x_0^1 must be discontinuous.

We will first determine an interval on L_h^e from which a trajectory will not intersect $\pm L_h^e$ again (no x_0^1) and will converge to the origin.

Since A is stable, there exists a positive definite matrix P such that

$$A^TPA - P < 0.$$

Define the Lyapunov function as

$$V(x) := x^TPx$$

then for every $x \in \mathbf{R}^2$, $V(A^kx) < V(x)$ for all $k > 1$.

Given a real number $\rho > 0$, denote the Lyapunov level set as

$$\mathcal{E}(\rho) := \{x \in \mathbf{R}^2 : x^TPx \leq \rho\}.$$

Let ρ_c be such that $\mathcal{E}(\rho_c) \subset \mathbf{S}^e$ and $\mathcal{E}(\rho_c)$ just touches $\pm L_h^e$. In this case, $\mathcal{E}(\rho_c)$ has only one intersection with L_h^e. Let this intersection be

$$p_c = \begin{bmatrix} \alpha_c \\ 1 \end{bmatrix}.$$

If $x_0 = p_c$, then the linear trajectory A^kx_0 will be inside $\mathcal{E}(\rho_c) \subset \mathbf{S}^e$. Hence, $\psi_2(k, x_0) = \text{sat}_2(A^kx_0) = A^kx_0$ for all $k > 0$ and will converge to the origin. Since

$$A^TPA - P < 0$$

there exists an interval around p_c in L_h^e, of nonzero length, such that for every x_0 in this interval, $\psi_2(k, x_0) = A^kx_0$, $k \geq 1$, will never touch $\pm L_h^e$ and will converge to the origin.

Here we will use a simple way to denote a line segment. Given two points $p_1, p_2 \in \mathbf{R}^2$, denote

$$[p_1, p_2] := \{\lambda p_1 + (1-\lambda)p_2 : 0 \leq \lambda \leq 1\}$$

and similarly

$$(p_1, p_2] = [p_1, p_2]\backslash\{p_1\}$$
$$[p_1, p_2) = [p_1, p_2]\backslash\{p_2\}$$
$$(p_1, p_2) = [p_1, p_2]\backslash\{p_1, p_2\}.$$

Define

$$\alpha_0 := \min\left\{\alpha < \alpha_c : A^k\begin{bmatrix}\alpha \\ 1\end{bmatrix} \in \mathbf{S}^e \qquad \forall k \geq 0\right\}$$

and

$$\beta_0 := \max\left\{\beta > \alpha_c : A^k\begin{bmatrix}\beta \\ 1\end{bmatrix} \in \mathbf{S}^e \qquad \forall k \geq 0\right\}.$$

Since $a_{21} \neq 0$, the line $AL_h^e := \{Ax : x \in L_h^e\}$ has intersections with both L_h^e and $-L_h^e$, so there exist points on both sides of p_c which will be mapped out of $\mathbf{S}^e$ under A. Hence α_0 and β_0 are finite numbers. Now, let

$$p_0 = \begin{bmatrix}\alpha_0 \\ 1\end{bmatrix} \quad q_0 = \begin{bmatrix}\beta_0 \\ 1\end{bmatrix}$$

then, for all $x_0 \in [p_0, q_0]$, $\psi_2(k, x_0) = A^kx_0$ will converge to the origin. Because of the extremal nature in the definition of α_0 and β_0, we must have $A^kp_0 \in \pm L_h^e$ for some k, otherwise α_0 would not be the minimum of the set. Therefore, define

$$m_0 := \min\{k : A^kp_0 \in \pm L_h^e\}$$

and similarly

$$n_0 := \min\{k : A^kq_0 \in \pm L_h^e\}.$$

If $m_0 > 1$, then by definition

$$\left|[0 \quad 1]A^kp_0\right| < 1 \qquad \forall k < m_0$$

and by continuity, there exists a neighborhood of p_0 such that $|[0\ 1]A^kx_0| < 1, \forall k < m_0$ for all x_0 in this neighborhood. Because of this, we can define

$$\alpha_1 := \min\left\{\alpha < \alpha_0 : A^k\begin{bmatrix}\alpha \\ 1\end{bmatrix} \in \mathbf{S}^e \qquad \forall k < m_0\right\}.$$

If $n_0 > 1$, then define

$$\beta_1 := \max\left\{\beta > \beta_0 : A^k\begin{bmatrix}\beta \\ 1\end{bmatrix} \in \mathbf{S}^e, \qquad \forall k < n_0\right\}.$$

Also, because $a_{21} \neq 0$, α_1 and β_1 are finite. Let

$$p_1 = \begin{bmatrix}\alpha_1 \\ 1\end{bmatrix} \quad q_1 = \begin{bmatrix}\beta_1 \\ 1\end{bmatrix}.$$

It follows from the extremal nature in the definition of α_1 and β_1 that there exists a $k < m_0$ such that $A^kp_1 \in \pm L_h^e$, so we define

$$m_1 := \min\{k : A^kp_1 \in \pm L_h^e\}$$

and similarly

$$n_1 := \min\{k: A^k q_1 \in \pm L_h^e\}.$$

For simplicity, we denote $\begin{bmatrix} -\infty \\ 1 \end{bmatrix}$ as p_∞ and $\begin{bmatrix} \infty \\ 1 \end{bmatrix}$ as q_∞. Implied by the definitions are the following:

$$p_1 \in (p_\infty, p_0) \quad q_1 \in (q_0, q_\infty),$$

and

$$m_1 < m_0 \quad n_1 < n_0.$$

Inductively, if $m_{i-1} > 1$, then define

$$\alpha_i := \min\left\{\alpha < \alpha_{i-1}: A^k \begin{bmatrix} \alpha \\ 1 \end{bmatrix} \in \mathrm{S}^e \qquad \forall\, k < m_{i-1}\right\}$$

and if $n_{j-1} > 1$, define

$$\beta_j := \max\left\{\beta > \beta_{j-1}: A^k \begin{bmatrix} \beta \\ 1 \end{bmatrix} \in \mathrm{S}^e \qquad \forall\, k < n_{j-1}\right\}.$$

Let

$$p_i = \begin{bmatrix} \alpha_i \\ 1 \end{bmatrix} \quad q_j = \begin{bmatrix} \beta_j \\ 1 \end{bmatrix}$$

and

$$m_i := \min\{k: A^k p_i \in \pm L_h^e\} \quad n_j := \min\{k: A^k q_j \in \pm L_h^e\}.$$

Then

$$p_i \in (p_\infty, p_{i-1}) \quad q_j \in (q_{j-1}, q_\infty),$$
$$m_i < m_{i-1} \quad n_j < n_{j-1}.$$

The induction procedure ends if both $m_i = 1$ and $n_j = 1$. We denote the maximum index of i as I and the maximum index of j as J. As an immediate consequence of these definitions, we have

$$\alpha_I < \alpha_{I-1} < \cdots < \alpha_1 < \alpha_0 < \alpha_c < \beta_0 < \beta_1 < \cdots < \beta_{J-1} < \beta_J$$

and

$$1 = m_I < m_{I-1} < \cdots < m_1 < m_0$$
$$1 = n_J < n_{J-1} < \cdots < n_1 < n_0.$$

We claim that this set of p_i, $i = 0, 1, 2, \ldots, I$, and q_j, $j = 0, 1, \ldots, J$, forms exactly the set of points where discontinuity occurs on the relation between the next intersection of a trajectory with $\pm L_h^e$ and the present one.

Lemma 4.1:

1) If $x_0 \in [p_0, q_0]$, then $\psi_2(k, x_0) = A^k x_0$ will be inside S^e for all $k > 0$ and will converge to the origin.
2) If $x_0 \in (p_{i+1}, p_i]$, then the next intersection of $\psi_2(k, x_0)$ with $\pm L_h^e$ is $\psi_2(m_i, x_0) = \mathrm{sat}_2(A^{m_i} x_0)$; If $x_0 \in (p_\infty, p_I]$, then the next intersection is $\psi_2(1, x_0) = \mathrm{sat}_2(Ax_0)$. Moreover, $A^{m_i} p_i \in \pm L_h^e$ and $A^{m_i} x_0 \notin \mathrm{S}^e$ for all $x_0 \in (p_\infty, p_i)$.
3) If $x_0 \in [q_j, q_{j+1})$, then the next intersection of $\psi_2(k, x_0)$ with $\pm L_h^e$ is $\psi_2(n_j, x_0) = \mathrm{sat}_2(A^{n_j} x_0)$; If $x_0 \in [q_J, q_\infty)$, then the next intersection is $\psi_2(1, x_0) = \mathrm{sat}_2(Ax_0)$. Moreover, $A^{n_j} q_j \in \pm L_h^e$ and $A^{n_j} x_0 \notin \mathrm{S}^e$ for all $x_0 \in (q_j, q_\infty)$.
4) $|[0\ 1] A^k p_{i+1}| \le 1$ for all $k < m_i$ and $|[0\ 1] A^{m_i} p_{i+1}| > 1$; $|[0\ 1] A^k q_{j+1}| \le 1$ for all $k < n_j$ and $|[0\ 1] A^{n_j} q_{j+1}| > 1$.

Proof: 1) This is a direct consequence of the definition of p_0 and q_0.

2) From the definition of m_i, $|[0\ 1] A^k p_i| < 1$ for all $k < m_i$ and $|[0\ 1] A^{m_i} p_i| = 1$. Since $A^{m_i} L_h^e$ is a straight line and $A^{m_i} p_c$ is in the interior of S^e, we have $|[0\ 1] A^{m_i} x_0| > 1$ for all $x_0 \in (p_\infty, p_i)$ (Note that p_c is to the right of p_i).

On the other hand, since

$$\alpha_{i+1} = \min\left\{\alpha < \alpha_i: A^k \begin{bmatrix} \alpha \\ 1 \end{bmatrix} \in \mathrm{S}^e \qquad \forall\, k < m_i\right\}$$

we have $|[0\ 1] A^k p_{i+1}| \le 1$ for all $k < m_i$. Also since $A^k p_c$ is in the interior of S^e, we have $|[0\ 1] A^k x_0| < 1$ for all $k < m_i$ and for all $x_0 \in (p_{i+1}, p_c)$.

Combining the above arguments we have, for all $x_0 \in (p_{i+1}, p_i]$, $|[0\ 1] A^{m_i} x_0| \ge 1$ and $|[0\ 1] A^k x_0| < 1$ for all $k < m_i$. This means that the next intersection with $\pm L_h^e$ is $\psi_2(m_i, x_0) = \mathrm{sat}_2(A^{m_i} x_0)$.

3) Similar to 2).

4) This is contained in the proof of 2). □

It is obvious that $\mathrm{sat}_2(A^{m_i} x)$ is a continuous function of x. Lemma 4.1 2) implies that for all $x_0 \in (p_{i+1}, p_i]$, the second coordinate of $\mathrm{sat}_2(A^{m_i} x_0)$, $[0\ 1]\mathrm{sat}_2(A^{m_i} x_0)$, is the constant 1 or -1, while the first coordinate remains linear on x_0. Similarly, for all $x_0 \in [q_j, q_{j+1})$, the second coordinate of $\mathrm{sat}_2(A^{n_j} x_0)$ is the constant 1 or -1 and the first coordinate is linear on x_0. Same relation holds for $x_0 \in (p_\infty, p_I]$ and $x_0 \in [q_I, q_\infty)$.

We will provide an easy way to compute p_i and q_j after revealing more properties about this set of points. In fact, the following properties will lead directly to the proof of Proposition 4.1. For $x_0 \in L_h^e$, the next intersection of $\psi_2(k, x_0)$ with $\pm L_h^e$ can be on L_h^e or on $-L_h^e$. For simplicity, we will assume that the next intersection is on L_h^e, otherwise we can replace the state $x(k)$ at the intersection with $-x(k)$, noting that we can multiply the state at any step with -1 without changing the convergence rate of a trajectory. Hence in the following, when we say that $x \in [p_i, q_j]$, we mean $x \in \pm[p_i, q_j]$; and when we say that $x \in \pm L_h^e$ is to the left (or right) of p_i, it could also be that x is to the right (or left) of $-p_i$.

Denote $p_i^1 = \psi_2(m_i, p_i) = A^{m_i} p_i$ and $q_j^1 = \psi_2(n_j, q_j) = A^{n_j} q_j$. We see that p_i^1 is the second intersection of $\psi_2(k, p_i)$ with $\pm L_h^e$ (the first one is p_i), and q_j^1 is the second intersection of $\psi_2(k, q_j)$ with $\pm L_h^e$.

Lemma 4.2:

1) $p_0^1, q_0^1 \in [p_0, q_0]$;
2) If $p_i^1 \in (q_{j-1}, q_j]$, then $m_{i-1} = m_i + n_{j-1}$; if $q_j^1 \in [p_i, p_{i-1})$, then $n_{j-1} = n_j + m_{i-1}$;
3) For $i \ge 1$, $p_i^1 \in (q_0, q_\infty)$ and for $j \ge 1$, $q_j^1 \in (p_\infty, p_0)$;
4) $p_i^1 \in (p_{i-1}^1, q_\infty)$, and $q_j^1 \in (p_\infty, q_{j-1}^1)$;
5) For $i, j \ge 1$, p_i^1 and q_j^1 cannot be both in $[p_i, q_j]$, nor both outside of $[p_i, q_j]$, i.e., there must be one of them inside $[p_i, q_j]$ and the other one outside of the interval.

Proof: First, we give a simple property arising from the Lyapunov function $V(x)$. Since $V(x)$ is a convex function and p_c takes the minimum value from all $x \in L_h^e$, we have, if both

s_1 and s_2 are to the left of p_c and s_1 is to the left of s_2, then $V(s_1) > V(s_2)$; if both s_1 and s_2 are to the right of p_c and s_1 to the right of s_2, then $V(s_1) > V(s_2)$.

1)Clearly, p_0^1 cannot be to the left of p_0, otherwise we would have $V(p_0^1) = V(A^{n_0}p_0) > V(p_0)$. Suppose on the contrary that $p_0^1 \in (q_0, q_\infty)$, then by Lemma 4.1 3), we would have $|[0\ 1]A^{n_0}A^{m_0}p_0| > 1$. A contradiction to the definition of p_0. Similar argument holds for q_0.

2) From Lemma 4.1 4), the first time $A^k p_i$ goes out of S^e is at $k = m_{i-1}$. And by Lemma 4.1 3) and 4), for $x_0 \in (q_{j-1}, q_j]$, the first time $A^k x_0$ goes out of S^e is $k = n_{j-1}$. Since $p_i^1 = A^{m_i}p_i \in (q_{j-1}, q_j]$ and $A^k p_i \in S^e$ for all $k < m_i$, we have $m_{i-1} = m_i + n_{j-1}$. Similarly, for q_j, we have $n_{j-1} = n_j + m_{i-1}$.

3) Similar to 1), p_i^1 cannot be to the left of p_i. Suppose on the contrary, that $p_i^1 \in [p_0, q_0]$, then $A^k p_i$ never goes out of S^e. This is a contradiction since $A^{n_0}p_i \notin S^e$. Also, suppose on the contrary that $p_i^1 \in [p_{i-l}, p_{i-l-1})$, $l \geq 0$, then similar to the argument in 2), we would have $m_{i-1} = m_i + m_{i-l-1}$ and hence $m_{i-1} > m_{i-l-1}$. This is a contradiction since m_i decreases as i is increased. So, we must have $p_i^1 \in (q_0, q_\infty)$ and similarly, $q_j^1 \in (p_\infty, p_0)$.

4) Since p_i is to the left of p_{i-1}, we have

$$V(p_i) > V(p_{i-1}) > V(p_c).$$

By 3), p_i^1 is to the right of p_c, this implies

$$V(p_c) < V(p_i^1) = V(A^{m_i}p_i)$$

and hence

$$V(A^{m_i}p_c) < V(p_c) < V(A^{m_i}p_i). \tag{10}$$

Since $p_{i-1} \in (p_i, p_c)$, the point $A^{m_i}p_{i-1}$ is on the line between $A^{m_i}p_c$ and $A^{m_i}p_i$. Also, since the function $V(x)$ is convex, it follows from (10) that

$$V(A^{m_i}p_{i-1}) < \max\{V(A^{m_i}p_c), V(A^{m_i}p_i)\} = V(A^{m_i}p_i).$$

Since $m_{i-1} > m_i$, we have $V(A^{m_{i-1}}p_{i-1}) < V(A^{m_i}p_{i-1})$ and

$$\begin{aligned} V(p_{i-1}^1) &= V(A^{m_{i-1}}p_{i-1}) < V(A^{m_i}p_{i-1}) \\ &< V(A^{m_i}p_i) = V(p_i^1). \end{aligned} \tag{11}$$

By 3), p_{i-1}^1 and p_i^1 are both to the right of p_c, hence the inequality (11) implies that p_i^1 is to the right of p_{i-1}^1, i.e., $p_i^1 \in (p_{i-1}^1, q_\infty)$.

Similarly, we have $q_j^1 \in (p_\infty, q_{j-1}^1)$.

5) Suppose that both $q_j^1, p_i^1 \in [p_i, q_j]$, then by 2) and 3), we have

$$n_{j-1} = n_j + m_{i-k1} \quad m_{i-1} = m_i + n_{j-k2}$$

where $k_1, k_2 \geq 1$. Since $m_{i-1} \leq m_{i-k1}$ and $n_{j-1} \leq n_{j-k2}$, it follows that

$$m_{i-1} < n_{j-1} \quad n_{j-1} < m_{i-1}$$

which is a contradiction.

On the other hand, suppose that both $q_j^1, p_i^1 \notin [p_i, q_j]$, then we must have q_j^1 to the left of p_i and p_i^1 to the right of q_j. Hence

$$V(q_j^1) > V(p_i) \quad V(p_i^1) > V(q_j).$$

Recall that $q_j^1 = A^{n_j}q_j$ and $p_i^1 = A^{m_i}p_i$, it follows that

$$V(q_j^1) > V(p_i) > V(p_i^1) > V(q_j) > V(q_j^1)$$

which is also a contradiction. □

From 1), 3), and 5) of Lemma 4.2, we can see that the only pair of p_i and q_j such that $p_i^1, q_j^1 \in [p_i, q_j]$ is p_0 and q_0. This fact can be used to generate the points p_i, $0, 1, \ldots, I$, and q_j, $j = 0, 1, \ldots, J$. Although it is possible to determine these points directly from the definition, it is hard to derive a computationally efficient method to generate the points from inside to outside, i.e., from p_0, q_0 to p_I, q_J. In the following, we provide an iterative method based on the properties in Lemmas 4.1 and 4.2 to generate the points from outside to inside, i.e., from p_I, q_J to p_0, q_0 and use the unique property that $p_0^1, q_0^1 \in [p_0, q_0]$ as a sign to stop the iteration.

Algorithm for Generating p_i, q_j, m_i, n_j and p_i^1, q_j^1

Step 1 Set $ii = 1$. Get the two intersections of AL_h^e with $\pm L_h^e$. They are p_I^1 and $-q_J^1$ (or $-p_I^1$ and q_J^1). Denote the line segment of AL_h^e between L_h^e and $-L_h^e$ as L_1. Multiply the two end points of L_1 from left with A^{-1}, then we get p_I and q_J. Clearly, the one to the left of p_c is p_I and the one to its right is q_J. If $p_I^1, q_J^1 \in [p_I, q_J]$, then $I = J = 0$ and stop the algorithm.

Step 2 $ii = ii + 1$. Check if AL_{ii-1} has intersections with $\pm L_h^e$. If not, let $L_{ii} = AL_{ii-1}$ and repeat this step. If there is, then cut off the part of AL_{ii-1} that is outside of S^e and let the remaining part be L_{ii}. The cut-off place is one of $\pm p_i^1$ and $\pm q_j^1$. Multiply the cut-off place from left with A^{-ii}. The result is so far the innermost p_i if it is to the left of p_c, or the innermost q_j if to the right of p_c. In the mean time, we also obtain $m_i = ii$ and\or $n_j = ii$. Let the innermost pair be p_i, q_j, if $p_i^1, q_j^1 \in [p_i, q_j]$, then we must have $i = j = 0$ and stop the algorithm since all the p_i, q_j have been computed. If $p_i^1, q_j^1 \in [p_i, q_j]$ is not true, then repeat this step.

We see from the above algorithm that the number of iterations equals $\max\{m_0, n_0\}$.

Item 4) in Lemma 4.2 shows that $[p_{i-1}, p_{i-1}^1] \subset [p_i, p_i^1]$ and $[q_{j-1}^1, q_{j-1}] \subset [q_j^1, q_j]$. Item 5) shows that either we have $[p_i, p_i^1] \subset [q_j^1, q_j]$, or $[q_j^1, q_j] \subset [p_i, p_i^1]$. Item 3) shows that all these intervals must include $[p_0, q_0]$. In summary, the facts in Lemma 4.2 jointly show that the intervals $[p_i, p_i^1]$ and $[q_j^1, q_j]$, $i = 0, 1, \ldots, I$, $j = 0, 1, \ldots J$ are ordered by inclusion. We can draw a figure for easy understanding of Lemma 4.2. If we draw arcs from p_i to p_i^1 and arcs from q_j to q_j^1, then these arcs

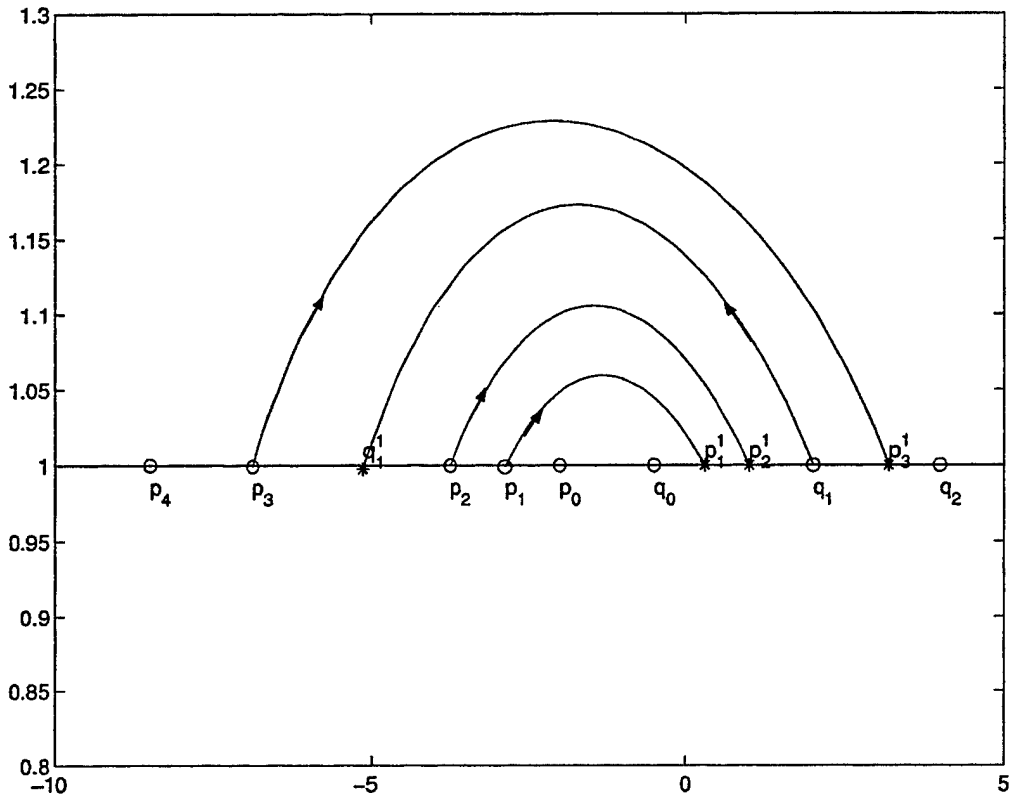


Fig. 2. Illustration for Lemma 4.2.

can be made not to intersect each other (see Fig. 2). In the figure, we have

$$[p_0, q_0] \subset [p_1, p_1^1] \subset [p_2, p_2^1] \subset [q_1^1, q_1] \subset [p_3, p_3^1].$$

Let $p_i^2 = \psi_2(m_{i-1}, p_i) = \text{sat}_2(A^{m_i-1}p_i)$, then p_i^2 is the third intersection of $\psi_2(k, p_i)$ with $\pm L_h^e$. By Lemma 4.1 4), we know that m_{i-1} is the smallest integer k such that $A^k p_i$ is outside of $\mathbf{S}^e$. Also let $q_j^2 = \psi_2(n_{j-1}, q_j) = \text{sat}_2(A^{n_j-1}q_j)$.

Lemma 4.3: Suppose that (9) is not true for any $x_1 \in \mathbf{R}$, then $p_i^2 \in (p_i, p_i^1)$ and $q_j^2 \in (q_j^1, q_j)$.

Proof: Suppose $p_i^1 \in (q_j, q_{j+1}]$, then by Lemma 4.2 2), the smallest k for $A^k p_i$ to go out of $\mathbf{S}^e$ is $k = m_{i-1} = m_i + n_j$. So $p_i^2 = \text{sat}_2(A^{m_i+n_j}p_i) = \text{sat}_2(A^{n_j}p_i^1)$. If on the contrary that $p_i^2 \in [p_i^1, q_\infty)$ (to the right of p_i^1), since $q_j^1 = A^{n_j}q_j \in \pm L_h^e$ is to the left of q_j, there must be a point $x \in (q_j, p_i^1) \subset (q_j, q_{j+1}]$ such that $A^{n_j}x$ is right above x, i.e., $\text{sat}_2(A^{n_j}x) = x$. By Lemma 4.1 3), the next intersection of $\psi_2(k, x)$ with $\pm L_h^e$ is $\text{sat}_2(A^{n_j}x)$, so we must have $A^k x \in \mathbf{S}^e$ for all $k < n_j$ and there exists x_1 such that (9) is true. A contradiction. On the other hand, if $p_i^2 = \text{sat}_2(A^{m_i-1}p_i) \in (p_\infty, p_i]$ (to the left of p_i), since $p_{i-1}^1 = A^{m_i-1}p_{i-1}$ is to the right of p_{i-1}, there must be a point $x \in [p_i, p_{i-1})$ such that $A^{m_i-1}x$ is right above x. Similar to the former case, we have a contradiction. Therefore, $p_i^2 \in (p_i, p_i^1)$, and similarly, $q_j^2 \in (q_j^1, q_j)$. □

This lemma says that if a trajectory starts from p_i or q_j, its third intersection with $\pm L_h^e$ will be closer to the central interval $[p_0, q_0]$ than the first intersection or the second one. We will show in the next lemma that this property can be actually extended to all $x \in L_h^e$.

Lemma 4.4: Assume that the condition (9) is not true for any $x_1 \in \mathbf{R}$. Given $x(0) = x_0 \in L_h^e$. Let x_0^1 and x_0^2 be the second and the third intersection of the trajectory $\psi_2(k, x_0)$ with L_h^e, (if the intersections are on $-L_h^e$, then get symmetric projections on L_h^e). If $x_0 \in (p_\infty, p_0]$, then $x_0^1 \in (x_0, q_\infty)$ and one of the following must be true.

1) $x_0^1 \in (p_0, q_0)$ and there is no third intersection x_0^2;
2) $x_0^1 \in (x_0, p_0]$;
3) $x_0^1 \in [q_0, q_\infty)$ and $x_0^2 \in (x_0, x_0^1)$.

Similarly, if $x_0 \in [q_0, q_\infty)$, then $x_0^1 \in (p_\infty, x_0)$ and one of the following must be true.

4) $x_0^1 \in (p_0, q_0)$ and there is no third intersection x_0^2;
5) $x_0^1 \in [q_0, x_0)$;
6) $x_0^1 \in (p_\infty, p_0]$ and $x_0^2 \in (x_0^1, x_0)$.

Also, if $x_0 \in (p_i, p_i^1)$ [or $x_0 \in (q_j^1, q_j)$], then x_0^1, x_0^2 and the subsequent intersections will all be in the interval (p_i, p_i^1) [or (q_j^1, q_j)]. Furthermore, for any $x_0 \in L_h^e$, there is a finite k_1 such that $\psi_2(k_1, x_0) \in (p_0, q_0)$. After that, $\psi_2(k, x_0)$ will have no more intersection with L_h^e and will converge to the origin.

Proof: Lemma 4.2 says that all the segments $[p_i, p_i^1]$ and $[q_j^1, q_j]$ are ordered by inclusion. We will prove the result of this lemma from the innermost segment to the outermost with an inductive procedure. Without loss of generality, assume that $[p_1, p_1^1]$ is the innermost segment (except for $[p_0, q_0]$), then we must have $p_1^1 \in (q_0, q_1]$ and $p_1^2 = \text{sat}_2(A^{n_0}p_1^1) = \text{sat}_2(A^{m_0}p_1)$.

By Lemma 4.3, $p_1^2 \in (p_1, p_1^1)$. There are three possibilities.

Case 1—$p_1^2 \in [q_0, p_1^1)$: (See Fig. 3.) For $x_0 \in [q_0, p_1^1)$, $x_0^1 = \text{sat}_2(A^{n_0}x_0)$ by Lemma 4.1 3). Since $q_0^1 = \text{sat}_2(A^{n_0}q_0)$ and $p_1^2 = \text{sat}(A^{n_0}p_1^1)$ are to the left of q_0 and p_1^1 respectively,

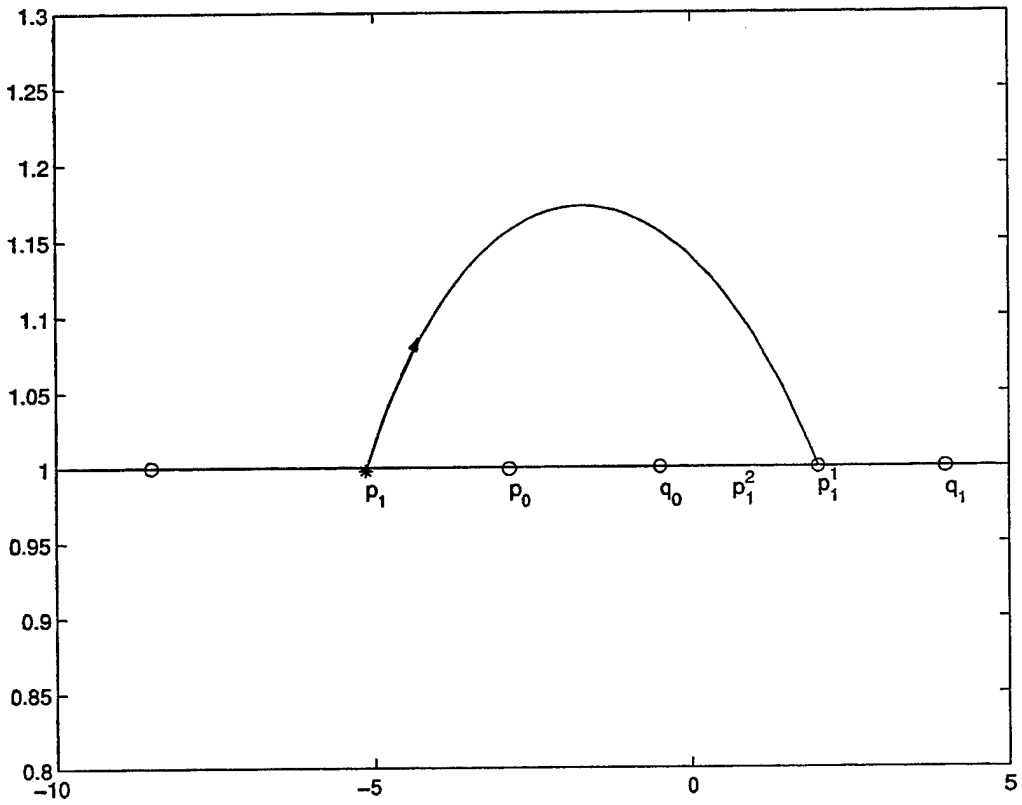


Fig. 3. Illustration for the proof of Lemma 4.4: Case 1.

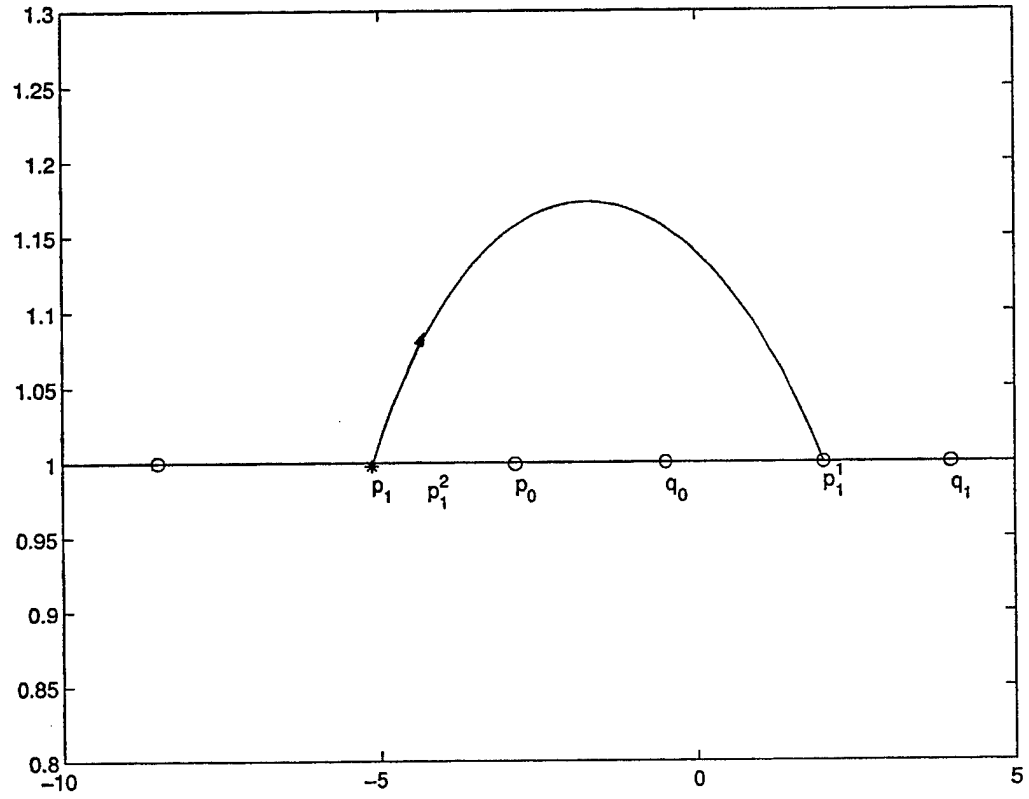


Fig. 4. Illustration for the proof of Lemma 4.4: Case 2.

x_0^1 must be to the left of x_0 and also, $x_0^1 \in [q_0^1, p_1^2)$. So we have $x_1^1 \in (q_0^1, x_0)$. This belongs to 4) or 5) of the lemma. If it is 4), then there will be no more intersection; If it is 5), then with the same argument, we have $x_0^2 \in (q_0^1, x_0^1)$, ... Moreover, the subsequent intersections will fall between p_0 and q_0 in a finite number of steps since there is no x_1 satisfying (9).

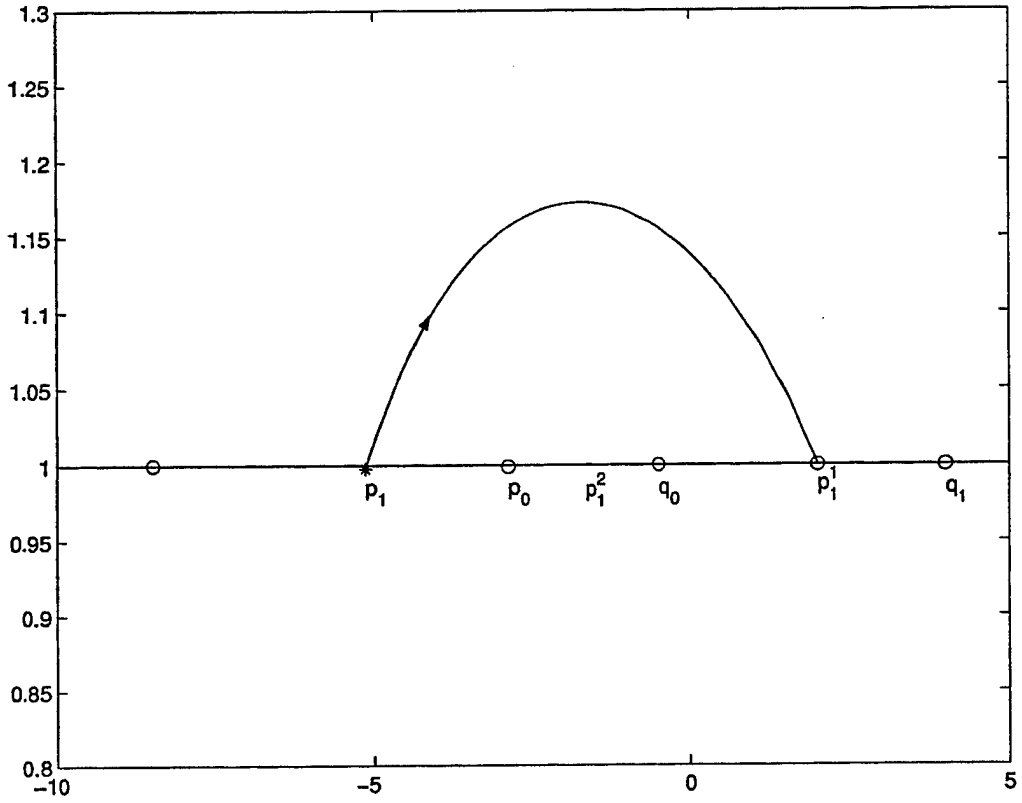


Fig. 5. Illustration for the proof of Lemma 4.4: Case 3.

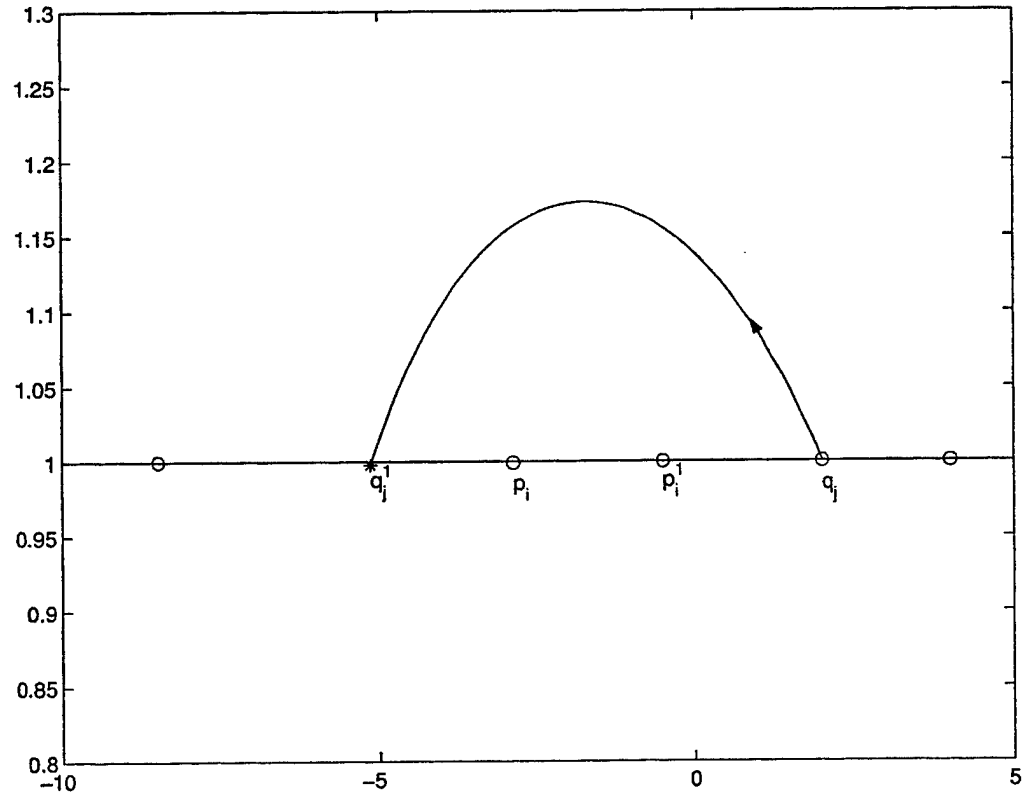


Fig. 6. Illustration for the proof of Lemma 4.4.

For $x_0 \in (p_1, p_0]$, $x_0^1 = \mathrm{sat}_2(A^{m_0}x_0) \in [p_0^1, p_1^2)$. If $x_0^1 \in (p_0, q_0)$, then we get 1) of the lemma. If $x_0^1 \in [q_0, p_1^2)$, then by the argument in the previous paragraph, we must have $x_0^2 \in [q_0^1, x_0^1) \subset (x_0, x_0^1)$ and we get 3) of the lemma.

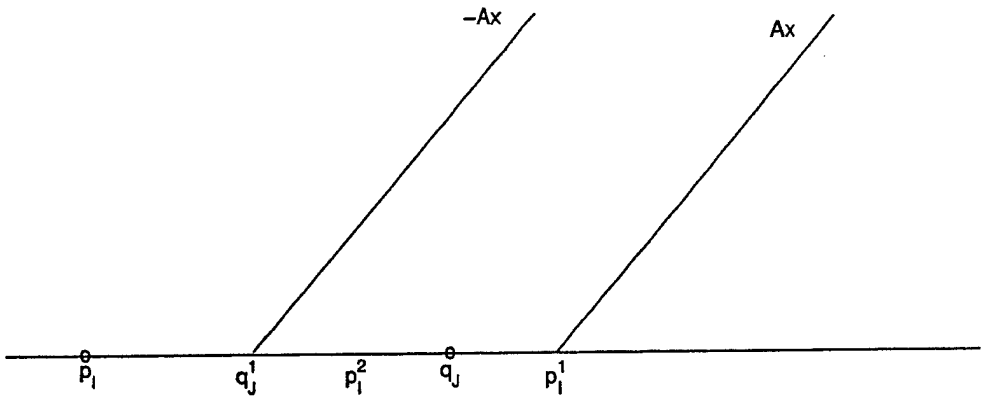
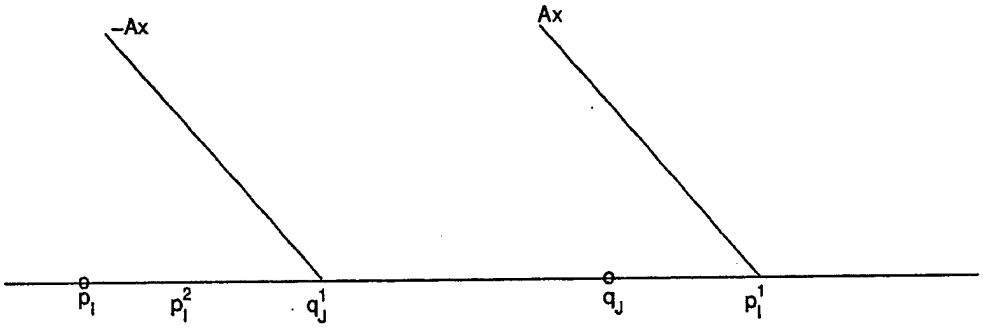


Fig. 7. Illustration for the proof of Lemma 4.4: Case i.



8. Illustration for the proof of Lemma 4.4: Case ii.

Case 2—$p_1^2 \in (p_1, p_0]$: (See Fig. 4.) For $x_0 \in (p_1, p_0]$, $x_0^1 = \text{sat}_2(A^{m_0}x_0) \in (p_1^2, p_0^1]$. Since p_1^2 and p_0^1 are to the right of p_1 and p_0 respectively, x_0^1 must also be to the right of x_0, i.e., $x_0^1 \in (x_0, p_0^1]$. If $x_0^1 \in (x_0, p_0)$, then we get 2) and the subsequent intersections, if any, will move rightward until falling between p_0 and q_0; if $x_0^1 \in (p_0, p_0^1) \subset (p_0, q_0)$, then we get 1).

For $x_0 \in [q_0, p_1^1)$, $x_0^1 = \text{sat}_2(A^{n_0}x_0) \in (p_1^2, q_0^1)$. If $x_0^1 \in [p_0, q_0^1) \subset (p_0, q_0)$ then we obtain 4). If $x_0^1 \in (p_1^2, p_0]$, then the argument in the foregoing paragraph applies and we have $x_0^2 \in (x_0^1, p_0^1] \subset (x_0^1, x_0)$, which belongs to 6).

Case 3—$p_1^2 \in (p_0, q_0)$: (See Fig. 5.) For $x_0 \in (p_1, p_0]$, we have $x_0^1 \in (p_1^2, p_0^1] \subset (p_0, q_0)$, which belongs to 1). For $x_0 \in [q_0, p_1^1)$, we have $x_0^1 \in (p_1^2, q_0^1] \subset (p_0, q_0)$, which belongs to 4).

So far, we have shown that one of 1)–6) holds for all $x_0 \in [p_1, p_1^1]$. And in each of the above three cases, we see that for all $x_0 \in (p_1, p_1^1)$, x_0^1, x_0^2 and the subsequent intersections are all in (p_1, p_1^1) and will fall between p_0 and q_0 in a finite number of steps.

Next, we assume that these properties hold for all $x_0 \in [p_i, p_i^1]$ and the next segment which includes $[p_i, p_i^1]$ is $[q_j^1, q_j]$ (see Fig. 6). We also have three cases: $q_j^2 \in [p_i^1, q_j)$, $q_j^2 \in (q_j^1, p_i]$ and $q_j^2 \in (p_i, p_i^1)$. By treating segment $[p_i, p_i^1]$ as $[p_0, q_0]$ in the proof for $[p_1, p_1^1]$, we can use the same argument to show that one of 1)–6) holds for all $x_0 \in (q_j^1, p_i] \cup [p_i^1, , q_j)$. Moreover, for all $x_0 \in (q_j^1, p_i] \cup [p_i^1, , q_j)$, the intersections will move toward $[p_i, p_i^1]$ and fall between $[p_i, p_i^1]$ in a finite number of steps.

Now, suppose that $[p_I, p_I^1]$ is the outermost segment. By induction, we have obtained the properties in the lemma for all $x_0 \in [p_I, p_I^1]$ and we would like to extend the properties to the whole line L_h^e.

Recall that $m_I = n_J = 1$, so $p_I^1 = Ap_I$ and $p_I^2 = \text{sat}_2(A^2p_I) = \text{sat}_2(Ap_I^1)$. The line AL_h^e will actually intersect with $\pm L_h^e$ at p_I^1 and $-q_J^1$ (or $-p_I^1$ and q_J^1). Assume that p_I^1 is on L_h^e. Then the ray $\{Ax\colon x \in [q_J, q_\infty)\}$ is below the line $-L_h^e$. Get a symmetric projection of this ray as $\{-Ax\colon x \in [q_J, q_\infty)\}$. Then the two rays $\{-Ax\colon x \in [q_J, q_\infty)\}$ and $\{Ax\colon x \in (p_\infty, p_I]\}$ are parallel and are both above the line L_h^e (see Figs. 7 and 8). Here, we have two cases.

Case i: The rays have a positive slope (see Fig. 7). Since $p_I^1 \in [q_J, q_\infty)$, we have $p_I^2 \in [q_J^1, q_\infty)$ and by Lemma 4.3, $p_I^2 \in (p_I, p_I^1)$. So, $p_I^2 \in (q_J^1, p_I^1)$.

For $x_0 \in (p_I^1, q_\infty)$, since both $q_J^1 = -\text{sat}_2(Aq_J)$ and $p_I^2 = -\text{sat}_2(Ap_I^1)$ are to the left of q_J and p_I^1, respectively, and since there exists no $x \in L_h^e$ such that $-\text{sat}_2(Ax) = x$, we must have, $x_0^1 = -\text{sat}_2(Ax_0)$ to the left of x_0 and $x_0^1 \in (p_I^2, x_0)$. If $x_0^1 \in (p_0, q_0)$, then we obtain 4). If $x_0^1 \in [q_0, x_0)$, then we obtain 5). If $x_0^1 \in (q_J^1, p_0) \subset (q_J^1, q_J)$, then by the established properties on $[q_J^1, q_J]$, we must have $x_0^2 \in (x_0^1, q_J)$ to the left of p_I^1 and hence to the left of x_0. Therefore, $x_0^2 \in (x_0^1, x_0)$ and we get 6).

For $x_0 \in (p_\infty, p_I)$, $x_0^1 = \text{sat}_2(Ax)$ is to the right of p_I^1, so the properties for $x_0 \in (p_I^1, p_\infty)$ applies. Also note that x_0^2 is to the right of p_I^2. Hence $x_0^2 \in (x_0, x_0^1)$ and we obtain 3).

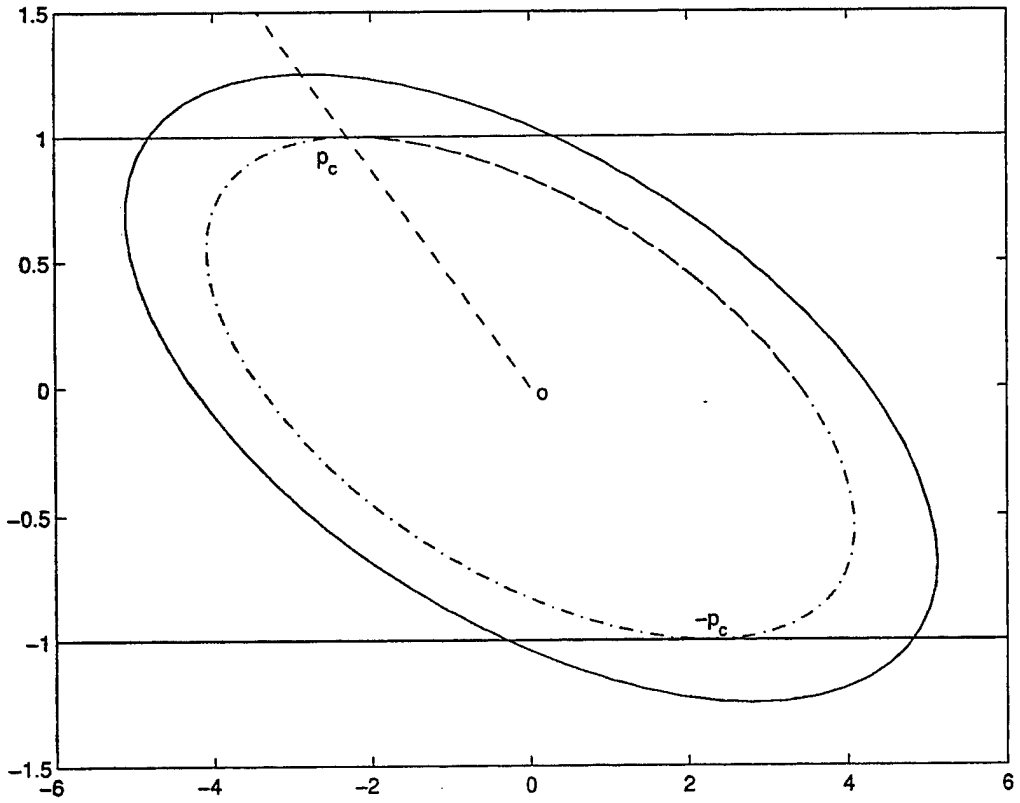


Fig. 9. Illustration for the proof of Lemma 4.5.

Case ii: The rays have a negative slope (see Fig. 8). In this case, $p_I^2 \in (p_I, q_J^1)$.

For $x_0 \in (p_\infty, p_I)$, since $p_I^1 = Ap_I$ is to the right of p_I and there exist no $x \in L_h^e$ such that $\text{sat}_2(Ax) = x$, we must have $x_0^1 = \text{sat}_2(Ax_0)$ to the right of x_0, in particular, $x_0 \in (x_0, p_I^1)$. If $x_0^1 \in (x_0, p_0]$, then we obtain 2). If $x_0^1 \in (p_0, q_0)$, then we have 1). If $x_0^1 \in (q_0, p_I^1)$, then by using the established property in the interval (p_I, p_I^1), we have $x_0^2 \in (p_I, x_0^1) \subset (x_0, x_0^1)$ and we obtain 3).

For $x_0 \in (p_I^1, q_\infty)$, $x_0^1 = -\text{sat}_2(Ax)$ is to the right of p_I^2. By applying the property for x_0 in (p_∞, p_I) and (p_I, p_I^1), we have $x_0^2 \in (x_0^1, x_0)$, which belongs to f).

Similar to the argument for the interval (p_1, p_1^1), it can be shown that the intersections will fall between (p_I, p_I^1) in a finite number of steps for all $x_0 \notin (p_I, p_I^1)$.

In summary, the intersections of a trajectory $\psi_2(k, x_0)$ with the lines $\pm L_h^e$ will move from the outer intervals to the inner intervals until falling into (p_0, q_0) in a finite number of steps. After that, it will not touch the lines $\pm L_h^e$ and will converge to the origin. □

Next, we suppose that the condition (9) is true for some $x_1 \in \mathbf{R}$. We would like to determine an interval in L_h^e such that a trajectory starting from this interval will converge to the origin.

Recall that p_c is defined to be the unique intersection of the Lyapunov ellipsoid $\mathcal{E}(\rho_c)$ with the line L_h^e (see Fig. 9). Also, α_c is the first coordinate of p_c, i.e., $p_c = \begin{bmatrix} \alpha_c \\ 1 \end{bmatrix}$.

Lemma 4.5: Assume that $\alpha_c \leq 0$. If there exist an integer $N > 0$ and a $d > 0$ such that $A^N \begin{bmatrix} x_1 \\ 1 \end{bmatrix} = \pm \begin{bmatrix} x_1 \\ 1+d \end{bmatrix}$, then we must have $x_1 < \alpha_c < 0$.

Proof: When $\alpha_c \leq 0$, an ellipsoid $\mathcal{E}(\rho)$ takes the shape in Fig. 9. Each ellipsoid $\mathcal{E}(\rho)$ has an intersection with the ray that starts from the origin and passes through p_c. This intersection is the highest point in the ellipsoid. Since $\alpha_c \leq 0$, it can be seen that

$$V\left(\begin{bmatrix} x_1 \\ 1 \end{bmatrix}\right) < V\left(\pm \begin{bmatrix} x_1 \\ 1+d \end{bmatrix}\right) \qquad \forall x_1 \geq \alpha_c, \quad d > 0.$$

Since

$$V\left(A^N \begin{bmatrix} x_1 \\ 1 \end{bmatrix}\right) < V\left(\begin{bmatrix} x_1 \\ 1 \end{bmatrix}\right)$$

it is impossible to have $A^N \begin{bmatrix} x_1 \\ 1 \end{bmatrix} = \pm \begin{bmatrix} x_1 \\ 1+d \end{bmatrix}$ for any $x_1 \geq \alpha_c$. □

Lemma 4.6: Let

$$\alpha_s = \min \{|x_1|: x_1 \text{ satisfies (9)}\}.$$

Case 1) $\alpha_c \leq 0$. Let

$$p_s = \begin{bmatrix} -\alpha_s \\ 1 \end{bmatrix}$$

then $p_s^1 = p_s \in (p_\infty, p_0)$. Suppose that $p_s \in [p_{i+1}, p_i)$. Then, for every $x_0 \in (p_s, p_i^1]$, the trajectory $\psi_2(k, x_0)$ will converge to the origin;

Case 2) $\alpha_c > 0$. Let

$$p_s = \begin{bmatrix} \alpha_s \\ 1 \end{bmatrix}$$

then $p_s^1 = p_s \in (q_0, q_\infty)$. Suppose that $p_s \in (q_j, q_{j+1}]$. Then, for every $x_0 \in [q_j^1, p_s]$, the trajectory $\psi_2(k, x_0)$ will converge to the origin.

In both cases, no limit trajectory can be formed completely inside the strip

$$\left\{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} : |x_1| < \alpha_s \right\}.$$

Proof: We only prove Case 1. Since $\alpha_c \leq 0$, by Lemma 4.5, there is no $x_1 \geq \alpha_c$ satisfying (9), so we have $-\alpha_s < \alpha_c < 0$ and p_s must be between p_{i+1} and p_i for some i, noting that p_s cannot be in $[p_0, q_0]$ by Lemma 4.1 1). Following the iterative procedure in the proof of Lemma 4.4, we can show that for all $x_0 \in [p_i, p_i^1]$, the trajectory will converge to the origin. Now we consider a point between p_s and p_i. For x_0 in this interval, the next intersection of the trajectory with the lines $\pm L_h^e$ is $x_0^1 = \text{sat}_2(A^{m_i}x_0)$. Since $\text{sat}_2(A^{m_i}p_s) = p_s$ (or $-p_s$) and $p_i^1 = \text{sat}_2(A^{m_i}p_i)$ is to the right of p_i, we must have $x_0^1 \in (x_0, p_i^1)$, and the subsequent intersections will move rightward and fall between p_i and p_i^1 in a finite number of steps. Therefore, the trajectory $\psi_2(k, x_0)$ will converge to the origin.

Now, consider $x_0 \in (p_i^1, q_\infty) \subset (p_c, q_\infty)$. Let k_1 be the minimal integer such that $A^{k_1}x_0$ goes out of $\mathbf{S}^e$, then by the shape of the Lyapunov ellipsoid, the point $A^{k_1}x_0$ must be to the left of x_0 (or to the right of $-x_0$ if $A^k x_0$ is below the line $-L_h^e$), otherwise we would have $V(A^{k_1}x_0) > V(x_0)$, which is impossible. Hence, $x_0^1 = \text{sat}_2(A^{k_1}x_0)$ must be to the left of x_0, and the subsequent intersections either fall between p_s and p_i^1 at a finite step, or go to the left of p_s. This shows that no limit trajectory can be formed completely to the right of p_s and symmetrically, to the left of $-p_s$. Hence, no limit trajectory can be formed completely inside the strip

$$\left\{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} : |x_1| < \alpha_s \right\}.$$

□

Proof of Proposition 4.1: It follows immediately from Lemmas 4.4 and 4.6. □

The number α_s and the point $p_s = \left[\begin{smallmatrix} \pm\alpha_s \\ 1 \end{smallmatrix}\right]$ can be easily computed by applying Lemma 4.1. Actually, all the x_1 satisfying (9) can be determined. Let $x_0 = \left[\begin{smallmatrix} x_1 \\ 1 \end{smallmatrix}\right]$, then x_1 satisfies (9) for some N if and only if x_0 satisfies

$$A^N x_0 \notin \mathbf{S}^e \quad A^k x_0 \in \mathbf{S}^e \qquad \forall\, k < N$$

and

$$\text{sat}_2(A^N x_0) = x_0. \tag{12}$$

Assume that $\alpha_c < 0$, then by Lemma 4.5, we only need to check if there is such an x_0 in the interval (p_∞, p_0). Clearly, no x_0 in $[p_0, q_0]$ satisfies (12) by Lemma 4.1 1). So we need to check over the intervals $[p_{i+1}, p_i)$ with i increased from 0 to $I-1$ and the interval (p_∞, p_I).

Consider a point x_0 in the interval $[p_{i+1}, p_i)$. By Lemma 4.1 2), the smallest integer N for $A^N x_0 \notin \mathbf{S}^e$ is $N = m_i$. By Lemma 4.2 3), $A^{m_i}p_i = p_i^1 \in (q_0, q_\infty)$ is to the right of p_i. So there exists $x_0 \in [p_{i+1}, p_i)$ satisfying (12) if and only if $A^{m_i}p_{i+1}$ is to the left of p_{i+1}, i.e.,

$$\text{sat}_2(A^{m_i}p_{i+1}) \in (p_\infty, p_{i+1}]. \tag{13}$$

If this is true, then x_1, the first coordinate of x_0, can be solved from

$$[1 \quad 0]A^{m_i}\begin{bmatrix} x_1 \\ 1 \end{bmatrix} = \pm x_1. \tag{14}$$

In summary, we have the following

Algorithm for Determining All the x_1 Satisfying Condition (9): Assume $\alpha_c \leq 0$. Initially set $i = 0$.

Step 1) $i = i + 1$. If (13) is satisfied, then compute x_1 from (14). Repeat this step until $i = I - 1$.

Step 2) Solve

$$[1 \quad 0]A\begin{bmatrix} x_1 \\ 1 \end{bmatrix} = \pm x_1$$

for x_1, if $x_1 \in (-\infty, \alpha_I)$, then x_1 satisfies (9) with $N = 1$.

V. Proof of the Main Results

Now, we turn back to the system (5),

$$x(k+1) = \text{sat}(Ax(k)). \tag{15}$$

For easy reference, we restate Theorem 2.1 as follows.

Theorem 5.1: The system (15) is globally asymptotically stable if and only if A is stable and none of the following statements are true.

1) There exists an $N \geq 1$ such that

$$\text{sat}(A^N v_1)) = \pm v_1 \quad \text{and} \quad A^k v_1 \in \mathbf{S} \qquad \forall\, k < N.$$

2) There exists an $N \geq 1$ such that

$$\text{sat}(A^N v_2)) = \pm v_2 \quad \text{and} \quad A^k v_2 \in \mathbf{S} \qquad \forall\, k < N.$$

3) There exists an $x_1 \in (-1, 1)$ and an $N \geq 1$ such that

$$\text{sat}\left(A^N \begin{bmatrix} x_1 \\ 1 \end{bmatrix}\right) = \pm \begin{bmatrix} x_1 \\ 1 \end{bmatrix}$$

and

$$A^k \begin{bmatrix} x_1 \\ 1 \end{bmatrix} \in \mathbf{S} \qquad \forall\, k < N.$$

4) There exists an $x_2 \in (-1, 1)$ and an $N \geq 1$ such that

$$\text{sat}\left(A^N \begin{bmatrix} 1 \\ x_2 \end{bmatrix}\right) = \pm \begin{bmatrix} 1 \\ x_2 \end{bmatrix}$$

and

$$A^k \begin{bmatrix} 1 \\ x_2 \end{bmatrix} \in \mathbf{S} \qquad \forall\, k < N.$$

Proof: We will exclude the possibility of the existence of limit trajectories (except for the trivial one at the origin) under the condition that none of statements 1)–4) in the theorem is true. In the following, when we say a limit trajectory, we mean a nontrivial one other than the origin. Clearly, every limit trajectory must include at least one point on the boundary of the unit square, i.e., a point in the set $\pm(L_h \cup L_v \cup \{v_1, v_2\})$. By Proposition 3.1, we know that a limit trajectory cannot have points in both $\pm L_h$ and $\pm L_v$. So we have two possibilities here, limit trajectories including points in $\pm(L_h \cup \{v_1, v_2\})$, and those including points in $\pm(L_v \cup \{v_1, v_2\})$. Because of the similarity,

we only exclude the first possibility under the condition that none of 1)–3) is true, the second possibility can be excluded under the condition that none of 1), 2) and 4) is true.

For a given initial state x_0, we denote the trajectory of the system (15) as $\psi(k, x_0)$ and the trajectory of (7) as $\psi_2(k, x_0)$.

Clearly, if $x_1 \in (-1, 1)$ satisfies 3), then this x_1 also satisfies (9). On the other hand, suppose that there is some x_1 that satisfies (9). Let p_s be as defined in Lemma 4.6 for the system (7) [if there is no x_1 that satisfies (9), then we can assume that

$$p_s = \begin{bmatrix} \pm\infty \\ 1 \end{bmatrix}$$

and the following argument also goes through]. Note that, if there is some $x_1 \in \mathbf{R}$, $|x_1| \le 1$, that satisfies (9), i.e.,

$$A^N \begin{bmatrix} x_1 \\ 1 \end{bmatrix} = \pm \begin{bmatrix} x_1 \\ 1+d \end{bmatrix}$$

and

$$\left| [0 \quad 1] A^k \begin{bmatrix} x_1 \\ 1 \end{bmatrix} \right| \le 1 \qquad \forall k < N$$

we must also have

$$\left| [1 \quad 0] A^k \begin{bmatrix} x_1 \\ 1 \end{bmatrix} \right| \le |x_1| \qquad \forall\, k < N$$

which indicates that x_1 satisfies 3). Otherwise, as in the proof of Proposition 3.1, the area of the convex hull of the set

$$\{\pm x_0, \pm A x_0, \ldots, \pm A^{N-1} x_0\}$$

would be less than the area of the convex hull of the set

$$\{\pm A x_0, \pm A^2 x_0, \ldots, \pm A^N x_0\}.$$

This would be a contradiction to the fact that $|\det(A)| < 1$.

Hence, if no x_1 satisfies 3), then p_s must be outside of $\mathbf{S}$. By Proposition 4.1, no limit trajectory of (7) can lie completely inside the strip

$$\left\{ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} : |x_1| < \alpha_s \right\}.$$

It follows that no limit trajectory of (15) can lie completely between $-L_v$ and L_v. Therefore, no limit trajectory of (15) can include only boundary points in $\pm L_h$. On the other hand, if a limit trajectory include only boundary points $\pm v_1$ (or $\pm v_2$, note that, by Proposition 3.1, no limit trajectory can include both $\pm v_1$ and $\pm v_2$), then 1) or 2) must be true, which contradicts our assumption. In short, if there is a limit trajectory that include points in $\pm(L_h \cup \{v_1, v_2\})$, it must include at least one point on $\pm L_h$ and one on $\pm v_1$ (or $\pm v_2$). Here, we assume that it includes v_2.

Let us consider the trajectories $\psi(k, v_2)$ and $\psi_2(k, v_2)$. Suppose that $\psi(k, v_2)$ has an intersection with $\pm L_h$ but does not include v_1 and any point in $\pm L_v$, we conclude that $\psi(k, v_2) = \psi_2(k, v_2)$ will converge to the origin. The argument goes as follows.

Let k_0 be the smallest k such that $\psi(k, v_2)$ intersects $\pm L_h$. Denote $v_2^1 = \psi(k_0, v_2)$. Since 2) is not true, k_0 must also be the smallest k such that

$$\left| [0 \quad 1] A^k v_2 \right| \ge 1.$$

So, we have $\psi_2(k, v_2) = \psi(k, v_2)$ for all $k \le k_0$. Here, we have two cases.

Case 1—$\alpha_c \le 0$: In this case, p_s is to the left of v_2. Since $v_2^1 = \mathrm{sat}(A^{k_0} v_2) = \mathrm{sat}_2(A^{k_0} v_2)$ goes to the right of v_2, by Lemma 4.5, v_2 must be to the left of p_0. It follows that $v_2 \in (p_s, p_i^1]$, where $(p_s, p_i^1]$ is the interval in Lemma 4.6 2). Hence, $\psi_2(k, v_2)$ will converge to the origin. Moreover, the subsequent intersections of $\psi_2(k, v_2)$ with $\pm L_h$ are between v_2 and v_2^1. Since $\psi(k, v_2)$ does not touch $\pm L_v$, we must have $\psi(k, v_2) = \psi_2(k, v_2)$ and hence $\psi(k, v_2)$ will also converge to the origin.

Case 2—$\alpha_c > 0$: In this case p_s is to the right of v_1. By the assumption that $\psi(k, v_2)$ does not include v_1, the intersections of $\psi_2(k, v_2)$ with $\pm L_h$ will stay to the left of v_1 (or to the right of $-v_1$). Since $\alpha_c > 0$, by Lemma 4.5, the intersections will move rightward until falling on $[q_j^1, p_s)$, where $[q_j^1, p_s)$ is the interval in Lemma 4.6 3). Similar to Case 1, we have that $\psi_2(k, v_2)$ converges to the origin and $\psi(k, v_2) = \psi_2(k, v_2)$.

So far, we have excluded the possibility that a limit trajectory includes any point in the set $\pm(L_h \cup \{v_1, v_2\})$. The possibility that a limit trajectory includes any point in the set $\pm(L_v \cup \{v_1, v_2\})$ can be excluded in a similar way. Thus, there exists no limit trajectory of any kind and the system (15) must be globally asymptotically stable. □

Here we provide a simple method to check the conditions 3) and 4) of Theorem 5.1 based on the algorithm to determine all the x_1 satisfying (9) and hence p_s in the previous section. From the proof of Theorem 5.1, we see that 3) is true if and only if $p_s \in \mathbf{S}$. To check 4), we can exchange x_1 and x_2, i.e., use a state transformation $y = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right] x$. The system (15) is then equivalent to

$$y(k+1) = \mathrm{sat}\left(\overline{A} y(k)\right) \tag{16}$$

where $\overline{A} = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right] A \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$. The condition 4) for the system (15) is equivalent to the condition 3) for the system (16).

VI. Conclusions

We gave a complete stability analysis of a planar discrete-time linear system under saturation. The analysis involves intricate investigation on the intersections of the trajectories with the lines $x_1 = \pm 1$ and $x_2 = \pm 1$. Our main result provides a necessary and sufficient condition for such a system to be globally asymptotically stable.

References

[1] F. Albertini and D. D'Alessandro, "Asymptotic stability of continuous-time systems with saturation nonlinearities," *Syst. Control Lett.*, vol. 29, no. 3, pp. 175–180, 1996.

[2] D. S. Bernstein and A. N. Michel, "A chronological bibliography on saturating actuators," *Int. J. Robust Nonlinear Control*, vol. 5, no. 5, pp. 375–380, 1995.

[3] L. Hou and A. N. Michel, "Asymptotic stability of systems with saturation constraints," *IEEE Trans. Automat. Contr.*, vol. 43, pp. 1148–1154, Aug. 1998.

[4] T. Hu and Z. Lin, "A complete stability analysis of planar linear systems under saturation," *IEEE Trans. Circuits Syst. I*, vol. 47, pp. 498–512, Apr. 2000.

[5] L. Jin, P. N. Nikiforuk, and M. M. Gupta, "Absolute stability conditions for discrete-time recurrent neural networks," *IEEE Trans. Neural Networks*, vol. 5, pp. 954–964, Nov. 1994.

[6] Z. Lin, *Low Gain Feedback*, ser. Lecture Notes in Control and Information Sciences. London, U.K.: Springer, 1998, vol. 240.

[7] D. Liu and A. N. Michel, "Asymptotic stability of systems operating on a closed hypercube," *Syst. Control Lett.*, vol. 19, no. 4, pp. 281–285, 1992.
[8] ——, "Sparsely interconnected neural networks for associative memories with applications to cellular neural networks," *IEEE Trans. Circuits Syst.*, vol. 41, pp. 295–307, Apr. 1994.
[9] ——, *Dynamical Systems with Saturation Nonlinearities*, ser. Lecture Notes in Control and Information Sciences. London: Springer, 1994, vol. 195.
[10] R. Mantri, A. Saberi, and V. Venkatasubramanian, "Stability analysis of continuous time planar systems with state saturation nonlinearity," *IEEE Trans. Circuits Syst. I*, vol. 45, pp. 989–993, Sept. 1998.
[11] J. H. F. Ritzerfeld, "A condition for the overflow stability of second-order digital filters that is satisfied by all scaled state-space structures using saturation," *IEEE Trans. Circuits Syst.*, vol. 36, pp. 1049–1057, Aug. 1989.

Tingshu Hu was born in Sichuan, China in 1966. She received the B.S. and M.S. degrees in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 1985 and 1988, respectively. She is currently working toward the Ph.D degree in electrical engineering, at the University of Virginia, Charlottesville.

Her research interests include systems with saturation nonlinearities and robust control theory, areas in which she has published several papers. She is also the co-author (with Zongli Lin) of the book *Control Systems with Actuator Saturation: Analysis and Design* (Birkhäuser, Boston, 2001).

Ms. Hu is an Associate Editor on the Conference Editorial Board of the IEEE Control Systems Society.

Zongli Lin (S'89–M'90–SM'98) was born in Fuqing, Fujian, China, in 1964. He received the B.S. degree in mathematics and computer science from Amoy University, Xiamen, China, in 1983, the Master of Engineering degree in automatic control from the Chinese Academy of Space Technology, Beijing, China, in 1989, and the Ph.D. degree in electrical and computer engineering, from Washington State University, Pullman, WA, in May 1994.

From July 1983 to July 1986, Dr. Lin worked as a Control Engineer at Chinese Academy of Space Technology. In January 1994, he joined the Department of Applied Mathematics and Statistics, State University of New York, Stoney Brook as a Visiting Assistant Professor, where he became an Assistant Professor in September 1994. Since July 1997, he has been an Assistant Professor in electrical engineering at University of Virginia, Charlottesville. His current research interests include nonlinear control, robust control, and control of systems with saturating actuators. In these areas he has published several papers. He is also the author of the recent book, *Low Gain Feedback* (Springer-Verlag, London, 1998) and a co-author (with Tingshu Hu) of the book *Control Systems with Actuator Saturation: Analysis and Design* (Birkhäuser, Boston, 2001).

Dr. Lin has been an Associate Editor on the Conference Editorial Board of the IEEE Control Systems Society and currently serves as an Associate Editor of IEEE TRANSACTIONS ON AUTOMATIC CONTROL. He is the recipient of a U.S. Office of Naval Research Young Investigator Award.

Publication 10

Stabilization of Exponentially Unstable Linear Systems with Saturating Actuators

Tingshu Hu, Zongli Lin, and Li Qiu

Abstract—**We study the problem of stabilizing exponentially unstable linear systems with saturating actuators. The study begins with planar systems with both poles exponentially unstable. For such a system, we show that the boundary of the domain of attraction under a saturated stabilizing linear state feedback is the unique stable limit cycle of its time-reversed system. A saturated linear state feedback is designed that results in a closed-loop system having a domain of attraction that is arbitrarily close to the null controllable region. This design is then utilized to construct state feedback laws for higher order systems with two exponentially unstable poles.**

Index Terms—**Actuator saturation, domain of attraction, null controllable region, semiglobal stabilization.**

I. Introduction

We consider the problem of stabilizing exponentially unstable linear systems subject to actuator saturation. For systems that are not exponentially unstable, this stabilization problem has been focus of study and is now well addressed. For example, it was shown in [13] that a linear system subject to actuator saturation can be globally asymptotically stabilized by nonlinear feedback if and only if the system is asymptotically null controllable with bounded controls (ANCBC), which, as shown in [11], is equivalent to the system being stabilizable in the usual linear sense and having open-loop poles in the closed left-half plane. A nested feedback design technique for designing nonlinear globally asymptotically stabilizing feedback laws was proposed in [16] for a chain of integrators and was fully generalized in [14]. Alternative solutions to the global stabilization problem consisting of scheduling a parameter in an algebraic Riccati equation according to the size of the state vector was later proposed in [12], [17]. The question of whether or not a general linear ANCBC system subject to actuator saturation can be globally asymptotically stabilized by linear feedback was answered in [3], [15], where it was shown that a chain of integrators of length greater than two cannot be globally asymptotically stabilized by saturated linear feedback.

The notion of semiglobal asymptotic stabilization on the null controllable region for linear systems subject to actuator saturation was introduced in [7], [8]. The semi-global framework for stabilization requires feedback laws that yield a closed-loop system which has an asymptotically stable equilibrium whose domain of attraction includes an *a priori* given (arbitrarily large) bounded subset of the null controllable region. In [7], [8], it was shown that, for linear ANCBC systems subject to actuator saturation, one can achieve semi-global asymptotic stabilization on the asymptotically null controllable region (the whole state space in this case) using linear feedback laws.

Manuscript received September 18, 1998; revised June 7, 1999, October 22, 1999, and October 13, 2000. Recommended by Associate Editor S. Weiland. The work of T. Hu and Z. Lin was supported in part by the U.S. Office of Naval Research Young Investigator Program under Grant N00014-99-1-0670. The work of L. Qiu was supported by Hong Kong Research Grant Council.

T. Hu and Z. Lin are with the Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903 USA (e-mail: th7f@virginia.edu; zl5y@virginia.edu).

L. Qiu is with the Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (e-mail: eeqiu@ee.ust.hk).

Publisher Item Identifier S 0018-9286(01)05119-4.

Despite the existing results (see [2] for an extensive chronological bibliography on the subject), the general picture of stabilizing exponentially unstable linear systems with saturating actuators remains not as clear as that of ANCBC systems. It is clear that this kind of systems cannot be globally stabilized in any way since they are not globally null controllable. The largest possible region on which a system can be stabilized is the null controllable region. In [4], we gave an explicit description of the null controllable region for a general linear system in terms of a set of extremal trajectories of the (time reversed) antistable subsystem. We recall that a linear system is said to be antistable if all its poles are in the open right-half plane and semistable if all its poles are in the closed left-half plane. For example, for a second order antistable system, the boundary of its null controllable region is covered by at most two extremal trajectories; and for a third order antistable system, the set of extremal trajectories can be described in terms of parameters in a real interval.

Based on the description of the null controllable region in [4], we begin our study of stabilization with planar antistable systems. We show that for such a system the boundary of the domain of attraction under any stabilizing saturated linear state feedback is the unique stable limit cycle of its time-reversed system. Moreover, the domain of attraction is convex. We next show that any second order antistable linear system can be semiglobally asymptotically stabilized on its null controllable region by saturated linear feedback. That is, for any *a priori* given set in the interior of the null controllable region, there exists a saturated linear feedback law that yields a closed-loop system which has an asymptotically stable equilibrium whose domain of attraction includes this given set. This design is then utilized to construct state feedback laws for higher order systems with two exponentially unstable poles.

The remainder of this note is organized as follows. Section II contains a brief summary of the description of the null controllable region which will be used in this note. Section III determines the domain of attraction for a second order antistable linear system under any saturated stabilizing linear feedback law. Section IV constructs saturated feedback laws that achieve semiglobal asymptotic stability on the null controllable region for any linear systems having two exponentially unstable poles. Finally, Section V draws some brief conclusions.

II. Results on the Null Controllable Region

Consider a linear system

$$\dot{x}(t) = Ax(t) + bu(t), \qquad |u| \le 1 \tag{1}$$

where $x(t) \in \mathbf{R}^n$ is the state and $u(t) \in \mathbf{R}$ is the control. Assume that (A, b) is stabilizable. The null controllable region of the system, denoted as $\mathcal{C}$, is defined to be the set of states that can be steered to the origin in a finite time by using a control u that is measurable and $|u(t)| \le 1$ for all t. If A is antistable, then $\mathcal{C}$ is a bounded convex open set. For a general unstable system, $\mathcal{C}$ is the Cartesian product of the null controllable region of its semistable subsystem, which is the whole subspace, and that of its antistable subsystem. It was shown in [4] that $\partial\mathcal{C}$ (the boundary of $\mathcal{C}$) of an anti-stable system is composed of a set of extremal trajectories of its time reversed system. The time reversed system of (1) is

$$\dot{z}(t) = -Az(t) - bv(t), \qquad |v| \le 1. \tag{2}$$

Suppose that A is anti-stable, denote

$$\mathcal{E} := \{v(t) = \operatorname{sgn}(c'e^{At}b),\ t \in \mathbf{R}\colon c \ne 0\} \tag{3}$$

0018–9286/01$10.00 © 2001 IEEE

and for a control v, $|v(t)| < 1$ for all $t \in \mathbf{R}$, denote the trajectory of (2) under the control of v as

$$\Phi(t, v) := \int_{-\infty}^{t} e^{-A(t-\tau)} b v(\tau)\, d\tau. \tag{4}$$

Since A is antistable, the integral in (4) exists for all $t \in \mathbf{R}$, so $\Phi(t, v)$ is well defined. It is shown in [4] that

$$\partial \mathcal{C} = \{\Phi(t, v): t \in \mathbf{R}, v \in \mathcal{E}\}. \tag{5}$$

In particular, for a second-order antistable system, if A has two real eigenvalues, then

$$\begin{aligned}\partial \mathcal{C} &= \left\{\pm\left[e^{-At} z_c^- - \int_0^t e^{-A(t-\tau)} b\, d\tau\right]: t \in [0, \infty]\right\} \\ &= \{\pm(-2e^{-At} + I)A^{-1}b: t \in [0, \infty]\}\end{aligned} \tag{6}$$

where $z_c^- = A^{-1}b$ is the equilibrium point under the constant control $u = -1$; if A has a pair of complex eigenvalues $\alpha \pm j\beta$, $\alpha, \beta > 0$, then

$$\begin{aligned}\partial \mathcal{C} &= \left\{\pm\left[e^{-At} z_s^- - \int_0^t e^{-A(t-\tau)} b\, d\tau\right]: t \in [0, T_p]\right\} \\ &= \{\pm[e^{-At} z_s^- - (I - e^{-At})A^{-1}b]: t \in [0, T_p]\}\end{aligned} \tag{7}$$

where $T_p = \pi/\beta$ and $z_s^- = (I + e^{-AT_p})^{-1}(I - e^{-AT_p})A^{-1}b$.

III. Domain of Attraction Under Saturated Linear State Feedback

Also consider the open-loop system (1). A saturated linear state feedback is given by $u = \sigma(fx)$, where $f \in \mathbf{R}^{1\times n}$ is the feedback gain and $\sigma(\cdot)$ is the saturation function $\sigma(s) = \operatorname{sgn}(s)\min\{1, |s|\}$. Such a feedback is said to be stabilizing if $A + bf$ is asymptotically stable. With a saturated linear state feedback applied, the closed-loop system is

$$\dot{x}(t) = Ax(t) + b\sigma(fx(t)). \tag{8}$$

Denote the state transition map of (8) by ϕ: $(t, x_0) \mapsto x(t)$. The domain of attraction $\mathcal{S}$ of the equilibrium $x = 0$ of (8) is defined by

$$\mathcal{S} := \left\{x_0 \in \mathbf{R}^n: \lim_{t\to\infty} \phi(t, x_0) = 0\right\}.$$

Obviously, $\mathcal{S}$ must lie within the null controllable region $\mathcal{C}$ of the system (1). Therefore, a design problem is to choose a state feedback gain so that $\mathcal{S}$ is arbitrarily close to $\mathcal{C}$. We refer to this problem as semiglobal stabilization on the null controllable region. We will first deal with antistable planar systems, then extend the results to higher order systems with only two antistable modes.

For the system (8), assume that $A \in \mathbf{R}^{2\times 2}$ is anti-stable. In [1], it was shown that the boundary of $\mathcal{S}$, denoted by $\partial\mathcal{S}$, is a closed orbit, but no method to find this closed orbit is provided. Generally, only a subset of $\mathcal{S}$ lying between $fx = 1$ and $fx = -1$ is detected as a level set of some Lyapunov function (see, e.g., [5]). Let P be a positive–definite matrix such that $(A+bf)'P + P(A+bf)$ is negative–definite. Since $\{z \in \mathbf{R}^2: -1 < fz < 1\}$ is an open neighborhood of the origin, it must contain

$$\mathcal{Q}_0 := \{z \in \mathbf{R}^2: z'Pz \le r_0\} \tag{9}$$

for some $r_0 > 0$. Clearly, $\mathcal{Q}_0$ is an invariant set inside $\mathcal{S}$. However, $\mathcal{Q}_0$ as an estimation of the domain of attraction can be very conservative (see, e.g., Fig. 1).

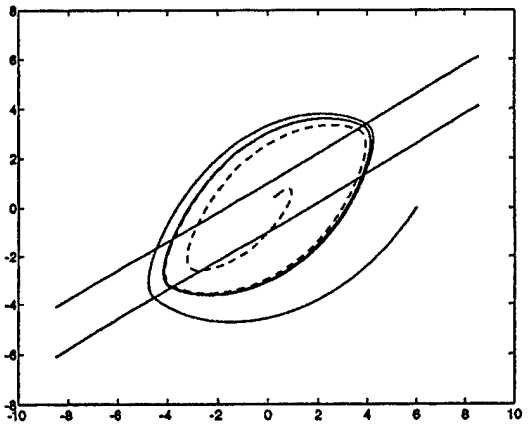


Fig. 1. Determination of $\partial\mathcal{S}$ from the limit cycle.

Lemma 3.1 [1]: The origin is the unique equilibrium point of system (8).

Let us introduce the time-reversed system of (8)

$$\dot{z}(t) = -Az(t) - b\sigma(fz(t)). \tag{10}$$

Clearly (10) also has only one equilibrium point, an unstable one, at the origin. Denote the state transition map of (10) by ψ: $(t, z_0) \mapsto z(t)$.

Theorem 3.1: $\partial\mathcal{S}$ is the unique limit cycle of planar systems (8) and (10). Furthermore, $\partial\mathcal{S}$ is the positive limit set of $\psi(\cdot, z_0)$ for all $z_0 \neq 0$.

This theorem says that $\partial\mathcal{S}$ is the unique limit cycle of (8) and (10). This limit cycle is a stable one for (10) (in a global sense), but an unstable one for (8). Therefore, it is easy to determine $\partial\mathcal{S}$ by simulating the time-reversed system (10). Shown in Fig. 1 is a typical result, where two trajectories, one starting from outside, the solid curve, and the other starting from inside, the dashed curve, both converge to the unique limit cycle. The straight lines in Fig. 1 are $fz = 1$ and $fz = -1$.

To prove Theorem 3.1, we need the following two lemmas, proofs of which can be found in [4].

Lemma 3.2: Suppose that $A \in \mathbf{R}^{2\times 2}$ is anti-stable and (f, A) is observable. Given a $c > 0$, let x_1, x_2, y_1 and y_2 ($x_1 \neq x_2$) be four points on the line $fx = c$, satisfying

$$y_1 = e^{AT_1}x_1, \qquad y_2 = e^{AT_2}x_2,$$

for some $T_1, T_2 > 0$ and

$$fe^{At_1}x_1 > c \quad fe^{At_2}x_2 > c, \quad \forall t_1 \in (0, T_1), \quad t_2 \in (0, T_2)$$

then, $\|y_1 - y_2\| > \|x_1 - x_2\|$.

Shown in Fig. 2 is an illustration of Lemma 3.2. The curve from x_i to y_i is $x(t) = e^{At}x_i$, $t \in [0, T_i]$, a segment of a trajectory of the autonomous system $\dot{x} = Ax$. Lemma 3.2 indicates that if any two different trajectories leave a straight line on the same side, they will be further apart when they return to it.

Lemma 3.3: Suppose that $A \in \mathbf{R}^{2\times 2}$ is asymptotically stable and (f, A) is observable. Given a $c > 0$, let x_1, x_2 be two points on the line $fx = c$ and y_1, y_2 be two points on $fx = -c$ such that

$$y_1 = e^{AT_1}x_1, \qquad y_2 = e^{AT_2}x_2$$

for some $T_1, T_2 > 0$, and

$$\begin{aligned}|fe^{At_1}x_1| < c, \qquad |fe^{At_2}x_2| < c, \\ \forall t_1 \in (0, T_1), \quad t_2 \in (0, T_2)\end{aligned}$$

then, $\|y_1 - y_2\| > \|x_1 - x_2\|$.

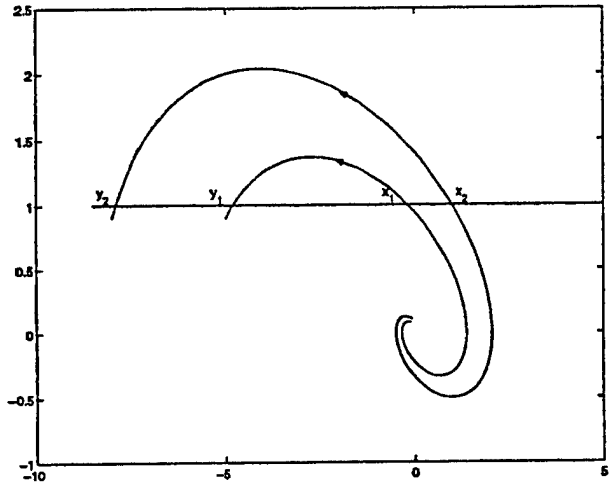


Fig. 2. Illustration of Lemma 3.2.

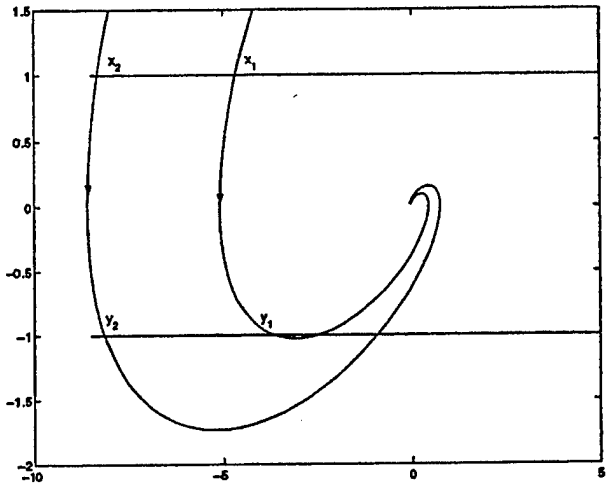


Fig. 3. Illustration of Lemma 3.3.

Shown in Fig. 3 is an illustration of Lemma 3.3. It says that if two different trajectories of the autonomous system $\dot{x} = Ax$ enter the region between $fx = c$ and $fx = -c$, they will be further apart when they leave the region. Notice that in Lemma 3.2, A is antistable, and in Lemma 3.3, A is asymptotically stable.

Proof of Theorem 3.1: We first prove that for the system (10), every trajectory $\psi(t, z_0)$, $z_0 \neq 0$, converges to a periodic orbit as $t \to \infty$. Recall that $\mathcal{Q}_0$ [defined in (9)] lies within the domain of attraction of the equilibrium $x = 0$ of (8) and is an invariant set. It follows that, for every state $z_0 \neq 0$ of (10), there is some $t_0 \geq 0$ such that $\psi(t, z_0)$ lies outside $\mathcal{Q}_0$ for all $t \geq t_0$. The state transition map of the system (10) is

$$\psi(t, z_0) = e^{-At}z_0 - \int_0^t e^{-A(t-\tau)}b\sigma(fz(\tau))\,d\tau. \qquad (11)$$

Since $-A$ is stable, the first term converges to the origin. Since $|\sigma(fz(\tau))| \leq 1$, the second term belongs to $\mathcal{C}$, the null controllable region of (1), for all t. It follows that there exists an $r_1 > r_0$ such that $\psi'(t, z_0)P\psi(t, z_0) \leq r_1 < \infty$ for all $t \geq t_0$. Let $\mathcal{Q} = \{z \in \mathbf{R}^2: r_0 \leq z'Pz \leq r_1\}$. Then $\psi(t, z_0)$, $t \geq t_0$, lies entirely in $\mathcal{Q}$. It follows from the Poincaré–Bendixon theorem that $\psi(t, z_0)$ converges to a periodic orbit.

The preceding paragraph shows that (8) and (10) have periodic orbits. We claim that the system (8) and (10) each has only one periodic orbit. For direct use of Lemma 3.2 and Lemma 3.3, we prove this claim through the original system (8).

First notice that a periodic orbit must enclose the unique equilibrium point $x = 0$ by the index theory, see e.g., [6], and must be symmetric to the origin ($-\Gamma$ is a periodic orbit if Γ is, hence if the periodic orbit is not symmetric, there will be two intersecting trajectories). Also, it cannot be completely contained in the linear region between $fx = 1$ and $fx = -1$. (Otherwise the asymptotically stable linear system $\dot{x} = (A + bf)x$ would have a closed trajectory in this region. This is impossible). Hence, it has to intersect each of the lines $fx = \pm 1$ at least twice. Assume without loss of generality that (f, A, b) is in the observer canonical form, i.e., $f = [0\ 1]$, $A = \left[\begin{smallmatrix} 0 & -a_1 \\ 1 & a_2 \end{smallmatrix}\right]$, $b = \left[\begin{smallmatrix} b_1 \\ b_2 \end{smallmatrix}\right]$, with a_1, $a_2 > 0$, and denote $x = \left[\begin{smallmatrix} \xi_1 \\ \xi_2 \end{smallmatrix}\right]$. In this case, $fx = \pm 1$ are horizontal lines. The stability of $A + bf$ requires that $-a_1 + b_1 < 0$ and $a_2 + b_2 < 0$. Observe that on the line $fx = 1$, we have $\xi_2 = 1$ and $\dot{\xi}_2 = \xi_1 + a_2 + b_2$. Hence, if $\xi_1 > -a_2 - b_2$, then $\dot{\xi}_2 > 0$, i.e., the trajectories go upwards; if $\xi_1 < -a_2 - b_2$, then $\dot{\xi}_2 < 0$, i.e., the trajectories go downwards. This implies that any periodic orbit crosses

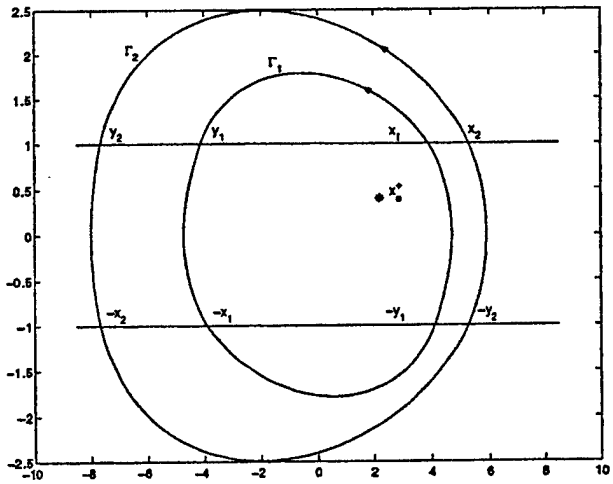


Fig. 4. Illustration for the proof of Theorem 3.1.

$fx = 1$ exactly twice and similarly for $fx = -1$. It also implies that a periodic orbit goes counter-clockwise.

Now, suppose on the contrary that (8) has two different periodic orbits Γ_1 and Γ_2, with Γ_1 enclosed by Γ_2, as illustrated in Fig. 4. Note that any periodic orbit must enclose the origin and any two trajectories cannot intersect. Hence, all the periodic orbits must be ordered by enclosement. Let x_1 and y_1 be the two intersections of Γ_1 with $fx = 1$, and x_2, y_2 be the two intersections of Γ_2 with $fx = 1$. Then, along Γ_1, the trajectory goes from x_1 to y_1, $-x_1$, $-y_1$ and returns to x_1; along Γ_2, the trajectory goes from x_2 to y_2, $-x_2$, $-y_2$ and returns to x_2.

Let $x_e^+ = -A^{-1}b$. Since $x_1 \to y_1$ along Γ_1 and $x_2 \to y_2$ along Γ_2 are on trajectories of $\dot{x} = Ax + b$ (or $d(x - x_e^+)/dt = A(x - x_e^+)$), we have

$$y_1 - x_e^+ = e^{AT_1}(x_1 - x_e^+) \quad y_2 - x_e^+ = e^{AT_2}(x_2 - x_e^+)$$

for some T_1, $T_2 > 0$. Furthermore, $f(x_1 - x_e^+) = f(x_2 - x_e^+) = f(y_1 - x_e^+) = f(y_2 - x_e^+) = 1 - fx_e^+ > 0$ (since $fx_e^+ = b_1/a_1 < 1$) and for all x on the two pieces of trajectories, $f(x - x_e^+) \geq 1 - fx_e^+$. It follows from Lemma 3.2 that

$$\|y_2 - y_1\| > \|x_2 - x_1\|.$$

On the other hand, $y_1 \to -x_1$ along Γ_1 and $y_2 \to -x_2$ along Γ_2 are on trajectories of $\dot{x} = (A + bf)x$ satisfying $-x_1 = e^{(A+bf)T_3}y_1$ and

$-x_2 = e^{(A+bf)T_4}y_2$ for some T_3, $T_4 > 0$. It follows from Lemma 3.3 that

$$\|x_2 - x_1\| > \|y_2 - y_1\|$$

which is a contradiction. Therefore, Γ_1 and Γ_2 must be the same periodic orbit. This shows that the systems have only one periodic orbit and, hence, it is a limit cycle.

We have so far proven that both (8) and (10) have a unique limit cycle and every trajectory $\psi(t, z_0)$, $z_0 \neq 0$, of (10) converges to this limit cycle. This implies that a trajectory $\phi(t, x_0)$ of (8) converges to the origin if and only if x_0 is inside the limit cycle. This shows that the limit cycle is ∂S. □

In the above proof, we also showed that ∂S is symmetric and has two intersections with $fx = 1$ and two with $fx = -1$. Another nice feature of S, as shown in [4], is that it is convex.

IV. Semiglobal Stabilization on the Null Controllable Region

A. Second Order Antistable Systems

In this subsection, we continue to assume that $A \in \mathbf{R}^{2\times 2}$ is antistable and (A, b) is controllable. We will show that the domain of attraction S of the equilibrium $x = 0$ of the closed-loop system (8) can be made arbitrarily close to the null controllable region C by judiciously choosing the feedback gain f. To state the main result of this section, we need to introduce the Hausdorff distance. Let $\mathcal{X}_1$, $\mathcal{X}_2$ be two bounded subsets of $\mathbf{R}^n$. Then, their Hausdorff distance is defined as

$$d(\mathcal{X}_1, \mathcal{X}_2) := \max\left\{\vec{d}(\mathcal{X}_1, \mathcal{X}_2), \vec{d}(\mathcal{X}_2, \mathcal{X}_1)\right\}$$

where

$$\vec{d}(\mathcal{X}_1, \mathcal{X}_2) = \sup_{x_1\in\mathcal{X}_1}\inf_{x_2\in\mathcal{X}_2}\|x_1 - x_2\|.$$

Here, the vector norm used is arbitrary.

Let P be the unique positive-definite solution of the following Riccati equation:

$$A'P + PA - Pbb'P = 0. \tag{12}$$

Note that this equation is associated with the minimum energy regulation, i.e., an LQR problem with cost

$$J = \int_0^\infty u'(t)u(t)\,dt.$$

The corresponding minimum energy state feedback gain is given by $f_0 = -b'P$. By the infinite gain margin and 50% gain reduction margin property of LQR regulators, the origin is a stable equilibrium of system

$$\dot{x}(t) = Ax(t) + b\sigma(kf_0x(t)) \tag{13}$$

for all $k > 0.5$. Let $S(k)$ be the domain of attraction of the equilibrium $x = 0$ of (13).

Theorem 4.1: $\lim_{k\to\infty} d(S(k), C) = 0$.

Proof: See the Appendix. □

Note that the use of high gain feedback is crucial here. The minimum energy feedback f_0 itself does not give a domain of attraction close to C. This is quite different from the related result in [8] and [9] for semistable open-loop systems. In these two papers, it was shown that if (A, B) is ANCBC, then low-gain feedback gives arbitrarily large domain of attraction.

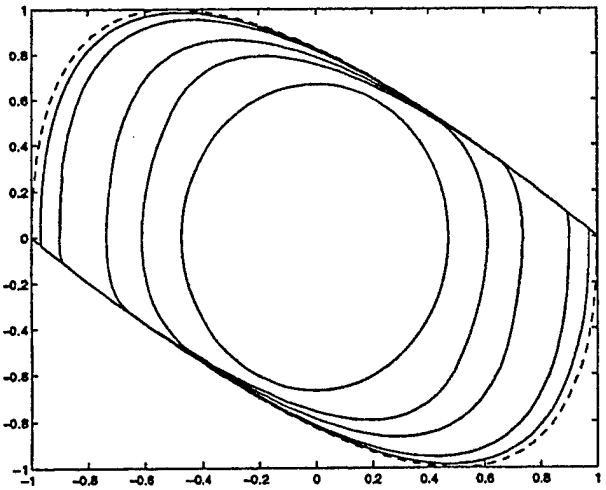


Fig. 5. Domains of attraction under different feedbacks.

Example 4.1: Let $A = \left[\begin{smallmatrix} 0 & -0.5 \\ 1 & 1.5 \end{smallmatrix}\right]$ and $b = \left[\begin{smallmatrix} 0 \\ -1 \end{smallmatrix}\right]$. Then $f_0 = [0\ 3]$. In Fig. 5, the boundaries of the domains of attraction corresponding to different $f = kf_0$, $k = 0.500\,05$, 0.65, 1, 3, are plotted. The regions do become bigger for greater k. The outermost boundary is ∂C. When $k = 3$, it can be seen that ∂S is already very close to ∂C.

B. Higher Order Systems with Two Exponentially Unstable Poles

Consider the following open-loop system:

$$\dot{x}(t) = Ax(t) + bu(t) = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}x(t) + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}u(t) \tag{14}$$

where $x = [x_1'\ x_2']'$, $x_1 \in \mathbf{R}^2$, $x_2 \in \mathbf{R}^n$, $A_1 \in \mathbf{R}^{2\times 2}$ is antistable and $A_2 \in \mathbf{R}^n$ is semi-stable. Assume that (A, b) is controllable. Denote the null controllable region of the subsystem

$$\dot{x}_1(t) = A_1x_1(t) + b_1u(t)$$

as C_1, then the null controllable region of (14) is $C_1 \times \mathbf{R}^n$[4]. Given γ_1, $\gamma_2 > 0$, denote

$$\begin{aligned}\Omega_1(\gamma_1) &:= \{\gamma_1x_1 \in \mathbf{R}^2 : x_1 \in \overline{C}_1\}\\ \Omega_2(\gamma_2) &:= \{x_2 \in \mathbf{R}^n : \|x_2\| \le \gamma_2\}.\end{aligned} \tag{15}$$

When $\gamma_1 = 1$, $\Omega_1(\gamma_1) = \overline{C}_1$ and when $\gamma_1 < 1$, $\Omega_1(\gamma_1)$ lies in the interior of C_1. In this section, we will show that given any $\gamma_1 < 1$ and $\gamma_2 > 0$, a state feedback can be designed such that $\Omega_1(\gamma_1) \times \Omega_2(\gamma_2)$ is contained in the domain of attraction of the equilibrium $x = 0$ of the closed-loop system.

For $\epsilon > 0$, let $P(\epsilon) = \left[\begin{smallmatrix} P_1(\epsilon) & P_2'(\epsilon) \\ P_2'(\epsilon) & P_3(\epsilon) \end{smallmatrix}\right] \in \mathbf{R}^{(2+n)\times(2+n)}$ be the unique positive-definite solution to the ARE

$$A'P + PA - Pbb'P + \epsilon^2 I = 0. \tag{16}$$

Clearly, as $\epsilon \downarrow 0$, $P(\epsilon)$ decreases. Hence, $\lim_{\epsilon\to 0} P(\epsilon)$ exists.

Let P_1 be the unique positive definite solution to the ARE

$$A_1'P_1 + P_1A_1 - P_1b_1b_1'P_1 = 0.$$

Then by the continuity property of the solution of the Riccati equation [19]

$$\lim_{\epsilon\to 0} P(\epsilon) = \begin{bmatrix} P_1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Let $f(\epsilon) := -b'P(\epsilon)$. First, consider the domain of attraction of the equilibrium $x = 0$ of the following closed-loop system:

$$\dot{x}(t) = Ax(t) + b\sigma(f(\epsilon)x(t)). \tag{17}$$

It is easy to see that

$$D(\epsilon) := \left\{x \in \mathbf{R}^{2+n}: x'P(\epsilon)x \le 1/\|b'P^{1/2}(\epsilon)\|^2\right\}$$

is contained in the domain of attraction of the equilibrium $x = 0$ of (17) and is an invariant set. Note that if $x_0 \in D(\epsilon)$, then $x(t) \in D(\epsilon)$ and $|f(\epsilon)x(t)| \le 1$ for all $t > 0$. That is, $x(t)$ will stay in the linear region of the closed-loop system, and in $D(\epsilon)$.

Lemma 4.1: Denote

$$r_1(\epsilon) := \frac{1}{2\left\|P_1^{1/2}(\epsilon)\right\| \|b'P^{1/2}(\epsilon)\|},$$

$$r_2(\epsilon) := \frac{-\|P_2(\epsilon)\| + \sqrt{\|P_2(\epsilon)\|^2 + 3\|P_1(\epsilon)\|\,\|P_3(\epsilon)\|}}{\|P_3(\epsilon)\|} r_1(\epsilon).$$

Then

$$D_1(\epsilon) := \left\{x \in \mathbf{R}^{2+n}: \|x_1\| \le r_1(\epsilon),\ \|x_2\| \le r_2(\epsilon)\right\} \subset D(\epsilon).$$

Moreover, $\lim_{\epsilon \to 0} r_2(\epsilon) = \infty$, and $r_1(\epsilon)$ increases with an upper bound as ϵ tends to zero.

Proof: It can be verified that

$$\|P_1(\epsilon)\| r_1^2(\epsilon) + 2\|P_2(\epsilon)\| r_1(\epsilon) r_2(\epsilon) + \|P_3(\epsilon)\| r_2^2(\epsilon) = \frac{1}{\|b'P^{1/2}(\epsilon)\|^2}. \tag{18}$$

So for all $x \in D_1(\epsilon)$, $x'P(\epsilon)x \le (1/\|b'P^{1/2}(\epsilon)\|^2)$, i.e., $D_1(\epsilon) \subset D(\epsilon)$. By the definition of $r_1(\epsilon)$ and $r_2(\epsilon)$, we have

$$r_2(\epsilon) = \frac{3\|P_1(\epsilon)\|}{\|P_2(\epsilon)\| + \sqrt{\|P_2(\epsilon)\|^2 + 3\|P_1(\epsilon)\|\,\|P_3(\epsilon)\|}} \cdot \frac{1}{2\left\|P_1^{1/2}(\epsilon)\right\| \|b'P^{1/2}(\epsilon)\|}.$$

Since as ϵ goes to zero, $P_2(\epsilon)$, $P_3(\epsilon) \to 0$, and $P_1(\epsilon) \to P_1$, so $r_1(\epsilon)$ is bounded whereas $r_2(\epsilon) \to \infty$. It follows from the monotonicity of $P(\epsilon)$ that r_1 is a monotonically decreasing function of ϵ. □

Theorem 4.2: Let $f_0 = -b_1'P_1$. For any $\gamma_1 < 1$ and $\gamma_2 > 0$, there exist a $k > 0.5$ and an $\epsilon > 0$ such that $\Omega_1(\gamma_1) \times \Omega_2(\gamma_2)$ is contained in the domain of attraction of the equilibrium $x = 0$ of the closed-loop system

$$\dot{x}(t) = Ax(t) + bu(t) \tag{19}$$

where

$$u(t) = \begin{cases} \sigma(kf_0x_1(t)), & x \notin D(\epsilon) \\ \sigma(f(\epsilon)x(t)), & x \in D(\epsilon). \end{cases} \tag{19}$$

Proof: Since $\gamma_1 < 1$, by Theorem 4.1, there exists a $k > 0.5$ such that $\Omega_1(\gamma_1)$ lies in the interior of the domain of attraction of the equilibrium $x_1 = 0$ of

$$\dot{x}_1(t) = A_1x_1(t) + b_1\sigma(kf_0x_1(t)). \tag{20}$$

Let $\epsilon_0 > 0$ be given. For an initial state $x_{10} \in \Omega_1(\gamma_1)$, denote the trajectory of (20) as $\psi(t, x_{10})$. Define

$$T(x_{10}) := \min\{t \ge 0: \|\psi(t, x_{10})\| \le r_1(\epsilon_0)\}$$

then $T(x_{10})$ is the time when $\psi(t, x_{10})$ first enters the ball $\{x_1 \in \mathbf{R}^2: \|x_1\| \le r_1(\epsilon_0)\}$. Let

$$T_M = \max\{T(x_{10}): x_{10} \in \partial\Omega_1(\gamma_1)\} \tag{21}$$

and

$$\gamma = \max_{t \in [0, T_M]} \|e^{A_2t}\|\gamma_2 + \int_0^{T_M} \left\|e^{A_2(T_M - \tau)}b_2\right\| d\tau \tag{22}$$

then by Lemma 4.1, there exists an $\epsilon < \epsilon_0$ such that $r_1(\epsilon) \ge r_1(\epsilon_0)$, $r_2(\epsilon) \ge \gamma$ and

$$D_1(\epsilon) = \{x \in \mathbf{R}^{2+n}: \|x_1\| \le r_1(\epsilon),\ \|x_2\| \le r_2(\epsilon)\} \subset D(\epsilon)$$

lies in the domain of attraction of the equilibrium $x = 0$ of (17).

Now consider an initial state of (19), $x_0 \in \Omega_1(\gamma_1) \times \Omega_2(\gamma_2)$. If $x_0 \in D(\epsilon)$, then $x(t)$ will go to the origin since $D(\epsilon)$ is an invariant set and is contained in the domain of attraction. If $x_0 \notin D(\epsilon)$, we conclude that $x(t)$ will enter $D(\epsilon)$ at some $T \le T_M$ under the control $u = \sigma(kf_0x_1(t))$. Observe that under this control, $x_1(t)$ goes along a trajectory of (20). If there is no switch, $x_1(t)$ will hit the ball $\{x_1 \in \mathbf{R}^2: \|x_1\| \le r_1(\epsilon_0)\}$ at $T(x_{10})$. Clearly $T(x_{10}) \le T_M$ and at this instant $\|x_2(T(x_{10}))\| \le \gamma \le r_2(\epsilon)$, so $x(T(x_{10})) \in D_1(\epsilon)$. Thus, we see that if there is no switch, $x(t)$ will be in $D_1(\epsilon)$ at $T(x_{10})$. Since $D_1(\epsilon) \subset D(\epsilon)$, $x(t)$ must have entered $D(\epsilon)$ at some earlier time $T \le T(x_{10}) \le T_M$. So we have that conclusion. With the switching control applied, once $x(t)$ enters the invariant set $D(\epsilon)$, it will remain in it and go to the origin asymptotically. □

V. Conclusion

We provided a simple semiglobal stabilization strategy for exponentially unstable linear systems with saturating actuators. For a planar antistable system, the controllers are saturated linear state feedbacks and for higher order systems with two antistable modes, the controllers are piecewise linear state feedbacks with only one switch.

Appendix
Proof of Theorem 4.1

For simplicity and without loss of generality, we assume that

$$A = \begin{bmatrix} 0 & -a_1 \\ 1 & a_2 \end{bmatrix}, \qquad a_1, a_2 > 0, \qquad b = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

Since A is anti-stable and (A, b) is controllable, A, b can always be transformed into this form. Suppose that A has already taken this form and $b = \left[\begin{smallmatrix} b_1 \\ b_2 \end{smallmatrix}\right]$. Let $V = [-A^{-1}b \ \ -b]$, then V is nonsingular and it can be verified that $V^{-1}AV = A$ and $V^{-1}b = \left[\begin{smallmatrix} 0 \\ -1 \end{smallmatrix}\right]$.

With this special form of A and b, we have

$$P = \begin{bmatrix} \frac{a_2}{a_1} & 0 \\ 0 & a_2 \end{bmatrix}$$

$f_0 = [0 \ \ 2a_2]$, $A + kbf_0 = \left[\begin{smallmatrix} 0 & -a_1 \\ 1 & a_2(1-2k) \end{smallmatrix}\right]$, $z_e^+ = -A^{-1}b = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ and $z_e^- = \left[\begin{smallmatrix} -1 \\ 0 \end{smallmatrix}\right]$. We also have $f_0A^{-1}b = 0$.

For a given $k > 0.5$, (13) has a unique limit cycle which is the boundary of $\mathcal{S}(k)$. To visualize the proof, $\partial\mathcal{C}$ and $\partial\mathcal{S}(k)$ for some k are plotted in Fig. 6, where the inner closed curve is $\partial\mathcal{S}(k)$, and the outer one is $\partial\mathcal{C}$.

We recall that when the eigenvalues of A are real [see (6)]

$$\partial\mathcal{C} = \left\{\pm\left[e^{-At}z_e^- - \int_0^t e^{-A(t-\tau)}b\,d\tau\right]: t \in [0, \infty]\right\} \tag{23}$$

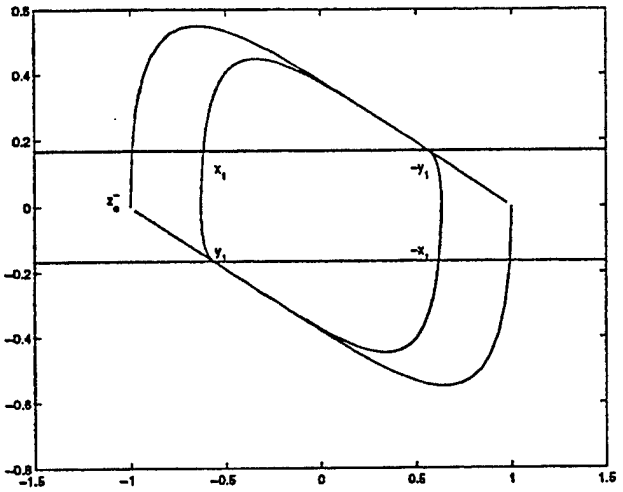


Fig. 6. The domain of attraction and the null controllable region.

and when the eigenvalues of A are complex [see (7)]

$$\partial\mathcal{C} = \left\{ \pm \left[e^{-At} z_s^- - \int_0^t e^{-A(t-\tau)} b \, d\tau \right] : t \in [0, T_p] \right\}. \tag{24}$$

On the other hand, $\partial\mathcal{S}(k)$ is the limit cycle of the time reversed system of (13),

$$\dot{z}(t) = -Az(t) - b\sigma(kf_0 z(t)). \tag{25}$$

Here, the limit cycle as a trajectory goes clockwise. From the proof of Theorem 3.1, we know that the limit cycle is symmetric and has two intersections with $kf_0 z = 1$ and two with $kf_0 z = -1$, see Fig. 6. Let T be the time required for the limit cycle trajectory to go from y_1 to x_1, and T_2 the time from x_1 to $-y_1$, then

$$\begin{aligned} \partial\mathcal{S}(k) = &\left\{ \pm e^{-(A+kbf_0)t} y_1 : t \in [0, T] \right\} \\ &\cup \left\{ \pm \left[e^{-At} x_1 - \int_0^t e^{-A(t-\tau)} b \, d\tau \right] : t \in [0, T_2] \right\}. \end{aligned} \tag{26}$$

Here and in the sequel, the dependence of x_1, y_1, T and T_2 on k is omitted for simplicity.

As $k \to \infty$, the distance between the line $kf_0 z = 1$ and $kf_0 z = -1$ approaches zero. By comparing (23), (24) and (26), we see that to prove the theorem, it suffices to show

$$\lim_{k\to\infty} T = 0, \qquad \lim_{k\to\infty} x_1 = \lim_{k\to\infty} y_1 = z_e^- \text{ (or } z_s^-)$$
$$\lim_{k\to\infty} T_2 = \infty \text{ (or } T_p).$$

In this case, the length of the part of the limit cycle between the lines $kf_0 z = 1$ and $kf_0 z = -1$ will tend to zero. We will first show that $\lim_{k\to\infty} T = 0$.

Let

$$x_1 = \begin{bmatrix} x_{11} \\ \dfrac{1}{2ka_2} \end{bmatrix}, \qquad y_1 = \begin{bmatrix} y_{11} \\ -\dfrac{1}{2ka_2} \end{bmatrix}$$

then $kf_0 x_1 = 1$, $kf_0 y_1 = -1$

$$\begin{bmatrix} x_{11} \\ \dfrac{1}{2ka_2} \end{bmatrix} = e^{-(A+kbf_0)T} \begin{bmatrix} y_{11} \\ -\dfrac{1}{2ka_2} \end{bmatrix} \tag{27}$$

and

$$\left| kf_0 e^{-(A+kbf_0)t} \begin{bmatrix} y_{11} \\ -\dfrac{1}{2ka_2} \end{bmatrix} \right| \le 1, \qquad \forall t \in [0, T].$$

We also note that the upward movement of the trajectory at x_1 and y_1 implies that $x_{11} < (2k-1)/2k$, $y_{11} < (1-2k)/2k$.

As $k \to \infty$, $A + kbf_0 = \left[\begin{smallmatrix} 0 & -a_1 \\ 1 & a_2(1-2k) \end{smallmatrix}\right]$ has two distinct real eigenvalues $-\lambda_1$ and $-\lambda_2$. (Their dependence on k is also omitted.) Assume $\lambda_2 > \lambda_1$. Since $\lambda_1\lambda_2 = a_1$ and $\lambda_1 + \lambda_2 = a_2(2k-1)$, we have $\lim_{k\to\infty} \lambda_1 = 0$, $\lim_{k\to\infty} \lambda_2 = +\infty$.

With the special form of $A + kbf_0$, it can be verified that

$$e^{(A+kbf_0)T} = \begin{bmatrix} \lambda_2 & \lambda_1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} e^{-\lambda_1 T} & 0 \\ 0 & e^{-\lambda_2 T} \end{bmatrix} \begin{bmatrix} \lambda_2 & \lambda_1 \\ 1 & 1 \end{bmatrix}^{-1}.$$

Hence, from (27), we obtain

$$x_{11} = \frac{1}{2ka_2} \frac{\lambda_2 - \lambda_1 + \lambda_2 e^{-\lambda_2 T} - \lambda_1 e^{-\lambda_1 T}}{e^{-\lambda_2 T} - e^{-\lambda_1 T}}$$
$$y_{11} = \frac{1}{2ka_2} \frac{\lambda_2 - \lambda_1 + \lambda_2 e^{\lambda_2 T} - \lambda_1 e^{\lambda_1 T}}{e^{\lambda_1 T} - e^{\lambda_2 T}}.$$

Since $y_{11} < (1-2k)/2k = -(\lambda_1+\lambda_2)/(2ka_2)$ and $e^{\lambda_1 T} - e^{\lambda_2 T} < 0$, we have

$$\lambda_1 e^{\lambda_2 T} < \lambda_2 - \lambda_1 + \lambda_2 e^{\lambda_1 T} < 2\lambda_2 e^{\lambda_1 T}$$

and

$$T < \frac{\ln \dfrac{2\lambda_2}{\lambda_1}}{\lambda_2 - \lambda_1} = \frac{1}{\lambda_2 - \lambda_1} \ln \frac{2\lambda_2^2}{a_1}$$

noting that $\lambda_1 = a_1/\lambda_2$. Since $\lim_{k\to\infty} \lambda_2 = \infty$, we get $\lim_{k\to\infty} T = 0$. It follows that

$$\begin{aligned} \lim_{k\to\infty} \frac{y_{11}}{x_{11}} &= \lim_{k\to\infty} \frac{\lambda_2 - \lambda_1 + \lambda_2 e^{\lambda_2 T} - \lambda_1 e^{\lambda_1 T}}{(\lambda_2 - \lambda_1) e^{(\lambda_1+\lambda_2)T} + \lambda_2 e^{\lambda_1 T} - \lambda_1 e^{\lambda_2 T}} \\ &= \lim_{k\to\infty} \frac{\lambda_2 - \lambda_1 \dfrac{1 + e^{\lambda_1 T}}{1 + e^{\lambda_2 T}}}{\lambda_2 e^{\lambda_1 T} - \lambda_1 \dfrac{e^{\lambda_2 T}(1 + e^{\lambda_1 T})}{1 + e^{\lambda_2 T}}} = 1 \end{aligned}$$

where we have used the fact that $\lim_{k\to\infty} \lambda_1 = 0$. Since x_1 and y_1 are bounded by the null controllable region, we have

$$\lim_{k\to\infty} (y_{11} - x_{11}) = 0. \tag{28}$$

On the limit cycle of (25), we also have

$$-y_1 = e^{-AT_2} x_1 - \int_0^{T_2} e^{-A(T_2-\tau)} b \, d\tau$$

i.e.,

$$\begin{bmatrix} y_{11} \\ -\dfrac{1}{2ka_2} \end{bmatrix} = -e^{-AT_2} \begin{bmatrix} x_{11} \\ \dfrac{1}{2ka_2} \end{bmatrix} + (I - e^{-AT_2}) A^{-1} b$$

$$(I + e^{-AT_2}) \begin{bmatrix} y_{11} \\ 0 \end{bmatrix} = (I - e^{-AT_2}) A^{-1} b + e^{-AT_2} \begin{bmatrix} y_{11} - x_{11} \\ -\dfrac{1}{2ka_2} \end{bmatrix} + \begin{bmatrix} 0 \\ \dfrac{1}{2ka_2} \end{bmatrix}.$$

It follows from (28) that

$$\lim_{k\to\infty}\left\{\begin{bmatrix} y_{11} \\ 0 \end{bmatrix} - (I+e^{-AT_2})^{-1}(I-e^{-AT_2})A^{-1}b\right\} = 0.$$

Hence

$$\lim_{k\to\infty}[0 \quad 1](I+e^{-AT_2})^{-1}(I-e^{-AT_2})A^{-1}b = 0.$$

For different cases, it can be shown from the above equality that

1) if the eigenvalues of A are real, then

$$\lim_{k\to\infty} T_2 = \infty, \qquad \lim_{k\to\infty} y_1 = \lim_{k\to\infty} x_1 = \lim_{k\to\infty}\begin{bmatrix} y_{11} \\ 0 \end{bmatrix} = z_e^-;$$

2) if the eigenvalues of A are complex, then

$$\lim_{k\to\infty} T_2 = T_p, \qquad \lim_{k\to\infty} y_1 = \lim_{k\to\infty} x_1 = \lim_{k\to\infty}\begin{bmatrix} y_{11} \\ 0 \end{bmatrix} = z_s^-.$$

This completes the proof. □

REFERENCES

[1] J. Alvarez, R. Suarez, and J. Alvarez, "Planar linear systems with single saturated feedback," *Syst. Control Lett.*, vol. 20, pp. 319–326, 1993.

[2] D. S. Bernstein and A. N. Michel, "A chronological bibliography on saturating actuators," *Int. J. Robust Nonlin. Control*, vol. 5, pp. 375–380, 1995.

[3] A. T. Fuller, "In-the-large stability of relay and saturating control systems with linear controller," *Int. J. Control*, vol. 10, pp. 457–480, 1969.

[4] T. Hu and Z. Lin, *Control Systems with Actuator Saturation: Analysis and Design.* Boston, MA: Birkhäuser, 2001.

[5] P.-O. Gutman and P. Hagander, "A new design of constrained controllers for linear systems," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 22–33, Jan. 1985.

[6] H. K. Khalil, *Nonlinear Systsems.* New York: MacMillan, 1992.

[7] Z. Lin, "Low gain feedback," in *Lecture Notes in Control and Information Sciences.* New York: Springer-Verlag, 1998, vol. 240.

[8] Z. Lin and A. Saberi, "Semi-global exponential stabilization of linear systems subject to "input saturation" via linear feedbacks," *Syst. Control Lett.*, vol. 21, pp. 225–239, 1993.

[9] ——, "A semi-global low-and-high design for linear systems with input saturation-stabilization and disturbance rejection," *Int. J. Robust Nonlin. Control*, vol. 5, pp. 381–398, 1995.

[10] D. Liu and A. N. Michel, "Dynamical systems with saturation nonlinearities," in *Lecture Notes in Control and Information Sciences.* New York: Springer-Verlag, 1994, vol. 195.

[11] E. D. Sontag, "An algebraic approach to bounded controllability of linear systems," *Int. J. Control*, vol. 39, pp. 181–188, 1984.

[12] R. Suarez, J. Alvarez-Ramirez, and J. Solis-Daun, "Linear systems with bounded inputs: Global stabilization with eigenvalue placement," *Int. J. Robust Nonlin. Control*, vol. 7, pp. 835–845, 1997.

[13] E. D. Sontag and H. J. Sussmann, "Nonlinear output feedback design for linear systems with saturating controls," in *Proc. 29th IEEE Conf. Decision Control*, 1990, pp. 3414–3416.

[14] H. J. Sussmann, E. D. Sontag, and Y. Yang, "A general result on the stabilization of linear systems using bounded controls," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 2411–2425, Dec. 1994.

[15] H. J. Sussmann and Y. Yang, "On the stabilizability of multiple integrators by means of bounded feedback controls," in *Proc. 30th IEEE Conf. Decision Control*, 1991, pp. 70–72.

[16] A. R. Teel, "Global stabilization and restricted tracking for multiple integrators with bounded controls," *Syst. Control Lett.*, vol. 18, pp. 165–171, 1992.

[17] ——, "Linear systems with input nonlinearities: Global stabilization by scheduling a family of H_∞-type controllers," *Int. J. Robust Nonlin. Control*, vol. 5, pp. 399–441, 1995.

[18] G. F. Wredenhagen and P. R. Belanger, "Piecewise-linear LQ control for systems with input constraints," *Automatica*, vol. 30, pp. 403–416, 1994.

[19] J. C. Willems, "Least squares stationary optimal control and the algebraic Riccati equations," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 621–634, Dec. 1971.

On Stabilization and Spectrum Assignment in Periodically Time-Varying Continuous-Time Systems

Joseph J. Yamé and Raymond Hanus

***Abstract*—This note discusses the stabilization and spectrum assignment problems in linear periodically time-varying (LPTV) *continuous-time* systems with *sampled* state or output feedback. The hybrid nature of the overall feedback system in this case imposes some carefulness in handling classical concepts related to purely LPTV continuous-time systems. In particular, this note points out the fact that the stabilization of such systems by periodic feedback gains with *sampled* state or output does not imply the relocation of the original characteristic exponents of the LPTV systems as stated previously in the literature. It is also shown that the concept of monodromy matrix as extended to LPTV hybrid systems has not all the features of a true monodromy matrix.**

***Index Terms*—Characteristic exponents, monodromy matrix, periodic systems, sampled-data feedback.**

I. INTRODUCTION

The foundations of the theory of periodic systems can be traced back to the work of Floquet who first brought the initial time-varying system into a transformed equivalent one with a time-invariant evolution matrix [9]. An important result of Floquet theory states that the stability of a linear periodically time-varying (LPTV) continuous-time system can be inferred from the location of the eigenvalues of the time-invariant matrix of its transformed equivalent. These eigenvalues are called the characteristic exponents or Poincaré exponents of the periodic system. In control engineering, the interest for continuous-time periodic systems is mainly motivated by numerous application-oriented problems such as sampled-data control, multirate digital control, generalized hold design, control of mechanical systems in rotation, etc. A central issue in the control literature centers around the stabilization of LPTV continuous-time systems by periodic controllers and essentially two different approaches have been used for the design of such stabilizing controllers. In the first approach [3], the input to the periodic controller is a continuous-time signal, whereas in the second approach, the input is a sampled signal. In this latter case, an important feature of the overall system is its hybrid continuous/discrete nature. It has been shown that when a LPTV system is controllable, the whole "monodromy matrix" is assignable by periodic feedback gains with sampled state or sampled output feedback [6], [1]. In [4], a further step has been taken by arguing that the *characteristic exponents* are all relocated with this feedback scheme. This note motivated by this last statement discusses issues pertaining to the *relationship* between the stabilization

Manuscript received September 17, 1999; revised April 10, 2000 and November 24, 2000. Recommended by Associate Editor T. Chen.

The authors are with the Service d'Automatique et d'Analyse des Systèmes, Faculté des Sciences Appliquées, Université libre de Bruxelles, 1050 Brussels, Belgium (e-mail: jyame@labauto.ulb.ac.be).

Publisher Item Identifier S 0018-9286(01)05116-9.

0018–9286/01$10.00 © 2001 IEEE

Publication 11

ELSEVIER

Systems & Control Letters 43 (2001) 203–210

SYSTEMS & CONTROL LETTERS

www.elsevier.com/locate/sysconle

Semi-global stabilization with guaranteed regional performance of linear systems subject to actuator saturation

Tingshu Hu[a,*,1], Zongli Lin[a], Yacov Shamash[b]

[a] *Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903, USA*
[b] *Department of Electrical Engineering, State University of New York, Stony Brook, NY 11794, USA*

Received 29 October 1999; received in revised form 15 January 2001

Abstract

For a linear system under a given saturated linear feedback, we propose feedback laws that achieve semi-global stabilization on the null controllable region while preserving the performance of the original feedback law in a fixed region. Here by semi-global stabilization on the null controllable region we mean the design of feedback laws that result in a domain of attraction that includes any a priori given compact subset of the null controllable region. Our design guarantees that the region on which the original performance is preserved would not shrink as the domain of attraction is enlarged by appropriately adjusting the feedback laws. Both continuous-time and discrete-time systems will be considered. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Actuator saturation; Semi-global stabilization; Performance

1. Introduction

We revisit the problem of semi-globally stabilizing a linear system on its null controllable region with saturating actuators. The null controllable region, denoted as $\mathscr{C}$, is the set of states that can be steered to the origin of the state space in a finite time using saturating actuators. The problem of semi-global stabilization on the null controllable region is, for any a priori given set $\mathscr{X}$ that is in the interior of the null controllable region $\mathscr{C}$, to find a stabilizing feedback law $u = F_{\mathscr{X}}(x)$ such that the resulting domain of attraction includes $\mathscr{X}$ as a subset.

This problem has been well studied for systems that are so-called asymptotically null controllable with bounded controls (ANCBC).[2] In particular, it is established in [6,7] that, in both continuous-time and discrete-time, a linear ANCBC system is semi-globally asymptotically stabilizable on its null controllable region by saturated *linear* feedback. We note that in this case, the null controllable region is the entire state space. The key to the possibility of achieving semi-global stabilization on $\mathscr{C}$ by *linear*

* Corresponding author.
E-mail addresses: th7f@virginia.edu (T. Hu), yshamash@notes.sunysb.edu (Y. Shamash).

[1] The work of Tingshu Hu and Zongli Lin was supported in part by the US Office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

[2] A continuous-time [resp. discrete-time] linear system is asymptotically null controllable with bounded controls if it is stabilizable in the usual linear sense and has all its open loop poles in the closed left-half plane [resp. the closed unit disc].

0167-6911/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0167-6911(01)00098-6

feedback is that the open loop system is ANCBC. In general saturated linear feedback cannot achieve semi-global stabilization on $\mathscr{C}$ if the open loop system is not ANCBC, although there have been many attempts to enlarge the domain of attraction by appropriately choosing the linear feedback gains (see, for example, [3] and the references therein).

Our objective in this paper is to construct nonlinear feedback laws that semi-globally stabilize a linear system (not necessarily ANCBC) subject to actuator saturation. This problem has been addressed before. In particular, it was established in [4,5] that, in both continuous-time and discrete-time, a linear system with only two exponentially unstable modes can be semi-globally stabilized on its null controllable region by controllers that switch between two linear feedback laws. By defining these two linear feedback laws on an appropriately constructed invariant set, it is guaranteed that switching would occur at most once. In discrete-time, general systems have been considered in [1] and feedback laws were constructed that achieve semi-global stabilization on the null controllable region. More specifically, a sequence of polygons are constructed that approaches the null controllable region as the number of vertices increases. The vertices divide the polygons into cones. The state feedback laws are then constructed based on the controls that drive the vertices of a polygon to the origin according to which cone the state belongs to.

In this paper we will first consider a general linear system subject to actuator saturation,

$$x(k+1)=Ax(k)+B\sigma(u(k)), \quad x\in\mathbb{R}^n,\ u\in\mathbb{R}^m, \quad (1)$$

where σ is the standard saturation function. With a slight abuse of notation, we use the same symbol to denote both the vector saturation function and the scalar saturation function, i.e., if $v\in\mathbb{R}^m$, then $\sigma(v)=[\sigma(v_1),\sigma(v_2),\ldots,\sigma(v_m)]^{\mathrm{T}}$ and $\sigma(v_i)=\mathrm{sgn}(v_i)\min\{1,|v_i|\}$. We also assume that a feedback law $u=F_0(x)$ has been designed such that the resulting closed-loop system in the absence of the saturation function

$$x(k+1)=Ax(k)+BF_0(x(k)) \quad (2)$$

has the desired performance. We need to study the stability and performance of the actual system in the presence of actuator saturation,

$$x(k+1)=Ax(k)+B\sigma(F_0(x(k))). \quad (3)$$

Let $\mathscr{D}_0$ be an invariant set of the closed-loop system and be inside the linear region of the saturation function: $\{x\in\mathbb{R}^n:\ \|F_0(x)\|_\infty\leqslant 1\}$. For example, a linear state feedback law $u=F_0x$ could be constructed that places the closed-loop poles at certain desired locations and $\mathscr{D}_0$ can be a level set of the form $\{x\in\mathbb{R}^n:\ x^{\mathrm{T}}P_0x\leqslant 1\}$, where $P_0>0$ satisfies

$$(A+BF_0)^{\mathrm{T}}P_0(A+BF_0)-P_0<0. \quad (4)$$

Suppose that $\mathscr{D}_0$ is in the linear region, then it is an invariant set and within $\mathscr{D}_0$, the saturation function does not have an effect and hence the desired closed-loop performance is preserved.

The objective of this paper is to construct feedback laws that semi-globally stabilize the system (1) on its null controllable region and in the mean time preserve the desired closed-loop performance in the region $\mathscr{D}_0$. The structure of our feedback laws is completely different from that of [1]. Instead of resorting to the cones of the polygons which are not invariant sets, we design our controller by combining a sequence of feedback laws $u=F_i(x)$, $i=0,1,\ldots,M$, in a way that the union of the invariant sets corresponding to each of the feedback laws is also an invariant set, which is shown to be in the domain of attraction. By appropriately selecting this sequence of feedback laws, the union of the invariant sets can then be made large enough to enclose any subset in the interior of the null controllable region. This idea was made feasible by the use of the lifting technique, which was used in [2] to provide an alternative proof of the results of [7] mentioned earlier. We will also extend the above results to continuous-time systems.

This paper is organized as follows. In Section 2 we propose a method for expanding the domain of attraction by switching between a finite sequence of feedback laws. This switching design is then used in Section 3 to show that the domain of attraction can be enlarged to include any subset in the interior of the null controllable region. Section 4 extends the results of Section 3 to continuous-time systems. An example is given in Section 5 to illustrate our design results. Finally, a brief concluding remark is made in Section 6.

2. Expansion of the domain of attraction

Let $u=F_i(x)$, $i=0,1,\ldots,M$, be a finite sequence of stabilizing feedback laws. Among these feedback laws, $u=F_0(x)$ can be viewed as the feedback law that

was originally designed to guarantee certain desired closed-loop performance in a given region and the remaining feedback laws have been introduced for the purpose of enlarging the domain of attraction while preserving the regional performance of the original feedback law $u = F_0(x)$.

For each $i = 0, 1, \ldots, M$, let $\mathscr{D}_i$ be an invariant set inside the domain of attraction of the equilibrium $x=0$ of the closed-loop system under feedback law $u = F_i(x)$,

$$x(k+1) = Ax(k) + B\sigma(F_i(x)). \tag{5}$$

Denote

$$\Omega_i = \bigcup_{j=0}^{i} \mathscr{D}_j, \quad i = 0, 1, \ldots, M.$$

Then, $\Omega_0 \subset \Omega_1 \subset \cdots \subset \Omega_M$.

Theorem 1. *For each $i = 0, 1, \ldots, M$, Ω_i is an invariant set inside the domain of attraction of $x=0$ of the closed-loop system*

$$x(k+1) = Ax(k) + B\sigma(G_i(x(k))), \tag{6}$$

where

$$G_i(x) := \begin{cases} F_0(x), & \text{if } x \in \Omega_0, \\ F_1(x), & \text{if } x \in \Omega_1 \backslash \Omega_0, \\ \vdots & \vdots \\ F_i(x), & \text{if } x \in \Omega_i \backslash \Omega_{i-1}. \end{cases} \tag{7}$$

Here we note that, for each $i=1,2,\ldots,M$, $\Omega_i \backslash \Omega_{i-1} = \mathscr{D}_i \backslash \bigcup_{j=0}^{i-1} \mathscr{D}_j$.

Proof. We prove the theorem by induction. The statement is trivially true for $i = 0$. Suppose it is true for $i \geqslant 0$, we need to show that it is also true for $i+1$. Let us write $G_{i+1}(x)$ as

$$G_{i+1}(x) = \begin{cases} G_i(x), & \text{if } x \in \Omega_i, \\ F_{i+1}(x), & \text{if } x \in \Omega_{i+1} \backslash \Omega_i. \end{cases} \tag{8}$$

If $x(0) \in \Omega_i$, then under the feedback $u = G_i(x)$, $x(k) \in \Omega_i$ for all k and $\lim_{k\to\infty} x(k) = 0$. If $x(0) \in \Omega_{i+1} \backslash \Omega_i = \mathscr{D}_{i+1} \backslash \Omega_i$, since $\mathscr{D}_{i+1}$ is inside the domain of attraction under the feedback $u = F_{i+1}(x)$ and Ω_i is a neighborhood of the origin, $x(k)$ will enter Ω_i at some $k_1 < \infty$. After that, the control is switched to $u = G_i(x)$ and by the foregoing argument, we also have $\lim_{k\to\infty} x(k) = 0$. This shows that Ω_{i+1} is inside the domain of attraction.

It is also easy to see that Ω_{i+1} is an invariant set under $u = G_{i+1}(x)$. □

From (7), we see that if $x \in \Omega_0 = \mathscr{D}_0$, then $u = F_0(x)$ is in effect and hence the pre-designed performance is guaranteed on $\mathscr{D}_0$.

For later use in Section 4, it can be verified in a similar way that Theorem 1 is also true for a continuous-time system $\dot{x}(t) = f(x, u, t)$, in particular,

$$\dot{x}(t) = Ax(t) + B\sigma(u(t)), \tag{9}$$

with a set of stabilizing feedback laws $u = F_i(x)$, $i = 0, 1, \ldots, M$. In the context of continuous-time systems, the existence and uniqueness of the solution of the closed-loop system equation is guaranteed by the fact that Ω_i's are invariant sets and nested to each other. In other words, a trajectory starting from a set Ω_i will remain in it. Once it enters a smaller set Ω_j, $j < i$, it will again remain in it.

3. A semi-global stabilization strategy

In this section, we utilize the lifting technique to design a sequence of ellipsoids that cover any prescribed compact subset of the null controllable region. Each ellipsoid is invariant and in the domain of attraction for the lifted closed-loop system under an appropriately chosen linear feedback. This, by Theorem 1, would achieve semi-global stabilization for the lifted system, and hence for the original system.

The null controllable region of (1) at step K, denoted as $\mathscr{C}(K)$, is the set of states that can be steered to the origin in K steps [5]. We see that $x_0 \in \mathscr{C}(K)$ if and only if there exists a control $u(\cdot)$, $\|u(k)\|_\infty \leqslant 1$, $k = 0, 1, \ldots, K-1$, such that

$$A^K x_0 + \sum_{i=0}^{K-1} A^{K-i-1} B u(i) = 0. \tag{10}$$

The null controllable region, denoted as $\mathscr{C}$, is the set of states that can be steered to the origin in a finite number of steps. Clearly, $\mathscr{C} = \bigcup_{K\geqslant 0} \mathscr{C}(K)$ and it can be shown by standard analysis that any compact subset of $\mathscr{C}$ is a subset of $\mathscr{C}(K)$ for some K. For simplicity, we assume that the pair (A, B) is controllable and A is nonsingular. Then there is an integer $n_0 \leqslant n$ such that, for all $K \geqslant n_0$, $\mathscr{C}(K)$ contains the origin in its interior and is bounded.

For a positive integer L, the lifted system of (1) with step L is given by

$$x_L(k+1) = A_L x_L(k) + B_L \sigma(u_L(k)), \tag{11}$$

where

$$x_L(k)=x(kL),\quad u_L(k)=\begin{bmatrix} u(kL) \\ u(kL+1) \\ u(kL+L-1) \end{bmatrix},$$

$$A_L=A^L,\ B_L=[A^{L-1}BA^{L-2}B\ \cdots\ AB\ \ B]. \quad (12)$$

We have more flexibility in the design of a system by using the lifting technique because it allows us to see further the effect of a control law and to consider the combined effect of the control action at several steps.

For a feedback matrix $F\in\mathbb{R}^{mL\times n}$, denote the unsaturated region (linear region) of the closed-loop system

$$x_L(k+1)=A_Lx_L(k)+B_L\sigma(Fx_L(k)) \quad (13)$$

as

$$\mathscr{L}(F):=\{x\in\mathbb{R}^n:\ |f_jx|\leqslant 1, j=1,2,\ldots,mL\},$$

where f_j is the jth row of F. If $L\geqslant n_0$, then there exists an F such that $A_L+B_LF=0$. For such an F, there is a corresponding $\mathscr{L}(F)$ and for all $x_{L0}=x_0\in\mathscr{L}(F)$, $A_Lx_0+B_L\sigma(Fx_0)=(A_L+B_LF)x_0=0$. Hence $\mathscr{L}(F)$ is an invariant set of the lifted system (13) and is inside the domain of attraction.

For a positive definite matrix $P\in\mathbb{R}^{n\times n}$, denote

$$\mathscr{E}(P)=\{x\in\mathbb{R}^n:\ x^{\mathrm{T}}Px\leqslant 1\}.$$

Suppose that $\mathscr{E}(P)\subset\mathscr{L}(F)$, then under the feedback law $u_L=Fx_L$, $\mathscr{E}(P)$ is also an invariant set inside the domain of attraction. Here we are interested in the ellipsoids because they can be generalized to the Lyapunov level sets for the case $A_L+B_LF\neq 0$. We will show that any compact subset of the null controllable region can be covered by the union of a finite set of such ellipsoids.

Lemma 1. *Given an integer $L\geqslant n_0$ and a positive number $\beta<1$, there exists a family of $F_i\in\mathbb{R}^{mL\times n}$, $i=1,2,\ldots,M$, with corresponding positive definite matrices P_i's, such that $A_L+B_LF_i=0$,*

$$\mathscr{E}(P_i)\subset\mathscr{L}(F_i),\quad i=1,2,\ldots,M,$$

and

$$\beta\mathscr{C}(L)\subset\bigcup_{i=1}^{M}\mathscr{E}(P_i),$$

where $\beta\mathscr{C}(L)=\{\beta x:\ x\in\mathscr{C}(L)\}$.

Proof. Let $\partial(\beta\mathscr{C}(L))$ be the boundary of $\beta\mathscr{C}(L)$. Firstly, we show that, there exists an $\varepsilon>0$ such that, for any $x_1\in\partial(\beta\mathscr{C}(L))$, there exist an $F\in\mathbb{R}^{mL\times n}$ and $P>0$ that satisfy

$$A_L+B_LF=0\quad\text{and}\quad\mathscr{B}(x_1,\varepsilon)\subset\mathscr{E}(P)\subset\mathscr{L}(F),$$

where $\mathscr{B}(x_1,\varepsilon)=\{x\in\mathbb{R}^n:\ \|x-x_1\|_2\leqslant\varepsilon\}$.

Let e_ℓ be the unit vector in $\mathbb{R}^n$ whose ℓth element is 1 and other elements are zeros. For simplicity, assume $x_1=\gamma e_1$, otherwise we can use a unitary transformation $x\to Vx$, $V^{\mathrm{T}}V=I$, to satisfy this. Note that a unitary transformation is equivalent to rotating the state space and does not change the shapes of $\mathscr{B}(x_1,\varepsilon)$, $\mathscr{E}(P)$ and $\mathscr{C}(L)$.

Since $x_1=\gamma e_1\in\beta\mathscr{C}(L)$, it follows from (10) and (12) that there exists a u_{L1}, $\|u_{L1}\|_\infty\leqslant\beta$, such that

$$A_L\gamma e_1+B_Lu_{L1}=0. \quad (14)$$

Define

$$\mu=\frac{\max\{\|x\|_2:\ x\in\partial\mathscr{C}(L)\}}{\min\{\|x\|_2:\ x\in\partial\mathscr{C}(L)\}}.$$

Since $L\geqslant n_0$, $\mathscr{C}(L)$ includes the origin in its interior and $\mu<\infty$. It follows that $\gamma e_\ell\in\mu\beta\mathscr{C}(L)$ for all $\ell\geqslant 2$. Therefore, for each $\ell\geqslant 2$, there exists a $u_{L\ell}$, $\|u_{L\ell}\|\leqslant\mu\beta$, such that

$$A_L\gamma e_\ell+B_Lu_{L\ell}=0. \quad (15)$$

Let $F=\{f_{j\ell}\}$ be chosen as

$$F=\frac{1}{\gamma}[u_{L1}\quad u_{L2}\quad\cdots\quad u_{Ln}],$$

then $|f_{j1}|\leqslant\beta/\gamma$ and $|f_{j\ell}|\leqslant\mu\beta/\gamma$ for $\ell=2,\ldots,n$ and $j=1,2,\ldots,mL$. From (14) and (15), we have

$$(A_L+B_LF)e_\ell=A_Le_\ell+\frac{1}{\gamma}B_Lu_{L\ell}=0,\quad\ell=1,2,\ldots,n.$$

This shows that $A_L+B_LF=0$.

Let

$$P=\begin{bmatrix} p_1 & 0 \\ 0 & p_2I_{n-1}\end{bmatrix},$$

where

$$p_1=\frac{1}{\gamma^2}\left(\frac{2\beta}{\beta+1}\right)^2,$$

$$p_2=(n-1)\left(\frac{\beta\mu}{\gamma}\right)^2\left(1-\frac{(\beta+1)^2}{4}\right)^{-1}.$$

Let $\gamma_{\min}=\min\{\|x\|:\ x\in\partial(\beta\mathscr{C}(L))\}$ and

$$\varepsilon=\left(1-\frac{2\beta}{\beta+1}\right)\gamma_{\min}\left\{\max\left(\frac{2\beta}{\beta+1},\frac{2\sqrt{n-1}\beta\mu}{\sqrt{4-(\beta+1)^2}}\right)\right\}^{-1}.$$

Then $\|P^{1/2}\|_2\varepsilon \leqslant 1-2\beta/(\beta+1)$. Note that ε is independent of γ and a particular x_1.

We also have

$$f_jP^{-1}f_j^{\mathrm{T}} = \frac{1}{p_1}f_{j1}^2 + \frac{1}{p_2}\sum_{\ell=2}^{n} f_{j\ell}^2$$

$$\leqslant \frac{1}{p_1}\left(\frac{\beta}{\gamma}\right)^2 + \frac{n-1}{p_2}\left(\frac{\beta\mu}{\gamma}\right)^2 = 1, \quad (16)$$

which implies that $\mathscr{E}(P) \subset \mathscr{L}(F)$. To see this, we verify that, for any $x \in \mathscr{E}(P)$,

$$|f_jx| = |f_jP^{-(1/2)}P^{1/2}x|$$
$$\leqslant (f_jP^{-1}f_j^{\mathrm{T}})^{1/2}(x^{\mathrm{T}}Px)^{1/2} \leqslant 1.$$

For $x \in \mathscr{B}(x_1,\varepsilon)$, we have

$$\|P^{1/2}x\|_2 \leqslant \|P^{1/2}x_1\|_2 + \|P^{1/2}(x-x_1)\|_2$$

$$\leqslant \frac{2\beta}{\beta+1} + \|P^{1/2}\|_2\varepsilon \leqslant 1.$$

This shows that $x^{\mathrm{T}}Px \leqslant 1$ and hence $\mathscr{B}(x_1,\varepsilon) \subset \mathscr{E}(P) \subset \mathscr{L}(F)$.

Because $\partial(\beta\mathscr{C}(L))$ is a compact set, there exists a finite set of $x_i \in \partial(\beta\mathscr{C}(L))$, $i=1,2,\ldots,M$, such that $\partial(\beta\mathscr{C}(L)) \subset \bigcup_{i=1}^{M}\mathscr{B}(x_i,\varepsilon)$. By the foregoing proof, we know that for each $x_i \in \partial(\beta\mathscr{C}(L))$, there exist an F_i and P_i such that $A_L+B_LF_i=0$ and

$$\mathscr{B}(x_i,\varepsilon) \subset \mathscr{E}(P_i) \subset \mathscr{L}(F_i).$$

Hence,

$$\partial(\beta\mathscr{C}(L)) \subset \bigcup_{i=1}^{M}\mathscr{E}(P_i).$$

It then follows that

$$\beta\mathscr{C}(L) \subset \bigcup_{i=1}^{M}\mathscr{E}(P_i).$$

To see this, for any $x \in \beta\mathscr{C}(L)$, let y be an intersection point of $\partial(\beta\mathscr{C}(L))$ with the straight line passing through the origin and x. Hence, $y \in \mathscr{E}(P_{i_0})$ for some i_0. Since $\mathscr{E}(P_{i_0})$ is convex and contains the origin, $x \in \mathscr{E}(P_{i_0})$. □

Remark 1. We would like to point out that, the family of F_i's may contain repeated members with different P_i's. This is the case, for example, when the system has a single input ($m=1$) and the lifting step L is the same as n, the dimension of the state space. In this case, we have only a unique $F_i=-B_L^{-1}A_L$ with $\mathscr{C}(L) \subset \mathscr{L}(F_i)$.

Lemma 1 shows that $\beta\mathscr{C}(L)$ can be covered by a finite number of ellipsoids and within each ellipsoid there is a corresponding linear feedback law such that the state of (11) will be steered to the origin the next step, or equivalently, the state of (1) will be steered to the origin in L steps. Because β can be made arbitrarily close to 1 and L can be made arbitrarily large, any compact subset of $\mathscr{C}$ can be covered by a family of such ellipsoids. It should be noted that as β gets closer to 1, ε will decrease and we need more ellipsoids to cover $\beta\mathscr{C}(L)$, although the determination of these ellipsoids could be technically involved for higher order systems. Also, in the above development, we need to lift the system by L steps to cover $\beta\mathscr{C}(L)$. Actually, the lifting step can be reduced if we replace the dead-beat condition $A_L+B_LF=0$ with a less restrictive one:

$$(A_L+B_LF)^{\mathrm{T}}P(A_L+B_LF) - cP \leqslant 0,$$

where $c \in (0,1)$ specifies the requirement of the convergence rate. A direct consequence of Lemma 1 is

Theorem 2. *Given any compact subset X_0 of $\mathscr{C}$ and a number $c \in (0,1)$, there exist an $L \geqslant 1$ and a family of $F_i \in \mathbb{R}^{mL\times n}$, $i=1,2,\ldots,M$, with corresponding positive definite matrices P_i's, such that*

$$(A_L+B_LF_i)^{\mathrm{T}}P_i(A_L+B_LF_i) - cP_i \leqslant 0, \quad (17)$$

$$\mathscr{E}(P_i) \subset \mathscr{L}(F_i), \quad i=1,2,\ldots,M, \quad (18)$$

and

$$X_0 \subset \bigcup_{i=1}^{M}\mathscr{E}(P_i). \quad (19)$$

Because of (17) and (18), $\mathscr{E}(P_i)$ is an invariant set inside the domain of attraction for the closed-loop system

$$x_L(k+1) = A_Lx_L(k) + B_L\sigma(F_ix_L(k)).$$

By Theorem 1, we can use a switching controller to make $\bigcup_{i=1}^{M}\mathscr{E}(P_i)$ inside the domain of attraction. Once the state enters the region $\mathscr{E}(P_0)$, the controller switches to the feedback law

$$u_L(k) = \bar{F}_0(x_L(k)) = \begin{bmatrix} F_0(x_L(k)) \\ F_0(x(kL+1)) \\ \vdots \\ F_0(x(kL+L-1)) \end{bmatrix}, \quad (20)$$

where the variables $x(kL+i)$, $i=1,2,\ldots,L-1$, can be recursively computed from the state $x_L(k)$ as follows:

$$x(kL+1)=Ax_L(k)+BF_0(x_L(k)),$$

$$\begin{aligned}x(kL+2)&=Ax(kL+1)+BF_0(x(kL+1)\\&=A(Ax_L(k)+BF_0(x_L(k)))\\&\quad+BF_0(Ax_L(k)+BF_0(x_L(k)))\end{aligned}$$

$$\vdots$$

$$x(kL+i+1)=Ax(kL+i)+BF_0(x(kL+i)).$$

Since feedback law (20) corresponds to $u=F_0(x)$ in the original time index, under which $\mathscr{E}(P_0)$ is an invariant set, $\mathscr{E}(P_0)$ is also an invariant set under feedback law (20) in the lifted time index and the desired performance in this region is preserved.

We also observe that, due to the switching and lifting that are involved in the construction of feedback laws, our final semi-globally stabilizing feedback laws, when implemented in the original system (1), are nonlinear and periodic in time.

4. Continuous-time systems

In this section, we consider the continuous-time counterpart of the system (1)

$$\dot{x}(t)=Ax(t)+B\sigma(u(t)),\quad x\in R^n,\ u\in\mathbb{R}^m. \tag{21}$$

The null controllable region at time T, denoted as $\mathscr{C}(T)$, is the set of states that can be steered to the origin in time T by a measurable control input u. The null controllable region, denoted as $\mathscr{C}$, is $\bigcup_{T\geqslant 0}\mathscr{C}(T)$.

Let $h>0$ be the lifting period. We are now interested in controlling the state of (21) at times kh, $k=1,2,\ldots$. Denote $x_h(k)=x(kh)$ and $u_h(k,\tau)=u(kh+\tau)$. Let $A_h=\mathrm{e}^{Ah}$; then the lifted system is

$$x_h(k+1)=A_hx_h(k)+\int_0^h \mathrm{e}^{A(h-\tau)}B\sigma(u_h(k,\tau))\,\mathrm{d}\tau. \tag{22}$$

Denote the set of $m\times n$ dimensional measurable functions defined on $[0,h)$ as $\mathscr{F}^{m\times n}$. With a matrix function $F\in\mathscr{F}^{m\times n}$, let the feedback control be $u_h(k,\tau)=F(\tau)x_h(k)$. Then the closed-loop system is

$$x_h(k+1)=A_hx_h(k)+\int_0^h \mathrm{e}^{A(h-\tau)}B\sigma(F(\tau)x_h(k))\,\mathrm{d}\tau. \tag{23}$$

The unsaturated region of the feedback law is then given by,

$$\begin{aligned}\mathscr{L}(F):=\{x\in\mathbb{R}^n:\ &|f_j(\tau)x|\leqslant 1,\\ &j=1,2,\ldots,m,\ \tau\in[0,h)\},\end{aligned}$$

where $f_j\in\mathscr{F}^{1\times n}$ is the jth row of F. If $x_h(k)\in\mathscr{L}(F)$, then $\sigma(F(\tau)x_h(k))=F(\tau)x_h(k)$ and

$$x_h(k+1)=\left(A_h+\int_0^h \mathrm{e}^{A(h-\tau)}BF(\tau)\,\mathrm{d}\tau\right)x_h(k). \tag{24}$$

The feedback $u_h(k,\tau)=F(\tau)x_h(k)$ is stabilizing if there exists $P>0$ such that

$$\left(A_h+\int_0^h \mathrm{e}^{A(h-\tau)}BF(\tau)\,\mathrm{d}\tau\right)^{\mathrm{T}}$$

$$\times P\left(A_h+\int_0^h \mathrm{e}^{A(h-\tau)}BF(\tau)\,\mathrm{d}\tau\right)-P\leqslant 0.$$

Note that P can be scaled such that $\mathscr{E}(P)\subset\mathscr{L}(F)$. In this case, $\mathscr{E}(P)$ is an invariant set inside the domain of attraction for the system (23). Since for all $x_h(k)\in\mathscr{E}(P)$, the control is linear in $x_h(k)$, so, when $x_h(k)$ tends to the origin, the control $u_h(k,\tau)=F(\tau)x_h(k)$ will gets smaller and hence the state of the original system (21) between $t=kh$ and $t=(k+1)h$ will stay close to $x_h(k)$. Similar to the discrete-time case, we have the following lemma.

Lemma 2. *Given $h>0$ and a positive number $\beta<1$, there exists a family of $F_i\in\mathscr{F}^{m\times n}$, $i=1,2,\ldots,M$, with corresponding positive definite matrices P_i's, such that*

$$A_h+\int_0^h \mathrm{e}^{A(h-\tau)}BF_i(\tau)\,\mathrm{d}\tau=0,$$

$$\mathscr{E}(P_i)\subset\mathscr{L}(F_i),\quad i=1,2,\ldots,M,$$

and

$$\beta\mathscr{C}(h)\subset\bigcup_{i=1}^{M}\mathscr{E}(P_i).$$

Proof. The idea of the proof is the same as that of Lemma 1. Here we just show how to construct ε, F and P for a given $x_1\in\partial(\beta\mathscr{C}(h))$. We also assume that $x_1=\gamma e_1$. Since $\gamma e_1\in\partial(\beta\mathscr{C}(h))$, there exists a $u_1\in\mathscr{F}^{m\times 1}$, $\|u_1(\tau)\|_\infty\leqslant\beta$ for all $\tau\in[0,h)$, such that

$$A_h\gamma e_1+\int_0^h \mathrm{e}^{A(h-\tau)}Bu_1(\tau)\,\mathrm{d}\tau=0,$$

and for $\ell \geqslant 2$, there exists a $u_\ell \in \mathscr{F}^{m\times 1}$, $\|u_\ell(\tau)\|_\infty \leqslant \beta\mu$ for all $\tau \in [0,h)$, such that

$$A_h\gamma e_\ell + \int_0^h e^{A(h-\tau)}Bu_\ell(\tau)\,d\tau = 0.$$

Let $F = 1/\gamma[u_1\ u_2\ \cdots\ u_n]$, and P, ε be the same as those in the proof of Lemma 1, the remaining part of the proof will be the same as that of Lemma 1 except that (16) is replaced with

$$f_j(\tau)P^{-1}f_j^{\mathrm{T}}(\tau) \leqslant 1, \quad \forall\tau \in [0,h),\ j = 1,2,\ldots,m. \ \square$$

The following is the counterpart of Theorem 2 for the discrete-time system (1).

Theorem 3. *Given any compact subset X_0 of $\mathscr{C}$ and a number $c \in (0,1)$, there exist an $h > 0$ and a family of $F_i \in \mathscr{F}^{m\times n}$, $i = 1,2,\ldots,M$, with corresponding positive definite matrices P_i's, such that*

$$\left(A_h + \int_0^h e^{A(h-\tau)}BF_i(\tau)\,d\tau\right)^{\mathrm{T}}$$

$$\times P_i\left(A_h + \int_0^h e^{A(h-\tau)}BF_i(\tau)\,d\tau\right) - cP_i \leqslant 0,$$

$$\mathscr{E}(P_i) \subset \mathscr{L}(F_i), \quad i = 1,2,\ldots,M,$$

and

$$X_0 \subset \bigcup_{i=1}^{M} \mathscr{E}(P_i).$$

Again, by Theorem 1, we can use a switching controller to make $\bigcup_{i=0}^{M}\mathscr{E}(P_i)$ inside the domain of attraction and hence semi-global stabilization can be achieved. Moreover, once the state enters the region $\mathscr{E}(P_0)$, the controller switches to the feedback law $u = F_0(x)$ and hence the desired performance in this region is preserved.

5. Example

Consider the system (1) with

$$A = \begin{bmatrix} 0.8876 & -0.5555 \\ 0.5555 & 1.5542 \end{bmatrix}, \quad B = \begin{bmatrix} -0.1124 \\ 0.5555 \end{bmatrix}.$$

The matrix A is exponentially unstable with a pair of eigenvalues $1.2209 \pm j0.4444$. The LQR controller corresponding to the cost function $J = \sum(x(k)^{\mathrm{T}}Qx(k) + u(k)^{\mathrm{T}}Ru(k))$, with $Q = I$, $R = 1$ is

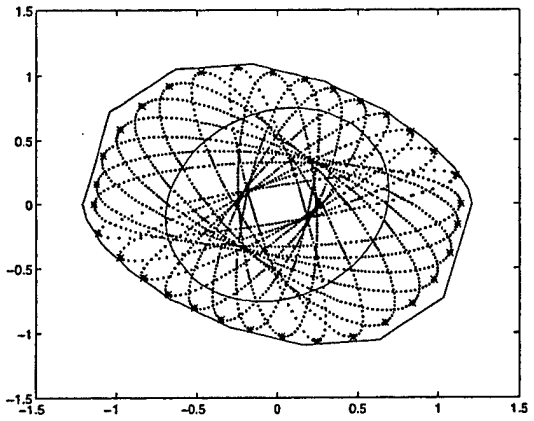


Fig. 1. The union of the invariant ellipsoids.

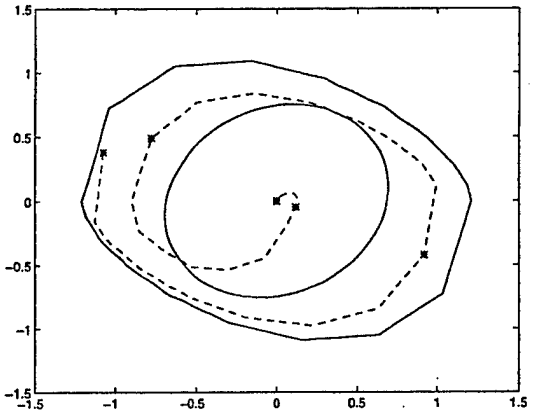


Fig. 2. A trajectory under the multiple switching control.

$u = F_0(x) = [\,-0.2630\ \ -2.1501\,]x$. Let $\mathscr{D}_0$ be obtained as

$$\mathscr{E}(P_0), \quad P_0 = \begin{bmatrix} 2.1367 & -0.2761 \\ -0.2761 & 1.7968 \end{bmatrix},$$

see the ellipsoid enclosed by the solid curve in Fig. 1.

To enlarge the domain of attraction, we take a lifting step of 8 and obtain 16 invariant ellipsoids with corresponding feedback controllers, see the ellipsoids enclosed by the dotted curves in Fig. 1. Each invariant ellipsoid is optimal with respect to certain x_i in the sense that it contains αx_i with $|\alpha|$ maximized, see the points marked with '*'. This is computed by using the LMI method [3]. The outermost curve in Fig. 1 is the boundary of the null controllable region $\mathscr{C}$. We see that the union of the ellipsoids covers a large portion of $\mathscr{C}$.

Figs. 2–4 show some simulation results of the closed-loop system under the multiple switching

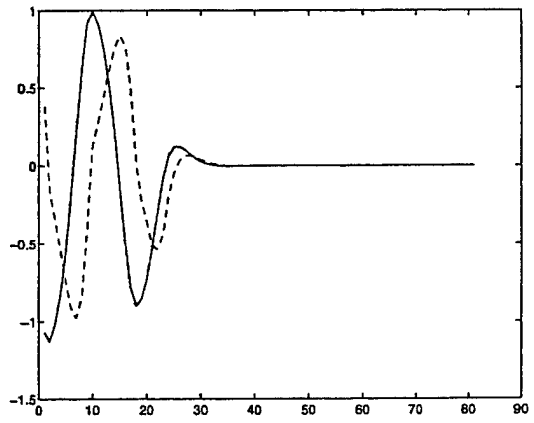


Fig. 3. Simulation: x_1, '-'; x_2, '–'.

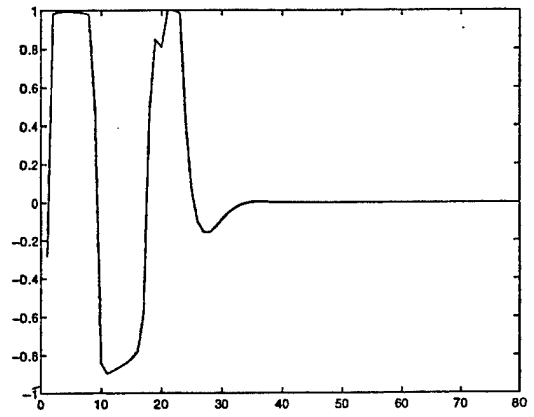


Fig. 4. Simulation: the control.

controls. The initial state is very close to the boundary of $\mathscr{C}$. In Fig. 2 the dashed trajectory is that of the unlifted system (1) under the switched control, and the trajectory of the lifted system is marked with '*'. Figs. 3 and 4 are the state and control of the original unlifted system.

6. Conclusions

In this paper, we have proposed a control design method for linear systems that are subject to actuator saturation. This design method applies to general (possibly exponentially unstable) systems in either continuous-time or discrete-time. The resulting feedback laws expand the domain of attraction achieved by an a priori designed feedback law to include any bounded set in the interior of the null controllable region, while preserving the desired performance of the original feedback law in a fixed region.

References

[1] H. Choi, On the stabilization of linear discrete-time systems subject to input saturation, Sys. & Contr. Lett. 36 (1999) 241–244.

[2] J. Collado, R. Lozano, A. Ailon, Semi-global stabilization of discrete-time systems with bounded inputs using periodic controller, Sys. & Contr. Lett. 36 (1999) 267–275.

[3] T. Hu, Z. Lin, On enlarging the basin of attraction for linear systems under saturated linear feedback, Sys. & Contr. Lett. 40 (1) (2000) 59–69.

[4] T. Hu, Z. Lin, L. Qiu, Stabilization of exponentially unstable linear systems with saturating actuators, IEEE Trans. Auto. Contr., to appear.

[5] T. Hu, D.E. Miller, L. Qiu, Null controllable region of LTI discrete-time systems with input saturation, Automatica, submitted.

[6] Z. Lin, A. Saberi, Semi-global exponential stabilization of linear systems subject to 'input saturation' via linear feedbacks, Sys. & Contr. Lett. 21 (1993) 225–239.

[7] Z. Lin, A. Saberi, Semi-global exponential stabilization of linear discrete-time systems subject to 'input saturation' via linear feedbacks, Sys. & Contr. Lett. 24 (1995) 125–132.

Publication 12

ELSEVIER

Systems & Control Letters 43 (2001) 211–217

SYSTEMS & CONTROL LETTERS

www.elsevier.com/locate/sysconle

Semi-global stabilization of linear systems subject to output saturation ☆

Zongli Lin *, Tingshu Hu

Department of Electrical Engineering, University of Virginia, Thornton Hall Room E313, Charlottesville, VA 22903-2442, USA

Received 3 March 2000; received in revised form 9 September 2000; accepted 23 January 2001

Abstract

It is established that a SISO linear stabilizable and detectable system subject to output saturation can be semi-globally stabilized by linear output feedback if all its *invariant zeros* are in the closed left-half plane, no matter where the open loop poles are. This result complements a recent result that such systems can always be globally stabilized by discontinuous nonlinear feedback laws, and can be viewed as dual to a well-known result: a linear stabilizable and detectable system subject to input saturation can be semi-globally stabilized by linear output feedback if all its *poles* are in the open left-half plane, no matter where the invariant zeros are. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Semi-global stabilization; Saturation; Low gain feedback; High gain feedback

1. Introduction

Physical limitations on actuators and sensors often cause the control input and/or measurement output to saturate. In control design, it is thus necessary to take the effects of input and/or output saturation into account.

While input saturation has been addressed in much detail in the literature (see, for example, [1] and the references therein), fewer results are available that deal with output saturation. For example, issues related to the observability of a linear system subject to output saturation were discussed in detail in [3]. A discontinuous dead beat controller was recently constructed for single input single output (SISO) linear systems in the presence of output saturation [2] that drives every initial state to the origin in a finite time.

In this paper, we consider the problem of semi-globally stabilizing linear systems using *linear* feedback of the saturated output measurement. Here, by semi-global stabilization we mean the construction of a stabilizing feedback law that yields a domain of attraction that contains any *a priori* given (arbitrarily large) bounded set. This problem was motivated by its counterpart for linear systems subject to input saturation [5]. More specifically, it was established in [5] that a linear system subject to input saturation can be semi-globally stabilized using linear feedback if the system is stabilizable and detectable in the usual linear sense and all its open loop *poles* are in the closed left-half plane, no matter where the invariant

☆ Work supported in part by the US office at Naval Research Young Investigator Program under grant N00014-99-1-0670.

* Corresponding author. Tel.: +1-804-924-6342; fax: +1-804-924-8818.

E-mail addresses: zl5y@virginia.edu (Z. Lin), th7f@virginia.edu (T. Hu).

0167-6911/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0167-6911(01)00099-8

zeros are. What we will show in this paper is that a single input single output linear system subject to output saturation can be semi-globally stabilized by linear output feedback if the system is stabilizable and detectable in the usual linear sense and all its *invariant zeros* are in the closed left-half plane, no matter where the open loop poles are. This result thus complements the results of [2] in the sense that, it requires an extra condition that the invariant zeros of the system be on the closed left-half plane to conclude semi-global stabilizability by *linear* feedback. It can also be viewed as dual to its input saturation counterpart in [5]. We, however, note that in the dual situation [5], the condition of all poles being in the closed left-half plane is necessary even with nonlinear feedback [7], while in the current situation, the condition of all invariant zero being in the closed left-half plane is not necessary with nonlinear feedback (by the result of [2]). It is not clear at this time if it would become necessary for linear feedback.

Although this result can be viewed as dual to its input saturation counterpart in [5], the mechanisms behind the stabilizing feedback laws are completely different. In the case of actuator saturation, we construct low gain feedback laws that avoid the saturation of the input signal for all initial states inside the *a priori* given set and the closed-loop system behaves linearly. Here in the case of output saturation, the output matrix is fixed and the output signal is always saturated for large initial states. Once the output is saturated, no information other than its sign is available for feedback. Our linear feedback laws are designed in such a way that they use the saturated output to cause the system output to oscillate into the linear region of output saturation function and remain in there in a finite time. The *same* linear feedback laws then stabilize the system at the origin. This is possible since all the invariant zeros are in the closed left-half plane and the feedback gains can be designed such that the overshoot of the output is arbitrarily small.

The precise problem formulation and the main results are presented in Section 2, which also concludes the paper with some simulation results.

2. Main results

Consider the following single input single output linear system subject to output saturation,

$$\begin{aligned} &\dot{x} = Ax + Bu, \quad x \in \mathbb{R}^n, \quad u \in \mathbb{R}, \\ &y = \sigma(Cx), \quad y \in \mathbb{R}, \end{aligned} \tag{1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is the standard saturation function, i.e., $\sigma(u) = \text{sign}(u)\min\{\alpha, |u|\}$. Our main results on semi-global stabilizability of the system (1) is given in the following theorem.

Theorem 1. *The system* (1) *is semi-globally asymptotically stabilizable by linear feedback of the saturated output if*

- *The pair* (A,B) *is stabilizable;*
- *The pair* (A,C) *is detectable; and*
- *All invariant zeros of the triple* (A,B,C) *are in the closed left-half plane.*

More specifically, for any a priori given bounded set $\mathscr{X}_0 \subset \mathbb{R}^{2n}$, *there exists a linear dynamic output feedback law of the form*

$$\begin{aligned} &\dot{z} = Fz + Gy, \quad z \in \mathbb{R}^n, \\ &u = Hz + H_0 y, \end{aligned} \tag{2}$$

such that the equilibrium $(x,z) = (0,0)$ *of the closed-loop system is asymptotically stable with* $\mathscr{X}_0$ *contained in its domain of attraction.*

Proof. We will establish this result in two steps. In the first step, we will construct a family of feedback laws of the form (2), parameterized in $\varepsilon \in (0,1]$. In the second step, we will show that, for any a priori given bounded set $\mathscr{X}_0 \subset \mathbb{R}^{2n}$, there exists an $\varepsilon^* \in (0,1]$ such that, for each $\varepsilon \in (0,\varepsilon^*]$, the equilibrium $(x,z)=(0,0)$ of the closed-loop system is asymptotically stable with $\mathscr{X}_0$ contained in its domain of attraction.

The construction of the feedback laws follows the following algorithm.

Step 1. Find a state transformation [6],

$$x = T\bar{x}, \quad \bar{x} = [x_0^{\mathrm{T}}\ x_1^{\mathrm{T}}]^{\mathrm{T}}, \quad x_1 = [x_{11}\ x_{12}\ \cdots\ x_{1r}]^{\mathrm{T}},$$

such that the system can be written in the following form:

$$\begin{aligned} &\dot{x}_0 = A_0 x_0 + B_0 x_{11}, \quad x_0 \in \mathbb{R}^{n_0}, \\ &\dot{x}_{11} = x_{12}, \\ &\dot{x}_{12} = x_{13}, \\ &\vdots \\ &\dot{x}_{1r} = C_0 x_0 + a_1 x_{11} + a_2 x_{12} + \cdots + a_r x_{1r} + u, \\ &y = \sigma(x_{11}), \end{aligned} \tag{3}$$

where (A_0, B_0) is stabilizable and the eigenvalues of A_0 are the invariant zeros of the triple (A, B, C) and hence are all in the closed left-half plane.

We note that the multiple input multiple output counterpart of the above canonical form will in general also require a transformation on the input and the output. The later cannot be performed due to the presence of output saturation.

Step 2. Let $F_0(\varepsilon)$ be such that

$$\lambda(A_0+B_0F_0(\varepsilon))=\{-\varepsilon+\lambda_0(A_0)\}\cup\lambda_-(A_0),\quad \varepsilon\in(0,1],$$

where $\lambda_0(A_0)$ and $\lambda_-(A_0)$ denote respectively the sets of eigenvalues of A_0 that are on the imaginary axis and in the open left-half plane. It is clear that $A_0+B_0F_0(\varepsilon)$ is Hurwitz for any $\varepsilon\in(0,1]$ and

$$\|F_0(\varepsilon)\|\leqslant\alpha_0\varepsilon,\quad \forall\varepsilon\in(0,1], \tag{4}$$

for some α_0 independent of ε.

Such an $F_0(\varepsilon)$ exists since (A_0, B_0) is stabilizable. We summarize some properties for the triple $(A_0, B_0, F_0(\varepsilon))$ from [4, Lemmas 2.2.3 and 2.2.4 and Theorem 3.3.1].

Lemma 1. *For the given triple $(A_0, B_0, F_0(\varepsilon))$, there exists a nonsingular matrix $T_0(\varepsilon)\in\mathbb{R}^{n_0\times n_0}$ such that*

$$\|T_0(\varepsilon)\|\leqslant\tau_0, \tag{5}$$

$$\|F_0(\varepsilon)T_0^{-1}(\varepsilon)\|\leqslant\beta_0\varepsilon, \tag{6}$$

$$\|F_0(\varepsilon)A_0T_0^{-1}(\varepsilon)\|\leqslant\beta_1\varepsilon, \tag{7}$$

$$T_0(\varepsilon)(A_0+B_0F_0(\varepsilon))T_0^{-1}(\varepsilon)=J_0(\varepsilon), \tag{8}$$

where τ_0, β_0 and β_1 are some constants independent of ε and $J_0(\varepsilon)\in\mathbb{R}^{n_0\times n_0}$ is a real matrix. Moreover, there exists a $P_0>0$, independent of ε, such that

$$J_0^{\mathrm{T}}(\varepsilon)P_0+P_0J_0(\varepsilon)\leqslant-\frac{\varepsilon}{2}I. \tag{9}$$

Step 3. Let L be such that $A+LC$ is Hurwitz. Such an L exists since the pair (A, C) is detectable.

Step 4. Construct the family of output feedback laws as follows:

$$\dot z=Az+Bu+L(Cz-y),$$

$$u=-C_0z_0-\sum_{i=1}^{r}a_iz_{1i}-\frac{\alpha_1}{\varepsilon^r}(y-F_0(\varepsilon)z_0) \tag{10}$$

$$-\frac{\alpha_2}{\varepsilon^{r-1}}z_{12}-\cdots-\frac{\alpha_r}{\varepsilon}z_{1r},$$

where z_0 and z_{1i}, $i=1,2,\ldots,r$, are defined as follows:

$$z=T\tilde z,\quad \tilde z=[z_0^{\mathrm{T}}\ \ z_1^{\mathrm{T}}]^{\mathrm{T}},\quad z_1=[z_{11}\ \ z_{12}\ \cdots\ z_{1r}]^{\mathrm{T}},$$

and α_i's are chosen such that

$$s^r+\alpha_rs^{r-1}+\alpha_{r-1}s^{r-2}+\cdots+\alpha_2s+\alpha_1=(s+1)^r,$$

i.e.,

$$\alpha_i=C_r^{i-1}=\frac{r!}{(i-1)!(r-i+1)!},\quad i=1,2,\ldots,r.$$

We now proceed with the second step of the proof: to show that, for any a priori given bounded set $\mathscr{X}_0\subset\mathbb{R}^{2n}$, there exists an $\varepsilon^*\in(0,1]$ such that, for each $\varepsilon\in(0,\varepsilon^*]$, the equilibrium $(x,z)=(0,0)$ of the closed-loop system is asymptotically stable with $\mathscr{X}_0$ contained in its domain of attraction. Without loss of generality, let us assume that the system is already in the form of (3), i.e., $T=I$. Letting $e=x-z$, we can write the closed-loop system as follows:

$$\begin{aligned}
&\dot x_0=A_0x_0+B_0x_{11},\\
&\dot x_{11}=x_{12},\\
&\dot x_{12}=x_{13},\\
&\vdots\\
&\dot x_{1r}=C_0(x_0-z_0)+a_1(x_{11}-z_{11})+a_2(x_{12}-z_{12})\\
&\qquad+\cdots+a_r(x_{1r}-z_{1r})-\frac{\alpha_1}{\varepsilon^r}(y-F_0(\varepsilon)z_0)\\
&\qquad-\frac{\alpha_2}{\varepsilon^{r-1}}z_{12}-\cdots-\frac{\alpha_r}{\varepsilon}z_{1r},\\
&\dot z=Az+L(Cz-y)+B[-C_0z_0-a_1z_{11}\\
&\qquad-\cdots-a_rz_{1r}-\frac{\alpha_1}{\varepsilon^r}(y-F_0(\varepsilon)z_0)\\
&\qquad-\frac{\alpha_2}{\varepsilon^{r-1}}z_{12}-\cdots-\frac{\alpha_r}{\varepsilon}z_{1r}],\\
&y=\sigma(x_{11}).
\end{aligned} \tag{11}$$

We next define a new set of state variables as follows:

$$\begin{aligned}
&\tilde x_0=T_0(\varepsilon)x_0,\\
&\tilde x_{11}=x_{11}-F_0(\varepsilon)x_0,\\
&\tilde x_{1i}=\varepsilon^{i-1}x_{1i}+C_{i-1}^1\varepsilon^{i-2}x_{1i-1}+C_{i-1}^2\varepsilon^{i-3}x_{1i-2}+\cdots\\
&\qquad+C_{i-1}^{i-2}\varepsilon x_{12}+C_{i-1}^{i-1}(x_{11}-F_0(\varepsilon)x_0),\\
&\qquad i=2,3,\ldots,r,\\
&e_0=x_0-z_0,\\
&e_{1i}=x_{1i}-z_{1i},\quad i=1,2,\ldots,r,
\end{aligned} \tag{12}$$

and denote

$$e = [e_0^{\mathrm{T}} \quad e_{11} \quad e_{12} \quad \cdots \quad e_{1r}]^{\mathrm{T}}.$$

With these new state variables, the closed-loop system can be written as follows:

$$\begin{aligned}
\dot{\tilde{x}}_0 &= J_0(\varepsilon)\tilde{x}_0 + T_0(\varepsilon)B_0\tilde{x}_{11},\\
\dot{\tilde{x}}_{11} &= -\frac{1}{\varepsilon}\tilde{x}_{11} + \frac{1}{\varepsilon}\tilde{x}_{12} - [F_0(\varepsilon)A_0T_0^{-1}(\varepsilon)\\
&\quad + F_0(\varepsilon)B_0F_0(\varepsilon)T_0^{-1}(\varepsilon)]\tilde{x}_0 - F_0(\varepsilon)B_0\tilde{x}_{11},\\
\dot{\tilde{x}}_{12} &= -\frac{1}{\varepsilon}\tilde{x}_{12} + \frac{1}{\varepsilon}\tilde{x}_{13} - [F_0(\varepsilon)A_0T_0^{-1}(\varepsilon)\\
&\quad + F_0(\varepsilon)B_0F_0(\varepsilon)T_0^{-1}(\varepsilon)]\tilde{x}_0 - F_0(\varepsilon)B_0\tilde{x}_{11},\\
&\vdots\\
\dot{\tilde{x}}_{1r-1} &= -\frac{1}{\varepsilon}\tilde{x}_{1r-1} + \frac{1}{\varepsilon}\tilde{x}_{1r} - [F_0(\varepsilon)A_0T_0^{-1}(\varepsilon)\\
&\quad + F_0(\varepsilon)B_0F_0(\varepsilon)T_0^{-1}(\varepsilon)]\tilde{x}_0 - F_0(\varepsilon)B_0\tilde{x}_{11},\\
\dot{\tilde{x}}_{1r} &= -\frac{1}{\varepsilon}\tilde{x}_{1r} + \frac{1}{\varepsilon}[x_{11} - \sigma(x_{11})] - \frac{1}{\varepsilon}F_0(\varepsilon)e_0\\
&\quad + \varepsilon^{r-1}[C_0e_0 + a_1e_{11} + a_2e_{12} + \cdots + a_re_{1r}]\\
&\quad + \alpha_2e_{12} + \alpha_3\varepsilon e_{13} + \cdots + \alpha_r\varepsilon^{r-2}e_{1r}\\
&\quad - [F_0(\varepsilon)A_0T_0^{-1}(\varepsilon) + F_0(\varepsilon)B_0F_0(\varepsilon)T_0^{-1}(\varepsilon)]\tilde{x}_0\\
&\quad - F_0(\varepsilon)B_0\tilde{x}_{11},\\
\dot{e} &= (A + LC)e - L[x_{11} - \sigma(x_{11})].
\end{aligned} \tag{13}$$

Choose a Lyapunov function candidate as follows:

$$V(\tilde{x}_0, \tilde{x}_{11}, \ldots, \tilde{x}_{1r}, e) = v\tilde{x}_0^{\mathrm{T}}P_0\tilde{x}_0 + \sum_{i=1}^{r}\tilde{x}_{1i}^2 + \sqrt{\varepsilon}e^{\mathrm{T}}Pe, \tag{14}$$

where $v \in (0, 1]$, independent of ε, is a constant whose value is to be determined later, P_0 is as defined in Lemma 1, and $P > 0$ is such that

$$(A + LC)^{\mathrm{T}}P + P(A + LC) = -I. \tag{15}$$

Let $c > 0$, independent of ε, be such that

$$c \geqslant \sup_{(x,z)\in\mathscr{X}_0, \varepsilon\in(0,1], v\in(0,1]} V(\tilde{x}_0, \tilde{x}_{11}, \ldots, \tilde{x}_{1r}, e). \tag{16}$$

Such a c exists due to the boundedness of $\mathscr{X}_0$ and the definition of the state variables as given by (12). With this choice of c, it is obvious that $(x, z) \in \mathscr{X}_0$ implies that $(\tilde{x}_0, \tilde{x}_{11}, \ldots, \tilde{x}_{1r}) \in L_V(c) := \{(\tilde{x}_0, \tilde{x}_{11}, \ldots, \tilde{x}_{1r}, e) \in \mathbb{R}^{2n}: V \leqslant c\}$.

Using Lemma 1, we can calculate the derivative of V inside the level set $L_V(c)$ along the trajectories of the closed-loop system (13) as follows:

$$\begin{aligned}
\dot{V} &= -v\tilde{x}_0^{\mathrm{T}}\tilde{x}_0 + 2v\tilde{x}_0^{\mathrm{T}}P_0T_0(\varepsilon)B_0\tilde{x}_{11}\\
&\quad + \sum_{i=1}^{r}\left[-\frac{2}{\varepsilon}\tilde{x}_{1i}^2 + \frac{2}{\varepsilon}\tilde{x}_{1i}\tilde{x}_{1i+1} - 2\tilde{x}_{1i}[F_0(\varepsilon)A_0T_0^{-1}(\varepsilon)\right.\\
&\quad \left. + F_0(\varepsilon)B_0F_0(\varepsilon)T_0^{-1}(\varepsilon)]\tilde{x}_0 - 2\tilde{x}_{1i}F_0(\varepsilon)B_0\tilde{x}_{11}\right]\\
&\quad + 2x_{1r}\left[\frac{1}{\varepsilon}[x_{11} - \sigma(x_{11})] - \frac{1}{\varepsilon}F_0(\varepsilon)e_0\right.\\
&\quad + \varepsilon^{r-1}[C_0e_0 + a_1e_{11} + \cdots + a_re_{1r}]\\
&\quad \left. + \alpha_2e_{12} + \alpha_3\varepsilon e_{13} + \cdots + \alpha_r\varepsilon^{r-2}e_{1r}\right] - \sqrt{\varepsilon}e^{\mathrm{T}}e\\
&\quad - 2\sqrt{\varepsilon}e^{\mathrm{T}}PL[x_{11} - \sigma(x_{11})]\\
&\leqslant -v\tilde{x}_0^{\mathrm{T}}\tilde{x}_0 + 2\delta_{01}v\|\tilde{x}_0\||\tilde{x}_{11}| + \sum_{i=1}^{r}\left[-\frac{2}{\varepsilon}\tilde{x}_{1i}^2\right.\\
&\quad \left. + \frac{2}{\varepsilon}\tilde{x}_{1i}\tilde{x}_{1i+1} + 2\delta_{i0}\varepsilon|\tilde{x}_{1i}|\|\tilde{x}_0\| + 2\delta_{i1}\varepsilon|\tilde{x}_{1i}|\|\tilde{x}_{11}|\right]\\
&\quad + \frac{2}{\varepsilon}|x_{1r}|\|x_{11} - \sigma(x_{11})| + 2\eta_1|\tilde{x}_{1r}|\|e\| - \sqrt{\varepsilon}e^{\mathrm{T}}e\\
&\quad + 2\eta_2\sqrt{\varepsilon}\|e\||x_{11} - \sigma(x_{11})|,
\end{aligned} \tag{17}$$

where δ_{ij}'s and η_i's are some constants, independent of ε.

We will continue our evaluation of $\dot{V}$ by considering two separated cases, $|x_{11}| \leqslant 1$ and $|x_{11}| > 1$.

Case 1. $|x_{11}| \leqslant 1$. In this case, we have,

$$\begin{aligned}
\dot{V} &\leqslant -v\tilde{x}_0^{\mathrm{T}}\tilde{x}_0 + 2\delta_{01}v\|\tilde{x}_0\||\tilde{x}_{11}|\\
&\quad - \frac{1}{\varepsilon}[\tilde{x}_{11} \quad \tilde{x}_{12} \quad \cdots \quad \tilde{x}_{1r-1} \quad \tilde{x}_{1r}]\\
&\quad \times \begin{bmatrix} 2 & -1 & \cdots & 0 & 0\\ -1 & 2 & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & 2 & -1\\ 0 & 0 & \cdots & -1 & 2 \end{bmatrix} \begin{bmatrix} \tilde{x}_{11}\\ \tilde{x}_{12}\\ \vdots\\ \tilde{x}_{1r-1}\\ \tilde{x}_{1r} \end{bmatrix}\\
&\quad + \sum_{i=1}^{r}[2\delta_{i0}\varepsilon|\tilde{x}_{1i}|\|\tilde{x}_0\| + 2\delta_{i1}\varepsilon|\tilde{x}_{1i}|\|\tilde{x}_{11}|]
\end{aligned}$$

$$
\begin{aligned}
&+2\eta_1|\tilde{x}_{1r}|\|e\| - \sqrt{\varepsilon}e^{\mathrm{T}}e \\
&\leqslant -\left[\nu - \left(\sum_{i=1}^{r}\delta_{i0}\right)\varepsilon^2 - \delta_{01}\nu^2\right]\|\tilde{x}_0\|^2 \\
&\quad -\left[\frac{\delta_1}{\varepsilon} - \delta_{01} - \delta_{10} - 2\delta_{11}\varepsilon - \left(\sum_{i=2}^{r}\delta_{i1}\right)\varepsilon\right]\tilde{x}_{11}^2 \\
&\quad -\sum_{i=2}^{r-1}\left[\frac{\delta_1}{\varepsilon} - \delta_{i0} - \delta_{i1}\varepsilon\right]|\tilde{x}_{1i}|^2 \\
&\quad -\left[\frac{\delta_1}{\varepsilon} - \delta_{r1}\varepsilon - \frac{2\eta_1^2}{\sqrt{\varepsilon}}\right]\tilde{x}_{1r}^2 - \frac{\sqrt{\varepsilon}}{2}\|e\|^2, \qquad (18)
\end{aligned}
$$

where we have used the fact that the matrix

$$
\begin{bmatrix}
2 & -1 & \cdots & 0 & 0 \\
-1 & 2 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 2 & -1 \\
0 & 0 & \cdots & -1 & 2
\end{bmatrix},
$$

is positive definite with its maximum eigenvalue denoted as $\delta_1 > 0$.

Let ν be such that $\nu \leqslant 1/2\delta_{01}$ and let $\varepsilon_1^* \in (0,1]$ be such that the following hold for all $\varepsilon \in (0,\varepsilon_1^*]$:

$$
\begin{aligned}
&\nu - \left(\sum_{i=1}^{r}\delta_{i0}\right)\varepsilon^2 - \delta_{01}\nu^2 \geqslant \frac{\nu}{4}, \\
&\frac{\delta_1}{\varepsilon} - \delta_{01} - \delta_{10} - 2\delta_{11}\varepsilon - \left(\sum_{i=2}^{2}\delta_{i1}\right)\varepsilon \geqslant \frac{\delta_1}{2\varepsilon}, \\
&\frac{\delta_1}{\varepsilon} - \delta_{i0} - \delta_{i1}\varepsilon \geqslant \frac{\delta_1}{2\varepsilon}, \quad i = 2,3,\ldots,r-1, \\
&\frac{\delta_1}{\varepsilon} - \delta_{r1}\varepsilon - \frac{2\eta_1^2}{\sqrt{\varepsilon}} \geqslant \frac{\delta_1}{2\varepsilon}. \qquad (19)
\end{aligned}
$$

With these choices of ν and ε_1^*, we conclude that, for any $|x_{11}| \leqslant 1$,

$$
\dot{V} \leqslant -\frac{\nu}{4}\|\tilde{x}_0\|^2 - \frac{\delta_1}{2\varepsilon}\sum_{i=1}^{r}|\tilde{x}_{1i}|^2 - \frac{\sqrt{\varepsilon}}{2}\|e\|^2, \quad \varepsilon \in (0,\varepsilon_1^*]. \qquad (20)
$$

Case 2. $|x_{11}| > 1$. In this case, we have,

$$
\begin{aligned}
\dot{V} &\leqslant -\nu\tilde{x}_0^{\mathrm{T}}\tilde{x}_0 + 2\delta_{01}\nu\|\tilde{x}_0\||\tilde{x}_{11}| \\
&\quad -\frac{1}{\varepsilon}[\tilde{x}_{11}^2 - (|x_{11}|-1)^2] - \sqrt{\varepsilon}e^{\mathrm{T}}e \\
&\quad +\sum_{i=1}^{r}[(\delta_{i0}+\delta_{i1})\varepsilon\tilde{x}_{1i}^2 + \delta_{i0}\varepsilon\|\tilde{x}_0\|^2 + \delta_{i1}\varepsilon\tilde{x}_{11}^2] \\
&\quad +\eta_1\tilde{x}_{1r}^2 + (\eta_1 + \eta_2\sqrt{\varepsilon})\|e\|^2 \\
&\quad +\eta_2\sqrt{\varepsilon}(|x_{11}|-1)^2. \qquad (21)
\end{aligned}
$$

Now let $\varepsilon_2^* \in (0,1]$ be such that, for all $\varepsilon \in (0,\varepsilon_2^*]$, $(\tilde{x}_0,\tilde{x}_{11},\ldots,\tilde{x}_{1r},e) \in L_V(c)$ implies that

$$
\begin{aligned}
&|F_0(\varepsilon)x_0| \leqslant \tfrac{1}{2}, \\
&2\delta_{01}\nu\|\tilde{x}_0\||\tilde{x}_{11}| + \sum_{i=1}^{r}[(\delta_{i0}+\delta_{i1})\varepsilon\tilde{x}_{1i}^2 + \delta_{i0}\varepsilon\|\tilde{x}_0\|^2 \\
&\quad +\delta_{i1}\varepsilon\tilde{x}_{11}^2] + \eta_1\tilde{x}_{1r}^2 + (\eta_1 + \eta_2\sqrt{\varepsilon})\|e\|^2 \\
&\quad +\eta_2\sqrt{\varepsilon}(|x_{11}|-1)^2 \leqslant \frac{1}{8\varepsilon}. \qquad (22)
\end{aligned}
$$

The first inequality is due to (4) and implies that

$$
\tilde{x}_{11}^2 - (|x_{11}|-1)^2 \geqslant \tfrac{1}{4}.
$$

With this choice of ε_2^*, we have that, for any $|x_{11}| > 1$,

$$
\dot{V} \leqslant -\nu\tilde{x}_0^{\mathrm{T}}\tilde{x}_0 - \sqrt{\varepsilon}e^{\mathrm{T}}e - \frac{1}{8\varepsilon}, \quad \varepsilon \in (0,\varepsilon_2^*]. \qquad (23)
$$

Combining Cases 1 and 2, we conclude that, for any $\varepsilon \in (0,\varepsilon^*]$ with $\varepsilon^* = \min\{\varepsilon_1^*,\varepsilon_2^*\}$,

$$
\dot{V} < 0, \quad \forall(\tilde{x}_0,\tilde{x}_{11},\tilde{x}_{12},\ldots,\tilde{x}_{1e},e) \in L_V(c) \setminus \{0\}, \qquad (24)
$$

which, in turn, shows that the equilibrium $(x,z)=(0,0)$ of the closed-loop system is asymptotically stable with $\mathscr{X}_0$ contained in its domain of attraction. □

In what follows, we will use a simple example to demonstrate the closed-loop system behavior. Consider the system (1) with

$$
A = \begin{bmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0
\end{bmatrix}, \quad
B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad
C = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}^{\mathrm{T}}.
$$

It can be easily verified that this system is controllable and observable with an invariant zero at $s = 0$. The open loop poles are located at $\{-1,\pm j,1\}$. Following the design algorithm we proposed above, we construct a family of parameterized output feedback laws as

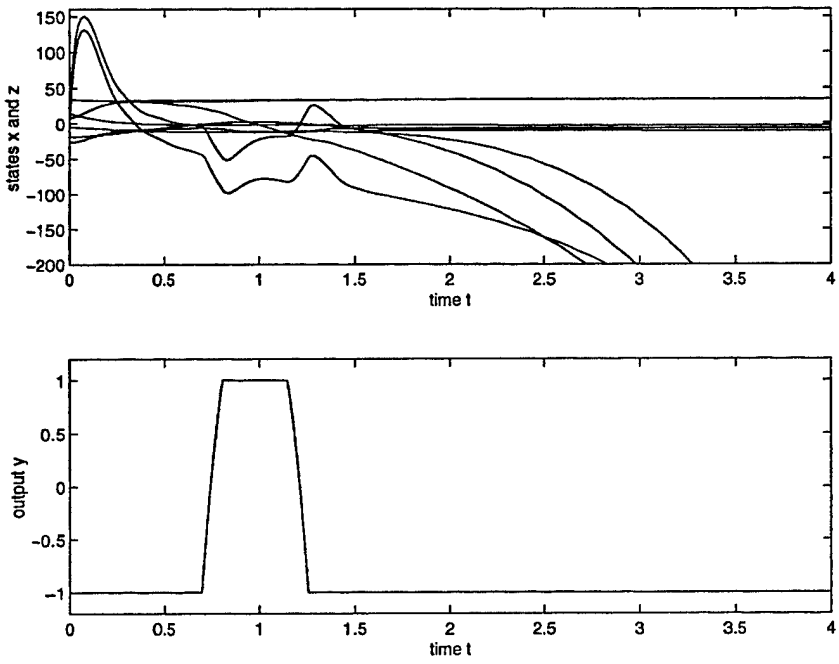


Fig. 1. $\varepsilon = 0.1$.

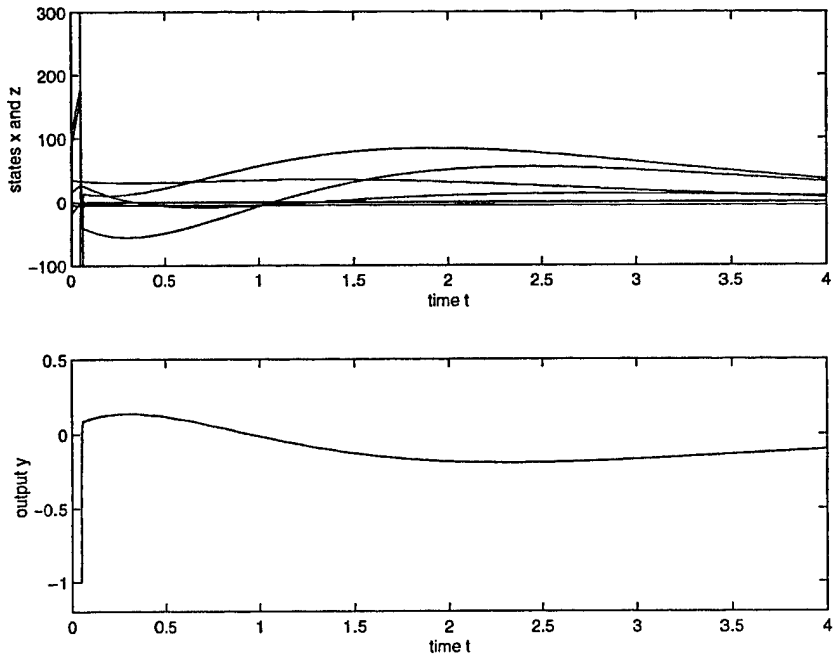


Fig. 2. $\varepsilon = 0.001$.

follows:

$$\begin{aligned}
\dot{z}_1 &= z_2 - 2(z_2 - y),\\
\dot{z}_2 &= z_3 - 4(z_2 - y),\\
\dot{z}_3 &= z_4 - 6(z_2 - y),\\
\dot{z}_4 &= z_1 - 4(z_2 - y) + u,\\
u &= -z_1 - \frac{1}{\varepsilon^3}(y + \varepsilon z_1) - \frac{3}{\varepsilon^2}z_3 - \frac{3}{\varepsilon}z_4.
\end{aligned} \tag{25}$$

Some simulation results are shown in Figs. 1 and 2. In the simulation, initial conditions are taken randomly as $[-4.5625, -18.8880, 8.2065, -4.0685, 34.5569, 15.4136, -25.4760, 15.8338]^T$. In Fig. 1, ε is chosen to be $\varepsilon = 0.1$. It is clear that with this choice of ε, the initial conditions are not inside the domain of attraction. In Fig. 2, ε is chosen to be $\varepsilon = 0.001$. It is clear that, the output is out of saturation after some time and the closed-loop system become linear and all its states converge to zero. This demonstrates that as ε decreases, the domain of attraction is enlarged.

References

[1] D.S. Bernstein, A.N. Michel, A chronological bibliography on saturating actuators, Internat. J. Robust Nonlinear Control 5 (1995) 375–380.

[2] G. Kreisselmeier, Stabilization of linear systems in the presence of output measurement saturation, Systems Control Lett. 29 (1996) 27–30.

[3] R.B. Koplon, M.L.J. Hautus, E.D. Sontag, Observability of linear systems with saturated outputs, Linear Algebra Appl. 205–206 (1994) 909–936.

[4] Z. Lin, Low Gain Feedback, Springer, London, 1998.

[5] Z. Lin, A. Saberi, Semi-global exponential stabilization of linear discrete-time systems subject to input saturation via linear feedbacks, Systems Control Lett. 24 (1995) 125–132.

[6] P. Sannuti, A. Saberi, A special coordinate basis of multivariable linear systems—Finite and infinite zero structure, squaring down and decoupling, Internat. J. Control 45 (1987) 1655–1704.

[7] E.D. Sontag, An algebraic approach to bounded controllability of linear systems, Internat. J. Control 39 (1984) 181–188.

Publication 13

In general of course, the objective is to have a short test interval. However, there is a tradeoff between the separability index and the length of the test period. As we have seen, γ^* is an increasing function of T.

VII. Conclusion

We have presented a methodology for error-free system identification in the situation where we have two candidate linear models subject to bounded energy noise, and where we have control over the input. The problem of selecting a best input signal over a test period (the minimum proper auxiliary signal design problem) has been solved and a solution given in terms of the solution to a boundary value system. The solution of this boundary value system also enables us to design a very efficient on-line identification scheme, the hyperplane test, that takes into account the fact that the input signal over the test period is known in advance.

A related problem which can be solved with the methodology presented here is the shortest test period problem where we fix the separability index and look for the shortest test period for perfect identification. The procedure is similar to the one presented in this paper, except the γ-iteration part should be replaced with T-iteration.

There are a number of possible extensions to the work presented here. One is to allow for additional inputs to the system, i.e., the models would have another input u, in addition to the auxiliary signal v, but this one measured online, just like the output y. This situation has been considered in [9].

Another possible extension is to allow for some nonlinearity. In particular, if the system is not linear in v, the Kalman filter implementation and the hyperplane test still apply. The problem is the optimization over v. An approach using optimization software is discussed in [3].

References

[1] P. Bolzern, P. Colaneri, and G. De Nicolao, "H_∞-differential Ricatti equations: Convergence properties and finite escape phenomena," *IEEE Trans. Automat. Contr.*, vol. 42, pp. 113–118, Jan. 1997.

[2] C. Bunks, J. P. Chancelier, F. Delebecque, M. Goursat, R. Nikoukhah, and S. Steer, *Engineering and Scientific Computing With Scilab*, C. Gomez, Ed. Boston, MA: Birkhauser, 1999.

[3] S. L. Campbell, K. G. Horton, R. Nikoukhah, and F. Delebecque, "Rapid model selection and the separability index," in *Proc. 4th IFAC Symp. Fault Detection Supervision Safety Technical Processes (SAFEPROCESS 2000)*, Budapest, Hungary, pp. 1187–1192.

[4] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*. Philadelphia, PA: SIAM, 1999.

[5] A. Gelb, *Applied Optimal Estimation*. Cambridge, MA: MIT Press, 1984.

[6] F. Kerestecioğlu, *Change Detection and Input Design in Dynamical Systems*. Taunton, U.K.: Research Studies, 1993.

[7] F. Kerestecioğlu and M. B. Zarrop, "Input design for detection of abrupt changes in dynamical systems," *Int. J. Control*, vol. 59, no. 4, pp. 1063–1084, 1994.

[8] R. Nikoukhah, "Guaranteed active failure detection and isolation for linear dynamical systems," *Automatica*, vol. 34, no. 11, pp. 1345–1358, 1998.

[9] R. Nikoukhah, S. L. Campbell, and F. Delebecque, "Detection signal design for failure detection: A robust approach," *Int. J. Adap. Control Signal Processing*, vol. 14, pp. 701–724, 2000.

[10] R. Nikoukhah, S. L. Campbell, K. G. Horton, and F. Delebecque, "Auxiliary signal design for robust multimodel identification,", INRIA Rep., num. 4000, 2000.

[11] I. R. Petersen and A. V. Savkin, *Robust Kalman Filtering for Signals and Systems With Large Uncertainties*. Boston, MA: Birkhauser, 1999.

[12] K. Uosaki, I. Tanaka, and H. Sugiyama, "Optimal input design for autoregressive model discrimination with constrained output variance," *IEEE Trans. Automat. Contr.*, vol. AC-29, pp. 348–350, Apr. 1984.

[13] X. J. Zhang, *Auxiliary Signal Design in Fault Detection and Diagnosis*. Heidelberg, Germany: Springer-Verlag, 1989.

Exact Characterization of Invariant Ellipsoids for Single Input Linear Systems Subject to Actuator Saturation

Tingshu Hu and Zongli Lin

***Abstract*—We present a necessary and sufficient condition for an ellipsoid to be an invariant set of a linear system under a saturated linear feedback. The condition is given in terms of linear matrix inequalities (LMIs) and can be easily used for optimization based analysis and design.**

***Index Terms*—Actuator saturation, contractive invariance, invariant ellipsoid.**

I. Introduction

In this paper, we will study the set invariance property for a linear system under saturated feedback

$$\dot{x} = Ax + B\text{sat}(Fx). \tag{1}$$

We will be interested in invariant ellipsoids due to the popularity of the quadratic Lyapunov function and the simplicity of the results. In the literature, invariant ellipsoids have been used to estimate the domain of attraction for nonlinear systems (see, e.g., [1]–[4], [12]–[14], [16], and the references therein). The problem of estimating the domain of attraction for (1) has been a focus of study in recent years.

For a matrix $F \in \mathbf{R}^{m\times n}$, denote the ith row of F as f_i and define

$$\mathcal{L}(F) := \{x \in \mathbf{R}^n : |f_i x| \le 1,\ i = 1, 2, \ldots, m\}.$$

If F is a feedback gain matrix, then $\mathcal{L}(F)$ is the region where the feedback control $u = \text{sat}(Fx)$ is linear in x. We call $\mathcal{L}(F)$ the linear region of the saturated feedback $\text{sat}(Fx)$, or simply, the linear region of saturation.

Let $P \in \mathbf{R}^{n\times n}$ be a positive–definite matrix. For a positive number ρ, denote

$$\mathcal{E}(P, \rho) = \left\{x \in \mathbf{R}^n : x^{\mathrm{T}} P x \le \rho\right\}.$$

If

$$(A + BF)^{\mathrm{T}} P + P(A + BF) < 0$$

and $\mathcal{E}(P, \rho) \subset \mathcal{L}(F)$, then $\mathcal{E}(P, \rho)$ is an invariant ellipsoid inside the domain of attraction. The largest of these $\mathcal{E}(P, \rho)$s was used as an estimate of the domain of attraction in the earlier literature (see e.g., [4], [15]). This estimation method is simple, yet could be very conservative. Recent efforts have been made to extend the ellipsoid beyond the linear region $\mathcal{L}(F)$ (see, e.g., [7], [5], [6], [12], and [14]). In particular, simple and general methods have been derived by applying the absolute stability analysis tools, such as the circle and Popov criteria, where the saturation is treated as a locally sector bounded nonlinearity.

More recently, we developed a new sufficient condition for an ellipsoid to be invariant in [10] (see also [8]). It was shown that this condition is less conservative than the existing conditions resulting from the circle criterion or the vertex analysis. The most important feature of this new condition is that it can be expressed as LMI's in terms of

Manuscript received February 11, 2001; revised September 1, 2001. Recommended by Associate Editor V. Balakrishnan. This work was supported in part by the US Office of Naval Research Young Investigator Program under Grant N00014-99-1-0670.

The authors are with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22903 USA (e-mail: th7f@virginia.edu; zl5y@virginia.edu).

Publisher Item Identifier S 0018-9286(02)01111-X.

0018–9286/02$17.00 © 2002 IEEE

all the varying parameters and hence can easily be used for controller synthesis. In this paper, we will further show that for the single input case, this condition is also necessary.

Notation: In this paper, we use sat: $\mathbf{R} \to \mathbf{R}$ to denote the standard saturation function, i.e., $\text{sat}(u) = \text{sign}(u)\min\{1, |u|\}$.

II. Review of the Existing Results

Consider the linear system with a single saturating input

$$\dot{x} = Ax + Bu, \quad x \in \mathbf{R}^n, \quad u \in \mathbf{R}, \quad |u|_\infty \le 1. \tag{2}$$

Under a saturated linear feedback $u = \text{sat}(Fx)$, the closed-loop system is

$$\dot{x} = Ax + B\text{sat}(Fx). \tag{3}$$

Given a positive definite matrix P, let $V(x) = x^{\mathrm{T}}Px$. The ellipsoid $\mathcal{E}(P, \rho)$ is said to be (contractively) invariant if

$$\dot{V}(x) = 2x^{\mathrm{T}}P(Ax + B\text{sat}(Fx)) \le (<)0$$

for all $x \in \mathcal{E}(P, \rho)\backslash\{0\}$. Clearly, if $\mathcal{E}(P, \rho)$ is contractively invariant, then it is inside the domain of attraction.

A. Conditions for Set Invariance

We developed a sufficient condition for set invariance in [10] (see also [8]) for multi-input systems. For single input systems, the condition can be stated as follows.

Theorem 1: Consider system (3). Given an ellipsoid $\mathcal{E}(P, \rho)$, suppose that

$$(A + BF)^{\mathrm{T}}P + P(A + BF) < 0. \tag{4}$$

If there exists an $H \in \mathbf{R}^{1\times n}$ such that

$$(A + BH)^{\mathrm{T}}P + P(A + BH) < 0 \tag{5}$$

and $\mathcal{E}(P, \rho) \subset \mathcal{L}(H)$, i.e., $|Hx| \le 1, \forall x \in \mathcal{E}(P, \rho)$, then $\mathcal{E}(P, \rho)$ is a contractively invariant set and hence inside the domain of attraction.

For ρ sufficiently small, we always have $\mathcal{E}(P, \rho) \subset \mathcal{L}(F)$. Hence, (4) is a necessary condition. Before determining the contractive invariance of an ellipsoid $\mathcal{E}(P, \rho)$, it is necessary to make sure that (4) is satisfied. Since (5) is very simple and easily tractable, it is desirable that (5) is also a necessary condition. Although we will finally show that this is indeed the case, the proof of the necessity is much more involved than the sufficiency. We have to approach it through some more existing results.

In [8], we obtained a necessary and sufficient condition for set invariance.

Theorem 2: Given an ellipsoid $\mathcal{E}(P, \rho)$ and an $F \in \mathbf{R}^{1\times n}$, suppose that

$$(A + BF)^{\mathrm{T}}P + P(A + BF) < 0. \tag{6}$$

Then $\mathcal{E}(P, \rho)$ is contractively invariant under $u = \text{sat}(Fx)$ if and only if there exists a function $h(x)$: $\mathbf{R}^n \to \mathbf{R}$, $|h(x)| \le 1$ for all $x \in \mathcal{E}(P, \rho)$, such that $\mathcal{E}(P, \rho)$ is contractively invariant under the control $u = h(x)$, i.e.,

$$x^{\mathrm{T}}P(Ax + Bh(x)) < 0, \qquad \forall x \in \mathcal{E}(P, \rho)\setminus\{0\}. \tag{7}$$

B. Existence of Control Law for Set Invariance

The condition of Theorem 2, the existence of $h(x)$ satisfying (7), was studied in [8], [11]. The results are summarized as follows.

Fact 1: The following two statements are equivalent:

a) the ellipsoid $\mathcal{E}(P, \rho)$ can be made contractively invariant for

$$\dot{x} = Ax + Bu \tag{8}$$

with a bounded control $u = h(x)$, $|h(x)| \le 1$;

b) the ellipsoid $\mathcal{E}(P, \rho)$ is contractively invariant for (8) under the control $u = -\text{sign}(B^{\mathrm{T}}Px)$, i.e., the following condition is satisfied,

$$\dot{V}(x) = x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2x^{\mathrm{T}}PB\,\text{sign}(B^{\mathrm{T}}Px) < 0, \quad \forall x \in \mathcal{E}(P, \rho)\setminus\{0\}. \tag{9}$$

For an arbitrary $P > 0$, there may exist no $\rho > 0$ such that $\mathcal{E}(P, \rho)$ can be made contractively invariant with a bounded control $u = h(x)$, $|h(x)| \le 1$. In the sequel, we simply say that $\mathcal{E}(P, \rho)$ can be made contractively invariant. Let P be given, the following proposition gives a condition for the existence of $\rho > 0$ such that $\mathcal{E}(P, \rho)$ can be made contractively invariant.

Proposition 1: For a given positive definite matrix P, $\mathcal{E}(P, \rho)$ can be made contractively invariant for some $\rho > 0$ if and only if there exists a $k > 0$ such that

$$A^{\mathrm{T}}P + PA - kPBB^{\mathrm{T}}P < 0. \tag{10}$$

Note that if (10) is satisfied for $k = k_0$, then it is satisfied for all $k > k_0$.

Theorem 3: Given $P > 0$, assume that $\mathcal{E}(P, \rho)$ can be made contractively invariant for some $\rho > 0$. Now we consider a given ρ. Let $\lambda_1, \lambda_2, \ldots, \lambda_{\mathcal{J}} > 0$ be real numbers such that

$$\det\begin{bmatrix} \lambda_j P - A^{\mathrm{T}}P - PA & P \\ \rho^{-1}PBB^{\mathrm{T}}P & \lambda_j P - A^{\mathrm{T}}P - PA \end{bmatrix} = 0 \tag{11}$$

and

$$B^{\mathrm{T}}P(A^{\mathrm{T}}P + PA - \lambda_j P)^{-1}PB > 0. \tag{12}$$

Then, $\mathcal{E}(P, \rho)$ is contractively invariant under the control $u = -\text{sign}(B^{\mathrm{T}}Px)$ if and only if

$$\lambda_j\rho - B^{\mathrm{T}}P(A^{\mathrm{T}}P + PA - \lambda_j P)^{-1}PB < 0, \quad \forall j = 1, 2, \ldots, \mathcal{J}. \tag{13}$$

If there exists no $\lambda_j > 0$ satisfying (11) and (12), then $\mathcal{E}(P, \rho)$ is contractively invariant.

Here we note that all the λ_js satisfying (11) are the eigenvalues of the matrix shown at the bottom of the page. Hence the condition of Theorem 3 can be easily checked. It is shown in [8] and [11] that there is a $\rho^* > 0$ such that $\mathcal{E}(P, \rho)$ is contractively invariant if and only if $\rho < \rho^*$. Therefore, the maximum value ρ^* can be obtained by checking the condition of Theorem 3 using a bisection method.

For the single input case, if P is fixed, then we can combine Theorems 2 and 3 to find the largest ρ such that $\mathcal{E}(P, \rho)$ is invariant under a given saturated feedback $u = \text{sat}(Fx)$. However, if P is an unknown parameter for optimization (for example, in the design problem of enlarging the domain of attraction), then the condition of Theorem 3 would not be easy to deal with. For this reason, we would like to use Theorem 1 since its condition leads to LMI constraints (see, e.g., [2]). The only concern is that we might not find the optimal invariant ellipsoid since the condition of Theorem 1 was only shown to be sufficient. This paper is intended to

$$\begin{bmatrix} P^{-(1/2)}A^{\mathrm{T}}P^{1/2} + P^{1/2}AP^{-(1/2)} & -I \\ -\rho^{-1}P^{1/2}BB^{\mathrm{T}}P^{1/2} & P^{-(1/2)}A^{\mathrm{T}}P^{1/2} + P^{1/2}AP^{-(1/2)} \end{bmatrix}.$$

close this gap. We will show in the next section that for the single input case, the condition in Theorem 1 is also necessary. If P is fixed, then the largest ρ satisfying the condition of Theorem 1 is the same as the largest ρ satisfying the condition of Theorem 3.

III. Main Results

Theorem 4: Given an ellipsoid $\mathcal{E}(P, \rho)$ and an $F \in \mathbf{R}^{1\times n}$, assume that

$$(A+BF)^{\mathrm{T}}P+P(A+BF)<0. \tag{14}$$

Then $\mathcal{E}(P, \rho)$ is contractively invariant under the feedback $u = \mathrm{sat}(Fx)$ if and only if there exists an $H \in \mathbf{R}^{1\times n}$ such that

$$(A+BH)^{\mathrm{T}}P+P(A+BH)<0 \tag{15}$$

and $\mathcal{E}(P, \rho) \subset \mathcal{L}(H)$.

Recall from [8], [9], $\mathcal{E}(P, \rho) \subset \mathcal{L}(H)$ is equivalent to

$$\rho HP^{-1}H^{\mathrm{T}} \le 1.$$

Lemma 1: Suppose that $\mathcal{E}(P, \rho)$ can be made contractively invariant for some $\rho > 0$. Let ρ^* be the supremum of all the ρs such that $\mathcal{E}(P, \rho)$ is contractively invariant under $u = -\mathrm{sign}(B^{\mathrm{T}}Px)$ (i.e., satisfying the conditions in Theorem 3). Then there exists a $\lambda > 0$ such that

$$\det\begin{bmatrix}\lambda P-A^{\mathrm{T}}P-PA & P\\ \dfrac{1}{\rho^*}PBB^{\mathrm{T}}P & \lambda P-A^{\mathrm{T}}P-PA\end{bmatrix}=0 \tag{16}$$

and

$$\lambda\rho^* = B^{\mathrm{T}}P(A^{\mathrm{T}}P+PA-\lambda P)^{-1}PB. \tag{17}$$

Proof: Let $V(x) = x^{\mathrm{T}}Px$. Under the control $u = -\mathrm{sign}(B^{\mathrm{T}}Px)$, the derivative of $V(x)$ along the trajectory of the closed-loop system is

$$\dot V(x) = x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x - 2x^{\mathrm{T}}PB\,\mathrm{sign}(B^{\mathrm{T}}Px).$$

Let

$$g(x) = x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x - 2x^{\mathrm{T}}PB.$$

Define

$$\gamma(\rho) = \max\left\{g(x): x^{\mathrm{T}}Px=\rho,\ B^{\mathrm{T}}Px\ge 0\right\}.$$

It is shown in [8] and [11] that if $\gamma(\rho) < 0$, then $\gamma(\rho_1) < 0$ for all $\rho_1 \in (0, \rho)$. Hence, $\mathcal{E}(P, \rho)$ is contractively invariant under $u = -\mathrm{sign}(B^{\mathrm{T}}Px)$ if and only if $\gamma(\rho) < 0$. Since the function $g(x)$ is uniformly continuous on any compact set, $\gamma(\rho)$ is continuous in ρ. By the definition of ρ^*, we have

$$\gamma(\rho) < 0, \qquad \forall \rho \in (0, \rho^*)$$

and $\gamma(\rho^*) = 0$. It is clear that $\mathcal{E}(P, \rho^*)$ is invariant but not contractively invariant. By Theorem 3, there exists a $\lambda > 0$ satisfying (16) and

$$B^{\mathrm{T}}P(A^{\mathrm{T}}P+PA-\lambda P)^{-1}PB>0 \tag{18}$$

and

$$\lambda\rho^* - B^{\mathrm{T}}P(A^{\mathrm{T}}P+PA-\lambda P)^{-1}PB\ge 0. \tag{19}$$

It can be shown with algebraic manipulation (see the Appendix) that (16) is equivalent to

$$B^{\mathrm{T}}P(A^{\mathrm{T}}P+PA-\lambda P)^{-1}P(A^{\mathrm{T}}P+PA-\lambda P)^{-1}PB^{\mathrm{T}}=\rho^*. \tag{20}$$

We claim that only "=" is possible in (19). Otherwise, if we let $x = (A^{\mathrm{T}}P+PA-\lambda P)^{-1}PB$, then from (20), we have $x^{\mathrm{T}}Px = \rho^*$, and from (18), we have $B^{\mathrm{T}}Px > 0$. Thus

$$\begin{aligned} g(x) &= x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x - 2x^{\mathrm{T}}PB\\ &= \lambda x^{\mathrm{T}}Px - x^{\mathrm{T}}PB\\ &= \lambda\rho^* - B^{\mathrm{T}}P(A^{\mathrm{T}}P+PA-\lambda P)^{-1}PB>0. \end{aligned}$$

This means that $\gamma(\rho^*) > 0$, which is a contradiction. Therefore, we must have

$$\lambda\rho^* - B^{\mathrm{T}}P(A^{\mathrm{T}}P+PA-\lambda P)^{-1}PB=0.$$

□

Lemma 2: Suppose that $\mathcal{E}(P, \rho)$ can be made contractively invariant for some $\rho > 0$. Let ρ^* be defined in Lemma 1 with λ satisfying

$$\det\begin{bmatrix}\lambda P-A^{\mathrm{T}}P-PA & P\\ \dfrac{1}{\rho^*}PBB^{\mathrm{T}}P & \lambda P-A^{\mathrm{T}}P-PA\end{bmatrix}=0 \tag{21}$$

and

$$\lambda\rho^* = B^{\mathrm{T}}P(A^{\mathrm{T}}P+PA-\lambda P)^{-1}PB. \tag{22}$$

Let

$$H_0 = -\frac{1}{\rho^*}B^{\mathrm{T}}P(A^{\mathrm{T}}P+PA-\lambda P)^{-1}P$$

then $\rho^* H_0P^{-1}H_0^{\mathrm{T}} = 1$, i.e., $\mathcal{E}(P, \rho^*) \subset \mathcal{L}(H_0)$ and

$$(A+BH_0)^{\mathrm{T}}P+P(A+BH_0)\le 0. \tag{23}$$

Proof: Since $\mathcal{E}(P, \rho)$ can be made contractively invariant for some $\rho > 0$, by Proposition 1, there exists a $k_0 > 0$ such that

$$A^{\mathrm{T}}P+PA-kPBB^{\mathrm{T}}P<0 \tag{24}$$

for all $k > k_0$.

From Fact 2 in the Appendix, (21) is equivalent to

$$B^{\mathrm{T}}P(A^{\mathrm{T}}P+PA-\lambda P)^{-1}P(A^{\mathrm{T}}P+PA-\lambda P)^{-1}PB=\rho^* \tag{25}$$

it follows that $\rho^* H_0P^{-1}H_0^{\mathrm{T}} = 1$, which implies that $\mathcal{E}(P, \rho^*) \subset \mathcal{L}(H_0)$.

We next proceed to prove (23). Consider a state transformation of the form $z = Tx$, where

$$T = U\left(\frac{P}{\rho^*}\right)^{1/2}$$

for some unitary matrix $U \in \mathbf{R}^{n\times n}$, $UU^{\mathrm{T}} = I$. U is to be specified later. Let

$$\overline{B} = TB, \quad \overline{A} = TAT^{-1}, \quad \overline{H}_0 = H_0T^{-1}$$

and

$$Q = \overline{A}^{\mathrm{T}} + \overline{A}.$$

Then

$$\begin{aligned}
&\overline{B}^{\mathrm{T}}(Q-\lambda I)^{-1}(Q-\lambda I)^{-1}\overline{B} \\
&\quad = B^{\mathrm{T}}\left(\frac{P}{\rho^*}\right)^{1/2}U^{\mathrm{T}} \\
&\qquad \cdot\left(U\left(\frac{P}{\rho^*}\right)^{-(1/2)}A^{\mathrm{T}}\left(\frac{P}{\rho^*}\right)^{1/2}U^{\mathrm{T}}\right. \\
&\qquad\quad \left.+\,U\left(\frac{P}{\rho^*}\right)^{1/2}A\left(\frac{P}{\rho^*}\right)^{-(1/2)}U^{\mathrm{T}}-\lambda I\right)^{-1}UU^{\mathrm{T}} \\
&\qquad \cdot\left(U\left(\frac{P}{\rho^*}\right)^{-(1/2)}A^{\mathrm{T}}\left(\frac{P}{\rho^*}\right)^{1/2}U^{\mathrm{T}}+U\left(\frac{P}{\rho^*}\right)^{1/2}\right. \\
&\qquad\quad \left.\cdot\,A\left(\frac{P}{\rho^*}\right)^{-(1/2)}U^{\mathrm{T}}-\lambda I\right)^{-1}U\left(\frac{P}{\rho^*}\right)^{1/2}B \\
&\quad = B^{\mathrm{T}}\left(\frac{P}{\rho^*}\right)^{1/2}\left(P^{-(1/2)}A^{\mathrm{T}}P^{1/2}\right. \\
&\qquad\qquad \left.+\,P^{1/2}AP^{-(1/2)}-\lambda I\right)^{-1} \\
&\qquad \cdot\left(P^{-(1/2)}A^{\mathrm{T}}P^{1/2}\right. \\
&\qquad\quad \left.+P^{1/2}AP^{-(1/2)}-\lambda I\right)^{-1}\left(\frac{P}{\rho^*}\right)^{1/2}B \\
&\quad = \frac{1}{\rho^*}B^{\mathrm{T}}P(A^{\mathrm{T}}P+PA-\lambda P)^{-1} \\
&\qquad \cdot P(A^{\mathrm{T}}P+PA-\lambda P)^{-1}PB=1,
\end{aligned}$$

where the last "=" follows from (25). For easy reference, we rewrite the above equation as

$$\overline{B}^{\mathrm{T}}(Q-\lambda I)^{-1}(Q-\lambda I)^{-1}\overline{B}=1. \tag{26}$$

Similarly, from (22), we obtain

$$\overline{B}^{\mathrm{T}}(Q-\lambda I)^{-1}\overline{B}=\lambda. \tag{27}$$

Also, we have

$$\overline{H}_0=-\overline{B}^{\mathrm{T}}(Q-\lambda I)^{-1}. \tag{28}$$

From (26) and (28), we have $\overline{H}_0\overline{H}_0^{\mathrm{T}}=1$. Recall that

$$\overline{H}_0=H_0T^{-1}=H_0\left(\frac{P}{\rho^*}\right)^{-(1/2)}U^{\mathrm{T}}.$$

By choosing U, we can make

$$-\overline{H}_0=\overline{B}^{\mathrm{T}}(Q-\lambda I)^{-1}=\begin{bmatrix}1 & 0_{1\times(n-1)}\end{bmatrix}. \tag{29}$$

Partition $\overline{B}$ and Q as follows

$$\overline{B}=\begin{bmatrix}b_1\\ b_2\end{bmatrix},\quad Q=\begin{bmatrix}q_1 & q_{21}^{\mathrm{T}}\\ q_{21} & Q_2\end{bmatrix},\qquad b_1,\,q_1\in\mathbf{R}.$$

From (29), we have

$$\overline{B}^{\mathrm{T}}=\begin{bmatrix}1 & 0_{1\times(n-1)}\end{bmatrix}(Q-\lambda I).$$

It follows that:

$$q_1-\lambda=b_1,\qquad q_{21}=b_2.$$

From (27) and (29), we have $b_1=\lambda$. In summary

$$\overline{B}=\begin{bmatrix}\lambda\\ b_2\end{bmatrix},\qquad Q=\begin{bmatrix}2\lambda & b_2^{\mathrm{T}}\\ b_2 & Q_2\end{bmatrix}.$$

Multiplying (24) from left with $(T^{-1})^{\mathrm{T}}$ and from right with T^{-1}, we obtain

$$Q-k\rho^*\overline{B}\overline{B}^{\mathrm{T}}<0$$

for all $k>k_0$. That is

$$\begin{bmatrix}k\rho^*\lambda^2-2\lambda & (k\rho^*\lambda-1)b_2^{\mathrm{T}}\\ (k\rho^*\lambda-1)b_2 & k\rho^*b_2b_2^{\mathrm{T}}-Q_2\end{bmatrix}>0,\qquad \forall k>k_0.$$

By Schur complements, this implies

$$\begin{aligned}
k\rho^*b_2b_2^{\mathrm{T}}-Q_2&>0,\\
k\rho^*b_2b_2^{\mathrm{T}}-Q_2-\frac{(k\rho^*\lambda-1)^2}{k\rho^*\lambda^2-2\lambda}b_2b_2^{\mathrm{T}}&>0,\qquad \forall k>k_0.
\end{aligned}$$

The second inequality can be rewritten as

$$-\frac{1}{k\rho^*\lambda^2-2\lambda}b_2b_2^{\mathrm{T}}-Q_2>0.$$

Since this inequality is true for all $k>k_0$, we must have $Q_2\le 0$.

Let's finally examine the term $(A+BH_0)^{\mathrm{T}}P+P(A+BH_0)$. Recall that in the new coordinate

$$\overline{H}_0=-\begin{bmatrix}1 & 0_{1\times(n-1)}\end{bmatrix}.$$

Hence

$$(\overline{A}+\overline{B}\,\overline{H}_0)^{\mathrm{T}}+\overline{A}+\overline{B}\,\overline{H}_0=Q-\begin{bmatrix}2\lambda & b_2^{\mathrm{T}}\\ b_2 & 0\end{bmatrix}=\begin{bmatrix}0 & 0\\ 0 & Q_2\end{bmatrix}\le 0$$

which is equivalent to

$$(T^{-1})^{\mathrm{T}}(A+BH_0)T^{\mathrm{T}}+T(A+BH_0)T^{-1}\le 0.$$

Multiplying both sides from left with T^{T} and from right with T, we have

$$(A+BH_0)T^{\mathrm{T}}T+T^{\mathrm{T}}T(A+BH_0)\le 0.$$

Noting that $T^{\mathrm{T}}T=P/\rho^*$, we finally obtain (23). □

Proof of Theorem 4: The sufficiency follows from Theorem 1. Now we prove the necessity. That is, suppose that $\mathcal{E}(P,\rho)$ is contractively invariant, then there exists an $H\in\mathbf{R}^{1\times n}$ such that

$$(A+BH)^{\mathrm{T}}P+P(A+BH)<0$$

and $\mathcal{E}(P,\rho)\subset\mathcal{L}(H)$.

Here, we have

$$(A+BF)^{\mathrm{T}}P+P(A+BF)<0.$$

Define $H(\alpha)=(1-\alpha)H_0+\alpha F$, where H_0 is defined in Lemma 2. It follows from (23) in Lemma 2 that:

$$(A+BH(\alpha))^{\mathrm{T}}P+P(A+BH(\alpha))<0 \tag{30}$$

for all $\alpha\in(0,1]$. Also from Lemma 2

$$\rho^*H_0P^{-1}H_0^{\mathrm{T}}=1.$$

Since $\mathcal{E}(P,\rho)$ is contractively invariant, we must have $\rho<\rho^*$. It follows that:

$$\rho H_0P^{-1}H_0^{\mathrm{T}}<1.$$

Since $H_0=H(0)$, by the continuity of $H(\alpha)$, there exists a sufficiently small $\alpha>0$ such that

$$\rho H(\alpha)P^{-1}H(\alpha)^{\mathrm{T}}<1 \tag{31}$$

i.e., $\mathcal{E}(P,\rho)\subset\mathcal{L}(H(\alpha))$. This completes the proof by observing (30) and letting $H=H(\alpha)$. □

The results in Theorems 2 and 4 and Fact 1 can be summarized in the following proposition.

Proposition 2: Given a $P>0$, assume that $\mathcal{E}(P,\rho)$ can be made contractively invariant for some $\rho>0$. Let $\rho>0$ be given. The following statements are equivalent:

$$\det(\rho - B^{\mathrm{T}}P\Phi^{-1}P\Phi^{-1}PB) = 0$$

$$\Updownarrow$$

$$\det\begin{bmatrix} \rho & -B^{\mathrm{T}}P\Phi^{-1} \\ -\Phi^{-1}PB & P^{-1} \end{bmatrix} = 0$$

$$\Updownarrow$$

$$\det\left(\begin{bmatrix} \rho & 0 \\ 0 & P^{-1} \end{bmatrix} - \begin{bmatrix} B^{\mathrm{T}}P & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} \Phi^{-1} & 0 \\ 0 & \Phi^{-1} \end{bmatrix}\begin{bmatrix} 0 & I \\ PB & 0 \end{bmatrix}\right) = 0$$

$$\Updownarrow$$

$$\det\left(\begin{bmatrix} \Phi & 0 \\ 0 & \Phi \end{bmatrix} - \begin{bmatrix} 0 & I \\ PB & 0 \end{bmatrix}\begin{bmatrix} \rho^{-1} & 0 \\ 0 & P \end{bmatrix}\begin{bmatrix} B^{\mathrm{T}}P & 0 \\ 0 & I \end{bmatrix}\right) = 0$$

$$\Updownarrow$$

$$\det\begin{bmatrix} \lambda P - A^{\mathrm{T}}P - PA & P \\ \rho^{-1}PBB^{\mathrm{T}}P & \lambda P - A^{\mathrm{T}}P - PA \end{bmatrix} = 0.$$

a) $\mathcal{E}(P, \rho)$ can be made contractively invariant with some $u = h(x)$, $|h(x)| \leq 1$;
b) $\mathcal{E}(P, \rho)$ is contractively invariant under the control $u = -\mathrm{sign}(B^{\mathrm{T}}Px)$;
c) $\mathcal{E}(P, \rho)$ is contractively invariant under $u = \mathrm{sat}(Fx)$, where F satisfies

$$(A + BF)^{\mathrm{T}}P + P(A + BF) < 0; \tag{32}$$

d) there exists $H \in \mathbf{R}^{1\times n}$ satisfying

$$(A + BH)^{\mathrm{T}}P + P(A + BH) < 0$$

and $\mathcal{E}(P, \rho) \subset \mathcal{L}(H)$.

The equivalence results in Proposition 2 are somewhat counter intuitive. Condition a) seems to be the weakest and condition d) seems to be the strongest. Yet they are all equal. The equivalence of a) and d) implies that if an ellipsoid can be made contractively invariant with a control $u = h(x)$, $|h(x)| \leq 1$, then there exists a control law linear in x in the ellipsoid to make it contractively invariant.

Of the four statements in Proposition 2, b) can be checked with Theorem 3 but there is no direct method to check c). The last condition d) is the most easily tractable and can be flexibly incorporated into analysis and design problems such as estimation of the domain of attraction and enlarging the domain of attraction (see [8]–[10]). It can also be extended for the purpose of disturbance rejection (see [8], [10]).

Let us illustrate the application of Proposition 2 with the following example. Suppose that we are given a shape reference set X_{R}. We want to design a controller $u = \mathrm{sat}(f(x))$ such that the scaled reference set αX_{R} is inside some invariant ellipsoid $\mathcal{E}(P, \rho)$ of the closed-loop system

$$\dot{x} = Ax + B\mathrm{sat}(f(x)).$$

The objective is to maximize the quantity α with design parameters P, ρ and the control law $u = \mathrm{sat}(f(x))$. This problem can be referred to as one of enlarging the domain of attraction as in [9]. In [9], we restricted the control law to be linear in the ellipsoid $\mathcal{E}(P, \rho)$. That is, $u = \mathrm{sat}(Fx)$ and $|Fx| \leq 1$ for all $x \in \mathcal{E}(P, \rho)$, which is equivalent to $\mathcal{E}(P, \rho) \subset \mathcal{L}(F)$. In view of Proposition 2, this restriction to linear control law will not affect the resulting maximal value of α. The great advantage of the restriction is that the optimization problem can be easily solved with LMI technique (see [8] and [9]).

IV. Conclusion

This paper presents a complete characterization of invariant ellipsoids for a single input linear system subject to actuator saturation. In particular, we obtained several equivalent conditions for an ellipsoid to be invariant under a certain saturated linear feedback. One of the condition is stated in terms of linear matrix inequality, which can be easily used for stability analysis and controller design.

Appendix
An Algebraic Fact

Fact 2: Assume that P and $(\lambda P - A^{\mathrm{T}}P - PA)$ are nonsingular, then

$$\det\begin{bmatrix} \lambda P - A^{\mathrm{T}}P - PA & P \\ \rho^{-1}PBB^{\mathrm{T}}P & \lambda P - A^{\mathrm{T}}P - PA \end{bmatrix} = 0 \tag{33}$$

is equivalent to

$$B^{\mathrm{T}}P(A^{\mathrm{T}}P + PA - \lambda P)^{-1}P(A^{\mathrm{T}}P + PA - \lambda P)^{-1}PB = \rho. \tag{34}$$

Proof: Denote

$$\Phi = \lambda P - A^{\mathrm{T}}P - PA,$$

then the equation (34) can be written as

$$B^{\mathrm{T}}P\Phi^{-1}P\Phi^{-1}PB = \rho. \tag{35}$$

By invoking the equalities

$$\det\begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \end{bmatrix} = \det(X_1)\det(X_4 - X_3X_1^{-1}X_2) \tag{36}$$

and

$$\det\begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \end{bmatrix} = \det(X_4)\det(X_1 - X_2X_1^{-1}X_3) \tag{37}$$

we see that (35) is equivalent to the equations shown at the top of the page. The last equation is (33).

References

[1] F. Blanchini, "Set invariance in control—A survey," *Automatica*, vol. 35, no. 11, pp. 1747–1767, 1999.
[2] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory*, ser. SIAM Studies in Appl. Mathematics Philadelphia, PA, 1994.
[3] E. J. Davison and E. M. Kurak, "A computational method for determining quadratic Lyapunov functions for nonlinear systems," *Automatica*, vol. 7, pp. 627–636, 1971.
[4] E. G. Gilbert and K. T. Tan, "Linear systems with state and control constraints: The theory and application of maximal output admissible sets," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 1008–1020, 1991.
[5] J. M. Gomes da Silva, Jr. and S. Tarbouriech, "Local stabilization of discrete-time linear systems with saturating controls: An LMI approach," *IEEE Trans. Automat. Contr.*, vol. 46, pp. 119–125, Jan. 2001.
[6] D. Henrion, S. Tarbouriech, and G. Garcia, "Output feedback robust stabilization of uncertain linear systems with saturating controls," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 2230–2237, Nov. 1999.

[7] H. Hindi and S. Boyd, "Analysis of linear systems with saturating using convex optimization," in *Proc. 37th IEEE Conf. Decision Control*, Florida, 1998, pp. 903–908.
[8] T. Hu and Z. Lin, *Control Systems With Actuator Saturation: Analysis and Design*. Boston, MA: Birkhäuser, 2001.
[9] ——, "On enlarging the basin of attraction for linear systems under saturated linear feedback," *Syst. Contr. Lett.*, vol. 40, no. 1, pp. 59–69, May 2000.
[10] T. Hu, Z. Lin, and B. M. Chen, "An analysis and design method for linear systems subject to actuator saturation and disturbance," *Automatica*, vol. 38, no. 2, pp. 351–359, 2002.
[11] T. Hu, Z. Lin, and Y. Shamash, "On maximizing the convergence rate of linear systems with input saturation," in *Proc. 2001 Amer. Control Conf.*, Arlington, VA, 2001, pp. 4896–4901.
[12] H. Khalil, *Nonlinear Systems*. Upper Saddle River, NJ: Prentice-Hall, 1996.
[13] K. A. Loparo and G. L. Blankenship, "Estimating the domain of attraction of nonlinear feedback systems," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 602–607, Apr. 1978.
[14] C. Pittet, S. Tarbouriech, and C. Burgat, "Stability regions for linear systems with saturating controls via circle and Popov criteria," in *Proc. 36th IEEE Conf. Decision Control*, San Diego, 1997, pp. 4518–4523.
[15] S. Weissenberger, "Application of results from the absolute stability to the computation of finite stability domains," *IEEE Trans. Automat. Control.*, vol. AC-13, pp. 124–125, 1968.
[16] G. F. Wredenhagen and P. R. Belanger, "Piecewise-linear LQ control for systems with input constraints," *Automatica*, vol. 30, pp. 403–416, 1994.

Solving a Nonlinear Output Regulation Problem: Zero Miss Distance of Pure PNG

Jae-Hyuk Oh

***Abstract*—This note presents a solution to the output regulation problem of a nonlinear system with time-varying disturbance: the system represents the well-known missile-target pursuit situation where the missile is guided by the pure proportional navigation guidance (PPNG) law while the target maneuvers with time-varying normal acceleration, and the problem is to prove the zero miss distance property of PPNG, which has been studied for decades without satisfactory success. To solve this problem, we construct a function by which a time sequence of the missile-to-target range is upper-bounded, and prove that the function is strictly decreasing, which is also proven to guarantee that there is always a subsequence that asymptotically converges to zero. The solution is given in the form of a necessary and sufficient condition guaranteeing zero miss distance of PPNG.**

***Index Terms*—Nonlinear output regulation, pure proportional navigation guidance (PPNG), time-varying disturbance, zero miss distance.**

NOMENCLATURE

$v_m(v_t)$	Missile (target) speed.
$a_m(a_t)$	Missile (target) acceleration.
θ_L	Euler angle from reference coordinate system to LOS coordinate system.
θ_m	Euler angle from LOS coordinate system to missile body coordinate system.
θ_t	Euler angle from LOS coordinate system to target body coordinate system.
N	Navigation constant.
r	Missile-to-target range.
ρ	v_t/v_m.
$s\theta_i$, $c\theta_i$	$\sin\theta_i$, $\cos\theta_i$.
a.e.	*almost everywhere*.

Manuscript received October 6, 1999; revised June 2, 2000 and February 22, 2001. Recommended by Associate Editor W. Lin.
The author was with Massachusetts Institute of Technology, Cambridge, MA 02139 USA. He is now with United Technologies Research Center, East Hartford, CT 06108 USA (e-mail: OhJ@utrc.utc.com).
Publisher Item Identifier S 0018-9286(02)01112-1.

I. INTRODUCTION

The pure proportional navigation guidance (PPNG) law has been widely adopted in many tactical missile systems because of its simplicity and high capturability, which has been proven in practice. As a consequence, a lot of research has been carried out to confirm mathematically the high capturability of the PPNG [1]–[6].

Specifically, it has been shown in [2], and [4]–[6] that a missile guided by PPNG can always intercept a target, which is maneuvering with time-varying normal acceleration, under some conditions on the navigation constant, the initial heading error, the initial missile-to-target range, the magnitude of target acceleration, and the ratio of missile speed to target speed. These conditions are equivalent to the conditions for the missile-to-target range to be *strictly decreasing* after a finite time. In other words, the prior research excludes the case when the missile-to-target range has a *fluctuating* time-profile caused by the target maneuver. This exclusion results in significant discrepancies between the mathematical analysis results and the actual capturability of the PPNG.

In fact, it is quite challenging to mathematically analyze the capturability of PPNG for the case when the missile-to-target range is doomed to fluctuate. Interestingly, the problem that we are dealing with can be viewed as an output regulation problem of a nonlinear system with time-varying perturbations: the nonlinear system is the missile-target pursuit dynamics, the feedback controller is the PPNG law, the output is the missile-to-target range, and the perturbation is the time-varying normal acceleration of the target. However, since our output zeroing manifold does not contain any equilibrium points, it is impossible to directly apply the results of earlier research [7]–[11] to this problem.

This note describes the construction of an asymptotic time-function by which the missile-to-target range is always upper-bounded. This approach can provide a necessary and sufficient condition under which a missile, which is launched toward the target with $\rho < 1$ and guided by the PPNG law, can always capture a target maneuvering arbitrarily with time-varying normal acceleration. Specifically, it is shown that a navigation constant larger than 1 is the only condition required to achieve zero miss distance. In our analysis, the nonlinear dynamics of pursuit situations are taken into full account.

II. PRELIMINARIES

The two-dimensional (2-D) pursuit situation can be represented as follows [2], [4]:

$$\dot{r} = (\rho c\theta_t - c\theta_m)v_m \tag{1}$$

$$\dot{\theta}_m = \frac{a_m}{v_m} - \dot{\theta}_L \tag{2}$$

$$\dot{\theta}_t = \frac{a_t}{v_t} - \dot{\theta}_L \tag{3}$$

where a_m, missile acceleration generated by the PPNG, and $\dot{\theta}_L$, LOS rate, are given by

$$\dot{\theta}_L = \frac{(\rho s\theta_t - s\theta_m)v_m}{r} \tag{4}$$

$$a_m = Nv_m\dot{\theta}_L. \tag{5}$$

0018–9286/02$17.00 © 2002 IEEE

Publication 14

ELSEVIER

Systems & Control Letters 45 (2002) 97–112

SYSTEMS & CONTROL LETTERS

www.elsevier.com/locate/sysconle

Analysis and design for discrete-time linear systems subject to actuator saturation ☆

Tingshu Hu[a,*], Zongli Lin[a], Ben M. Chen[b]

[a] *Department of Electrical & Computer Engineering, University of Virginia, Charlottesville, VA 22903, USA*
[b] *Department of Electrical & Computer Engineering, National University of Singapore, Singapore 117576, Singapore*

Received 14 February 2001; received in revised form 7 July 2001; accepted 24 August 2001

Abstract

We present a method to estimate the domain of attraction for a discrete-time linear system under a saturated linear feedback. A simple condition is derived in terms of an auxiliary feedback matrix for determining if a given ellipsoid is contractively invariant. Moreover, the condition can be expressed as linear matrix inequalities (LMIs) in terms of all the varying parameters and hence can easily be used for controller synthesis. The following surprising result is revealed for systems with single input: suppose that an ellipsoid is made invariant with a linear feedback, then it is invariant under the saturated linear feedback if and only if there exists a saturated (nonlinear) feedback which makes the ellipsoid invariant. Finally, the set invariance condition is extended to determine invariant sets for systems with persistent disturbances. LMI based methods are developed for constructing feedback laws that achieve disturbance rejection with guaranteed stability requirements. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Actuator saturation; Stability analysis; Disturbance rejection; Set invariance

1. Introduction

In this paper, we are interested in the control of linear systems subject to actuator saturation and persistent disturbances,

$$x(k+1) = Ax(k) + B\text{sat}(u(k)) + Ew(k), \quad x \in \mathbb{R}^n, \ u \in \mathbb{R}^m, \ w \in \mathbb{R}^q, \tag{1}$$

where x is the state, u is the control, w is the disturbance and $\text{sat}(\cdot)$ is the standard saturation function. First, we will consider the closed-loop stability under a given linear state feedback $u = Fx$ in the absence of the disturbance. There has been a lot of work on this topic (see, e.g. [3–5,9–14] and the references therein). For the continuous-time case, various simple and general methods for estimating the domain of attraction have been developed by applying the absolute stability analysis tools, such as the circle and Popov criteria (see, e.g. [5,9,10,12], where the saturation is treated as a locally sector bounded nonlinearity and the domain of attraction is estimated by use of quadratic and Lur'e type Lyapunov functions). The multivariable circle criterion in [9]

☆ This work was supported in part by the US Office of Naval Research Young Investigator Program under grant N00014-99-1-0670.
* Corresponding author.
E-mail addresses: th7f@virginia.edu (T. Hu), zl5y@virginia.edu (Z. Lin), bmchen@nus.edu.sg (B.M. Chen).

0167-6911/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-6911(01)00168-2

was restated in [12], in terms of (nonlinear) matrix inequalities in controller parameters and other auxiliary optimization parameters, such as the positive definite matrix P in the Lyapunov function $V(x)=x^{\mathrm{T}}Px$ and the saturation levels. By fixing some of the parameters, these matrix inequalities simplify to linear matrix inequalities (LMIs) and can be treated with the LMI software. A nice feature of these analysis tools is that they can be adapted for controller synthesis by simply considering the feedback gain matrix as an additional optimization parameter.

In [8], a simpler criterion is derived in terms of an auxiliary feedback matrix for determining if a given ellipsoid is contractively invariant under a given feedback law. This condition is shown to be less conservative than the existing conditions which are based on the circle criterion or the vertex analysis. The most important feature of this new condition is that it can be expressed as LMIs in terms of all the varying parameters and hence can easily be used for controller synthesis.

In this paper, the set invariance criterion in [8] will be extended to discrete-time systems although the approach has to be quite different. In [8], the set invariance criterion is proven by expanding the derivative of the Lyapunov function and examining the terms that include the saturated feedback $\mathrm{sat}(Fx)$. However, for the discrete-time case, the terms of the increment of the Lyapunov function cannot be examined separately. A new approach by placing the saturated control $\mathrm{sat}(Fx)$ in the convex hull of a group of linear controls is derived to establish the main results. By further exploiting the idea, we will reveal a surprising fact for the single input systems: Given a feedback matrix F, assume that an ellipsoid is invariant for the linear system

$$x(k+1)=Ax(k)+BFx(k).$$

Then, it is invariant for the system

$$x(k+1)=Ax(k)+B\mathrm{sat}(Fx(k))$$

if and only if there exists a feedback law $u=h(x)$, $|h(x)|\leqslant 1$, such that the ellipsoid is invariant for the system

$$x(k+1)=Ax(k)+Bh(x(k)).$$

This means that the set invariance property under a group of saturated linear feedback laws is in some sense independent of a particular feedback in this group as long as all the corresponding linear feedback laws make the ellipsoid invariant.

Based on the stability analysis result, some disturbance rejection problems will be considered, such as, set invariance property in the presence of disturbance, invariant set enlargement, disturbance rejection and disturbance rejection with guaranteed stability region.

This paper is organized as follows. Section 2 addresses the analysis of and design for closed-loop stability. Section 3 addresses the issues related to disturbance rejection. In particular, Section 2.1 presents conditions for an ellipsoid to be invariant. Section 2.2 derives a necessary and sufficient condition for an ellipsoid to be invariant in the case of single input. Section 2.3 proposes an optimization approach to estimating the domain of attraction. Section 2.4 presents a controller design approach to enlarging the domain of attraction. A brief concluding remark is given in Section 4.

2. Estimation of the domain of attraction

2.1. Condition for set invariance — multiple input case

Consider the open-loop system

$$x(k+1)=Ax(k)+B\mathrm{sat}(u(k)),\quad x\in\mathbb{R}^n,\ u\in\mathbb{R}^m, \tag{2}$$

where $\mathrm{sat}(\cdot)$ is the standard saturation function of appropriate dimensions. In the above system, $\mathrm{sat}:\mathbb{R}^m\to\mathbb{R}^m$, and $\mathrm{sat}(u)=[\mathrm{sat}(u_1),\mathrm{sat}(u_2),\ldots,\mathrm{sat}(u_m)]^{\mathrm{T}}$, where $\mathrm{sat}(u_i)=\mathrm{sgn}(u_i)\min\{1,|u_i|\}$. Here we have slightly abused

the notation by using sat(·) to denote both the scalar valued and the vector valued saturation functions. Suppose that a state feedback $u=Fx$ has been designed such that $A+BF$ is Schur stable and that the closed-loop linear system satisfies some performance requirement. We would like to know how the closed-loop system behaves in the presence of saturation nonlinearity, in particular, to what extent the stability is preserved. Our first objective of this paper is to obtain an estimate of the domain of attraction of the origin for the closed-loop system

$$x(k+1)=Ax(k)+B\text{sat}(Fx(k)). \tag{3}$$

For a matrix $F\in\mathbb{R}^{m\times n}$, denote the jth row of F as f_j and define

$$\mathscr{L}(F):=\{x\in\mathbb{R}^n: |f_jx|\leqslant 1,\ j\in[1,m]\}.$$

If F is the feedback matrix, then $\mathscr{L}(F)$ is the region where the feedback control $u=\text{sat}(Fx)$ is linear in x.

For $x(0)=x_0\in\mathbb{R}^n$, denote the state trajectory of the system (3) as $\psi(k,x_0)$. The *domain of attraction* of the origin is

$$\mathscr{S}:=\left\{x_0\in\mathbb{R}^n: \lim_{k\to\infty}\psi(k,x_0)=0\right\}.$$

A set is said to be *invariant* if all the trajectories starting from it will remain in it.

Let $P\in\mathbb{R}^{n\times n}$ be a positive-definite matrix. For $\rho>0$, denote

$$\mathscr{E}(P,\rho)=\{x\in\mathbb{R}^n: x^{\mathrm{T}}Px\leqslant\rho\}.$$

Let $V(x)=x^{\mathrm{T}}Px$. The set $\mathscr{E}(P,\rho)$ is said to be *contractively invariant* if

$$\Delta V(x):=(Ax+B\text{sat}(Fx))^{\mathrm{T}}P(Ax+B\text{sat}(Fx))-x^{\mathrm{T}}Px<0$$

for all $x\in\mathscr{E}(P,\rho)\setminus\{0\}$. Clearly, if $\mathscr{E}(P,\rho)$ is contractively invariant, then it is inside the domain of attraction.

We will develop conditions under which $\mathscr{E}(P,\rho)$ is contractively invariant and thus obtain an estimate of the domain of attraction.

Let $\mathscr{D}$ be the set of $m\times m$ diagonal matrices whose diagonal elements are either 1 or 0. There are 2^m elements in $\mathscr{D}$. Suppose that each element of $\mathscr{D}$ is labeled as D_i, $i=1,2,\ldots,2^m$. Then, $\mathscr{D}=\{D_i: i\in[1,2^m]\}$. Denote $D_i^-=I-D_i$. Clearly, D_i^- is also an element of $\mathscr{D}$ if $D_i\in\mathscr{D}$. Given two vectors, $u,v\in\mathbb{R}^m$, $\{D_iu+D_i^-v: i\in[1,2^m]\}$ is the set of vectors formed by choosing some elements from u and the remaining from v. Given two matrices $F,H\in\mathbb{R}^{m\times n}$, $\{D_iF+D_i^-H: i\in[1,2^m]\}$ is the set of matrices formed by choosing some rows from F and the remaining from H.

With these D_i and D_i^- matrices, a discrete-time counterpart of Theorem 10.4 in [9] (when applied to saturation nonlinearities) can be derived with some standard technique in robustness analysis for systems with varying parameters.

Proposition 1. *Given an ellipsoid $\mathscr{E}(P,\rho)$, if there exists a positive diagonal matrix $K\in\mathbb{R}^{m\times m}$, $K<I$ such that*

$$(A+B(D_iF+D_i^-KF))^{\mathrm{T}}P(A+B(D_iF+D_i^-KF))-P<0,\quad\forall i\in[1,2^m], \tag{4}$$

and $\mathscr{E}(P,\rho)\subset\mathscr{L}(KF)$, then $\mathscr{E}(P,\rho)$ is a contractively invariant set.

Here, the varying gain of each control channel (due to saturation) is viewed as an uncertain parameter varying between K_{ii} and 1, and the quadratic stability (within $\mathscr{E}(P,\rho)$) of the systems corresponding to this box of uncertain parameters is guaranteed by those on the vertices of the box, $\{D_iF+D_i^-KF: i\in[1,2^m]\}$. Similar to [8], we have the following less conservative criterion for an ellipsoid to be contractively invariant.

Theorem 1. *Given an ellipsoid $\mathscr{E}(P,\rho)$, if there exists an $H\in\mathbb{R}^{m\times n}$ such that*

$$(A+B(D_iF+D_i^-H))^{\mathrm{T}}P(A+B(D_iF+D_i^-H))-P<0,\quad \forall i\in[1,2^m], \tag{5}$$

and $\mathscr{E}(P,\rho)\subset\mathscr{L}(H)$, then $\mathscr{E}(P,\rho)$ is a contractively invariant set.

Although a natural discrete-time counterpart of Theorem 1 in [8], the above theorem cannot be proven in a similar way. Before starting the proof of Theorem 1, we need some simple facts about the convex hull of a set of points. Recall that for a group of points, $u^1,u^2,\ldots,u^{\mathscr{I}}$, the convex hull of these points is defined as,

$$\mathrm{co}\{u^i\colon i\in[1,\mathscr{I}]\}:=\left\{\sum_{i=1}^{\mathscr{I}}\alpha_iu^i\colon \sum_{i=1}^{\mathscr{I}}\alpha_i=1,\ \alpha_i\geqslant 0\right\}.$$

Lemma 1. *Let $u,u^1,u^2,\ldots,u^{\mathscr{I}}\in\mathbb{R}^{m_1}$, $v,v^1,v^2,\ldots,v^{\mathscr{I}}\in\mathbb{R}^{m_2}$. If $u\in\mathrm{co}\{u^i\colon i\in[1,\mathscr{I}]\}$ and $v\in\mathrm{co}\{v^j\colon j\in[1,\mathscr{J}]\}$, then*

$$\begin{bmatrix}u\\v\end{bmatrix}\in\mathrm{co}\left\{\begin{bmatrix}u^i\\v^j\end{bmatrix}\colon i\in[1,\mathscr{I}],\ j\in[1,\mathscr{J}]\right\}. \tag{6}$$

Proof. Since $u\in\mathrm{co}\{u^i\colon i\in[1,\mathscr{I}]\}$ and $v\in\mathrm{co}\{v^j\colon j\in[1,\mathscr{J}]\}$, there exist $\alpha_i,\beta_j\geqslant 0$, $i=1,2,\ldots,\mathscr{I}$, $j=1,2,\ldots,\mathscr{J}$, such that

$$\sum_{i=1}^{\mathscr{I}}\alpha_i=\sum_{j=1}^{\mathscr{J}}\beta_j=1,\quad u=\sum_{i=1}^{\mathscr{I}}\alpha_iu^i,\quad v=\sum_{j=1}^{\mathscr{J}}\beta_jv^j.$$

Therefore,

$$\begin{bmatrix}u\\v\end{bmatrix}=\begin{vmatrix}\sum_{i=1}^{\mathscr{I}}\alpha_iu^i\\ \sum_{j=1}^{\mathscr{J}}\beta_jv^j\end{vmatrix}=\begin{vmatrix}\sum_{i=1}^{\mathscr{I}}\alpha_iu^i(\sum_{j=1}^{\mathscr{J}}\beta_j)\\ \sum_{j=1}^{\mathscr{J}}\beta_jv^j(\sum_{i=1}^{\mathscr{I}}\alpha_i)\end{vmatrix}=\begin{vmatrix}\sum_{i=1}^{\mathscr{I}}\sum_{j=1}^{\mathscr{J}}\alpha_i\beta_ju^i\\ \sum_{i=1}^{\mathscr{I}}\sum_{j=1}^{\mathscr{J}}\alpha_i\beta_jv^j\end{vmatrix}=\sum_{i=1}^{\mathscr{I}}\sum_{j=1}^{\mathscr{J}}\alpha_i\beta_j\begin{bmatrix}u^i\\v^j\end{bmatrix}.$$

Noting that

$$\sum_{i=1}^{\mathscr{I}}\sum_{j=1}^{\mathscr{J}}\alpha_i\beta_j=\sum_{i=1}^{\mathscr{I}}\alpha_i\sum_{j=1}^{\mathscr{J}}\beta_j=1,$$

we obtain (6). □

Lemma 2. *Let $u,v\in\mathbb{R}^m$,*

$$u=\begin{bmatrix}u_1\\u_2\\\vdots\\u_m\end{bmatrix},\quad v=\begin{bmatrix}v_1\\v_2\\\vdots\\v_m\end{bmatrix}.$$

Suppose that $|v_j|\leqslant 1$ for all $j\in[1,m]$, then

$$\mathrm{sat}(u)\in\mathrm{co}\{D_iu+D_i^-v\colon i\in[1,2^m]\}.$$

Proof. Since $|v_j|\leqslant 1$, we have

$$\mathrm{sat}(u_j)\in\mathrm{co}\{u_j,v_j\},\quad \forall j\in[1,m].$$

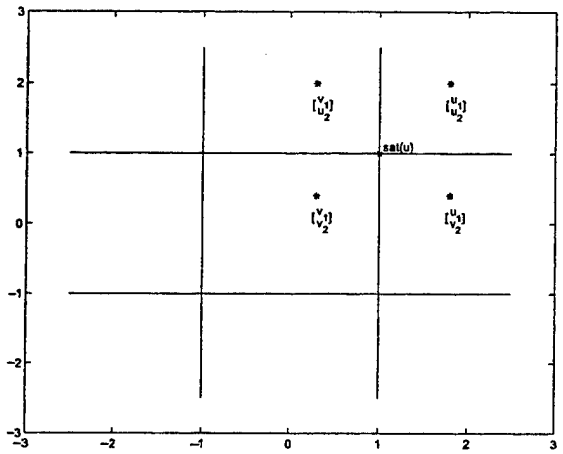


Fig. 1. Illustration for Lemma 2.

By applying Lemma 1 inductively, we have

$$\mathrm{sat}(u_1) \in \mathrm{co}\{u_1, v_1\},$$

$$\mathrm{sat}\left(\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}\right) \in \mathrm{co}\left\{\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \begin{bmatrix} u_1 \\ v_2 \end{bmatrix}, \begin{bmatrix} v_1 \\ u_2 \end{bmatrix}, \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\right\},$$

$$\mathrm{sat}\left(\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}\right) \in \mathrm{co}\left\{\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}, \begin{bmatrix} u_1 \\ u_2 \\ v_3 \end{bmatrix}, \begin{bmatrix} u_1 \\ v_2 \\ u_3 \end{bmatrix}, \begin{bmatrix} u_1 \\ v_2 \\ v_3 \end{bmatrix}, \begin{bmatrix} v_1 \\ u_2 \\ u_3 \end{bmatrix}, \begin{bmatrix} v_1 \\ u_2 \\ v_3 \end{bmatrix}, \begin{bmatrix} v_1 \\ v_2 \\ u_3 \end{bmatrix}, \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}\right\},$$

$$\vdots$$

and finally,

$$\mathrm{sat}(u) \in \mathrm{co}\{D_i u + D_i^- v\colon\ i \in [1, 2^m]\}. \qquad \square$$

Lemma 2 is illustrated in Fig. 1 for the case where $m = 2$.

Given two feedback matrices $F, H \in \mathbb{R}^{m\times n}$, suppose that $|h_j x| \leqslant 1$ for all $j \in [1, m]$, then by Lemma 2, we have

$$\mathrm{sat}(Fx) \in \mathrm{co}\{D_i F x + D_i^- H x\colon\ i \in [1, 2^m]\}.$$

In this way, we have placed $\mathrm{sat}(Fx)$ into the convex hull of a group of linear feedbacks.

Proof of Theorem 1. Let $V(x) = x^{\mathrm{T}} P x$, we need to show that

$$\Delta V(x) = (Ax + B\mathrm{sat}(Fx))^{\mathrm{T}} P (Ax + B\mathrm{sat}(Fx)) - x^{\mathrm{T}} P x < 0, \quad \forall x \in \mathscr{E}(P, \rho)\backslash\{0\}. \tag{7}$$

Since $\mathscr{E}(P, \rho) \subset \mathscr{L}(H)$, i.e., $|h_j x| \leqslant 1$ for all $j \in [1, m]$ and $x \in \mathscr{E}(P, \rho)$, by Lemma 2, for every $x \in \mathscr{E}(P, \rho)$,

$$\mathrm{sat}(Fx) \in \mathrm{co}\{D_i F x + D_i^- H x\colon\ i \in [1, 2^m]\}.$$

It follows that

$$Ax + B\mathrm{sat}(Fx) \in \mathrm{co}\{Ax + B(D_i F + D_i^- H)x\colon\ i \in [1, 2^m]\}.$$

By the convexity of the function $V(z)=z^{\mathrm{T}}Pz$, we have

$$(Ax+B\mathrm{sat}(Fx))^{\mathrm{T}}P(Ax+B\mathrm{sat}(Fx)) \leqslant \max_{i\in[1,2^m]} x^{\mathrm{T}}(A+B(D_iF+D_i^-H))^{\mathrm{T}}P(A+B(D_iF+D_i^-H))x$$

for every $x\in\mathscr{E}(P,\rho)$. Since condition (5) is satisfied, we have

$$\max_{i\in[1,2^m]} x^{\mathrm{T}}(A+B(D_iF+D_i^-H))^{\mathrm{T}}P(A+B(D_iF+D_i^-H))x < x^{\mathrm{T}}Px$$

for all $x\neq 0$. Therefore, for every $x\in\mathscr{E}(P,\rho)\backslash\{0\}$,

$$(Ax+B\mathrm{sat}(Fx))^{\mathrm{T}}P(Ax+B\mathrm{sat}(Fx)) < x^{\mathrm{T}}Px.$$

This verifies (7). □

We see that Proposition 1 is a special case of Theorem 1 by setting $H=KF$. Clearly the condition in Theorem 1 is less conservative than that in Proposition 1. This will be illustrated in Example 1. Another important advantage of Theorem 1 is that, when optimization is concerned, it leads to constraints in the form of linear matrix inequality while from Proposition 1 we can only get bilinear matrix inequalities. This will be investigated later.

2.2. *The necessary and sufficient condition — single input case*

For the single input case ($m=1$), $\mathscr{D}=\{0,1\}$. So the condition in Theorem 1 for $\mathscr{E}(P,\rho)$ to be contractively invariant simplifies to: there exists an $H\in\mathbb{R}^{1\times n}$ such that

$$(A+BF)^{\mathrm{T}}P(A+BF)-P<0, \quad (A+BH)^{\mathrm{T}}P(A+BH)-P<0$$

and $\mathscr{E}(P,\rho)\in\mathscr{L}(H)$. In fact, we can go one step further to obtain the following surprising result.

Theorem 2. *Assume $m=1$. Given an ellipsoid $\mathscr{E}(P,\rho)$, suppose that*

$$(A+BF)^{\mathrm{T}}P(A+BF)-P<0. \tag{8}$$

Then, $\mathscr{E}(P,\rho)$ is contractively invariant under $u=\mathrm{sat}(Fx)$ if and only if there exists a function $h(x):\mathbb{R}^m\to\mathbb{R}$, $|h(x)|\leqslant 1$ for all $x\in\mathscr{E}(P,\rho)$, such that $\mathscr{E}(P,\rho)$ is contractively invariant under the control $u=h(x)$, i.e.,

$$(Ax+Bh(x))^{\mathrm{T}}P(Ax+Bh(x))-x^{\mathrm{T}}Px<0, \quad \forall x\in\mathscr{E}(P,\rho)\setminus\{0\}. \tag{9}$$

Proof. The "only if" part is obvious. Now we show the "if" part. Here we have $|h(x)|\leqslant 1$ for all $x\in\mathscr{E}(P,\rho)$. It follows from Lemma 2 that for every $x\in\mathscr{E}(P,\rho)$, $\mathrm{sat}(Fx)\in\mathrm{co}\{Fx,h(x)\}$.

By the convexity of the function $V(z)=z^{\mathrm{T}}Pz$, we have

$$(Ax+B\mathrm{sat}(Fx))^{\mathrm{T}}P(Ax+B\mathrm{sat}(Fx)) \leqslant \max\{x^{\mathrm{T}}(A+BF)^{\mathrm{T}}P(A+BF)x,(Ax+Bh(x))^{\mathrm{T}}P(Ax+Bh(x))\}.$$

By (8) and (9), we obtain

$$(Ax+B\mathrm{sat}(Fx))^{\mathrm{T}}P(Ax+B\mathrm{sat}(Fx))-x^{\mathrm{T}}Px<0, \quad \forall x\in\mathscr{E}(P,\rho)\backslash\{0\}.$$

This shows that $\mathscr{E}(P,\rho)$ is contractively invariant under $u=\mathrm{sat}(Fx)$. □

Theorem 2 implies that, for the single input case, the invariance of an ellipsoid $\mathscr{E}(P,\rho)$ under a saturated linear control $u=\mathrm{sat}(Fx)$ is in some sense independent of F as long as the condition $(A+BF)^{\mathrm{T}}P(A+BF)-P<0$ is satisfied. In other words, suppose that both F_1 and F_2 satisfy the condition $(A+BF_i)^{\mathrm{T}}P(A+BF_i)-P<0$, $i=1,2$, then the maximal invariant ellipsoid $\mathscr{E}(P,\rho)$ (with ρ maximized under a fixed P) is the same under either $u=\mathrm{sat}(F_1x)$ or $u=\mathrm{sat}(F_2x)$. In [6, Chapter 11], we developed a computational method to determine the largest ρ such that $\mathscr{E}(P,\rho)$ can be made invariant.

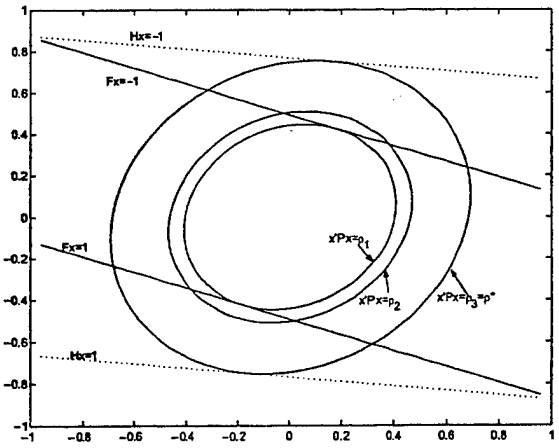


Fig. 2. Invariant ellipsoids determined with different methods.

Example 1. Consider the closed-loop system (3) with

$$A=\begin{bmatrix} 0.8876 & -0.5555 \\ 0.5555 & 1.5542 \end{bmatrix}, \quad B=\begin{bmatrix} -0.1124 \\ 0.5555 \end{bmatrix}, \quad F=[\,-0.7651 \quad -2.0299\,].$$

Given

$$P=\begin{bmatrix} 5.0127 & -0.6475 \\ -0.6475 & 4.2135 \end{bmatrix}.$$

By combining Theorem 2 and Hu and Lin's method [6], the maximal $\mathscr{E}(P,\rho)$ is $\mathscr{E}(P,\rho^*)$ with $\rho^*=2.3490$.

Let's compare the largest invariant ellipsoid, $\mathscr{E}(P,\rho^*)$, with those obtained by other methods.

(1) The maximal ρ such that $\mathscr{E}(P,\rho)\subset\mathscr{L}(F)$ is $\rho_1=0.8237$;
(2) The maximal ρ satisfying the condition in Proposition 1 is $\rho_2=1.0710$;
(3) The maximal ρ satisfying the condition in Theorem 1 is $\rho_3=\rho^*=2.3490$, with

$$H=[\,-0.1389 \quad -1.3018].$$

Shown in Fig. 2 is a comparison of the invariant ellipsoids obtained with different methods. It is very interesting to note that $\rho_3=\rho^*$. In this case, the largest invariant ellipsoid obtained by Theorem 1 is not conservative at all. In fact, as we have proven for continuous-time systems in [7], if $m=1$, then the condition in Theorem 1 (in its continuous-time form) is also necessary.

2.3. *Estimation of the domain of attraction — an LMI approach*

With all the ellipsoids satisfying the set invariance condition in Theorem 1, we would like to choose from among them the "largest" one to get a least conservative estimate of the domain of attraction. In the literature (e.g., see [2,3,5]), the largeness of a set is usually measured by its volume. Here, we will follow the idea in [8] and take the shape of a set into consideration. Let $X_R\subset\mathbb{R}^n$ be a prescribed bounded convex set. For a set $S\subset\mathbb{R}^n$, define

$$\alpha_R(S):=\sup\{\alpha>0:\ \alpha X_R\subset S\}.$$

If $\alpha_R(S)\geqslant 1$, then $X_R\subset S$. Two typical types of X_R are the ellipsoids

$$X_R=\{x\in\mathbb{R}^n:\ x^TRx\leqslant 1\}, \quad R>0$$

and the polyhedrons

$$X_R = \mathrm{co}\{x_1, x_2, \ldots, x_l\}.$$

We can choose the reference set X_R according to the available information on the initial conditions. For instance, if some possible initial conditions are known, we can choose X_R as a polyhedron containing all these initial conditions. In the extreme case, we may choose X_R to be $\mathrm{co}\{x_0, -x_0\}$ when we want to know if x_0 is in the domain of attraction.

Now we would like to choose from all the $\mathscr{E}(P,\rho)$'s that satisfy the condition in Theorem 1 such that the quantity $\alpha_R(\mathscr{E}(P,\rho))$ is maximized. For this reason, we call X_R the shape reference set. The problem of maximizing the contractively invariant ellipsoid with respect to a shape reference set can be formulated as

$$\begin{aligned} &\sup_{P>0,\rho,H} \alpha \\ &\text{s.t.} \quad \text{(a) } \alpha X_R \subset \mathscr{E}(P,\rho), \\ &\quad \text{(b) } (A + B(D_iF + D_i^- H))^T P (A + B(D_iF + D_i^- H)) - P < 0, \quad \forall i \in [1, 2^m], \\ &\quad \text{(c) } |h_j x| \leqslant 1, \quad \forall x \in \mathscr{E}(P,\rho), \ j \in [1,m]. \end{aligned} \tag{10}$$

We will transform the above optimization constraints into LMIs. If X_R is a polyhedron, then by Schur complement, (a) is equivalent to

$$\alpha^2 x_k^T \left(\frac{P}{\rho}\right) x_k \leqslant 1 \quad \Leftrightarrow \quad \begin{bmatrix} 1/\alpha^2 & x_k^T \\ x_k & (P/\rho)^{-1} \end{bmatrix} \geqslant 0, \quad k \in [1, l]. \tag{11}$$

If X_R is an ellipsoid $\{x: x^T R x \leqslant 1\}$, then (a) is equivalent to

$$\frac{R}{\alpha^2} \geqslant \frac{P}{\rho} \quad \Leftrightarrow \quad \begin{bmatrix} 1/\alpha^2 R & I \\ I & (P/\rho)^{-1} \end{bmatrix} \geqslant 0. \tag{12}$$

Also by Schur complement and some manipulation, the constraint (b) is equivalent to

$$\begin{bmatrix} (P/\rho)^{-1} & (P/\rho)^{-1}(A + B(D_iF + D_i^- H))^T \\ (A + B(D_iF + D_i^- H))(P/\rho)^{-1} & (P/\rho)^{-1} \end{bmatrix} > 0, \quad i \in [1, 2^m]. \tag{13}$$

From [8], the constraint (c) is equivalent to

$$\rho h_j P^{-1} h_j^T \leqslant 1 \Leftrightarrow \begin{bmatrix} 1 & h_j(P/\rho)^{-1} \\ (P/\rho)^{-1} h_j^T & (P/\rho)^{-1} \end{bmatrix} \geqslant 0, \quad j \in [1, m]. \tag{14}$$

Let $\gamma = 1/\alpha^2$, $Q = (P/\rho)^{-1}$ and $Z = H(P/\rho)^{-1}$. Also let the jth row of Z be z_j, i.e., $z_j = h_j(P/\rho)^{-1}$. If X_R is a polyhedron, then from (11), (13) and (14), the optimization problem (10) can be rewritten as

$$\begin{aligned} &\inf_{Q,Z} \gamma \\ &\text{s.t.} \quad \text{(a1)} \begin{bmatrix} \gamma & x_k^T \\ x_k & Q \end{bmatrix} \geqslant 0, \quad k \in [1, l], \\ &\quad \text{(b)} \begin{bmatrix} Q & (AQ + B(D_iFQ + D_i^- Z))^T \\ AQ + B(D_iFQ + D_i^- Z) & Q \end{bmatrix} > 0, \quad i \in [1, 2^m], \\ &\quad \text{(c)} \begin{bmatrix} 1 & z_j \\ z_j^T & Q \end{bmatrix} \geqslant 0, \quad j \in [1, m], \end{aligned} \tag{15}$$

where all the constraints are given in LMIs.

If X_R is an ellipsoid, we just need to replace (a1) with another LMI,

$$\text{(a2)} \begin{bmatrix} \gamma R & I \\ I & Q \end{bmatrix} \geqslant 0.$$

2.4. Controller design

Our objective in this section is to choose a feedback matrix $F \in \mathbb{R}^{m\times n}$ such that the estimated domain of attraction as obtained by the method of Section 2.3 is maximized with respect to X_R. This can be simply done by taking the F in (15) as an extra optimization parameter. To make the optimization easy, we use a new parameter Y to replace FQ in (15(b)) and the resulting LMI problem is

$$\begin{aligned} &\inf_{Q,Y,Z} \gamma \\ &\text{s.t.} \quad (15(a1)),\ (15(c)), \\ &\quad (b) \begin{bmatrix} Q & (AQ + B(D_iY + D_i^- Z))^T \\ AQ + B(D_iY + D_i^- Z) & Q \end{bmatrix} > 0, \quad i \in [1, 2^m]. \end{aligned} \tag{16}$$

The optimal F will be recovered from YQ^{-1}. Denote the optimal value of the above optimization problem as γ^*.

Let's consider a simpler optimization problem

$$\begin{aligned} &\inf_{Q,Z} \gamma \\ &\text{s.t.} \quad (15(a1)),\ (15(c)), \\ &\quad (b1) \begin{bmatrix} Q & (AQ + BZ)^T \\ AQ + BZ & Q \end{bmatrix} > 0. \end{aligned} \tag{17}$$

Denote its optimal value as γ_1^*. We claim that $\gamma^* = \gamma_1^*$. The argument goes as follows. Since (b1) is only one of the inequality constraints in (b) (when $D_i = 0$), (17) can be viewed as a problem resulting from dropping the other $2^m - 1$ constraints in (16(b)), hence we have $\gamma^* \geqslant \gamma_1^*$. On the other hand, we can also see (b1) as a result of (b) by restricting $Y = Z$ (recall that $D_i + D_i^- = I$). This means that the constraints of (17) are more restrictive than that of (16). Hence (16) admits a less infimum than (17), i.e., $\gamma^* \leqslant \gamma_1^*$. In summary, we must have $\gamma^* = \gamma_1^*$.

In view of the above observation, we might as well solve the simpler optimization problem (17) if our objective is to enlarge the domain of attraction. If we solve (17) and let $H = ZQ^{-1}$, then the resulting invariant ellipsoid is in the linear region of the state feedback $u = \text{sat}(Hx)$, i.e., $\mathscr{E}(P, \rho) \subset \mathscr{L}(H)$. In this case, the convergence of trajectories would be generally very slow inside the ellipsoid. Suppose that we have another feedback $u = \text{sat}(Fx)$ such that F and H satisfy (5), then by Theorem 1, the ellipsoid is also invariant under $u = \text{sat}(Fx)$. There could be infinitely many such F's. We may choose among these F's to optimize other performances such as convergence rate.

3. Disturbance rejection

3.1. Problem statement

Consider the open-loop system

$$x(k+1) = Ax(k) + B\text{sat}(u(k)) + Ew(k), \quad x \in \mathbb{R}^n,\ u \in \mathbb{R}^m,\ w \in \mathbb{R}^q, \tag{18}$$

where, without loss of generality, we assume that the bounded disturbance w belongs to the set $\mathscr{W} := \{w: w(k)^T w(k) \leqslant 1,\ \forall k \geqslant 0\}$. Let the state feedback be $u = Fx$. The closed-loop system is

$$x(k+1) = Ax(k) + B\text{sat}(Fx(k)) + Ew(k). \tag{19}$$

For an initial state $x(0) = x_0$, denote the state trajectory of the closed-loop system under w as $\psi(k, x_0, w)$.

Our primary concern is the boundedness of the trajectories. A set in $\mathbb{R}^n$ is said to be *invariant* if all the trajectories starting from it will remain in it regardless of $w \in \mathscr{W}$. An ellipsoid $\mathscr{E}(P,\rho)$ is said to be *strictly invariant* if

$$(Ax + B\text{sat}(Fx) + Ew)^{\mathrm{T}} P(Ax + B\text{sat}(Fx) + Ew) < \rho$$

for all $x \in \mathscr{E}(P,\rho)$ and $w, w^{\mathrm{T}}w \leqslant 1$.

The notion of invariant set plays an important role in studying the stability and other performances of a system, see [1,2,9] and the references therein. To keep the state trajectory bounded for a large range of initial conditions, it is desired to have a large invariant set. On the other hand, a small invariant set indicates that the system is insensitive to the disturbance. Suppose that we have an invariant set containing the origin in its interior, then all the trajectories starting from the origin will remain inside the invariant set regardless of the disturbance. Hence for the purpose of disturbance rejection, we would also like to have a small invariant set containing the origin in its interior.

To formally state the objectives of this section, we need to extend the notion of the domain of attraction of an equilibrium to that of an invariant set as follows.

Definition 1. Let $\mathscr{B}$ be a bounded invariant set of (19). The domain of attraction of $\mathscr{B}$ is

$$\mathscr{S}(\mathscr{B}) := \left\{ x_0 \in \mathbb{R}^n : \lim_{k\to\infty} d(\psi(k,x_0,w),\mathscr{B}) = 0,\ \forall w \in \mathscr{W} \right\},$$

where $d(\psi(k,x_0,w),\mathscr{B}) = \inf_{x\in\mathscr{B}} ||\psi(k,x_0,w) - x||$ is the distance from $\psi(k,x_0,w)$ to $\mathscr{B}$.

In the above definition, $||\cdot||$ can be any norm. The problems we are to address in this section are formulated as follows.

Problem 1 (*Set invariance analysis*). Let F be known. Given an ellipsoid $\mathscr{E}(P,\rho)$, determine if $\mathscr{E}(P,\rho)$ is (strictly) invariant.

Problem 2 (*Invariant set enlargement*). Given a shape reference set $X_0 \subset \mathbb{R}^n$, design an F such that the closed-loop system has an invariant set $\mathscr{E}(P,\rho) \supset \alpha_2 X_0$ with α_2 maximized.

Problem 3 (*Disturbance rejection*). Given a shape reference set $X_\infty \subset \mathbb{R}^n$, design an F such that the closed-loop system has an invariant set $\mathscr{E}(P,\rho) \subset \alpha_3 X_\infty$ with α_3 minimized. Here we can also take X_∞ to be the (possibly unbounded) polyhedron $\{x \in \mathbb{R}^n : |c_i x| \leqslant 1,\ i \in [1,p]\}$. In this case, the minimization of α_3 leads to the minimization of the L_∞-norm of the output $y = Cx \in \mathbb{R}^p$.

Problem 4 (*Disturbance rejection with guaranteed domain of attraction*). Given two shape reference sets, X_∞ and X_0. Design an F such that the closed-loop system has an invariant set $\mathscr{E}(P,1) \supset X_0$, and for all $x_0 \in \mathscr{E}(P,1)$, $\psi(k,x_0,w)$ will enter a smaller invariant set $\mathscr{E}(P,\rho_1) \subset \alpha_4 X_\infty$ with α_4 minimized.

3.2. Condition for set invariance

We consider the closed-loop system (19) with a given F.

Theorem 3. *For a given ellipsoid $\mathscr{E}(P,\rho)$, if there exist an $H \in \mathbb{R}^{m\times n}$ and a positive number η such that*

$$(1+\eta)(A + B(D_i F + D_i^- H))^{\mathrm{T}} P (A + B(D_i F + D_i^- H)) + \left(\frac{1+\eta}{\rho\eta}\lambda_{\max}(E^{\mathrm{T}}PE) - 1\right) P \leqslant (<) 0 \qquad (20)$$

for all $i \in [1, 2^m]$, and $\mathscr{E}(P,\rho) \subset \mathscr{L}(H)$, then $\mathscr{E}(P,\rho)$ is a (strictly) invariant set for system (19).

Proof. We prove the strict invariance. That is, we will show that

$$(Ax + B\text{sat}(Fx) + Ew)^{\mathrm{T}} P (Ax + B\text{sat}(Fx) + Ew) < \rho, \quad \forall x \in \mathscr{E}(P,\rho),\ w^{\mathrm{T}}w \leqslant 1.$$

Since $\mathscr{E}(P,\rho)\subset\mathscr{L}(H)$, similar to the proof of Theorem 1, we can show that

$$Ax+B\mathrm{sat}(Fx)+Ew\in\mathrm{co}\{Ax+B(D_iF+D_i^-H)x+Ew:\ i\in[1,2^m]\}$$

for every $w\in\mathbb{R}^q$ and $x\in\mathscr{E}(P,\rho)$. By the convexity of the function $V(z)=z^{\mathrm{T}}Pz$, for every $x\in\mathscr{E}(P,\rho)$ and every w, $w^{\mathrm{T}}w\leqslant 1$,

$$\begin{aligned}&(Ax+B\mathrm{sat}(Fx)+Ew)^{\mathrm{T}}P(Ax+B\mathrm{sat}(Fx)+Ew)\\&\quad\leqslant\max_{i\in[1,2^m]}(Ax+B(D_iF+D_i^-H)x+Ew)^{\mathrm{T}}P(Ax+B(D_iF+D_i^-H)x+Ew).\end{aligned}$$

Using the fact that $(a+b)^{\mathrm{T}}(a+b)\leqslant(1+\eta)a^{\mathrm{T}}a+(1+1/\eta)b^{\mathrm{T}}b$ for any $\eta>0$, we have

$$\begin{aligned}&(Ax+B\mathrm{sat}(Fx)+Ew)^{\mathrm{T}}P(Ax+B\mathrm{sat}(Fx)+Ew)\\&\quad\leqslant\max_{i\in[1,2^m]}(1+\eta)x^{\mathrm{T}}(A+B(D_iF+D_i^-H))^{\mathrm{T}}P(A+B(D_iF+D_i^-H))x+\left(1+\frac{1}{\eta}\right)w^{\mathrm{T}}E^{\mathrm{T}}PEw\\&\quad\leqslant\max_{i\in[1,2^m]}(1+\eta)x^{\mathrm{T}}(A+B(D_iF+D_i^-H))^{\mathrm{T}}P(A+B(D_iF+D_i^-H))x+\left(1+\frac{1}{\eta}\right)\lambda_{\max}(E^{\mathrm{T}}PE).\end{aligned}$$

To prove the strict invariance, it suffices to show that there exists an $\eta>0$ such that for all $x\in\partial\mathscr{E}(P,\rho)$ and for all $i\in[1,2^m]$,

$$(1+\eta)x^{\mathrm{T}}(A+B(D_iF+D_i^-H))^{\mathrm{T}}P(A+B(D_iF+D_i^-H))x+\left(1+\frac{1}{\eta}\right)\lambda_{\max}(E^{\mathrm{T}}PE)<\rho. \tag{21}$$

Noticing that $1=x^{\mathrm{T}}(P/\rho)x$ on $\partial\mathscr{E}(P,\rho)$, we see that (21) is guaranteed by (20). □

Theorem 3 deals with Problem 1 and can be easily used for controller design in Problem 2 and Problem 3. For Problem 2, we can solve the following optimization problem:

$$\begin{aligned}&\sup_{P>0,\rho,\eta>0,F,H}\ \alpha_2\\&\text{s.t.}\quad\text{(a) }\alpha_2X_0\subset\mathscr{E}(P,\rho),\\&\quad\text{(b) }(A+B(D_iF+D_i^-H))^{\mathrm{T}}P(A+B(D_iF+D_i^-H))\\&\qquad\frac{1}{1+\eta}\left(\frac{1+\eta}{\eta}\lambda_{\max}\left(\frac{E^{\mathrm{T}}PE}{\rho}\right)-1\right)P\leqslant 0,\quad i\in[1,2^m],\\&\quad\text{(c) }|h_jx|\leqslant 1,\quad\forall x\in\mathscr{E}(P,\rho),\quad j\in[1,m].\end{aligned} \tag{22}$$

Let $\gamma=1/\alpha_2^2$, $Q=(P/\rho)^{-1}$, $Y=FQ$ and $Z=HQ$. Similar to Section 2.3, the above optimization problem can be transformed into one with LMI constraints (providing two scalars are fixed). Here, constraint (b) is equivalent to the existence of $\lambda\in(0,\eta/(1+\eta))$ such that

$$\begin{bmatrix}\dfrac{1}{1+\eta}\left(1-\dfrac{1+\eta}{\eta}\lambda\right)Q & (AQ+BD_iY+BD_i^-Z)^{\mathrm{T}}\\ AQ+BD_iY+BD_i^-Z & Q\end{bmatrix}\geqslant 0,\quad i\in[1,2^m] \tag{23}$$

and

$$\begin{bmatrix}\lambda & E^{\mathrm{T}}\\ E & Q\end{bmatrix}\geqslant 0. \tag{24}$$

Hence, the optimization problem (22) is equivalent to

$$\begin{aligned} &\inf_{\eta>0,\lambda,Q,Y,Z} \gamma \\ &\text{s.t.} \qquad (15(a1)),\ (15(c)),\ (23),\ (24). \end{aligned} \tag{25}$$

We see that (15(a1)) and (15(c)) are LMIs. If we fix η and λ, then (23) and (24) are also LMIs. The global infimum of γ can be obtained by running η from 0 to ∞ and λ from 0 to $\eta/(1+\eta)$.

In fact, the computation can be simplified by reducing the number of parameters fixed beforehand (η and λ) from two to one. Denote

$$g=\frac{1}{1+\eta}\left(1-\frac{1+\eta}{\eta}\lambda\right).$$

We see that as η varies from 0 to ∞ and λ from 0 to $\eta/(1+\eta)$, g varies from 0 to 1. If we fix g, then

$$\lambda=1-\frac{1}{1+\eta}-\eta g.$$

It can be shown with standard analysis that as η varies from 0 to ∞, the maximal value of λ is

$$\lambda^*=(1-\sqrt{g})^2,$$

obtained at $\eta^*=(1/\sqrt{g})-1$. Since the constraint (24) is the least restrictive by taking $\lambda=\lambda^*$, the optimal solution to (25) will be obtained with $\lambda=\lambda^*$. In view of these arguments, we can solve (25) by running g from 0 to 1, taking $\lambda=\lambda^*=(1-\sqrt{g})^2$, solving the resulting LMI problems and picking the minimal γ. In this case, we only need to fix the parameter g before solving an LMI problem.

For Problem 3, we have

$$\begin{aligned} &\inf_{P>0,\rho,\eta>0,F,H} \alpha_3 \\ &\text{s.t.} \qquad (a)\ \mathscr{E}(P,\rho)\subset\alpha_3 X_\infty,\ (22(b)),\ (22(c)), \end{aligned} \tag{26}$$

which can be solved similarly as (22).

3.3. Disturbance rejection with guaranteed domain of attraction

Given $X_0\subset\mathbb{R}^n$, if the optimal solution of Problem 2 is $\alpha_2^*>1$, then there are infinitely many choices of the feedback matrices F's such that X_0 is contained in some invariant ellipsoid. We will use this extra freedom for disturbance rejection. That is, to construct another invariant set $\mathscr{E}(P,\rho_1)$ which is as small as possible with respect to some X_∞. Moreover, X_0 is inside the domain of attraction of $\mathscr{E}(P,\rho_1)$. In this way, all the trajectories starting from X_0 will enter $\mathscr{E}(P,\rho_1)\subset\alpha_4 X_\infty$ for some $\alpha_4>0$. Here the number α_4 is a measure of the degree of disturbance rejection.

Before addressing Problem 4, we need to answer the following question: Suppose that for given F and P, both $\mathscr{E}(P,\rho_1)$ and $\mathscr{E}(P,\rho_2)$, $\rho_1<\rho_2$, are strictly invariant sets, then under what conditions will the other ellipsoids $\mathscr{E}(P,\rho)$, $\rho\in(\rho_1,\rho_2)$ also be strictly invariant? If they are, then all the trajectories starting from within $\mathscr{E}(P,\rho_2)$ will enter $\mathscr{E}(P,\rho_1)$ and remain inside it.

Theorem 4. *Given two ellipsoids, $\mathscr{E}(P,\rho_1)$ and $\mathscr{E}(P,\rho_2)$, $\rho_2>\rho_1>0$, if there exist $H_1,H_2\in\mathbb{R}^{m\times n}$ and a positive η such that*

$$(1+\eta)(A+B(D_iF+D_i^-H_1))^{\mathrm{T}}P(A+B(D_iF+D_i^-H_1))+\left(\frac{1+\eta}{\rho_1\eta}\lambda_{\max}(E^{\mathrm{T}}PE)-1\right)P<0, \tag{27}$$

$$(1+\eta)(A+B(D_iF+D_i^-H_2))^{\mathrm{T}}P(A+B(D_iF+D_i^-H_2))+\left(\frac{1+\eta}{\rho_2\eta}\lambda_{\max}(E^{\mathrm{T}}PE)-1\right)P<0, \tag{28}$$

for all $i\in[1,2^m]$, *and* $\mathscr{E}(P,\rho_1)\subset\mathscr{L}(H_1)$, $\mathscr{E}(P,\rho_2)\subset\mathscr{L}(H_2)$, *then for every* $\rho\in[\rho_1,\rho_2]$, *there exists an* $H\in\mathbb{R}^{m\times n}$ *such that*

$$(1+\eta)(A+B(D_iF+D_i^-H))^{\mathrm{T}}P(A+B(D_iF+D_i^-H))+\left(\frac{1+\eta}{\rho\eta}\lambda_{\max}(E^{\mathrm{T}}PE)-1\right)P<0 \tag{29}$$

and $\mathscr{E}(P,\rho)\in\mathscr{L}(H)$. *This implies that* $\mathscr{E}(P,\rho)$ *is also strictly invariant.*

Proof. Let $h_{1,j}$ and $h_{2,j}$ be the jth rows of H_1 and H_2, respectively. The conditions $\mathscr{E}(P,\rho_1)\subset\mathscr{L}(H_1)$ and $\mathscr{E}(P,\rho_2)\subset\mathscr{L}(H_2)$ are equivalent to

$$\begin{bmatrix}1/\rho_1 & h_{1,j}\\ h_{1,j}^{\mathrm{T}} & P\end{bmatrix}\geqslant 0,\quad \begin{bmatrix}1/\rho_2 & h_{2,j}\\ h_{2,j}^{\mathrm{T}} & P\end{bmatrix}\geqslant 0,\quad j\in[1,m].$$

Since $\rho\in[\rho_1,\rho_2]$, there exists an $\alpha\in[0,1]$ such that $1/\rho=\alpha/\rho_1+(1-\alpha)/\rho_2$. Let $H=\alpha H_1+(1-\alpha)H_2$. Clearly

$$\begin{bmatrix}1/\rho & h_j\\ h_j^{\mathrm{T}} & P\end{bmatrix}\geqslant 0,$$

which implies that $\mathscr{E}(P,\rho)\subset\mathscr{L}(H)$. Since (27) and (28) are equivalent to

$$\begin{bmatrix}\dfrac{1}{1+\eta}\left(1-\dfrac{1+\eta}{\rho_1\eta}\lambda_{\max}(E^{\mathrm{T}}PE)\right)P & (A+B(D_iF+D_i^-H_1))^{\mathrm{T}}\\ A+B(D_iF+D_i^-H_1) & P^{-1}\end{bmatrix}>0$$

and

$$\begin{bmatrix}\dfrac{1}{1+\eta}\left(1-\dfrac{1+\eta}{\rho_2\eta}\lambda_{\max}(E^{\mathrm{T}}PE)\right)P & (A+B(D_iF+D_i^-H_2))^{\mathrm{T}}\\ A+B(D_iF+D_i^-H_2) & P^{-1}\end{bmatrix}>0$$

by convexity, we have

$$\begin{bmatrix}\dfrac{1}{1+\eta}\left(1-\dfrac{1+\eta}{\rho\eta}\lambda_{\max}(E^{\mathrm{T}}PE)\right)P & (A+B(D_iF+D_i^-H))^{\mathrm{T}}\\ A+B(D_iF+D_i^-H) & P^{-1}\end{bmatrix}>0,$$

which is equivalent to (29). □

In view of Theorem 4, to solve Problem 4, we can construct two invariant ellipsoids $\mathscr{E}(P,\rho_1)$ and $\mathscr{E}(P,\rho_2)$ satisfying the condition of Theorem 4 such that $X_0\subset\mathscr{E}(P,\rho_2)$ and $\mathscr{E}(P,\rho_1)\subset\alpha_4X_\infty$ with α_4 minimized. Since ρ_2 can be absorbed into other parameters, we assume for simplicity that $\rho_2=1$ and $\rho_1<1$. Problem 4 can then be formulated as

$$\begin{aligned}&\inf_{P>0,0<\rho_1<1,\eta>0,F,H_1,H_2}\alpha_4\\ &\text{s.t. (a) } X_0\subset\mathscr{E}(P,1),\quad \mathscr{E}(P,\rho_1)\subset\alpha_4X_\infty,\\ &\text{(b) }\begin{bmatrix}\dfrac{1}{1+\eta}\left(1-\dfrac{1+\eta}{\rho_1\eta}\lambda_{\max}(E^{\mathrm{T}}PE)\right)P & (A+B(D_iF+D_i^-H_1))^{\mathrm{T}}\\ A+B(D_iF+D_i^-H_1) & P^{-1}\end{bmatrix}>0,\quad i\in[1,2^m],\\ &\text{(c) }\begin{bmatrix}\dfrac{1}{1+\eta}\left(1-\dfrac{1+\eta}{\eta}\lambda_{\max}(E^{\mathrm{T}}PE)\right)P & (A+B(D_iF+D_i^-H_2))^{\mathrm{T}}\\ A+B(D_iF+D_i^-H_2) & P^{-1}\end{bmatrix}>0,\quad i\in[1,2^m],\\ &\text{(d) } |h_{1,j}x|\leqslant 1,\quad \forall x\in\mathscr{E}(P,\rho_1),\quad j\in[1,m],\\ &\text{(e) } |h_{2,j}x|\leqslant 1,\quad \forall x\in\mathscr{E}(P,1),\quad j\in[1,m].\end{aligned} \tag{30}$$

If we fix ρ_1, λ and η, then the constraints of the optimization problem can be transformed into LMIs. To obtain the global infimum, we may vary ρ_1 from 0 to 1, η from 0 to ∞ and λ from 0 to $\rho_1\eta/(1+\eta)$. Similar

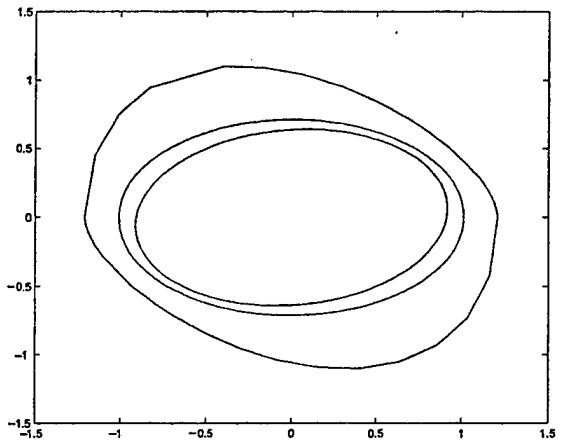


Fig. 3. The invariant ellipsoids and the null controllable region.

to the treatment of the optimization problem (25), we can reduce the number of parameters fixed beforehand (ρ_1, λ and η) from three to two.

3.4. An example

Consider the system (18) with

$$A=\begin{bmatrix}0.9741 & -0.2474\\ 0.2474 & 1.2710\end{bmatrix},\quad B=\begin{bmatrix}-0.0259\\ 0.2474\end{bmatrix},\quad E=\begin{bmatrix}0.0057\\ 0.0082\end{bmatrix}.$$

Let's first consider Problem 2 of enlarging the invariant ellipsoid. Here we choose the shape reference set as a unit ball, i.e., $X_0=\mathscr{E}(I,1)$. By solving (22), we obtain $\alpha_2^*=0.6337$, along with $\eta^*=0.0143$, $\lambda^*=1.9879*10^{-4}$ and

$$P^*=\begin{bmatrix}1.2148 & -0.1667\\ -0.1667 & 2.4681\end{bmatrix},\quad F^*=[-0.5726\quad -1.2574].$$

The invariant set $\mathscr{E}(P^*,1)$ is the smaller ellipsoid in Fig. 3. The larger ellipsoid is obtained as the maximal invariant ellipsoid in the absence of disturbance ($E=0$). The outmost closed curve is the boundary of the null controllable region of the system in the absence of disturbance, which is the largest possible invariant set that can be achieved with any control law (see [6]).

Next, we consider Problem 3. We take the reference set X_∞ also to be the unit ball. The optimal α_3 is found to be $\alpha_3^*=0.0825$ with $\eta^*=0.2126$, $\lambda^*=0.0307$ and

$$P^*=\begin{bmatrix}1920.2 & -1884.6\\ -1884.6 & 2150.6\end{bmatrix},\quad F^*=[-0.3314\quad -2.4721].$$

For Problem 4, we take X_∞ to be the unit ball and $X_0=\alpha_0\mathscr{E}(I,1)$. From the solution to Problem 2, we know that α_0 must be less than $\alpha_2^*=0.6337$. From the solution to Problem 3, we know that α_4^* must be greater than $\alpha_3^*=0.0825$. First, we choose $\alpha_0=0.5$. By solving (30), we obtain $\alpha_4^*=0.2960$, along with $\eta^*=0.0431$, $\lambda^*=2.4565*10^{-4}$, $\rho_1^*=0.1440$ and

$$P^*=\begin{bmatrix}1.9000 & -0.7335\\ -0.7335 & 3.7429\end{bmatrix},\quad F^*=[-0.5707\quad -1.4722].$$

In Fig. 4, the smaller dotted ellipsoid is $\mathscr{E}(P^*,\rho_1)$ and the bigger one is $\mathscr{E}(P^*,1)$. A trajectory starting from the boundary of $\mathscr{E}(P^*,1)$ is plotted in Fig. 4. In the simulation, the disturbance is chosen as

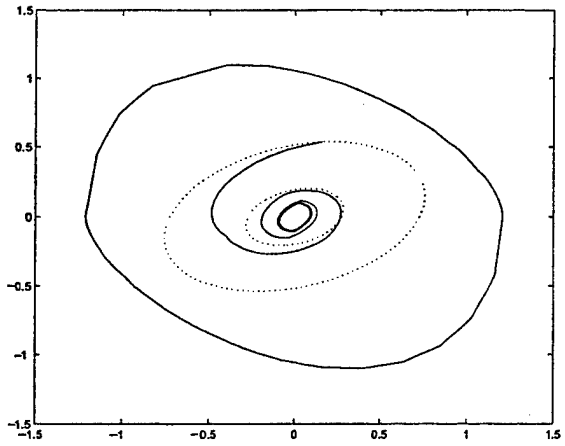


Fig. 4. The invariant ellipsoids and a trajectory, $\alpha_0 = 0.5$, $\alpha_4^* = 0.2960$.

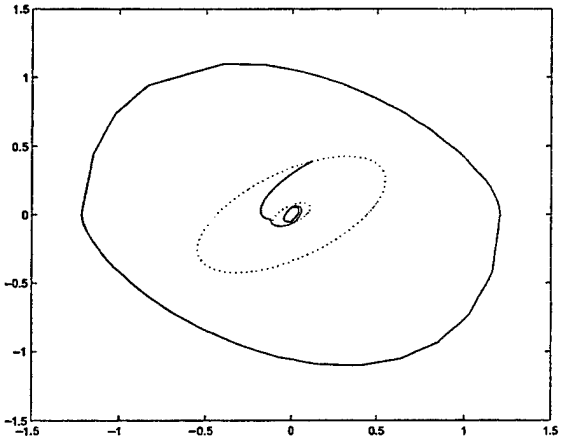


Fig. 5. The invariant ellipsoids and a trajectory, $\alpha_0 = 0.3$, $\alpha_4^* = 0.1262$.

$w(k) = \text{sign}(\sin(0.2k))$. We see that the trajectory enters $\mathscr{E}(P, \rho_1)$ and stays inside of it. However, the disturbance may not be rejected to a satisfactory level. This is because enlarging the outer ellipsoid and reducing the inner ellipsoid are conflicting objectives. To obtain a better disturbance rejection performance, we have to choose smaller X_0. For example, if we choose $\alpha_0 = 0.3$, then we obtain $\alpha_4^* = 0.1262$, along with

$$P^* = \begin{bmatrix} 5.1710 & -3.9277 \\ -3.9277 & 8.5125 \end{bmatrix}, \quad F^* = [-0.5123 \quad -1.8880].$$

Fig. 5 shows the invariant ellipsoids $\mathscr{E}(P^*, 1)$ and $\mathscr{E}(P^*, \rho_1)$, and a trajectory starting from the boundary of $\mathscr{E}(P^*, 1)$.

4. Conclusions

We considered linear systems subject to actuator saturation and persistent disturbance. Simple criteria for determining if a given ellipsoid is contractively invariant have been derived. With the aid of these criteria, we developed analysis and design methods for closed-loop stability and disturbance rejection. Examples were used to demonstrate the effectiveness of these methods.

References

[1] F. Blanchini, Ultimate boundedness control for uncertain discrete-time systems via set-induced Lyapunov function, IEEE Trans. Automat. Control 39 (1994) 428–433.
[2] S. Boyd, L. El Ghaoui, E. Feron, V. Balakrishnan, Linear Matrix Inequalities in Systems and Control Theory, SIAM Studies in Applied Mathematics, Philadelphia, 1994.
[3] E.J. Davison, E.M. Kurak, A computational method for determining quadratic Lyapunov functions for non-linear systems, Automatica 7 (1971) 627–636.
[4] E.G. Gilbert, K.T. Tan, Linear systems with state and control constraints: the theory and application of maximal output admissible sets, IEEE Trans. Automat. Control 36 (1991) 1008–1020.
[5] H. Hindi, S. Boyd, Analysis of linear systems with saturation using convex optimization, Proceedings of the 37th IEEE CDC, Florida, 1998, pp. 903–908.
[6] T. Hu, Z. Lin, Control Systems with Actuator Saturation: Analysis and Design, Birkhäuser, Boston, 2001.
[7] T. Hu, Z. Lin, Exact characterization of invariant ellipsoids for linear systems with saturating actuators, IEEE Trans. Automat. Control, to appear.
[8] T. Hu, Z. Lin, B.M. Chen, An analysis and design method for linear systems subject to actuator saturation and disturbance, Proceedings of the American Control Conferences, 2000, pp. 725–729; also in Automatica, to appear.
[9] H. Khalil, Nonlinear Systems, Prentice-Hall, Upper Saddle River, NJ, 1996.
[10] R.L. Kosut, Design of linear systems with saturating linear control and bounded states, IEEE Trans. Automat. Control 28(1) (1983) 121–124.
[11] K.A. Loparo, G.L. Blankenship, Estimating the domain of attraction of nonlinear feedback systems, IEEE Trans. Automat. Control 23 (4) (1978) 602–607.
[12] C. Pittet, S. Tarbouriech, C. Burgat, Stability regions for linear systems with saturating controls via circle and Popov criteria, Proceedings of the 36th IEEE CDC, San Diego, 1997, pp. 4518–4523.
[13] B.G. Romanchuk, Computing regions for attraction with polytopes: planar case, Automatica 32 (12) (1996) 1727–1732.
[14] A. Vanelli, M. Vidyasagar, Maximal Lyapunov functions and domain of attraction for autonomous nonlinear systems, Automatica 21 (1) (1985) 69–80.

Publication 15

PERGAMON

Automatica 38 (2002) 351–359

automatica

www.elsevier.com/locate/automatica

Brief Paper

An analysis and design method for linear systems subject to actuator saturation and disturbance☆

Tingshu Hu[a,1], Zongli Lin[a,*,1], Ben M. Chen[b]

[a]*Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22903, USA*
[b]*Department of Electrical and Computer Engineering, National University of Singapore, Singapore, 117576, Singapore*

Received 2 November 1999; revised 6 November 2000; received in final form 17 August 2001

Abstract

We present a method for estimating the domain of attraction of the origin for a system under a saturated linear feedback. A simple condition is derived in terms of an auxiliary feedback matrix for determining if a given ellipsoid is contractively invariant. This condition is shown to be less conservative than the existing conditions which are based on the circle criterion or the vertex analysis. Moreover, the condition can be expressed as linear matrix inequalities (LMIs) in terms of all the varying parameters and hence can easily be used for controller synthesis. This condition is then extended to determine the invariant sets for systems with persistent disturbances. LMI based methods are developed for constructing feedback laws that achieve disturbance rejection with guaranteed stability requirements. The effectiveness of the developed methods is illustrated with examples. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Actuator saturation; Stability; Domain of attraction; Invariant set; Disturbance rejection

1. Introduction

In this paper, we are interested in the control of linear systems subject to actuator saturation and persistent disturbances:

$$\dot{x} = Ax + B\sigma(u) + Ew, \quad x \in \mathbf{R}^n, \quad u \in \mathbf{R}^m, \quad w \in \mathbf{R}^q, \quad (1)$$

where x is the state, u is the control, w is the disturbance and $\sigma(\cdot)$ is the standard saturation function. Our first concern is the closed-loop stability (when $w = 0$) under a given linear state feedback $u = Fx$. There has been a lot of work on this topic (see, e.g., Davison & Kurak, 1971; Gilbert & Tan, 1991; Hindi & Boyd, 1998; Khalil, 1996; Loparo & Blankenship, 1978; Pittet, Tarbouriech, & Burgat, 1997; Vanelli & Vidyasagar, 1985; Weissenberger, 1968 and the references therein). In particular, various simple and general methods for estimating the domain of attraction have been developed by applying the absolute stability analysis tools, such as the circle and Popov criteria (see, e.g., Hindi & Boyd, 1998; Khalil, 1996; Pittet et al., 1997 ; Weissenberger, 1968), where the saturation is treated as a locally sector bounded nonlinearity and the domain of attraction is estimated by use of quadratic and Lur'e type Lyapunov functions. In Hindi and Boyd (1998) and Pittet et al. (1997), the condition for local stability and some performance problems are expressed in terms of (nonlinear) matrix inequalities in system parameters and other auxiliary optimization parameters. By fixing some of the parameters, these matrix inequalities simplify to linear matrix inequalities (LMIs) and can be treated with the LMI software.

Since the circle criterion is applicable to general memoryless sector bounded nonlinearities, we can expect the conservatism in estimating the domain of attraction when it is applied to the saturation nonlinearity. In this paper, a less conservative estimation of the domain of attraction is obtained by using a quadratic Lyapunov function. This is made possible by exploring the special property of saturation. Moreover, since this condition is given in terms of LMIs, it is very easy to handle in both analysis and design.

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Andrew R. Teel under the direction of Editor Hassan Khalil.

* Corresponding author. Tel.: +1-434-9246342; fax: +1-434-9248818.

E-mail addresses: th7f@virginia.edu (T. Hu), zl5y@virginia.edu (Z. Lin), bmchen@nus.edu.sg (B.M. Chen).

[1] Supported in part by the US Office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

0005-1098/02/$ - see front matter © 2001 Elsevier Science Ltd. All rights reserved.
PII: S0005-1098(01)00209-6

In the presence of disturbance, we are interested in knowing if there exists a bounded invariant set such that all the trajectories starting from inside of it will remain in it. This problem was addressed by Blanchini (1990) and Blanchini (1994). In this paper, we would further like to synthesize feedback laws that have the ability to reject the disturbance. Here disturbance rejection is in the sense that, there is a small (as small as possible) neighborhood of the origin such that all the trajectories starting from the origin will remain in it. This performance was analyzed by Hindi and Boyd (1998) for the class of disturbances with finite energy. In this paper, we will deal with persistent disturbances and propose a controller design method.

Furthermore, we are also interested in the problem of asymptotic disturbance rejection with nonzero initial states. A related problem was addressed by Hu and Lin (2001b) and Saberi, Lin, and Teel (1996), where the disturbances are input additive and enter the system before the saturating actuator, i.e., the system has a state equation: $\dot{x}=Ax+B\sigma(u+w)$. It is shown in these papers that given any positive number D, any compact subset X_0 of the null controllable region and any arbitrarily small neighborhood X_∞ of the origin, there is a feedback control such that any trajectory starting from within X_0 will enter X_∞ in a finite time for all disturbances w:$\|w\|_\infty \leqslant D$. We, however, could not expect to have this nice result for system (1), where the disturbance enters the system after the saturating actuator. If w or E is sufficiently large, it may even be impossible to keep the state bounded. What we can expect is to have a set X_0 (as large as we can get) and a set X_∞ (as small as we can get) such that all the trajectories starting from X_0 will enter X_∞ in a finite time and remain in it thereafter.

This paper is organized as follows. Section 2 addresses the analysis of and design for closed-loop stability. Section 3 addresses issues related to disturbance rejection. A brief concluding remark is given in Section 4.

2. Stability analysis

2.1. Problem statement

Consider the open-loop system

$$\dot{x}=Ax+B\sigma(u), \quad x\in\mathbf{R}^n, \quad u\in\mathbf{R}^m, \tag{2}$$

where $\sigma(\cdot)$ is the standard saturation function of appropriate dimensions. In the above system, $\sigma:\mathbf{R}^m\to\mathbf{R}^m$, and $\sigma(u)=[\sigma(u_1)\ \sigma(u_2)\ \cdots\ \sigma(u_m)]^{\mathrm{T}}$, where $\sigma(u_i)=\mathrm{sgn}(u_i)\min\{1,|u_i|\}$. Here we have slightly abused the notation by using σ to denote both the scalar valued and the vector valued saturation functions. Suppose that a state feedback $u=Fx$ has been designed such that $A+BF$ is Hurwitz. We would like to know how the closed-loop system behaves in the presence of saturation nonlinearity, in particular, to what extent the stability is preserved. Our main objective in this section is to obtain an estimate of the domain of attraction of the origin for the closed-loop system

$$\dot{x}=Ax+B\sigma(Fx). \tag{3}$$

Denote the ith column of B as b_i and the ith row of F as f_i. Then $BF=b_1f_1+\cdots+b_mf_m$. For a matrix $F\in\mathbf{R}^{m\times n}$, define

$$\mathscr{L}(F):=\{x\in\mathbf{R}^n:\ |f_ix|\leqslant 1, i\in[1,m]\}.$$

If F is the feedback matrix, then $\mathscr{L}(F)$ is the region in the state space where the control is linear in x.

For $x(0)=x_0\in\mathbf{R}^n$, denote the state trajectory of system (3) as $\psi(t,x_0)$. Then the *domain of attraction* of the origin is

$$\mathscr{S}:=\left\{x_0\in\mathbf{R}^n:\ \lim_{t\to\infty}\psi(t,x_0)=0\right\}.$$

Let $P\in\mathbf{R}^{n\times n}$ be a positive-definite matrix. Denote

$$\mathscr{E}(P,\rho)=\{x\in\mathbf{R}^n:\ x^{\mathrm{T}}Px\leqslant\rho\}.$$

Let $V(x)=x^{\mathrm{T}}Px$. The ellipsoid $\mathscr{E}(P,\rho)$ is said to be *contractively invariant* if $\dot{V}(x)=2x^{\mathrm{T}}P(Ax+B\sigma(Fx))<0$ for all $x\in\mathscr{E}(P,\rho)\backslash\{0\}$. Clearly, if $\mathscr{E}(P,\rho)$ is contractively invariant, then it is inside the domain of attraction. We will develop conditions under which $\mathscr{E}(P,\rho)$ is contractively invariant and hence obtain an estimate of the domain of attraction.

2.2. A set invariance condition based on circle criterion

A multivariable circle criterion is presented in Khalil (1996, Theorem 10.1) and is applied to estimate the domain of attraction for system (3), with a given feedback gain F, in Hindi and Boyd (1998) and Pittet et al. (1997).

Proposition 1 (Khalil, 1996; Pittet et al., 1997). *Assume that (F,A,B) is controllable and observable. Given an ellipsoid $\mathscr{E}(P,\rho)$, if there exist positive diagonal matrices $K_1,K_2\in\mathbf{R}^{n\times n}$ with $K_1<I$, $K_1+K_2\geqslant I$ such that*

$$(A+BK_1F)^{\mathrm{T}}P+P(A+BK_1F)+\tfrac{1}{2}(F^{\mathrm{T}}K_2+PB)(K_2F+B^{\mathrm{T}}P)<0 \tag{4}$$

and $\mathscr{E}(P,\rho)\subset\mathscr{L}(K_1F)$, then $\mathscr{E}(P,\rho)$ is a contractively invariant set and hence inside the domain of attraction.

A similar condition based on circle criterion is given in Hindi and Boyd (1998). These conditions are then used for stability and performance analysis with LMI software in Hindi and Boyd (1998) and Pittet et al. (1997). Since inequality (4) is not jointly convex in K_1, K_2 and P, these

parameters need to be optimized separately and there is no guarantee that the global optimal solutions can be obtained for related problems.

2.3. An improved condition for set invariance

We will develop a less conservative set invariance condition by exploring the special property of the saturation nonlinearity. It is based on direct Lyapunov function analysis in terms of an auxiliary feedback matrix $H \in \mathbf{R}^{m\times n}$. This condition turns out to be equivalent to some LMIs. Denote the ith row of H as h_i. For two matrices $F, H \in \mathbf{R}^{m\times n}$ and a vector $v \in \mathbf{R}^m$, denote

$$M(v,F,H) = \begin{bmatrix} v_1 f_1 + (1-v_1)h_1 \\ \vdots \\ v_m f_m + (1-v_m)h_m \end{bmatrix}. \tag{5}$$

Let $\mathscr{V} = \{v \in \mathbf{R}^m: v_i = 1 \text{ or } 0\}$. There are 2^m elements in $\mathscr{V}$. We will use a $v \in \mathscr{V}$ to choose from the rows of F and H to form a new matrix $M(v,F,H)$: if $v_i = 1$, then the ith row of $M(v,F,H)$ is f_i and if $v_i = 0$, then the ith row of $M(v,F,H)$ is h_i. For example, suppose $m = 2$, then

$$\{M(v,F,H): v \in \mathscr{V}\} = \left\{H, \begin{bmatrix} h_1 \\ f_2 \end{bmatrix}, \begin{bmatrix} f_1 \\ h_2 \end{bmatrix}, F\right\}.$$

Theorem 1. *Given an ellipsoid $\mathscr{E}(P,\rho)$, if there exists an $H \in \mathbf{R}^{m\times n}$ such that*

$$(A + BM(v,F,H))^{\mathrm{T}}P + P(A + BM(v,F,H)) < 0 \tag{6}$$

for all $v \in \mathscr{V}$ and $\mathscr{E}(P,\rho) \subset \mathscr{L}(H)$, i.e., $|h_i x| \leqslant 1$ for all $x \in \mathscr{E}(P,\rho)$, $i \in [1,m]$, then $\mathscr{E}(P,\rho)$ is a contractively invariant set.

Proof. Let $V(x) = x^{\mathrm{T}}Px$, we need to show that

$$\dot{V}(x) = 2x^{\mathrm{T}}P(Ax + B\sigma(Fx)) < 0 \quad \forall x \in \mathscr{E}(P,\rho)\backslash\{0\}.$$

Here we have

$$\dot{V}(x) = 2x^{\mathrm{T}}A^{\mathrm{T}}Px + 2x^{\mathrm{T}}PB\sigma(Fx)$$

$$= 2x^{\mathrm{T}}A^{\mathrm{T}}Px + \sum_{i=1}^{m} 2x^{\mathrm{T}}Pb_i\sigma(f_i x).$$

For each term $2x^{\mathrm{T}}Pb_i\sigma(f_i x)$,

(1) If $x^{\mathrm{T}}Pb_i \geqslant 0$ and $f_i x \leqslant -1$, then $2x^{\mathrm{T}}Pb_i\sigma(f_i x) = -2x^{\mathrm{T}}Pb_i \leqslant 2x^{\mathrm{T}}Pb_i h_i x$. Here we note that $-1 \leqslant h_i x$ $\forall x \in \mathscr{E}(P,\rho)$.
(2) If $x^{\mathrm{T}}Pb_i \geqslant 0$ and $f_i x \geqslant -1$, then $\sigma(f_i x) \leqslant f_i x$ and $2x^{\mathrm{T}}Pb_i\sigma(f_i x) \leqslant 2x^{\mathrm{T}}Pb_i f_i x$.
(3) If $x^{\mathrm{T}}Pb_i \leqslant 0$ and $f_i x \geqslant 1$, then $2x^{\mathrm{T}}Pb_i\sigma(f_i x) = 2x^{\mathrm{T}}Pb_i \leqslant 2x^{\mathrm{T}}Pb_i h_i x$. Here we note that $1 \geqslant h_i x$ $\forall x \in \mathscr{E}(P,\rho)$.
(4) If $x^{\mathrm{T}}Pb_i \leqslant 0$ and $f_i x \leqslant 1$, then $\sigma(f_i x) \geqslant f_i x$ and $2x^{\mathrm{T}}Pb_i\sigma(f_i x) \leqslant 2x^{\mathrm{T}}Pb_i f_i x$.

Combining all the four cases, we have

$$2x^{\mathrm{T}}Pb_i\sigma(f_i x) \leqslant \max\{2x^{\mathrm{T}}Pb_i h_i x,\ 2x^{\mathrm{T}}Pb_i f_i x\}$$

for every $x \in \mathscr{E}(P,\rho)$ and each $i \in [1,m]$. Therefore, for every $x \in \mathscr{E}(P,\rho)$,

$$\dot{V}(x) \leqslant 2x^{\mathrm{T}}A^{\mathrm{T}}Px + \sum_{i=1}^{m} \max\{2x^{\mathrm{T}}Pb_i h_i x,\ 2x^{\mathrm{T}}Pb_i f_i x\}.$$

Now we associate every $x \in \mathscr{E}(P,\rho)$ with a vector $v(x) \in \mathscr{V}$ as follows: if $2x^{\mathrm{T}}Pb_i h_i x < 2x^{\mathrm{T}}Pb_i f_i x$, then we set $v_i = 1$, otherwise we set $v_i = 0$. It follows that

$$\dot{V}(x) \leqslant 2x^{\mathrm{T}}A^{\mathrm{T}}Px + 2\sum_{i=1}^{m}(v_i x^{\mathrm{T}}Pb_i f_i x + (1-v_i)x^{\mathrm{T}}Pb_i h_i x)$$

$$= 2x^{\mathrm{T}}A^{\mathrm{T}}Px + 2x^{\mathrm{T}}P\left(\sum_{i=1}^{m} b_i(v_i f_i + (1-v_i)h_i)\right)x$$

$$= 2x^{\mathrm{T}}(A + BM(v,F,H))^{\mathrm{T}}Px.$$

In view of (6), we have that $\dot{V}(x) < 0$ for all $x \in \mathscr{E}(P,\rho)\backslash\{0\}$. □

A geometric interpretation of Theorem 1 can be found in Hu and Lin (2001a). If we restrict H to be K_1F, where K_1 is the same as that in Proposition 1, then we have

Corollary 1. *Given an ellipsoid $\mathscr{E}(P,\rho)$, if there exists a positive diagonal matrix $K_1 \in \mathbf{R}^{n\times n}$ such that*

$$(A + BM(v,F,K_1F))^{\mathrm{T}}P + P(A + BM(v,F,K_1F)) < 0$$

for all $v \in \mathscr{V}$ and $\mathscr{E}(P,\rho) \subset \mathscr{L}(K_1F)$, then $\mathscr{E}(P,\rho)$ is a contractively invariant set.

This corollary is equivalent to Theorem 10.4 in Khalil (1996) when applied to saturation nonlinearity. Obviously, the condition in Corollary 1 is more conservative than that in Theorem 1 because the latter provides more freedom in choosing the H matrix. However, it is evident from Khalil (1996) that the condition in Proposition 1 is even more conservative than that in Corollary 1. Computations show that in general, for a fixed P, Theorem 1 allows for a larger ρ than Corollary 1. Therefore, Theorem 1 offers a wider choice of invariant ellipsoids for optimization and will lead to less conservative estimation of the domain of attraction.

2.4. Estimation of the domain of attraction

With all the ellipsoids satisfying the set invariance condition, we would like to choose from among them the "largest" one to get a least conservative estimation of the domain of attraction. In the literature (see, e.g., Boyd, El Ghaoui, Feron, & Balakrishnan, 1994; Davison &

Kurak, 1971; Pittet et al., 1997), the largeness of an ellipsoid is usually measured by its volume. Here, we will take its shape into consideration. Let $X_R \subset \mathbf{R}^n$ be a prescribed bounded convex set. For a set $S \subset \mathbf{R}^n$, define

$$\alpha_R(S) := \sup\{\alpha > 0: \alpha X_R \subset S\}.$$

If $\alpha_R(S) \geqslant 1$, then $X_R \subset S$. Two typical types of X_R are the ellipsoids

$$X_R = \mathscr{E}(R, 1) = \{x \in \mathbf{R}^n: x^T R x \leqslant 1\}, \quad R > 0$$

and the polyhedrons

$$X_R = \mathrm{co}\{x_1, x_2, \ldots, x_l\},$$

where "co" denotes the convex hull.

Theorem 1 gives a condition for an ellipsoid to be inside the domain of attraction. Now we would like to choose from all the $\mathscr{E}(P,\rho)$'s that satisfy the condition such that the quantity $\alpha_R(\mathscr{E}(P,\rho))$ is maximized. This problem can be formulated as

$$\begin{aligned} \sup_{P>0,\rho,H} \quad & \alpha \\ \text{s.t.} \quad & \text{(a) } \alpha X_R \subset \mathscr{E}(P,\rho), \\ & \text{(b) } (A + BM(v,F,H))^T P \\ & \qquad + P(A + BM(v,F,H)) < 0 \ \ \forall v \in \mathscr{V}, \\ & \text{(c) } \mathscr{E}(P,\rho) \subset \mathscr{L}(H). \end{aligned} \tag{7}$$

If we replace α with $\log\det(P/\rho)^{-1}$ and remove constraint (a), then we obtain the problem of maximizing the volume of $\mathscr{E}(P,\rho)$. Similar modification can be made to other optimization problems to be formulated in this paper. Moreover, the following procedure to transform (7) into a convex optimization problem with LMI constraints can be adapted to the corresponding volume maximization (or minimization) problems.

Now we transform the constraints of (7) into LMIs. If X_R is a polyhedron, then by Schur complement, (a) is equivalent to

$$\alpha^2 x_i^T \left(\frac{P}{\rho}\right) x_i \leqslant 1 \ \Leftrightarrow \ \begin{bmatrix} \frac{1}{\alpha^2} & x_i^T \\ x_i & \left(\frac{P}{\rho}\right)^{-1} \end{bmatrix} \geqslant 0 \tag{8}$$

for all $i \in [1, l]$. If X_R is an ellipsoid $\mathscr{E}(R, 1)$, then (a) is equivalent to

$$\frac{R}{\alpha^2} \geqslant \left(\frac{P}{\rho}\right) \ \Leftrightarrow \ \begin{bmatrix} \frac{1}{\alpha^2} R & I \\ I & \left(\frac{P}{\rho}\right)^{-1} \end{bmatrix} \geqslant 0. \tag{9}$$

Constraint (b) is equivalent to

$$\left(\frac{P}{\rho}\right)^{-1} (A + BM(v,F,H))^T + (A + BM(v,F,H)) \left(\frac{P}{\rho}\right)^{-1} < 0 \quad \forall v \in \mathscr{V}. \tag{10}$$

From Hindi and Boyd (1998), constraint (c) is equivalent to

$$\rho h_i P^{-1} h_i^T \leqslant 1 \Leftrightarrow \begin{bmatrix} 1 & h_i \left(\frac{P}{\rho}\right)^{-1} \\ \left(\frac{P}{\rho}\right)^{-1} h_i^T & \left(\frac{P}{\rho}\right)^{-1} \end{bmatrix} \geqslant 0 \tag{11}$$

for all $i \in [1, m]$. Let $\gamma = 1/\alpha^2, Q = (P/\rho)^{-1}$ and $G = H(P/\rho)^{-1}$. Also let the ith row of G be g_i, i.e., $g_i = h_i(P/\rho)^{-1}$. Note that $M(v,F,H)Q = M(v,FQ,HQ) = M(v,FQ,G)$. If X_R is a polyhedron, then from (8),(10) and (11) optimization problem (7) can be rewritten as

$$\begin{aligned} \inf_{Q>0,G} \quad & \gamma \\ \text{s.t.} \quad & \text{(a1)} \begin{bmatrix} \gamma & x_i^T \\ x_i & Q \end{bmatrix} \geqslant 0, \ i \in [1, l], \\ & \text{(b) } QA^T + AQ + M(v,FQ,G)^T B^T \\ & \qquad + BM(v,FQ,G) < 0 \ \ \forall v \in \mathscr{V}, \\ & \text{(c)} \begin{bmatrix} 1 & g_i \\ g_i^T & Q \end{bmatrix} \geqslant 0, \ i \in [1, m], \end{aligned} \tag{12}$$

where all the constraints are given in LMIs. If X_R is an ellipsoid, we just need to replace (a1) with

$$\text{(a2)} \quad \begin{bmatrix} \gamma R & I \\ I & Q \end{bmatrix} \geqslant 0.$$

Note that there are 2^m matrix inequalities in constraint (b) corresponding to all $v \in \mathscr{V}$.

Example 1. We use an example of Pittet et al. (1997) to illustrate our results. The system is described by (3) with

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 5 \end{bmatrix}, \quad F = [-2 \ \ -1].$$

With $x_1 = [-1 \ \ 0.8]^T$ and $X_R = \mathrm{co}\{x_1, -x_1\}$, we solve (12) and get $\alpha^* = 1/(\gamma^*)^{1/2} = 4.3711$. The maximal ellipsoid is $\mathscr{E}(P^*, 1)$,

$$P^* = \begin{bmatrix} 0.1170 & 0.0627 \\ 0.0627 & 0.0558 \end{bmatrix}$$

(see the solid ellipsoid in Fig. 1). The inner dashed ellipsoid is an invariant set obtained by the circle criterion method in Pittet et al. (1997) and the region bounded by the dash–dotted curve is obtained by the Popov method, also in Pittet et al. (1997). We see that both the regions obtained by the circle criterion and by the Popov method can be actually enclosed in a single invariant ellipsoid.

To get a better estimation, we vary x_1 over a unit circle, and solve (12) for each x_1. Let the optimal α be $\alpha^*(x_1)$. The outermost dotted boundary in Fig. 1 is formed by the points $\alpha^*(x_1)x_1$ as x_1 varies along the unit circle.

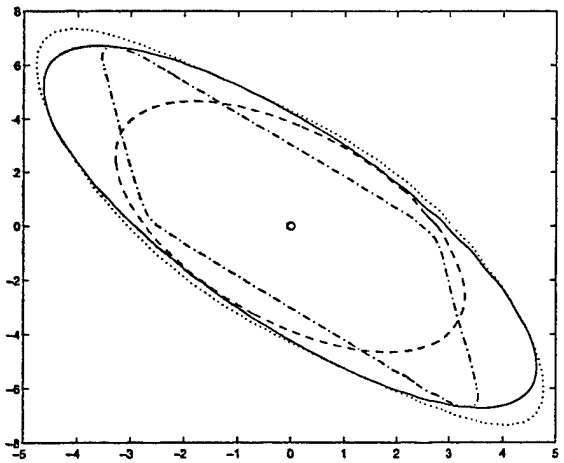


Fig. 1. The invariant sets obtained with different methods.

2.5. Controller design

Now our objective is to choose a feedback matrix $F \in \mathbf{R}^{m\times n}$ such that the estimated domain of attraction as obtained by the method of the last subsection is maximized with respect to X_R. This can be simply done by taking the F in (12) as an extra optimization parameter. To make the optimization easy, we use a new parameter Y to replace FQ in (12) and the resulting LMI problem is

$$\begin{aligned} \inf_{Q>0,Y,G} \quad & \gamma \\ \text{s.t.} \quad & (12a1),(12c) \text{ and} \\ & \text{(b) } QA^{\mathrm{T}} + AQ + M(v,Y,G)^{\mathrm{T}}B^{\mathrm{T}} \\ & \quad + BM(v,Y,G) < 0, \ \forall v \in \mathscr{V}. \end{aligned} \tag{13}$$

The optimal F will be recovered from YQ^{-1}. Consider a simpler optimization problem

$$\begin{aligned} \inf_{Q>0,G} \quad & \gamma \\ \text{s.t.} \quad & (12a1),(12c) \quad \text{and} \\ & \text{(b)} \quad QA^{\mathrm{T}} + AQ + G^{\mathrm{T}}B^{\mathrm{T}} + BG < 0. \end{aligned} \tag{14}$$

If $Y = G$, then all the 2^m inequalities in (13b) are the same as (14b). Hence the new problem (14) can be considered as a result from forcing $Y = G$ in (13). On this account, (14) should have an infimum no less than (13). On the other hand, since the 2^m inequalities in (13b) include (14b), problem (14) can also be considered as a result from discarding $2^m - 1$ inequality constraints of (13b). Because of this, (14) should have an infimum no larger than (13). These arguments show that the optimal values of (14) and (13) must be the same. From the above analysis, we see that, if our only purpose is to enlarge the domain of attraction, we might as well solve the simpler optimization problem (14). The freedom in choosing Y can be used to improve other performances beyond large domain of attraction.

3. Disturbance rejection

3.1. Problem statement

Consider the open-loop system

$$\dot{x} = Ax + B\sigma(u) + Ew, \quad x \in \mathbf{R}^n, \quad u \in \mathbf{R}^m, \quad w \in \mathbf{R}^q, \tag{15}$$

where, without loss of generality, we assume that the bounded disturbance w belongs to the set

$$\mathscr{W} := \{w: w(t)^{\mathrm{T}}w(t) \leqslant 1 \ \forall t \geqslant 0\}.$$

Let the state feedback be $u = Fx$. The closed-loop system is

$$\dot{x} = Ax + B\sigma(Fx) + Ew. \tag{16}$$

For an initial state $x(0) = x_0$, denote the state trajectory of the closed-loop system under w as $\psi(t, x_0, w)$. A set in $\mathbf{R}^n$ is said to be *invariant* if all the trajectories starting from it will remain in it regardless of $w \in \mathscr{W}$. An ellipsoid $\mathscr{E}(P,\rho)$ is said to be *strictly invariant* if $\dot{V} = 2x^{\mathrm{T}}P(Ax + B\sigma(Fx) + Ew) < 0$ for all w such that $w^{\mathrm{T}}w \leqslant 1$ and all $x \in \partial\mathscr{E}(P,\rho)$, the boundary of $\mathscr{E}(P,\rho)$. The notion of invariant set plays an important role in studying the stability and other performances of a system (see, e.g., Blanchini, 1994; Boyd et al., 1994 and the references therein).

Our primary concern is the boundedness of the trajectories for some set of initial states (may be as large as possible). This requires a large invariant set. On the other hand, for the purpose of disturbance rejection, we would also like to have a small invariant set containing the origin in its interior so that a trajectory starting from the origin will stay close to the origin.

To formally state the objectives of this section, we need to extend the notion of the domain of attraction as follows.

Definition 1. Let $\mathscr{B}$ be a bounded invariant set of (16). The domain of attraction of $\mathscr{B}$ is

$$\mathscr{S}(\mathscr{B}) := \left\{x_0 \in \mathbf{R}^n: \lim_{t\to\infty} d(\psi(t,x_0,w),\mathscr{B}) = 0 \ \forall w \in \mathscr{W}\right\},$$

where $d(\psi(t, x_0, w), \mathscr{B}) = \inf_{x\in\mathscr{B}} \|\psi(t, x_0, w) - x\|$ is the distance from $\psi(t, x_0, w)$ to $\mathscr{B}$.

The problems we are to address in this section are given as follows:

Problem 1 (Set invariance analysis).
Let F be known. Given an ellipsoid $\mathscr{E}(P,\rho)$, determine if $\mathscr{E}(P,\rho)$ is (strictly) invariant.

Problem 2 (Invariant set enlargement).
Given a bounded set $X_0 \subset \mathbf{R}^n$, design F such that the closed-loop system has an invariant set $\mathscr{E}(P,\rho) \supset \alpha_2 X_0$ with α_2 maximized.

Problem 3 (Disturbance rejection).

Given a set $X_\infty \subset \mathbf{R}^n$, design F such that the closed-loop system has an invariant set $\mathscr{E}(P,\rho) \subset \alpha_3 X_\infty$ with α_3 minimized.

Problem 4 (Disturbance rejection with guaranteed domain of attraction).

Given two reference sets, X_∞ and X_0, design F such that the closed-loop system has an invariant set $\mathscr{E}(P,1) \supset X_0$, and for all $x_0 \in \mathscr{E}(P,1)$, $\psi(t,x_0,w)$ will enter a smaller invariant set $\mathscr{E}(P,\rho_1) \subset \alpha_4 X_\infty$ with α_4 minimized.

3.2. Condition for set invariance

We consider closed-loop system (16) with a given F. The following theorem gives a sufficient condition for the invariance of a set $\mathscr{E}(P,\rho)$.

Theorem 2. *For a given set $\mathscr{E}(P,\rho)$, if there exist an $H \in \mathbf{R}^{m\times n}$ and a positive number η such that*

$$(A+BM(v,F,H))^{\mathrm{T}}P + P(A+BM(v,F,H)) + \frac{1}{\eta}PEE^{\mathrm{T}}P + \frac{\eta}{\rho}P \leqslant 0 (<0) \quad \forall v \in \mathscr{V} \tag{17}$$

and $\mathscr{E}(P,\rho) \subset \mathscr{L}(H)$, then $\mathscr{E}(P,\rho)$ is a (strictly) invariant set for system (16).

Proof. We prove the strict invariance. That is, for $V(x)=x^{\mathrm{T}}Px$, we will show that

$$\dot{V} = 2x^{\mathrm{T}}P(Ax+B\sigma(Fx)+Ew) < 0$$

for all $x \in \partial\mathscr{E}(P,\rho)$ and all w such that $w^{\mathrm{T}}w \leqslant 1$. Following the procedure of the proof of Theorem 1, we can show that for every $x \in \mathscr{E}(P,\rho)$, there exists a $v \in \mathscr{V}$ such that

$$2x^{\mathrm{T}}P(Ax+B\sigma(Fx)) \leqslant 2x^{\mathrm{T}}(A+BM(v,F,H))^{\mathrm{T}}Px.$$

Since

$$2x^{\mathrm{T}}PEw \leqslant \frac{1}{\eta}x^{\mathrm{T}}PEE^{\mathrm{T}}Px + \eta w^{\mathrm{T}}w \leqslant \frac{1}{\eta}x^{\mathrm{T}}PEE^{\mathrm{T}}Px + \eta$$

we have

$$\dot{V} \leqslant x^{\mathrm{T}}\left(2(A+BM(v,F,H))^{\mathrm{T}}P + \frac{1}{\eta}PEE^{\mathrm{T}}P\right)x + \eta.$$

It follows from (17) that for all $x \in \mathscr{E}(P,\rho)$,

$$\dot{V} < -\frac{\eta}{\rho}x^{\mathrm{T}}Px + \eta.$$

Observing that on the boundary of $\mathscr{E}(P,\rho)$, $x^{\mathrm{T}}Px=\rho$, hence $\dot{V} < 0$. It follows that $\mathscr{E}(P,\rho)$ is a strictly invariant set. □

Theorem 2 deals with Problem 1 and can be easily used for controller design in Problems 2 and 3. For Problem 2, we can solve the following optimization problem:

$$\sup_{P>0,\rho,\eta>0,F,H} \alpha_2 \quad \text{s.t.} \quad \alpha_2 X_0 \subset \mathscr{E}(P,\rho), \quad \mathscr{E}(P,\rho) \subset \mathscr{L}(H) \text{ and } (17). \tag{18}$$

Let $Q=(P/\rho)^{-1}$, $Y=FQ$ and $G=HQ$, then (17) is equivalent to

$$QA^{\mathrm{T}} + AQ + M(v,Y,G)^{\mathrm{T}}B^{\mathrm{T}} + BM(v,Y,G) + \frac{\rho}{\eta}EE^{\mathrm{T}} + \frac{\eta}{\rho}Q < 0 \quad \forall v \in \mathscr{V}.$$

If we fix ρ/η, then the original optimization constraints can be transformed into LMIs as with (7). The global maximum of α_2 will be obtained by running ρ/η from 0 to ∞. For Problem 3, we have

$$\inf_{P>0,\rho,\eta>0,F,H} \alpha_3 \quad \text{s.t.} \quad \mathscr{E}(P,\rho) \subset \alpha_3 X_\infty, \quad \mathscr{E}(P,\rho) \subset \mathscr{L}(H) \text{ and } (17), \tag{19}$$

which can be solved similarly as Problem 2.

3.3. Disturbance rejection with guaranteed domain of attraction

Given $X_0 \subset \mathbf{R}^n$, if the optimal solution of Problem 2 is $\alpha_2^* > 1$, then there are infinitely many choices of the feedback matrices F's such that X_0 is contained in some invariant ellipsoid. We will use this extra freedom for disturbance rejection, that is, to construct another invariant set $\mathscr{E}(P,\rho_1)$ which is as small as possible with respect to some X_∞. Moreover, X_0 is inside the domain of attraction of $\mathscr{E}(P,\rho_1)$. In this way, all the trajectories starting from X_0 will enter $\mathscr{E}(P,\rho_1) \subset \alpha_4 X_\infty$ for some $\alpha_4 > 0$. Here the number α_4 is a measure of the degree of disturbance rejection.

Before addressing Problem 4, we need to answer the following question: Suppose that for given F and P, both $\mathscr{E}(P,\rho_1)$ and $\mathscr{E}(P,\rho_2), \rho_1 < \rho_2$ are strictly invariant, then under what conditions will the other ellipsoids $\mathscr{E}(P,\rho), \rho \in (\rho_1,\rho_2)$ also be strictly invariant? If they are, then all the trajectories starting from within $\mathscr{E}(P,\rho_2)$ will enter $\mathscr{E}(P,\rho_1)$ and remain inside it.

Theorem 3. *Given two ellipsoids, $\mathscr{E}(P,\rho_1)$ and $\mathscr{E}(P,\rho_2)$, $\rho_2 > \rho_1 > 0$, if there exist $H_1, H_2 \in \mathbf{R}^{m\times n}$ and a positive number η such that*

$$(A+BM(v,F,H_1))^{\mathrm{T}}P + P(A+BM(v,F,H_1)) + \frac{1}{\eta}PEE^{\mathrm{T}}P + \frac{\eta}{\rho_1}P < 0 \quad \forall v \in \mathscr{V}, \tag{20}$$

$$(A+BM(v,F,H_2))^{\mathrm{T}}P+P(A+BM(v,F,H_2))$$
$$+\frac{1}{\eta}PEE^{\mathrm{T}}P+\frac{\eta}{\rho_2}P<0 \quad \forall v\in\mathscr{V} \tag{21}$$

and $\mathscr{E}(P,\rho_1)\subset\mathscr{L}(H_1)$, $\mathscr{E}(P,\rho_2)\subset\mathscr{L}(H_2)$, *then for every* $\rho\in[\rho_1,\rho_2]$, *there exists an* $H\in\mathbf{R}^{m\times n}$ *such that*

$$(A+BM(v,F,H))^{\mathrm{T}}P+P(A+BM(v,F,H))$$
$$+\frac{1}{\eta}PEE^{\mathrm{T}}P+\frac{\eta}{\rho}P<0 \quad \forall v\in\mathscr{V} \tag{22}$$

and $\mathscr{E}(P,\rho)\in\mathscr{L}(H)$. *This implies that* $\mathscr{E}(P,\rho)$ *is also strictly invariant.*

Proof. Let $h_{1,i}$ and $h_{2,i}$ be the ith row of H_1 and H_2 respectively. The conditions $\mathscr{E}(P,\rho_1)\subset\mathscr{L}(H_1)$ and $\mathscr{E}(P,\rho_2)\subset\mathscr{L}(H_2)$ are equivalent to

$$\begin{bmatrix}\frac{1}{\rho_1} & h_{1,i}\\ h_{1,i}^{\mathrm{T}} & P\end{bmatrix}\geqslant 0,\quad \begin{bmatrix}\frac{1}{\rho_2} & h_{2,i}\\ h_{2,i}^{\mathrm{T}} & P\end{bmatrix}\geqslant 0,\quad i\in[1,m].$$

Since $\rho\in[\rho_1,\rho_2]$, there exists a $\lambda\in[0,1]$ such that $1/\rho=\lambda(1/\rho_1)+(1-\lambda)/\rho_2$. Let $H=\lambda H_1+(1-\lambda)H_2$. Clearly

$$\begin{bmatrix}\frac{1}{\rho} & h_i\\ h_i^{\mathrm{T}} & P\end{bmatrix}\geqslant 0.$$

From (20) and (21), and by convexity, we have (22). □

In view of Theorem 3, to solve Problem 4, we only need to construct two invariant ellipsoids $\mathscr{E}(P,\rho_1)$ and $\mathscr{E}(P,\rho_2)$ satisfying the condition of Theorem 3 such that $X_0\subset\mathscr{E}(P,\rho_2)$ and $\mathscr{E}(P,\rho_1)\subset\alpha_4X_\infty$ with α_4 minimized. Since ρ_2 can be absorbed into other parameters, we assume for simplicity that $\rho_2=1$ and $\rho_1<1$. Problem 4 can then be formulated as

$$\begin{aligned}\inf_{P>0,0<\rho_1<1,\eta>0,F,H_1,H_2,}\quad & \alpha_4\\ \text{s.t.}\quad & \text{(a) } X_0\subset\mathscr{E}(P,1),\\ & \qquad \mathscr{E}(P,\rho_1)\subset\alpha_4X_\infty,\\ & \text{(b) (20), (21)}\\ & \text{(c) } \mathscr{E}(P,\rho_1)\subset\mathscr{L}(H_1),\\ & \text{(d) } \mathscr{E}(P,1)\subset\mathscr{L}(H_2).\end{aligned} \tag{23}$$

If we fix ρ_1 and η, then (23) can also be transformed into a convex optimization problem with LMI constraints. To obtain the global infimum, we may vary ρ_1 from 0 to 1 and η from 0 to ∞.

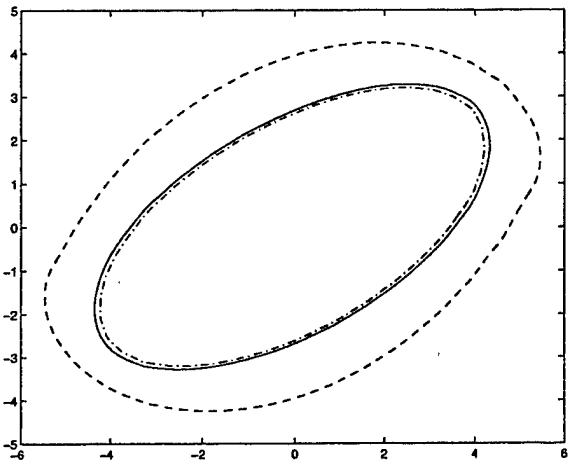


Fig. 2. The invariant ellipsoids and the null controllable region.

Example 2. The open-loop system is described by (15) with

$$A=\begin{bmatrix}0.6 & -0.8\\ 0.8 & 0.6\end{bmatrix},\quad B=\begin{bmatrix}2\\ 4\end{bmatrix},\quad E=\begin{bmatrix}0.1\\ 0.1\end{bmatrix}.$$

The system has a pair of unstable complex poles. We first ignore the disturbance and solve (13) for a feedback with the objective of maximizing the domain of attraction with respect to the unit ball, $X_{\mathrm{R}}=\mathscr{E}(I,1)$. The result is,

$$\alpha_1^*=1/(\gamma^*)^{1/2}=2.4417,$$
$$P_1^*=\begin{bmatrix}0.0752 & -0.0566\\ -0.0566 & 0.1331\end{bmatrix},$$
$$F_1^*=[-0.0025\quad -0.2987]$$

and the invariant ellipsoid is $\mathscr{E}(P_1^*,1)$ (see the larger ellipsoid in Fig. 2). As a comparison, we also plotted the boundary of the null controllable region (Hu & Lin, 2001a) of the open-loop system (see the dashed outer curve).

We next deal with Problem 2. By solving (18) with X_0 being a unit ball, we obtain $\alpha_2^*=2.3195$, with $\eta_2^*=0.019$. The resulting invariant ellipsoid is $\mathscr{E}(P_2^*,1)$, with

$$P_2^*=\begin{bmatrix}0.0835 & -0.0639\\ -0.0639 & 0.1460\end{bmatrix}$$

(see the inner dash–dotted ellipsoid in Fig. 2).

To deal with Problem 3, we solve (19), with X_∞ also being a unit ball. We obtain $\alpha_3^*=0.0606$, which shows that the disturbance can be rejected to a very small level. Now we turn to Problem 4. The optimization result by solving Problem 2 gives us some guide in choosing X_0. Here we choose $X_0=\mathscr{E}(I,2^2)$, $X_\infty=\mathscr{E}(I,1)$. The optimal solution is $\alpha_4^*=0.9725,\eta^*=0.006,\rho_1^*=0.0489$,

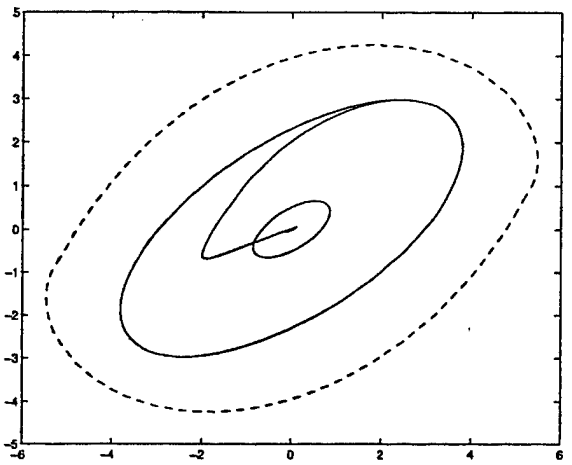


Fig. 3. The invariant ellipsoids and a trajectory.

$F_4^* = [0.2844 \quad -1.4430]$, and

$$P_4^* = \begin{bmatrix} 0.1145 & -0.0922 \\ -0.0922 & 0.1872 \end{bmatrix}.$$

In Fig. 3, the larger ellipsoid is $\mathscr{E}(P_4^*, 1)$, the smaller ellipsoid is $\mathscr{E}(P_4^*, \rho_1)$ and the outermost dashed closed curve is the boundary of the null controllable region. A trajectory is plotted with $x_0 \in \partial\mathscr{E}(P_4^*, 1)$ and $w = \text{sign}(\sin(0.3t))$.

4. Conclusions

We considered linear systems subject to actuator saturation and disturbance. A condition for determining if a given ellipsoid is contractively invariant was derived and shown to be less conservative than the existing conditions that are based on the circle criterion or the vertex analysis. With the aid of this condition, we developed analysis and design methods, both for closed-loop stability and disturbance rejection. Examples were used to demonstrate the effectiveness of these methods.

References

Blanchini, F. (1990). Feedback control for LTI systems with state and control bounds in the presence of disturbances. *IEEE Transaction on Automatic Control, AC-35*, 1231–1243.

Blanchini, F. (1994). Ultimate boundedness control for uncertain discrete-time systems via set-induced Lyapunov function. *IEEE Transaction on Automatic Control, AC-39*, 428–433.

Boyd, S., El Ghaoui, L., Feron, E., & Balakrishnan, V. (1994). *Linear matrix inequalities in systems and control theory*. SIAM Studies in Appl. Mathematics, Philadelphia.

Davison, E. J., & Kurak, E. M. (1971). A computational method for determining quadratic Lyapunov functions for non-linear systems. *Automatica, 7*, 627–636.

Gilbert, E. G., & Tan, K. T. (1991). Linear systems with state and control constraints: The theory and application of maximal output admissible sets. *IEEE Transaction on Automatic Control, AC-36*, 1008–1020.

Hindi, H., & Boyd, S. (1998). Analysis of linear systems with saturating using convex optimization. *Proceedings of the 37th IEEE Conference on Decision and Control*, Florida (pp. 903–908).

Hu, T., & Lin, Z. (2001a). *Control systems with actuator saturation: Analysis and design*. Boston: Birkhäuser.

Hu, T., & Lin, Z. (2001b). Practical stabilization of exponentially unstable linear systems subject to actuator saturation nonlinearity and disturbance. *International Journal of Robust Nonlinear Control, 11*, 555–588.

Khalil, H. (1996). *Nonlinear systems*. Upper Saddle River, NJ: Prentice-Hall.

Loparo, K. A., & Blankenship, G. L. (1978). Estimating the domain of attraction of nonlinear feedback systems. *IEEE Transaction on Automatic Control, AC-23*, 602–607.

Pittet, C., Tarbouriech, S., & Burgat, C. (1997). Stability regions for linear systems with saturating controls via circle and Popov criteria. *Proceedings of the 36th IEEE Conference on Decision and Control*, San Diego, CA (pp. 4518–4523).

Saberi, A., Lin, Z., & Teel, A. R. (1996). Control of linear systems with saturating actuators. *IEEE Transaction on Automatic Control, AC-41*, 368–378.

Vanelli, A., & Vidyasagar, M. (1985). Maximal Lyapunov functions and domain of attraction for autonomous nonlinear systems. *Automatica, 21*, 69–80.

Weissenberger, S. (1968). Application of results from the absolute stability to the computation of finite stability domains. *IEEE Transactions on Automatic Control, AC-13*, 124–125.

Tingshu Hu was born in Sichuan, China in 1966. She received her B.S. and M.S. degrees in Electrical Engineering from Shanghai Jiao Tong University, Shanghai, China, in 1985 and 1988, respectively, and a Ph.D degree in Electrical Engineering from University of Virginia, USA, in May 2001. Her research interests include systems with saturation nonlineari ties and robust control theory. She has published several papers in these areas. She is also a co-author (wiht Zongli Lin) of the book *Control Systems with Actuator Saturation: Analysis and Design* (Birkhäuser, Boston, 2001). She is currently an associate editor on the Conference Editorial Board of the IEEE Control Systems Society.

Zongli Lin was born in Fuqing, Fujian, China on February 24, 1964. He received his B.S. degree in Mathematics and Computer Science from Amoy University, Xiamen, China, in 1983, his Master of Engineering degree in automatic control from Chinese Academy of Space Technology, Beijing, China, in 1989, and his Ph.D. degree in Electrical and Computer Engineering from Washington State University, Pullman, Washington, in May 1994.

From July 1983 to July 1986, Dr. Lin worked as a control engineer at Chinese Academy of Space Technology. In January 1994, he joined the Department of Applied Mathematics and Statistics, State University of New York at Stony Brook as a visiting assistant professor, where he became an assistant professor in September 1994. Since July 1997, he has been with the Department of Electrical and Computer Engineering at University of Virginia, where he is currently an associate professor.

His current research interests include nonlinear control, robust control, and control of systems with saturating actuators. He has published several papers in these areas. He is also the author of the

book, *Low Gain Feedback* (Springer-Verlag, London, 1998) and a co-author (with Tingshu Hu) of the recent book *Control Systems with Actuator Saturation: Analysis and Design* (Birkhäuser, Boston, 2001).

A senior member of IEEE, Dr. Lin was an associate editor on the Conference Editorial Board of the IEEE Control Systems Society and currently serves as an Associate Editor of *IEEE Transactions on Automatic Control*. He is also a member of the IEEE Control Systems Society's Technical Committee on Nonlinear Systems and Control and heads its Working Group on Control with Constraints. He is the recipient of a US Office of Naval Research Young Investigator Award.

Ben M. Chen, born on November 25, 1963, in Fuqing, Fujian, China, received his B.S. degree in mathematics and computer science from Amoy University, Xiamen, China, in 1983, an M.S. degree in electrical engineering from Gonzaga University, Spokane, Washington, in 1988, and a Ph.D. degree in Electrical and Computer Engineering from Washington State University, Pullman, Washington, in 1991.

He worked as a software engineer from 1983 to 1986 in South-China Computer Corporation, China, and was an assistant professor from 1992 to 1993 in the Electrical Engineering Department of State University of New York at Stony Brook. Since 1993, he has been with the Department of Electrical and Computer Engineering, National University of Singapore, where he is currently an associate professor. His current research interests are in linear control and system theory, control applications, development of internet-based virtual laboratories and internet security systems.

He is an author or co-author of five monographs, *Hard Disk Drive Servo Systems* (London: Springer, 2001), *Robust and H_∞ Control* (London: Springer, 2000), *H_∞ Control and Its Applications* (London: Springer, 1998), *H_2 Optimal Control* (London: Prentice Hall, 1995), *Loop Transfer Recovery: Analysis and Design* (London: Springer, 1993), and one textbook, *Basic Circuit Analysis* (Singapore: Prentice Hall, 1st Ed., 1996; 2nd Ed., 1998). He was an associate editor in 1997–1998 on the Conference Editorial Board of IEEE Control Systems Society. He currently serves as an associate editor of IEEE Transactions on Automatic Control.

Publication 16

Transactions Briefs

Stability Analysis of Linear Time-Delay Systems Subject to Input Saturation

Yong-Yan Cao, Zongli Lin, and Tingshu Hu

Abstract—**This paper is devoted to stability analysis of linear systems with state delay and input saturation. The domain of attraction resulting from an *a priori* designed state feedback law is analyzed using Lyapunov–Razumikhin and Lyapunov–Krasovskii functional approach. Both delay-independent and delay-dependent estimation of the domain of attraction are presented using the linear matrix inequality technique. The problem of designing linear state feedback laws such that the domain of attraction is enlarged is formulated and solved as an optimization problem with LMI constraints. Numerical examples are used to demonstrate the effectiveness of the proposed design technique.**

Index Terms—**Actuator saturation, domain of attraction, linear matrix inequality, time-delay.**

I. Introduction

Nonlinear systems with time-delay constitute basic mathematical models of real phenomena, for instance, in circuits theory, economics and mechanics. Not only dynamical systems with time-delay are common in chemical processes and long transmission lines in pneumatic, hydraulic, or rolling mill systems, but computer controlled systems requiring numerical computation have time-delays in control loops. The presence of time-delays in control loops usually degrades system performance and complicates the analysis and design of feedback controllers. Stability analysis and synthesis of retarded systems is an important issue addressed by many authors and for which surveys can be found in several monographs (see *e.g.*, [7], [9], [10], [13], [17], [20]).

Another common, but difficult, control problem is to deal with actuator saturation since all control devices are subject to saturation (limited in force, torque, current, flow rate, etc.). The analysis and synthesis of controllers for dynamic systems subject to actuator saturation have been attracting increasingly more attention (see, for example, [1], [11], [14], [15] and the references therein).

Actuator saturation and time-delays are often observed together in control systems. To deal with both problems effectively, appropriate design methods are required. Up to now, only a few methods were reported to deal with these problems simultaneously. Chen *et al.* [5] studied the stabilization problem of saturating time-delay system with state feedback and sampled-state feedback and they derived several sufficient conditions to ensure the system stability in terms of norm inequalities. Chou *et al.* [6] exploited a sufficient condition to stabilize a linear uncertain time-delay system containing input saturation. The problem of robust stabilization of uncertain time-delay systems containing a saturating actuator was addressed by Niculescu *et al.* [16] by a high gain approach. Oucheriah [18] considered a method to synthesize a globally stabilizing state feedback controller by means of an asymptotic observer for time-delay systems. In [19], a dynamic anti-windup method was presented for the systems with input delay and saturation. All of these works have mainly focused on the stabilizability of the systems.

In this paper, we will first analyze the stability and domain of attraction for linear systems with time-delay in state and actuator saturation. A less conservative estimate of the domain of attraction will be derived based on the Lyapunov–Razumikin and Lypunov–Krasovskii functional approaches. This estimate is then maximized over the choice of the feedback gains. It is known that the estimates of the domain of attraction made by small gain theorem, Popov criterion or circle criterion are sometimes very conservative. In [12], a less conservative analysis approach is proposed to analyze the stability and the domain of attraction for systems with actuator saturation. The idea is to formulate the analysis problem into a constrained optimization problem with constraints given by a set of linear matrix inequalities (LMI's). In this paper, we will further exploit the idea in [12] to arrive at an estimate of the domain of attraction for the linear systems subject to both delay in state and actuator saturation. An LMI optimization approach will be proposed to design the state feedback gain which maximizes this estimate of the domain of attraction.

The paper is organized as follows. Section II gives some preliminary results and states more precisely our problem formulation. Delay-dependent and delay-independent stability and domain of attraction of the closed-loop system with input saturation and state delay will be analyzed in Sections III and IV respectively. Numerical examples illustrating our design procedure and its effectiveness are given in Section V. The paper is concluded in Section VI.

Notations: The following notations will be used throughout the paper. $\mathbf{R}$ denotes the set of real numbers, $\mathbf{R}^+$ denotes the set of nonnegative real numbers, $\mathbf{R}^n$ denotes the n dimensional Euclidean space and $\mathbf{R}^{m\times n}$ denotes the set of all $m \times n$ real matrices. The notation $X \geq Y$ (respectively, $X > Y$), where X and Y are symmetric matrices, means that $X - Y$ is positive semidefinite (respectively, positive definite). $\mathcal{C}_{n,\tau} = \mathcal{C}([-\tau, 0], \mathbf{R}^n)$ denotes the Banach space of continuous vector functions mapping the interval $[-\tau, 0]$ into $\mathbf{R}^n$ with the topology of uniform convergence. The following norms will be used: $\|\cdot\|$ refers to either the Euclidean vector norm or the induced matrix 2-norm; $\|\phi\|_c = \sup_{-\tau\leq t\leq 0}\|\phi(t)\|$ stands for the norm of a function $\phi \in \mathcal{C}_{n,\tau}$. Moreover, we denote by $\mathcal{C}^v_{n,\tau}$ the set $\mathcal{C}^v_{n,\tau} = \{\phi \in \mathcal{C}_{n,\tau}: \|\phi\|_c < v\}$, where v is a positive real number.

II. Problem Statement and Preliminaries

A. Problem Statement

Let us consider the linear system with time-delay in state and input saturation

$$\dot{x}(t) = Ax(t) + A_d x(t-\tau) + B\sigma(u(t)) \tag{1}$$

$$x(t) = \psi(t), \quad t \in [-\tau, 0] \tag{2}$$

where $x \in \mathbf{R}^n$ is the state, $u \in \mathbf{R}^m$ the control input, τ a constant and A, A_d and B are known matrices. Assume that the initial condition ψ is a continuous vector-valued function, *i.e.*, $\psi \in \mathcal{C}_{n,\tau}$. The function $\sigma: \mathbf{R}^m \to \mathbf{R}^m$ is the standard saturation function defined as follows:

$$\sigma(u) = [\sigma(u_1) \quad \sigma(u_2) \cdots \sigma(u_m)]^T$$

Manuscript received March 8, 2001; revised August, 6 2001 and August13, 2001. This work was supported in part by the U.S. Office of Naval Research Young Investigator Program under Grant N00014-99-1-0670. This paper was commended by Associate Editor G. Chen.

The authors are with the Department of Electrical & Computer Engineering, University of Virginia, Charlottesville, VA 22903 USA (e-mail: yycao@virginia.edu; zl5y@virginia.edu; th7f@virginia.edu).

Publisher Item Identifier S 1057-7122(02)01179-0.

1057–7122/02$17.00 © 2002 IEEE

where $\sigma(u_i) = \text{sign}(u_i)\min\{1, |u_i|\}$. Here we have slightly abused the notation by using σ to denote both the scalar valued and the vector valued saturation functions. Also, note that it is without loss of generality to assume unity saturation level. We use $x_t \in \mathcal{C}_{n,\tau}$ to denote the restriction of $x(t)$ to the interval $[t-\tau, t]$ translated to $[-\tau, 0]$, that is, $x_t(\theta) = x(t+\theta), \quad \theta \in [-\tau, 0]$.

In this paper, we consider the control of the system (1) using a linear state feedback $u = Fx$. The closed-loop system under this feedback is given by

$$\dot{x}(t) = Ax(t) + A_d x(t-\tau) + B\sigma(Fx(t)), \quad x_0 = \psi \in \mathcal{C}_{n,\tau}. \quad (3)$$

We will be interested in the stability analysis and design for (3). For an initial condition $x_0 \in \mathcal{C}_{n,\tau}$, denote the state trajectory of the system (3) as $x(t, x_0)$. Suppose that the solution $x(t) \equiv 0$ is asymptotically stable, then the domain of attraction of the origin is

$$\mathcal{S} := \left\{ x_0 \in \mathcal{C}_{n,\tau}: \quad \lim_{t\to\infty} x(t, x_0) = 0 \right\}.$$

A set $\mathcal{X} \subset \mathcal{C}_{n,\tau}$ is said to be invariant if

$$x_0 \in \mathcal{X} \Longrightarrow x_t \in \mathcal{X} \qquad \forall t \geq 0.$$

In general, given a stabilizing state feedback $u = Fx$, it is impossible to determine exactly the domain of attraction of the origin. The objective of this paper is to obtain an estimate of the domain of attraction for (3). The problems to be studied in this paper are the following.

Problem 1: Given a state feedback matrix F and a set of initial conditions $\mathcal{D}$, determine if $\mathcal{D} \subset \mathcal{S}$.

Problem 2: Design an F such that an estimate of the domain of attraction is maximized.

B. Razumikhin Theorem and Krasovskii Theorem

For stability analysis of systems with time-delay, the Razumikhin Theorem and Krasovskii Theorem are used extensively. In what follows, we give a brief summary of the two theorems simplified to autonomous systems.

Consider the retarded functional differential equation

$$\dot{x}(t) = f(x_t), \quad t \geq 0 \quad (4)$$

$$x(t) = \psi(t), \quad t \in [-\tau, 0]. \quad (5)$$

Assume that $\psi \in \mathcal{C}_{n,\tau}$ and the map $f(\psi)$: $\mathcal{C}_{n,\tau} \longmapsto \mathbf{R}^n$ is continuous and Lipschitzian in ψ and $f(0) = 0$. Also denote the solution of the functional differential (4) with the initial condition $x_0 \in \mathcal{C}_{n,\tau}$ as $x(t, x_0)$.

Definition 1: The trivial solution $x(t) \equiv 0$ of (4) and (5) is said to be asymptotically stable if

1) for every $\delta > 0$ there exists an $\epsilon = \epsilon(\delta)$ such that for any $\psi \in \mathcal{B}(0, \epsilon)$ the solution $x(t, \psi)$ of (4) and (5) satisfies $x_t \in \mathcal{B}(0, \delta)$ for all $t \geq 0$.
2) for every $\eta > 0$ there exist a $T(\eta)$ and a $v_0 > 0$ independent of η such that $\psi \in \mathcal{B}(0, v_0)$ implies that $\|x_t\|_c < \eta, \forall t \geq T(\eta)$.

The Krasovskii Theorem and the Razumikhin Theorem give conditions for $x(t) \equiv 0$ to be asymptotically stable. Actually, more information about invariant set and regional stability is contained in the proofs for these theorems in [9]. The additional information is incorporated in the following statement of these theorems.

Theorem 1 (Krasovskii Stability Theorem): Suppose that the function f: $\mathcal{C}_{n,\tau} \longmapsto \mathbf{R}^n$ takes bounded sets of $\mathcal{C}_{n,\tau}$ in bounded sets of $\mathbf{R}^n$ and suppose that $u(s)$, $v(s)$ and $w(s)$ are scalar, continuous, positive and nondecreasing functions. If there is a continuous function V: $\mathcal{C}_{n,\tau} \longmapsto \mathbf{R}^+$ and a positive number ρ such that for all $x_t \in L_V(\rho) := \{\psi \in \mathcal{C}_{n,\tau} : V(\psi) \leq \rho\}$, the following conditions hold.

1) $u(\|x_t(0)\|) \leq V(x_t) \leq v(\|x_t\|_c)$.
2) $\dot{V}(x_t) \leq -w(\|x_t(0)\|)$.

Then, the solution $x(t) \equiv 0$ of the (4) and (5) is asymptotically stable. Moreover, the set $L_V(\rho)$ is an invariant set inside the domain of attraction.

Theorem 2 (Razumikhin Stability Theorem): Suppose that $u(s)$, $v(s)$, $w(s)$ and $p(s) \in \mathbf{R}^+ \longmapsto \mathbf{R}^+$ are scalar, continuous and nondecreasing functions, $u(s)$, $v(s)$, $w(s)$ positive for $s > 0$, $u(0) = v(0) = 0$ and $p(s) > s$ for $s > 0$. If there is a continuous function V: $\mathbf{R}^n \longmapsto \mathbf{R}$ and a positive number ρ, such that for all $x_t \in M_V(\rho) := \{\psi \in \mathcal{C}_{n,\tau}: V(\psi(\theta)) \leq \rho, \forall \theta \in [-\tau, 0]\}$, the following conditions hold.

1) $u(\|x\|) \leq V(x) \leq v(\|x\|)$.
2) $\dot{V}(x(t)) \leq -w(\|x(t)\|)$, *if* $V(x(t+\theta)) < p(V(x(t))), \forall \theta \in [-\tau, 0]$.

Then, the solution $x(t) \equiv 0$ of the (4) and (5) is asymptotically stable. Moreover, the set $M_V(\rho)$ is an invariant set inside the domain of attraction.

C. Some Mathematical Tools

Let f_i be the i-th row of the matrix F. We define the symmetric polyhedron

$$\mathcal{L}(F) = \{x \in \mathbf{R}^n: |f_i x| \leq 1, \quad i = 1, \ldots, m\}.$$

If the control u does not saturate for all $i = 1, \ldots, m$, that is $x \in \mathcal{L}(F)$, then the nonlinear system (3) admits the following linear representation:

$$\dot{x}(t) = (A + BF)x(t) + A_d x(t-\tau). \quad (6)$$

Let $P \in \mathbf{R}^{n\times n}$ be a positive-definite matrix. For a number $\rho > 0$, the ellipsoid $\Omega(P, \rho)$ is defined by

$$\Omega(P, \rho) := \left\{ x \in \mathbf{R}^n: x^T P x \leq \rho \right\}.$$

Let $\mathcal{V}$ be the set of $m \times m$ diagonal matrices whose diagonal elements are either 1 or 0. Then there are 2^m elements in $\mathcal{V}$. Suppose that each element of $\mathcal{V}$ is labeled as $D_i, i = 1, 2, \ldots, 2^m$ and denote $D_i^- = I - D_i$. Clearly, D_i^- is also an element of $\mathcal{V}$ if $D_i \in \mathcal{V}$.

Lemma 1 [11]: Let $F, H \in \mathbf{R}^{m\times n}$ be given. For $x \in \mathbf{R}^n$, if $\|Hx\|_\infty \leq 1$, then

$$\sigma(Fx) \in \text{co}\{D_i Fx + D_i^- Hx: i \in [1, 2^m]\}$$

where co$\{\cdot\}$ denotes the convex hull of a set.

Lemma 2 [3]: For any $x, y \in \mathbf{R}^n$ and a matrix $M > 0$ with compatible dimensions, the following inequality holds

$$2x^T y \leq x^T M x + y^T M^{-1} y.$$

III. Delay-Independent Analysis

In this section, we will give methods for estimating the domain of attraction for the system (3) with invariant sets. We will first give conditions for a set to be an invariant set inside the domain of attraction and then use optimization approach to enlarge the invariant set by choosing the feedback gain matrix F and the Lyapunov function.

A. Razumikhin Functional Approach

Theorem 3: Let $F \in \mathbf{R}^{m\times n}$ be given. For a positive definite matrix $P \in \mathbf{R}^{n\times n}$ and a number $\rho > 0$, consider the set

$$M_V(\rho) = \{\psi \in \mathcal{C}_{n,\tau}: \psi(\theta) \in \Omega(P, \rho) \qquad \forall \theta \in [-\tau, 0]\}.$$

If there exist two matrices $H \in \mathbf{R}^{m\times n}$ and $W \in \mathbf{R}^{n\times n}, W > 0$ such that

$$\begin{aligned} &(A + B(D_i F + D_i^- H))^T P \\ &\quad + P(A + B(D_i F + D_i^- H)) \\ &\quad + P A_d W A_d^T P + P \\ &< 0, \ i \in [1, 2^m], \end{aligned} \quad (7)$$

$$P \geq W^{-1} \quad (8)$$

and $\Omega(P,\rho) \subset \mathcal{L}(H)$, i.e., $|h_i x| \leq 1$ for all $x \in \Omega(P,\rho)$, $i = 1,2,\ldots,m$, then the solution $x(t) \equiv 0$ is asymptotically stable for the system (3) and the set $M_V(\rho)$ is an invariant set inside the domain of attraction.

Proof: Given $P > 0$, consider a quadratic Lyapunov function candidate $V(x) = x^T P x$. First, we have $\varepsilon_1\|x\|^2 \leq V(x) \leq \varepsilon_2\|x\|^2$, where $\varepsilon_1 = \lambda_{\min}(P)$, $\varepsilon_2 = \lambda_{\max}(P)$. The derivative of V along the solutions of (3) is

$$\dot{V}(x(t)) = 2x(t)^T PAx(t) + 2x^T(t)PA_d x(t-\tau) + 2x^T(t)PB\sigma(Fx(t)).$$

In what follows, we will be interested in $x_t \in M_V(\rho)$. In this case, $x(t) \in \Omega(P,\rho)$. Since $|h_i x| \leq 1$ for all $x \in \Omega(P,\rho), i = 1,2,\ldots,m$, by Lemma 1, for every $x(t) \in \Omega(P,\rho)$

$$\sigma(Fx(t)) \in \operatorname{co}\left\{\left(D_i F + D_i^- H\right) x(t) : i = 1,\ldots,2^m\right\}.$$

It follows that for every $x(t) \in \Omega(P,\rho)$, we have

$$\dot{V}(x(t)) \leq \max_{i\in[1,2^m]} 2x^T(t)P\left(A + B\left(D_i F + D_i^- H\right)\right)x(t) + 2x^T(t)PA_d x(t-\tau).$$

From Lemma 2 and (8), we further have

$$\dot{V}(x(t)) \leq \max_{i\in[1,2^m]} x^T(t)\Big(\left(A + B\left(D_i F + D_i^- H\right)\right)^T P + P\left(A + B\left(D_i F + D_i^- H\right)\right) + PA_d W A_d^T P\Big)x(t) + V(x(t-\tau)). \quad (9)$$

By Razumikhin Theorem, to prove that $M_V(\rho)$ is an invariant set inside the domain of attraction, it suffices to show that there exist an $\varepsilon > 1$ and a $\delta > 0$ such that

$$\dot{V}(x(t)) \leq -\delta V(x(t)), \quad \text{if} \quad V(x(t+\theta)) < \varepsilon V(x(t)) \quad \forall\theta \in [-\tau, 0]. \quad (10)$$

In the remainder of the proof, we will construct such ε and δ and show that they satisfy (10).

From (7), we see that there exists a $\delta > 0$ such that

$$\left(A + B\left(D_i F + D_i^- H\right)\right)^T P + P\left(A + B\left(D_i F + D_i^- H\right)\right) + PA_d W A_d^T P + (1+2\delta)P < 0, \quad i \in [1, 2^m].$$

Let $\varepsilon = 1 + \delta$. Now suppose that $V(x(t+\theta)) < \varepsilon V(x(t))$, $\forall\theta \in [-\tau, 0]$. Then from (9), we have

$$\dot{V}(x) \leq \max_{i\in[1,2^m]} x^T\Big(\left(A + B\left(D_i F + D_i^- H\right)\right)^T P + P\left(A + B\left(D_i F + D_i^- H\right)\right) + PA_d W A_d^T P + \varepsilon P\Big)x < -\delta V(x(t)).$$

This completes the proof. ∎

Note that the condition of Theorem 3 does not include any information of time-delay, *i.e.*, the theorem provides a delay-independent condition for regional stability of linear time-delay systems with input saturation in terms of the feasibility of several linear matrix inequalities. This result can also be easily extended to systems with multiple time-varying time-delays in state [2].

Remark 1: In practice, we may be interested in the stability region in which the asymptotic stability of closed-loop system (3) is guaranteed under saturation and the linear closed-loop system (6) (i.e., unsaturated closed-loop system) is β-stable. As shown in [17], β-stability is equivalent to

$$\dot{y}(t) = (A + BF + \beta I)y(t) + e^{\beta\tau}A_d y(t-\tau)$$

which is stable. This can be guaranteed by the following matrix inequality:

$$(A+BF)^T P + P(A+BF) + e^{2\beta\tau}PA_d W A_d^T P + 2\beta P < 0. \quad (11)$$

Remark 2: If the matrix A_d is rank deficiency, i.e. there exists a decomposition $A_d = D_d E_d$, where $D_d \in \mathbb{R}^{n\times p}$, $E_d \in \mathbb{R}^{p\times n}$, $p < n$, then we can prove with similar arguments that $M_V(\rho)$ is an invariant set inside the domain of attraction if there exist two matrices $H \in \mathbb{R}^{m\times n}$ and $W \in \mathbb{R}^{p\times p} > 0$ satisfying the matrix inequalities

$$\left(A + B\left(D_i F + D_i^- H\right)\right)^T P + P\left(A + B\left(D_i F + D_i^- H\right)\right) + PD_d W D_d^T P + P < 0, \quad i \in [1, 2^m], \qquad E_d^T W^{-1} E_d \leq P.$$

With all the $M_V(\rho)$ satisfying the set invariance condition, we would like to choose the "largest" one to obtain the least conservative estimate of the domain of attraction by the method introduced in [12]. We see that the "size" of the set $M_V(\rho)$ is proportional to the size of $\Omega(P,\rho)$. Here we would like to take the shape of $\Omega(P,\rho)$ into consideration. For this purpose, we introduced the notion of shape reference set as in [12]. Let $\mathcal{X}_R \subset \mathbb{R}^n$ be a prescribed bounded convex set containing origin. For a set $S \subset \mathbb{R}^n$ containing origin, define the size of S with respect to $\mathcal{X}_R$ as

$$\alpha_R(S) := \sup\{\alpha > 0 : \alpha\mathcal{X}_R \subset S\}.$$

Obviously, if $\alpha_R(S) \geq 1$, then $\mathcal{X}_R \subset S$. Two typical types of $\mathcal{X}_R$ are the ellipsoid

$$\mathcal{X}_R = \{x \in \mathbb{R}^n : x^T R x \leq 1\} \quad R > 0$$

and the polyhedron

$$\mathcal{X}_R = \operatorname{co}\{x_1, x_2, \ldots, x_l\}$$

where $x_1, x_2, \ldots, x_l$ are some given points in $\mathbb{R}^n$.

Theorem 3 gives a condition for a set $M_V(\rho)$ to be inside the domain of attraction for the closed-loop time-delay system subject to input saturation (3). With a given shape reference set, we can choose from all the $\Omega(P,\rho)$'s that satisfy the condition such that the quantity $\alpha_R(\Omega(P,\rho))$ is maximized. This problem can be formulated as

$$\sup_{P\geq W^{-1}>0,\rho,H} \alpha, \quad \text{s.t.}$$

$$a)\ \alpha\mathcal{X}_R \subset \Omega(P,\rho),$$

$$b)\ \left(A + B\left(D_i F + D_i^- H\right)\right)^T P + P\left(A + B\left(D_i F + D_i^- H\right)\right) + PA_d W A_d^T P + P < 0, \quad i \in [1, 2^m],$$

$$c)\ |h_i x| \leq 1,\ \forall x \in \Omega(P,\rho), \quad i \in [1, m]. \quad (12)$$

Let $Q = (\rho^{-1}P)^{-1}$, $\gamma = 1/\alpha^2$ and $G = HQ$. With similar procedure as in [12], we can transform the above optimization problem to an LMI problem. That is, if we substitute ρW with W, then for the case that $\mathcal{X}_R$ is a polyhedron, the optimization problem (12) can be rewritten as follows:

$$\inf_{W\geq Q>0,G} \gamma, \quad \text{s.t.}$$

$$a)\ \begin{bmatrix} \gamma & x_i^T \\ x_i & Q \end{bmatrix} \geq 0,\ i \in [1, l].$$

$$b)\ QA^T + AQ + B\left(D_i FQ + D_i^- G\right) + \left(D_i FQ + D_i^- G\right)^T B^T + A_d W A_d^T + Q < 0, i \in [1, 2^m].$$

$$c)\ \begin{bmatrix} 1 & g_i \\ g_i^T & Q \end{bmatrix} \geq 0, \quad i \in [1, m]. \quad (13)$$

If $\mathcal{X}_R$ is an ellipsoid, then we need to replace $a)$ in (13) with

$$\alpha^{-2}R \geq \rho^{-1}P \Longleftrightarrow R^{-1} \leq \gamma Q.$$

As is proven in [12], for systems without delay, *i.e.*, $A_d = 0$, solving the above LMI optimization problem will give a less conservative estimate of the domain of attraction than other methods resulting from, for example, the circle criterion.

If the unsaturated system is required to have some stability margin, *i.e.*, it is required to be β-stable, based on Remark 1, the additional LMI constraint (11) needs to be added to optimization problem (13), leading to the following LMI optimization problem:

$$\inf_{W \geq Q > 0, G} \gamma, \quad \text{s.t.}$$
$$a)\ b) \text{ and } c) \text{ in (13).}$$
$$d)\ QA^T + AQ + BFQ + (BFQ)^T + e^{2\beta\tau} A_d W A_d^T + 2\beta Q < 0. \tag{14}$$

The problem of designing a feedback matrix F such that the estimate of the domain of attraction is enlarged can be formulated by simply taking the parameter F in (13) as a variable for optimization. To do so, we just need to replace $Y = FQ$ in (13b) with a new varaiable Y.

B. Krasovskii Functional Approach

In this subsection, we will consider the following Lyapunov-Krasovskii functional:

$$V(x_t) := x^T(t)Px(t) + \int_{t-\tau}^{t} x^T(s)Wx(s)ds \tag{15}$$

where $P > 0$ and $W > 0$. This type of functional has been widely used for stability analysis of time-delay systems (see, *e.g.*, [9]).

Theorem 4: Let the feedback gain $F \in \mathbb{R}^{m\times n}$ be given. For given $P, W > 0$ and $\rho > 0$, consider the set

$$L_V(\rho) = \left\{ \psi \in \mathcal{C}_{n,\tau} : \psi^T(0)P\psi(0) + \int_{-\tau}^{0} \psi^T(s)W\psi(s)ds \leq \rho \right\}. \tag{16}$$

If there exists a matrix $H \in \mathbb{R}^{m\times n}$ such that we get (17) shown at the bottom of the page and $\Omega(P,\rho) \subset \mathcal{L}(H)$, then the solution $x(t) \equiv 0$ of the system (3) is asymptotically stable. Moreover, the set $L_V(\rho)$ is an invariant set inside the domain of attraction.

Proof: Consider the Lyapunov functional given by (15). First, we have

$$\varepsilon_1 \|x_t(0)\|^2 \leq V(x_t) \leq \varepsilon_2 \|x_t\|_c^2$$

where $\varepsilon_1 = \lambda_{\min}(P)$, $\varepsilon_2 = \lambda_{\max}(P) + \tau\lambda_{\max}(W)$. Then

$$\dot{V}(x_t) = x^T(t)(A^TP + PA + W)x(t) + 2x^T(t)PA_dx(t-\tau) + 2x^T(t)PB\sigma(Fx(t)) - x^T(t-\tau)Wx(t-\tau).$$

We will be interested in $x_t \in L_V(\rho)$. In this case, $x(t) \in \Omega(P,\rho) \subset \mathcal{L}(H)$ and we have

$$\sigma(Fx(t)) \in \operatorname{co}\left\{ \left(D_iF + D_i^- H\right) x(t) : i \in [1, 2^m] \right\}.$$

With similar arguments as in the proof of Theorem 3, we get the second equation shown at the bottom of the page where $\xi^T(t) = \left[x^T(t)\ x^T(t-\tau)\right]$. Under the condition (17), there exists a $\delta > 0$ such that we get the third equation shown at the bottom of the page. It follows that

$$\dot{V}(x_t) < -\delta x(t)^T Px(t) \leq -\delta\varepsilon_1 \|x(t)\|_2^2.$$

By Krasovskii Stability Theorem, $L_V(\rho)$ is an invariant set inside the domain of attraction. ∎

As an estimate of the domain of attraction, the invariant set $L_V(\rho)$ in Theorem 4 depends not only on the P matrix, but also on an integration over $[-\tau, 0]$. This makes the structure of the set $L_V(\rho)$ much more complicated than the invariant set $M_V(\rho)$ in Theorem 3 based on Lyapunov-Razumikhin functional approach. Hence, it is not easy to measure the size of the set $L_V(\rho)$. Because of this, we would like to determine a subset of $L_V(\rho)$ which is of a more regular shape, say, like $M_V(\rho)$ in Theorem 3.

Let $z(t) = P^{1/2}x(t)$. Then

$$\|z_t\|_c = \sup_{-\tau\leq\theta\leq 0} \|z(t)\| = \sup_{-\tau\leq\theta\leq 0} \left(x^T(t)Px(t)\right)^{1/2}$$

and $$V(x_t) \leq \left(1 + \tau\lambda_{\max}\left(P^{-1/2}WP^{-1/2}\right)\right)\|z_t\|_c^2.$$

Let

$$\rho_1 = \frac{\rho}{1 + \tau\lambda_{\max}\left(P^{-1/2}WP^{-1/2}\right)}.$$

Then, we have

$$M(\rho_1) = \left\{ \psi \in \mathcal{C}_{n,\tau} : \psi(\theta)^T P\psi(\theta) \leq \rho_1, \ \forall\theta \in [-\tau, 0] \right\} \subset L_V(\rho).$$

On the other hand, let

$$\delta = \frac{\rho}{\lambda_{\max}(P) + \tau\lambda_{\max}(W)}$$

then the ball $\mathcal{B}(\delta) = \{\psi \in \mathcal{C}_{n,\tau} : \|\psi\|_c^2 < \delta\}$ is inside the domain of attraction. We see that the size of $M(\rho_1)$ is proportional to the size of $\Omega(P,\rho_1)$. With a given $\mathcal{X}_R$, we can choose from all the $\Omega(P,\rho_1)$'s

$$\begin{bmatrix} \left(A + B\left(D_iF + D_i^-H\right)\right)^T P + P\left(A + B\left(D_iF + D_i^-H\right)\right) + W & PA_d \\ A_d^TP & -W \end{bmatrix} < 0,\ i \in [1, 2^m] \tag{17}$$

$$\dot{V}(x_t) \leq \max_{i\in[1,2^m]} \xi(t)^T \begin{bmatrix} \left(A + B\left(D_iF + D_i^-H\right)\right)^T P + P\left(A + B\left(D_iF + D_i^-H\right)\right) + W & PA_d \\ A_d^TP & -W \end{bmatrix} \xi(t)$$

$$\begin{bmatrix} \left(A + B\left(D_iF + D_i^-H\right)\right)^T P + P\left(A + B\left(D_iF + D_i^-H\right)\right) + W & PA_d \\ A_d^TP & -W \end{bmatrix} < -\begin{bmatrix} \delta P & 0 \\ 0 & 0 \end{bmatrix},\quad i \in [1, 2^m].$$

such that the quantity $\alpha_R(\Omega(P,\rho_1))$ is maximized. This problem can be formulated as shown in (18) at the bottom of the page.

As in Section III-A, we can cast the problem into the LMI framework. Let $Q=(\rho^{-1}P)^{-1}, \gamma=1/\alpha^2, G=HQ$ and substitute ρW^{-1} with X, we can reduce the optimization problem (18) to the one with LMI constraints as shown in (19) at the bottom of the page. If $\mathcal{X}_R$ is an ellipsoid, then $a)$ should be replaced with

$$\alpha^{-2}R-\left(1+\tau\lambda_{\max}\left(P^{-1/2}WP^{-1/2}\right)\right)\frac{P}{\rho}>0 \Leftrightarrow (1+\tau\varepsilon)R^{-1}<\gamma Q.$$

Also, as in Section III-A, a controller design problem can be readily formulated by taking F in (19) as an optimizing parameter.

IV. Delay-Dependent Analysis

To reduce conservativeness in the analysis when the information on the delay is available, in this section, we will establish a delay-dependent stability result for the time-delay system (3) with input saturation. For simplicity, denote $\hat{A}_i := A+A_d+B(D_iF+D_i^-H)$.

Theorem 5: Let the state feedback gain F be given. Consider the ellipsoid $\Omega(P,\rho)$. If there exist matrices $H\in\mathbf{R}^{m\times n}, P_1, P_2\in\mathbf{R}^{n\times n}$, $P_1>0$, $P_2>0$ and $\tau_0>0$ such that

$$\hat{A}_i^TP+P\hat{A}_i+\tau_0PA_d(P_1+P_2)A_d^TP+2\tau_0P<0,\quad i\in[1,2^m] \tag{20}$$

$$\left[A+B(D_iF+D_i^-H)\right]^TP_1^{-1}\left[A+B\left(D_iF+D_i^-H\right)\right]\le P,\quad i\in[1,2^m] \tag{21}$$

$$A_d^TP_2^{-1}A_d\le P \tag{22}$$

and $\Omega(P,\rho)\subset\mathcal{L}(H)$, then $x(t)\equiv 0$ of the system (3) is delay-dependent asymptotically stable. Moreover, for any time-delay $\tau\le\tau_0$ and any initial condition ψ, $\psi(\theta)\in\Omega(P,\rho)$, $\forall\theta\in[-\tau,0]$, we have $\lim_{t\to\infty}x(t)=0$.

Proof: Since $x(t)$ is continuously differentiable for $t\ge 0$, using the Leibniz-Newton formula, one can write

$$\begin{aligned}x(t-\tau)=&x(t)-\int_{t-\tau}^{t}\dot{x}(s)ds\\ =&x(t)-\int_{t-\tau}^{t}\Big[Ax(s)+A_dx(s-\tau)\\ &+B\sigma(Fx(s))\Big]ds\end{aligned} \tag{23}$$

for $t\ge\tau$. Thus the system (3) can be rewritten as

$$\begin{aligned}\dot{x}(t)=&(A+A_d)x(t)\\ &-A_d\int_{t-\tau}^{t}\Big[Ax(s)+A_dx(s-\tau)+B\sigma(Fx(s))\Big]ds\\ &+B\sigma(Fx(t))\\ x(t)=&\psi(t),\quad\in[-2\tau,0]\end{aligned} \tag{24}$$

where $\psi\in C_{n,2\tau}$. By [9], [21], the asymptotic stability of the above system will ensure the asymptotic stability of the original time-delay system (3).

Choose the Lyapunov functional candidate as $V(x(t))=x^T(t)Px(t)$. To prove the theorem, it suffices to show that $x(t)\equiv 0$ is asymptotically stable for the system (24) and that the set

$$M_V(\rho)=\{\psi\in C_{n,2\tau}:\psi(\theta)\in\Omega(P,\rho),\ \forall\theta\in[-2\tau,0]\}$$

is an invariant set inside the domain of attraction. Here we use x_t to denote the restriction of $x(t)$ to the interval $[t-2\tau,t]$ translated to $[-2\tau,0]$, that is, $x_t(\theta)=x(t+\theta),\quad\theta\in[-2\tau,0]$.

We are interested in $x_t\in M_V(\rho)$. In this case, $x(t)\in\Omega(P,\rho)$ and we have

$$\begin{aligned}\dot{V}(x(t))\le&2\max_{i\in[1,2^m]}x^T(t)P\hat{A}_ix(t)\\ &+\tau x^T(t)PA_d(P_1+P_2)A_d^TPx(t)\\ &+\int_{-\tau}^{0}[Ax(t+s)+B\sigma(Fx(t+s))]^TP_1^{-1}\\ &\times[Ax(t+s)+B\sigma(Fx(t+s))]ds\\ &+\int_{-\tau}^{0}\Big[x^T(t-\tau+s)A_d^TP_2^{-1}A_dx(t-\tau+s)ds.\end{aligned}$$

By the convexity of the function $x^TP_1^{-1}x$ and Lemma 1, we have

$$\begin{aligned}&[Ax(t)+B\sigma(Fx(t))]^TP_1^{-1}[Ax(t)+B\sigma(Fx(t))]\\ &\le\max_{i\in[1,2^m]}x(t)^T(A+B(D_iF+D_i^-H))^T\\ &\cdot P_1^{-1}(A+B(D_iF+D_i^-H))x(t).\end{aligned}$$

It follows from (21) that

$$[Ax(t)+B\sigma(Fx(t))]^T\cdot P_1^{-1}[Ax(t)+B\sigma(Fx(t))]\le x^T(t)Px(t)$$

and from (22), we have

$$x^T(t)A_d^TP_2^{-1}A_dx(t)\le x^T(t)Px(t).$$

$$\sup_{P>0,\rho,H}\ \alpha,\quad\text{s.t.}$$

$a)$ $\alpha\mathcal{X}_R\subset\Omega(P,\rho_1)$.

$b)$ $\begin{bmatrix}\left(A+B\left(D_iF+D_i^-H\right)\right)^TP+P\left(A+B\left(D_iF+D_i^-H\right)\right)+W & PA_d\\ A_d^TP & -W\end{bmatrix}<0,\quad i\in[1,2^m]$.

$c)$ $|h_ix|\le 1,\ \forall x\in\Omega(P,\rho),\ i\in[1,m]$. (18)

$$\inf_{Q>0,X>0,G,\rho}\ \gamma,\ \text{s.t.}$$

$a)$ $\begin{bmatrix}\gamma & (1+\tau\varepsilon)x_i^T\\ (1+\tau\varepsilon)x_i & (1+\tau\varepsilon)Q\end{bmatrix}>0,\ Q\le\varepsilon X,\ i\in[1,l]$.

$b)$ $\begin{bmatrix}QA^T+AQ+B\left(D_iFQ+D_i^-G\right)+\left(D_iFQ+D_i^-G\right)^TB^T+A_dXA_d^T & Q\\ Q & -X\end{bmatrix}<0, i\in[1,2^m]$.

$c)$ Constraint (13c). (19)

Hence

$$\begin{aligned}\dot{V}(x(t)) \le & 2 \max_{i\in[1,2^m]} x^T(t)P\hat{A}_i x(t) \\ & + \tau x^T(t)PA_d(P_1+P_2)A_d^T P x(t) \\ & + \int_{-\tau}^{0} V(x(t+s))ds \\ & + \int_{-\tau}^{0} V(x(t-\tau+s))ds. \end{aligned} \tag{25}$$

By Razumikhin Theorem, to show that $M_V(\rho)$ is an invariant set inside the domain of attraction, it suffices to construct an $\varepsilon > 1$ and a $\delta > 0$ such that

$$\begin{aligned} & \dot{V}(x(t)) < -\delta V(x(t)), \\ & \text{if } V(x(t+\theta)) < \varepsilon V(x(t)) \qquad \forall\theta \in [-2\tau, 0]. \end{aligned} \tag{26}$$

Under the condition of (20), there exists a $\delta_1 > 0$ such that

$$\hat{A}_i^T P + P\hat{A}_i + \tau_0 PA_d(P_1+P_2)A_d^T P + 2\tau_0(1+2\delta_1)P < 0, \qquad i \in [1, 2^m].$$

Let $\varepsilon = 1+\delta_1$. Suppose that $V(x(t-\theta)) < \varepsilon V(x(t)) \quad \forall\theta \in [-2\tau, 0]$. Then from (25), we have

$$\begin{aligned}\dot{V}(x(t)) \le & 2 \max_{i\in[1,2^m]} x^T(t)P\hat{A}_i x(t) \\ & + \tau_0 x^T(t)PA_d(P_1+P_2)A_d^T P x(t) \\ & + 2\tau_0 \varepsilon x(t)^T P x(t) \\ = & \max_{i\in[1,2^m]} x^T \Big(\hat{A}_i^T P + P\hat{A}_i + \tau_0 PA_d(P_1+P_2)A_d^T P \\ & + 2\tau_0\varepsilon P \Big) x(t) \\ < & -2\tau_0\delta_1 x(t)^T P x(t). \end{aligned}$$

This completes our proof. ■

Remark 3: By letting $Q = \rho P^{-1}$, $G = HQ$ and $\bar{A}_i = A + A_d$, we see that the matrix inequalities (20) to (22) are equivalent to the LMIs shown in (27)–(29) at the bottom of the page where we have replaced P_1 and P_2 with P_1/ρ and P_2/ρ.

Theorem 5 provides a delay-dependent condition for regional stability of linear time-delay systems with input saturation in terms of the feasibility of several linear matrix inequalities. This result can also be easily extended to systems with multiple time-varying time-delays in state [3]. Note that in the proof the transformation (23) is used to transform the time-delay system with single time delay to a system with distributed delay. It is shown in [8] that such a transformation may incur some additional dynamics that can be characterized by appropriate additional eigenvalues. And hence, if the smallest of such delays is less than the stability delay limit of the original system, then any stability criteria obtained using such transformation will be conservative.

Remark 4: Theorems 4 and 5 can also be strengthened when A_d is rank deficiency as in Remark 2.

As in Section III-A, we can propose an LMI optimization method for estimating the domain of attraction for any given time-delay for the system (3). If $\mathcal{X}_R$ is a polyhedron, then we have the following optimization problem for estimating the domain of attraction for systems with $\tau \le \tau_0$

$$\begin{aligned} & \inf_{Q>0,P_1>0,P_2>0,G} \gamma, \quad \text{s.t.} \\ & a) \begin{bmatrix} \gamma x_i^T & \\ x_i & Q \end{bmatrix} \ge 0, \quad i = 1, 2, \ldots, l \\ & b) \text{ LMI (27)–(29)}, \quad i \in [1, 2^m], \\ & c) \text{ Constrain (13c)}. \end{aligned} \tag{30}$$

As usual, the analysis problem can be easily modified for controller design by taking F as an optimizing parameter.

V. Numerical Examples

Example 1: Consider the example given in [22]. The system is described by (1) with

$$A = \begin{bmatrix} 1 & 1.5 \\ 0.3 & -2 \end{bmatrix}, \quad A_d = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}$$
$$B = \begin{bmatrix} 10 \\ 1 \end{bmatrix}, \quad \tau = 1, \quad u_{\max} = 15.$$

In [22], a feedback matrix

$$F = [-0.3592 \quad -0.1421]$$

is obtained with local stability in the ball $\mathcal{B}(\delta) = \{x \in \mathbb{R}^n : \|x\|^2 \le \delta\}$ with $\delta = 1.7919 \times 10^3$. As in [22], we require that the origin of the saturated system be asymptotically stable and that the unsaturated system is β-stable with $\beta = 1$. By Theorem 3 and solving optimization problem (14) with the above control law and a unit ball as the reference set, we obtain

$$\alpha = 47.0626, \quad P = \begin{bmatrix} 0.1324 & 0.0283 \\ 0.0283 & 0.4489 \end{bmatrix} \times 10^{-3}.$$

This means that the asymptotic stability of the saturated system and β-stability ($\beta = 1$) of the unsaturated system are guaranteed in the ellipsoid $\Omega(P, 1)$ which include the ball $\mathcal{B}(\delta)$ with $\delta = \alpha^2 = 2.2149 \times 10^3$. Obviously, this estimation is less conservative than the result of [22].

If we only require that the saturated system be asymptotically stable, *i.e.*, $\beta = 0$, by Theorem 3, we obtain

$$\alpha = 67.0618 \quad P = \begin{bmatrix} 0.2223 & 0.0000 \\ 0.0000 & 0.2223 \end{bmatrix} \times 10^{-3}.$$

This means that the asymptotic stability of the saturated time-delay system is guaranteed in the ellipsoid $\Omega(P, 1)$ which includes the ball $\mathcal{B}(\delta)$ with $\delta = \alpha^2 = 4.4973 \times 10^3$. This is an estimate of the domain of attraction of the saturated time-delay system. Note that this estimate of the domain of attraction is delay-independent, *i.e.*, it holds for any size of time-delay. This ellipsoid $\Omega(P, 1)$ is shown in Fig. 1. The dot-

$$Q\bar{A}_i^T + \bar{A}_i Q + B(D_i FQ + D_i^- G) + (D_i FQ + D_i^- G)^T B^T + \tau_0 A_d(P_1+P_2)A_d^T + 2\tau_0 Q < 0 \tag{27}$$

$$\begin{bmatrix} Q & [AQ + B(D_i FQ + D_i^- G)]^T \\ AQ + B(D_i FQ + D_i^- G) & P_1 \end{bmatrix} \ge 0 \tag{28}$$

$$\begin{bmatrix} Q & QA_d^T \\ A_d Q & P_2 \end{bmatrix} \ge 0 \tag{29}$$

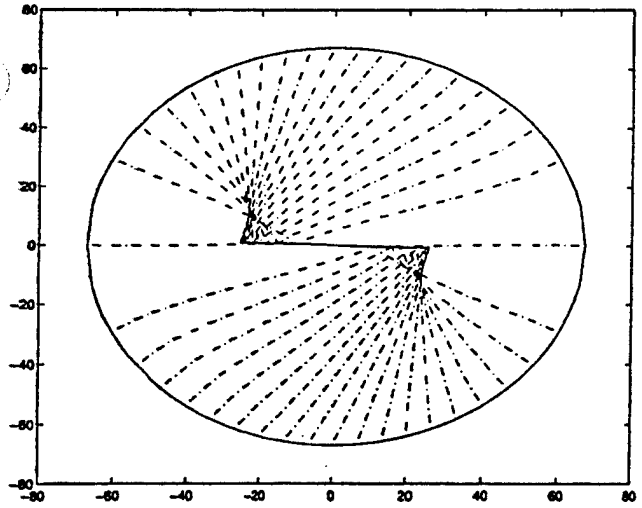


Fig. 1. Estimation of domain of attraction of Example 1.

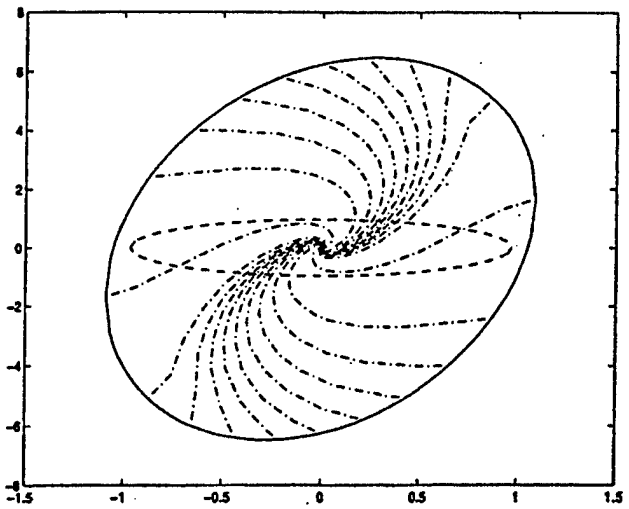


Fig. 2. Estimate of domain of attraction of Example 2.

TABLE I
COMPUTATION RESULTS OF EXAMPLE 1 BY LYAPUNOV-KRASOVSKII APPROACH

β	τ	α	ε	P	
1	0.01	94.7215	2.3225	0.8703 0.2046 0.2046 0.8980	$\times 10^{-4}$
1	0.1	86.5436	2.1114	0.8895 0.1949 0.1949 0.9240	$\times 10^{-4}$
1	1	54.8073	1.7449	0.0835 0.0166 0.0166 0.1139	$\times 10^{-3}$
1	10	no solution			
0	0.01	97.6418	4.1145	0.7231 0.3103 0.3103 0.6684	$\times 10^{-4}$
0	0.1	86.5569	2.3225	0.8567 0.2143 0.2143 0.8801	$\times 10^{-4}$
0	1	59.8417	0.9850	0.1305 0.0060 0.0060 0.1371	$\times 10^{-3}$
0	10	27.8434	0.4177	0.1823 0.0155 0.0155 0.2456	$\times 10^{-3}$

TABLE II
COMPUTATION RESULTS OF EXAMPLE 2

τ_0	α	P	F
0.1	1.5685	0.2418 −0.0097 −0.0097 0.0530	[−1.6523 0.8092]
0.2	1.2597	0.3631 −0.0260 −0.0260 0.0431	[−2.0026 0.8110]
0.3	1.2557	0.5067 −0.0373 −0.0373 0.0311	[−2.3284 0.7827]
0.35	0.9680	0.9033 −0.0393 −0.0393 0.0256	[−2.6383 0.7204]

dashed curves are the state trajectories with the initial conditions on this ellipsoid and $\tau = 10$. Obviously, all trajectories converge to the origin.

If we use the LMI optimization (19) by Lyapunov-Krasovskii approach with the above control law, the computational results are shown Table I. From this table, we find that our result is less conservative than that of [22] because our estimate of domain of attraction when $\tau = 1$ and $\beta = 1$ includes the ball $\mathcal{B}(\delta)$ with $\delta = \alpha^2 = 3.004 \times 10^3$ which is much bigger than the ball given in [22]. We can also find that the estimation of the domain of attraction by Lyapunov-Krasovskii approach becomes smaller as the size of time-delay becomes larger.

Example 2: Consider the following delay system (1) with

$$A = \begin{bmatrix} 0.5 & -1 \\ 0.5 & -0.5 \end{bmatrix}$$
$$A_d = \begin{bmatrix} 0.6 & 0.4 \\ 0 & -0.5 \end{bmatrix}$$
$$B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad u_{\max} = 5.$$

For simplicity, we also use the unit ball as our reference set. We are not able to obtain a feasible solution to LMI optimization problem (13). This means that this system may not be delay-independently stabilizable by a saturated memoryless state feedback law. Fortunately, the optimization problem (30) is feasible for $\tau_0 \leq 0.35$. This means that this saturated system is delay-dependently stabilizable with a memoryless state feedback. Table II shows the computational results with different time-delay. From Table II, we find that α increases when the system time-delay τ_0 decreases. Fig. 2 illustrates the estimate of the domain of attraction and the state trajectories for $\tau = 0.35$. The outer ellipsoid is $\Omega(P, 1)$ and the inner ellipsoid is the ball $\mathcal{B}(\alpha)$. The dot-dashed curves are the state trajectories with initial conditions inside this ellipsoid $\Omega(P, 1)$.

VI. CONCLUSIONS

In this paper, the domain of attraction of time-delay system subject to input saturation is addressed by applying Lyapunov-Razumihkin and -Krasovskii functional approach. An estimation of the domain of attraction is proposed by using the linear matrix inequality optimization. We also proposed a memoryless state feedback design method for the systems with time-delay in state and subject to input saturation to enlarge the domain of attraction. Both the delay-independent and delay-dependent local stabilizing controllers are discussed. Numerical examples show the effectiveness of the proposed method.

REFERENCES

[1] D. S. Bernstein and A. N. Michel, "A chronological bibliography on saturating actuators," *Int. J. Robust Nonlinear Cont.*, vol. 5, pp. 375–380, 1995.

[2] Y.-Y. Cao and Y.-X. Sun, "Robust stabilization of uncertain systems with time-varying multi-state-delay," *IEEE Trans. Automat. Control*, vol. 43, pp. 1484–1488, Oct. 1998.

[3] Y.-Y. Cao, Y.-X. Sun, and C. Cheng, "Delay-dependent robust stabilization of uncertain systems with multiple state delays," *IEEE Trans. Automat. Control*, vol. 43, pp. 1608–1612, Nov. 1998.

[4] Y.-Y. Cao, Y.-X. Sun, and J. Lam, "Delay-dependent robust H_∞ control for uncertain systems with time-varying delays," *IEE Proc. D: Contr. Theory Appl.*, vol. 145, no. 3, pp. 338–344, 1998.
[5] B.-S. Chen, S.-S. Wang, and H.-C. Lu, "Stabilization of time-delay systems containing saturating actuators," *Int. J. Contr.*, vol. 47, pp. 867–881, 1988.
[6] J.-H. Chou, I.-R. Horng, and B.-S. Chen, "Dynamical feedback compensator for uncertain time-delay systems containing saturating actuator," *Int. J. Contr.*, vol. 49, pp. 961–968, 1989.
[7] K. Gu, "Discretized Lyapunov functional for uncertain systems with multiple time-delay," *Int. J. Contr.*, vol. 72, no. 16, pp. 1436–1445, 1999.
[8] K. Gu and S. I. Niculescu, "Additional dynamics in transformed time-delay systems," *IEEE Trans. Automat. Control*, vol. 45, pp. 572–575, Mar. 2000.
[9] J. Hale, *Theory of Functional Differential Equations.* New York: Springer, 1977.
[10] Q. L. Han and B. Ni, "Delay-dependent robust stabilization for uncertain constrained systems with pointwise and distributed time-varying delays," in *Proc. IEEE 38th Conf. DC*, 1999, pp. 215–220.
[11] T. Hu and Z. Lin, *Control Systems with Actuator Saturation: Analysis and Design.* Boston, MA: Birkhäuser, 2001.
[12] T. Hu, Z. Lin, and B. M. Chen, "An analysis and design method for linear systems subject to actuator saturation and disturbance," in *Proc. ACC'00*, 2000, pp. 725–729.
[13] B. Lehman, J. Bentsman, S. V. Lunel, and E. I. Verriest, "Vibrational control of nonlinear time lag systems with bounded delay: Averaging theory, stabilizability and transient behavior," *IEEE Trans. Automat. Control*, vol. 39, pp. 898–912, May 1994.
[14] Z. Lin, *Low Gain Feedback.* London, U.K.: Springer, 1998.
[15] D. Liu and A. Michel, *Dynamical Systems With Saturation Nonlinearities: Analysis and Design.* London, U.K.: Springer-Verlog, 1994.
[16] S. I. Niculescu, J. M. Dion, and L. Dugard, "Robust stabilization for uncertain time-delay systems containing saturating actuators," *IEEE Trans. Automat. Control*, vol. 41, pp. 742–747, May 1996.
[17] S.-I. Niculescu, E. I. Verriest, L. Dugard, and J.-D. Dion, "Stability and robust stability of time-delay systems: A guided tour," in *Stability and Control of Time-Delay Systems*, L. Dugard and E. I. Verriest, Eds. London, U.K.: Springer-Verlag, 1997, vol. 228, pp. 1–71.
[18] S. Oucheriah, "Global stabilization of a class of linear continuous time-delay systems with saturating controls," *IEEE Trans. Circuits Syst. I*, vol. 43, pp. 1012–1015, 1996.
[19] J.-K. Park, C.-H. Choi, and H. Choo, "Dynamic anti-windup method for a class of time-delay control systems with input saturation," *Int. J. Robust Nonlin. Contr.*, vol. 10, no. 6, pp. 457–488, 2000.
[20] G. Stepan, *Retarded Dynamical Systems: Stability and Characteristic Functions* Harlow, U.K, 1989, Pitman Research Notes in Mathematics, Longman Scientific and Technical.
[21] T.-J. Su and C.-G. Huang, "Robust stability of delay dependence for linear uncertain systems," *IEEE Trans. Automat. Control*, vol. 37, pp. 1656–1659, Oct. 1992.
[22] S. Tarbouriech and J. M. Gomes da Silva Jr., "Synthesis of controllers for continuous-time delay systems with saturating controls via LMI's," *IEEE Trans. Automat. Contr.*, vol. 45, pp. 105–111, Jan. 2000.

Multiple Resonance Networks

Antonio Carlos M. de Queiroz

***Abstract*—This brief shows how "multiple resonance networks" of any order and with many possible structures can be systematically designed using standard lossless impedance synthesis techniques. These networks are composed of linear lumped or distributed capacitors, inductors, and transformers, with a switch separating one of the capacitors from the remaining circuit. They have the property of transferring completely the energy initially stored in the capacitor insulated by the switch, to another, much smaller, capacitor in the circuit, through a linear transient when the switch is closed. These circuits find applications in the production of very high voltages for pulsed power systems.**

***Index Terms*—Linear network synthesis, power converters, resonance.**

I. Introduction

"Multiple resonance networks" [1] is a name that generalizes the "double resonance" [2], [3], "triple resonance" [4]–[6], and the higher order networks discussed in this brief. These circuits are usually composed of a transformer and some extra capacitors and inductors and work by transferring the energy initially stored in a capacitor at one side of the transformer to another, much smaller, capacitor at the other side of the transformer, through a linear transient composed (in the ideal lossless case) of a sum of several cosinusoidal waveforms (Fig. 1).

The "double resonance" case is long known [2], [7] as the "Tesla coil" [3]. In this case, only two capacitors and one transformer are used, resulting in a fourth-order system with a transient formed by two oscillatory modes (Fig. 2). With the system properly designed, after some cycles all the initial energy in C_1 is transferred to C_2, and the obtained voltage is given, by energy conservation, by

$$v_{\text{out}}\max = v_{\text{in}}(0)\sqrt{\frac{C_1}{C_p}} \tag{1}$$

(with $p = 2$). This same equation fixes the maximum output voltage for all the systems of this type.

More recently, triple resonance systems were developed [4]–[6] for instrumentation used in high-energy physics. An additional capacitor and an inductor were added to the output side (Fig. 3), with the aim of reducing the voltage stress over the transformer and of taking into consideration the output capacitance of the transformer. With only the extra inductor added, the system is still a double resonance system, long known as the "Tesla magnifier." With the extra capacitor the system is of sixth order and the transient has three oscillatory modes, but operation with complete energy transfer is equally possible.

In all the cases found in the literature, the design of these systems is based on the analysis of a fixed structure. The following sections show that the design can be made by synthesis, can be applied to a wide range of structures, and can be extended to systems of any order.

II. Synthesis Approach

The transformer can be left out of the problem, because it can be inserted after the synthesis of a "ladder" structure composed of series

Manuscript received November 29, 2000; revised July 26, 2001, and September 19, 2001. This paper was recommended by Associate Editor P. K. Rajan.

The author is with the Electrical Engineering Program – COPPE and the Electronic and Computer Engineering Department, Federal University of Rio de Janeiro, Rio de Janeiro 21945-970, Brazil (e–mail: acmq@coe.ufrj.br).

Publisher Item Identifier S 1057-7122(02)01185-6.

1057–7122/02$17.00 © 2002 IEEE

Publication 17

ELSEVIER

Systems & Control Letters 46 (2002) 137–151

SYSTEMS & CONTROL LETTERS

www.elsevier.com/locate/sysconle

Set invariance analysis and gain-scheduling control for LPV systems subject to actuator saturation

Yong-Yan Cao[a,*], Zongli Lin[a], Yacov Shamash[b]

[a] *Department of Electrical and Computer Engineering, P.O. Box 400743, University of Virginia, Charlottesville, VA 22903, USA*
[b] *Department of Electrical and Computer Engineering, State University of New York, Stony Brook, NY 11794, USA*

Received 15 November 2001; received in revised form 21 January 2002

Abstract

In this paper, a set invariance analysis and gain scheduling control design approach is proposed for the polytopic linear parameter-varying systems subject to actuator saturation. A set invariance condition is first established. By utilizing this set invariance condition, the design of a time-invariant state feedback law is formulated and solved as an optimization problem with LMI constraints. A gain-scheduling controller is then designed to further improve the closed-loop performance. Numerical examples are presented to demonstrate the effectiveness of the proposed analysis and design method. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Set invariance; Linear parameter-varying systems; Gain-scheduling; Actuator saturation; Linear matrix inequality

1. Introduction

In recent years there has been significant interest in the study of linear parameter-varying (LPV) systems, which is motivated by the gain scheduling control design methodology [9–11]. LPV systems are systems that depend on unknown but measurable time-varying parameters. The measurement of these parameters provides real-time information on the variations of the plant's characteristics. Hence, it is desirable to design controllers that are scheduled based on this information. LPV control theory has proven to be useful to simplify the interpolation and realization problems associated with the conventional gain-scheduling. The analysis and synthesis of LPV systems have been investigated recently in [1,8,13,14] by the linear matrix inequality approach. The approach involves the design of several linear time-invariant (LTI) controllers for a parameterized family of linear time-invariant system models and the interpolation of these controller gains.

In control system design, actuator saturation is inevitable. It can severely degrade the closed-loop system performance and sometimes even make the otherwise stable closed-loop system unstable by some large perturbation. The analysis and synthesis of control systems with actuator saturation nonlinearities have been receiving increasing attention recently (see, for example, [2,5,7] and the references therein). Very often, actuator saturation is dealt with by either designing low gain control laws that, for a given bound on the initial conditions, avoid the saturation limits, or estimating the region of attraction in the presence of actuator saturation. In this paper, we will analyze the stability of LPV systems with actuator saturation. The recent analysis

* Corresponding author. Fax: +1-804-924-8818.
E-mail addresses: yycao@virginia.edu (Yong-Yan Cao), zl5y@virginia.edu (Zongli Lin).

0167-6911/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-6911(02)00128-7

approach proposed in [5,6] is used to analyze the set invariance and then a gain-scheduled optimal control design is proposed. The resulting closed-loop system not only possesses a large domain of attraction that contains *a priori* given set of initial conditions, but also guarantees a minimal performance index.

The paper is organized as follows. Problem statement and the preliminaries will be given in Section 2. Set invariance of LPV systems subject to actuator saturation will be analyzed in Section 3. A linear time-invariant controller design method based on the LMI optimization will also be introduced in this section. A gain-scheduled controller design method will be proposed in Section 4. In Section 5, numerical examples will be used to illustrate the proposed analysis and design method. The paper will be concluded in Section 6.

2. Problem statement and preliminary

2.1. Problem statement

We consider the polytopic LPV systems, whose system matrices are affine functions of a parameter vector $p(t)$, subject to actuator saturation,

$$\dot{x}(t) = A(p(t))x(t) + B(p(t))\sigma(u(t)), \tag{1}$$

$$z(t) = C(p(t))x(t) + D(p(t))\sigma(u(t)), \tag{2}$$

where

$$A(p(t)) = \sum_{j=1}^{r} p_j(t)A_j, \quad B(p(t)) = \sum_{j=1}^{r} p_j(t)B_j,$$

$$C(p(t)) = \sum_{j=1}^{r} p_j(t)C_j, \quad D(p(t)) = \sum_{j=1}^{r} p_j(t)D_j$$

with $x \in \mathbb{R}^n$ denoting the state vector, $u \in \mathbb{R}^m$ the control input vector, $z \in \mathbb{R}^p$ the control output vector and $p(t) = [p_1(t)\ p_2(t)\ \dots\ p_r(t)]^{\mathrm{T}} \in \mathbb{R}^r$ the time-varying parameter vector. It is assumed that the time-varying parameter vector $p(t)$ belongs to the unit simplex $\mathscr{P}$, where

$$\mathscr{P} := \left\{ \sum_{j=1}^{r} p_j = 1,\ 0 \leqslant p_j \leqslant 1 \right\}. \tag{3}$$

Therefore, when $p_i(t) = 1$ and $p_j(t) = 0$ for $j \in [1, r]$, $j \neq i$, the LPV model (1)–(2) reduces to its ith linear time-invariant "local" model, i.e., $(A(p), B(p), C(p), D(p)) = (A_i, B_i, C_i, D_i)$. That is, the LPV system matrices vary inside a corresponding polytope Ω whose vertices consist of r local system matrices

$$\Omega = \mathrm{co}\{(A_i, B_i, C_i, D_i), \quad i \in [1, r]\}, \tag{4}$$

where co denotes the convex hull.

The function $\sigma : \mathbb{R}^m \to \mathbb{R}^m$ is the standard saturation function of appropriate dimensions defined as follows

$$\sigma(u) = [\sigma(u_1)\ \sigma(u_2)\ \cdots\ \sigma(u_m)]^{\mathrm{T}},$$

where $\sigma(u_i) = \mathrm{sign}(u_i)\min\{1, |u_i|\}$. Here we have slightly abused the notation by using σ to denote both the scalar valued and the vector valued saturation functions. Also note that it is without loss of generality to assume unity saturation level. The non-unity saturation level can be absorbed into the input matrices by

applying the following substitution

$$\hat{B} = BU, \quad \hat{D} = DU, \quad \hat{u} = U^{-1}u,$$

where $U = \mathrm{diag}(u_{\max,i})$, and $u_{\max,i}$ is the saturation amplitude of the ith input.

The aim of this paper is to study the design of a linear state feedback law

$$u(t) = Fx(t) \tag{5}$$

or a time-varying parameter-dependent linear state feedback law

$$u(t) = \sum_{i=1}^{r} p_i(t)F_i x(t), \tag{6}$$

which asymptotically stabilizes the LPV system subject to actuator saturation (1). Control law (5) is a constant feedback law, while (6) is a time-varying feedback law. Control law (6) is the so-called gain-scheduled controller.

In this paper, we will consider the optimal control problem of the LPV plants subject to actuator saturation. That is, we will design a control u, which minimizes the following worst-case performance subject to the LPV model (1)–(2):

$$\min_{u(t)} \max_{(A(p),B(p),C(p),D(p))\in\Omega} \left\{ J = \int_0^\infty z^{\mathrm{T}}(t)z(t) \right\}. \tag{7}$$

2.2. Some mathematical tools

Let f_i be the ith row of the matrix F. We define the symmetric polyhedron

$$\mathscr{L}(F) = \{x \in \mathbb{R}^n : |f_i x| \leqslant 1,\ i = 1,2,\ldots,m\}.$$

If the control u does not saturate for all $i = 1,2,\ldots,m$, that is $x \in \mathscr{L}(F)$, then nonlinear system (1) admits the following linear representation

$$\dot{x}(t) = (A(p(t)) + B(p(t))F)x(t). \tag{8}$$

Let $P \in \mathbb{R}^{n\times n}$ be a positive-definite matrix. For a positive number ρ, denote

$$\Omega(P,\rho) = \{x \in \mathbb{R}^n : x^{\mathrm{T}}Px \leqslant \rho\}.$$

An ellipsoid $\Omega(P,\rho)$ is inside $\mathscr{L}(F)$ if and only if

$$f_i(P/\rho)^{-1}f_i^{\mathrm{T}} \leqslant 1, \quad i = 1,2,\ldots,m.$$

Let $\mathscr{V}$ be the set of $m \times m$ diagonal matrices whose diagonal elements are either 1 or 0. For example, if $m = 2$, then

$$\mathscr{V} = \left\{ \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\}.$$

There are 2^m elements in $\mathscr{V}$. Suppose that each element of $\mathscr{V}$ is labeled as E_i, $i = 1,2,\ldots,2^m$, and denote $E_i^- = I - E_i$. Clearly, E_i^- is also an element of $\mathscr{V}$ if $E_i \in \mathscr{V}$.

Lemma 1 (Hu and Lin [5]). *Let $F,H \in \mathbb{R}^{m\times n}$ be given. For $x \in \mathbb{R}^n$, if $x \in \mathscr{L}(H)$, then*

$$\sigma(Fx) \in \mathrm{co}\{E_iFx + E_i^-Hx : i \in [1,2^m]\},$$

This means that we can rewrite $\sigma(Fx)$ *as*

$$\sigma(Fx) = \sum_{i=1}^{2^m} \eta_i (E_i F + E_i^- H)x,$$

where $0 \leqslant \eta_i \leqslant 1$, $\sum_{i=1}^{2^m} \eta_i = 1$.

Lemma 2. *Suppose that matrices* $M_i \in \mathbb{R}^{m\times n}$, $i = 1, 2, \ldots, r$, *and a positive semi-definite matrix* $P \in \mathbb{R}^{m\times m}$ *are given. If* $\sum_{i=1}^{r} p_i = 1$ *and* $0 \leqslant p_i \leqslant 1$, *then*

$$\left(\sum_{i=1}^{r} p_i M_i\right)^{\mathrm{T}} P \left(\sum_{i=1}^{r} p_i M_i\right) \leqslant \sum_{i=1}^{r} p_i M_i^{\mathrm{T}} P M_i. \tag{9}$$

Proof.

$$\begin{aligned}\left(\sum_{i=1}^{r} p_i M_i\right)^{\mathrm{T}} P \left(\sum_{i=1}^{r} p_i M_i\right) &= \sum_{i=1}^{r} p_i^2 M_i^{\mathrm{T}} P M_i + \sum_{i=1}^{r}\sum_{j<i}^{r} p_i p_j (M_i^{\mathrm{T}} P M_j + M_j^{\mathrm{T}} P M_i) \\ &\leqslant \sum_{i=1}^{r} p_i^2 M_i^{\mathrm{T}} P M_i + \sum_{i=1}^{r}\sum_{j<i}^{r} p_i p_j (M_i^{\mathrm{T}} P M_i + M_j^{\mathrm{T}} P M_j) \\ &= \sum_{i=1}^{r} p_i M_i^{\mathrm{T}} P M_i. \quad \square\end{aligned}$$

3. A set invariance condition

For a given LPV system subject to actuator saturation and a given linear control law $u = Fx$, we first need to establish a set invariance condition. For simplicity, we will denote

$$\hat{A}_{i,j} = A_i + B_i(E_j F + E_j^- H),$$

$$\hat{C}_{i,j} = C_i + D_i(E_j F + E_j^- H).$$

Theorem 3. *For a given system* (1) *and a given state feedback control matrix* F, *the ellipsoid* $\Omega(P,\gamma)$ *is an invariant set of the closed-loop system under linear state feedback control law* (5) *if there exists a matrix* $H \in \mathbb{R}^{m\times n}$ *satisfying the following matrix inequalities*

$$\begin{aligned}&(A_i + B_i(E_j F + E_j^- H))^{\mathrm{T}} P + P(A_i + B_i(E_j F + E_j^- H)) \\ &\quad + (C_i + D_i(E_j F + E_j^- H))^{\mathrm{T}} (C_i + D_i(E_j F + E_j^- H)) < 0, \quad i \in [1, r],\ j \in [1, 2^m]\end{aligned} \tag{10}$$

and $\Omega(P,\gamma) \subset \mathscr{L}(H)$. *Moreover, for any initial condition* $x_0 \in \Omega(P,\gamma)$, *the performance objective function* (7) *satisfies*

$$J \leqslant x_0^{\mathrm{T}} P x_0 \leqslant \gamma.$$

Proof. Choose a Lyapunov function

$$V(x) = x^{\mathrm{T}}Px.$$

Then,

$$\dot{V} = [A(p)x + B(p)\sigma(Fx)]^{\mathrm{T}}Px + x^{\mathrm{T}}P[A(p)x + B(p)\sigma(Fx)].$$

By Lemma 1, we have

$$\dot{V} = x^{\mathrm{T}}\left[\sum_{i=1}^{r} p_iA_i + \sum_{i=1}^{r} p_iB_i\sum_{j=1}^{2^m}\eta_j(E_jF + E_j^-H)\right]^{\mathrm{T}} Px$$

$$+ x^{\mathrm{T}}P\left[\sum_{i=1}^{r} p_iA_i + \sum_{i=1}^{r} p_iB_i\sum_{j=1}^{2^m}\eta_j(E_jF + E_j^-H)\right] x$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{2^m} p_i\eta_j x^{\mathrm{T}}[(A_i + B_i(E_jF + E_j^-H))^{\mathrm{T}}P + P(A_i + B_i(E_jF + E_j^-H))]x.$$

On the other hand, (10) implies that the following matrix inequalities hold:

$$(A_i + B_i(E_jF + E_j^-H))^{\mathrm{T}}P + P(A_i + B_i(E_jF + E_j^-H)) < 0 \quad \forall i \in [1,r],\ j \in [1,2^m],$$

which implies

$$\dot{V} < 0 \quad \forall x \in \Omega(P,\gamma)\setminus\{0\}.$$

Thus, if $x_0^{\mathrm{T}}Px_0 \leqslant \gamma$, then $x^{\mathrm{T}}(t)Px(t) \leqslant \gamma$ for $t \geqslant 0$, i.e., $\Omega(P,\gamma)$ is a positively invariant set. This also implies that system (1) is asymptotically stable at the origin with $\Omega(P,\gamma)$ contained in the domain of attraction.

To complete the proof, we note that

$$J = \int_0^\infty (z^{\mathrm{T}}z + \dot{V}(x))\,\mathrm{d}t + x_0^{\mathrm{T}}Px_0 = \int_0^\infty \bar{J}(t)\,\mathrm{d}t + x_0^{\mathrm{T}}Px_0,$$

where

$$\bar{J}(t) = z^{\mathrm{T}}(t)z(t) + \dot{V}(x).$$

By Lemma 1, we can rewrite (2) as

$$z(t) = \sum_{i=1}^{r} p_iC_ix(t) + \sum_{i=1}^{r} p_iD_i\sigma(Fx(t))$$

$$= \sum_{i=1}^{r} p_i\left(C_ix(t) + D_i\sum_{j=1}^{2^m}\eta_j(E_jF + E_j^-H)x(t)\right)$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{2^m} p_i\eta_j\hat{C}_{i,j}x(t).$$

Hence, by Lemma 2,

$$\begin{aligned}\dot{\bar{J}}(t) &= z^{\mathrm{T}}(t)z(t) + \dot{V}(x)\\ &= \sum_{i=1}^{r}\sum_{j=1}^{2^m} p_i\eta_j x^{\mathrm{T}}(\hat{A}_{i,j}^{\mathrm{T}}P + P\hat{A}_{i,j})x + \Big(\sum_{i=1}^{r}\sum_{j=1}^{2^m} p_i\eta_j\hat{C}_{i,j}x\Big)^{\mathrm{T}}\Big(\sum_{i=1}^{r}\sum_{j=1}^{2^m} p_i\eta_j\hat{C}_{i,j}x\Big)\\ &\leqslant \sum_{i=1}^{r}\sum_{j=1}^{2^m} p_i\eta_j x^{\mathrm{T}}(\hat{A}_{i,j}^{\mathrm{T}}P + P\hat{A}_{i,j})x + \sum_{i=1}^{r}\sum_{j=1}^{2^m} p_i\eta_j x^{\mathrm{T}}\hat{C}_{i,j}^{\mathrm{T}}\hat{C}_{i,j}x\\ &= \sum_{i=1}^{r}\sum_{j=1}^{2^m} p_i\eta_j x^{\mathrm{T}}(\hat{A}_{i,j}^{\mathrm{T}}P + P\hat{A}_{i,j} + \hat{C}_{i,j}^{\mathrm{T}}\hat{C}_{i,j})x.\end{aligned}$$

It is easy to see that if matrix inequalities (10) hold, i.e.,

$$\hat{A}_{i,j}^{\mathrm{T}}P + P\hat{A}_{i,j} + \hat{C}_{i,j}^{\mathrm{T}}\hat{C}_{i,j} < 0,$$

then $\dot{\bar{J}}(t) \leqslant 0$, which implies $J \leqslant x_0^{\mathrm{T}}Px_0 \leqslant \gamma$. □

Remark 1. If we do not consider the optimal performance index (7), Theorem 3 is a set invariance condition of LPV system subject to actuator saturation. For the special case of $r = 1$, Theorem 3 recovers the set invariance condition for linear time-invariant systems subject to actuator saturation [5]. Additionally, Theorem 3 also addressed the quadratic performance problem for linear systems subject to actuator saturation.

Based on Theorem 3, we can present the following optimization problem minimizing the upper bound of the performance function (7) for a given initial condition set $\mathscr{X}_0$:

$$\begin{aligned}\min_{P>0,F,H}\quad & \gamma,\\ \text{s.t.}\quad & \text{(a) } \mathscr{X}_0 \subset \Omega(P,\gamma),\\ & \text{(b) inequalities (10)}\quad \forall i\in[1,r],\ j\in[1,2^m],\\ & \text{(c) } |h_i x| \leqslant 1\quad \forall x\in\Omega(P,\gamma),\ i=[1,m],\end{aligned}\tag{11}$$

where h_i denotes the ith row of H.

The feasibility of the above optimization problem (11) ensures the existence of a stabilizing state feedback matrix F such that the given initial condition set $\mathscr{X}_0$ is contained in the domain of attraction of the system (1)–(2), and the performance index $J \leqslant \gamma$. On the other hand, for a given constant control matrix F designed for the systems without considering actuator saturation, (11) can also be used to determine if an initial condition set $\mathscr{X}_0$ is contained in the domain attraction of the origin when the system is subject to actuator saturation. In what follows, we will show that the optimization problem (11) can be solved as an LMI optimization problem.

For simplicity, we assume that the initial condition set $\mathscr{X}_0$ is the combination of some given points,

$$\mathscr{X}_0 := \mathrm{co}\{x_0^1, x_0^2, \ldots, x_0^l\},$$

where $x_0^i \in \mathbb{R}^n$, $i = 1,2,\ldots,l$, are some given points. Let

$$Q = (P/\gamma)^{-1},\quad Y = FQ,\quad Z = HQ.$$

Then, Condition (a) is equivalent to

$$(x_0^i)^{\mathrm{T}} P x_0^i \leqslant \gamma \quad \Leftrightarrow \quad \begin{bmatrix} 1 & (x_0^i)^{\mathrm{T}} \\ x_0^i & Q \end{bmatrix} \geqslant 0, \quad i = 1,2,\ldots,l.$$

Condition (b) is equivalent to

$$\begin{bmatrix} (A_iQ + B_i(E_jY + E_j^- Z))^{\mathrm{T}} + (A_iQ + B_i(E_jY + E_j^- Z)) & * \\ C_iQ + D_i(E_jY + E_j^- Z) & -\gamma I \end{bmatrix} < 0 \tag{12}$$

for $\forall i \in [1,r]$, $\forall j \in [1,2^m]$. Condition (c) is equivalent to

$$h_i \left(\frac{P}{\gamma}\right)^{-1} h_i^{\mathrm{T}} \leqslant 1 \quad \Leftrightarrow \quad \begin{bmatrix} 1 & h_iQ \\ Qh_i^{\mathrm{T}} & Q \end{bmatrix} \geqslant 0 \quad \forall i \in [1,m].$$

Also let the ith row of Z be z_i, i.e., $z_i = h_iQ$. The optimization problem (11) can then be reduced to the following one with LMI constraints,

$$\begin{aligned} \min_{Q>0,Y,Z} \quad & \gamma, \text{s.t.} \quad \text{(a)} \begin{bmatrix} 1 & (x_0^i)^{\mathrm{T}} \\ x_0^i & Q \end{bmatrix} \geqslant 0 \quad \forall i \in [1,l], \\ & \text{(b) LMI (12)} \quad \forall i \in [1,r], \ j \in [1,2^m], \\ & \text{(c)} \begin{bmatrix} 1 & z_i \\ z_i^{\mathrm{T}} & Q \end{bmatrix} \geqslant 0 \quad \forall i \in [1,m]. \end{aligned} \tag{13}$$

Theorem 4. *For a given system* (1), *the state feedback control matrix* F *that minimizes the upper bound of performance function* (7) *can be solved by*

$$F = YQ^{-1},$$

where $(Q > 0, Y)$ *is the solution of the LMI optimization problem* (13).

In the optimization problem (13), the amplitude of control law (5) is not constrained, i.e., there is no control amplitude constraint on the control law. In [6], the authors proved that this controller design method is less conservative than the approaches based on circle criterion and Popov criterion [4]. On the other hand, to avoid the controller gain being too large, we may constrain it to be bounded by $\mu_0 > 1$, i.e., $|f_i x| \leqslant \mu_0$, which is equivalent to the following LMI

$$\begin{bmatrix} \mu_0^2 & y_i \\ y_i^{\mathrm{T}} & Q \end{bmatrix} \geqslant 0 \quad \forall i \in [1,m],$$

where y_i denotes ith row of Y.

If we require $Y = Z$, then we recover the design algorithm which constrains the optimal control law to be unsaturated [3]. The unsaturated control algorithm can be described as

$$\begin{aligned} \min_{Q>0,Y} \quad & \gamma, \\ \text{s.t.} \quad & \text{(a)} \begin{bmatrix} 1 & (x_0^i)^{\mathrm{T}} \\ x_0^i & Q \end{bmatrix} \geqslant 0 \quad \forall i \in [1,l], \\ & \text{(b)} \begin{bmatrix} (A_iQ + B_iY)^{\mathrm{T}} + (A_iQ + B_iY) & * \\ C_iQ + D_iY & -\gamma I \end{bmatrix} < 0 \quad \forall i \in [1,r], \\ & \text{(c)} \begin{bmatrix} 1 & y_i \\ y_i^{\mathrm{T}} & Q \end{bmatrix} \geqslant 0 \quad \forall i \in [1,m]. \end{aligned} \tag{14}$$

Note that the constraints in (14) imply that $\Omega(Q^{-1},1) \subset \mathscr{L}(F)$ and hence the control $u=Fx$ will never reach saturation limits. In (13), we permit the control to be saturated and hence our algorithm will result in a larger domain of attraction. It is known that low-gain controllers that avoid saturation will often result in low levels of performance, especially for the cases where the disturbance is intermediate or small amplitude.

4. Gain-scheduled control law design

The approach to gain-scheduling involves the design of several LTI controllers for a parameterized family of time-invariant system models and the interpolation of these controller gains. If the time-varying parameter vector $p(t)$ can be measured or estimated on-line, then we can design a gain-scheduled control law

$$u(t):=\tilde{F}(t)x(t)=\left(\sum_{j=1}^{r} p_j(t)F_j\right)x(t), \tag{15}$$

where F_j is the "local" state feedback matrix for the local model (A_j,B_j). It is reasonable to expect that this kind of control laws can result in a larger domain of attraction and better performance. Note that F in (5) is a constant matrix, while $\tilde{F}$ in (15) is a time-varying matrix function of time-varying parameter $p(t)$ although matrices F_j's are constant for all $j=1,2,\ldots,r$.

With control law (15), the closed-loop system (1)–(2) can be rewritten as

$$\dot{x}(t)=\sum_{i=1}^{r} p_iA_ix(t)+\sum_{i=1}^{r} p_iB_i\sigma(\tilde{F}x(t)),$$

$$z(t)=\sum_{i=1}^{r} p_iC_ix(t)+\sum_{i=1}^{r} p_iD_i\sigma(\tilde{F}x(t)).$$

By Lemma 1, we have that for any matrix $\tilde{H}$ of the same dimensions of $\tilde{F}$ such that $x\in\mathscr{L}(\tilde{H})$,

$$\dot{x}(t)=\sum_{i=1}^{r} p_i\left[A_ix(t)+B_i\sum_{s=1}^{2^m}\eta_s(t)(E_s\tilde{F}+E_s^-\tilde{H})x(t)\right]$$

$$=\sum_{i=1}^{r} p_i\sum_{s=1}^{2^m}\eta_s[A_i+B_i(E_s\tilde{F}+E_s^-\tilde{H})]x(t)$$

and

$$z(t)=\sum_{s=1}^{2^m}\eta_s\sum_{i=1}^{r} p_i[C_i+D_i(E_s\tilde{F}+E_s^-\tilde{H})]x(t),$$

where $0\leqslant\eta_s(t)\leqslant 1$, $\sum_{s=1}^{2^m}\eta_s(t)=1$, for all $s=1,2,\ldots,2^m$. If we let

$$\tilde{H}=\sum_{j=1}^{r} p_jH_j,$$

then

$$\dot{x}(t)=\sum_{s=1}^{2^m}\eta_s(t)\sum_{i=1}^{r} p_i\sum_{j=1}^{r} p_j\tilde{A}_{s,i,j}x(t), \tag{16}$$

$$z(t)=\sum_{s=1}^{2^m}\eta_s(t)\sum_{i=1}^{r} p_i\sum_{j=1}^{r} p_j\tilde{C}_{s,i,j}x(t), \tag{17}$$

where

$$\tilde{A}_{s,i,j} := A_i + B_i(E_s F_j + E_s^- H_j),$$

$$\tilde{C}_{s,i,j} := C_i + D_i(E_s F_j + E_s^- H_j), \quad s \in [1, 2^m], \ i, j \in [1, r].$$

Remark 2. It is easy to find that the closed-loop system described by (16)–(17) can be further simplified if the subsystems (A_i, B_i, C_i, D_i) possess common input matrices B and D, namely $B_i = B$, $D_i = D$ for all i. In this case, the closed-loop system (1)–(2) can be simplified as ,

$$\dot{x}(t) = \sum_{s=1}^{2^m} \eta_s(t) \sum_{i=1}^{r} p_i(A_i + B(E_s F_i + E_s^- H_i))x(t),$$

$$z(t) = \sum_{s=1}^{2^m} \eta_s(t) \sum_{i=1}^{r} p_i(C_i + D(E_s F_i + E_s^- H_i))x(t).$$

Theorem 5. *Suppose that system* (1)–(2) *and the local state feedback control matrices* F_j, $j = 1, 2, \ldots, r$, *are given. The ellipsoid* $\Omega(P, \gamma)$ *is an invariant set of the closed-loop system under the gain-scheduled state feedback law* (15) *if there exist matrices* $H_j \in \mathbb{R}^{m \times n}$, $j = 1, 2, \ldots, r$, *satisfying*

$$(A_i + B_i(E_s F_j + E_s^- H_j))^{\mathrm{T}} P + P(A_i + B_i(E_s F_j + E_s^- H_j)) + (C_i + D_i(E_s F_j + E_s^- H_j))^{\mathrm{T}}(C_i + D_i(E_s F_j + E_s^- H_j)) < 0, \quad i, j \in [1, r], \ s \in [1, 2^m], \tag{18}$$

and $\Omega(P, \gamma) \subset \bigcap_{i=1}^{r} \mathcal{L}(H_i)$. *Moreover, for any* $x_0 \in \Omega(P, \gamma)$, *the performance objective function* (7) *satisfies* $J \leqslant x_0^{\mathrm{T}} P x_0 \leqslant \gamma$.

Proof. Choose a Lyapunov function $V(x(t)) = x^{\mathrm{T}}(t)Px(t)$. We note that

$$x \in \bigcap_{j=1}^{r} \mathcal{L}(H_j),$$

implies

$$x \in \mathcal{L}\left(\sum_{j=1}^{r} p_j H_j \right),$$

since $\sum_{j=1}^{r} p_j = 1$ and $0 \leqslant p_j \leqslant 1$. Let

$$\tilde{H} = \sum_{j=1}^{r} p_j H_j.$$

Then, by Lemma 1,

$$\dot{V} = x^{\mathrm{T}}(t) \left\{ \left[\sum_{i=1}^{r} p_i \sum_{s=1}^{2^m} \eta_s (A_i + B_i(E_s \tilde{F} + E_s^- \tilde{H})) \right]^{\mathrm{T}} P + P \left[\sum_{i=1}^{r} p_i \sum_{s=1}^{2^m} \eta_s (A_i + B_i(E_s \tilde{F} + E_s^- \tilde{H})) \right] \right\} x(t)$$

$$= x^{\mathrm{T}}(t) \left\{ \left[\sum_{s=1}^{2^m} \eta_s \sum_{i=1}^{r} \sum_{j=1}^{r} p_i p_j \tilde{A}_{s,i,j} \right]^{\mathrm{T}} P + P \left[\sum_{s=1}^{2^m} \eta_s \sum_{i=1}^{r} \sum_{j=1}^{r} p_i p_j \tilde{A}_{s,i,j} \right] \right\} x(t). \tag{19}$$

It is easy to see that (18) implies

$$\tilde{A}_{s,i,j}^{\mathrm{T}}P + P\tilde{A}_{s,i,j} < 0, \quad i,j \in [1,r],\ s \in [1,2^m].$$

Hence, we have $\dot{V}(x) < 0$ for all $x \in \Omega(P,\gamma)\setminus\{0\}$. That is, the system is asymptotically stable at the origin with $\Omega(P,\gamma)$ contained in the domain of attraction.

Similar to the proof of Theorem 3, we can prove that

$$\tilde{J}(t) = z^{\mathrm{T}}(t)z(t) + \dot{V}(x(t)) \leqslant 0$$

if matrix inequalities (18) holds and hence $J \leqslant x_0^{\mathrm{T}}Px_0 \leqslant \gamma$. □

Corollary 6. *For the special case of $B_i = B$ and $D_i = D$ for all i, the ellipsoid $\Omega(P,\gamma)$ is an invariant set of the closed-loop system under the gain-scheduled state feedback control law* (15), *if there exist r matrices $H_i \in \mathbb{R}^{m\times n}$, satisfying*

$$(Ai + B(E_sF_i + E_s^- H_i))^{\mathrm{T}}P + P(A_i + B(E_sF_i + E_s^- H_i)) + (C_i + D(E_sF_i + E_s^- H_i))^{\mathrm{T}}(C_i + D(E_sF_i + E_s^- H_i)) < 0, \quad i \in [1,r],\ s \in [1,2^m], \tag{20}$$

and $\Omega(P,\gamma) \subset \bigcap_{i=1}^{r} \mathscr{L}(H_i)$. Moreover, for any $x_0 \in \Omega(P,\gamma)$, the performance objective function (7) *satisfies $J \leqslant x_0^{\mathrm{T}}Px_0 \leqslant \gamma$.*

In what follows, we present a less conservative set invariance condition.

Note that system (16)–(17) can be rewritten as

$$\dot{x}(t) = \sum_{s=1}^{2^m} \eta_s \left\{ \sum_{i=1}^{r} p_i^2 \tilde{A}_{s,i,i} + \sum_{i=1}^{r}\sum_{j<i}^{r} 2p_ip_j \left(\frac{\tilde{A}_{s,i,j} + \tilde{A}_{s,j,i}}{2} \right) \right\} x(t),$$

$$z(t) = \sum_{s=1}^{2^m} \eta_s \left\{ \sum_{i=1}^{r} p_i^2 \tilde{C}_{s,i,i} + \sum_{i=1}^{r}\sum_{j<i}^{r} 2p_ip_j \left(\frac{\tilde{C}_{s,i,j} + \tilde{C}_{s,j,i}}{2} \right) \right\} x(t).$$

Let

$$\bar{p}_l := \begin{cases} p_i^2 & l = i^2 \\ 2p_ip_j & l = ij \end{cases} \quad \text{for } i < j = 1,2,\ldots,r,$$

$$\bar{A}_{s,l} := \begin{cases} \tilde{A}_{s,i,i} & l = i^2 \\ (\tilde{A}_{s,i,j} + \tilde{A}_{s,j,i})/2 & l = ij \end{cases} \quad \text{for } i < j = 1,2,\ldots,r,$$

$$\bar{C}_{s,l} := \begin{cases} \tilde{C}_{s,i,i} & l = i^2 \\ (\tilde{C}_{s,i,j} + \tilde{C}_{s,j,i})/2 & l = ij \end{cases} \quad \text{for } i < j = 1,2,\ldots,r.$$

We then have

$$0 \leqslant \bar{p}_l \leqslant 1, \quad \sum_{l=1}^{r(r+1)/2} \bar{p}_l = 1$$

and

$$\dot{x} = \sum_{s=1}^{2^m} \eta_s \sum_{l=1}^{r(r+1)/2} \bar{p}_l \bar{A}_{s,l} x,$$

$$z = \sum_{s=1}^{2^m} \eta_s \sum_{l=1}^{r(r+1)/2} \bar{p}_l \bar{C}_{s,l} x.$$

Hence,

$$\dot{V}(x)=x^{\mathrm{T}}\left[\sum_{s=1}^{2^m}\eta_s\sum_{l=1}^{r(r+1)/2}\bar{p}_l(\bar{A}_{s,l}^{\mathrm{T}}P+P\bar{A}_{s,l})\right]x,$$

$$\bar{J}(t)=x^{\mathrm{T}}\left[\sum_{s=1}^{2^m}\eta_s\sum_{l=1}^{r(r+1)/2}\bar{p}_l(\bar{A}_{s,l}^{\mathrm{T}}P+P\bar{A}_{s,l})\right]x+x^{\mathrm{T}}\left(\sum_{s=1}^{2^m}\eta_s\sum_{l=1}^{r(r+1)/2}\bar{p}_l\bar{C}_{s,l}\right)^{\mathrm{T}}\left(\sum_{s=1}^{2^m}\eta_s\sum_{l=1}^{r(r+1)/2}\bar{p}_l\bar{C}_{s,l}\right)x.$$

Theorem 7. *Suppose that system* (1)–(2) *and the local state feedback control matrices* F_j, $j=1,2,\ldots,r$, *are given. The ellipsoid* $\Omega(P,\gamma)$ *is an invariant set of the closed-loop system under the gain-scheduled control law* (15), *if there exist matrices* $H_j\in\mathbb{R}^{m\times n}$, $j=1,2,\ldots,r$, *satisfying*

$$\tilde{A}_{s,i,i}^{\mathrm{T}}P+P\tilde{A}_{s,i,i}+\tilde{C}_{s,i,i}^{\mathrm{T}}\tilde{C}_{s,i,i}<0,\quad i\in[1,r],\ s\in[1,2^m],\tag{21}$$

$$\begin{aligned}&(\tilde{A}_{s,i,j}+\tilde{A}_{s,j,i})^{\mathrm{T}}P+P(\tilde{A}_{s,i,j}+\tilde{A}_{s,j,i})\\&\quad+\tfrac{1}{2}(\tilde{C}_{s,i,j}+\tilde{C}_{s,j,i})^{\mathrm{T}}(\tilde{C}_{s,i,j}+\tilde{C}_{s,j,i})<0,\quad i<j\in[1,r],\ s\in[1,2^m]\end{aligned}\tag{22}$$

and $\Omega(P,\gamma)\subset\bigcap_{i=1}^{r}\mathcal{L}(H_i)$. *Moreover, for any* $x_0\in\Omega(P,\gamma)$, *the performance objective function* (7) *satisfies*

$$J\leqslant x_0^{\mathrm{T}}Px_0\leqslant\gamma.$$

Remark 3. In comparison with Theorem 5, the number of matrix inequalities in Theorem 7 is reduced by $r(r-1)\times 2^{m-1}$. In comparison with Corollary 6, another $r(r-1)\times 2^{m-1}$ matrix inequalities can be removed for the special case of $B_i=B$ and $D_i=D$, $\forall i$.

Let

$$Q=(P/\gamma)^{-1},\quad Y_j=F_jQ,\quad Z_j=H_jQ,\ j\in[1,r].$$

Denote the ith row of the matrix Z_j as z_i^j. Then (21) and (22) are equivalent to the following LMIs

$$\begin{bmatrix}(A_iQ+B_i(E_sY_i+E_s^-Z_i))^{\mathrm{T}}+(A_iQ+B_i(E_sY_i+E_s^-Z_i)) & *\\ C_iQ+D_i(E_sY_i+E_s^-Z_i) & -\gamma I\end{bmatrix}<0,\quad i\in[1,r],\ s\in[1,2^m]\tag{23}$$

and

$$\begin{bmatrix}A_iQ+B_i(E_sY_j+E_s^-Z_j)+A_jQ+B_j(E_sY_i+E_s^-Z_i)\\ \quad+(A_iQ+B_i(E_sY_j+E_s^-Z_j)+A_jQ+B_j(E_sY_i+E_s^-Z_i))^{\mathrm{T}} & *\\ C_iQ+D_i(E_sY_j+E_s^-Z_j)+C_jQ+D_j(E_sY_i+E_s^-Z_i) & -2\gamma I\end{bmatrix}<0,\quad i<j\in[1,r],\ s\in[1,2^m],\tag{24}$$

respectively. Then, we have the following theorem.

Theorem 8. *Suppose that system* (1)–(2) *and local state feedback control matrices* F_j, $j=1,2,\ldots,r$, *are given. Then gain-scheduled state feedback control law* (15) *that minimizes the upper bound of performance function* (7) *can be solved by*

$$F_j=Y_jQ^{-1}\quad\forall j\in[1,r],$$

where $(Q>0, Y_j)$ *is a solution of the following LMI optimization problem*:

$$\begin{aligned}
&\min_{Q>0,Y_j,Z_j} \quad \gamma,\\
&\text{s.t.} \quad \text{(a)} \begin{bmatrix} 1 & (x_0^i)^{\mathrm{T}} \\ x_0^i & Q \end{bmatrix} \geqslant 0 \quad \forall i \in [1,l],\\
&\qquad \text{(b) LMI (23), (24),}\\
&\qquad \text{(c)} \begin{bmatrix} 1 & z_i^j \\ (z_i^j)^{\mathrm{T}} & Q \end{bmatrix} \geqslant 0 \quad \forall i \in [1,m],\ j \in [1,r].
\end{aligned} \tag{25}$$

5. Numerical examples

Example 1. First, we consider a simple LPV system with the following system matrices

$$A_1 = \begin{bmatrix} 0 & 1 \\ 0.1 & -0.1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 1 \\ 0.1 & 1 \end{bmatrix}, \quad B_1 = B_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

$$C_1 = C_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad D_1 = D_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The system output is $y = x_1$. The input is subject to saturation $u_{\max} = 1$. We are interested in designing an optimal controller such that the initial condition $x_0 = [4 \ \ -1]^{\mathrm{T}}$ is contained in the domain of attraction of the origin. The unsaturated control algorithm (14), the constant control algorithm (13) and the gain-scheduling control algorithm (25) are used to design the control law respectively. Table 1 shows the computation results by the different control algorithms.

Fig. 1 shows the outputs and inputs of the system by the different controllers with

$$p_1 = 0.5 - 0.5\sin(0.1\pi t - 0.5\pi),$$

$$p_2 = 0.5 + 0.5\sin(0.1\pi t - 0.5\pi).$$

The dotted curves correspond to the controller computed by algorithm (14), the dotted dash curves correspond to algorithm (13) and the solid curves correspond to algorithm (25). It is obvious that the gain-scheduled controller has the shortest rising time and the smallest performance cost while the unsaturated controller has the longest rising time with the largest cost.

If the initial condition $x_0 = 1.3 \times [4 \ \ -1]^{\mathrm{T}}$, we find the unsaturated control algorithm (14) and the constant control algorithm (13) cannot obtain a feasible solution while the gain-scheduling control algorithm (25) can still work well. This implies that the gain-scheduled controller can result in a larger domain of attraction.

Table 1
Computation results by different algorithms

	Algorithm (14)	Algorithm (13)	Algorithm (25)
F	$[-0.2703 \ \ -1.6093]$	$[-2.4772 \ \ -12.1166]$	$F_1 = [-4.0303 \ \ -15.4821]$ $F_2 = [-3.8787 \ \ -15.7463]$
γ	120.4165	120.3905	32.6565

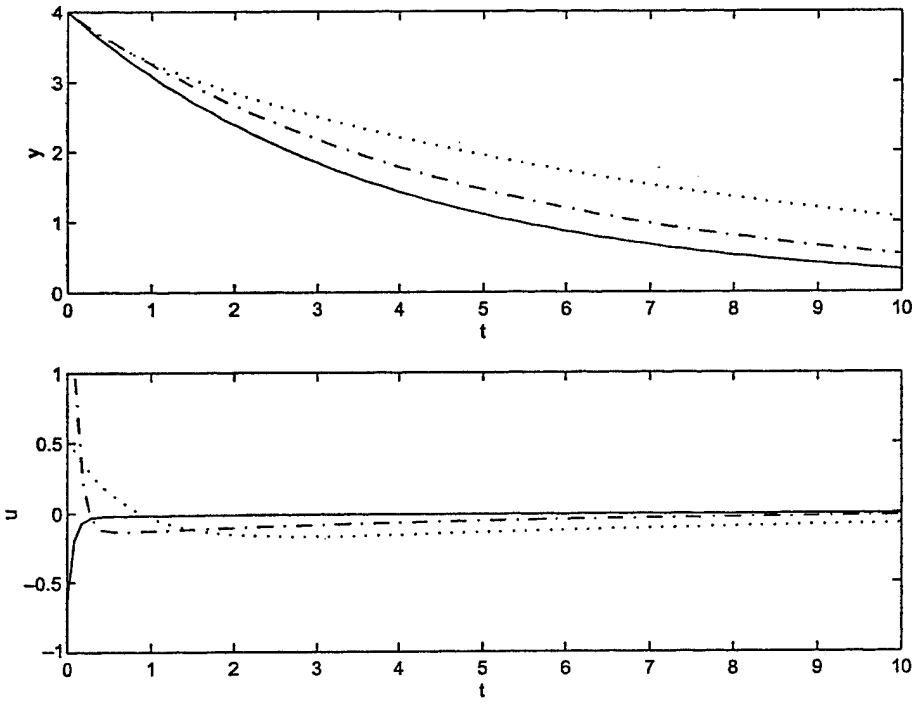


Fig. 1. The outputs and inputs of Example 1 with different controllers solved by: algorithm (14) (dotted curves), algorithm (13) (dotted dash curves) and algorithm (25) (solid curves).

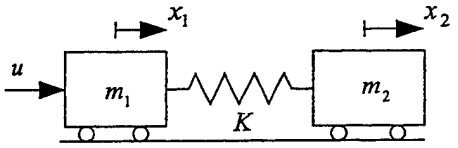


Fig. 2. Coupled spring-mass systems.

Example 2. The second example is about the control of a two-mass-spring system shown in Fig. 2. The system is given by the following state space equations [12]

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\dfrac{K}{m_1} & \dfrac{K}{m_1} & 0 & 0 \\ \dfrac{K}{m_2} & -\dfrac{K}{m_2} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \dfrac{1}{m_1} \\ 0 \end{bmatrix} u,$$

$$y = x_2.$$

Here, x_1 and x_2 are the positions of the two carts respectively, and x_3 and x_4 are their respective velocities. m_1 and m_2 are the masses of the two bodies and K is the spring constant. For the nominal system $m_1 = m_2 = 1$ with appropriate units. The spring constant is assumed to be uncertain in the range $K_{\min} = 0.5 \leqslant K \leqslant K_{\max} = 2$. It is assumed that exact measurement of the state is available. In the simulation, we set the performance

Table 2
Performance bounds computed by different algorithms

	Algorithm (14)	Algorithm (13)	Algorithm (25)
J	6.0788	6.0780	5.9481

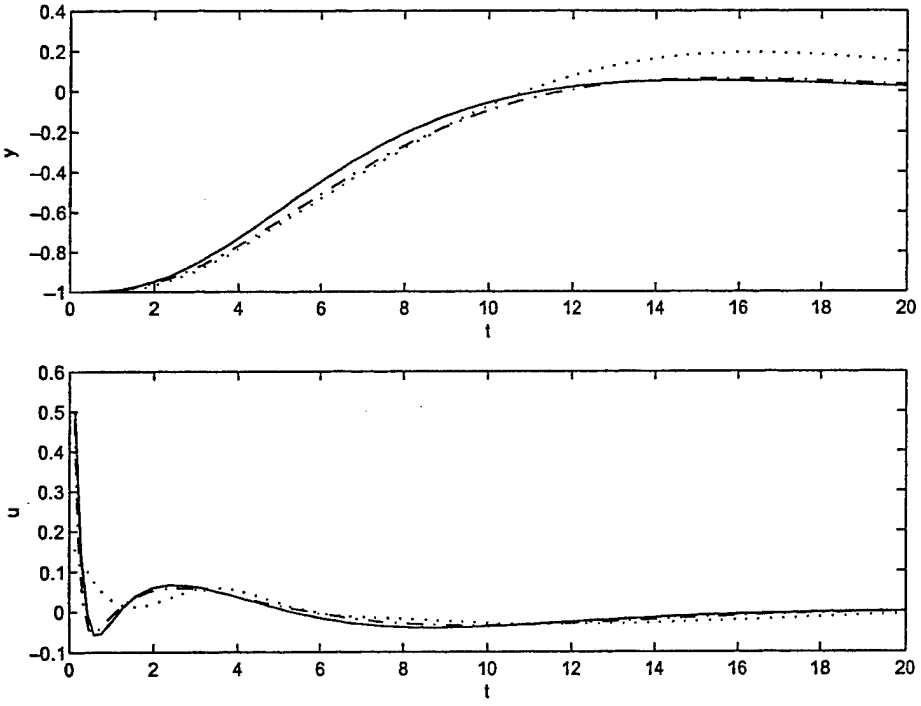


Fig. 3. The outputs and inputs of the two-mass-spring system with different controllers solved by: algorithm (14) (dotted curves), algorithm (13) (dotted dash curves) and algorithm (25) (solid curves).

output as

$$z = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \sigma(u)$$

and we assume that the saturation limit is $u_{\max} = 0.5$ and that the time-varying parameter is given as

$$K = K_{\min}(0.5 + 0.5\sin(0.01\pi t)) + K_{\max}(0.5 - 0.5\sin(0.01\pi t)).$$

We compare the simulation results by the different control algorithms: unsaturated control algorithm (14), the constant control algorithm (13) and the gain-scheduling control algorithm (25). Table 2 shows the computed performance bounds γ by using the different algorithms. Fig. 3 shows the computed outputs and inputs, where the dotted curves are obtained by using the unsaturated control algorithm (14), dotted dash curves by using the constant control algorithm (13) and the solid curves by the gain-scheduling control algorithm (25). We see that the controller computed by the unsaturated control algorithm (14) leads to the largest overshoot and the longest rising time. This is because the input is not able to reach the saturation limit due to the conservatism of algorithm (14). On the other hand, it is observed that the gain-scheduling control algorithm can result in a faster response than the constant control algorithm (13). From the simulation results, we can conclude that

the gain-scheduling control algorithm (25) improve the performance with both output and input reaching the set-point in shorter time.

6. Conclusions

In this paper, we have addressed the set invariance and the gain-scheduling control for LPV systems subject to actuator saturation. The positively invariant set of LPV systems subject to actuator saturation is analyzed by vertex system analysis approach. The optimal control problem for the given initial condition, i.e., steering it to the origin with the minimal performance cost, is solved with the LMI optimization approach. A gain-scheduled controller design method is proposed to reduce the conservativeness. The numerical examples also demonstrate the effectiveness of our design.

Acknowledgements

Work supported in part by the US office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

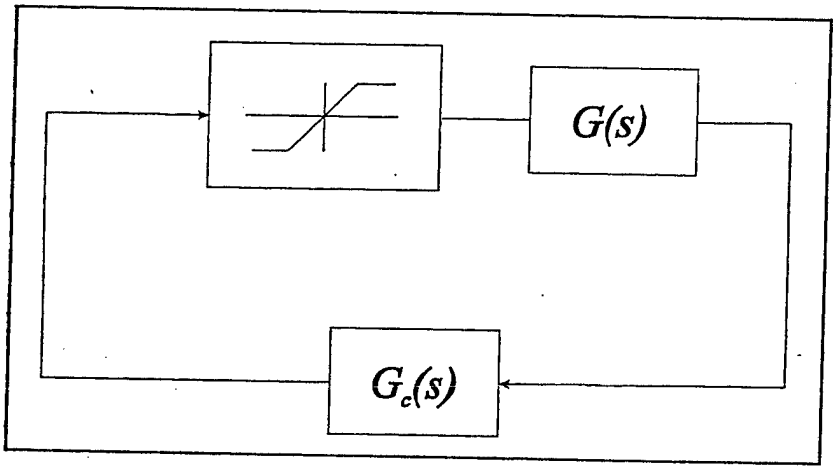
References

[1] P. Apkarian, P. Gahinet, G. Becker, Self-scheduled H_∞ control of linear parameter-varying systems, Automatica 31 (1995) 1251–1261.
[2] D.S. Bernstein, A.N. Michel, A chronological bibliography on saturating actuators, Int. J. Robust Nonlinear Control 5 (1995) 375–380.
[3] S. Boyd, L. El Ghaoui, E. Feron, V. Balakrishnan, Linear Matrix Inequalities in System and Control Theory, SIAM, Philadelphia, 1994.
[4] H. Hindi, S. Boyd, Analysis of linear systems with saturating using convex optimization, Proceedings of the 37th IEEE Conference on Decision Control, 1998, pp. 903–908.
[5] T. Hu, Z. Lin, Control Systems with Actuator Saturation: Analysis and Design, Birkhäuser, Boston, 2001.
[6] T. Hu, Z. Lin, B.M. Chen, An analysis and design method for linear systems subject to actuator saturation and disturbance, Automatica 38 (2) (2002) 351–359.
[7] Z. Lin, Low Gain Feedback, Springer, London, 1998.
[8] A.K. Packard, Gain scheduling via linear fractional transformations, Systems Control Lett. 22 (1994) 79–92.
[9] W.J. Rugh, Analytical framework for gain scheduling, IEEE Control Systems Mag. 11 (1991) 74–84.
[10] J.S. Shamma, M. Athans, Analysis of nonlinear gain-scheduled control systems, IEEE Trans. Automat. Control 35 (1990) 898–907.
[11] J.S. Shamma, M. Athans, Gain scheduling: potential hazards and possible remedies, IEEE Control Systems Mag. 12 (1992) 101–107.
[12] B. Wie, D.S. Bernstein, Benchmark problems for robust control design, J. Guidance, Control, Dynamics 15 (1992) 1057–1059.
[13] F. Wu, A. Packard, LQG control design for LPV systems, Proceedings of the American Control Conference, 1995, pp. 4440–4444.
[14] F. Wu, X.H. Yang, A.K. Packard, G. Becker, Induced L_2 norm control for LPV systems with bounded parameter variation rates, Int. J. Robust Nonlinear Control 6 (1996) 983–998.

Publication 18

Control Engineering Series

ACTUATOR SATURATION CONTROL



edited by
Vikram Kapila
Karolos M. Grigoriadis

Chapter 3

Null Controllability and Stabilization of Linear Systems Subject to Asymmetric Actuator Saturation[1]

T. Hu, A. N. Pitsillides, and Z. Lin

University of Virginia, Charlottesville, Virginia

3.1. Introduction

We consider the problem of controlling exponentially unstable linear systems subject to asymmetric actuator saturation. This control problem involves basic issues such as characterization of the null controllable region by bounded controls and stabilizability on the null controllable region. These issues have been focuses of study of and are now well-addressed for linear systems that are not exponentially unstable. For example, it is well-known [10,11] that such systems are globally null controllable with bounded controls as long as they are controllable in the usual linear system sense.

In regard to stabilizability, it is shown in [12] that a linear system subject to actuator saturation can be globally asymptotically stabilized by smooth feedback if and only if the system is asymptotically null controllable with bounded controls (ANCBC), which, as shown in [10, 11], is equivalent to the system being stabilizable in the usual linear sense and having open loop

[1]Work supported in part by the US Office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

poles in the closed left-half plane. A nested feedback design technique for designing nonlinear globally asymptotically stabilizing feedback laws was proposed in [14] for a chain of integrators and was fully generalized in [13].

The notion of semiglobal asymptotic stabilization on the null controllable region for linear systems subject to actuator saturation was introduced in [7]. The semiglobal framework for stabilization requires feedback laws that yield a closed-loop system which has an asymptotically stable equilibrium whose domain of attraction includes an *a priori* given (arbitrarily large) bounded subset of the null controllable region. In [7], it was shown that, for linear ANCBC systems subject to actuator saturation, one can achieve semiglobal asymptotic stabilization on the null controllable region using linear feedback laws.

On the other hand, the counterparts of the above mentioned results for exponentially unstable linear systems are less understood. Recently, we made an attempt to systematically study issues related to null controllable regions and the stabilizability on them of exponentially unstable linear systems subject to actuator saturation and gave a rather clear understanding of these issues [4]. Specifically, we gave a simple exact description of the null controllable region for a general anti-stable linear system in terms of a set of extremal trajectories of its time-reversed system. We also constructed feedback laws that semiglobally asymptotically stabilize any linear time invariant system with two exponentially unstable poles on its null controllable region. This is in the sense that, for any *a priori* given set in the interior of the null controllable region, there exists a linear feedback law that yields a closed-loop system which has an asymptotically stable equilibrium whose domain of attraction includes the given set. One critical assumption made in [4] is that the actuator saturation is symmetric. The symmetry of the saturation function to a large degree simplifies the analysis of the closed-loop system, it, however, excludes the application of the results to many practical systems.

The goal of this chapter is to generalize the results of [4] to the case where the actuator saturation is asymmetric. We will first characterize the null controllable region and then study the problem of stabilization. We take a similar approach as in [4] to characterize the null controllable region. In studying the problem of stabilization, we found the methods used in [4] to derive the main results not applicable to the asymmetric case, since the methods rely mainly on the symmetric property of the saturation function. For a planar anti-stable system under a given saturated linear feedback, we showed in [4] that the boundary of the domain of attraction is the unique limit cycle of the closed-loop system. The uniqueness of the limit cycle was established on the symmetric property of the vector field

and the trajectories. We further showed that if the gain is increased along the direction of the LQR feedback, then the domain of attraction can be made arbitrarily close to the null controllable region. This result was also obtained by applying the symmetric property of the trajectories.

In this chapter, we propose a quite different approach to solving these problems for the case of asymmetric saturation. In particular, we will construct a Lyapunov function from the closed trajectory, and show that under certain condition, the Lyapunov function is decreasing within the closed trajectory, thus verifying that the closed trajectory forms the boundary of the domain of attraction. If the state feedback is obtained from the LQR method, then there is a unique closed trajectory (a limit cycle). We will also show that if the gain is increased along the direction of the LQR feedback, then the domain of attraction can be made arbitrarily close to the null controllable region. This result will be developed by a careful examination of the vector field of the closed-loop system.

For higher order systems with two anti-stable modes, we have similar results as in the symmetric case: given any compact subset of the null controllable region, there is a controller (switching between two saturated linear feedback laws) that achieves a domain of attraction which includes the given compact subset of the null controllable region.

3.2. Preliminaries and Notation

Consider a linear system

$$\dot{x}(t) = Ax(t) + bu(t), \tag{3.1}$$

where $x(t) \in \mathbf{R}^n$ is the state and $u(t) \in \mathbf{R}$ is the control. Given real numbers $u^- < 0$ and $u^+ > 0$, define

$$\mathcal{U}_a := \left\{ u : u \text{ is measurable and } u^- \leq u(t) \leq u^+, \ \forall t \in \mathbf{R} \right\}. \tag{3.2}$$

A control signal u is said to be *admissible* if $u \in \mathcal{U}_a$. In this chapter, we are interested in the control of the system (3.1) by using admissible controls. Our first concern is the set of states that can be steered to the origin by an admissible control.

Definition 3.1. A state x_0 is said to be null controllable if there exist a $T \in [0, \infty)$ and an admissible control u such that the state trajectory $x(t)$ of the system satisfies $x(0) = x_0$ and $x(T) = 0$. The set of all null controllable states is called the null controllable region of the system and is denoted by $\mathcal{C}$.

With the above definition, we have

$$\mathcal{C} = \bigcup_{T\in[0,\infty)} \left\{ x = -\int_0^T e^{-A\tau} b u(\tau) d\tau : \ u \in \mathcal{U}_a \right\}. \tag{3.3}$$

Remark 3.1. For a linear system with multiple inputs,

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{3.4}$$

where $u \in \mathbf{R}^m$, and $u_i \in [u_i^-, u_i^+]$, $u_i^- < 0, u_i^+ > 0$, let $\mathcal{C}_i$ be the null controllable region of the system

$$\dot{x}(t) = Ax(t) + b_i u_i(t),$$

then it is easy to verify that the null controllable region of the system (3.4) is

$$\mathcal{C} = \sum_{i=1}^m \mathcal{C}_i = \left\{ x_1 + x_2 + \cdots + x_m : \ x_i \in \mathcal{C}_i, \ i = 1, 2, \cdots, m \right\}.$$

Hence, it is without loss of generality that we consider the single input system (3.1).

For simplicity, a linear system and the matrix A are said semistable if all the eigenvalues of A are in the closed left half plane; and anti-stable if all the eigenvalues of A are in the open right half plane.

We recall a fundamental result from the literature [2, 10, 11]:

Proposition 3.1. Assume that (A, b) is controllable.

a) If A is semistable, then $\mathcal{C} = \mathbf{R}^n$.

b) If A is anti-stable, then $\mathcal{C}$ is a bounded convex open set containing the origin.

c) If $A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$ with $A_1 \in \mathbf{R}^{n_1 \times n_1}$ anti-stable and $A_2 \in \mathbf{R}^{n_2 \times n_2}$ semistable, and b is partitioned as $\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ accordingly, then

$$\mathcal{C} = \mathcal{C}_1 \times \mathbf{R}^{n_2},$$

where $\mathcal{C}_1$ is the null controllable region of the anti-stable system

$$\dot{x}_1(t) = A_1 x_1 + b_1 u(t).$$

Because of this proposition, we can concentrate on the study of null controllable regions of anti-stable systems. For this kind of systems,

$$\overline{\mathcal{C}} = \left\{ x = -\int_0^\infty e^{-A\tau} b u(\tau) d\tau : \ u \in \mathcal{U}_a \right\}, \tag{3.5}$$

where $\overline{\mathcal{C}}$ denotes the closure of $\mathcal{C}$. We also use "∂" to denote the boundary of a set. In this chapter, we will derive a method for explicitly describing $\partial\mathcal{C}$ in Section 3.

In the study of null controllable regions we will assume, without loss of generality, that (A, b) is controllable and A is anti-stable.

For a general system

$$\dot{x} = f(x, u), \tag{3.6}$$

its time-reversed system is

$$\dot{z} = -f(z, v). \tag{3.7}$$

Consider the time-reversed system of (3.1):

$$\dot{z}(t) = -Az(t) - bv(t). \tag{3.8}$$

Definition 3.2. A state z_f is said to be reachable if there exist $T \in [0, \infty)$ and an admissible control v such that the state trajectory $z(t)$ of the system (3.8) satisfies $z(0) = 0$ and $z(T) = z_f$. The set of all reachable states is called the reachable region of the system (3.8) and is denoted by $\mathcal{R}$.

It is known that $\mathcal{C}$ of (3.1) is the same as $\mathcal{R}$ of (3.8) (see, e.g., [8]). To avoid confusion, we will continue to use the notation x, u and $\mathcal{C}$ for the original system (3.1), and z, v and $\mathcal{R}$ for the time-reversed system (3.8).

3.3. Null Controllable Regions

In Section 3.3.1, we show that the boundary of the null controllable region of a general anti-stable linear system with saturating actuator is composed of a set of extremal trajectories of the time-reversed system. The descriptions of this set are further simplified for systems with only real poles and for systems with complex poles in Sections 3.3.2 and 3.3.3, respectively.

3.3.1. General Description of Null Controllable Regions

We will characterize the null controllable region $\mathcal{C}$ of the system (3.1) through studying the reachable region $\mathcal{R}$ of its time-reversed system (3.8).

Since A is anti-stable, we have

$$\overline{\mathcal{R}} = \left\{ z = -\int_0^\infty e^{-A\tau} b v(\tau) d\tau : \ v \in \mathcal{U}_a \right\}$$
$$= \left\{ z = -\int_{-\infty}^0 e^{A\tau} b v(\tau) d\tau : \ v \in \mathcal{U}_a \right\}.$$

Noticing that $e^{A\tau} = e^{-A(0-\tau)}$, we see that a point z in $\overline{\mathcal{R}}$ is a state of the time-reversed system (3.8) at $t = 0$ by applying an admissible control v from $-\infty$ to 0.

Define the asymmetric sign function $\mathrm{sgn}_a(\cdot)$ as

$$\mathrm{sgn}_a(r) := \begin{cases} u^+, & r > 0, \\ 0, & r = 0, \\ u^-, & r < 0. \end{cases}$$

Theorem 3.1.

$$\partial\mathcal{R} = \left\{ z = -\int_{-\infty}^0 e^{A\tau} b \ \mathrm{sgn}_a \left(c' e^{A\tau} b \right) d\tau : \ c \neq 0 \right\}. \tag{3.9}$$

$\overline{\mathcal{R}}$ is strictly convex. Moreover, for each $z^* \in \partial\mathcal{R}$, there exists a unique admissible control v^* such that

$$z^* = -\int_{-\infty}^0 e^{A\tau} b v^*(\tau) d\tau. \tag{3.10}$$

Proof. This can be proved similarly as Theorem 2.3.1 in [4]. □

Theorem 3.1 says that for $z^* \in \partial\mathcal{R}$, there is a unique admissible control v^* satisfying (3.10). From (3.9), this implies $v^*(t) = \mathrm{sgn}_a(c' e^{At} b)$ for some $c \neq 0$ (such c, $\|c\| = 1$, may be nonunique, where $\|\cdot\|$ is the Euclidean norm). So, if v is an admissible control and there is no c such that $v(t) = \mathrm{sgn}_a(c' e^{At} b)$ for $t \leq 0$, then

$$-\int_{-\infty}^0 e^{A\tau} b v(\tau) d\tau \notin \partial\mathcal{R}$$

and must be in the interior of $\mathcal{R}$.

Since $\mathrm{sgn}_a(kc' e^{A\tau} b) = \mathrm{sgn}_a(c' e^{A\tau} b)$ for any positive number k, equation (3.9) shows that $\partial\mathcal{R}$ can be determined from the surface of a unit ball in $\mathbf{R}^n$. In what follows, we will simplify (3.9) and describe $\partial\mathcal{R}$ in terms of a set of trajectories of the time-reversed system (3.8).

Denote

$$\mathcal{E}_c := \left\{ v(t) = \operatorname{sgn}_a \left(c' e^{At} b \right), t \in \mathbf{R} : \ c \neq 0 \right\}, \tag{3.11}$$

and for an admissible control v, denote

$$\Phi(t, v) := - \int_{-\infty}^{t} e^{-A(t-\tau)} b v(\tau) d\tau. \tag{3.12}$$

Since A is anti-stable, the integral in (3.12) exists for all $t \in \mathbf{R}$, so $\Phi(t, v)$ is well defined.

If $v(t) = \operatorname{sgn}_a(c' e^{At} b)$, then

$$\begin{aligned} \Phi(t, v) &= - \int_{-\infty}^{t} e^{-A(t-\tau)} b v(\tau) d\tau \\ &= - \int_{-\infty}^{0} e^{A\tau} b \operatorname{sgn}_a \left(c' e^{At} e^{A\tau} b \right) d\tau \\ &\in \partial \mathcal{R}, \end{aligned}$$

for all $t \in \mathbf{R}$, i.e., $\Phi(t, v)$ lies entirely on $\partial \mathcal{R}$. An admissible control v such that $\Phi(t, v)$ lies entirely on $\partial \mathcal{R}$ is said to be *extremal* and such $\Phi(t, v)$ an *extremal trajectory*. On the other hand, given an admissible control $v(t)$, if there exists no c such that $v(t) = \operatorname{sgn}_a(c' e^{At} b)$ for all $t \leq 0$, then by Theorem 3.1, $\Phi(0, v) \notin \partial \mathcal{R}$ and must be in the interior of $\mathcal{R}$. By the time invariance property of the system, if there exists no c such that $v(t) = \operatorname{sgn}_a(c' e^{At} b)$ for all $t \leq t_0$, $\Phi(t, v)$ must be in the interior of $\mathcal{R}$ for all $t \geq t_0$. Consequently, $\mathcal{E}_c$ is the set of extremal controls.

Definition 3.3. $v_1, v_2 \in \mathcal{E}_c$ are said to be equivalent, denoted by $v_1 \sim v_2$, if there exists an $h \in \mathbf{R}$ such that $v_1(t) = v_2(t - h)$ for all $t \in \mathbf{R}$.

The following theorem shows that $\partial \mathcal{R}$ is covered by a minimal subset of the extremal trajectories.

Theorem 3.2. Let $\mathcal{E}_c^m \subset \mathcal{E}_c$ be such that for every $v \in \mathcal{E}_c$, there exists a unique $v_1 \in \mathcal{E}_c^m$ such that $v \sim v_1$. Then

$$\partial \mathcal{R} = \left\{ \Phi(t, v) : \ t \in \mathbf{R}, \ v \in \mathcal{E}_c^m \right\}. \tag{3.13}$$

Proof. For any fixed $t \in \mathbf{R}$, it follows from (3.9) that

$$\begin{aligned} \partial \mathcal{R} &= \left\{ - \int_{-\infty}^{t} e^{-A(t-\tau)} b \operatorname{sgn}_a \left(c' e^{-At} e^{A\tau} b \right) d\tau : \ c \neq 0 \right\} \\ &= \left\{ - \int_{-\infty}^{t} e^{-A(t-\tau)} b \operatorname{sgn}_a \left(c' e^{A\tau} b \right) d\tau : \ c \neq 0 \right\}, \end{aligned}$$

i.e., $\partial\mathcal{R} = \{\Phi(t,v) : \ v \in \mathcal{E}_c\}$, for any fixed $t \in \mathbf{R}$. So $\partial\mathcal{R}$ can be viewed as the set of extremal trajectories at any frozen time. Now let t vary, then each point on $\partial\mathcal{R}$ moves along a trajectory but the whole set is invariant. So we can also write $\partial\mathcal{R} = \{\Phi(t,v) : \ v \in \mathcal{E}_c, \ t \in \mathbf{R}\}$, which is equivalent to

$$\partial\mathcal{R} = \left\{\Phi(t,v) : \ t \in \mathbf{R}, \ v \in \mathcal{E}_c\right\}. \tag{3.14}$$

Noting that a shift in time of the control corresponds to the same shift of the state trajectory, we see that, if $v_1 \sim v_2$, then

$$\left\{\Phi(t,v_1) : \ t \in \mathbf{R}\right\} = \left\{\Phi(t,v_2) : \ t \in \mathbf{R}\right\}.$$

And (3.13) follows from the property of $\mathcal{E}_c^m$. □

It turns out that for some classes of systems, $\mathcal{E}_c^m$ can be easily described. For second order systems, $\mathcal{E}_c^m$ contains only one or two elements, so $\partial\mathcal{R}$ can be covered by no more than two trajectories; and for third order systems, $\mathcal{E}_c^m$ can be described using a parameter that varies within a real interval. We will see later that for systems of different eigenvalue structures, the descriptions of $\mathcal{E}_c^m$ can be quite different. All the results in the following subsections are easy extension of the counterparts in [4], hence the proofs are omitted.

3.3.2. Systems with Only Real Eigenvalues

It follows from, for example, [8, p. 77], that if A has only real eigenvalues and $c \neq 0$, then $c'e^{At}b$ has at most $n-1$ zeros. This implies that an extremal control can have at most $n-1$ switches. It was shown in [4] that the converse is also true.

Theorem 3.3. For the system (3.8), assume that A has only real eigenvalues, then

a) an extremal control has at most $n-1$ switches;

b) any bang-bang control with $n-1$ or less switches is an extremal control.

By Theorem 3.3, the set of extremal controls can be described as follows,

$$\mathcal{E}_c = \left\{ \mathrm{sgn}_a(\pm u) : u(t) = \begin{cases} 1, & t \in [-\infty, t_1), \\ (-1)^i, & t \in [t_i, t_{i+1}), \\ (-1)^{n-1}, & t \in [t_{n-1}, \infty), \end{cases} \right.$$
$$\left. -\infty < t_1 < t_2 \leq \cdots \leq t_{n-1} \leq \infty \right\} \cup \left\{u^+, u^-\right\},$$

where u^+ (or u^-) denotes a constant control $v(t) \equiv u^+$ (or u^-). Here we allow $t_i = t_{i+1}$ $(i \neq 1)$ and $t_{n-1} = \infty$, so the above description of $\mathcal{E}_c$ consists of all bang-bang controls with $n-1$ or less switches.

By setting $t_1 = 0$, we immediately get $\mathcal{E}_c^m$,

$$\mathcal{E}_c^m = \left\{ \mathrm{sgn}_a(\pm u) : u(t) = \begin{cases} 1, & t \in [-\infty, t_1), \\ (-1)^i, & t \in [t_i, t_{i+1}), \\ (-1)^{n-1}, & t \in [t_{n-1}, \infty), \end{cases} \right.$$
$$\left. 0 = t_1 < t_2 \leq \cdots \leq t_{n-1} \leq \infty \right\} \cup \left\{ u^+, u^- \right\}.$$

For each $v \in \mathcal{E}_c^m$, we have $v(t) = u^-$ (or u^+) for all $t \leq 0$. Hence, for $t \leq 0$,

$$\Phi(t, v) = -\int_{-\infty}^{t} e^{-A(t-\tau)} b v(t) d\tau = -A^{-1} b u^- \ (\text{ or } \ -A^{-1} b u^+).$$

And for $t > 0$, $v(t)$ is a bang-bang control with $n-2$ or less switches. Denote $z_e^+ = -A^{-1}bu^+$ and $z_e^- = -A^{-1}bu^-$, then from Theorem 3.2 we have,

Observation 3.3.1. $\partial\mathcal{R} = \partial\mathcal{C}$ is covered by two bunches of trajectories. The first bunch consists of trajectories of (3.8) when the initial state is z_e^+ and the input is a bang-bang control that starts at $t = 0$ with $v = u^-$ and has $n-2$ or less switches. The second bunch consists of the trajectories of (3.8) when the initial state is z_e^- and the input is a bang-bang control that starts at $t = 0$ with $v = u^+$ and has $n-2$ or less switches.

Remark 3.2. Since the trajectories of the time-reversed system (3.8) and those of the original system are the same except that their directions are opposite, we can also say that $\partial\mathcal{R} = \partial\mathcal{C}$ is covered by the trajectories of the original system under the same controls. The fundamental difference is that it is quite easy to generate the trajectories with the time-reversed system, while it is unrealistic to get the trajectories from the original system. For example, suppose that we have a trajectory of the time-reversed system that starts at z_e^+ under the control $v = u^-$, then it goes toward z_e^- since z_e^- is a stable equilibrium under the control $v = u^-$. On the contrary, if we apply $u = u^-$ to the original system with initial condition z_e^-, we cannot get a reversed trajectory because z_e^- is an (unstable) equilibrium under the control $u = u^-$. The trajectory of the time-reversed system from z_e^+ to z_e^- under the control $v = u^-$ could be partly recovered by the original system if we know one point (except z_e^-) on the trajectory. But this is unrealistic.

Furthermore, $\partial\mathcal{R}$ can be simply described in terms of the open-loop transition matrix.

$$\partial\mathcal{R} = \left\{ \left[\sum_{i=1}^{n-1} \pm(u^- - u^+)(-1)^i e^{-A(t-t_i)} - \mathrm{sgn}_a\left(\pm(-1)^n\right) I \right] A^{-1}b : \right.$$
$$\left. 0 = t_1 \le t_2 \cdots \le t_{n-1} \le t \le \infty \right\}.$$

Here, we allow $t_1 = t_2$ to include $\pm z_e^+$. For second order systems,

$$\begin{aligned}
\partial\mathcal{R} &= \left\{ e^{-At} z_e^+ - \int_0^t e^{-A(t-\tau)} b\, u^- d\tau : \ t \in [0,\infty] \right\} \\
&\quad \cup \left\{ e^{-At} z_e^- - \int_0^t e^{-A(t-\tau)} b\, u^+ d\tau : \ t \in [0,\infty] \right\} \\
&= \left\{ \left((u^- - u^+)\, e^{-At} - I\, u^- \right) A^{-1} b : \ t \in [0,\infty] \right\} \\
&\quad \cup \left\{ \left((u^+ - u^-)\, e^{-At} - I\, u^+ \right) A^{-1} b : \ t \in [0,\infty] \right\}. \quad (3.15)
\end{aligned}$$

If $n = 3$, then one half of $\partial\mathcal{R} = \partial\mathcal{C}$ can be formed by the trajectories of (3.8) starting from z_e^+ with the control initially being $v = u^-$ and then switching at any time to $v = u^+$. So the trajectories go toward z_e^- at first then turn back toward z_e^+. The other half is formed by the trajectories of (3.8) starting from z_e^- with the control initially being $v = u^+$ and then switching at any time to $v = u^-$. So the trajectories go toward z_e^+ at first then turn back toward z_e^-. That is,

$$\begin{aligned}
\partial\mathcal{R} = &\left\{ e^{-At} z_e^+ - \int_0^{t_2} e^{-A(t-\tau)} b\, u^- d\tau - \int_{t_2}^{t} e^{-A(t-\tau)} b\, u^+ d\tau : \ 0 \le t_2 \le t \le \infty \right\} \\
\cup &\left\{ e^{-At} z_e^- - \int_0^{t_2} e^{-A(t-\tau)} b\, u^+ d\tau - \int_{t_2}^{t} e^{-A(t-\tau)} b\, u^- d\tau : \ 0 \le t_2 \le t \le \infty \right\}.
\end{aligned}$$

3.3.3. Systems with Complex Eigenvalues

For a system with complex eigenvalues, the set $\mathcal{E}_c^m$ is harder to determine. In what follows, we consider two important cases.

Case 1. $A \in \mathbf{R}^{2\times 2}$ has a pair of complex eigenvalues $\alpha \pm j\beta$, $\alpha, \beta > 0$.

It can be verified that

$$\left\{ \mathrm{sgn}_a\left(c' e^{At} b\right) : \ c \neq 0 \right\} = \left\{ \mathrm{sgn}_a(\sin(\beta t + \theta)) : \ \theta \in [0, 2\pi) \right\}.$$

Hence the set of extremal controls is

$$\mathcal{E}_c = \left\{ v(t) = \mathrm{sgn}_a(\sin(\beta t + \theta)),\ t \in \mathbf{R} :\ \theta \in [0, 2\pi) \right\}.$$

It is easy to see that

$$\mathcal{E}_c^m = \left\{ v(t) = \mathrm{sgn}_a(\sin(\beta t)),\ t \in \mathbf{R} \right\}$$

contains only one element. Denote $T_p = \frac{\pi}{\beta}$, then $e^{-AT_p} = -e^{-\alpha T_p} I$. Let

$$z_s^- = \left(1 - e^{-\alpha T_p}\right)^{-1} \left(-u^- + e^{-\alpha T_p} u^+\right) A^{-1} b,$$

and

$$z_s^+ = \left(1 - e^{-\alpha T_p}\right)^{-1} \left(-u^+ + e^{-\alpha T_p} u^-\right) A^{-1} b.$$

It can be verified that the extremal trajectory corresponding to $v(t) = \mathrm{sgn}_a(\sin(\beta t))$ is periodic with period $2T_p$: in the first half period it goes from z_s^- to z_s^+ under the control $v = u^+$ and in the second half period it goes from z_s^+ to z_s^- under the control $v = u^-$. That is,

$$\begin{aligned}
\partial\mathcal{R} &= \left\{ e^{-At} z_s^- - \int_0^t e^{-A(t-\tau)} b\, u^+ d\tau \ :\ t \in [0, T_p] \right\} \\
&\quad \cup \left\{ e^{-At} z_s^+ - \int_0^t e^{-A(t-\tau)} b\, u^- d\tau \ :\ t \in [0, T_p] \right\} \\
&= \left\{ e^{-At} z_s^- - \left(I - e^{-At}\right) A^{-1} b\, u^+ :\ t \in [0, T_p] \right\} \\
&\quad \cup \left\{ e^{-At} z_s^+ - \left(I - e^{-At}\right) A^{-1} b\, u^- :\ t \in [0, T_p] \right\}. \qquad (3.16)
\end{aligned}$$

Case 2. $A \in \mathbf{R}^{3\times 3}$ has eigenvalues $\alpha \pm j\beta$ and α_1, with $\alpha, \beta, \alpha_1 > 0$.

a) $\alpha = \alpha_1$. Then similar to Case 1,

$$\mathcal{E}_c = \left\{ v(t) = \mathrm{sgn}_a(k + \sin(\beta t + \theta)),\ t \in \mathbf{R} :\ k \in \mathbf{R},\ \theta \in [0, 2\pi) \right\}.$$

Since $\mathrm{sgn}_a(k + \sin(\beta t + \theta))$ is the same for all $k \geq 1$ (or $k \leq -1$), we have

$$\mathcal{E}_c^m = \left\{ v(t) = \mathrm{sgn}_a(k + \sin(\beta t)),\ t \in \mathbf{R} :\ k \in [-1, 1] \right\}.$$

Each $v \in \mathcal{E}_c^m$ is periodic with period $2T_p$, but the lengths of positive and negative parts vary with k. $\Phi(t, v)$ can be easily determined from simulation. A formula can also be derived for $\Phi(t, v)$.

b) $\alpha \neq \alpha_1$. Then

$$\mathcal{E}_c = \left\{ v(t) = \mathrm{sgn}_a\left(k_1 e^{(\alpha_1-\alpha)t} + k_2 \sin(\beta t + \theta)\right),\ t \in \mathbf{R}: \right.$$
$$\left. (k_1, k_2) \neq (0,0), \theta \in [0, 2\pi) \right\}.$$

It can be shown that

$$\mathcal{E}_c^m = \left\{u^+, u^-\right\} \cup \left\{v(t) = \mathrm{sgn}_a(\sin(\beta t))\right\} \cup \mathcal{E}_{c3}^m.$$

where u^+ (or u^-) denotes a constant control $v(t) \equiv u^+$ (or u^-) and

$$\mathcal{E}_{c3}^m = \left\{ v(t) = \mathrm{sgn}_a\left(\pm e^{(\alpha_1-\alpha)t} + \sin(\beta t + \theta)\right), t \in \mathbf{R}:\ \theta \in [0, 2\pi) \right\}. \tag{3.17}$$

When $\alpha_1 < \alpha$, for each $v \in \mathcal{E}_{c3}^m$, $v(t) = u^+$ (or u^-) for all $t \leq 0$, so the corresponding extremal trajectories stay at $z_e^+ = -A^{-1}bu^+$ (or $z_e^- = -A^{-1}bu^-$) before $t = 0$. And after some time, they go toward a periodic trajectory since as t goes to infinity, $v(t)$ becomes periodic; When $\alpha_1 > \alpha$, for each $v \in \mathcal{E}_{c3}^m$, $v(t) = u^+$ (or u^-) for all $t \geq 0$, and the corresponding extremal trajectories start from near periodic and go toward z_e^+ or z_e^-.

Plotted in Figure 1 are some extremal trajectories on $\partial\mathcal{R}$ of the time-reversed system (3.8) with

$$A = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.8 & -2 \\ 0 & 2 & 0.8 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad u^+ = 1, \quad u^- = -0.5.$$

For higher order systems, the relative locations of the eigenvalues are more diversified and the analysis will be technically much more involved. It can, however, be expected that in the general case, the number of parameters used to describe $\mathcal{E}_c^m$ is $n - 2$.

3.4. Domain of Attraction under Saturated Linear State Feedback

Consider the open loop system,

$$\dot{x}(t) = Ax(t) + bu(t), \tag{3.18}$$

with admissible control $u \in \mathcal{U}_a$. A saturated linear state feedback is given by $u = \mathrm{sat}_a(fx)$, where $f \in \mathbf{R}^{1\times n}$ is the feedback gain and $\mathrm{sat}_a(\cdot)$ is the

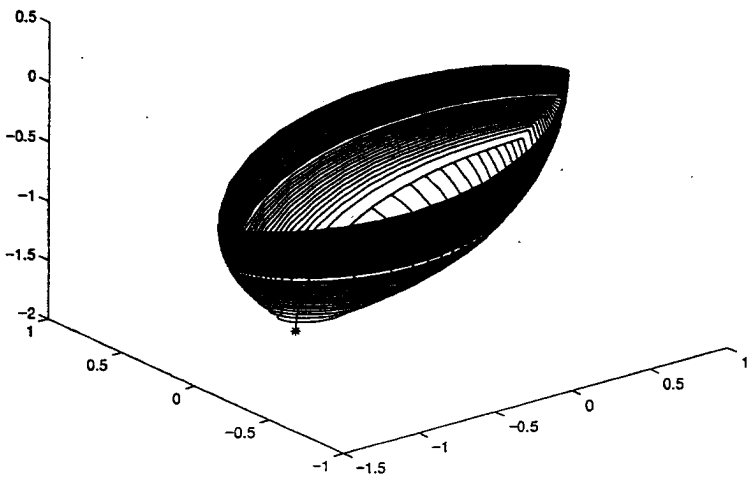


Figure 1: Extremal trajectories on $\partial\mathcal{R}$, $\alpha_1 < \alpha$.

asymmetric saturation function

$$\mathrm{sat}_a(r) = \begin{cases} u^+, & r > u^+, \\ r, & r \in [u^-, u^+], \\ u^-, & r < u^-. \end{cases}$$

Such a feedback is said to be stabilizing if $A + bf$ is asymptotically stable. With a saturated linear state feedback applied, the closed loop system is

$$\dot{x}(t) = Ax(t) + b\,\mathrm{sat}_a(fx(t)). \tag{3.19}$$

Denote the state transition map of (3.19) by $\phi : (t, x_0) \mapsto x(t)$. The domain of attraction $\mathcal{S}$ of the equilibrium $x = 0$ of (3.19) is defined by

$$\mathcal{S} := \left\{ x_0 \in \mathbf{R}^n : \lim_{t\to\infty} \phi(t, x_0) = 0 \right\}.$$

Obviously, $\mathcal{S}$ must lie within the null controllable region $\mathcal{C}$ of the system (3.18). Therefore, a design problem is to choose a state feedback gain so that $\mathcal{S}$ is close to $\mathcal{C}$. We refer to this problem as semiglobal stabilization on the null controllable region.

We will first deal with anti-stable planar systems, then extend the results to higher order systems with only two anti-stable modes. Consider the

system (3.19). Assume that $A \in \mathbf{R}^{2\times 2}$ is anti-stable. For the symmetric case where $u^- = -u^+$, it was shown in [4] that $\partial\mathcal{S}$ is the unique limit cycle of the system (3.19). This limit cycle is unstable for (3.19) but is stable for the time-reversed system of (3.19). So it can be easily obtained by simulating the time-reversed system.

However, the method used in [4] to prove the uniqueness of the limit cycle relies on the symmetric property of the vector field. There is no obvious way to generalize the method to the asymmetric case. In this section, we will present a quite different approach to this problem. Actually, we will construct a Lyapunov function from the closed trajectory, and show that the Lyapunov function decreases in time as long as the trajectory starts from within the closed trajectory. Therefore, the open set enclosed by the closed trajectory is the domain of attraction.

Lemma 3.1. The origin is the unique equilibrium point of the system (3.19) and there is a closed-trajectory.

Proof. The other two candidate equilibrium points are $x_e^+ = -A^{-1}bu^+$ and $x_e^- = -A^{-1}bu^-$. For x_e^+ to be an equilibrium, we must have $fx_e^+ \geq u^+$ so that $u = \mathrm{sat}_a(fx_e^+) = u^+$. Since A is anti-stable and (A, b) is controllable, we can assume without loss of generality that

$$A = \begin{bmatrix} 0 & 1 \\ -a_1 & a_2 \end{bmatrix}, \quad a_1, a_2 > 0, \quad b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

This implies that if $f = \begin{bmatrix} f_1 & f_2 \end{bmatrix}$ is stabilizing, then $f_1/a_1 < 1$. It can be easily shown that

$$fx_e^+ = -fA^{-1}bu^+ = (f_1/a_1)u^+ < u^+.$$

This rules out x_e^+ to be an equilibrium point. The other candidate x_e^- can be ruled out similarly.

The existence of a closed-trajectory can be proven similarly as in [4]. □

Suppose that Γ is a closed trajectory. By the index theory (see, e.g., [6]), Γ must enclose the origin. Also, Γ must have two intersections with one of the lines $fx = u^+$ or $fx = u^-$, or both of them. Otherwise there would be a closed trajectory completely in the linear region of the vector field.

Proposition 3.2. Denote the region enclosed by a closed trajectory Γ as Ω, then Ω is convex.

Proof. We start the proof by showing some general properties of second order linear autonomous systems,

$$\dot{x}(t) = Ax(t), \quad \det(A) \neq 0. \tag{3.20}$$

Denote $\theta(t) = \angle x(t)$, $r(t) = \|x(t)\|$, then $x(t) = r(t)\begin{bmatrix} \cos\theta(t) \\ \sin\theta(t) \end{bmatrix}$, and it can be verified that

$$\dot{\theta} = \begin{bmatrix} \cos\theta & \sin\theta \end{bmatrix}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} A \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}. \tag{3.21}$$

Noting that

$$\begin{bmatrix} \cos\theta & \sin\theta \end{bmatrix}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} = 0,$$

equation (3.21) has at most four equilibriums on $[0, 2\pi)$, which correspond to the directions of the real eigenvectors of A. If $\theta(0)$ is an equilibrium, then $\theta(t)$ is a constant, and $x(t)$ is a straight line. If $\theta(0)$ is not an equilibrium, then $\theta(t)$ will never reach an equilibrium at finite time. If it reaches, the trajectory $x(t)$ will intersect with a straight line trajectory, which is impossible. Hence, if x_0 is not an eigenvector of A, $\dot{\theta}(t)$ will never be equal to zero, thus $\dot{\theta}(t)$ will be sign definite. This shows that $\angle x(t)$ is strictly monotonously increasing (or decreasing).

Let's now consider the direction angle of the trajectory, $\angle\dot{x}(t) =: \gamma(t)$. Since $\ddot{x}(t) = A\dot{x}(t)$, by the same argument, $\gamma(t)$ also increases (or decreases) monotonically. We claim that if the system is asymptotically stable or anti-stable, then $\dot{\theta}(t)$ and $\dot{\gamma}(t)$ have the same sign, i.e., the trajectories bend toward the origin; if the signs of the two eigenvalues are different, then $\dot{\theta}(t)$ and $\dot{\gamma}(t)$ have opposite signs. This can be simply shown as follows. Rewrite (3.21) as,

$$\dot{\theta} = \frac{1}{\|x\|^2} x' \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} Ax. \tag{3.22}$$

Similarly,

$$\dot{\gamma} = \frac{1}{\|\dot{x}\|^2}\dot{x}'\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} A\dot{x} = \frac{1}{\|\dot{x}\|^2} x'A'\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} AAx.$$

It is trivial to verify that $A'\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} A = \det(A)\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ for a 2×2 matrix A. So

$$\dot{\gamma} = \frac{\det(A)}{\|\dot{x}\|^2} x' \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} Ax = \frac{\|x\|^2 \det(A)}{\|\dot{x}\|^2}\dot{\theta}. \tag{3.23}$$

This shows that the claim is true.

Now we can apply the above result to the closed trajectory Γ. We refer to Figure 2. Assume that Γ goes anti-clockwise and intersects both the lines $fx = u^+$ and $fx = u^-$ (the arguments also apply if Γ only intersects one

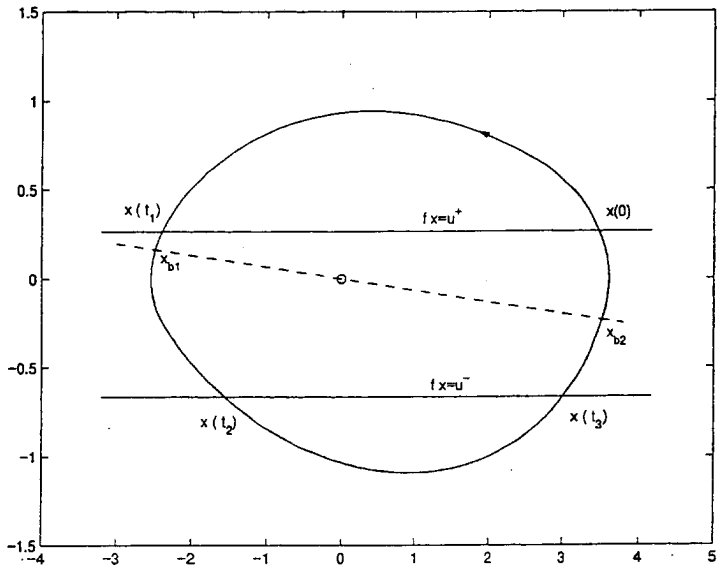


Figure 2: Illustration for the proof of Proposition 3.2.

of the straight lines). Let the intersections be $x(0), x(t_1), x(t_2)$ and $x(t_3)$. From $x(0)$ to $x(t_1)$, we have $\dot{x} = Ax + bu^+$, which can be rewritten as

$$d(x - x_e^+)/dt = A(x - x_e^+).$$

By Lemma 3.1, x_e^+ is below the line $fx = u^+$. So from $x(0)$ to $x(t_1)$, $\angle[x(t) - x_e^+]$ increases, and hence $\angle\dot{x}(t)$ increases (since A is anti-stable). From $x(t_1)$ to $x(t_2)$, we have $\dot{x} = (A + bf)x$, so $\angle x(t)$ and $\angle\dot{x}(t)$ also increase (since $A + bf$ is stable). Similarly, $\angle\dot{x}(t)$ increases from $x(t_2)$ to $x(t_3)$, and from $x(t_3)$ to $x(0)$. It is straightforward to verify that $\dot{x}(t)$ is continuous at $x(0), x(t_1), x(t_2)$ and $x(t_3)$. Hence, $\angle\dot{x}(t)$, the direction angle, is monotonically increasing along Γ. This implies that the region Ω enclosed by Γ is convex. □

The following theorem shows that under certain condition, a closed trajectory Γ is the boundary of the domain of attraction.

Theorem 3.4. Let Γ be a closed-trajectory of the system (3.19). Let the intersections of Γ with the line $\{\mu A^{-1}b : \mu \in \mathbf{R}\}$ be x_{b1} and x_{b2} (see Figure 2). If $fx_{b1}, fx_{b2} \in [u^-, u^+]$, i.e., the two intersections x_{b1} and x_{b2} are between the two lines $fx = u^-$ and $fx = u^+$, then $\partial\mathcal{S} = \Gamma$.

Proof. Without loss of generality, we assume that

$$A = \begin{bmatrix} 0 & -a_1 \\ 1 & a_2 \end{bmatrix}, \quad a_1, a_2 > 0, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

and $f = \begin{bmatrix} 0 & 1 \end{bmatrix}$. Then $fx = u^-$ and $fx = u^+$ are two horizontal lines (see Figure 3). Since $A + bf$ is Hurwitz, we have $b_1 < a_1, b_2 < -a_2$ and that the trajectories go anticlockwise.

Denote the region enclosed by Γ as Ω. Since Ω contains the origin in its interior, we can define a Minkowski functional

$$\kappa(x) := \min\left\{\gamma \geq 0 : x \in \gamma\Omega\right\}.$$

(If Ω is symmetric and convex, $\kappa(x)$ is a norm). Clearly, $\kappa(x) = 1$ for all $x \in \Gamma$. Since Γ is a trajectory and the vector field $\dot{x}$ in (3.19) is continuous, $\frac{\partial\kappa(x)}{\partial x}$ exists and is continuous along Γ. Since Ω is bounded and convex, it follows that $\frac{\partial\kappa(x)}{\partial x} \neq 0$ for all $x \in \Gamma$. Note that $\frac{\partial\kappa(x)}{\partial x}$ is the gradient of the function $\kappa(x)$, so it is perpendicular to the tangent of the curve $\Gamma = \{x \in \mathbf{R}^2 : \kappa(x) = 1\}$, which is $\dot{x}$. Therefore,

$$\left(\frac{\partial\kappa(x)}{\partial x}\right)' \dot{x} = 0, \quad \forall x \in \Gamma. \tag{3.24}$$

Define a Lyapunov function as $V(x) := \frac{1}{2}\kappa^2(x)$. It can be verified that for any constant $\alpha > 0$,

$$\kappa(\alpha x) = \alpha\kappa(x), \quad V(\alpha x) = \alpha^2 V(x),$$

and

$$\left.\frac{\partial\kappa(x)}{\partial x}\right|_{x=\alpha x_0} = \left.\frac{\partial\kappa(x)}{\partial x}\right|_{x=x_0}.$$

Since $\frac{\partial V(x)}{\partial x} = \kappa(x)\frac{\partial\kappa(x)}{\partial x}$, $\frac{\partial V(x)}{\partial x}$ exists and is continuous for all $x \in \mathbf{R}^2$. It follows that

$$\left.\frac{\partial V(x)}{\partial x}\right|_{x=\alpha x_0} = \alpha \left.\frac{\partial V(x)}{\partial x}\right|_{x=x_0}, \tag{3.25}$$

and

$$\left(\frac{\partial V(x)}{\partial x}\right)' \dot{x} = 0, \quad \frac{\partial V(x)}{\partial x} \neq 0, \quad \forall x \in \Gamma. \tag{3.26}$$

We conclude that for all $x \in \Omega$, along the trajectory of the system (3.19),

$$\dot{V}(x) = \left(\frac{\partial V(x)}{\partial x}\right)' \dot{x} = \left(\frac{\partial V(x)}{\partial x}\right)' (Ax + b\,\mathrm{sat}_a(fx)) \leq 0. \tag{3.27}$$

This will be proved in the following.

With the special form of A, b and f, the trajectory Γ goes anti-clockwise. Suppose that it starts at the righthand side intersection $x(0)$ with $fx = u^+$ and intersects $fx = u^+$ and $fx = u^-$ at $x(t_1), x(t_2)$ and $x(t_3)$. We partition the curve Γ into four parts, $\Gamma_1 : x(t_3) \to x(0)$; $\Gamma_2 : x(0) \to x(t_1)$; $\Gamma_3 : x(t_1) \to x(t_2)$ and $\Gamma_4 : x(t_2) \to x(t_3)$ (see Figure 3). Here we note that it might happen that Γ only intersects one of the two straight lines, say $fx = u^+$. In this case, Γ_1 and Γ_3 are merged into one connected curve and we don't have Γ_4. We will see that the following proof can be easily adapted for this case.

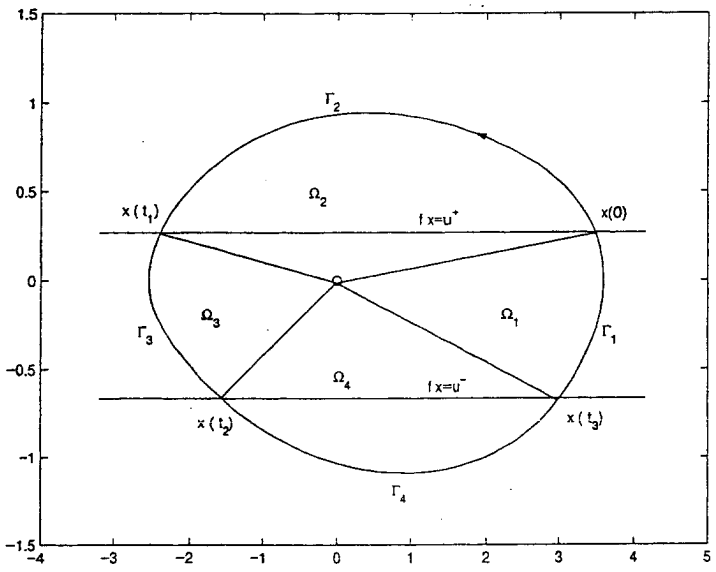


Figure 3: Illustration for the proof of Theorem 3.4.

Let $\Omega_i, i = 1, 2, 3, 4$ be the region enclosed by Γ_i and the two straight lines that connect the origin and the two end points of Γ_i.

We now consider $\dot{V}(x)$ for x in the interior of Ω. In other words, we consider $\dot{V}(\alpha x_r)$ for $\alpha \in (0, 1)$ and $x_r \in \Gamma$. In the sequal, we use $\frac{\partial V(x_r)}{\partial x}$ to denote $\frac{\partial V(x)}{\partial x}\Big|_{x=x_r}$ for simplicity.

1) If $x_r \in \Gamma_1 \cup \Gamma_3$, then $\alpha x_r \in \Omega_1 \cup \Omega_3$. By (3.25), (3.26) and (3.27),

$$\begin{aligned}\frac{\dot{V}(\alpha x_r)}{\alpha} &= \left(\frac{\partial V(x_r)}{\partial x}\right)' (\alpha A x_r + b\,\mathrm{sat}_a(\alpha f x_r)) \\ &= \left(\frac{\partial V(x_r)}{\partial x}\right)' (A x_r + b f x_r)\alpha = \left(\frac{\partial V(x_r)}{\partial x}\right)' \dot{x}_r \alpha = 0.\end{aligned}$$

2) If $x_r \in \Gamma_2 \cup \Gamma_4$, then $\alpha x_r \in \Omega_2 \cup \Omega_4$ and we have

$$\begin{aligned}\frac{\dot{V}(\alpha x_r)}{\alpha} &= \left(\frac{\partial V(x_r)}{\partial x}\right)' (\alpha A x_r + b\,\mathrm{sat}_a(\alpha f x_r)) \\ &= \alpha \left(\frac{\partial V(x_r)}{\partial x}\right)' A x_r + \left(\frac{\partial V(x_r)}{\partial x}\right)' b\,\mathrm{sat}_a(\alpha f x_r) \\ &=: g(\alpha, x_r).\end{aligned}$$

Since for a fixed x_r, $\left(\frac{\partial V(x_r)}{\partial x}\right)' A x_r$, $\left(\frac{\partial V(x_r)}{\partial x}\right)' b$ and $f x_r$ are all constants, it follows that $g(\alpha, x_r)$, as a function of α, is a bended line. It bends at the point where $\alpha f x_r = u^+$ (or u^-). Also, we have $g(0, x_r) = g(1, x_r) = 0$. ($g(1, x_r) = 0$ is from (3.26)). Therefore, for a fixed x_r, we have either $g(\alpha, x_r) \geq 0$ for all $\alpha \in (0,1)$ or $g(\alpha, x_r) \leq 0$ for all $\alpha \in (0,1)$.

Denote

$$k_1(x_r) = \left(\frac{\partial V(x_r)}{\partial x}\right)' (A x_r + b f x_r), \quad k_2(x_r) = \left(\frac{\partial V(x_r)}{\partial x}\right)' A x_r.$$

Since $\frac{\partial V(x_r)}{\partial x}$ is continuous in x_r, both $k_1(x_r)$ and $k_2(x_r)$ are continuous functions of x_r. For a fixed x_r, $k_1(x_r)$ is the slope of $g(\alpha, x_r)$ for small α when $\mathrm{sat}_a(\alpha f x_r) = \alpha f x_r$, and $k_2(x_r)$ is the slope of $g(\alpha, x_r)$ for large α when $\mathrm{sat}_a(\alpha f x_r) = u^+$ (or u^-). Since $g(\alpha, x_r)$ is a bended line for a fixed x_r and $g(0, x_r) = g(1, x_r) = 0$, $k_1(x_r)$ and $k_2(x_r)$ must have opposite signs.

If $k_2(x_r) > 0$, then $g(\alpha, x_r) < 0$ for all $\alpha \in (0,1)$; If $k_2(x_r) < 0$, then $g(\alpha, x_r) > 0$ for all $\alpha \in (0,1)$; If $k_2(x_r) = 0$, then $k_1(x_r)$ must also be 0, (otherwise we would not have $g(0, x_r) = 0$). In the last case, we have

$$\left(\frac{\partial V(x_r)}{\partial x}\right)' \begin{bmatrix} A x_r & b \end{bmatrix} = 0, \quad (\text{ since } f x_r \neq 0 \text{ on } \Gamma_2 \cup \Gamma_4),$$

i.e., $A x_r$ and b are perpendicular to the same nonzero vector in $\mathbf{R}^2$, which means that there is some nonzero number μ such that $A x_r = \mu b$, or $x_r = \mu A^{-1} b$. This will not happen since we have assumed that the two intersections of Γ with the line $\{\mu A^{-1} b : \mu \in \mathbf{R}\}$ are between the two

straight lines $fx = u^-$ and $fx = u^+$. In other words, the two intersections are on Γ_1 and Γ_3.

Therefore, $k_2(x_r) \neq 0$ for all $x_r \in \Gamma_2 \cup \Gamma_4$. If $k_2(x_r)$ has the same sign '+' on Γ_2 and Γ_4, then

$$\dot{V}(\alpha x_r) = \alpha g(\alpha, x_r) < 0, \quad \forall \alpha \in (0,1), x_r \in (\Gamma_2 \cup \Gamma_4) \setminus (\Gamma_1 \cup \Gamma_3),$$

i.e., $\dot{V}(x) < 0$ for all x in the interior of Ω_2 and Ω_4. The next step is to show that this is indeed the case. Actually, by the continuity of $k_2(x_r)$ and that $k_2(x_r) \neq 0$ for all $x_r \in \Gamma_2 \cup \Gamma_4$, we only need to find one point on Γ_2 and one point on Γ_4 such that $k_2(x_r) > 0$.

With the special form of A and f, the line above the origin is $fx = u^+$ and the one below the origin is $fx = u^-$. Also a trajectory goes along Γ anti-clockwise. Hence, on Γ_2, there is a point x_r such that $\dot{x}_r = \begin{bmatrix} -d_1 \\ 0 \end{bmatrix}, d_1 > 0$ and

$$\frac{\partial V(x_r)}{\partial x} = \begin{bmatrix} 0 \\ c_1 \end{bmatrix}, \quad c_1 > 0.$$

Note that the gradient points outward of Γ. Let $x_r = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, then

$$\dot{x}_r = \begin{bmatrix} -a_1 x_2 + b_1 u^+ \\ x_1 + a_2 x_2 + b_2 u^+ \end{bmatrix} = \begin{bmatrix} -d_1 \\ 0 \end{bmatrix}.$$

So we have $x_1 + a_2 x_2 = -b_2 u^+$. From the stability of

$$A + bf = \begin{bmatrix} 0 & -a_1 + b_1 \\ 1 & a_2 + b_2 \end{bmatrix},$$

we also have, $b_1 < a_1, b_2 < -a_2 < 0$. Hence $x_1 + a_2 x_2 > 0$. It follows that

$$k_2(x_r) = \begin{bmatrix} 0 & c_1 \end{bmatrix} \begin{bmatrix} -a_1 x_2 \\ x_1 + a_2 x_2 \end{bmatrix} = c_1(x_1 + a_2 x_2) > 0.$$

Similarly, on Γ_4, there is a point x_r such that $\dot{x}_r = \begin{bmatrix} d_2 \\ 0 \end{bmatrix}, d_2 > 0$ and

$$\frac{\partial V(x_r)}{\partial x} = \begin{bmatrix} 0 \\ -c_2 \end{bmatrix}, \quad c_2 > 0.$$

In particular,

$$\dot{x}_r = \begin{bmatrix} -a_1 x_2 + b_1 u^- \\ x_1 + a_2 x_2 + b_2 u^- \end{bmatrix} = \begin{bmatrix} d_2 \\ 0 \end{bmatrix}.$$

So we have $x_1 + a_2 x_2 = -b_2 u^- < 0$. Note that $b_2 < 0$ and $u^- < 0$. It follows that

$$k_2(x_r) = \begin{bmatrix} 0 & -c_2 \end{bmatrix} \begin{bmatrix} -a_1 x_2 \\ x_1 + a_2 x_2 \end{bmatrix} > 0.$$

These show that there exist one point on Γ_2 and one point on Γ_4 such that $k_2(x_r) > 0$.

In summary of the the above analysis, we have $\dot{V}(x) < 0$, for all x in the interior of Ω_2 and Ω_4, and $\dot{V}(x) = 0$ for all $x \in \Omega_1 \cup \Omega_3$. It follows that no trajectory starting from within Ω will approach $\Gamma = \partial\Omega$.

We next show that there exists no closed trajectory within Ω. Let E be the line on the common boundary of Ω_1 and Ω_2. Suppose that there is a closed trajectory Γ_1 that intersects E at x_0. Note that Γ_1 must enclose the origin. Let the trajectory start at x_0, then it goes through $\Omega_2, \Omega_3, \Omega_4, \Omega_1$ and returns to x_0 at some t. Since $\dot{V} < 0$ in the interior of Ω_2 and Ω_4, we must have $V(x_0, t) < V(x_0, 0)$. This is a contradiction since V is independent of t.

Therefore, all the trajectories starting from within Ω will converge to the origin. Since the trajectories do not intersect each other, all the trajectories starting from outside of Γ will stay outside of it. We hence conclude that the interior of Ω is the domain of attraction. That is, $\partial\mathcal{S} = \partial\Omega = \Gamma$. □

The condition $fx_{b1}, fx_{b2} \in [u^-, u^+]$ in Theorem 3.4 is always true in a special case when the line $\{\mu A^{-1}b : \mu \in \mathbf{R}\}$ is in parallel to the straight lines $fx = u^-$ and $fx = u^+$. This is the case if $b_1 = 0$ in the special form of A, b, f in the proof of the theorem. So in this case, any closed-trajectory is the boundary of the domain of attraction. Thus, we can further conclude that there is a unique closed trajectory (and hence a unique limit cycle).

In the next section, we will show that if f is designed by the LQR method, then the line $\{\mu A^{-1}b : \mu \in \mathbf{R}\}$ is in parallel to the straight lines $fx = u^-$ and $fx = u^+$. Moreover, the domain of attraction $\mathcal{S}$ can be made arbitrarily close to the null controllable region $\mathcal{C}$ by simply increasing the feedback gain.

3.5. Semiglobal Stabilization on the Null Controllable Region

3.5.1. Second Order Anti-stable Systems

In this subsection, we continue to assume that $A \in \mathbf{R}^{2\times 2}$ is anti-stable and (A, b) is controllable. We will show that the domain of attraction $\mathcal{S}$ of the equilibrium $x = 0$ of the closed-loop system (3.19) can be made

arbitrarily close to the null controllable region $\mathcal{C}$ by judiciously choosing the feedback gain f. To state the main result of this section, we need to introduce the Hausdorff distance. Let $\mathcal{X}_1, \mathcal{X}_2$ be two bounded subsets of $\mathbf{R}^n$. Then their Hausdorff distance is defined as,

$$d(\mathcal{X}_1, \mathcal{X}_2) := \max\left\{\vec{d}(\mathcal{X}_1, \mathcal{X}_2), \vec{d}(\mathcal{X}_2, \mathcal{X}_1)\right\},$$

where

$$\vec{d}(\mathcal{X}_1, \mathcal{X}_2) = \sup_{x_1 \in \mathcal{X}_1} \inf_{x_2 \in \mathcal{X}_2} \|x_1 - x_2\|.$$

Here the vector norm used is arbitrary.

Let P be the unique positive definite solution to the following Riccati equation,

$$A'P + PA - Pbb'P = 0. \tag{3.28}$$

Note that this equation is associated with the minimum energy regulation, i.e., an LQR problem with cost

$$J = \int_0^\infty u'(t)u(t)dt.$$

The corresponding minimum energy state feedback gain is given by $f_0 = -b'P$. By the infinite gain margin and 50% gain reduction margin property of LQR regulators, the origin is a stable equilibrium of the system,

$$\dot{x}(t) = Ax(t) + b\,\mathrm{sat}_a(kf_0x(t)), \tag{3.29}$$

for all $k > 0.5$. Let $\mathcal{S}(k)$ be the domain of attraction of the equilibrium $x = 0$ of (3.29).

The following lemma is a simple generalization of the result of [3].

Lemma 3.2. Let $u_m = \min\{-u^-, u^+\}$. Define

$$\Omega_0 = \left\{x \in \mathbf{R}^2 : x'Px \le \frac{4u_m^2}{b'Pb}\right\}.$$

Then Ω_0 is in the domain of attraction for the system (3.29) for all $k > 0.5$.

Theorem 3.5. $\lim_{k\to\infty} d(\mathcal{S}(k), \mathcal{C}) = 0$.

Proof. For simplicity and without loss of generality, we assume that

$$A = \begin{bmatrix} 0 & -a_1 \\ 1 & a_2 \end{bmatrix}, \quad a_1, a_2 > 0, \quad b = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

Since A is anti-stable and (A, b) is controllable, A, b can always be transformed into this form. Suppose that A has already taken this form and $b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$. Let $V = \begin{bmatrix} -A^{-1}b & -b \end{bmatrix}$, then V is nonsingular and it can be verified that $V^{-1}AV = A$ and $V^{-1}b = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$. With this special form of A and b, we have,

$$P = 2\begin{bmatrix} \frac{a_2}{a_1} & 0 \\ 0 & a_2 \end{bmatrix}, \quad f_0 = \begin{bmatrix} 0 & 2a_2 \end{bmatrix},$$

and

$$A + kbf_0 = \begin{bmatrix} 0 & -a_1 \\ 1 & a_2(1-2k) \end{bmatrix}, \quad A^{-1}b = \begin{bmatrix} -1 \\ 0 \end{bmatrix}.$$

Hence, the line $\{\mu A^{-1}b : \mu \in \mathbf{R}\}$ is actually the line $x_2 = 0$ and it is between the two lines $kf_0x = u^+$ and $kf_0x = u^-$ (i.e., $x_2 = \frac{u^+}{2a_2k}$ and $x_2 = \frac{u^-}{2a_2k}$) for all $k > 0.5$. Therefore, the condition in Theorem 3.4 is satisfied for all $k > 0.5$ and the closed-loop system has a unique limit cycle which is the boundary of $\mathcal{S}(k)$. Also, by Lemma 3.2, the limit cycle always encloses the fixed ellipsoid Ω_0. To visualize the proof, $\partial\mathcal{C}, \Omega_0$ and $\partial\mathcal{S}(k)$ for some k, are plotted in Figure 4, where the inner closed curve is $\partial\mathcal{S}(k) = \Gamma$, and the outer dashed one is $\partial\mathcal{C}$.

For convenience, we proceed the proof with the time-reversed system of (3.29),

$$\dot{z}(t) = -Az(t) - b\,\mathrm{sat}_a(kf_0z(t)). \tag{3.30}$$

Observe that Γ is also the unique limit cycle of this system.

Recall from (3.15) and (3.16) that $\partial\mathcal{C}$ is formed by the trajectories of the system $\dot{z} = -Az - bv$: one from z_e^+(or z_s^+) to z_e^-(or z_s^-) under the control $v = u^-$ and the other from z_e^- (or z_s^-) to z_e^+ (or z_s^+) under the control $v = u^+$. On the other hand, when k is sufficiently large, the limit cycle must have two intersections with each of the lines $kf_0z = u^+$ and $kf_0z = u^-$. Suppose that the limit cycle trajectory starts at the righthand side intersection with $kf_0z = u^-$, goes clockwise and intersects the two lines successively at time t_1, t_2 and t_3 (see the points $z(0), z(t_1), z(t_2)$ and $z(t_3)$ in Figure 4). We also note that from $z(0)$ to $z(t_1)$, $v = \mathrm{sat}_a(kf_0z) = u^-$ for the closed-loop system (3.30) and from $z(t_2)$ to $z(t_3)$, $v = u^+$. By comparing the two closed trajectories Γ and $\partial\mathcal{C}$, we see that the proof can be completed by showing that as $k \to \infty$, $z(0), z(t_3) \to z_e^+$(or z_s^+), and $z(t_1), z(t_2) \to z_e^-$ (or z_s^-).

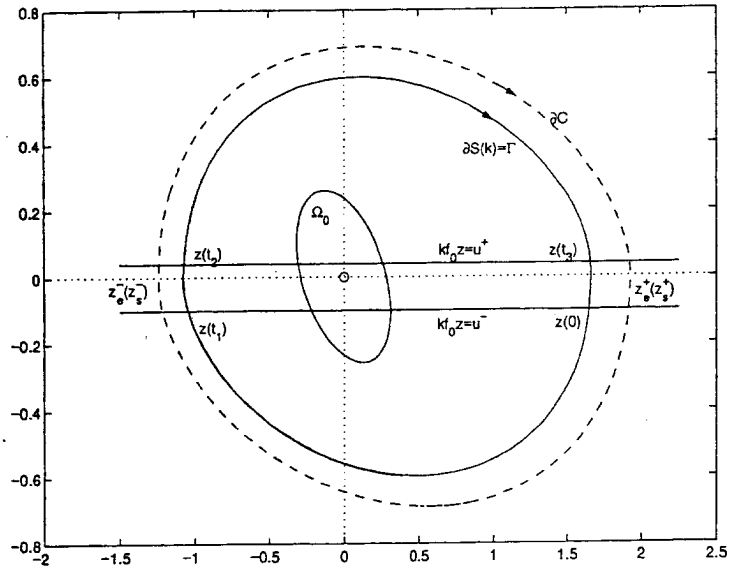


Figure 4: Illustration for the proof of Theorem 3.5.

Note that $kf_0z = 2ka_2z_2$, we can rewrite the closed-loop system (3.30) as,

$$\dot{z}_1 = a_1 z_2, \tag{3.31}$$

$$\dot{z}_2 = -z_1 - a_2 z_2 + \mathrm{sat}_a(2ka_2 z_2). \tag{3.32}$$

Since the trajectory goes clockwise and by (3.31), we have

$$\dot{z}_1(t_1) < 0, \quad \dot{z}_2(t_1) > 0, \quad \dot{z}_1(t_2) > 0, \quad \dot{z}_2(t_2) > 0. \tag{3.33}$$

With the particular form of A, b, we have $z_e^+ = -A^{-1}bu^+ = \begin{bmatrix} u^+ \\ 0 \end{bmatrix}$ and $z_e^- = -A^{-1}bu^- = \begin{bmatrix} u^- \\ 0 \end{bmatrix}$. Let $h = \max\{|z_2| : z \in \partial\Omega_0\}$. In the following proof, we will consider k such that $\left|\frac{u^+}{2ka_2}\right|, \left|\frac{u^-}{2ka_2}\right| < \frac{1}{2}h$. This means that the height of the part of Ω_0 above (below) the line $kf_0z = u^+$ ($kf_0z = u^-$) is greater than $\frac{1}{2}h$. Since $\mathcal{S}(k)$ is convex, since it must enclose Ω_0 and be inside $\mathcal{C}$, it follows from (3.33) that there exists a constant $\eta > 0$ such that the slope of Γ at $z(t_1)$ and $z(t_2)$ satisfy:

$$\frac{\dot{z}_2(t_1)}{\dot{z}_1(t_1)} < -\eta, \quad \frac{\dot{z}_2(t_2)}{\dot{z}_1(t_2)} > \eta.$$

Suppose that we draw a line that is tangent to Γ at $z(t_2)$; then by the convexity of $\mathcal{S}(k)$, $z(t_1)$ must be to the right of this line. This implies that

$$z_1(t_1) \geq z_1(t_2) - \frac{\dot{z}_1(t_2)}{\dot{z}_2(t_2)}(z_2(t_2) - z_2(t_1)),$$

i.e.,

$$\begin{aligned} z_1(t_2) &\leq z_1(t_1) + \frac{\dot{z}_1(t_2)}{\dot{z}_2(t_2)}(z_2(t_2) - z_2(t_1)) \\ &\leq z_1(t_1) + \frac{u^+ - u^-}{2\eta k a_2}, \end{aligned} \tag{3.34}$$

and similarly, $z(t_2)$ is to the right of the line tangent to Γ at $z(t_1)$,

$$\begin{aligned} z_1(t_2) &\geq z_1(t_1) + \frac{\dot{z}_1(t_1)}{\dot{z}_2(t_1)}(z_2(t_2) - z_2(t_1)) \\ &\geq z_1(t_1) - \frac{u^+ - u^-}{2\eta k a_2}. \end{aligned} \tag{3.35}$$

It follows from (3.34) and (3.35) that $\lim_{k\to\infty}(z_1(t_1) - z_1(t_2)) = 0$ and similarly $\lim_{k\to\infty}(z_1(t_3) - z_1(0)) = 0$. Since $\lim_{k\to\infty} z_2(0) = \lim_{k\to\infty} z_2(t_1) = \lim_{k\to\infty} z_2(t_2) = \lim_{k\to\infty} z_2(t_3) = 0$, these imply that

$$\lim_{k\to\infty}(z(t_1) - z(t_2)) = 0, \quad \lim_{k\to\infty}(z(t_3) - z(0)) = 0. \tag{3.36}$$

From (3.33), we also have

$$\dot{z}_2(t_1) = -z_1(t_1) - a_2 z_2(t_1) + u^- = -z_1(t_1) - \frac{u^-}{2k} + u^- > 0.$$

It follows that

$$z_1(t_1) < u^- - \frac{u^-}{2k}. \tag{3.37}$$

Now we break the proof into two cases.

Case 1. A has two real eigenvalues.

In this case, $z_e^+ = \begin{bmatrix} u^+ \\ 0 \end{bmatrix}$ and $z_e^- = \begin{bmatrix} u^- \\ 0 \end{bmatrix}$ are on the boundary of $\mathcal{C}$. For the particular structure of A and b, it can be verified that every point in $\mathcal{C}$ is to the right of z_e^- and to the left of z_e^+. Since $z(t_1)$ must be in $\mathcal{C}$, we have $z_1(t_1) > u^-$. It follows from (3.37) that $\lim_{k\to\infty} z_1(t_1) = u^-$. With (3.36), we finally have $\lim_{k\to\infty} z(t_1) = \lim_{k\to\infty} z(t_2) = z_e^-$ and similarly, $\lim_{k\to\infty} z(0) = \lim_{k\to\infty} z(t_3) = z_e^+$.

Case 2. A has a pair of complex eigenvalues $\alpha \pm j\beta$.

Denote $T_p = \frac{\pi}{\beta}$, then $e^{-AT_p} = -e^{-\alpha T_p} I$. First, we claim that as $k \to \infty$, $t_1 \to T_p$. To prove this claim, we recall some simple facts about a second-order linear system with a pair of complex eigenvalues,

$$\dot{v} = -Av. \tag{3.38}$$

For this system, suppose that $v(0) \neq 0$, then $\angle v(t)$ is monotonically increasing (or decreasing). Consider $v(t_1) = e^{-At_1} v(0)$. If the trajectory $\{e^{-At} v(0) : t \in [0, t_1]\}$ can be separated from the origin with a straight line, then $t_1 < T_p$. Now suppose $0 < t_1 \leq T_p$. If $v(t_1)$ and $v(0)$ are aligned, then we must have $t_1 = T_p$; If $v(t_1)$ and $v(0)$ tend to be aligned, then t_1 will approach T_p.

From $z(0)$ to $z(t_1)$, $\dot{z} = -Az - bu^-$. If we let $v = z - z_e^-$, then the part of Γ from $z(0)$ to $z(t_1)$ is a trajectory of (3.38). From Lemma 3.1 we know that z_e^- does not belong to the half plane $kf_0 z \leq u^-$, so this part of trajectory is below z_e^- (the origin in the v coordinate). Hence we must have $0 < t_1 < T_p$. Since $z(0)$ must be the right of Ω_0, we have $z_1(0) > 0$ for sufficiently large k. It follows that $\|v(0)\| = \left\| \begin{bmatrix} z_1(0) - u^- \\ z_2(0) \end{bmatrix} \right\|$ is greater than a constant. So $\|v(t_1)\|$ is also greater than a constant. Note that as $k \to \infty$, $v_2(0), v_2(t_1) \to 0$. Therefore, $v(0)$ and $v(t_1)$ tend to be aligned, so we get $\lim_{k\to\infty} t_1 = T_p$. Similarly, $\lim_{k\to\infty}(t_3 - t_2) = T_p$.

Now we have

$$\lim_{k\to\infty} e^{-At_1} = \lim_{k\to\infty} e^{-A(t_3 - t_2)} = e^{-AT_p} = -e^{-\alpha T_p} I.$$

It follows that, as $k \to \infty$,

$$\begin{aligned} &\left(z(t_1) - z_e^-\right) + e^{-\alpha T_p}\left(z(0) - z_e^-\right) \\ &= \left(e^{-At_1} + e^{-\alpha T_p} I\right)\left(z(0) - z_e^-\right) \to 0, \end{aligned} \tag{3.39}$$

and

$$\begin{aligned} &\left(z(t_3) - z_e^+\right) + e^{-\alpha T_p}\left(z(t_2) - z_e^+\right) \\ &= \left(e^{-A(t_3 - t_2)} + e^{-\alpha T_p} I\right)\left(z(t_2) - z_e^+\right) \to 0. \end{aligned} \tag{3.40}$$

Recall from (3.36) that

$$z(t_1) - z(t_2) \to 0, \quad z(0) - z(t_3) \to 0,$$

the property in (3.40) implies that

$$\left(z(0) - z_e^+\right) + e^{-\alpha T_p}\left(z(t_1) - z_e^+\right) \to 0. \tag{3.41}$$

Let

$$\xi_1 = \left(z(t_1) - z_e^-\right) + e^{-\alpha T_p}\left(z(0) - z_e^-\right),$$
$$\xi_2 = \left(z(0) - z_e^+\right) + e^{-\alpha T_p}\left(z(t_1) - z_e^+\right).$$

It can be solved that

$$z(t_1) = \left(1 - e^{-\alpha T_p}\right)^{-1}\left(z_e^- - e^{-\alpha T_p} z_e^+\right) + O(\xi_1) + O(\xi_2).$$

From (3.39) and (3.41), we know that $\lim_{k\to\infty}\xi_1 = \lim_{k\to\infty}\xi_2 = 0$, so we have

$$\lim_{k\to\infty} z(t_1) = -\left(1 - e^{-\alpha T_p}\right)^{-1}\left(u^- - e^{-\alpha T_p} u^+\right) A^{-1} b = z_s^-.$$

Similarly, we have $\lim_{k\to\infty} z(0) = z_s^+$. It follows from (3.36) that

$$\lim_{k\to\infty} z(t_2) = z_s^-, \quad \lim_{k\to\infty} z(t_3) = z_s^+.$$

□

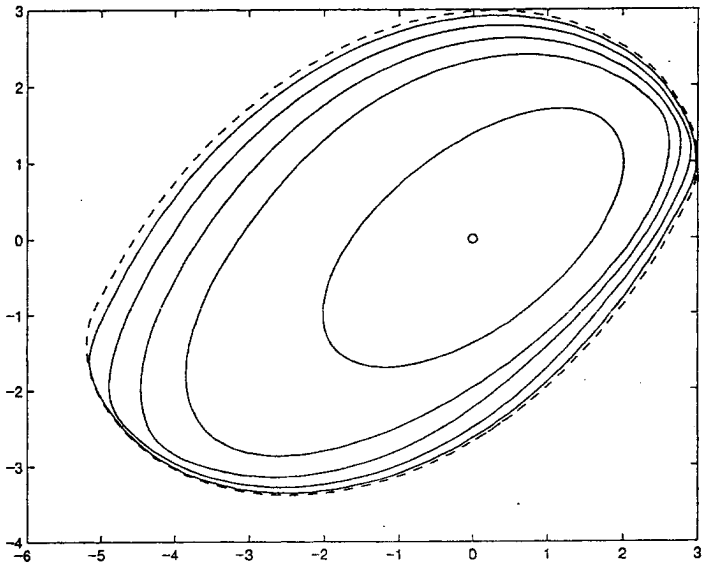


Figure 5: The domains of attraction under different feedback gains.

Example 3.1. Consider the open-loop system (3.1) with

$$A = \begin{bmatrix} 0.6 & -0.8 \\ 0.8 & 0.6 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 4 \end{bmatrix},$$

$u^- = -0.5$ and $u^+ = 1$. Then we have

$$f_0 = \begin{bmatrix} 0.12 & -0.66 \end{bmatrix}.$$

In Figure 5, the boundaries of the domains of attraction corresponding to different $f = kf_0$, $k = 0.50005, 0.6, 0.7, 1, 2$, are plotted from the inner to the outer. It is clear from the figure that the domain of attraction becomes larger as k is increased. The outermost dashed closed curve is $\partial\mathcal{C}$.

3.5.2. Higher Order Systems with Two Exponentially Unstable Poles

Consider the following open-loop system

$$\dot{x}(t) = Ax(t) + bu(t) = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} x(t) + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} u(t), \tag{3.42}$$

where $x = \begin{bmatrix} x_1' & x_2' \end{bmatrix}'$, $x_1 \in \mathbf{R}^2, x_2 \in \mathbf{R}^n$, $A_1 \in \mathbf{R}^{2\times 2}$ is anti-stable and $A_2 \in \mathbf{R}^n$ is semistable. Assume that (A, b) is controllable. Denote the null controllable region of the subsystem

$$\dot{x}_1(t) = A_1 x_1(t) + b_1 u(t)$$

as $\mathcal{C}_1$, then the null controllable region of (3.42) is $\mathcal{C}_1 \times \mathbf{R}^n$. Given $\gamma_1, \gamma_2 > 0$, denote

$$\Omega_1(\gamma_1) := \left\{ \gamma_1 x_1 \in \mathbf{R}^2 : \ x_1 \in \overline{\mathcal{C}}_1 \right\},$$

and

$$\Omega_2(\gamma_2) := \left\{ x_2 \in \mathbf{R}^n : \ \|x_2\| \le \gamma_2 \right\}.$$

When $\gamma_1 = 1$, $\Omega_1(\gamma_1) = \overline{\mathcal{C}}_1$ and when $\gamma_1 < 1$, $\Omega_1(\gamma_1)$ lies in the interior of $\mathcal{C}_1$. In this section, we will show that given any $\gamma_1 < 1$ and $\gamma_2 > 0$, a state feedback can be designed such that $\Omega_1(\gamma_1) \times \Omega_2(\gamma_2)$ is contained in the domain of attraction of the equilibrium $x = 0$ of the closed-loop system.

For $\epsilon > 0$, let $P(\epsilon) = \begin{bmatrix} P_1(\epsilon) & P_2(\epsilon) \\ P_2'(\epsilon) & P_3(\epsilon) \end{bmatrix} \in \mathbf{R}^{(2+n)\times(2+n)}$ be the unique positive definite solution to the ARE

$$A'P + PA - Pbb'P + \epsilon^2 I = 0. \tag{3.43}$$

Clearly, as $\epsilon \downarrow 0$, $P(\epsilon)$ decreases. Hence $\lim_{\epsilon\to 0} P(\epsilon)$ exists.

Let P_1 be the unique positive definite solution to the ARE

$$A_1'P_1 + P_1A_1 - P_1b_1b_1'P_1 = 0.$$

Then by the continuity property of the solution of the Riccati equation [15],

$$\lim_{\epsilon \to 0} P(\epsilon) = \begin{bmatrix} P_1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Let $f(\epsilon) := -b'P(\epsilon)$. First, consider the domain of attraction of the equilibrium $x = 0$ of the following closed-loop system

$$\dot{x}(t) = Ax(t) + b\,\mathrm{sat}_a(f(\epsilon)x(t)). \tag{3.44}$$

Let $u_m = \min\{-u^-, u^+\}$. It is easy to see that

$$D(\epsilon) := \left\{ x \in \mathbf{R}^{2+n} : \; x'P(\epsilon)x \le 4u_m / \|b'P^{\frac{1}{2}}(\epsilon)\|^2 \right\}$$

is contained in the domain of attraction of the equilibrium $x = 0$ of (3.44) and is an invariant set. Note that if $x_0 \in D(\epsilon)$, then $x(t) \in D(\epsilon)$ and $|f(\epsilon)x(t)| \le u_m$ for all $t > 0$. That is, $x(t)$ will stay in the linear region of the closed-loop system, and in $D(\epsilon)$.

Theorem 3.6. Let $f_0 = -b_1'P_1$. For any $\gamma_1 < 1$ and $\gamma_2 > 0$, there exist $k > 0.5$ and $\epsilon > 0$ such that $\Omega_1(\gamma_1) \times \Omega_2(\gamma_2)$ is contained in the domain of attraction of the equilibrium $x = 0$ of the closed-loop system

$$\dot{x}(t) = Ax(t) + bu(t), \;\; u(t) = \begin{cases} \mathrm{sat}_a(kf_0x_1(t)), & x \notin D(\epsilon), \\ \mathrm{sat}_a(f(\epsilon)x(t)), & x \in D(\epsilon). \end{cases} \tag{3.45}$$

Proof. Similar to Theorem 4.4.1 in [4] and Theorem 4.2 in [5]. □

3.6. Conclusions

In this chapter we have studied the problem of controlling a linear system subject to asymmetric actuator saturation. The null controllable region of such a system is first characterized. Simple feedback laws are constructed to stabilize a system with no more than two exponentially unstable open-loop poles. The feedback law guarantees a domain of attraction that includes any given compact set inside the null controllable region.

References

[1] J. Alvarez, R. Suarez and J. Alvarez. Planar Linear Systems with Single Saturated Feedback, *Systems & Control Letters*, 20 (1993) 319–326.

[2] O. Hájek. *Control Theory in the Plane*, Springer-Verlag, (1991).

[3] P. -O. Gutman and P. Hagander. A New Design of Constrained Controllers for Linear Systems, *IEEE Trans. Automat. Contr.*, 30 (1985) 22–33.

[4] T. Hu and Z. Lin. *Control Systems with Actuator Saturation: Analysis and Design*, Birkhäuser, Boston, (2001).

[5] T. Hu, Z. Lin and L. Qiu. Stabilization of Exponentially Unstable Linear Systems with Saturating Actuators, *IEEE Transaction on Automatic Control*, to appear.

[6] H. K. Khalil. *Nonlinear Systsems*, MacMillan, New York, (1992).

[7] Z. Lin and A. Saberi. Semiglobal Exponential Stabilization of Linear Systems Subject to 'Input Saturation' via Linear Feedbacks, *Systems and Control Letters*, 21 (1993) 225–239.

[8] J. Macki and M. Strauss. *Introduction to Optimal Control*, Springer-Verlag, (1982).

[9] A. Saberi, Z. Lin and A. R. Teel. Control of Linear Systems with Saturating Actuators, *IEEE Trans. Automat. Contr.*, 41 (1996) 368–378.

[10] W. E. Schmitendorf and B. R. Barmish. Null Controllability of Linear Systems with Constrained Controls, *SIAM J. Control and Optimization*, 18 (1980) 327–345.

[11] E. D. Sontag. An Algebraic Approach to Bounded Controllability of Linear Systems, *Int. J. Control*, 39 (1984) 181–188.

[12] E. D. Sontag and H. J. Sussmann. Nonlinear Output Feedback Design for Linear Systems with Saturating Controls, in: *Proc. 29th IEEE Conf. on Dec. and Control*, (1990) 3414–3416.

[13] H. J. Sussmann, E. D. Sontag, and Y. Yang. A General Result on the Stabilization of Linear Systems Using Bounded Controls, *IEEE Trans. Automat. Contr.*, 39 (1994) 2411–2425.

[14] A. R. Teel. Global Stabilization and Restricted Tracking for Multiple Integrators with Bounded Controls, *System and Control Letters*, 18 (1992) 165–171.

[15] J. C. Willems. Least Squares Stationary Optimal Control and the Algebraic Riccati Equations, *IEEE Trans. Automat. Contr.*, 16 (1971) 621–634.

Publication 19

INTERNATIONAL JOURNAL OF ROBUST AND NONLINEAR CONTROL
Int. J. Robust Nonlinear Control 2002; (in press) (DOI: 10.1002/rnc.647)

Output regulation of general discrete-time linear systems with saturation nonlinearities

Tingshu Hu*,† and Zongli Lin

Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22903, USA

SUMMARY

This paper studies the classical problem of output regulation for linear discrete-time systems subject to actuator saturation and extends the recent results on continuous-time systems to discrete-time systems. The asymptotically regulatable region, the set of all initial conditions of the plant and the exosystem for which the asymptotic output regulation is possible, is characterized in terms of the null controllable region of the anti-stable subsystem of the plant. Feedback laws are constructed that achieve regulation on the asymptotically regulatable region. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: saturation nonlinearities; output regulation; asymptotically regulatable region; discrete-time systems

1. INTRODUCTION

There has been considerable research on the problem of stabilization and output regulation of linear systems subject to actuator saturation. The problem of stabilization involves issues ranging from the characterization of the null controllable region (or, asymptotically null controllable region), the set of all initial conditions that can be driven to the origin by the saturating actuators in some finite time (respectively, asymptotically), to the construction of feedback laws that achieve stabilization on the entire or a large portion of the asymptotically null controllable region. Recent years have witnessed extensive research that addresses these issues. In particular, for an open loop system that are stabilizable and have all its poles in the closed left-half plane, it was established in Reference [1] that the asymptotically null controllable region is the entire state space. For this reason, a linear system that is stabilizable in the usual

*Correspondence to: Tingshu Hu, Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22903, U.S.A.
†E-mail: th7f@virginia.edu

Contract/grant sponsor: US Office of Naval Research Young Investigator Program; contract/grant number: N00014-99-1-0670

Copyright © 2002 John Wiley & Sons, Ltd.

linear sense and has all its poles in the closed left-half plane is referred to as asymptotically null controllable with bounded controls, or ANCBC. For ANCBC systems subject to actuator saturation, various feedback laws that achieve global or semi-global stabilization on the asymptotically null controllable region have been constructed (see, for example, References [2–5]). For exponentially unstable open-loop systems subject to actuator saturation, the asymptotically null controllable regions were recently characterized and feedback laws were constructed that achieve semi-global stabilization on the asymptotically null controllable region (see References [6–8]).

In comparison with the problem of stabilization, the problem of output regulation for linear systems subject to actuator saturation, however, has received relatively less attention. The few works that motivated our recent research [9] on continuous-time systems were References [10,5,11,2]. In References [5,2], the problem of output regulation was studied for ANCBC systems subject to actuator saturation. Necessary and sufficient conditions on the plant/exosystem and their initial conditions were derived under which output regulation can be achieved. Under these conditions, feedback laws that achieve output regulation were constructed based on the semi-global stabilizing feedback laws of Reference [4]. The recent work [10] made an attempt to address the problem of output regulation for exponentially unstable linear systems subject to actuator saturation. The attempt was to enlarge the set of initial conditions of the plant and the exosystem under which output regulation can be achieved. In particular, for plants with only one positive pole and exosystems that contain only one frequency component, feedback laws were constructed that achieve output regulation on what was later characterized in Reference [9] as the asymptotically regulatable region.

In Reference [9], we systematically studied the problem of output regulation for general continuous-time linear systems subject to actuator saturation. In particular, we characterized the regulatable region, the set of plant and exosystem initial conditions for which output regulation is possible with the saturating actuators. We then constructed feedback laws that achieve regulation on the regulatable region.

The objective of this paper is to extend the above results to discrete-time systems. In Section 2, we formulate the problem of output regulation for linear systems with saturating actuators. Section 3 characterizes the regulatable region. Sections 4 and 5, respectively, construct state feedback and error feedback laws that achieve output regulation on the regulatable region. Finally, Section 6 gives a brief concluding remark to our current work.

Throughout the paper, we will use standard notation. For a vector $u \in \mathbf{R}^m$, we use $|u|_\infty$ and $|u|_2$ to denote the vector ∞-norm and the 2-norm. For a vector sequence $u(k) \in \mathbf{R}^m$, $k = 0, 1, 2, \ldots$; we define $\|u\|_\infty = \sup_{k \geqslant 0} |u(k)|_\infty$. We use $\mathrm{sat}(\cdot)$ to denote the standard saturation function $\mathrm{sat}(s) = \mathrm{sgn}(s) \min\{1, |s|\}$. With a slight abuse of notation and for simplicity, for a vector $u \in \mathbf{R}^m$, we also use the same $\mathrm{sat}(u)$ to denote the vector saturation function, i.e. $\mathrm{sat}(u) = [\mathrm{sat}(u_1)\ \mathrm{sat}(u_2)\ \cdots\ \mathrm{sat}(u_m)]^{\mathrm{T}}$.

2. PRELIMINARIES AND PROBLEM STATEMENT

In this section, we state the discrete-time version of the classical results on the problem of output regulation for continuous-time linear systems [12] (see also Reference [13]). These results will

Copyright © 2002 John Wiley & Sons, Ltd.

motivate our formulation of as well as the solution to the problem of output regulation for discrete-time linear systems subject to actuator saturation.

2.1. *Review of output regulation for linear systems*

Consider a linear system

$$x(k+1) = Ax(k) + Bu(k) + Pw(k)$$

$$w(k+1) = Sw(k)$$

$$e(k) = Cx(k) + Qw(k) \tag{1}$$

The first equation of this system describes a plant, with state $x \in \mathbf{R}^n$ and input $u \in \mathbf{R}^m$, subject to the effect of a disturbance represented by Pw. The third equation defines the error $e \in \mathbf{R}^q$ between the actual plant output Cx and a reference signal $-Qw$ that the plant output is required to track. The second equation describes an autonomous system, often called the exosystem, with state $w \in \mathbf{R}^r$. The exosystem models the class to disturbances and references taken into consideration.

The control action to the plant, u, can be provided either by state feedback or by error feedback. The objective is to achieve internal stability and output regulation. Internal stability means that if we disconnect the exosystem and set w equal to zero then the closed-loop system is asymptotically stable. Output regulation means that for any initial conditions of the closed-loop system, we have that $e(k) \to 0$ as $k \to \infty$.

The solution to these problems is based on the following three assumptions.

A1. The eigenvalues of S are on or outside of the unit circle;
A2. The pair $(A, B)$is stabilizable;
A3. The pair

$$\left(\begin{bmatrix} C & Q \end{bmatrix}, \begin{bmatrix} A & P \\ 0 & S \end{bmatrix} \right)$$

is detectable.

For continuous-time systems, complete solutions to the output regulation problems were established in Reference [12] by Francis. These solutions can be adapted for discrete-time systems as follows:

Proposition 1
Suppose Assumptions A1 and A2 hold. Then, the problem of output regulation by state feedback is solvable if and only if there exist matrices Π that Γ that solve the linear matrix equations

$$\Pi S = A\Pi + B\Gamma + P$$

$$C\Pi + Q = 0 \tag{2}$$

Moreover, if in addition Assumption A3 also holds, the solvability of the above linear matrix equations is also a necessary and sufficient condition for the solvability of the problem of output regulation by error feedback.

Copyright © 2002 John Wiley & Sons, Ltd.

2.2. Output regulation for linear systems subject to actuator saturation

Motivated by the classical formulation of output regulation for linear systems, we consider the following plant and the exosystem:

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) + Pw(k) \\ w(k+1) &= Sw(k) \\ e(k) &= Cx(k) + Qw(k) \end{aligned} \tag{3}$$

where u is the output of saturating actuators and is constrained by $\|u\|_\infty \leqslant 1$. A control u that satisfies this constraint is referred to as an *admissible control.* Because of the bound on the control input, both the plant and the exosystem cannot operate in the entire state space. For this reason, we assume that $(x_0, w_0) \in \mathscr{Y}_0$ for some $\mathscr{Y}_0 \subset \mathbf{R}^n \times \mathbf{R}^r$. Let

$$\mathscr{X}_0 = \{x_0 \in \mathbf{R}^n : (x_0, 0) \in \mathscr{Y}_0\}$$

The problem to be addressed in this paper is the following.

Problem 1
The problem of output regulation by state feedback for the system (3) is to find, if possible, a state feedback law $u = \phi(x, w)$, with $|\phi(x, w)|_\infty \leqslant 1$ and $\phi(0,0) = 0$, such that

1. the equilibrium $x = 0$ of the system

$$x(k+1) = Ax(k) + B\phi(x(k), 0)$$

   is asymptotically stable with $\mathscr{X}_0$ contained in its domain of attraction;
2. for all $(x_0, w_0) \in \mathscr{Y}_0$, the interconnection of (3) and the feedback law $u = \phi(x, w)$ results in bounded state trajectories $x(k)$an d $\lim_{k\to\infty} e(k) = 0$.

If only the error e is available, the state (x, w) can be reconstructed after a finite number of steps if we further assume that the pair in A3 is observable. But the initial condition (x_0, w_0)mi ght have to be constrained in a subset of $\mathscr{Y}_0$.

Our objective is to characterize the maximal set of initial conditions (x_0, w_0), the largest possible $\mathscr{Y}_0$, on which the above problem is solvable and to explicitly construct feedback law that actually solves the problem for $\mathscr{Y}_0$ as large as possible.

We will assume that (A, B) is stabilizable. We will also assume that S is neutrally stable and all its eigenvalues are on the unit circle. The stabilizability of (A, B) is clearly necessary for the stabilization of the plant. The assumption on $S$is without loss of generality. Since the components corresponding to the asymptotically stable modes of the exosystem will tend to zero, they will not affect the regulation of the output. On the other hand, if the exosystem has unstable modes, either the disturbance Pw or the signal Qw will go unbounded. It is generally impossible to drive the error e to zero asymptotically with a bounded control (see Reference [11]).

Copyright © 2002 John Wiley & Sons, Ltd.

3. THE REGULATABLE REGION

In this section, we will characterize the set of all initial states of the plant and the exosystem on which the problem of output regulation is solvable under the restriction that $\|u\|_\infty \leqslant 1$. We will refer to this set as the asymptotically regulatable region.

To begin with, we observe from the classical output regulation theory (see Section 2.1) that for this problem to be solvable, there must exist matrices $\Pi \in \mathbf{R}^{n\times r}$ and $\Gamma \in \mathbf{R}^{m\times r}$ that solve the matrix equations (2). Given the matrices Π and Γ, we define a new state $z = x - \Pi w$ and rewrite the system equations as

$$\begin{aligned} z(k+1) &= Az(k) + Bu(k) - B\Gamma w(k) \\ w(k+1) &= Sw(k) \\ e(k) &= Cz(k) \end{aligned} \tag{4}$$

From these new equations, it is clear that $e(k)$ goes to zero asymptotically if $z(k)$ goes to zero asymptotically. The latter is possible only if (see Reference [11] for the continuous-time case)

$$\sup_{k\geqslant 0} |\Gamma S^k w_0|_\infty < 1 \tag{5}$$

For this reason, we will restrict our attention to exosystem initial conditions in the following compact set

$$\mathscr{W}_0 = \{w_0 \in \mathbf{R}^r : |\Gamma w(k)|_\infty = |\Gamma S^k w_0|_\infty \leqslant \rho,\ \forall k \geqslant 0\} \tag{6}$$

for some $\rho \in [0, 1)$. For later use, we also denote $\delta = 1 - \rho$. We note that the compactness of $\mathscr{W}_0$ can be guaranteed by the observability of (Γ, S). Indeed, if (Γ, S) is not observable, then the exosystem can be reduced to make it so.

We can now precisely define the notion of asymptotically regulatable region as follows.

Definition 1

1. Given $K > 0$, a pair $(z_0, w_0) \in \mathbf{R}^n \times \mathscr{W}_0$ is regulatable in K steps if there exists an admissible control u, such that the response of (4) satisfies $z(K) = 0$. The set of all (z_0, w_0) regulatable in K steps is denoted as $\mathscr{R}_g(K)$.
2. A pair (z_0, w_0) is regulatable if $(z_0, w_0) \in \mathscr{R}_g(K)$ for some $K < \infty$. The set of all regulatable $(z_0, w_0)$is referred to as the regulatable region and is denoted as $\mathscr{R}_g$.
3. The set of all (z_0, w_0) for which there exist admissible controls such that the response of (4) satisfies $\lim_{k\to\infty} z(k) = 0$ is referred to as the asymptotically regulatable region and is denoted as $\mathscr{R}_g^a$.

Remark 1

Note that the regulatable region is defined in terms of $\lim_{k\to\infty} z(k) = 0$ rather than $\lim_{k\to\infty} e(k) = 0$. Requiring the former instead of the latter will also guarantee the closed-loop stability in the absence of w. Like the continuous-time case [6,9], this will result in essentially the same description of the regulatable region.

Copyright © 2002 John Wiley & Sons, Ltd.

We will describe $\mathscr{R}_g(K)$, $\mathscr{R}_g$ and $\mathscr{R}_g^a$ in terms of the asymptotically null controllable region of the plant

$$v(k+1) = Av(k) + Bu(k), \quad \|u\|_\infty \leqslant 1$$

Definition 2
The null controllable region at step K, denoted as $\mathscr{C}(K)$, is the set of $v_0 \in \mathbf{R}^n$th at can be driven to the origin in K steps and the null controllable region, denoted as $\mathscr{C}$, is the set of $v_0 \in \mathbf{R}^n$th at can be driven to the origin in finite number of steps by admissible controls. The asymptotically null controllable region, denoted as $\mathscr{C}^a$, is the set of all v_0 that can be driven to the origin asymptotically by admissible controls.

Clearly,

$$\mathscr{C} = \bigcup_{K\in[0,\infty)} \mathscr{C}(K)$$

and

$$\mathscr{C}(K) = \left\{ \sum_{i=0}^{K-1} A^{-i-1}Bu(i) : \|u\|_\infty \leqslant 1 \right\} \tag{7}$$

Some simple methods to describe $\mathscr{C}$ and $\mathscr{C}^a$ were developed in Reference [8] (see also Reference [6]).

To simplify the characterization of $\mathscr{R}_g$ and $\mathscr{R}_g^a$ and without loss of generality, let us assume that

$$z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \quad z_1 \in \mathbf{R}^{n_1}, \quad z_2 \in \mathbf{R}^{n_2}$$

and

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \tag{8}$$

where $A_1 \in \mathbf{R}^{n_1 \times n_1}$ is semi-stable (i.e. all its eigenvalues are on or inside the unit circle) and $A_2 \in \mathbf{R}^{n_2 \times n_2}$ is anti-stable (i.e. all its eigenvalues are outside of the unit circle). The anti-stable subsystem

$$\begin{aligned} z_2(k+1) &= A_2 z_2(k) + B_2 u(k) - B_2 \Gamma w(k) \\ w(k+1) &= Sw(k) \end{aligned} \tag{9}$$

is of crucial importance. Denote its regulatable regions as $\mathscr{R}_{g_2}(K)$ and $\mathscr{R}_{g_2}$, and the null controllable regions for the system

$$v_2(k+1) = A_2 v_2(k) + B_2 u(k)$$

as $\mathscr{C}_2(K)$ and $\mathscr{C}_2$. Then, the asymptotically null controllable region of the system

$$v(k+1) = Av(k) + Bu(k)$$

Copyright © 2002 John Wiley & Sons, Ltd.

is given by $\mathscr{C}^a = \mathbf{R}^{n_1} \times \mathscr{C}_2$ [14], where $\mathscr{C}_2$ is a bounded convex open set. Denote the closure of $\mathscr{C}_2$ as $\bar{\mathscr{C}}_2$, then

$$\bar{\mathscr{C}}_2 = \left\{ \sum_{i=0}^{\infty} A_2^{-i-1} B_2 u(i) : \|u\|_\infty \leqslant 1 \right\}$$

Theorem 1
Let $V_2 \in \mathbf{R}^{n_2 \times r}$ be the unique solution to the matrix equation

$$-A_2 V_2 + V_2 S = -B_2 \Gamma \tag{10}$$

and let

$$V(K) = V_2 - A^{-K} V_2 S^K$$

Then,

(a) $\mathscr{R}_{g_2}(K) = \{(z_2, w) \in \mathbf{R}^{n_2} \times \mathscr{W}_0 : z_2 - V(K)w \in \mathscr{C}_2(K)\}$ (11)

(b) $\mathscr{R}_{g_2} = \{(z_2, w) \in \mathbf{R}^{n_2} \times \mathscr{W}_0 : z_2 - V_2 w \in \mathscr{C}_2\}$ (12)

(c) $\mathscr{R}_g^a = R^{n_1} \times \mathscr{R}_{g_2}$ (13)

Proof
(a) Given $(z_{20}, w_0) \in \mathbf{R}^{n_2} \times \mathscr{W}_0$ and an admissible control u, the solution of (9) at $k = K$ is

$$z_2(K) = A^K \left(z_{20} + \sum_{i=0}^{K-1} A_2^{-i-1} B_2 u(i) - \sum_{i=0}^{K-1} A_2^{-i-1} B_2 \Gamma S^i w_0 \right) \tag{14}$$

Applying (10), we have

$$\begin{aligned} -\sum_{i=0}^{K-1} A_2^{-i-1} B_2 \Gamma S^i &= \sum_{i=0}^{K-1} A_2^{-i-1} (-A_2 V_2 + V_2 S) S^i \\ &= \sum_{i=0}^{K-1} (-A_2^{-i} V_2 S^i + A_2^{-i-1} V_2 S^{i+1}) \\ &= -V_2 + A_2^{-K} V_2 S^K \\ &= -V(K) \end{aligned} \tag{15}$$

where the third "=" is simply obtained by expanding the terms in the summation and cancelling all the middle terms. Thus,

$$A^{-K} z_2(K) = z_{20} - V(K) w_0 + \sum_{i=0}^{K-1} A^{-i-1} B_2 u(i)$$

By setting $z_2(K) = 0$, we immediately obtain (a) from the definition of $\mathscr{R}_g(K)$an d (7).
(b) and (c). The proof is lengthy and can be found in Reference [6]. □

Remark 2
Given (z_0, w_0), there exists an admissible control u such that $\lim_{k \to \infty} z(k) = 0$ if and only if $(z_0, w_0) \in \mathscr{R}_g^a$. Recalling that $z = x - \Pi w$, we observe that, for a given pair of initial states in the

Copyright © 2002 John Wiley & Sons, Ltd.

original co-ordinates, (x_0, w_0), there is an admissible control u such that $\lim_{k\to\infty}(x(k) - \Pi w(k)) = 0$ if and only if $x_{20} - (\Pi_2 + V_2)w_0 \in \mathscr{C}_2$, where $\Pi_2 = [0\ I_{n_2}]\Pi$.

4. STATE FEEDBACK CONTROLLER

In this section, we will construct a feedback law that solves the problem of output regulation by state feedback for linear systems subject to actuator saturation. Our feedback law will be based on a stabilizing feedback law $u = f(v)$, $|f(v)|_\infty \leqslant 1$ for all $v \in \mathbf{R}^n$, which makes the system

$$v(k+1) = Av(k) + Bf(v(k)) \tag{16}$$

have an asymptotically stable equilibrium at the origin. Actually, any feedback of the form $u = f(v) = \text{sat}(Fu)$ will stabilize the system locally at the origin as long as $A + BF$ is asymptotically stable. In References [6–8], we presented some methods for designing $f(v)$ to enlarge the domain of attraction of the origin. Here, we assume that a stabilizing feedback law $u = f(v)$ has been designed and the equilibrium $v = 0$ of the closed-loop system (16) has a domain of attraction $\mathscr{S} \subset \mathscr{C}^a$.

Now consider the system (4). Given a state feedback $u = g(z, w), |g(z, w)|_\infty \leqslant 1$ for all $(z, w) \in \mathbf{R}^n \times \mathscr{W}_0$, we have the closed-loop system

$$\begin{aligned} z(k+1) &= Az(k) + Bg(z(k), w(k)) - B\Gamma w(k) \\ w(k+1) &= Sw(k) \end{aligned} \tag{17}$$

Denote the time response of $z(k)$ to the initial state (z_0, w_0) as $z(k, z_0, w_0)$ and define

$$\mathscr{S}_{zw} := \left\{(z_0, w_0) \in \mathbf{R}^n \times \mathscr{W}_0 : \lim_{k\to\infty} z(k, z_0, w_0) = 0\right\}$$

Since $\mathscr{R}_g^a$ is the set of all (z_0, w_0)fo r which $z(k)$can be driven to the origin asymptotically, we must have $\mathscr{S}_{zw} \subset \mathscr{R}_g^a$. Our objective is to design a control law $u = g(z, w)$ such that $\mathscr{S}_{zw}$ is as large as possible, or as close to $\mathscr{R}_g^a$ as possible.

First we need a mild assumption which can be removed by modifying the controller (see Reference [6]). Assume that there exists a matrix $V \in \mathbf{R}^{n\times r}$ such that

$$-AV + VS = -B\Gamma \tag{18}$$

This will be the case if A and S have no common eigenvalues (see e.g. p. 26 of Reference [16]). With the decomposition in (8), if we partition V accordingly as

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

then V_2 satisfies $-A_2V_2 + V_2S = -B_2\Gamma$. Denote

$$D_{zw} := \{(z, w) \in \mathbf{R}^n \times \mathscr{W}_0 : z - Vw \in \mathscr{S}\} \tag{19}$$

on which the following observation can be made.

Copyright © 2002 John Wiley & Sons, Ltd.

Observation 1
(a) The set D_{zw} increases as $\mathscr{S}$ increases, and if $\mathscr{S} = \mathscr{C}^a$, then $D_{zw} = \mathscr{R}_g^a$; (b) In the absence of w, $x_0 \in \mathscr{S} \Rightarrow (z_0, 0) \in D_{zw}$.

Proof
The fact that D_{zw} increases as $\mathscr{S}$ increases is easy to see. To see the rest of (a), we note that, for a general plant, $\mathscr{C}^a = \mathbf{R}^{n_1} \times \mathscr{C}_2$. If $\mathscr{S} = \mathscr{C}^a$, then $\mathscr{S} = \mathbf{R}^{n_1} \times \mathscr{C}_2$, and

$$\begin{aligned} D_{zw} &= \{(z, w) \in \mathbf{R}^n \times \mathscr{W}_0 : z - Vw \in \mathbf{R}^{n_1} \times \mathscr{C}_2\} \\ &= \{(z_1, z_2, w) \in \mathbf{R}^{n_1} \times \mathbf{R}^{n_2} \times \mathscr{W}_0 : z_1 - V_1 w \in \mathbf{R}^{n_1}, z_2 - V_2 w \in \mathscr{C}_2\} \\ &= \mathbf{R}^{n_1} \times \mathscr{R}_{g_2} = \mathscr{R}_g^a \end{aligned}$$

Part (b) is also clear if we note that $z_0 = x_0 - \Pi w_0 = x_0$ for $w_0 = 0$. □

With this observation, we see that our objective of enlarging $\mathscr{S}_{zw}$ is simply to design a feedback law such that $D_{zw} \subset \mathscr{S}_{zw}$. We will reach this objective through a series of technical lemmas.

Lemma 1
Let $u = f(z - Vw)$. Consider the closed-loop system

$$\begin{aligned} z(k+1) &= Az(k) + Bf(z(k) - Vw(k)) - B\Gamma w(k) \\ w(k+1) &= Sw(k) \end{aligned} \tag{20}$$

For this system, D_{zw} is an invariant set and for all $(z_0, w_0) \in D_{zw}$, $\lim_{k\to\infty} (z(k) - Vw(k)) = 0$.

Proof
Substitute (18) into system (20), we obtain

$$\begin{aligned} z(k+1) &= Az(k) + Bf(z(k) - Vw(k)) - AVw(k) + VSw(k) \\ &= A(z(k) - Vw(k)) + Bf(z(k) - Vw(k)) + Vw(k+1) \end{aligned}$$

Define the new state $v := z - Vw$, we have

$$v(k+1) = Av(k) + Bf(v(k))$$

which has a domain of attraction $\mathscr{S}$. This also implies that $\mathscr{S}$ is an invariant set for the v-system.

If $(z_0, w_0) \in D_{zw}$, then $v_0 = z_0 - Vw_0 \in \mathscr{S}$. It follows that

$$v(k) = z(k) - Vw(k) \in \mathscr{S}$$

for all $k \geqslant 0$ and $\lim_{k\to\infty} (z(k) - Vw(k)) = \lim_{k\to\infty} v(k) = 0$. □

Lemma 1 says that, in the presence of w, the simple feedback $u = f(z - Vw)$ will cause $z(k) - Vw(k)$ to approach zero and $z(k)$ to approach $Vw(k)$, which is bounded. Our next step is to construct a finite sequence of controllers

$$u = f_\ell(z, w, \alpha), \quad \ell = 0, 1, 2, \ldots, N,$$

Copyright © 2002 John Wiley & Sons, Ltd.

all parameterized in $\alpha \in (0,1)$. By judiciously switching between these controllers, we can cause $z(k)$ to approach $\alpha^\ell Vw(k)$ for any ℓ. By choosing N large enough, $z(k)$ will become arbitrarily small in a finite number of steps. Once $z(k)$ becomes small enough, we will use the controller

$$u = \Gamma w + \delta \,\mathrm{sat}\left(\frac{Fz}{\delta}\right)$$

(F to be specified later) to make $z(k)$ converge to the origin.

Let $F \in \mathbf{R}^{m\times n}$ be such that

$$v(k+1) = Av(k) + B\,\mathrm{sat}\,(Fv(k)) \tag{21}$$

is asymptotically stable. Let $X > 0$ be such that

$$(A+BF)^{\mathrm{T}}X(A+BF) - X < 0$$

and the ellipsoid $\mathscr{E} := \{v \in \mathbf{R}^n : v^{\mathrm{T}}Xv \leqslant 1\}$ be in the linear region of the saturation function, i.e. $|Fv|_\infty \leqslant 1$ for all $v \in \mathscr{E}$. Then $\mathscr{E}$ is an invariant set and is in the domain of attraction for the closed-loop system (21). Similar to the continuous-time case, we have the following lemma.

Lemma 2

Suppose that $D \subset \mathbf{R}^n$ is an invariant set in the domain of attraction for the system

$$v(k+1) = Av(k) + Bf(v(k)) \tag{22}$$

then for any $\alpha > 0, \alpha D$ is an invariant set in the domain of attraction for the system

$$v(k+1) = Av(k) + \alpha Bf\left(\frac{v(k)}{\alpha}\right) \tag{23}$$

For any $\alpha \in (0,1)$, there exists a positive integer N such that

$$\alpha^N |X^{1/2}Vw|_2 < \delta, \quad \forall w \in \mathscr{W}_0 \tag{24}$$

i.e. $\alpha^N Vw \in \delta\mathscr{E}$, for all $w \in \mathscr{W}_0$. Define a sequence of subsets in $\mathbf{R}^n \times \mathscr{W}_0$ as

$$D_{zw}^\ell = \{(z,w) \in \mathbf{R}^n \times \mathscr{W}_0 : z - \alpha^\ell Vw \in \alpha^\ell \mathscr{E}\}, \quad \ell = 0,1,\ldots,N$$

$$D_{zw}^{N+1} = \{(z,w) \in \mathbf{R}^n \times \mathscr{W}_0 : z \in \delta\mathscr{E}\}$$

and, on each of these sets, define a state feedback law as follows:

$$f_\ell(z,w,\alpha) = (1-\alpha^\ell)\Gamma w + \alpha^\ell \,\mathrm{sat}\left(\frac{F(z-\alpha^\ell Vw)}{\alpha^\ell}\right), \quad \ell = 0,1,\ldots,N$$

$$f_{N+1}(z,w) = \Gamma w + \delta\,\mathrm{sat}\left(\frac{Fz}{\delta}\right)$$

It can be verified that, for each $\ell = 0,1,\ldots,N+1$, $|f_\ell(z,w,\alpha)|_\infty \leqslant 1$ for all $(z,w) \in \mathbf{R}^n \times \mathscr{W}_0$.

Copyright © 2002 John Wiley & Sons, Ltd.

Lemma 3
Let $u = f_\ell(z, w, \alpha)$. Consider the closed-loop system

$$\begin{aligned} z(k+1) &= Az(k) + Bf_\ell(z(k), w(k), \alpha) - B\Gamma w(k) \\ w(k+1) &= Sw(k) \end{aligned} \tag{25}$$

For this system, D_{zw}^ℓ is an invariant set. Moreover, if $\ell = 0, 1, \ldots, N$, then for all $(z_0, w_0) \in D_{zw}^\ell$, $\lim_{k\to\infty}(z(k) - \alpha^\ell V w(k)) = 0$; if $\ell = N+1$, then, for all $(z_0, w_0) \in D_{zw}^{N+1}$, $\lim_{k\to\infty} z(k) = 0$.

Proof
With $u = f_\ell(z, w, \alpha)$, $\ell = 0, 1, \ldots, N$, we have

$$\begin{aligned} z(k+1) &= Az(k) + (1-\alpha^\ell)B\Gamma w(k) + \alpha^\ell B \operatorname{sat}\left(\frac{F(z(k) - \alpha^\ell V w(k))}{\alpha^\ell}\right) - B\Gamma w(k) \\ &= Az(k) + \alpha^\ell B \operatorname{sat}\left(\frac{F(z(k) - \alpha^\ell V w(k))}{\alpha^\ell}\right) - \alpha^\ell B\Gamma w(k) \end{aligned} \tag{26}$$

Let $v_\ell = z - \alpha^\ell V w$, then by (18)

$$v_\ell(k+1) = Av_\ell(k) + \alpha^\ell B \operatorname{sat}\left(\frac{Fv_\ell(k)}{\alpha^\ell}\right) \tag{27}$$

It follows from Lemma 2 that $\alpha^\ell \mathscr{E}$ is an invariant set in the domain of attraction for the v_ℓ-system. Hence D_{zw}^ℓ is invariant for the system (25) and if $(z_0, w_0) \in D_{zw}^\ell$, i.e.

$$v_{\ell 0} = z_0 - \alpha^\ell V w_0 \in \alpha^\ell \mathscr{E}$$

then

$$\lim_{k\to\infty}(z(k) - \alpha^\ell V w(k)) = \lim_{k\to\infty} v_\ell(k) = 0$$

With $u = f_{N+1}(z, w) = \Gamma w + \delta \operatorname{sat}(Fz/\delta)$, we have

$$z(k+1) = Az(k) + \delta B \operatorname{sat}\left(\frac{Fz(k)}{\delta}\right)$$

and the same argument applies. □

Based on the technical lemmas established above, we construct our final state feedback law as follows:

$$\begin{aligned} u &= g(z, w, \alpha, N) \\ &= \begin{cases} f_{N+1}(z, w) & \text{if } (z, w) \in \Omega^{N+1} := D_{zw}^{N+1} \\ f_\ell(z, w, \alpha) & \text{if } (z, w) \in \Omega^\ell := D_{zw}^\ell \setminus \bigcup_{j=\ell+1}^{N+1} D_{zw}^j, \quad \ell = 0, 1, \ldots, N \\ f(z - Vw) & \text{if } (z, w) \in \Omega := \mathbf{R}^n \times \mathscr{W}_0 \setminus \bigcup_{j=0}^{N+1} D_{zw}^j \end{cases} \end{aligned} \tag{28}$$

Copyright © 2002 John Wiley & Sons, Ltd.

Since $\Omega, \Omega^0, \ldots, \Omega^{N+1}$ are disjoint and their union is $\mathbf{R}^n + \mathscr{W}_0$, the controller is well defined on $\mathbf{R}^n \times \mathscr{W}_0$. What remains to be shown is that this controller will accomplish our objective if the parameter α is properly chosen.

Let

$$\alpha_0 = \max_{w \in \mathscr{W}_0} \frac{|X^{1/2} V w|_2}{|X^{1/2} V w|_2 + 1}$$

It is obvious that $\alpha_0 \in (0, 1)$.

Theorem 2

Choose any $\alpha \in (\alpha_0, 1)$ and let N be specified as in (24). Then for all $(z_0, w_0) \in D_{zw}$, the solution of the closed-loop system

$$\begin{aligned} z(k+1) &= Az(k) + Bg(z(k), w(k), \alpha, N) - B\Gamma w(k) \\ w(k+1) &= Sw(k) \end{aligned} \tag{29}$$

satisfies $\lim_{k\to\infty} z(k) = 0$, i.e. $D_{zw} \subset \mathscr{S}_{zw}$

Proof

The control $u = g(z, w, \alpha, N)$ is executed by choosing one from $f_\ell(z, w, \alpha)$, $\ell = 0, 1, \ldots, N+1$, and $f(z - Vw)$. The crucial point is to guarantee that (z, w) will move successively from Ω, to Ω^0, $\Omega^1, \ldots$, finally entering Ω^{N+1}, in which $z(k)$will converge to the origin.

Without loss of generality, we assume that $(z_0, w_0) \in \Omega \cap D_{zw}$, so the control $u = f(z - Vw)$ is in effect at the beginning. By Lemma 1, $\lim_{k\to\infty} (z(k) - Vw(k)) = 0$. Hence there is a finite step $k_0 \geqslant 0$ such that $z(k_0) - Vw(k_0) \in \mathscr{E}$, i.e. $(z(k_0), w(k_0)) \in D^0_{zw}$. The condition $(z(k), w(k)) \in D^\ell_{zw}$, $\ell > 0$, might be satisfied at a smaller step $k_1 \leqslant k_0$. In any case, there is a finite step $k_1 \geqslant 0$ such that

$$(z(k_1), w(k_1)) \in \Omega^\ell = D^\ell_{zw} \setminus \bigcup_{j=\ell+1}^{N+1} D^j_{zw}$$

for some $\ell = 0, 1, \ldots, N+1$. After that, the control $u = f_\ell(z, w, \alpha)$ will be in effect.

We claim that, for any $(z(k_1), w(k_1)) \in \Omega^\ell$, under the control $u = f_\ell(z, w, \alpha)$, there is a finite integer $k_2 > k_1$ such that $(z(k_2), w(k_2)) \in D^{\ell+1}_{zw}$.

Since $\Omega^\ell \subset D^\ell_{zw}$, by Lemma 3, we have that, under the control $u = f_\ell(z, w, \alpha)$,

$$\lim_{k\to\infty} (z(k) - \alpha^\ell V w(k)) = 0$$

Since $\alpha \in (\alpha_0, 1)$, we have

$$(1 - \alpha)|X^{1/2} V w| < \alpha, \quad \forall w \in \mathscr{W}_0$$

Therefore, for $\ell < N$

$$\begin{aligned} |X^{1/2}(z - \alpha^{\ell+1} V w)| &\leqslant |X^{1/2}(z - \alpha^\ell V w)| + \alpha^\ell (1 - \alpha)|X^{1/2} V w| \\ &< |X^{1/2}(z - \alpha^\ell V w)| + \alpha^{\ell+1} \end{aligned} \tag{30}$$

Copyright © 2002 John Wiley & Sons, Ltd.

Since the first term on the right-hand side goes to zero asymptotically, there exists a finite $k_2 > k_1$ such that

$$|X^{1/2}(z(k_2) - \alpha^{\ell+1} Vw(k_2))| \leqslant \alpha^{\ell+1}$$

This implies that $z(k_2) - \alpha^{\ell+1} Vw(k_2) \in \alpha^{\ell+1}\mathscr{E}$, i.e. $(z(k_2), w(k_2)) \in D_{zw}^{\ell+1}$.

If $\ell = N$, then, by (24)

$$|X^{1/2}z| \leqslant |X^{1/2}(z - \alpha^N Vw)| + \alpha^N |X^{1/2}Vw|$$

$$< |X^{1/2}(z - \alpha^N Vw)| + \delta$$

Also, the first term goes to zero asymptotically, so there exists a finite integer k_2 such that $|X^{1/2}z(k_2)| \leqslant \delta$, i.e. $(z(k_2), w(k_2)) \in D_{zw}^{N+1}$.

Just as before, (z, w) might have entered $D_{zw}^{\ell+q}$, $q > 1$, before it enters $D_{zw}^{\ell+1}$. In any case, there is a finite k such that

$$(z(k), w(k)) \in \Omega^{\ell+q} = D_{zw}^{\ell+q} \setminus \bigcup_{j=\ell+q+1}^{N+1} D_{zw}^{j}$$

for some $q \geqslant 1$. After that, the controller will be switched to $f_{\ell+q}(z, w, \alpha)$.

It is also important to note that, by Lemma 3, D_{zw}^{ℓ} is invariant under the control $u = f_\ell(z, w, \alpha)$. Once $(z, w) \in \Omega^\ell \subset D_{zw}^{\ell}$, it will never go back to $\Omega^q, q < \ell$ (or Ω) since $\Omega^q, q < \ell$ and Ω have no intersection with D_{zw}^{ℓ}, (but Ω^q, $q > \ell$, might have intersection with D_{zw}^{ℓ}). In summary, for any $(z_0, w_0) \in D_{zw}$, suppose $(z_0, w_0) \in \Omega^\ell$, the control will first be $f_\ell(z, w, \alpha)$ and then switch successively to $f_{\ell_1}, f_{\ell_2}, \ldots$, with $\ell_1, \ell_2, \ldots$, strictly increasing until (z, w) enters D_{zw}^{N+1} and remains there. Hence, $\lim_{k\to\infty} z(k) = 0$. □

From the proof of Theorem 2, we see that for all $(z_0, w_0) \in D_{zw}$, the number of switches is at most $N + 2$.

5. ERROR FEEDBACK

In the continuous-time case [9], the set of initial conditions on which $\lim_{k\to\infty} z(k) = 0$ is achieved by an error feedback can be made arbitrarily close to that by a state feedback. This is because the observer error can be made arbitrarily small in an arbitrarily short time interval. However, for discrete-time systems, this is impossible. Suppose that the pair in Assumption A3 is observable, then there is a minimal number of steps for any observer to reconstruct all the states. Let this minimal number of steps be n_0. We also assume that a stabilizing state feedback law $u = f(v)$, $|f(v)|_\infty \leqslant 1$ for all $v \in \mathbf{R}^n$, has been designed such that the origin of the closed-loop system (16) has a domain of attraction $S \subset \mathscr{C}^a$. By using the design of Section 4, the set

$$D_{zw} := \{(z, w) \in \mathbf{R}^n \times \mathscr{W}_0 : z - Vw \in \mathscr{S}\}$$

can be made a subset of $\mathscr{S}_{zw}$ with a state feedback $u = g(z, w, \alpha, N)$.

Copyright © 2002 John Wiley & Sons, Ltd.

Now for the case where only the tracking error e is available for feedback, a simple strategy is to let the control u be zero before the states are completely recovered, and after that we let $u = g(z, w, \alpha, N)$ as in (28) i.e.

$$u = \begin{cases} 0 & \text{if } k < n_0 \\ g(z, w, \alpha, N) & \text{if } k \geqslant n_0 \end{cases} \tag{31}$$

The question is: what is the set of initial states on which $\lim_{k\to\infty} z(k) = 0$ under the control of (31)? The answer is very simple. With $u = 0$, we have

$$z(k+1) = Az(k) - B\Gamma w(k)$$

$$w(k+1) = Sw(k)$$

By applying (18), we have

$$(z - Vw)(k+1) = A(z - Vw)(k)$$

Hence,

$$(z - Vw)(n_0) = A^{n_0}(z - Vw)(0) = A^{n_0}(z_0 - Vw_0)$$

For (z_0, w_0) to be in $\mathcal{S}_{zw}$, it suffices to have $(z(n_0), w(n_0)) \in D_{zw}$, i.e. $z(n_0) - Vw(n_0) \in \mathcal{S}$. This is in turn equivalent to

$$(z_0, w_0) \in \hat{D}_{zw} := \{(z, w) \in \mathbf{R} \times \mathcal{W}_0 : A^{n_0}(z - Vw) \in \mathcal{S}\}$$

In summary, we have $\hat{D}_{zw} \subset \mathcal{S}_{zw}$ under the control of (31).

The set $\hat{D}_{zw}$ is close to D_{zw} if A^{n_0} is close to the identity matrix.

6. CONCLUSIONS

In this paper, we have studied the problem of output regulation for linear discrete-time systems subject to actuator saturation. The plants considered here are general and can be exponentially unstable. We first characterized the regulatable region, the set of initial conditions of the plant and the exosystem for which output regulation can be achieved. We then constructed feedback laws, of both state feedback and error feedback type, that achieve output regulation on the regulatable region.

REFERENCES

1. Sontag ED. An algebraic approach to bounded controllability of linear systems. *International Journal on Control* 1984; **39**:181–188.
2. Teel AR. Feedback Stabilization: Nonlinear solutions to inherently nonlinear problems. *Ph.D. dissertation*, California, Berkeley, CA, 1992.
3. Sussmann HT, Sontag ED, Yang Y. A general result on the stabilization of linear systems using bounded controls. *IEEE Transactions on Automatic Control* 1994; **39**:2411–2425.
4. Lin Z, Saberi A. Semi-global exponential stabilization of linear systems subject to 'input saturation' via linear feedbacks. *Systems & Control Letters* 1993; **21**:225–239.
5. Lin Z. *Low Gain Feedback*. Lecture Notes in Control and Information Sciences, Vol. 240. Springer: London, 1998.
6. Hu T, Lin Z. *Control Systems with Actuator Saturation: Analysis and Design*. Birkhäuser: Boston, 2001.
7. Hu T, Lin Z, Shamash Y. Semi-global stabilization with guaranteed regional performance of linear systems subject to actuator saturation. *Systems & Control Letters*, 2001; **43**:203–210.
8. Hu T, Miller DE, Qiu L. Null controllable region of LTI discrete-time systems with input saturation, submitted for publication.

Copyright © 2002 John Wiley & Sons, Ltd.

9. Hu T, Lin Z. Output regulation of general linear systems with saturating actuators. *Proceedings of the 39th IEEE Conference on Decision and Control*, Sydney, Australia, 2000; 3242–3247.
10. De Santis R, Isidori A. Output regulation for linear systems with anti-stable eigenvalues in the presence of input saturation. *Proceedings of the 38th IEEE CDC*, 1999; 2100–2111.
11. Lin Z, Stoorvogol AA, Saberi A. Output regulation for linear systems subject to input saturation. *Automatica* 1996; **32**:29–47.
12. Francis BA. The linear multivariable regulator problem. *SIAM Journal on Control and Optimization* 1975; **15**:480–505.
13. Isidori A, Byrnes CI. Output regulation for nonlinear systems. *IEEE Transactions on Automatic Control* 1990; **35**:131–140.
14. Hájek O. *Control Theory in the Plane*. Springer: Berlin, 1991.
15. Zhou K, Doyle JC, Glover K. *Robust and Optimal Control*. Prentice-Hall: Englewood Cliffs, NJ, 1996.
16. Lin Z, Saberi A. Semi-global exponential stabilization of linear discrete-time systems subject to 'input saturation' via linear feedbacks. *Systems & Control Letters* 1995; **24**:125–132.

Copyright © 2002 John Wiley & Sons, Ltd.

Publication 20

On Semiglobal Stabilizability of Antistable Systems by Saturated Linear Feedback

Tingshu Hu and Zongli Lin

Abstract—**It was recently established that a second-order antistable linear system can be semiglobally stabilized on its null controllable region by saturated linear feedback and a higher order linear system with two or more antistable poles can be semiglobally stabilized on its null controllable region by more general bounded feedback laws. We will show in this note that a system with three real-valued antistable poles cannot be semiglobally stabilized on its null controllable region by the simple saturated linear feedback.**

Index Terms—**Actuator saturation, antistable systems, semiglobal stabilizability.**

I. Introduction

<AUTHOR PLEASE NOTE: REFERENCE [13] IS NOT USED IN PAPER. PLEASE CITE REFERENCE [13] WITHIN TEXT OR DELETE IT FROM YOUR REFERENCE LIST. THANK YOU.>There has been a long history of exploring global or semiglobal stabilizability for linear systems with saturating actuators. In 1969, Fuller [1] studied global stabilizability of a chain of integrators of length greater than two by saturated linear feedback and obtained a negative result. This important problem also attracted the attention of Sussmann and Yang [9]. They obtained similar results independently in 1991. Because of the negative result on global stabilizability with saturated linear feedback, the only choice is to use general nonlinear feedback. In 1992, Teel [11] proposed a nested feedback design technique for designing nonlinear globally asymptotically stabilizing feedback laws for a chain of integrators. This technique was fully generalized by Sussman, Sontag and Yang [8] in 1994. Alternative solutions to global stabilization problem consisting of scheduling a parameter in an algebraic Riccati equation according to the size of the state vector were later proposed in [7], [10], and [12].

Another trend in the development, motivated by the objective of designing simple controllers, is semiglobal stabilizability with saturated linear feedback laws. The notion of semiglobal asymptotic stabilization for linear systems subject to actuator saturation was introduced in [5] and [6]. The semiglobal framework for stabilization requires feedback laws that yield a closed-loop system which has an asymptotically stable equilibrium whose domain of attraction includes an *a priori* given (arbitrarily large) bounded subset of the state space. In [5] and [6], it was shown that, a linear system can be semiglobally stabilized by saturated linear feedback if it is stabilizable in the usual linear sense and has all its poles in the closed left-half plane.

It is notable that all the results mentioned above pertain to systems whose open-loop poles are all in the closed left-half plane. Such systems are said to be semistable. If a system has some open-loop poles in the open right-half plane, then it is exponentially unstable. A system with all its poles in the open right-half plane is said to be antistable. It is evident that the domain of attraction has to be a subset of the asymptotically null controllable region, the set of initial states that can be driven to the origin asymptotically with bounded controls delivered by the saturating actuators. Since the asymptotically null controllable region of a semistable system is the whole state space (if it is stabilizable in the linear sense), it is possible to stabilize it globally/semiglobally. However, the asymptotically null controllable region of an exponentially unstable system is not the whole state space, hence it cannot be globally/semiglobally stabilized with saturated feedback. For this reason, we generalized the notion of global/semiglobal stabilization, which was only suitable for semistable systems, by giving it a new meaning [2], [3]. A linear system subject to actuator saturation is globally stabilizable if there is a saturated feedback law such that the closed-loop system has a stability region which is equal to the asymptotically null controllable region; it is semiglobally stabilizable if, given an arbitrary compact subset of the asymptotically null controllable region, there is a saturated feedback law under which the closed-loop system has a stability region that includes this given compact set.

A first step toward global/semiglobal stabilization, which cannot be bypassed, is the characterization of the asymptotically null controllable region. We made this first step in [2], and then proceeded to construct stabilizing feedback laws for semiglobal stabilization. In [2] and [3], we developed simple feedback laws for systems with two antistable poles. For a second-order antistable system the controllers proposed are a family of saturated linear feedbacks of the form $u = \text{sat}(kF_0x)$ and for a high-order system with only two antistable poles, each controller in the family switches between two saturated linear feedbacks. In [2] and [4], we proposed a nonlinear switching feedback laws for more general systems. The controllers are more complicated than those of [3].

Given the results of [2]–[4], it is interesting to ask if a system with three or more antistable poles can be semiglobally stabilized with saturated linear feedback. In contrast to semistable systems, which can be semiglobally stabilized by saturated linear feedback, this note will show that a system with three or more antistable poles cannot even be *semiglobally* stabilized by saturated linear feedback.

The remaining of this note is organized as follows. Section II reviews some results on the asymptotically null controllable region and develops some algebraic tools. Section III establishes the fact that a third-order antistable system cannot be semiglobally stabilized by saturated linear feedback. Some concluding remarks are given in Section IV.

II. Preliminary Results and Some Algebraic Tools

We recall from [2] a description of the asymptotically null controllable region. Consider a single-input linear system subject to actuator saturation

$$\dot{x} = Ax + bu, \quad x \in \mathbf{R}^n, u \in \mathbf{R}, |u| \leq 1. \tag{1}$$

Assume that (A, b) is controllable in the usual linear sense. Then the asymptotically null controllable region is the same as the null controllable region, which is the set of initial states that can be driven to the origin in finite time. We use $\mathcal{C}$ to denote the null controllable region of system (1). It is known that if A is semistable, then $\mathcal{C} = \mathbf{R}^n$ and if A is antistable, then $\mathcal{C}$ is a bounded convex open set containing the origin in its interior. In this note, we will restrict our attention to third-order antistable systems with only real poles. For such a system, the boundary of the null controllable region is (see [2])

$$\partial\mathcal{C} = \left\{ \pm \left(-2e^{-At_2} + 2e^{-At_1} - I \right) A^{-1}b : 0 \leq t_1 \leq t_2 \leq \infty \right\}. \tag{2}$$

Manuscript received September 28, 2001; revised January 13, 2002. Recommended by Associate Editor J. Huang. This work was supported in part by the US Office of Naval Research Young Investigator Program under Grant N00014-99-1-0670.

The authors are with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22903 USA (e-mail: th7f@virginia.edu; zl5y@virginia.edu).

Publisher Item Identifier S 0018-9286(02)XXXXX-X.

0018-9286/02$17.00 © 2002 IEEE

Fig. 1 illustrates a typical shape of $\mathcal{C}$ with a bunch of curves on $\partial\mathcal{C}$.

Denote

$$\partial\mathcal{C}^+ = \left\{\left(-2e^{-At_2} + 2e^{-At_1} - I\right)A^{-1}b : 0 < t_1 < t_2 < \infty\right\}$$
$$\partial\mathcal{C}^- = -\partial\mathcal{C}^+$$
$$\partial\mathcal{C}^0 = \left\{\left(\pm(-2e^{-At_2} + I\right)A^{-1}b : 0 \le t_2 \le \infty\right\}.$$

It can be verified that $\partial\mathcal{C} = \partial\mathcal{C}^+ \cup \partial\mathcal{C}^- \cup \partial\mathcal{C}^0$ and $\partial\mathcal{C}^+$, $\partial\mathcal{C}^-$ and $\partial\mathcal{C}^0$ are disjoint. In Fig. 1, the solid curves are on $\partial\mathcal{C}^+$ and the dashed ones are on $\partial\mathcal{C}^-$. The two smooth curves connecting z_e^+ and z_e^- (one on the highest boundary and the other on the lowest boundary) form $\partial\mathcal{C}^0$. It has been shown in [2] that all the curves are trajectories of the system (1) under controls that only take values of 1 and -1. They are called extremal trajectories. The solid curves on $\partial\mathcal{C}^+$ are trajectories of the system (1) under the control $u = 1$ and those on $\partial\mathcal{C}^-$ are trajectories under the control $u = -1$. The curve on the highest boundary is a trajectory going from z_e^- to z_e^+ under the control $u = -1$ and the one on the lowest boundary is a trajectory going from z_e^+ to z_e^- under the control $u = 1$.

From the definition, we see that $\partial\mathcal{C}^+$ and $\partial\mathcal{C}^-$ are smooth surfaces and $\partial\mathcal{C}^0$ is a closed curve connecting $\partial\mathcal{C}^+$ and $\partial\mathcal{C}^-$, i.e., $\partial\mathcal{C}^0$ is composed of all the common limit points of $\partial\mathcal{C}^+$ and $\partial\mathcal{C}^-$.

Since $\mathcal{C}$ is an open set, $\partial\mathcal{C} \cap \mathcal{C}$ is empty. Here, we summarize some facts from [2].

Fact 2.1: Under the constraint that $|u| \le 1$, the following hold true.

1) All the states in $\mathcal{C}$ can be driven to the origin.
2) All the states outside of $\mathcal{C} \cup \partial\mathcal{C}$ will grow unbounded no matter what control is applied.
3) All the states on $\partial\mathcal{C}$ cannot be driven to the interior of $\mathcal{C}$. The only way to keep them bounded is to make them stay on $\partial\mathcal{C}$ with a control $u = 1$ or $u = -1$. For $x(0) \in \partial\mathcal{C}^+$, the only control to keep $x(t), t \in [0, \varepsilon)$ for some $\varepsilon > 0$ on $\partial\mathcal{C}$ is $u = 1$ and for $x(0) \in \partial\mathcal{C}^-$, the only control is $u = -1$.

From Fact 2.1, we know that $\mathcal{C} \cup \partial\mathcal{C}$ is the largest bounded set that can be rendered invariant by means of admissible controls.

The basic fact we will use to prove our main result is that any segment on $\partial\mathcal{C}^0$ is three dimensional, i.e., it cannot be fit into any plane. Before proving this fact, we need an algebraic result which will be used several times in this note.

Lemma 2.1: Suppose that (A, b) is controllable, $A \in \mathbf{R}^{3\times3}$ is anti-stable and A has no complex eigenvalues. Let t_1, t_2, t_3 be distinct real numbers. Then, for all $(k_1, k_2, k_3) \neq (0, 0, 0)$

$$\left(k_1 e^{At_1} + k_2 e^{At_2} + k_3 e^{At_3}\right) b \neq 0. \tag{3}$$

Proof: See Appendix A.

We note that, if A has complex eigenvalues $\alpha \pm j\beta$ and $(t_1, t_2, t_3) = (\pi N_1/\beta, \pi N_2/\beta, \pi N_3/\beta)$, where N_1, N_2 and N_3 are integers, there may exist $(k_1, k_2, k_3) \neq (0, 0, 0)$ that satisfy (3). For instance, suppose that A has eigenvalues $1, 1 \pm j1$. Let $t_1 = 0, t_2 = \pi, t_3 = 2\pi$, $k_1 = 1, k_2 = 0, k_3 = -e^{-2\pi}$, then $k_1 e^{At_1} + k_2 e^{At_2} + k_3 e^{At_3} = 0$.

Proposition 2.1: Let

$$x(t) = \left(-2e^{-At} + I\right)A^{-1}b.$$

If t_1, t_2, t_3 and t_4 are distinct numbers, then $x(t_i), i = 1, 2, 3, 4$, are not in the same plane.

Proof: For simplicity, assume that $t_1 < t_2 < t_3 < t_4$. We first show that $x(t_i), i = 1, 2, 3$, are not on the same straight line. Suppose, on the contrary, that they are then

$$x(t_3) - x(t_2) = c(x(t_2) - x(t_1))$$

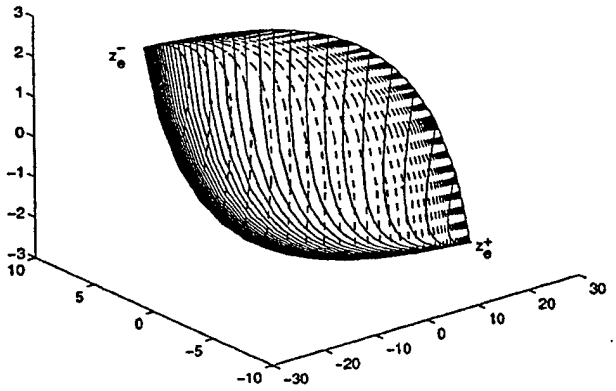


Fig. 1. $\partial\mathcal{C}$ of a third-order system.

for some c, i.e.,

$$\left(e^{-At_3} - e^{-At_2}\right)A^{-1}b - c\left(e^{-At_2} - e^{-At_1}\right)A^{-1}b = 0.$$

This can be written as

$$A^{-1}e^{-At_3}\left(I - (1+c)e^{-A(t_3-t_2)} + ce^{A(t_3-t_1)}\right)b = 0$$

which contradicts Lemma 2.1. Here, we note that A and e^{At} commute.

Now that $x(t_i), i = 1, 2, 3$, are not on the same straight line, they uniquely determine a plane. Let this plane be $fx = 1$. Suppose, on the contrary, that $x(t_4)$ is also in this plane, then

$$fx(t_1) = fx(t_2) = fx(t_3) = fx(t_4) = 1.$$

By mean value theorem, there exist $t_1' \in (t_1, t_2)$, $t_2' \in (t_2, t_3)$ and $t_3' \in (t_3, t_4)$ such that

$$f\dot{x}(t_1') = f\dot{x}(t_2') = f\dot{x}(t_3') = 0$$

which is equivalent to

$$fe^{At_1'}b = fe^{-At_2'}b = fe^{-At_3'}b = 0.$$

This equality can be written as

$$f\left[e^{-At_1'}b \quad e^{-At_2'}b \quad e^{-At_3'}b\right] = 0$$

which implies that the 3×3 matrix

$$M = \left[e^{-At_1'}b \quad e^{-At_2'}b \quad e^{-At_3'}b\right]$$

is singular. This again contradicts Lemma 2.1. Hence we conclude that $x(t_i), i = 1, 2, 3, 4$, are not in the same plane. □

Recalling that

$$\partial\mathcal{C}^0 = \left\{\pm\left(-2e^{-At} + I\right)A^{-1}b : 0 \le t \le \infty\right\}$$

Proposition 2.1 implies that any segment of $\partial\mathcal{C}^0$ is three dimensional and cannot be placed in one plane.

III. Main Results

For an $x_0 \in \mathbf{R}^3$ and a positive number r, denote

$$\mathcal{B}(x_0, r) = \left\{x \in \mathbf{R}^3 : \|x - x_0\| \le r\right\}$$

where $\|\cdot\|$ is the Euclidean norm. We use $\text{sat}(\cdot)$ to denote the standard saturation function, i.e., $\text{sat}(u) = \text{sign}(u)\min\{1, |u|\}$. Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two subsets of $\mathbf{R}^n$. Then their distance is defined as

$$d(\mathcal{X}_1, \mathcal{X}_2) := \inf_{x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2} \|x_1 - x_2\|.$$

In terms of $\partial\mathcal{C}^+, \partial\mathcal{C}^-$ and $\partial\mathcal{C}^0$, the space $\mathbf{R}^3$ can be divided into three subsets

$$\mathbf{E}^+ = \{\gamma x : \gamma \in (0, \infty), x \in \partial\mathcal{C}^+\}$$
$$\mathbf{E}^- = -\mathbf{E}^+$$
$$\mathbf{E}^0 = \{\gamma x : \gamma \in [0, \infty), x \in \partial\mathcal{C}^0\}.$$

We see that $\mathbf{E}^+$ and $\mathbf{E}^-$ are simply connected open sets and $\mathbf{E}^0$ is a surface consisting of the common limit points of $\mathbf{E}^+$ and $\mathbf{E}^-$. In other words, $\mathbf{E}^+$ and $\mathbf{E}^-$ are connected by $\mathbf{E}^0$.

Theorem 3.1: A third-order antistable system with real poles cannot be semiglobally stabilized by saturated linear feedback on its null controllable region $\mathcal{C}$. Specifically, for any saturated linear feedback law $u = \text{sat}(Fx)$, there exists a ball $\mathcal{B}(x_*, r_*) \subset \mathcal{C}$, where $r_* > 0$ is independent of F, such that all the trajectories starting from $\mathcal{B}(x_*, r_*)$ will grow unbounded.

To prove Theorem 3.1, we first examine the difference between the control under a saturated linear feedback and the control that is required to keep the state bounded. Unlike the case for a second order system, here for a third order system, there is always a ball of fixed size where the difference between the two controls is greater than a fixed positive number. The result is stated in the following lemma.

Lemma 3.1: There exist positive numbers r_0 and d_0 such that for any $F \in \mathbf{R}^{1\times 3}$, there is a ball $\mathcal{B}(x_0, r_0)$, where $x_0 \in \partial\mathcal{C}^+$, such that $d(\mathcal{B}(x_0, r_0), \mathbf{E}^0) \geq d_0$ and $\text{sat}(Fx) \leq 1/2$ for all $x \in \mathcal{B}(x_0, r_0)$.

Proof: For an $F \in \mathbf{R}^{1\times 3}$, denote the distance between the two planes $Fx = 1$ and $Fx = -1$ by $g(F)$. By Proposition 2.1, $\partial\mathcal{C}^0$ is a three dimensional closed curve. So there exists a minimal distance g_0 between $Fx = 1$ and $Fx = -1$ such that $\partial\mathcal{C}^0$ lies completely between these two planes. In other words, if $g(F) = g_1 < g_0$, then the total length of the segments of $\partial\mathcal{C}^0$ which are in the half space $Fx \leq -1$ must be greater than a positive number depending only on g_1. Also by Proposition 2.1, there are no more than two segments of $\{x(t) = (-2e^{-At} + I)A^{-1}b : 0 \leq t \leq \infty\}$ which are in the half space, otherwise, there would be four points $x(t_i), i = 1, 2, 3, 4$, in the plane $Fx = -1$. Hence, there is a segment of $\partial\mathcal{C}^0$, with length greater than a fixed positive number, that is completely in the half space $Fx \leq -1$.

Let us first consider an F such that $g(F) \leq 1/2g_0$. By the foregoing arguments, there is a segment of $\partial\mathcal{C}^0$, with length greater than $l_0 > 0$, that is completely in the half space $Fx \leq -1$. Since $\partial\mathcal{C}^0$ is a compact set and any segment on it is three dimensional, the largest distance from a point of the segment to the plane $Fx = -1$ is greater than a fixed positive number. Recalling that $\partial\mathcal{C}^0$ is the closed curve that connects $\partial\mathcal{C}^+$ and $\partial\mathcal{C}^-$, we know that there is a simply connected region on the surface of $\partial\mathcal{C}^+$ that is in the half space $Fx \leq -1$ and the surface area of this region is greater than a fixed-positive number. It follows that there exists a ball $\mathcal{B}(x_1, r_1)$ with $x_1 \in \partial\mathcal{C}^+$ and r_1 a fixed-positive number such that $\mathcal{B}(x_1, r_1)$ is in the half space $Fx \leq -1$ and $d(\mathcal{B}(x_1, r_1), \mathbf{E}^0) \geq d_1$, where d_1 is also a fixed positive number. Here, for $x \in \mathcal{B}(x_1, r_1)$, $\text{sat}(Fx) = -1$.

Next, we consider an F such that $g(F) > 1/2g_0$. Then there exists a segment of $\partial\mathcal{C}^0$, of length greater than a fixed-positive number, which is between the two planes $Fx = -1/2$ and $Fx = 1/2$. Following similar arguments as in the previous paragraph, there is a ball $\mathcal{B}(x_2, r_2)$ with $x_2 \in \partial\mathcal{C}^+$ and r_2 a fixed-positive number such that $\mathcal{B}(x_2, r_2)$ is between the two planes $Fx = -1/2$ and $Fx = 1/2$ and $d(\mathcal{B}(x_2, r_2), \mathbf{E}^0) \geq d_2$, where d_2 is also a fixed positive number. Here, for $x \in \mathcal{B}(x_2, r_2)$, $|\text{sat}(Fx)| \leq 1/2$.

If we let $d_0 = \min\{d_1, d_2\}$ and $r_0 = \min\{r_1, r_2\}$, then the result of Lemma 3.1 readily follows. □

To prove Theorem 3.1, we need to show that, under the control of any feedback $u = \text{sat}(Fx)$, there exists a ball in $\mathcal{C}$ of radius greater than a fixed positive number, such that all the trajectories starting from the ball will go out of $\mathcal{C}$ and diverge. We will use Lyapunov function analysis to show this result. The Lyapunov function is defined in terms of $\partial\mathcal{C}$ as follows:

$$V(x) := \gamma \geq 0, \text{ such that } \frac{x}{\gamma} \in \partial\mathcal{C} \quad (\text{or } x \in \gamma\partial\mathcal{C}). \tag{4}$$

Since $\mathcal{C}$ is a bounded convex open set, any ray starting from the origin has a unique intersection with $\partial\mathcal{C}$ and hence any vector in the state space can be uniquely scaled to be exactly on $\partial\mathcal{C}$. Therefore, $V(x)$ is a well-defined positive–definite function.

Clearly, $V(x) = 1$ for all $x \in \partial\mathcal{C}$ and $V(x) < 1$ for all $x \in \mathcal{C}$. We also see that $V(\alpha x) = \alpha V(x)$ for any $\alpha > 0$. Moreover, if $\partial V/\partial x$ exists at some x_0, then

$$\left.\frac{\partial V}{\partial x}\right|_{x=\alpha x_0} = \left.\frac{\partial V}{\partial x}\right|_{x=x_0}. \tag{5}$$

To see this, note that

$$\frac{V\left(\alpha x_0 + \begin{bmatrix}\Delta\\0\\0\end{bmatrix}\right) - V(\alpha x_0)}{\Delta} = \frac{V\left(x_0 + \begin{bmatrix}\frac{1}{\alpha}\Delta\\0\\0\end{bmatrix}\right) - V(x_0)}{\frac{1}{\alpha}\Delta}$$

and the aforementioned equality is also true if we replace $\begin{bmatrix}\Delta\\0\\0\end{bmatrix}$ with $\begin{bmatrix}0\\\Delta\\0\end{bmatrix}$ or $\begin{bmatrix}0\\0\\\Delta\end{bmatrix}$.

Lemma 3.2: The Lyapunov function $V(x)$ is continuously differentiable in x for $x \in \mathbf{E}^+$. For all $x \in \mathbf{E}^+$, $b^T(\partial V/\partial x) \neq 0$.

Proof: Every $x \in \mathbf{E}^+$ can be expressed as

$$x = \gamma\left(-2e^{-At_2} + 2e^{-At_1} - I\right)A^{-1}b$$

for some $\gamma \in (0, \infty), 0 \leq t_1 < t_2 < \infty$. It follows from the definition that $V(x) = \gamma$. We see that x is analytic in γ, t_1 and t_2. Moreover

$$dx = T\begin{bmatrix}d\gamma\\dt_1\\dt_2\end{bmatrix}$$

where

$$T = \left[\frac{\partial x}{\partial\gamma}\ \frac{\partial x}{\partial t_1}\ \frac{\partial x}{\partial t_2}\right]$$
$$= \left[\left(-2e^{-At_2} + 2e^{-At_1} - I\right)A^{-1}b \vdots -2\gamma e^{-At_1}b \vdots 2\gamma e^{-At_2}b\right].$$

We claim that T is nonsingular. This can be seen as follows. For simplicity, consider an $x_0 \in \partial\mathcal{C}^+$, then $\gamma = 1$. Applying Lemma 2.1, we see that $\partial x/\partial t_1 = -2e^{-At_1}b$ and $\partial x/\partial t_2 = 2e^{-At_2}b$ are independent and they determine a plane $fx = 1$ that contains x_0. Since $\partial\mathcal{C}^+$ is smooth, this plane is tangential to $\partial\mathcal{C}^+$ at x_0. Since $\mathcal{C}$ is a bounded convex set containing the origin in its interior, this plane $fx = 1$ does not contain the origin. Hence, the vector from x_0 to the origin, $-x_0 = -(-2e^{-At_2} + 2e^{-At_1} - I)A^{-1}b$, must be independent of the

two vectors that determine the plane. That is, the three column vectors in T are independent.

Now that T is nonsingular, we have

$$\begin{bmatrix} d\gamma \\ dt_1 \\ dt_2 \end{bmatrix} = T^{-1} dx.$$

It is also clear that T and T^{-1} are continuous. Hence, γ, t_1 and t_2 are continuously differentiable in x for $x \in \mathbf{E}^+$. Therefore, $V(x) = \gamma$ is continuously differentiable in x for $x \in \mathbf{E}^+$.

Noting that $V(x) = \gamma$, we have

$$\frac{\partial V}{\partial x} = \left([1 \quad 0 \quad 0] T^{-1} \right)^T$$

and

$$\left(\frac{\partial V}{\partial x}\right)^T b = [1 \quad 0 \quad 0] T^{-1} b.$$

Suppose on the contrary that $(\partial V/\partial x)^T b = 0$, then

$$T^{-1} b = \begin{bmatrix} 0 \\ k_1 \\ k_2 \end{bmatrix}$$

for some $(k_1, k_2) \neq (0, 0)$. It follows that:

$$b = T \begin{bmatrix} 0 \\ k_1 \\ k_2 \end{bmatrix} = -2k_1 \gamma e^{-At_1} b + 2k_2 \gamma e^{-At_2} b$$

which contradicts Lemma 2.1. Therefore, we must have $(\partial V/\partial x)^T b \neq 0$ for all $x \in \mathbf{E}^+$. □

Since $(\partial V/\partial x)^T b$ is continuous in $\mathbf{E}^+$, from Lemma 3.2, we conclude that $(\partial V/\partial x)^T b$ either "> 0" or "< 0" in $\mathbf{E}^+$ and for any compact subset of $\mathbf{E}^+$, $|(\partial V/\partial x)^T b|$ is greater than a positive number. Now we are ready to prove the main result of this note.

Proof of theorem 3.1: Let $F \in \mathbf{R}^{1\times 3}$ be an arbitrary feedback gain matrix. From Lemma 3.1, we know that there always exists a ball $\mathcal{B}(x_0, r_0)$, $x_0 \in \partial\mathcal{C}^+$, such that $\text{sat}(Fx) \leq 1/2$ for all $x \in \mathcal{B}(x_0, r_0)$ and $d(\mathcal{B}(x_0, r_0), \mathbf{E}^0) \geq d_0$. Here, d_0 and r_0 are independent of F.

Since $\mathcal{B}(x_0, r_0)$ contains one point x_0 in $\mathbf{E}^+$ and has a distance greater than d_0 from $\mathbf{E}^0$, we have $\mathcal{B}(x_0, r_0) \subset \mathbf{E}^+$. Without loss of generality, assume that $\mathcal{B}(x_0, r_0) \subset \mathbf{E}^+ \cap 5/4\mathcal{C} \setminus 3/4\mathcal{C}$. Otherwise, we can choose a smaller r_0. Let $\mathcal{M}$ be the maximal compact set in $\mathbf{E}^+ \cap 5/4\mathcal{C} \setminus 3/4\mathcal{C}$ such that $d(\mathcal{M}, \mathbf{E}^0) = d_0$. By Lemma 3.2, there is a positive number η such that

$$\min \left\{ \left| \left(\frac{\partial V}{\partial x}\right)^T b \right| : x \in \mathcal{M} \right\} \geq \eta.$$

Since $d(\mathcal{B}(x_0, r_0), \mathbf{E}^0) \geq d_0$, we must have $\mathcal{B}(x_0, r_0) \subset \mathcal{M}$. Therefore

$$\left| \left(\frac{\partial V}{\partial x}\right)^T b \right| \geq \eta, \ \forall x \in \mathcal{B}(x_0, r_0). \tag{6}$$

Consider the derivative of the Lyapunov function $V(x)$ along the trajectory of the system

$$\dot{x} = Ax + b\,u.$$

We have

$$\dot{V}(x, u) = \left(\frac{\partial V}{\partial x}\right)^T Ax + \left(\frac{\partial V}{\partial x}\right)^T bu.$$

From Fact 2.1, we know that if a control $u = 1$ is applied at $x \in \partial\mathcal{C}^+$, the trajectory will stay on $\partial\mathcal{C}^+$ and $V(x)$ will remain to be 1. Hence, for $x \in \partial\mathcal{C}^+$,

$$\dot{V}(x, 1) = \left(\frac{\partial V}{\partial x}\right)^T Ax + \left(\frac{\partial V}{\partial x}\right)^T b = 0. \tag{7}$$

Since $(\partial V/\partial x)^T b \neq 0$ and any control $u < 1$ is unable to bring a state on $\partial\mathcal{C}^+$ into a smaller level set $\gamma_1 \mathcal{C}$, $\gamma_1 < 1$, we must have

$$\dot{V}(x, u) = \left(\frac{\partial V}{\partial x}\right)^T Ax + \left(\frac{\partial V}{\partial x}\right)^T bu > 0$$

for all $u < 1$. This implies that $(\partial V/\partial x)^T b < 0$ for $x \in \partial\mathcal{C}^+$ and also for $x \in \mathbf{E}^+$. It follows from (6) that:

$$\left(\frac{\partial V}{\partial x}\right)^T b < -\eta, \quad \forall x \in \mathcal{B}(x_0, r_0). \tag{8}$$

By (5) and (7), if a control $u = \gamma = V(x)$ is applied at $x \in \gamma\partial\mathcal{C}^+$, we will have

$$\left(\frac{\partial V}{\partial x}\right)^T Ax + \left(\frac{\partial V}{\partial x}\right)^T bV(x) = 0. \tag{9}$$

Now consider the system under the saturated linear feedback $u = \text{sat}(Fx)$. Recalling from Lemma 3.1 that $\text{sat}(Fx) \leq 1/2$ and $V(x) \geq 3/4$ for all $x \in \mathcal{B}(x_0, r_0)$, we have

$$\text{sat}(Fx) - V(x) \leq \frac{1}{2} - \frac{3}{4} = -\frac{1}{4}. \tag{10}$$

It follows from (9), (8), and (10) that:

$$\begin{aligned} \dot{V}(x, \text{sat}(Fx)) &= \left(\frac{\partial V}{\partial x}\right)^T Ax + \left(\frac{\partial V}{\partial x}\right)^T b\,\text{sat}(Fx) \\ &= \left(\frac{\partial V}{\partial x}\right)^T Ax + \left(\frac{\partial V}{\partial x}\right)^T bV(x) \\ &\quad + \left(\frac{\partial V}{\partial x}\right)^T b(\text{sat}(Fx) - V(x)) \\ &= \left(\frac{\partial V}{\partial x}\right)^T b(\text{sat}(Fx) - V(x)) \\ &> \frac{\eta}{4} \end{aligned}$$

for all $x \in \mathcal{B}(x_0, r_0)$. We see that there exists a positive number N such that $\|\dot{x}\| \leq N$ for all $x \in 5/4\mathcal{C}$ under any saturated feedback control. Hence, there exists a $t_0 \in (0, 4/\eta)$ and an $r_1 \in (0, r_0)$ independent of x_0, such that all the trajectories starting from $\mathcal{B}(x_0, r_1)$ will stay inside $\mathcal{B}(x_0, r_0)$ for $t \in [0, t_0]$. Therefore, $V(x(t_0)) - V(x(0)) \geq \eta/4t_0$. Also, there exists a $r_2 \in (0, r_1)$ such that

$$\mathcal{B}(x_0, r_2) \subset \mathbf{E}^+ \setminus \left(1 - \frac{\eta t_0}{4}\right)\mathcal{C}.$$

Clearly, for all $x(0) \in \mathcal{B}(x_0, r_2)$, $V(x(0)) > 1 - \eta t_0/4$. Hence, for any trajectory starting from $\mathcal{B}(x_0, r_2)$, we will have $V(x(t_0)) > V(x(0)) + \eta/4t_0 > 1$, which means that the trajectory has gone out of $\partial\mathcal{C}$ at t_0 and will diverge by Fact 2.1.

It is easy to see that there exists a ball $\mathcal{B}(x_*, r_*) \subset \mathcal{B}(x_0, r_2) \cap \mathcal{C}$, with r_* greater than a fixed positive number. In summary, no matter what F is, there always exists a ball $\mathcal{B}(x_*, r_*) \subset \mathcal{C}$ from which the trajectories will diverge under the saturated linear feedback $u = \text{sat}(Fx)$. This completes the proof. □

IV. Conclusion

We have shown in this note that a third-order antistable system with real eigenvalues cannot be semiglobally stabilized with saturated linear feedback. The study is based on examining a Lyapunov function defined in terms of the null controllable region. The level sets of the Lyapunov function are the null controllable region scaled by positive numbers. The main idea is to show the existence of a ball inside the null controllable region, with radius greater than a fixed-positive number, from which the Lyapunov function will grow unbounded. The increasing of the Lyapunov function is caused by the difference between the control $u = \text{sat}(Fx)$ and the one that is required to keep the state within a level set. The difference between the two controls cannot be reduced to an arbitrarily small level because the two surfaces $\partial\mathcal{C}^+$ and $\partial\mathcal{C}^-$ cannot be separated with a plane, or, the closed curve that connects these two surfaces is three dimensional. For systems with complex eigenvalues, if we define $\partial\mathcal{C}^+$ to be the set of states on $\partial\mathcal{C}$ which can only be kept on $\partial\mathcal{C}$ by $u = 1$ and $\partial\mathcal{C}^-$ to be the set of states on $\partial\mathcal{C}$ which can only be kept on $\partial\mathcal{C}$ by $u = -1$, then intuitively, these two surfaces are not separable with a plane. With a similar procedure, the negative result in this note can be extended to systems with complex eigenvalues, although it is somewhat harder to characterize the curve that separates $\partial\mathcal{C}^+$ and $\partial\mathcal{C}^-$.

Appendix
Proof of Lemma 2.1

For simplicity, we assume that the smallest t_i is t_3 and $t_3 = 0$. Otherwise, we can multiply (3) from left with e^{-At_3}. We also assume that $t_2 > t_1 > 0$.

We will first show that

$$k_1 e^{At_1} + k_2 e^{At_2} + k_3 I \neq 0,\ \forall (k_1, k_2, k_3) \neq (0,0,0). \tag{11}$$

We assume that A has three distinct eigenvalues $\lambda_1, \lambda_2, \lambda_3$, with $0 < \lambda_1 < \lambda_2 < \lambda_3$. For the case where A has two or three identical eigenvalues, we can prove the result in a simpler way using similar ideas. We further assume that $A = \text{diag}[\lambda_1, \lambda_2, \lambda_3]$. Then (11) can be reorganized as

$$\begin{bmatrix} e^{\lambda_1 t_1} & e^{\lambda_1 t_2} & 1 \\ e^{\lambda_2 t_1} & e^{\lambda_2 t_2} & 1 \\ e^{\lambda_3 t_1} & e^{\lambda_3 t_2} & 1 \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ k_3 \end{bmatrix} \neq 0,\ \forall (k_1, k_2, k_3) \neq (0,0,0). \tag{12}$$

This is equivalent to

$$\det \begin{bmatrix} e^{\lambda_1 t_1} & e^{\lambda_1 t_2} & 1 \\ e^{\lambda_2 t_1} & e^{\lambda_2 t_2} & 1 \\ e^{\lambda_3 t_1} & e^{\lambda_3 t_2} & 1 \end{bmatrix} \neq 0. \tag{13}$$

Direct computation shows that

$$\det \begin{bmatrix} e^{\lambda_1 t_1} & e^{\lambda_1 t_2} & 1 \\ e^{\lambda_2 t_1} & e^{\lambda_2 t_2} & 1 \\ e^{\lambda_3 t_1} & e^{\lambda_3 t_2} & 1 \end{bmatrix} = \left(e^{\lambda_2 t_1} - e^{\lambda_1 t_1}\right)\left(e^{\lambda_3 t_2} - e^{\lambda_1 t_2}\right) - \left(e^{\lambda_2 t_2} - e^{\lambda_1 t_2}\right)\left(e^{\lambda_3 t_1} - e^{\lambda_1 t_1}\right).$$

We claim that

$$\frac{e^{\lambda_3 t_2} - e^{\lambda_1 t_2}}{e^{\lambda_3 t_1} - e^{\lambda_1 t_1}} > \frac{e^{\lambda_2 t_2} - e^{\lambda_1 t_2}}{e^{\lambda_2 t_1} - e^{\lambda_1 t_1}} \tag{14}$$

from which (13) and (11) will follow.

We now proceed to prove (14). Define

$$f(\lambda) := \frac{e^{\lambda t_2} - e^{\lambda_1 t_2}}{e^{\lambda t_1} - e^{\lambda_1 t_1}} \times \frac{e^{\lambda_1 t_1}}{e^{\lambda_1 t_2}} = \frac{e^{(\lambda - \lambda_1) t_2} - 1}{e^{(\lambda - \lambda_1) t_1} - 1}.$$

It suffices to show that $f(\lambda)$ is an increasing function of λ for $\lambda > \lambda_1$, or, equivalently, that

$$f_1(\lambda) = \frac{e^{\lambda t_2} - 1}{e^{\lambda t_1} - 1}$$

is an increasing function of λ for $\lambda > 0$. Let $g(\lambda) = df_1/d\lambda \left(e^{\lambda t_1} - 1\right)^2$, then

$$\begin{aligned} g(\lambda) &= (t_2 - t_1)e^{\lambda(t_2+t_1)} - t_2 e^{\lambda t_2} + t_1 e^{\lambda t_1} \\ &= e^{\lambda t_1}\left((t_2 - t_1)e^{\lambda t_2} - t_2 e^{\lambda(t_2 - t_1)} + t_1\right). \end{aligned}$$

Let

$$g_1(\lambda) = g(\lambda)e^{-\lambda t_1} = (t_2 - t_1)e^{\lambda t_2} - t_2 e^{\lambda(t_2 - t_1)} + t_1.$$

Then

$$\frac{dg_1}{d\lambda} = (t_2 - t_1)t_2\left(e^{\lambda t_2} - e^{\lambda(t_2 - t_1)}\right) > 0,\ \forall \lambda > 0.$$

Since $g_1(0) = 0$, it follows that $g_1(\lambda) > 0$ and hence $g(\lambda) > 0$ for all $\lambda > 0$. Therefore, $df_1/d\lambda > 0$ for all $\lambda > 0$ and hence $f_1(\lambda)$ is an increasing function of λ. It follows that (14) and (11) are true.

We next show (3). Suppose, on the contrary, that there exist $(k_1, k_2, k_3) \neq (0,0,0)$ such that

$$\left(k_1 e^{At_1} + k_2 e^{At_2} + k_3 I\right) b = 0.$$

Noting that

$$k_1 e^{At_1} + k_2 e^{At_2} + k_3 I = \gamma_1 A^2 + \gamma_2 A + \gamma_3 I$$

for some $(\gamma_1, \gamma_2, \gamma_3) \neq (0,0,0)$, we would have

$$[A^2 b \quad Ab \quad b]\begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} = 0$$

which contradicts the assumption that (A, b) is controllable. □

References

[1] A. T. Fuller, "In-the-large stability of relay and saturating control systems with linear controller," *Int. J. Control*, vol. 10, pp. 457–480, 1969.

[2] T. Hu and Z. Lin, *Control Systems With Actuator Saturation: Analysis and Design*. Boston, MA: Birkhäuser, 2001.

[3] T. Hu, Z. Lin, and L. Qiu, "Stabilization of exponentially unstable linear systems with saturating actuators," *IEEE Trans. Automat. Contr.*, vol. 46, pp. 973–979, June 2001.

[4] T. Hu, Z. Lin, and Y. Shamash, "Semiglobal stabilization with guaranteed regional performance of linear systems subject to actuator saturation," *Syst. Control Lett.*, vol. 43, no. 3, pp. 203–210, 2001.

[5] Z. Lin, *Low Gain Feedback*. London, U.K.: Springer-Verlag, 1998, vol. 240, Lecture Notes in Control and Information Sciences.

[6] Z. Lin and A. Saberi, "Semiglobal exponential stabilization of linear systems subject to 'input saturation' via linear feedbacks," *Syst. Control Lett.*, vol. 21, pp. 225–239, 1993.

[7] A. Megretski, "L_2 BIBO output feedback stabilization with saturated control," in *Proc. 13th IFAC World Congress*, vol. D, 1996, pp. 435–440.

[8] H. J. Sussmann, E. D. Sontag, and Y. Yang, "A general result on the stabilization of linear systems using bounded controls," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 2411–2425, <AUTHOR: PLEASE PROVIDE MONTH> 1994.

[9] H. J. Sussmann and Y. Yang, "On the stabilizability of multiple integrators by means of bounded feedback controls," in *Proc. 30th IEEE Conf. Decision and Control*, 1991, pp. 70–72.

[10] R. Suarez, J. Alvarez-Ramirez, and J. Solis-Daun, "Linear systems with bounded inputs: Global stabilization with eigenvalue placement," *Int. J. Robust Nonlin. Control*, vol. 7, pp. 835–845, 1997.

[11] A. R. Teel, "Global stabilization and restricted tracking for multiple integrators with bounded controls," *Syst. Control Lett.*, vol. 18, pp. 165–171, 1992.

[12] ——, "Linear systems with input nonlinearities: Global stabilization by scheduling a family of H_∞ -type controllers," *Int. J. Robust Nonlin. Control*, vol. 5, pp. 399–441, 1995.
[13] ——, "A nonlinear small gain theorem for the analysis of control systems," *IEEE Trans. Automat. Contr.*, vol. 42, pp. 1256–1270, <AUTHOR: PLEASE PROVIDE MONTH> 1996.

IEEE Proof

Publication 21

SCL 2276
pp: 1--14 (col.fig.: Nil)
PROD. TYPE: COM
ED: Mamatha
PAGN: Vidya - SCAN: Shobha

ARTICLE IN PRESS

ELSEVIER

Systems & Control Letters ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

SYSTEMS & CONTROL LETTERS

www.elsevier.com/locate/sysconle

An explicit description of null controllable regions of linear systems with saturating actuators

Tingshu Hu[a,*,1], Zongli Lin[a,1], Li Qiu[b,2]

[a] *Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903, USA*
[b] *Department of Electrical & Electronic Engineering, Hong Kong University of Science & Technology, Clear Water Bay, Kowloon, Hong Kong*

Received 25 September 2000; received in revised form 13 March 2002; accepted 16 April 2002

Abstract

We give simple exact descriptions of the null controllable regions for general linear systems with saturating actuators. The description is in terms of a set of extremal trajectories of the anti-stable subsystem. For lower order systems or systems with only real eigenvalues, this description is further simplified to result in explicit formulae for the boundaries of the null controllable regions. © 2002 Published by Elsevier Science B.V.

Keywords: ■; ■; ■

1. Introduction

One of the most fundamental issues associated with the control of a system is its controllability. Since all practical control inputs are bounded (due to actuator saturation), the constrained controllability was formulated earlier than the unconstrained one. While the unconstrained controllability has been well understood for several decades, there have been continual efforts towards full understanding of the constrained controllability (see, e.g., [1–3,7,8,10,12,16–19] and the references therein).

For a linear system with a constrained input, the null controllable region at a time $T \in (0, \infty)$, denoted as $\mathscr{C}(T)$, is defined to be the set of states that can be steered to the origin in time T with a constrained control. The union of $\mathscr{C}(T)$ for all $T \in (0, \infty)$, denoted as $\mathscr{C}$, is called the null controllable region. In the earlier studies, the null controllable region, also called the controllable set or the reachable set (of the time reversed system), was closely related to the time optimal control (e.g., [2,7,11,13]): for a given initial state x_0, the time optimal control problem has a solution if and only if $x_0 \in \mathscr{C}$; If x_0 is on the boundary of $\mathscr{C}(T)$, then the minimal time to steer x_0 to the origin is T. It is well-known that the time optimal controls are bang–bang controls. For discrete-time systems, the time optimal control can be computed through linear programming and $\mathscr{C}(T)$ can be exactly obtained, although the computational burden increases as T increases. Also closely related to the time optimal control is the model predictive control or the receding horizon control. The model predictive

* Corresponding author.
E-mail addresses: th7f@virginia.edu (Tingshu Hu), zl5y@virginia.edu (Zongli Lin), eeqiu@ee.ust.hk (Li Qiu).

[1] Work supported in part by the US Office of Naval Research Young Investigator Program under grant N00014-99-1-0670.
[2] Work supported by Hong Kong Research Grant Council under grant HKUST6046/00E.

0167-6911/02/$ - see front matter © 2002 Published by Elsevier Science B.V.
PII: S0167-6911(02)00176-7

control has been extensively studied and has found wide applications in slow processes (see, e.g., [14,15] for a survey). The development of the model predictive control also contributes to the characterization of $\mathscr{C}(T)$ for discrete-time systems. On the contrary, for continuous-time systems, since the time optimal control is generally impossible to compute except by numerical approximation, there has been no result on the explicit or analytical characterization of $\mathscr{C}(T)$ or $\mathscr{C}$ of exponentially unstable systems. There are however numerical algorithms available to obtain approximations of $\mathscr{C}$ based on some partial properties about the boundary of $\mathscr{C}$ for second-order systems (e.g., [18,19]). There are also numerical methods for testing if a particular point in the state space is inside $\mathscr{C}$ (e.g., [6]). In this paper, we will focus on the analytic characterization of $\mathscr{C}$ for general linear systems.

We recall that a linear system is said to be anti-stable if all its poles are in the open right-half plane and semi-stable if all its poles are in the closed left-half plane.

For a semi-stable linear system, it is well-known [13,16,17] that the null controllable region is the whole state-space as long as the system is controllable in the usual linear system sense. For a general system with exponentially unstable modes, there exists a nice decomposition result concerning the null controllable region [5]. Suppose such a system is decomposed into the sum of a controllable semi-stable subsystem and an anti-stable subsystem, then the null controllable region of the whole system is the Cartesian product of the null controllable region of the first subsystem, which is its whole state space, and that of the second subsystem, which is a bounded convex open set.

However, little was known about the null controllable region of an anti-stable system. This paper is dedicated to solving this problem. We will give simple exact descriptions of the null controllable region of a general anti-stable linear system with saturating actuators in terms of a set of extremal trajectories of its time reversed system. This set of extremal trajectories is particularly easy to describe for low order systems or systems with only real eigenvalues. For example, for a second-order system, the boundary of its null controllable region is covered by at most two extremal trajectories; and for a third-order system, the set of extremal trajectories can be described in terms of parameters in a real interval.

The remainder of the paper is organized as follows. Section 2 contains some preliminaries and definitions of notation. Section 3 gives a simple exact description of the null controllable regions of anti-stable linear systems with bounded controls. Section 4 draws a brief conclusion to this paper.

2. Preliminaries and notation

Consider a linear system

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{1}$$

where $x(t) \in \mathbf{R}^n$ is the state and $u(t) \in \mathbf{R}^m$ is the control. Let

$$\mathscr{U}_a = \{u: u \text{ is measurable and } \|u(t)\|_\infty \leqslant 1, \forall t \in \mathbf{R}\}, \tag{2}$$

where $\|u(t)\|_\infty = \max_i |u_i(t)|$. A control signal u is said to be *admissible* if $u \in \mathscr{U}_a$. In this paper, we are interested in the control of system (1) by using admissible controls. Our concern is the set of states that can be steered to the origin by admissible controls.

Definition 2.1. A state x_0 is said to be null controllable if there exist a $T \in [0, \infty)$ and an admissible control u such that the state trajectory $x(t)$ of the system satisfies $x(0) = x_0$ and $x(T) = 0$. The set of all null controllable states is called the null controllable region of the system and is denoted by $\mathscr{C}$.

With the above definition, we see that $x_0 \in \mathscr{C}$ if and only if there exist $T \in [0,\infty)$ and a $u \in \mathscr{U}_a$ such that

$$0 = x(T) = e^{AT}x_0 + \int_0^T e^{A(T-\tau)}Bu(\tau)\,d\tau = e^{AT}\left(x_0 + \int_0^T e^{-A\tau}Bu(\tau)\,d\tau\right).$$

It follows that

$$\mathscr{C} = \bigcup_{T\in[0,\infty)} \left\{x = -\int_0^T e^{-A\tau}Bu(\tau)\,d\tau:\ u \in \mathscr{U}_a\right\}. \tag{3}$$

The minus sign "−" before the integral can be removed since $\mathscr{U}_a$ is a symmetric set. In what follows we recall from the literature some existing results on the characterization of the null controllable region.

Proposition 2.1. *Assume that (A,B) is controllable.*

(a) *If A is semi-stable, then $\mathscr{C} = \mathbf{R}^n$.*

(b) *If A is anti-stable, then $\mathscr{C}$ is a bounded convex open set containing the origin.*

(c) *If*

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$$

with $A_1 \in \mathbf{R}^{n_1\times n_1}$ anti-stable and $A_2 \in \mathbf{R}^{n_2\times n_2}$ semi-stable, and B is partitioned as

$$\begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$$

accordingly, then $\mathscr{C} = \mathscr{C}_1 \times \mathbf{R}^{n_2}$ where $\mathscr{C}_1$ is the null controllable region of the anti-stable system $\dot{x}_1(t) = A_1x_1 + B_1u(t)$.

Statement (a) is well-known [13,16,17]. Statements (b) and (c) are proven in [5]. Because of this proposition, we can concentrate on the study of null controllable regions of anti-stable systems. For this kind of systems,

$$\bar{\mathscr{C}} = \left\{x = \int_0^\infty e^{-A\tau}Bu(\tau)\,d\tau:\ u \in \mathscr{U}_a\right\}, \tag{4}$$

where $\bar{\mathscr{C}}$ denotes the closure of $\mathscr{C}$. We will use "∂" to denote the boundary of a set. In [6], a nonlinear programming based algorithm is proposed to test if a point in the state space belongs to $\mathscr{C}$. In Section 3, we will derive a method for explicitly describing $\partial\mathscr{C}$. To this end, we will need some more preliminaries.

If $B = [b_1, b_2, \ldots, b_m]$ and the null controllable region of the system $\dot{x}(t) = Ax(t) + b_iu_i(t)$ is $\mathscr{C}_i$, $i = 1,\ldots,m$, then

$$\mathscr{C} = \sum_{i=1}^m \mathscr{C}_i = \{x_1 + x_2 + \cdots + x_m:\ x_i \in \mathscr{C}_i,\ i = 1,2,\ldots,m\}. \tag{5}$$

In view of (5) and Proposition 2.1, in the study of the null controllable regions we will assume, without loss of generality, that (A,B) is controllable, A is anti-stable, and $m = 1$. For clarity, we rename B as b.

For a general system

$$\dot{x} = f(x,u), \tag{6}$$

its time reversed system is

$$\dot{z} = -f(z,v). \tag{7}$$

It is easy to see that $x(t)$ solves (6) with $x(0) = x_0, x(t_1) = x_1$, and certain u if and only if $z(t) = x(t_1 - t)$ solves (7) with $z(0) = x_1$, $z(t_1) = x_0$, and $v(t) = u(t_1 - t)$. The two systems have the same curves as trajectories, but traverse in opposite directions.

Consider the time reversed system of (1):

$$\dot{z}(t) = -Az(t) - bv(t). \tag{8}$$

Definition 2.2. A state z_f is said to be reachable if there exist $T \in [0, \infty)$ and an admissible control v such that the state trajectory $z(t)$ of system (8) satisfies $z(0) = 0$ and $z(T) = z_f$. The set of all reachable states is called the reachable region of system (8) and is denoted by $\mathscr{R}$.

It is known that $\mathscr{C}$ of (1) is the same as $\mathscr{R}$ of (8) (see, e.g., [13]). To avoid confusion, we will continue to use the notation x, u and $\mathscr{C}$ for the original system (1), and z, v and $\mathscr{R}$ for the time-reversed system (8).

3. Null controllable regions

In Section 3.1, we show that the boundary of the null controllable region of a general anti-stable linear system with saturating actuator is composed of a set of extremal trajectories of the time reversed system. The descriptions of this set are further simplified for systems with only real poles and for systems with complex poles in Sections 3.2 and 3.3, respectively.

3.1. Description of the null controllable regions

We will characterize the null controllable region $\mathscr{C}$ of system (1) through studying the reachable region $\mathscr{R}$ of its time reversed system (8).

Since A is anti-stable, we have

$$\bar{\mathscr{R}} = \left\{ z = \int_0^\infty e^{-A\tau} b v(\tau)\, d\tau : v \in \mathscr{U}_a \right\} = \left\{ z = \int_{-\infty}^0 e^{A\tau} b v(\tau)\, d\tau : v \in \mathscr{U}_a \right\}.$$

The change of integration interval from $[0, \infty]$ to $[-\infty, 0]$ is crucial to our development, as will be clear from Eq. (17). Noticing that $e^{A\tau} = e^{-A(0-\tau)}$, we see that a point z in $\bar{\mathscr{R}}$ is a state of the time-reversed system (8) at $t = 0$ by applying an admissible control v from $-\infty$ to 0.

Theorem 3.1.

$$\partial\mathscr{R} = \left\{ z = \int_{-\infty}^0 e^{A\tau} b \operatorname{sgn}(c' e^{A\tau} b)\, d\tau : c \neq 0 \right\}. \tag{9}$$

$\bar{\mathscr{R}}$ is strictly convex. Moreover, for each $z^* \in \partial\mathscr{R}$, there exists a unique admissible control v^* such that

$$z^* = \int_{-\infty}^0 e^{A\tau} b v^*(\tau)\, d\tau. \tag{10}$$

Remark 3.1. We give some simple facts about convex sets in this remark. Consider a closed set S. If S is convex and $z^* \in \partial S$, then by separation theorem, there exists a hyperplane $c'z = k$ that is tangential to ∂S at z^* and the set S lies completely to one side of the hyperplane, i.e.,

$$c'z \leqslant k = c'z^*, \quad \forall z \in S.$$

A set S is said to be strictly convex if it is convex and for any two points $z_1, z_2 \in \partial S$, $\alpha z_1 + (1-\alpha) z_2 \notin \partial S$ for all $\alpha \in (0, 1)$. This is equivalent to saying that any hyperplane that is tangential to ∂S has only one intersection point with ∂S, or, for any $c \neq 0$, there exists a unique $z^* \in \partial S$ such that $c'z^* = \max_{z \in S} c'z$.

Proof of Theorem 1. First, the convexity of $\bar{\mathscr{R}}$ can easily be verified by definition. Let $z^* \in \partial\mathscr{R}$. Then, there exists a nonzero vector $c \in \mathbf{R}^n$ such that

$$c'z^* = \max_{z\in\bar{\mathscr{R}}} c'z = \max_{v\in\mathscr{U}_a} \int_{-\infty}^{0} c'e^{A\tau}bv(\tau)\,d\tau. \tag{11}$$

Since $c \neq 0$ and (A, b) is controllable, $c'e^{At}b \not\equiv 0$. Since $c'e^{At}b$ has a finite number of zeros in any finite interval,

$$\mu(\{t\colon c'e^{At}b = 0\}) = 0, \tag{12}$$

where $\mu(\cdot)$ denotes the measure of a set.

It is easy to see that

$$v^*(t) = \operatorname{sgn}(c'e^{At}b)$$

maximizes the right-hand side of (11). We maintain that v^* is the unique optimal solution of (11). To verify this, we need to show that for any $v \in \mathscr{U}_a, v \neq v^*$,

$$\int_{-\infty}^{0} c'e^{A\tau}bv^*(\tau)\,d\tau > \int_{-\infty}^{0} c'e^{A\tau}bv(\tau)\,d\tau. \tag{13}$$

Since $v \neq v^*$, there are a set $E_1 \subset [-\infty, 0]$ with nonzero measure, i.e., $\mu(E_1) = \delta_1 > 0$, and a number $\varepsilon_1 > 0$ such that

$$|v(t) - v^*(t)| \geqslant \varepsilon_1, \quad \forall t \in E_1.$$

By (12), there exist a set $E \subset E_1$, with $\mu(E) = \delta > 0$, and a positive number $\varepsilon > 0$ such that

$$|c'e^{At}b| \geqslant \varepsilon, \quad \forall t \in E.$$

Noting that $v \in \mathscr{U}_a$, we have

$$c'e^{At}b\,(v^*(t) - v(t)) \geqslant 0, \quad \forall t \in [-\infty, 0].$$

It then follows that

$$\int_{-\infty}^{0} c'e^{A\tau}b(v^*(\tau) - v(\tau))\,d\tau$$

$$\geqslant \int_{E} c'e^{A\tau}b(v^*(\tau) - v(t))\,d\tau = \int_{E} |c'e^{A\tau}b|\,|v^*(\tau) - v(\tau)|\,d\tau \geqslant \delta\varepsilon\varepsilon_1 > 0.$$

This shows that $v^*(t)$ is the unique optimal solution of (11) and hence the unique admissible control satisfying

$$z^* = \int_{-\infty}^{0} e^{A\tau}bv^*(\tau)\,d\tau. \tag{14}$$

On the other hand, if

$$z^* = \int_{-\infty}^{0} e^{A\tau}b\operatorname{sgn}(c'e^{A\tau}b)\,d\tau$$

for some nonzero c, then obviously

$$c'z^* = \max_{z\in\bar{\mathscr{R}}} c'z.$$

This shows that $z^* \in \partial\mathscr{R}$ and we have (9).

Since for each $c \neq 0$, the optimal solution $v^*(t)$ and z^* of (11) is unique, we see that $\bar{\mathscr{R}}$ is strictly convex. □

UNCORRECTED PROOF

Theorem 3.1 says that for $z^* \in \partial\mathscr{R}$, there is a unique admissible control v^* satisfying (10). From (9), this implies $v^*(t) = \operatorname{sgn}(c'e^{At}b)$ for some $c \neq 0$ (such c, $\|c\| = 1$ may be nonunique, where $\|\cdot\|$ is the Euclidean norm). So, if v is an admissible control and there is no c such that $v(t) = \operatorname{sgn}(c'e^{At}b)$ for $t \leqslant 0$, then

$$\int_{-\infty}^{0} e^{A\tau}bv(\tau)\,d\tau \notin \partial\mathscr{R}$$

and must be in the interior of $\mathscr{R}$.

Since $\operatorname{sgn}(kc'e^{A\tau}b) = \operatorname{sgn}(c'e^{A\tau}b)$ for any positive number k, Eq. (9) shows that $\partial\mathscr{R}$ can be determined from the surface of a unit ball in $\mathbf{R}^n$. In what follows, we will simplify (9) and describe $\partial\mathscr{R}$ in terms of a set of trajectories of the time-reversed system (8).

Denote

$$\mathscr{E} := \{v(t) = \operatorname{sgn}(c'e^{At}b),\ t \in \mathbf{R} : c \neq 0\} \tag{15}$$

and for an admissible control v, denote

$$\Phi(t, v) := \int_{-\infty}^{t} e^{-A(t-\tau)}bv(\tau)\,d\tau. \tag{16}$$

Since A is anti-stable, the integral in (16) exists for all $t \in \mathbf{R}$, so $\Phi(t, v)$ is well defined.

If $v(t) = \operatorname{sgn}(c'e^{At}b)$, then

$$\Phi(t, v) = \int_{-\infty}^{t} e^{-A(t-\tau)}bv(\tau)\,d\tau = \int_{-\infty}^{0} e^{A\tau}b\operatorname{sgn}(c'e^{At}e^{A\tau}b)\,d\tau \in \partial\mathscr{R} \tag{17}$$

for any $t \in \mathbf{R}$, i.e., $\Phi(t, v)$ lies entirely on $\partial\mathscr{R}$. An admissible control v such that $\Phi(t, v)$ lies entirely on $\partial\mathscr{R}$ is said to be *extremal* and such $\Phi(t, v)$ an *extremal trajectory*. On the other hand, given an admissible control $v(t)$, if there exists no c such that $v(t) = \operatorname{sgn}(c'e^{At}b)$ for all $t \leqslant 0$, then by Theorem 3.1, $\Phi(0, v) \notin \partial\mathscr{R}$ and must be in the interior of $\mathscr{R}$. By the time invariance property of the system, if there exists no c such that $v(t) = \operatorname{sgn}(c'e^{At}b)$ for all $t \leqslant t_0$, $\Phi(t, v)$ must be in the interior of $\mathscr{R}$ for all $t \geqslant t_0$. Consequently, $\mathscr{E}$ is the set of extremal controls.

The following lemma shows that $\partial\mathscr{R}$ is covered by the set of extremal trajectories.

Lemma 3.1.

$$\partial\mathscr{R} = \{\Phi(t, v)\colon t \in \mathbf{R}, v \in \mathscr{E}\}. \tag{18}$$

Proof. For any fixed $t \in \mathbf{R}$, it follows from (9) that

$$\partial\mathscr{R} = \left\{\int_{-\infty}^{t} e^{-A(t-\tau)}b\operatorname{sgn}(c'e^{-At}e^{A\tau}b)\,d\tau\colon c \neq 0\right\} = \left\{\int_{-\infty}^{t} e^{-A(t-\tau)}b\operatorname{sgn}(c'e^{A\tau}b)\,d\tau\colon c \neq 0\right\},$$

i.e., $\partial\mathscr{R} = \{\Phi(t, v)\colon v \in \mathscr{E}\}$, for any fixed $t \in \mathbf{R}$. So $\partial\mathscr{R}$ can be viewed as the set of extremal trajectories at any frozen time. Now let t vary, then each point on $\partial\mathscr{R}$ moves along a trajectory but the whole set is invariant. So we can also write $\partial\mathscr{R} = \{\Phi(t, v)\colon v \in \mathscr{E},\ t \in \mathbf{R}\}$, which is equivalent to (18). □

Unlike (9), Eq. (18) shows that $\partial\mathscr{R}$ is covered by extremal trajectories. It, however, introduces redundancy by repeating the same set $\{\Phi(t, v)\colon v \in \mathscr{E}\}$ for all $t \in \mathbf{R}$. This redundancy can be removed by a careful examination of the set $\mathscr{E}$. Indeed, the set $\{\Phi(t, v)\colon t \in \mathbf{R}\}$ can be identical for a class of $v \in \mathscr{E}$.

Definition 3.1.

(a) Two extremal controls $v_1, v_2 \in \mathscr{E}$ are said to be equivalent, denoted by $v_1 \sim v_2$, if there exists an $h \in \mathbf{R}$ such that $v_1(t) = v_2(t - h)$ for all $t \in \mathbf{R}$.

UNCORRECTED PROOF

(b) Two vectors $c_1, c_2 \in \mathbf{R}^n$ are said to be equivalent, denoted by $c_1 \sim c_2$, if there exist a $k > 0$ and an $h \in \mathbf{R}$ such that $c_1 = ke^{A'h}c_2$.

Noting that a shift in time of the control corresponds to the same shift of the state trajectory, we see that, if $v_1 \sim v_2$, then $\{\Phi(t, v_1)\colon t \in \mathbf{R}\} = \{\Phi(t, v_2)\colon t \in \mathbf{R}\}$; and if $c_1 \sim c_2$, then $\operatorname{sgn}(c_1' e^{At} b) \sim \operatorname{sgn}(c_2' e^{At} b)$.

Definition 3.2.
(a) A set $\mathscr{E}_{\min} \subset \mathscr{E}$ is called a minimal representative of $\mathscr{E}$ if for any $v \in \mathscr{E}$, there exists a unique $v_1 \in \mathscr{E}_{\min}$ such that $v \sim v_1$.
(b) A set $M \subset \mathbf{R}^n$ is called a minimal representative of $\mathbf{R}^n$ if for any $c \in \mathbf{R}^n$, there exists a unique $c_1 \in M$ such that $c \sim c_1$.

With this definition, there will be no pair of distinct elements in $\mathscr{E}_{\min}$ or in M that are equivalent. It should be noted that the minimal representative of $\mathscr{E}$ or $\mathbf{R}^n$ is unique up to equivalence and $\mathscr{E}_{\min}$ and M always exist. An immediate consequence of these definitions and Lemma 3.1 is

Theorem 3.2. *If $\mathscr{E}_{\min}$ is a minimal representative of $\mathscr{E}$, then*

$$\partial\mathscr{R} = \{\Phi(t, v)\colon t \in \mathbf{R}, v \in \mathscr{E}_{\min}\};$$

If M is a minimal representative of $\mathbf{R}^n$, then

$$\partial\mathscr{R} = \{\Phi(t, \operatorname{sgn}(c' e^{At} b))\colon t \in \mathbf{R}, c \in M \backslash \{0\}\}.$$

It turns out that for some classes of systems, $\mathscr{E}_{\min}$ can be easily described. For second order systems, $\mathscr{E}_{\min}$ contains only one or two elements, so $\partial\mathscr{R}$ can be covered by no more than two trajectories; and for third-order systems, $\mathscr{E}_{\min}$ corresponds to some real intervals. We will see later that for systems of different eigenvalue structures, the descriptions of $\mathscr{E}_{\min}$ can be quite different.

3.2. Systems with only real eigenvalues

It follows from, for example, [13, p. 77], that if A has only real eigenvalues and $c \neq 0$, then $c' e^{At} b$ has at most $n-1$ zeros. This implies that an extremal control can have at most $n-1$ switches. We will show that the converse is also true.

Theorem 3.3. *For system* (8), *assume that A has only real eigenvalues, then*
(a) *an extremal control has at most $n-1$ switches;*
(b) *any bang–bang control with $n-1$ or less switches is an extremal control.*

Proof. See Appendix A. □

By Theorem 3.3, the set of extremal controls can be described as follows:

$$\mathscr{E} = \left\{ \pm v\colon v(t) = \begin{cases} 1 & -\infty \leqslant t < t_1, \\ (-1)^i & t_i \leqslant t < t_{i+1}, \\ (-1)^{n-1}, & t_{n-1} \leqslant t < \infty, \end{cases} \quad -\infty < t_1 < t_2 \leqslant \cdots \leqslant t_{n-1} \leqslant \infty \right\} \cup \{v(t) \equiv \pm 1\},$$

where $t_i, i = 1, \ldots, n-1$, are the switching times. If $v(t)$ has a switch, then the first switch occurs at $t = t_1$. Here we allow $t_i = t_{i+1}$ $(i \neq 1)$ and $t_{n-1} = \infty$, so the above description of $\mathscr{E}$ consists of all bang–bang controls with $n-1$ or less switches.

To obtain a minimal representative of $\mathscr{E}$, we can simply set $t_1 = 0$, that is,

$$\mathscr{E}_{\min} = \left\{ \pm v\colon v(t) = \begin{cases} 1, & -\infty \le t < t_1, \\ (-1)^i, & t_i \leqslant t < t_{i+1}, \quad 0 = t_1 < t_2 \leqslant \cdots \leqslant t_{n-1} \leqslant \infty \\ (-1)^{n-1}, & t_{n-1} \leqslant t < \infty. \end{cases} \right\} \cup \{v(t) \equiv \pm 1\}.$$

For every $v \in \mathscr{E}_{\min}$, we have $v(t) = 1$ (or -1) for all $t < 0$. Hence, for $t \leqslant 0$,

$$\Phi(t, v) = \int_{-\infty}^{t} e^{-A(0-\tau)} b \, d\tau = A^{-1}b \text{ (or } -A^{-1}b).$$

Afterwards, $v(t)$ is a bang–bang control with $n-2$ or less switches. Denote $z_e^+ = -A^{-1}b$ and $z_e^- = A^{-1}b$, then from Theorem 3.2 we have,

Observation 3.1. *$\partial\mathscr{R} = \partial\mathscr{C}$ is covered by two bunches of trajectories. The first bunch consists of trajectories of* (8) *whose initial state is z_e^+ and the input is a bang–bang control that starts at $t = 0$ with -1 and has $n-2$ or less switches. The second bunch consists of the trajectories of* (8) *whose initial state is z_e^- and the input is a bang–bang control that starts at $t = 0$ with 1 and has $n-2$ or less switches.*

Furthermore, $\partial\mathscr{R}$ can be simply described in terms of the open-loop transition matrix. Note that for a fixed $t \geqslant 0$,

$$\begin{aligned} &\{\Phi(t, v)\colon v \in \mathscr{E}_{\min}\} \\ &\quad = \left\{ \pm \left[e^{-At} z_e^+ - \sum_{i=1}^{n-1} \int_{t_i}^{t_{i+1}} e^{-A(t-\tau)} b(-1)^i \, d\tau \right] : 0 = t_1 < t_2 < \cdots \leqslant t_{n-1} \leqslant t_n = t \right\} \cup \{\pm z_e^+\} \\ &\quad = \left\{ \pm \left[\sum_{i=1}^{n-1} 2(-1)^i e^{-A(t-t_i)} + (-1)^n I \right] A^{-1}b\colon 0 = t_1 < t_2 < \cdots \leqslant t_{n-1} \leqslant t \right\} \cup \{\pm z_e^+\}. \end{aligned}$$

Hence,

$$\begin{aligned} \partial\mathscr{R} &= \{\Phi(t, v)\colon t \in \mathbf{R}, v \in \mathscr{E}_{\min}\} \\ &= \left\{ \pm \left[\sum_{i=1}^{n-1} 2(-1)^i e^{-A(t-t_i)} + (-1)^n I \right] A^{-1}b\colon 0 = t_1 \leqslant t_2 \leqslant \cdots \leqslant t_{n-1} \leqslant t \leqslant \infty \right\}. \end{aligned}$$

Here, we allow $t_1 = t_2$ to include $\pm z_e^+$. For second-order systems,

$$\partial\mathscr{R} = \left\{ \pm \left[e^{-At} z_e^- - \int_0^t e^{-A(t-\tau)} b \, d\tau \right] : t \in [0, \infty] \right\} = \{\pm(-2e^{-At} + I)A^{-1}b\colon t \in [0, \infty]\}. \tag{19}$$

Plotted in Fig. 1 is the $\partial\mathscr{R}$ of a second-order system with

$$A = \begin{bmatrix} 0 & -0.5 \\ 1 & 1.5 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

We see that $\partial\mathscr{R}$ consists of a trajectory from z_e^+ to z_e^- under the constant control $v = -1$ and a trajectory from z_e^- to z_e^+ under the constant control $v = 1$.

If $n = 3$, then one half of $\partial\mathscr{R} = \partial\mathscr{C}$ can be formed by the trajectories of (8) starting from z_e^+ with the control initially being -1 and then switching at any time to 1. So the trajectories go toward z_e^- at first then turn back toward z_e^+. The other half is just symmetric to the first half. That is

$$\partial\mathscr{R} = \left\{ \pm \left[e^{-At} z_e^+ + \int_0^{t_2} e^{-A(t-\tau)} b \, d\tau - \int_{t_2}^{t} e^{-A(t-\tau)} b \, d\tau \right] : 0 \leqslant t_2 \leqslant t \leqslant \infty \right\}. \tag{20}$$

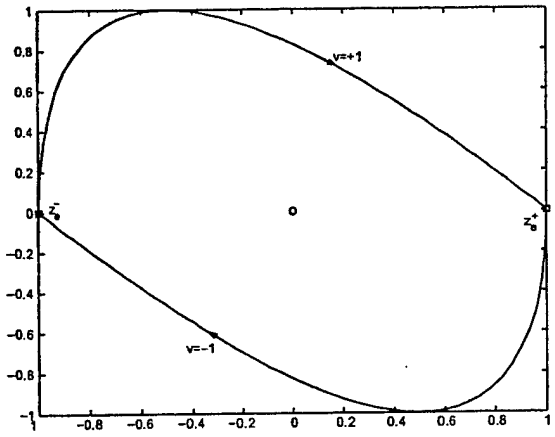


Fig. 1. $\partial\mathscr{R}$ of a second order system.

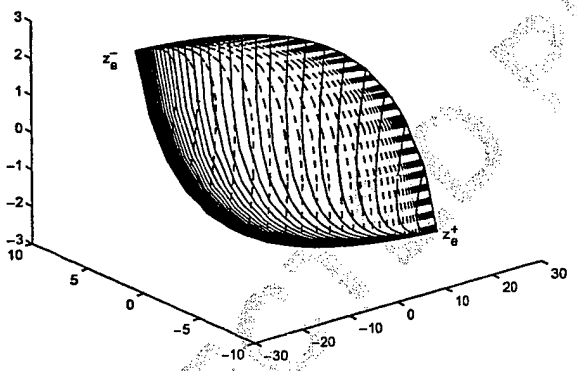


Fig. 2. $\partial\mathscr{R}$ of a third-order system.

Plotted in Fig. 2 is the $\partial\mathscr{R}$ of a third-order system with

$$A=\begin{bmatrix}0.2 & 1 & 0\\ 0 & 0.2 & 0\\ 0 & 0 & 0.4\end{bmatrix},\quad B=\begin{bmatrix}1\\1\\1\end{bmatrix}.$$

Since the trajectories of the original system and those of the time-reversed system are the same but traverse in opposite directions, we can also say that $\partial\mathscr{R}=\partial\mathscr{C}$ is covered by a set of trajectories of the original system. While all the trajectories of the time-reversed system start at z_e^+ or z_e^- and are very easy to generate by simulation, it is impossible to obtain the same trajectories from the original system. For example, when $n=2$, one half of $\partial\mathscr{R}$ is formed by the trajectory of the time-reversed system that starts at z_e^- under a constant control $v=+1$. The trajectory goes from z_e^- toward z_e^+ asymptotically but never reaches z_e^+ at a finite time. It seems that if we apply $u=+1$ at z_e^+ to the original system, the trajectory will go from z_e^+ to z_e^- along the same trajectory of the time-reversed system. However, this is not the case. The trajectory of the original system will stay at z_e^+ under the constant control $u=+1$. The boundary $\partial\mathscr{R}$ can only be partially generated from the original system if we know one point on it other than $\pm z_e^+$. But this point is not easy to determine.

3.3. Systems with complex eigenvalues

For a system with complex eigenvalues, the minimal representative set $\mathscr{E}_{\min}$ is harder to determine. In what follows, we consider two important cases.

Case 1. $A \in \mathbf{R}^{2\times 2}$ *has a pair of complex eigenvalues* $\alpha \pm \mathrm{j}\beta$, $\alpha, \beta > 0$.

In order to arrive at an explicit formula for $\mathscr{C}$, we need to simplify $c'\mathrm{e}^{At}b$. To this end, let V be the nonsingular matrix such that

$$A = V \begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix} V^{-1}$$

and let

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = V'c, \quad \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = V^{-1}b,$$

then

$$c'\mathrm{e}^{At}b = [c_1 \quad c_2] \begin{bmatrix} \cos(\beta t) & -\sin(\beta t) \\ \sin(\beta t) & \cos(\beta t) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \mathrm{e}^{\alpha t}$$

$$= [\cos(\beta t) \quad \sin(\beta t)] \begin{bmatrix} b_1 & b_2 \\ -b_2 & b_1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \mathrm{e}^{\alpha t}.$$

Since

$$\begin{bmatrix} b_1 & b_2 \\ -b_2 & b_1 \end{bmatrix}$$

is nonsingular, it follows that

$$\left\{ \begin{bmatrix} b_1 & b_2 \\ -b_2 & b_1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} : c \neq 0 \right\} = \left\{ r \begin{bmatrix} \sin(\theta) \\ \cos(\theta) \end{bmatrix} : r \neq 0, \theta \in [0, 2\pi) \right\}.$$

Hence

$$\{\mathrm{sgn}(c'\mathrm{e}^{At}b) : c \neq 0\} = \{\mathrm{sgn}(\sin(\beta t + \theta)) : \theta \in [0, 2\pi)\},$$

and the set of extremal controls is

$$\mathscr{E} = \{v(t) = \mathrm{sgn}(\sin(\beta t + \theta)), t \in \mathbf{R} : \theta \in [0, 2\pi)\}.$$

It is easy to see that

$$\mathscr{E}_{\min} = \{v(t) = \mathrm{sgn}(\sin(\beta t)), t \in \mathbf{R}\}$$

contains only one element. Denote $T_p = \pi/\beta$, then $\mathrm{e}^{-AT_p} = -\mathrm{e}^{-\alpha T_p} I$. Let

$$z_s^- = (I + \mathrm{e}^{-AT_p})^{-1}(I - \mathrm{e}^{-AT_p})A^{-1}b = \frac{1 + \mathrm{e}^{-\alpha T_p}}{1 - \mathrm{e}^{-\alpha T_p}} z_e^-. \tag{21}$$

It can be verified that the extremal trajectory corresponding to $v(t) = \mathrm{sgn}(\sin(\beta t))$ is periodic with period $2T_p$ and,

$$\partial\mathscr{R} = \left\{ \pm \left[\mathrm{e}^{-At} z_s^- - \int_0^t \mathrm{e}^{-A(t-\tau)} b \, \mathrm{d}\tau \right] : t \in [0, T_p) \right\}$$

$$= \{\pm[\mathrm{e}^{-At} z_s^- - (I - \mathrm{e}^{-At})A^{-1}b] : t \in [0, T_p)\}. \tag{22}$$

CORRECTED PROOF

Case 2: $A \in \mathbf{R}^{3\times 3}$ *has eigenvalues* $\alpha \pm j\beta$ *and* α_1, *with* $\alpha, \beta, \alpha_1 > 0$. (a) $\alpha = \alpha_1$. Then similar to Case 1,

$$\mathscr{E} = \{v(t) = \operatorname{sgn}(k + \sin(\beta t + \theta)), t \in \mathbf{R} : k \in \mathbf{R}, \theta \in [0, 2\pi)\}.$$

Since $\operatorname{sgn}(k + \sin(\beta t + \theta))$ is the same for all $k \geqslant 1$ (or $k \leqslant -1$), we have

$$\mathscr{E}_{\min} = \{v(t) = \operatorname{sgn}(k + \sin(\beta t)), t \in \mathbf{R} : k \in [-1, 1]\}.$$

Each $v \in \mathscr{E}_{\min}$ is periodic with period $2T_p$, but the lengths of positive and negative parts vary with k. $\Phi(t, v)$ can be easily determined from simulation.

(b) $\alpha_1 \neq \alpha$. Then

$$\mathscr{E} = \{v(t) = \operatorname{sgn}(k_1 e^{(\alpha_1 - \alpha)t} + k_2 \sin(\beta t + \theta)), t \in \mathbf{R} : (k_1, k_2) \neq (0, 0), \theta \in [0, 2\pi)\}.$$

$\mathscr{E}$ can be decomposed as $\mathscr{E} = \mathscr{E}_1 \cup \mathscr{E}_2 \cup \mathscr{E}_3$, where

$$\mathscr{E}_1 = \{v(t) \equiv \pm 1\} \quad (k_2 = 0),$$

$$\mathscr{E}_2 = \{v(t) = \pm \operatorname{sgn}(\sin(\beta t + \theta)) : \theta \in [0, 2\pi)\} \quad (k_1 = 0),$$

$$\mathscr{E}_3 = \{v(t) = \pm \operatorname{sgn}(k e^{(\alpha_1 - \alpha)t} + \sin(\beta t + \theta)) : k > 0, \theta \in [0, 2\pi)\}.$$

We will show that a minimal representative of $\mathscr{E}_3$ is

$$\mathscr{E}_{3\,\min} = \{v(t) = \pm \operatorname{sgn}(e^{(\alpha_1 - \alpha)t} + \sin(\beta t + \theta)) : \theta \in [0, 2\pi)\}. \tag{23}$$

Let

$$v(t) = \operatorname{sgn}(k e^{(\alpha_1 - \alpha)t} + \sin(\beta t + \theta)) \in \mathscr{E}_3.$$

Since $k > 0$, there is a number $h \in \mathbf{R}$ such that $e^{(\alpha_1 - \alpha)h} = k$. So

$$v(t) = \operatorname{sgn}(e^{(\alpha_1 - \alpha)(t+h)} + \sin(\beta(t + h) - \beta h + \theta)) = v_1(t + h)$$

for some $v_1(t) \in \mathscr{E}_{3\,\min}$. On the other hand, given $v_1, v_2 \in \mathscr{E}_{3\,\min}$, suppose

$$v_1(t) = \operatorname{sgn}(e^{(\alpha_1 - \alpha)t} + \sin(\beta t + \theta_1))$$

$$v_2(t) = \operatorname{sgn}(e^{(\alpha_1 - \alpha)t} + \sin(\beta t + \theta_2))$$

and $v_1 \sim v_2$, i.e., $v_1(t) = v_2(t - h)$ for some h, then

$$\operatorname{sgn}(e^{(\alpha_1 - \alpha)t} + \sin(\beta t + \theta_1)) = \operatorname{sgn}(e^{(\alpha_1 - \alpha)(t-h)} + \sin(\beta(t - h) + \theta_2)).$$

If $\alpha_1 < \alpha$ (or $\alpha_1 > \alpha$), both $e^{(\alpha_1 - \alpha)t}$ and $e^{(\alpha_1 - \alpha)(t-h)}$ go to zero as t goes to ∞ (or $-\infty$). For $v_1(t)$ and $v_2(t - h)$ to change signs at the same time, we must have $\beta t + \theta_1 = \beta(t - h) + \theta_2 + l\pi$, for some integer l. Since at any switching time of $v_1(t)$ and $v_2(t)$, $\sin(\beta t + \theta_1) < 0$, $\sin(\beta(t - h) + \theta_2) < 0$, we conclude that $\sin(\beta t + \theta_1) = \sin(\beta(t - h) + \theta_2)$ and $e^{(\alpha_1 - \alpha)t} = e^{(\alpha_1 - \alpha)(t-h)}$. So we get $h = 0, \theta_1 = \theta_2$. These shows that $\mathscr{E}_{3\,\min}$ is a minimal representative of $\mathscr{E}_3$.

The minimal representative of $\mathscr{E}_2$ is the same as $\mathscr{E}_{\min}$ in Case 1. It follows that

$$\mathscr{E}_{\min} = \{v(t) \equiv \pm 1\} \cup \{v(t) = \operatorname{sgn}(\sin(\beta t))\} \cup \mathscr{E}_{3\,\min}.$$

If $\alpha_1 < \alpha$, for each $v \in \mathscr{E}_{3\,\min}$, $v(t) = 1$(or -1) for all $t \leqslant 0$, so the corresponding extremal trajectories stay at $z_e^+ = -A^{-1}b$ or z_e^- before $t = 0$. And after some time, they go toward the periodic trajectory since as t goes

Corrected Proof

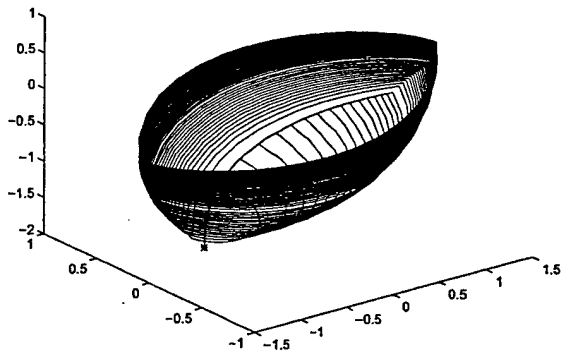


Fig. 3. Extremal trajectories on $\partial\mathscr{R}$, $\alpha_1 < \alpha$.

to infinity, $v(t)$ becomes periodic; When $\alpha_1 > \alpha$, for each $v \in \mathscr{E}_{3\,\min}$, $v(t) = 1(\text{or } -1)$ for all $t \geqslant 0$, and the corresponding extremal trajectories start from near periodic and go toward z_e^+ or z_e^-.

Plotted in Fig. 3 are some extremal trajectories on $\partial\mathscr{R}$ of the time-reversed system (8) with

$$A = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.8 & -2 \\ 0 & 2 & 0.8 \end{bmatrix}; \quad B = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

which has two complex poles.

For higher order systems, the relative location of the eigenvalues are more diversified and the analysis will be technically much more involved. It can, however, be expected that in the general case, the number of parameters used to describe $\mathscr{E}_{\min}$ is $n-2$.

4. Conclusions

We gave a clear understanding of the null controllable regions of general linear systems with saturating actuators. We showed that the boundary of the null controllable region of an anti-stable linear system is composed of a set of extremal trajectories of its time-reversed system. The description of the boundary of the null controllable region is further simplified for lower-order systems and systems that have only real eigenvalues.

5. Uncited references

[4,9]

Appendix A. Proof of Theorem 3.3

First we present a lemma. Let us use $\mathscr{P}_k$ to denote the set of real polynomials with degree less than integer k. The number 0 is considered a polynomial with arbitrary degree or with degree -1.

UNCORRECTED PROOF

Lemma A.1. *Given N positive integers, $k_1, k_2, \ldots, k_N$, define a set of functions*

$$\mathscr{G}_N := \left\{ g(t) = \sum_{i=1}^{N} e^{a_i t} f_i(t) \colon f_i \in \mathscr{P}_{k_i}, a_i \in \mathbf{R}, g(t) \not\equiv 0 \right\}.$$

Then $g(t) \in \mathscr{G}_N$ has at most $\sum_{i=1}^{N} k_i - 1$ zeros.

Proof. We prove this lemma by induction. It is easy to see that the statement is true when $N = 1$. Now assume that it is true when N is replaced by $N-1$. Let $g(t) \in \mathscr{G}_N$. Suppose on the contrary that g has $\sum_{i=1}^{N} k_i$ or more zeros. Then $\tilde{g}(t) = g(t)e^{-a_N t}$ also has $\sum_{i=1}^{N} k_i$ or more zeros. Therefore, the k_Nth derivative of $\tilde{g}$,

$$[\tilde{g}(t)]^{(k_N)} = \left[\sum_{i=1}^{N-1} e^{(a_i - a_N)t} f_i(t) + f_N(t)\right]^{(k_N)} = \left[\sum_{i=1}^{N-1} e^{(a_i - a_N)t} f_i(t)\right]^{(k_N)} \in \mathscr{G}_{N-1},$$

has at least $\sum_{i=1}^{N-1} k_i$ zeros, which is a contradiction. □

Proof of Theorem 3.3. . The proof of (a) was sketched in [13, p. 77], where an additional assumption of normality was required. This assumption is satisfied for system (8) since it is single input and (A, b) is controllable. To show (b), assume that A has N distinct real eigenvalues $\lambda_i, i = 1, 2, \ldots, N$, each with a multiplicity of k_i ($\sum_{i=1}^{N} k_i = n$). It is well-known that $c'e^{At}b = \sum_{i=1}^{N} e^{\lambda_i t} f_i(t)$ for some $f_i \in \mathscr{P}_{k_i}$. If $c \neq 0$, then $c'e^{At}b \not\equiv 0$ by the controllability of (A, b). (Thus (a) follows from Lemma A.1). To complete the proof of (b), we first show that any bang–bang control v with $n-1$ switches is an extremal control.

Let $t_1, t_2, \ldots, t_{n-1} \in \mathbf{R}$ be distinct switching times of v. From the following $n-1$ linear equations

$$c'e^{At_i}b = 0, \quad i = 1, 2, \ldots, n-1$$

at least one nonzero vector $c \in \mathbf{R}^n$ can be solved. With any such c, (a) implies $g(t) = c'e^{At}b \not\equiv 0$ has no other zeros than the $n-1$ zeros at $t_i, i = 1, 2, \ldots, n-1$.

Now the question is whether $g(t)$ indeed changes signs at each t_i. If it does, then $v(t) = \operatorname{sgn}(c'e^{At}b)$ (or $\operatorname{sgn}(-c'e^{At}b)$) and v is an extremal control.

We now show that g does change signs at each t_i. If g does not change sign at a certain t_i, then $g(t)$ must have a local extremum at t_i, so $\dot{g}(t_i) = 0$. We argue that there is at most one t_i such that $\dot{g}(t_i) = 0$, otherwise $\dot{g}$ will have at least n zeros, counting the at least $n-2$ ones lying within the intervals (t_i, t_{i+1}), which is impossible by Lemma A.1, since $\dot{g}$ has the same structure as g.

We further conclude that g, however, cannot have a local extremum at any of these t_i's.

Let $g(t) = \sum_{i=1}^{N} e^{\lambda_i t} f_i(t)$. Assume without loss of generality that $f_N(t) \not\equiv 0$. Suppose on the contrary that g has a local minimum (or maximum) at t_1, then $\tilde{g}(t) = g(t)e^{-\lambda_N t}$ also has a local minimum (or maximum) at t_1, furthermore, $\tilde{g}(t_i) = 0, \dot{\tilde{g}}(t_i) \neq 0, i = 2, 3, \ldots, n-1$. Hence, there exists an $\varepsilon > 0$ (or $\varepsilon < 0$) such that $\tilde{g}(t) - \varepsilon = \sum_{i=1}^{N-1} e^{(\lambda_i - \lambda_N)t} f_i(t) + f_N(t) - \varepsilon$ has n zeros, which contradicts with Lemma A.1. Therefore, g changes signs at all t_i. This shows that $v(t) = \operatorname{sgn}(c'e^{At}b)$ (or $\operatorname{sgn}(-c'e^{At}b)$) is an extremal control.

Now consider the case that v has less than $n-1$ switches, say $n-1-j$ switches, $t_i, i = 1, 2, \ldots, n-1-j$. For simplicity and without loss of generality, assume that A is in the Jordan canonical form (the state transformation matrix can be absorbed in c' and (b). Partition A, b as

$$A = \begin{bmatrix} \star & \star \\ 0 & A_1 \end{bmatrix}, \quad b = \begin{bmatrix} \star \\ b_1 \end{bmatrix}$$

where A_1 is of size $n-j$. It is easy to see that A_1 is also of the Jordan canonical form and (A_1, b_1) is controllable. Furthermore,

$$e^{At} = \begin{bmatrix} \star & \star \\ 0 & e^{A_1 t} \end{bmatrix}.$$

Partition $c = [0 \quad c_1]'$ accordingly, then $c' e^{At} b = c_1' e^{A_1 t} b_1$. By the foregoing proof for the full dimensional case, we see that there exists c_1 such that $v(t) = \operatorname{sgn}(c_1' e^{A_1 t} b_1)$ is a bang–bang control with switching times exactly at $t_i, i = 1, 2, \ldots, n-1-j$.

Therefore, we conclude that any bang–bang control with less than $n-1$ switches is also extremal. □

References

[1] D.S. Bernstein, A.N. Michel, A chronological bibliography on saturating actuators, Internat. J. Robust Nonlinear Control 5 (1995) 375–380.
[2] C.A. Desoer, J. Wing, An optimal strategy for a saturating sampled data system, IRE Trans. Automat. Control AC-6 (1961) 5–15.
[3] M.E. Fisher, J.E. Gayek, Estimating reachable sets for two-dimensional linear discrete systems, J. Opt. Theory Appl. 56 (1987) 67–88.
[4] P.-O. Gutman, P. Hagander, A new design of constrained controllers for linear systems, IEEE Trans. Automat. Control 30 (1985) 22–33.
[5] O. Hájek, Control Theory in the Plane, Springer, Berlin, 1991.
[6] T. Hu, L. Qiu, Controllable regions of linear systems with bounded inputs, Systems Control Lett. 33 (1998) 55–61.
[7] R.E. Kalman, Optimal nonlinear control of saturating systems by intermittent action, IRE Wescon Convention Record Part 4, 1957, pp. 130–135.
[8] S.S. Keerthi, E.G. Gilbert, Computation of minimum-time feedback control laws for discrete-time systems with state-control constraints, IEEE trans. Automat. Control 32 (1987) 432–435.
[9] J.B. Lasserre, On reachable and controllable sets for two-dimensional linear discrete-time systems, J. Opt. Theory Appl. 70 (1991) 583–595.
[10] J.B. Lasserre, Reachable, controllable sets and stabilizing control of constrained linear systems, Automatica 29 (1993) 531–536.
[11] E. Lee, L. Markus, Foundations of Optimal Control, New York, Wiley, 1967.
[12] J.N. Lin, Determination of reachable set for a linear discrete system, IEEE Trans. Automat. Control AC-15 (1970) 339–342.
[13] J. Macki, M. Strauss, Introduction to Optimal Control, Springer, Berlin, 1982.
[14] D.Q. Mayne, Control of constrained dynamic systems, European J. Control 7 (2001) 87–99.
[15] D.Q. Mayne, J.B. Rawlings, C.V. Rao, P.O.M. Scokaert, Constrained model predictive control: stability and optimality Automatica 36 (2000) 789–814.
[16] W.E. Schmitendorf, B.R. Barmish, Null controllability of linear systems with constrained controls, SIAM J. Control Optim. 18 (1980) 327–345.
[17] E.D. Sontag, An algebraic approach to bounded controllability of linear systems, Internat. J. Control 39 (1984) 181–188.
[18] J. Stephan, M. Bodson, J. Lehocsky, Properties of recoverable sets for input and state constrained systems, Proceedings of the American Control Conference, Seattle, 1995, pp. 3912–3913.
[19] J. Stephan, M. Bodson, J. Lehocsky, Calculation of recoverable sets for 2-dimensional systems with input and state constraints, Proceedings of the Conference on Decision and Control, New Orleans, 1995, pp. 631–636.

Publication 22

To appear in *IEEE Transactions on Automatic Control*

On Improving the Performance with Bounded Continuous Feedback Laws

Tingshu Hu† Zongli Lin†

† Department of Electrical and Computer Engineering
P.O. Box 400743, University of Virginia
Charlottesville, VA 22903, U.S.A.
Email: th7f, zl5y@virginia.edu

Abstract

We present controller design methods to smoothen the discontinuity resulting from a piecewise linear control (PLC) law which was proposed to improve the convergence performance for systems with input constraints. The continuous control laws designed in this paper are explicit functions of the state and are easily implementable. We also show that the convergence performance can be further improved by using a saturated high gain feedback law. The efficiency of the proposed methods is illustrated with the PUMA 560 robot model.

Keywords: Switching, invariant ellipsoid, convergence rate, constrained control

[1]This work was supported in part by the US Office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

1 Introduction

We consider a linear system subject to input saturation and state constraint,

$$\dot{x} = Ax + Bu, \quad |u|_\infty \le 1, \quad x \in \Omega_0, \tag{1}$$

where $x \in \mathbf{R}^n, u \in \mathbf{R}^m, |u|_\infty = \max\{|u_i|,\ i \in [1,m]\}$ and Ω_0 contains the origin in its interior. To achieve a large domain of attraction, we may try to find a large ellipsoid (see [2, 7])

$$\mathcal{E}(P,\rho) := \{x \in \mathbf{R}^n :\ x^{\mathrm{T}} P x \le \rho\} \subset \Omega_0,$$

with $0 < P \in \mathbf{R}^{n\times n}$, such that this ellipsoid is invariant under $u = Fx$ and

$$\mathcal{E}(P,\rho) \subset \mathcal{L}(F) := \left\{x \in \mathbf{R}^n :\ |Fx|_\infty \le 1\right\}.$$

For simplicity, we use $\mathcal{E}(P)$ to denote $\mathcal{E}(P,1)$. Generally, the maximization of $\mathcal{E}(P,\rho)$ would result in low feedback gain F and slow convergence rate, i.e., some eigenvalues of $A+BF$ are close to the imaginary axis. In [7], Wredenhagen and Belanger proposed a PLC design method to reconcile large domain of attraction and good convergence performance. The basic idea is to use LQ method to construct a sequence of nested ellipsoids

$$\mathcal{E}(P_0,\rho_0) \supset \mathcal{E}(P_1,\rho_1) \supset \cdots \supset \mathcal{E}(P_N,\rho_N)$$

along with corresponding feedback gain matrices $F_i, i = 0,1,\cdots,N$, such that $\mathcal{E}(P_0,\rho_0) \subset \Omega_0$, $\mathcal{E}(P_i,\rho_i) \subset \mathcal{L}(F_i)$, and each $\mathcal{E}(P_i,\rho_i)$ is invariant under the feedback $u = F_i x, i = 0,1,\cdots,N$. Also, as the index i is increased, the convergence rate under the feedback $u = F_i x$ increases. The final controller takes the following form:

$$u = \begin{cases} F_N x, & \text{if } x \in \mathcal{E}(P_N,\rho_N), \\ F_{N-1}x, & \text{if } x \in \mathcal{E}(P_{N-1},\rho_{N-1}) \setminus \mathcal{E}(P_N,\rho_N), \\ \quad\vdots & \\ F_0 x, & \text{if } x \in \mathcal{E}(P_0,\rho_0) \setminus \mathcal{E}(P_1,\rho_1). \end{cases} \tag{2}$$

In this way, the domain of attraction is ensured to include the largest ellipsoid $\mathcal{E}(P_0,\rho_0)$ and as a state trajectory moves from an outer ellipsoid to an inner ellipsoid, the convergence rate is increased. Since each ellipsoid is invariant under the corresponding feedback law, the switch is safe (no chattering) and the existence and uniqueness of the solution to the closed-loop differential equation is ensured. Such a control law is referred to as PLC in [7]. Because $\mathcal{E}(P_i,\rho_i) \subset \mathcal{L}(F_i), i = 0,1,\cdots,N$, the control u will never exceed the saturation bound if the initial state $x_0 \in \mathcal{E}(P_0,\rho_0)$.

Since F_i is generally different from F_{i-1}, the control u in (2) is discontinuous at the switching surface $\partial\mathcal{E}(P_i,\rho_i)$, the boundary of the ellipsoid $\mathcal{E}(P_i,\rho_i)$. Effort has been made to smoothen the

discontinuity in [3, 4, 5, 6], etc. In [6], a continuous feedback law was constructed from the linear combination of F_k and F_{k+1}. Since this simple interpolation may cause the control to exceed the constraint, smaller bounds on the control were imposed and the ellipsoids were required to be "tightly" nested ($P_k - P_{k+1}$ should be sufficiently small). By using the gain scheduling methods in [3, 4, 5], it has been shown that the discontinuity can be smoothened by using a continuum of ellipsoids $\mathcal{E}(P(\varepsilon))$, where ε is a scheduling variable. The essence of these gain scheduling methods is the following: For every $\varepsilon > 0$, $P(\varepsilon)$ is solved from a parameter dependent Riccati equation along with an LQR gain matrix $F(\varepsilon)$. As ε is increased, the ellipsoid $\mathcal{E}(P(\varepsilon))$ becomes smaller and the convergence rate within $\mathcal{E}(P(\varepsilon))$ is increased. The gain scheduling idea is to associate each $x \in \mathbf{R}^n$ with a parameter ε, or, to define a function $\varepsilon(x)$: $x \mapsto \varepsilon$. The final controller has the form of $u = F(\varepsilon(x))x$. Since the function $\varepsilon(x)$ is generally very hard to compute, technical issues are involved in controller implementation. These issues were considered in [3] and a method to simplify the computation of $\varepsilon(x)$ was proposed. The proposed method involves solving a convex optimization problem for every point x in the state space.

This note is intended to propose explicit controller structures which would achieve the objective of improving the convergence performance using continuous control laws. For easy reference, here we collect some simple mathematical facts as follows.

Fact 1 *For two ellipsoids $\mathcal{E}(P_1)$ and $\mathcal{E}(P_2)$,*

$$\mathcal{E}(P_1) \subset \mathcal{E}(P_2) \iff P_1 \geq P_2;$$

$$\mathcal{E}(P_1) \subset \mathrm{int}(\mathcal{E}(P_2)) \iff P_1 > P_2;$$

where $\mathrm{int}(\mathcal{E}(P_2)) = \{x \in \mathbf{R}^n : x^{\mathrm{T}} P_2 x < 1\}$ is the interior of $\mathcal{E}(P_2)$. For an ellipsoid $\mathcal{E}(P)$ and a matrix $F \in \mathbf{R}^{m \times n}$,

$$\begin{aligned} \mathcal{E}(P) \subset \mathcal{L}(F) &\iff f_i^{\mathrm{T}} f_i \leq P, \ \ i \in [1, m] \iff f_i P^{-1} f_i^{\mathrm{T}} \leq 1, \ \ i \in [1, m] \\ &\iff \begin{bmatrix} 1 & f_i P^{-1} \\ P^{-1} f_i^{\mathrm{T}} & P^{-1} \end{bmatrix} \geq 0, \quad i \in [1, m]. \end{aligned}$$

2 A continuous feedback law for improving the performance

We consider the system

$$\dot{x} = Ax + Bu, \quad |u|_\infty \leq 1, \quad x \in \Omega_0, \tag{3}$$

with a two stage switching feedback law

$$u = \begin{cases} F_1 x, & \text{if } x \in \mathcal{E}(P_1), \\ F_0 x, & \text{if } x \in \mathcal{E}(P_0) \setminus \mathrm{int}(\mathcal{E}(P_1)), \end{cases} \tag{4}$$

where

$$P_0 < P_1, \quad \Omega_0 \supset \mathcal{E}(P_0), \quad \mathcal{E}(P_0) \subset \mathcal{L}(F_0), \quad \mathcal{E}(P_1) \subset \mathcal{L}(F_1), \tag{5}$$

$$(A + BF_0)^{\mathrm{T}} P_0 + P_0(A + BF_0) \leq -\alpha_0 P_0, \tag{6}$$

$$(A + BF_1)^{\mathrm{T}} P_1 + P_1(A + BF_1) \leq -\alpha_1 P_1, \tag{7}$$

and $0 < \alpha_0 < \alpha_1$. Assume that α_0 and α_1 are the maximal positive numbers that satisfy (6) and (7), respectively. The inequality $\alpha_0 < \alpha_1$ implies that the convergence rate of the Lyapunov function $V_1(x) = x^{\mathrm{T}} P_1 x$ under $u = F_1 x$ is greater than that of $V_0(x) = x^{\mathrm{T}} P_0 x$ under $u = F_0 x$. We consider a feedback law (4) of only one switch because the method to be proposed can be readily extended to smoothen the discontinuity of a controller with multiple switches. Actually, because the proposed continuous feedback law guarantees a progressively increasing convergence rate, we only need to use the outmost and the innermost ellipsoids $\mathcal{E}(P_0, \rho_0)$ and $\mathcal{E}(P_N, \rho_N)$ along with their corresponding feedback gain matrices F_0 and F_N. Without loss of generality, we have assumed that $\rho_0 = \rho_1 = 1$. Otherwise, ρ_0 and ρ_1 can be absorbed into the matrices P_0 and P_1.

The control (4) is discontinuous at the surface of the inner ellipsoid, $\partial\mathcal{E}(P_1)$. The main idea for smoothening this discontinuity is to construct a continuum of ellipsoids $\mathcal{E}(P(\gamma)), \gamma \in [0, 1]$, between $\mathcal{E}(P_0)$ and $\mathcal{E}(P_1)$, progressively shrinking, along with a continuum of feedback matrices $F(\gamma)$, such that

$$\mathcal{E}(P(\gamma)) \subset \mathcal{L}(F(\gamma))$$

and

$$(A + BF(\gamma))^{\mathrm{T}} P(\gamma) + P(\gamma)(A + BF(\gamma)) \leq -\alpha(\gamma) P(\gamma),$$

with $\alpha(\gamma)$ monotonically increasing as γ changes from 0 to 1. For $x \in \partial\mathcal{E}(P(\gamma))$, we use the control $u = F(\gamma)x$. Suppose that for every $x \in \mathcal{E}(P_0) \backslash \mathrm{int}(\mathcal{E}(P_1))$, there exists a unique $\gamma \in [0, 1]$ such that $x^{\mathrm{T}} P(\gamma) x = 1$, then we can define

$$\gamma(x) := \{\gamma \in [0, 1] : \ x^{\mathrm{T}} P(\gamma) x = 1\}, \tag{8}$$

and the feedback law can be simply written as

$$u = F(\gamma(x))x. \tag{9}$$

The control law (9) is implementable if the function $\gamma(x)$ and the feedback matrix $F(\gamma(x))$ can be computed efficiently on line. That is, we should be able to tell which surface of the ellipsoids the state x is on. This depends on how we design the functions $P(\gamma)$ and $F(\gamma)$.

The following are the functions we propose. Let

$$Q_0 = P_0^{-1}, \quad Q_1 = P_1^{-1}, \quad H_0 = F_0 Q_0, \quad H_1 = F_1 Q_1.$$

Define

$$Q(\gamma) := (1-\gamma)Q_0 + \gamma Q_1, \quad H(\gamma) := (1-\gamma)H_0 + \gamma H_1 \tag{10}$$

and

$$P(\gamma) := Q(\gamma)^{-1}, \quad F(\gamma) := H(\gamma)P(\gamma). \tag{11}$$

It is clear that $Q(\gamma) > 0$ for all $\gamma \in [0,1]$. Hence $Q(\gamma)$, $H(\gamma)$, $P(\gamma)$ and $F(\gamma)$ are all continuous in γ over the interval $[0,1]$. Same function $Q(\gamma)$ was used in [3], where $F(\gamma)$ was the solution to a Riccati equation. In what follows, we show that, with $F(\gamma)$ defined as above, the continuous feedback law $u = F(\gamma(x))x$ possesses all the desired properties. We will also provide an explicit formula to compute $\gamma(x)$.

Theorem 1 *With $P(\gamma)$ defined in (11), there exists a unique $\gamma \in [0,1]$ such that $x^{\mathrm{T}}P(\gamma)x = 1$ for every $x \in \mathcal{E}(P_0) \setminus \mathrm{int}(\mathcal{E}(P_1))$. With $\gamma(x)$ defined in (8), we have*

$$\gamma(x) = \lambda_{\min}\left[(Q_0 - Q_1)^{-\frac{1}{2}}(Q_0 - xx^{\mathrm{T}})(Q_0 - Q_1)^{-\frac{1}{2}}\right], \tag{12}$$

for $x \in \mathcal{E}(P_0) \setminus \mathrm{int}(\mathcal{E}(P_1))$. Let $\gamma(x) = 1$ for $x \in \mathrm{int}(\mathcal{E}(P_1))$. Then the control $u = F(\gamma(x))x$ is continuous in x and $|F(\gamma(x))x| \le 1$ for all $x \in \mathcal{E}(P_0)$. Moreover, each ellipsoid $\mathcal{E}(P(\gamma)), \gamma \in [0,1]$ is invariant and every trajectory starting from $x_0 \in \mathcal{E}(P_0)$ will converge to the origin with increasing rate.

Before proving Theorem 1, we present two lemmas. Define

$$\alpha(\gamma) := \max\{\alpha > 0 : \ (A + BF(\gamma))^{\mathrm{T}}P(\gamma) + P(\gamma)(A + BF(\gamma)) \le -\alpha P(\gamma)\}.$$

Then $\alpha(\gamma)$ is the convergence rate of $V(x) = x^{\mathrm{T}}P(\gamma)x$ under the linear control $u = F(\gamma)x$.

Lemma 1

1. *$\mathcal{E}(P(\gamma))$ shrinks as γ increases, namely, if $\gamma_1 < \gamma_2$, then $\mathcal{E}(P(\gamma_2)) \subset \mathrm{int}(\mathcal{E}(P(\gamma_1)))$;*
2. *For all $\gamma \in [0,1]$, $\mathcal{E}(P(\gamma)) \subset \mathcal{L}(F(\gamma))$;*
3. *$\alpha(\gamma)$ is strictly increasing for $\gamma \in [0,1]$.*

Proof. From (5), we have $Q_0 - Q_1 = P_0^{-1} - P_1^{-1} > 0$.

1. If $\gamma_1 < \gamma_2$, then

$$Q(\gamma_1) - Q(\gamma_2) = (\gamma_2 - \gamma_1)(Q_0 - Q_1) > 0.$$

Hence $P(\gamma_1) < P(\gamma_2)$ and it follows from Fact 1 that $\mathcal{E}(P(\gamma_2)) \subset \mathrm{int}(\mathcal{E}(P(\gamma_1))$.

2. Since $\mathcal{E}(P_0) \subset \mathcal{L}(F_0)$ and $\mathcal{E}(P_1) \subset \mathcal{L}(F_1)$, it follows from Fact 1 that

$$\begin{bmatrix} 1 & f_{0i}P_0^{-1} \\ P_0^{-1}f_{0i}^{\mathrm{T}} & P_0^{-1} \end{bmatrix} \geq 0, \quad \begin{bmatrix} 1 & f_{1i}P_1^{-1} \\ P_1^{-1}f_{1i}^{\mathrm{T}} & P_1^{-1} \end{bmatrix} \geq 0, \quad i \in [1,m],$$

i.e.,

$$\begin{bmatrix} 1 & h_{0i} \\ h_{0i}^{\mathrm{T}} & Q_0 \end{bmatrix} \geq 0, \quad \begin{bmatrix} 1 & h_{1i} \\ h_{1i}^{\mathrm{T}} & Q_1 \end{bmatrix} \geq 0, \quad i \in [1,m].$$

By convexity, we have

$$\begin{bmatrix} 1 & h_i(\gamma) \\ h_i^{\mathrm{T}}(\gamma) & Q(\gamma) \end{bmatrix} \geq 0, \quad i \in [1,m], \quad \gamma \in [0,1].$$

That is,

$$\begin{bmatrix} 1 & f_i(\gamma)P^{-1}(\gamma) \\ P^{-1}(\gamma)f_i^{\mathrm{T}}(\gamma) & P^{-1}(\gamma) \end{bmatrix} \geq 0, \quad i \in [1,m], \quad \gamma \in [0,1].$$

Therefore, by Fact 1, we have $\mathcal{E}(P(\gamma)) \subset \mathcal{L}(F(\gamma))$ for all $\gamma \in [0,1]$.

3. By multiplying both sides of (6) and (7) with Q_0 and Q_1 respectively, we obtain

$$\begin{aligned} Q_0A^{\mathrm{T}} + AQ_0 + H_0^{\mathrm{T}}B^{\mathrm{T}} + BH_0 &\leq -\alpha_0 Q_0, \\ Q_1A^{\mathrm{T}} + AQ_1 + H_1^{\mathrm{T}}B^{\mathrm{T}} + BH_1 &\leq -\alpha_1 Q_1. \end{aligned}$$

By convexity, we have

$$Q(\gamma)A^{\mathrm{T}} + AQ(\gamma) + H(\gamma)^{\mathrm{T}}B^{\mathrm{T}} + BH(\gamma) \leq -(1-\gamma)\alpha_0 Q_0 - \gamma\alpha_1 Q_1, \quad \gamma \in [0,1].$$

Since $\alpha_1 > \alpha_0$ and $Q_1 > 0$, we see that

$$(1-\gamma)\alpha_0 Q_0 + \gamma\alpha_1 Q_1 = \alpha_0 Q(\gamma) + \gamma(\alpha_1 - \alpha_0)Q_1 > \alpha_0 Q(\gamma).$$

It follows that $\alpha(\gamma) > \alpha_0$ for all $\gamma \in (0,1]$. To show that $\alpha(\gamma_2) > \alpha(\gamma_1)$ for $\gamma_2 > \gamma_1$, we observe that

$$\begin{aligned} \gamma_2 &= \left(1 - \frac{\gamma_2 - \gamma_1}{1-\gamma_1}\right)\gamma_1 + \frac{\gamma_2 - \gamma_1}{1-\gamma_1}, \\ Q(\gamma_2) &= \left(1 - \frac{\gamma_2 - \gamma_1}{1-\gamma_1}\right)Q(\gamma_1) + \frac{\gamma_2 - \gamma_1}{1-\gamma_1}Q_1, \end{aligned}$$

and

$$H(\gamma_2) = \left(1 - \frac{\gamma_2 - \gamma_1}{1-\gamma_1}\right)H(\gamma_1) + \frac{\gamma_2 - \gamma_1}{1-\gamma_1}H_1.$$

Following the same procedure by replacing Q_0 with $Q(\gamma_1)$, H_0 with $H(\gamma_1)$, and γ with $\frac{\gamma_2-\gamma_1}{1-\gamma_1}$, we can show that $\alpha(\gamma_2) > \alpha(\gamma_1)$ for $\gamma_2 > \gamma_1$. □

Lemma 2 *For every $x \in \mathcal{E}(P_0)\backslash\mathrm{int}(\mathcal{E}(P_1))$, there exists a unique $\gamma \in [0,1]$ such that $x^{\mathrm{T}}P(\gamma)x = 1$. Let $\gamma(x)$ be defined in (8). Then, for $x \in \mathcal{E}(P_0) \setminus \mathrm{int}(\mathcal{E}(P_1))$,*

$$\gamma(x) = \lambda_{\min}\left[(Q_0 - Q_1)^{-\frac{1}{2}}(Q_0 - xx^{\mathrm{T}})(Q_0 - Q_1)^{-\frac{1}{2}}\right]. \tag{13}$$

Proof. For each $x \in \mathcal{E}(P_0) \setminus \mathrm{int}(\mathcal{E}(P_1))$, we have $x^{\mathrm{T}}P_0x \le 1$ and $x^{\mathrm{T}}P_1x \ge 1$, i.e., $x^{\mathrm{T}}P(0)x \le 1$ and $x^{\mathrm{T}}P(1)x \ge 1$. Since $P(\gamma)$ is continuous in γ for $\gamma \in [0,1]$, there exists a $\gamma \in [0,1]$ such that $x^{\mathrm{T}}P(\gamma)x = 1$. By Lemma 1 item 1, there exists a unique $\gamma \in [0,1]$ such that $x^{\mathrm{T}}P(\gamma)x = 1$. Hence the function $\gamma(x)$ is well defined by (8).

Let γ be the unique number in $[0,1]$ such that $x^{\mathrm{T}}P(\gamma)x = 1$, i.e., $x \in \partial\mathcal{E}(P(\gamma))$. By Lemma 1 item 1, $x \in \mathrm{int}(\mathcal{E}(P(\gamma_1)))$ for all $\gamma_1 \in [0,\gamma)$, i.e.,

$$x^{\mathrm{T}}P(\gamma_1)x < 1, \quad \forall\gamma_1 \in [0,\gamma).$$

It follows from Schur complement that

$$\begin{aligned}
&\begin{bmatrix} 1 & x^{\mathrm{T}} \\ x & Q(\gamma_1) \end{bmatrix} = \begin{bmatrix} 1 & x^{\mathrm{T}} \\ x & Q_0 - \gamma_1(Q_0 - Q_1) \end{bmatrix} > 0 \\
\iff\ & Q_0 - \gamma_1(Q_0 - Q_1) - xx^{\mathrm{T}} > 0 \\
\iff\ & \gamma_1 I < (Q_0 - Q_1)^{-\frac{1}{2}}(Q_0 - xx^{\mathrm{T}})(Q_0 - Q_1)^{-\frac{1}{2}}, \quad \gamma_1 \in [0,\gamma).
\end{aligned}$$

By continuity, we have

$$\gamma I \le (Q_0 - Q_1)^{-\frac{1}{2}}(Q_0 - xx^{\mathrm{T}})(Q_0 - Q_1)^{-\frac{1}{2}}. \tag{14}$$

From $x^{\mathrm{T}}P(\gamma)x = 1$, we have,

$$\begin{aligned}
0 &= \det\begin{bmatrix} 1 & x^{\mathrm{T}} \\ x & Q(\gamma) \end{bmatrix} \\
&= \det[Q(\gamma) - xx^{\mathrm{T}}] \\
&= \det[Q_0 - \gamma(Q_0 - Q_1) - xx^{\mathrm{T}}].
\end{aligned}$$

Hence,

$$\det\left[\gamma I - (Q_0 - Q_1)^{-\frac{1}{2}}(Q_0 - xx^{\mathrm{T}})(Q_0 - Q_1)^{-\frac{1}{2}}\right] = 0, \tag{15}$$

which implies that γ is an eigenvalue of the matrix $(Q_0 - Q_1)^{-\frac{1}{2}}(Q_0 - xx^{\mathrm{T}})(Q_0 - Q_1)^{-\frac{1}{2}}$. In view of (14), we obtain (13). □

Proof of Theorem 1. The first statement and equation (12) have been proved in Lemma 2. From the continuity of the eigenvalues of a matrix in its elements, it follows that $\gamma(x)$ is continuous in x at every $x \in \mathcal{E}(P_0) \setminus \mathrm{int}(\mathcal{E}(P_1))$. Since $\gamma(x) = 1$ for all $x \in \partial\mathcal{E}(P_1)$, the function $\gamma(x)$

can be extended continuously to all $x \in \mathcal{E}(P_0)$ by letting $\gamma(x) = 1$ for $x \in \mathcal{E}(P_1)$. Since $F(\gamma)$ is continuous in γ, the control $u = F(\gamma(x))x$ is continuous in x. The claim that $|F(\gamma(x))x|_\infty \leq 1$ for all $x \in \mathcal{E}(P_0)$ follows directly from Lemma 1 item 2.

Now we consider the convergence of trajectories. Under the control $u = F(\gamma(x))x$, on the boundary of each $\mathcal{E}(P(\gamma))$, $\gamma \in [0, 1]$,

$$x^{\mathrm{T}} P(\gamma)\dot{x} = \frac{1}{2}x^{\mathrm{T}}((A + BF(\gamma))^{\mathrm{T}}P + P(A + BF(\gamma))x \leq -\frac{1}{2}\alpha(\gamma)x^{\mathrm{T}}P(\gamma)x < 0.$$

We also have $x^{\mathrm{T}}P_1\dot{x} < 0$ for all $x \in \mathcal{E}(P_1)$. Hence all the trajectories starting from the boundary of $\mathcal{E}(P(\gamma))$ will be in the interior of $\mathcal{E}(P(\gamma))$ for all $t > 0$. This proves the invariance of $\mathcal{E}(P(\gamma))$. Moreover, the convergence rate $\alpha(\gamma)$ increases as the trajectory enters the inner ellipsoids. Therefore, if a trajectory strats from $x_0 \in \mathcal{E}(P_0)$, it will converge to the origin. □

3 Further improvement of the convergence performance

As can be seen from Lemma 1 item 2, the control law constructed in Section 2 satisfies the control constraint by avoiding saturation. Since $\mathcal{E}(P(\gamma)) \subset \mathcal{L}(F(\gamma))$, there are at most two intersections between the ellipsoid $\mathcal{E}(P(\gamma))$ and a pair of hyperplanes, $f_i(\gamma)x = \pm 1$. Hence the control $u_i = f_i(\gamma)x$ may take the maximal value ± 1 only at two points on $\partial\mathcal{E}(P(\gamma))$. Along a trajectory, the control signal could be well below the saturation level most of the time. This means that the capacity of the actuators is not fully utilized and we still have much potential to improve the convergence performance. In [1], it is shown that the control law that maximizes the convergence rate of a Lyapunov function $V(x) = x^{\mathrm{T}}Px$ under actuator saturation is simply $u_i = -\mathrm{sign}(b_i^{\mathrm{T}}Px), i = 1, 2, \cdots, m$. Due to the discontinuity of this bang bang control, a saturated high gain linear feedback law of the form $u = -\mathrm{sat}(kB^{\mathrm{T}}Px)$ is proposed to achieve a suboptimal convergence rate. Here, $\mathrm{sat}(\cdot)$ is the standard vector-valued saturation function: $\{\mathrm{sat}(u)\}_i = \mathrm{sign}(u_i)\min\{|u_i|, 1\}$. It is also shown that the maximal convergence rate depends on the choice of the P matrix (see Chapter 11 of [1] for more detail). Generally, the objective of producing a high convergence rate conflicts with the objective of achieving a large invariant ellipsoid. In other words, if P is chosen such that the maximal convergence rate is high, then the largest ellipsoid $\mathcal{E}(P, \rho)$ that can be made invariant would be small. A simple way to reconcile the objective of producing a large invariant ellipsoid and that of achieving a high convergence rate with a control $u = -\mathrm{sat}(kB^{\mathrm{T}}Px)$ is also to adjust the P matrix according to the size of the state. Using the method in Section 2, a state dependent matrix $P(\gamma(x))$ can be determined. While the state feedback $u = F(\gamma(x))x$ increases the convergence rate as the trajectory enters smaller ellipsoid $\mathcal{E}(P(\gamma))$, a state feedback of the form $u = -\mathrm{sat}(kB^{\mathrm{T}}P(\gamma(x))x)$ with a high gain k can be used for further improvement of the convergence performance.

Theorem 2 *Let $F_0 = -G_0 B^{\mathrm{T}} P_0$ and $F_1 = -G_1 B^{\mathrm{T}} P_1$ be the feedback matrices of two LQ controllers such that $\mathcal{E}(P_0) \subset \mathcal{L}(F_0)$, $\mathcal{E}(P_1) \subset \mathcal{L}(F_1)$ and $P_1 < P_0$. Assume that $G_0 = \mathrm{diag}\{g_{01}, g_{02}, g_{03}\} > 0$ and $G_1 = \mathrm{diag}\{g_{11}, g_{12}, g_{13}\} > 0$. Let $k_0 = \max\{g_{01}, g_{02}, g_{03}, g_{11}, g_{12}, g_{13}\}$. Then for any $k > k_0$, under the control of*

$$u = -\mathrm{sat}(kB^{\mathrm{T}} P(\gamma(x))x), \tag{16}$$

all the trajectories starting from $\mathcal{E}(P_0)$ will converge to the origin. The convergence rate increases as k increases.

Proof. Denote $G(\gamma) = (1-\gamma)G_0 + \gamma G_1$. Then $0 < G(\gamma) \leq k_0 I$ for all $\gamma \in [0,1]$ and $F(\gamma) = -G(\gamma)B^{\mathrm{T}}P(\gamma)$. It follows from Theorem 1 that under the control of $u = F(\gamma(x))x$, all the trajectories starting from $\mathcal{E}(P_0)$ will converge to the origin. Also we have $|F(\gamma(x))x|_\infty \leq 1$ for all $x \in \mathcal{E}(P_0)$. The convergence rate is faster under (16) for $k \geq k_0$ can be seen from

$$-x^{\mathrm{T}}P(\gamma)b_i\,\mathrm{sat}(kb_i^{\mathrm{T}}P(\gamma)x) \leq -x^{\mathrm{T}}P(\gamma)b_i\,\mathrm{sat}(g_i(\gamma)b_i^{\mathrm{T}}P(\gamma)x) = x^{\mathrm{T}}P(\gamma)b_i f_i(\gamma)x, \tag{17}$$

where $g_i(\gamma)$ is the ith diagonal element of $G(\gamma)$. As k increases, the left-hand side of (17) decreases. Hence the convergence rate increases as k increases. □

Since both $\mathrm{sat}(\cdot)$ and $P(\gamma(\cdot))$ are continuous functions, the control (16) is continuous in x. We call (16) the fast continuous control law. As will be seen in the example, the feedback law (16) can improve the convergence rate significantly over $u = F(\gamma(x))x$.

Example: We consider the Puma 560 robot model that was used in [7]. The robot has three joints: the trunk (Joint 1), the shoulder (Joint 2) and the elbow (Joint 3). These joints are controlled by three actuators which are subject to different saturation bounds, 97.8Nm, 136.4Nm, and 89.4Nm, respectively. The linear model was calculated about $(\theta_1, \theta_2, \theta_3) = (57°, 115°, 172°)$ with resulting system matrices

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & -0.0451 & -0.0451 & 0 & 0 & 0 \\ 0 & -0.0457 & -0.0457 & 0 & 0 & 0 \\ 0 & -4.5551 & -4.5551 & 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.0925 & 0.0000 & 0.0026 \\ 0.0000 & 0.0979 & -0.0952 \\ 0.0026 & -0.0952 & 0.3616 \end{bmatrix}.$$

In [7], a PLC control law with 5 switches was designed. Based on the outmost and the innermost ellipsoids of [7], with corresponding feedback matrices, we designed a continuous feedback law of the form $u = F(\gamma(x))x$ (Here we need to take into account the non-unity saturation bounds of the three actuators). Figs. 1 and 2 illustrate the simulation results under the PLC law and the continuous feedback law. The initial condition is $x_0 = 10[1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1]^{\mathrm{T}}/\sqrt{6}$, which is the

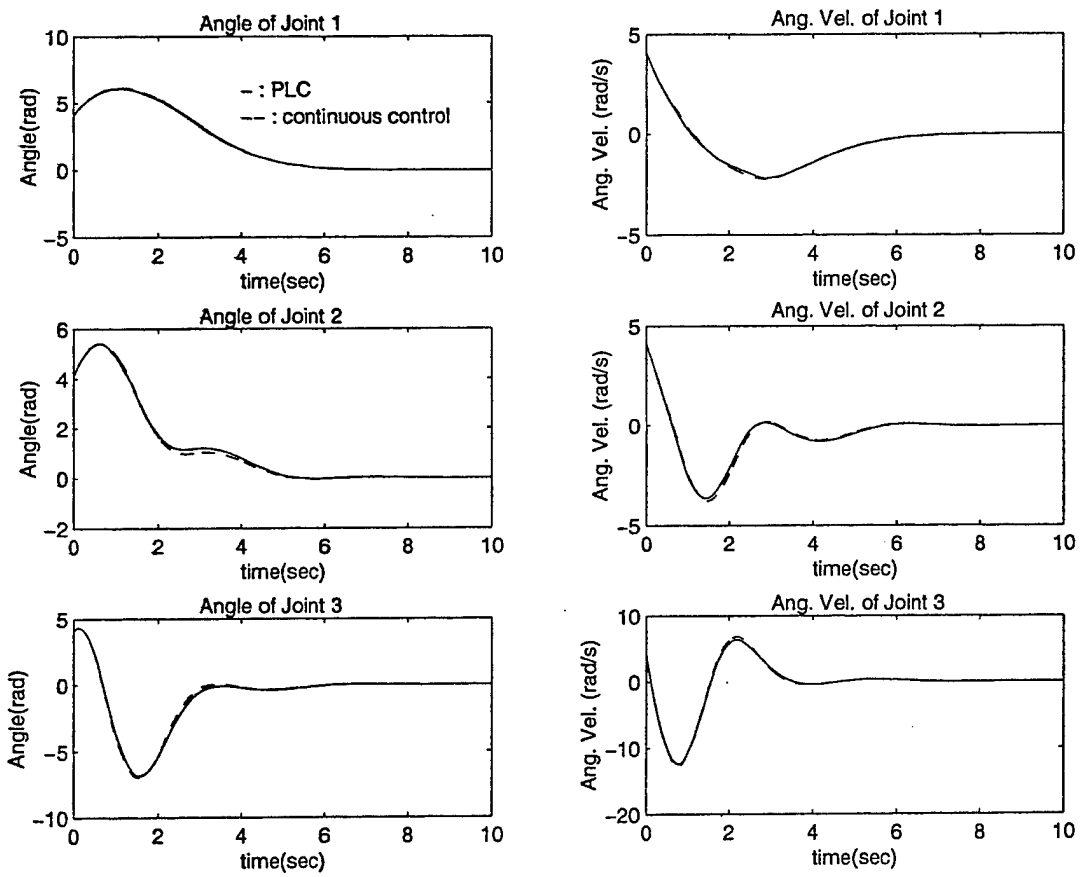


Figure 1: Time responses of the states: PLC vs continuous control

same as that in [7]. From Fig. 1, we can see that the time response of the states under the PLC control and that under the continuous feedback control are almost identical. The control signals under the PLC law are however discontinuous and display big spikes (see Fig. 2). These spikes can be reduced by increasing the number of the nested ellipsoids and the controllers used in switching. However, this would increase the numerical burden when determining the smallest ellipsoid that includes a given state and would also increase the data storage for the controller. The control signals under the continuous control law are continuous, as expected.

From Fig. 2, we also see that the control signals are well below the saturation level. This indicates that there is a potential for further improvement of the performance. We use the controller (16) for this purpose. Recall that we have assumed unity saturation level in (1) and (16) is only suitable for systems with unity saturation level. To transform the system into the standard form of (1), let $\Lambda = \text{diag}\{97.8, 136.4, 89.4\}$, $\bar{B} = B\Lambda$ and $\bar{u} = \Lambda^{-1}u$. Then the system $\dot{x} = Ax + \bar{B}\bar{u}$ has a unity saturation level. For this system, the fast continuous control law is $\bar{u} = -\text{sat}(k\bar{B}^{\mathrm{T}}P(\gamma(x))x$. Equivalently, $u = \Lambda\bar{u} = -\Lambda\,\text{sat}(k\Lambda B^{\mathrm{T}}P(\gamma(x))x)$. The actual controllers

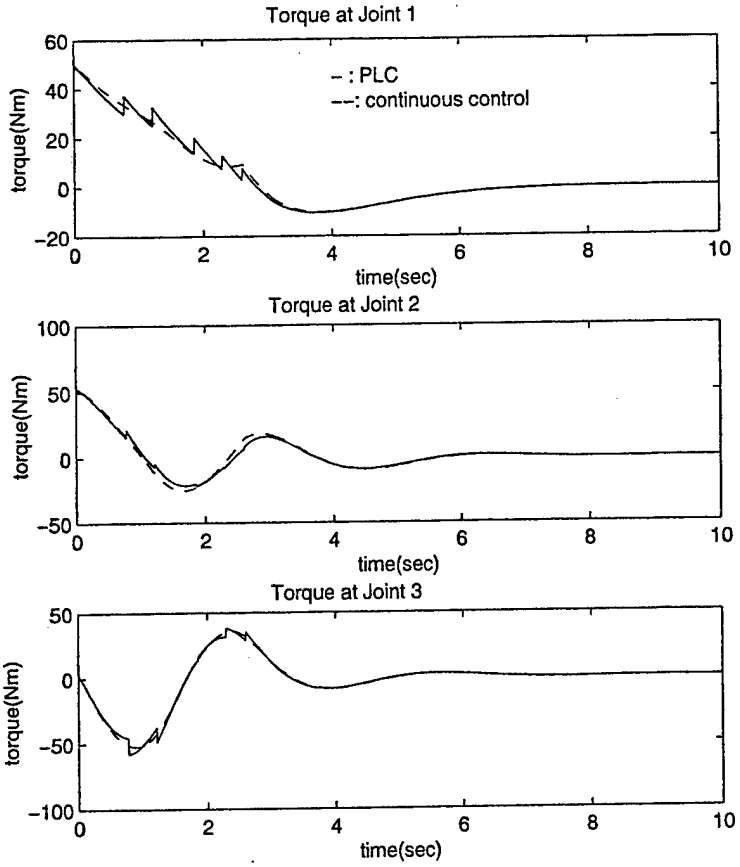


Figure 2: The control signals: PLC vs continuous control

have the following forms:

$$\begin{cases} u_1 = -97.8\,\mathrm{sat}(97.8kb_1^{\mathrm{T}}P(\gamma(x))x), \\ u_2 = -136.4\,\mathrm{sat}(136.4kb_2^{\mathrm{T}}P(\gamma(x))x), \\ u_3 = -89.4\,\mathrm{sat}(89.4kb_3^{\mathrm{T}}P(\gamma(x))x). \end{cases} \tag{18}$$

Here, in the construction of $P(\gamma)$, we have used the outmost and the innermost ellipsoids in [7]. To make full use of the actuator capacities, we have taken $k = 6$ in simulation. Figs. 3 and 4 illustrate the simulation results under the control (18) (the solid curves) as compared with those under the control $u = F(\gamma(x))x$ (the dashed curves). From Fig. 3, we see that the performance of the state response is significantly improved by using the fast continuous control law (18). Fig. 4 shows that this control law has utilized more potential of the actuator capacities. We notice that there is a sharp turn (not discontinuity) in the torque at Joint 1. This may happen when the state trajectory enters the smaller ellipsoid $\mathcal{E}(P_1)$, since at the intersection between a trajectory with the boundary of this ellipsoid, the function $\gamma(x)$ is continuous but not differentiable in x.

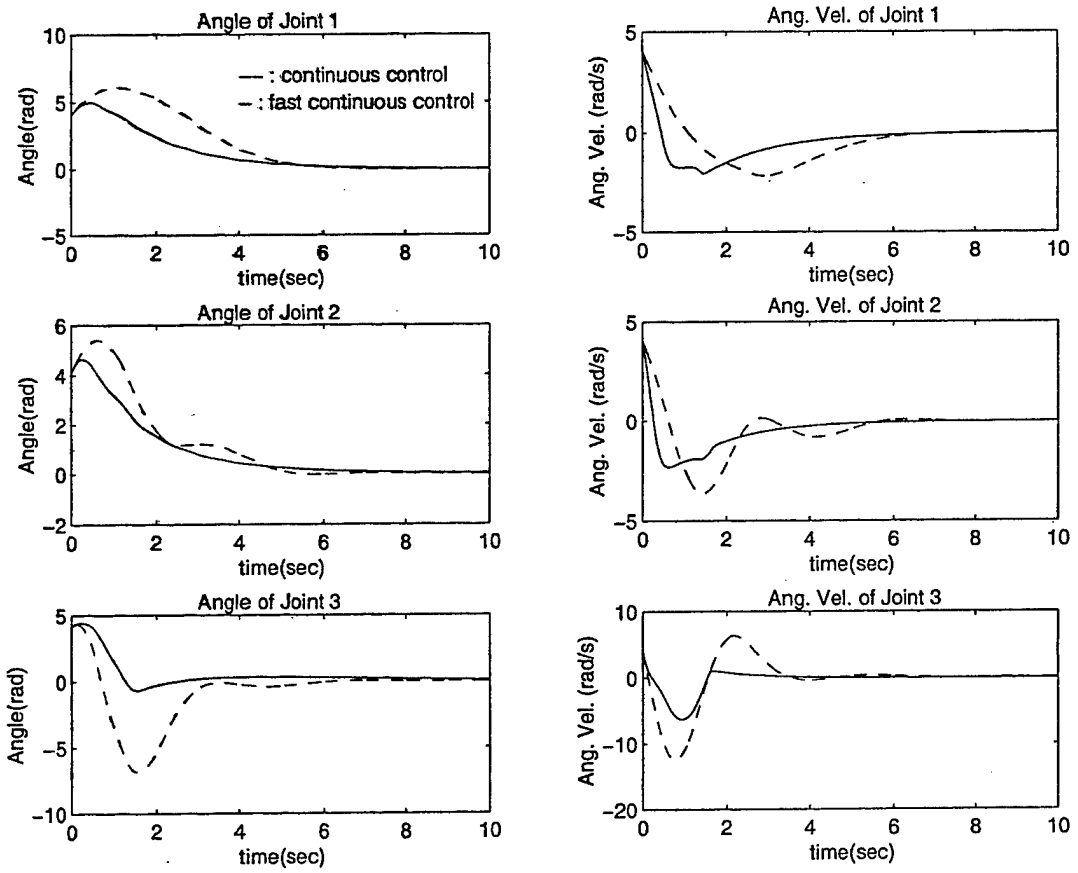


Figure 3: Time response of the states: continuous control vs fast continuous control

4 Conclusions

We developed simple continuous feedback laws for improving the convergence performance of linear systems subject to actuator and state constraints. The control laws are expressed as explicit functions of the state and are easily implementable. The efficiency of the proposed methods is illustrated with a PUMA 560 robot model.

References

[1] T. Hu and Z. Lin. *Control Systems with Actuator Saturation: Analysis and Design*, Birkhäuser, Boston, 2001.

[2] T. Hu and Z. Lin, "On enlarging the basin of attraction for linear systems under saturated linear feedback," *Sys. & Contr. Lett.*, Vol. 40, No. 1, pp. 59-69, May, 2000.

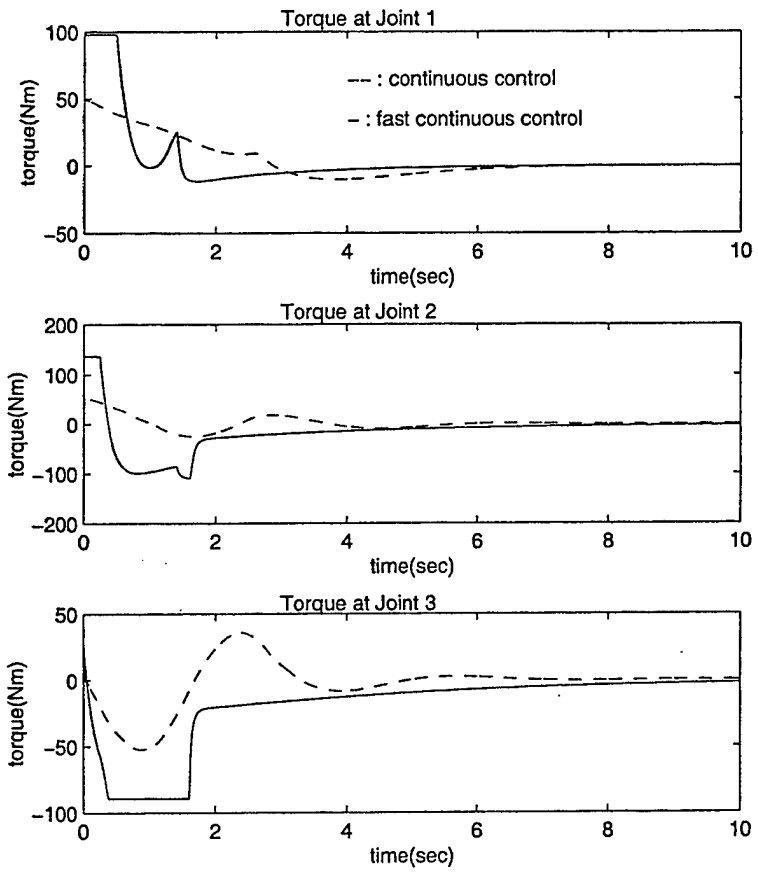


Figure 4: The control signals: continuous control vs fast continuous control

[3] A. Megretski, "L_2 BIBO output feedback stabilization with saturated control," *13th IFAC World Congress*, Vol. D, pp. 435-440, 1996.

[4] R. Suarez, J. Alvarez-Ramirez and J. Solis-Daun, "Linear systems with bounded inputs: global stabilization with eigenvalue placement," *Int. J. Robust Nonlin. Contr.*, Vol. 7, pp. 835-845, 1997.

[5] A. R. Teel, "Linear systems with input nonlinearities: global stabilization by scheduling a family of H_∞-type controllers," *Int. J. of Robust and Nonlinear Control*, Vol. 5, pp. 399-441, 1995.

[6] G. F. Wredenhagen, "A new method of controller design for systems with input constraints using interpolation functions," *Proceedings of the 33rd Conference on Decision and Control*, pp. 1024-1029, Lake Buena Vista, FL, 1994.

[7] G. F. Wredenhagen and P. R. Belanger, "Piecewise-linear LQ control for systems with input constraints," *Automatica*, Vol. 30, pp. 403-416, 1994.

Publication 23

Submitted to IEEE Transaction on Automatic Control

Composite Quadratic Lyapunov Functions for Constrained Control Systems

Tingshu Hu Zongli Lin

Department of Electrical and Computer Engineering
University of Virginia
Charlottesville, VA 22903
th7f, zl5y@virginia.edu

Abstract

A Lyapunov function based on a group of quadratic functions is introduced in this paper. We call this Lyapunov function a composite quadratic function. Some important properties of this Lyapunov function are revealed. We show that this function is continuously differentiable and its level set is the convex hull of a group of ellipsoids. These results are used to study the set invariance properties of linear systems with input and state constraints. We show that for a system under a given saturated linear feedback, the convex hull of a group of invariant ellipsoids is also invariant. If each ellipsoid in a group can be made invariant with a bounded control of the saturating actuator, then their convex hull can also be made invariant by the same actuator. For a group of ellipsoids, each invariant under a separate saturated linear feedback, we also present a method for constructing a nonlinear continuous feedback law which makes their convex hull invariant.

Keywords: Quadratic functions, invariant set, constrained control

[1]This work was supported in part by the US Office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

1 Introduction

We consider linear systems subject to input saturation and state constraint. Control problems for these systems have attracted tremendous attention in recent years because of their practical significance and the theoretical challenges. For linear systems with input saturation, global and semi-global stabilization results have been obtained for semi-stable systems[1] (see, e.g., [15, 16, 19, 21, 22, 23]) and systems with two anti-stable poles (see [12, 8]). For more general systems with both input saturation and state constraint, there are numerous research reports on their stability analysis and design (see [1, 5, 6, 8, 18, 25] and the references therein). While analytical characterizations of the domain of attraction and the maximal (controlled) invariant set have been attempted and are believed to be extremely hard except for some special cases (see, e.g., [12]), most of the literature is dedicated to obtaining an estimation of the domain of attraction with reduced conservatism or to enlarge some invariant set inside the domain of attraction. Along this direction, the notion of set invariance has played a very important role (see, e.g., [2, 3, 25]). The most commonly used invariant sets for continuous-time systems are invariant ellipsoids, resulting from the level sets of quadratic Lyapunov functions. The problem of estimating the domain of attraction by using invariant ellipsoids has been extensively studied, e.g., in [4, 5, 7, 14, 17, 24, 25]. More recently, we developed a new sufficient condition for an ellipsoid to be invariant in [11] (see also [8]). It was shown that this condition is less conservative than the existing conditions resulting from the circle criterion or the vertex analysis. The most important feature of this new condition is that it can be expressed as LMIs in terms of all the varying parameters and hence can easily be used for controller synthesis. A recent discovery makes this condition even more attractive. In [9], we showed that for the single input case, this condition is also necessary, thus the largest ellipsoid obtained with the LMI approach is actually the largest one.

In this paper, we will introduce a new type of Lyapunov function which is based on a group of quadratic functions. This is motivated by problems arising from estimating the domain of attraction and constructing controllers to enlarge the domain of attraction. Suppose that there are a group of invariant ellipsoids of the closed-loop system under a saturated feedback law. It is clear that the union of this group of ellipsoids is also an invariant set of the closed system. The

[1] A linear system is said to be semi-stable if all its poles are in the closed left-half plane.

question whether the convex hull of this group of ellipsoids, a set potentially much larger than the union, is invariant remains unclear. Another problem is related to enlarging the domain of attraction by merging two or more feedback laws. Suppose that we have two ellipsoids, each invariant under different feedback laws. In [13], we have shown that a switching feedback law can be constructed to make the union of the two ellipsoids invariant. We would further like to make the convex hull invariant, possibly with a continuous feedback law. Although the discontinuity of the switching feedback law in [13] does not cause chattering and guarantees the existence and uniqueness of solution, a continuous feedback law would be more appealing.

In this paper, we will study the invariance of the convex hull of a group of ellipsoids by using a new Lyapunov function which is based on a group of quadratic functions. The Lyapunov function is defined in such a way that its level set is the convex hull of a group of ellipsoids. A nice feature of this function is that it is continuously differentiable in the state x. This makes it possible to construct continuous feedback laws based on the gradient of the function or on a given set of linear feedback laws.

We first use the Lyapunov function to study the invariant sets of a linear system under the control of a saturated linear feedback. We will show that if each ellipsoid in a group is invariant by the condition of [11], then their convex hull is also invariant. We then extend this result to systems with a class of saturation-like nonlinearities. We will further show that if each ellipsoid in a group can be made invariant with saturating actuators, then their convex hull can also be made invariant. Finally, we will construct, from a group of saturated linear feedback laws, a continuous nonlinear feedback law which makes the convex hull invariant. For the convex hull of two ellipsoids, an explicit formula for the feedback law is derived.

The composite quadratic function is motivated from the study of control systems with saturating actuators and state constraints, it is a potential tool to handle more general nonlinearities, as will be demonstrated in this paper.

This paper is organized as follows. In Section 2, we introduce the composite quadratic Lyapunov function and show that this function is continuously differentiable and its level set is the convex hull of a group of ellipsoids. In Sections 3 – 5, we use these properties of the Lyapunov function to study the set invariance of linear systems with input and state constraints. In particular, we will show in Section 3 that under a given saturated linear feedback, the convex

hull of a group of invariant ellipsoids is also invariant. In Section 4, we will study the controlled invariance of the convex hull. In Section 5, we will present a method for constructing a nonlinear continuous controller which makes the convex hull invariant. Section 6 draws the conclusions to this paper.

Notation: We use $\text{sat}(\cdot)$ to denote the standard vector valued saturation function. For $u \in \mathbf{R}^m$, the ith component of $\text{sat}(u)$ is $\{\text{sat}(u)\}_i = \text{sign}(u_i)\min\{1, |u_i|\}$. We use $|u|_\infty$ and $|u|_2$ to denote respectively the infinity norm and the 2-norm.

Let $P \in \mathbf{R}^{n\times n}$ be a positive-definite matrix and ρ be a positive number. Denote

$$\mathcal{E}(P,\rho) := \{x \in \mathbf{R}^n : x^{\mathrm{T}}Px \le \rho\}.$$

For simplicity, we use $\mathcal{E}(P)$ to denote $\mathcal{E}(P,1)$. For a matrix $F \in \mathbf{R}^{m\times n}$, denote the ith row of F as f_i and define

$$\mathcal{L}(F) := \{x \in \mathbf{R}^n : \ |f_i x| \le 1, i \in [1,m]\}.$$

If F is the feedback matrix, then $\mathcal{L}(F)$ is the region in the state space where the control $u = \text{sat}(Fx)$ is linear in x. For $x_0 \in \mathbf{R}^n$ and $r \ge 0$, denote $\mathcal{B}(x_0,r) = \{x \in \mathbf{R}^n : \ |x - x_0|_2 \le r\}$.

2 The composite quadratic function

With a positive definite matrix $P \in \mathbf{R}^{n\times n}$, a quadratic function can be defined as $V(x) = x^{\mathrm{T}}Px$. For a positive number ρ, a level set of $V(x)$, denoted $L_V(\rho)$, is

$$L_V(\rho) := \left\{x \in \mathbf{R}^n : V(x) \le \rho\right\} = \mathcal{E}(P,\rho).$$

In this paper, we will be interested in a function determined by a group of positive definite matrices $P_1, P_2, \cdots, P_N \in \mathbf{R}^{n\times n}$. Let $Q_j = P_j^{-1}, j \in [1,N]$. For a vector $\gamma \in \mathbf{R}^N$, define

$$Q(\gamma) := \sum_{j=1}^{N} \gamma_j Q_j, \quad P(\gamma) := Q^{-1}(\gamma).$$

Let

$$\Gamma = \left\{\gamma \in \mathbf{R}^N : \ \gamma_1 + \gamma_2 + \cdots + \gamma_N = 1, \gamma_j \ge 0, j \in [1,N]\right\}.$$

It is easy to see that $Q(\gamma), P(\gamma) > 0$ for all $\gamma \in \Gamma$ and these two matrix functions are analytic in $\gamma \in \Gamma$. The composite quadratic function is defined as

$$V_c(x) := \min_{\gamma\in\Gamma} x^{\mathrm{T}}P(\gamma)x. \tag{1}$$

For $\rho > 0$, the level set of $V_c(x)$ is

$$L_{V_c}(\rho) := \left\{ x \in \mathbf{R}^n : \ V_c(x) \le \rho \right\}.$$

A very useful property of this composite quadratic function is that its level set is the convex hull of the level sets of $x^{\mathrm{T}} P_j x$, the ellipsoids $\mathcal{E}(P_j, \rho), j \in [1, N]$. Another nice property of $V_c(x)$ is that it is continuously differentiable. In order to establish these results, we need some simple preliminaries.

Fact 1 [8, 10]. *For a vector $f_0 \in \mathbf{R}^{1\times n}$ and a matrix $P > 0$, $\mathcal{E}(P) \subset \mathcal{L}(f_0)$ if and only if*

$$f_0 P^{-1} f_0^{\mathrm{T}} \le 1 \iff \begin{bmatrix} 1 & f_0 P^{-1} \\ P^{-1} f_0^{\mathrm{T}} & P^{-1} \end{bmatrix} \ge 0.$$

1) The equality $f_0 P^{-1} f_0^{\mathrm{T}} = 1$ holds if and only if the ellipsoid $\mathcal{E}(P)$ touches the hyperplane $f_0 x = 1$ at $x_0 = P^{-1} f_0^{\mathrm{T}}$ (the only intersection), i.e.,

$$1 = f_0 x_0 > f_0 x \quad \forall x \in \mathcal{E}(P) \setminus \{x_0\}.$$

2) If $f_0 P^{-1} f_0^{\mathrm{T}} < 1$, then

$$\begin{bmatrix} 1 & f_0 P^{-1} \\ P^{-1} f_0^{\mathrm{T}} & P^{-1} \end{bmatrix} > 0$$

and the ellipsoid $\mathcal{E}(P)$ lies strictly between the hyperplanes $f_0 x = 1$ and $f_0 x = -1$ without touching them.

A dual result, which will be useful, can be obtained by exchanging the roles of f_0 and x_0^{T}. Given x_0 and suppose that $x_0^{\mathrm{T}} P x_0 = 1$, $f_0^{\mathrm{T}} = P x_0$, then

$$1 = f_0 x_0 > f x_0 \quad \forall f^{\mathrm{T}} \in \mathcal{E}(P^{-1}) \setminus \{f_0^{\mathrm{T}}\}.$$

For $F \in \mathbf{R}^{m\times n}$, $\mathcal{L}(F) = \cap_{i=1}^m \mathcal{L}(f_i)$ and $\mathcal{E}(P) \subset \mathcal{L}(F)$ if and only if $f_i P^{-1} f_i^{\mathrm{T}} \le 1$ for all $i \in [1, m]$. Denote the convex hull of the ellipsoids $\mathcal{E}(P_j, \rho), j \in [1, N]$ as

$$\mathrm{co}\left\{ \mathcal{E}(P_j, \rho), j \in [1, N] \right\} := \left\{ \sum_{j=1}^N \gamma_j x_j : \ x_j \in \mathcal{E}(P_j, \rho), \gamma \in \Gamma \right\}.$$

Then we have

Theorem 1 $L_{V_c}(\rho) = \mathrm{co}\left\{ \mathcal{E}(P_j, \rho), j \in [1, N] \right\} = \bigcup_{\gamma\in\Gamma} \mathcal{E}(P(\gamma), \rho)$.

Proof. It is obvious that $V_c(kx) = k^2 V_c(x)$, so $L_{V_c}(\rho) = \sqrt{\rho}\, L_{V_c}(1)$. Since $\mathcal{E}(P,\rho) = \sqrt{\rho}\,\mathcal{E}(P)$, it suffices to show that

$$L_{V_c}(1) = \mathrm{co}\left\{\mathcal{E}(P_j), j \in [1,N]\right\} = \bigcup_{\gamma\in\Gamma} \mathcal{E}(P(\gamma)).$$

We first show that $\mathrm{co}\left\{\mathcal{E}(P_j), j \in [1,N]\right\} \subset L_{V_c}(1)$. Suppose that $x \in \mathrm{co}\left\{\mathcal{E}(P_j), j \in [1,N]\right\}$, then there exists a $\gamma \in \Gamma$ and $x_j \in \mathcal{E}(P_j), j \in [1,N]$, such that

$$x = \gamma_1 x_1 + \gamma_2 x_2 + \cdots + \gamma_N x_N.$$

Since $x_j \in \mathcal{E}(P_j)$, we have $x_j^{\mathrm{T}} P_j x_j \le 1$, which is equivalent (by Schur complement) to

$$\begin{bmatrix} 1 & x_j^{\mathrm{T}} \\ x_j & Q_j \end{bmatrix} \ge 0, \quad j \in [1,N].$$

Recalling that $Q(\gamma) = \gamma_1 Q_1 + \gamma_2 Q_2 + \cdots + \gamma_N Q_N$, we have, by the convexity,

$$\begin{bmatrix} 1 & x^{\mathrm{T}} \\ x & Q(\gamma) \end{bmatrix} \ge 0,$$

which implies that $x^{\mathrm{T}} P(\gamma) x \le 1$. It follows that $V_c(x) \le 1$ and $x \in L_{V_c}(1)$.

We next prove that $\bigcup_{\gamma\in\Gamma} \mathcal{E}(P(\gamma)) \subset \mathrm{co}\left\{\mathcal{E}(P_j), j \in [1,N]\right\}$. It suffices to show that $\mathcal{E}(P(\gamma)) \subset \mathrm{co}\left\{\mathcal{E}(P_j), j \in [1,N]\right\}$ for every $\gamma \in \Gamma$. Let γ be any vector in the set Γ. One way for the proof is to show that for any $x \in \mathcal{E}(P(\gamma))$, there exist a $\hat{\gamma} \in \Gamma$ and a group of $x_j \in \mathcal{E}(P_j), j \in [1,N]$, such that $x = \hat{\gamma}_1 x_1 + \cdots + \hat{\gamma}_N x_N$. But this approach seems to be not easy. Instead, we will prove $\mathcal{E}(P(\gamma)) \subset \mathrm{co}\left\{\mathcal{E}(P_j), j \in [1,N]\right\}$ by using Fact 1. Suppose on the contrary that there exists an $x_0 \in \mathcal{E}(P(\gamma))$ and $x_0 \notin \mathrm{co}\left\{\mathcal{E}(P_j), j \in [1,N]\right\}$. Then, there exists a vector $h \in \mathbf{R}^{1\times n}$ such that

$$hx_0 > hx \quad \forall x \in \mathrm{co}\left\{\mathcal{E}(P_j), j \in [1,N]\right\}.$$

Let $x_* \in \mathcal{E}(P(\gamma))$ be the point such that $hx_* = \max\{hx : \ x \in \mathcal{E}(P(\gamma))\}$ and let $h_* = h/(hx_*)$, then $\max\{h_* x : \ x \in \mathcal{E}(P(\gamma))\} = 1$ and the hyperplane $h_* x = 1$ touches the ellipsoid $\mathcal{E}(P(\gamma))$ at only one point x_*. It is obvious that

$$h_* x < h_* x_0 \le h_* x_* = 1 \quad \forall x \in \mathrm{co}\left\{\mathcal{E}(P_j), j \in [1,N]\right\}.$$

This implies that each ellipsoid $\mathcal{E}(P_j)$ is between the hyperplanes $h_* x = 1$ and $h_* x = -1$ without touching them. By Fact 1, we have

$$\begin{bmatrix} 1 & h_* Q_j \\ Q_j h_*^{\mathrm{T}} & Q_j \end{bmatrix} > 0, \quad j \in [1,N]. \tag{2}$$

By the convexity, we should have

$$\begin{bmatrix} 1 & h_* Q(\gamma) \\ Q(\gamma) h_*^{\mathrm{T}} & Q(\gamma) \end{bmatrix} > 0 \iff h_* Q(\gamma) h_*^{\mathrm{T}} < 1. \tag{3}$$

However, since the hyperplane $h_* x = 1$ touches the ellipsoid $\mathcal{E}(P(\gamma))$, we also have

$$h_* Q(\gamma) h_*^{\mathrm{T}} = 1,$$

which contradicts (3). Therefore, we conclude that there exists no $x_0 \in \mathcal{E}(P(\gamma))$ such that $x_0 \notin \mathrm{co}\left\{\mathcal{E}(P_j), j \in [1, N]\right\}$. This proves that $\mathcal{E}(P(\gamma)) \subset \mathrm{co}\left\{\mathcal{E}(P_j), j \in [1, N]\right\}$.

Finally, we show that $L_{V_c}(1) \subset \bigcup_{\gamma \in \Gamma} \mathcal{E}(P(\gamma))$. Suppose that $x \in L_{V_c}(1)$. Then there exists a $\gamma \in \Gamma$ such that $x^{\mathrm{T}} P(\gamma) x \leq 1$. It follows that $x \in \mathcal{E}(P(\gamma)) \subset \bigcup_{\gamma \in \Gamma} \mathcal{E}(P(\gamma))$.

Combining the above set inclusion results, we obtain

$$L_{V_c}(1) \subset \bigcup_{\gamma \in \Gamma} \mathcal{E}(P(\gamma)) \subset \mathrm{co}\left\{\mathcal{E}(P_j), j \in [1, N]\right\} \subset L_{V_c}(1).$$

Therefore,

$$L_{V_c}(1) = \bigcup_{\gamma \in \Gamma} \mathcal{E}(P(\gamma)) = \mathrm{co}\left\{\mathcal{E}(P_j), j \in [1, N]\right\}.$$

□

Theorem 2 *The function $V_c(x)$ is continuously differentiable in x. Let γ^* be an optimal γ such that $x^{\mathrm{T}} P(\gamma^*) x = \min_{\gamma \in \Gamma} x^{\mathrm{T}} P(\gamma) x$, then*

$$\frac{\partial V_c}{\partial x} = 2 P(\gamma^*) x.$$

Proof. Let us first establish a property about the differentiability which will simplify the proof. Suppose that $V_c(x)$ is differentiable at x_0 with partial derivative $\frac{\partial V_c}{\partial x}\Big|_{x=x_0}$ and let k be given. Since $V_c(kx) = k^2 V_c(x)$ for all $x \in \mathrm{R}^n$, we have

$$\begin{aligned} V_c(kx_0 + \Delta x) - V_c(kx_0) &= k^2 \left(V_c\left(x_0 + \Delta x/k\right) - V_c(x_0)\right) \\ &= k^2 \left(\left(\frac{\partial V_c}{\partial x}\Big|_{x=x_0} \right)^{\mathrm{T}} \Delta x/k + o(|\Delta x/k|) \right) \\ &= k \left(\frac{\partial V_c}{\partial x}\Big|_{x=x_0} \right)^{\mathrm{T}} \Delta x + o(|\Delta x|), \end{aligned}$$

where $|\cdot|$ can be any norm. It follows that

$$\frac{\partial V_c}{\partial x}\Big|_{x=kx_0} = k \, \frac{\partial V_c}{\partial x}\Big|_{x=x_0}.$$

In view of this equality, we only need to consider those x on the boundary of $L_{V_c}(1)$, $\partial L_{V_c}(1)$. Here we use "∂" to denote both the boundary of a set and the partial derivative.

Since the set $L_{V_c}(1)$ is convex, for each $x_0 \in \partial L_{V_c}(1)$, there exists a vector $h_0 \in \mathbf{R}^{1\times n}$ such that

$$1 = h_0 x_0 \geq h_0 x \ \ \forall x \in L_{V_c}(1), \tag{4}$$

which implies that $L_{V_c}(1) \subset \mathcal{L}(h_0)$. Let γ^* be an optimal γ such that

$$x_0^{\mathrm{T}} P(\gamma^*) x_0 = \min_{\gamma\in\Gamma} x_0^{\mathrm{T}} P(\gamma) x_0 = 1.$$

Since $\mathcal{E}(P(\gamma^*)) \subset L_{V_c}(1)$, it follows that $\mathcal{E}(P(\gamma^*)) \subset \mathcal{L}(h_0)$ and

$$h_0 x_0 \geq h_0 x \ \ \forall x \in \mathcal{E}(P(\gamma^*)).$$

By Fact 1, the hyperplane $h_0 x = 1$ is tangential to the ellipsoid $\mathcal{E}(P(\gamma^*))$ at x_0. Therefore, such a vector h_0 is uniquely determined by x_0. Denote $P_0 = P(\gamma^*)$. Combining the above results, we have

$$\mathcal{E}(P_0) \subset L_{V_c}(1) \subset \mathcal{L}(h_0). \tag{5}$$

By Fact 1, we also have

$$x_0 = P_0^{-1} h_0^{\mathrm{T}}, \quad h_0 P_0^{-1} h_0^{\mathrm{T}} = 1, \tag{6}$$

and

$$h_0 x_0 > h x_0 \quad \forall h^{\mathrm{T}} \in \mathcal{E}(P_0^{-1}) \setminus \{h_0^{\mathrm{T}}\}. \tag{7}$$

Now we show that

$$\left.\frac{\partial V_c}{\partial x}\right|_{x=x_0} = 2h_0^{\mathrm{T}} = 2P_0 x_0.$$

From (5), it follows that for all $x \in \partial L_{V_c}(1)$, $V_c(x) = 1$, $h_0 x \leq 1$ and $x^{\mathrm{T}} P_0 x \geq 1$, i.e.,

$$h_0 x \leq 1 = V_c^{\frac{1}{2}}(x) = 1 \leq (x^{\mathrm{T}} P_0 x)^{\frac{1}{2}} \quad \forall x \in \partial L_{V_c}(1).$$

Since $V_c^{\frac{1}{2}}(kx) = kV_c^{\frac{1}{2}}(kx), h_0(kx) = kh_0 x$ and $((kx)^{\mathrm{T}} P_0 (kx))^{\frac{1}{2}} = k(x^{\mathrm{T}} P_0 x)^{\frac{1}{2}}$, we have

$$h_0(kx) \leq V_c^{\frac{1}{2}}(kx) \leq ((kx)^{\mathrm{T}} P_0 (kx))^{\frac{1}{2}} \quad \forall k \geq 0, \ \ x \in \partial L_{V_c}(1).$$

Since every point in $\mathbf{R}^n$ can be written as kx for some $k \geq 0$ and $x \in \partial L_{V_c}(1)$, we have

$$h_0 x \leq V_c^{\frac{1}{2}}(x) \leq (x^{\mathrm{T}} P_0 x)^{\frac{1}{2}} \quad \forall x \in \mathbf{R}^n. \tag{8}$$

Recalling that $(x_0^{\mathrm{T}} P_0 x_0)^{\frac{1}{2}} = 1$ and from (6), we have

$$\left.\frac{\partial (x^{\mathrm{T}} P_0 x)^{\frac{1}{2}}}{\partial x}\right|_{x=x_0} = P_0 x_0 = h_0^{\mathrm{T}}.$$

Hence,

$$((x_0 + \Delta x)^{\mathrm{T}} P_0 (x_0 + \Delta x))^{\frac{1}{2}} = 1 + h_0 \Delta x + o(|\Delta x|). \tag{9}$$

Recalling from (4) that $h_0 x_0 = 1$, we have

$$h_0(x_0 + \Delta x) = 1 + h_0 \Delta x. \tag{10}$$

Combining (8), (9), (10) and that $V_c^{\frac{1}{2}}(x_0) = 1$, we obtain

$$V_c^{\frac{1}{2}}(x_0 + \Delta x) = 1 + h_0 \Delta x + o(|\Delta x|) = V_c^{\frac{1}{2}}(x_0) + h_0 \Delta x + o(|\Delta x|),$$

which implies that $V_c^{\frac{1}{2}}(x)$ is differentiable at $x = x_0$ and the partial derivative is given by h_0^{T}. It follows that $V_c(x)$ is differentiable at $x = x_0$ with the partial derivative given by

$$\left.\frac{\partial V_c}{\partial x}\right|_{x=x_0} = 2V_c^{\frac{1}{2}}(x_0) h_0^{\mathrm{T}} = 2h_0^{\mathrm{T}} = 2P_0 x_0 = 2P(\gamma^*) x_0.$$

In the rest of the proof, we show that $\partial V_c / \partial x$ is continuous in x. Since

$$\left.\frac{\partial V_c}{\partial x}\right|_{x=kx_0} = k \left.\frac{\partial V_c}{\partial x}\right|_{x=x_0},$$

it suffices to prove the continuity on the surface $\partial L_{V_c}(1)$. Let $x_0 \in \partial L_{V_c}(1)$ with h_0 and P_0 defined as above. Then we have $h_0^{\mathrm{T}} = P_0 x_0$ and $h_0 P_0^{-1} h_0^{\mathrm{T}} = h_0 x_0 = 1$. Consider $v \in \partial L_{V_c}(1)$. Let

$$h(v) = \frac{1}{2} \left(\left.\frac{\partial V_c}{\partial x}\right|_{x=v} \right)^{\mathrm{T}}.$$

Then $h(x_0) = h_0$ and by the foregoing proof, we have $h(v)v = 1$, $L_{V_c}(1) \subset \mathcal{L}(h(v))$ and hence $\mathcal{E}(P_0) \subset \mathcal{L}(h(v))$. By Fact 1,

$$h(v) P_0^{-1} h^{\mathrm{T}}(v) \leq 1 \iff h^{\mathrm{T}}(v) \in \mathcal{E}(P_0^{-1}) \quad \forall v \in \partial L_{V_c}(1). \tag{11}$$

It follows that there exists a positive number d_0 such that $|h(v)|_2 \leq d_0$ for all $v \in \partial L_{V_c}(1)$. Now suppose on the contrary that $h(v)$ is not continuous at $v = x_0$ on the surface $\partial L_{V_c}(1)$. Then there exists a positive number ε such that for any arbitrarily small number $\delta > 0$, there exists a

$v \in \mathcal{B}(x_0, \delta) \cap \partial L_{V_c}(1)$ satisfying $h^{\mathrm{T}}(v) \notin \mathcal{B}(h_0, \varepsilon)$. Note that ε is fixed and δ can be arbitrarily small. What we will show next is that the assumption of the existence of such ε will cause contradiction.

From above, we have

$$h_0 P_0^{-1} h_0^{\mathrm{T}} = 1, \quad h(v) P_0^{-1} h(v)^{\mathrm{T}} \leq 1, \quad h_0 x_0 = 1, \quad h(v) v = 1.$$

By Fact 1, we know that

$$h_0 x_0 > h x_0 \quad \forall\, h^{\mathrm{T}} \in \mathcal{E}(P_0^{-1}) \setminus h_0^{\mathrm{T}}.$$

Then it is clear that

$$\sup_{h^{\mathrm{T}} \in \mathcal{E}(P_0^{-1}) \setminus \mathcal{B}(h_0, \varepsilon)} h x_0 =: k^* < 1. \tag{12}$$

Hence for all $h^{\mathrm{T}} \in \mathcal{E}(P_0^{-1}) \setminus \mathcal{B}(h_0, \varepsilon)$, $h x_0 \leq k^* < 1$. On the other hand, for all $v \in \mathcal{B}(x_0, \delta) \cap \partial L_{V_c}(1)$, we have $h(v) v = 1$ and

$$|h(v) x_0 - h(v) v| \leq |h(v)|_2 \delta \leq d_0 \delta.$$

Hence

$$h(v) x_0 \geq h(v) v - d_0 \delta = 1 - d_0 \delta.$$

Let δ be chosen such that $d_0 \delta < 1 - k^*$. Then

$$h(v) x_0 \geq 1 - d_0 \delta > k^* \quad \forall\, v \in \mathcal{B}(x_0, \delta) \cap \partial L_{V_c}(1). \tag{13}$$

By assumption, there exists a $v \in \mathcal{B}(x_0, \delta) \cap \partial L_{V_c}(1)$ such that $h^{\mathrm{T}}(v) \in \mathcal{E}(P_0^{-1}) \setminus \mathcal{B}(h_0, \varepsilon)$ (note that $h^{\mathrm{T}}(v) \in \mathcal{E}(P_0^{-1})$ is from (11)). It follows from (12) that $h(v) x_0 \leq k^*$, which contradicts (13). Therefore, $h(v)$ must be continuous at x_0.

Finally, we note that the partial derivative is continuous at $x = 0$ with $\partial V_c / \partial x|_{x=0} = 0$. □

We next consider some computational issues with regard to the function $V_c(x)$. The function $V_c(x)$ and an optimal γ can be computed with LMI toolbox:

$$\begin{aligned} V_c(x) \;=\; & \min_{\gamma} \alpha \\ & \text{s.t.} \quad \begin{bmatrix} \alpha & x^{\mathrm{T}} \\ x & \sum_{j=1}^{N} \gamma_j Q_j \end{bmatrix} \geq 0, \\ & \sum_{j=1}^{N} \gamma_j = 1, \;\; \gamma_j \geq 0. \end{aligned} \tag{14}$$

If we only have two ellipsoids, there exists a more efficient way to obtain $V_c(x)$ through computing the generalized eigenvalues of certain matrices. In this case, we have

$$V_c(x) = \min_{\lambda \in [0,1]} x^{\mathrm{T}}(\lambda Q_1 + (1-\lambda)Q_2)^{-1}x.$$

Denote $\alpha(\lambda, x) = x^{\mathrm{T}}(\lambda Q_1 + (1-\lambda)Q_2)^{-1}x$.

Proposition 1 *Assume that $Q_1 - Q_2$ is nonsingular. For every $x \in \mathbf{R}^n$, the function $\alpha(\lambda, x)$ is strictly convex in $\lambda \in [0,1]$ and there exists a unique $\lambda^* \in [0,1]$ such that $\alpha(\lambda^*, x) = \min_{\lambda \in [0,1]} \alpha(\lambda, x)$. Moreover, λ^* is a continuous function of x.*

Remark 1 The assumption that $Q_1 - Q_2$ is nonsingular is without loss of generality. For the case where $n = 2$, $\det(Q_1 - Q_2) = 0$ implies that either $Q_1 \geq Q_2$ or $Q_1 \leq Q_2$. If $Q_1 \geq Q_2$, then $P_1 \leq P_2$ and $V_c(x) = V_1(x)$, which is not meaningful.

Proof of Proposition 1. The first and the second partial derivatives of $\alpha(\lambda, x)$ with respect to λ are

$$\frac{\partial \alpha}{\partial \lambda} = x^{\mathrm{T}}(\lambda Q_1 + (1-\lambda)Q_2)^{-1}(Q_2 - Q_1)(\lambda Q_1 + (1-\lambda)Q_2)^{-1}x$$

and

$$\frac{\partial^2 \alpha}{\partial \lambda^2} = 2x^{\mathrm{T}}(\lambda Q_1 + (1-\lambda)Q_2)^{-1}(Q_2 - Q_1)(\lambda Q_1 + (1-\lambda)Q_2)^{-1}(Q_2 - Q_1)(\lambda Q_1 + (1-\lambda)Q_2)^{-1}x.$$

Since $(\lambda Q_1 + (1-\lambda)Q_2)^{-1} > 0$ for all $\lambda \in [0,1]$ and $Q_2 - Q_1$ is nonsingular, we have $\partial^2\alpha/\partial\lambda^2 > 0$ for all $\lambda \in [0,1]$ and $x \in \mathbf{R}^n$. This shows that $\alpha(\lambda, x)$ is strictly convex in λ and establishes the uniqueness of $\lambda^* \in [0,1]$ such that $\alpha(\lambda^*, x) = \min_{\lambda \in [0,1]} \alpha(\lambda, x)$. Denote the function of λ^* in x as $\lambda^*(x)$. Consider $x_0 \in \mathbf{R}^n$. If $\lambda^*(x_0) \in (0,1)$, then for x in a neighborhood of x_0, $\alpha(\lambda, x)$ has a unique minimum at $\lambda^* \in (0,1)$ satisfying

$$\left.\frac{\partial \alpha}{\partial \lambda}\right|_{\lambda=\lambda^*} = x^{\mathrm{T}}(\lambda^* Q_1 + (1-\lambda^*)Q_2)^{-1}(Q_2 - Q_1)(\lambda^* Q_1 + (1-\lambda^*)Q_2)^{-1}x = 0.$$

Since $\frac{\partial^2\alpha}{\partial\lambda^2} \neq 0$, by implicit function theorem, $\lambda^*(x)$ is continuously differentiable at x_0. For those x_0 such that $\lambda^*(x_0) = 0$ (or 1), we have two possibilities

1. $\partial\alpha/\partial\lambda|_{\lambda=0, x=x_0} = 0$. Then as x varies in a neighborhood of x_0, $\partial\alpha/\partial\lambda = 0$ occurs in a neighborhood of $\lambda = 0$. In this neighborhood of x_0, if $\partial\alpha/\partial\lambda = 0$ for some $\lambda < 0$, then we must have $\lambda^*(x) = 0$ and, if $\partial\alpha/\partial\lambda = 0$ for some $\lambda > 0$, then $\lambda^*(x) > 0$ and must be in a neighborhood of $\lambda = 0$. These show the continuity of $\lambda^*(x)$ for this case.

2. $\partial\alpha/\partial\lambda|_{\lambda=0,x=x_0} > 0$. Then we have $\partial\alpha/\partial\lambda|_{\lambda=0} > 0$ for all x in a neighborhood of x_0. By convexity, we have $\lambda^*(x) = 0$ for all x in this neighborhood of x_0, which also confirms the continuity of $\lambda^*(x)$. □

Let $\gamma^* = (\lambda^*, 1-\lambda^*)$. By Proposition 1, γ^* is the unique value such that $x^{\mathrm{T}}(P(\gamma^*))x = V_c(x)$ and it depends on x continuously. This will be useful in Section 5 to our construction of continuous feedback laws from two given feedback laws.

Here we provide a method for computing the value λ such that $\partial\alpha/\partial\lambda = 0$ for a given x. By Proposition 1, this will give us λ^* and γ^*.

Proposition 2 *Let $x \in \mathbf{R}^n$ and $Q_1, Q_2 > 0$ be given. Assume that $Q_1 - Q_2$ is nonsingular. Let $U \in \mathbf{R}^{n\times n}$ be such that $U^{\mathrm{T}}U = I$ and $U^{\mathrm{T}}xx^{\mathrm{T}}U = \mathrm{diag}\{c, 0, \cdots, 0\}$. Let $\hat{Q}_1 = U^{\mathrm{T}}Q_1U, \hat{Q}_2 = U^{\mathrm{T}}Q_2U$ and partition $\hat{Q}_1$ and $\hat{Q}_2$ as*

$$\hat{Q}_1 = [\hat{q}_1 \;\; \hat{Q}_{12}], \quad \hat{Q}_2 = [\hat{q}_2 \;\; \hat{Q}_{22}], \quad \hat{q}_1, \hat{q}_2 \in \mathbf{R}^{n\times 1}.$$

Then $\partial\alpha/\partial\lambda = 0$ at $\lambda \in [0,1]$ if and only if

$$\det \begin{bmatrix} \lambda(\hat{Q}_{12} - \hat{Q}_{22}) + \hat{Q}_{22} & \hat{Q}_1 - \hat{Q}_2 \\ 0_{(n-1)\times(n-1)} & \lambda(\hat{Q}_{12} - \hat{Q}_{22}^{\mathrm{T}})^{\mathrm{T}} + \hat{Q}_{22}^{\mathrm{T}} \end{bmatrix} = 0. \tag{15}$$

The proof of Proposition 2 can be found in Appendix A. All the λ's satisfying (15) can be obtained by computing the generalized eigenvalues of the matrix pair (X, Y) where

$$X = \begin{bmatrix} \hat{Q}_{12} - \hat{Q}_{22} & 0_{n\times n} \\ 0_{(n-1)\times(n-1)} & \hat{Q}_{12}^{\mathrm{T}} - \hat{Q}_{22}^{\mathrm{T}} \end{bmatrix}, \quad Y = \begin{bmatrix} \hat{Q}_{22} & \hat{Q}_1 - \hat{Q}_2 \\ 0_{(n-1)\times(n-1)} & \hat{Q}_{22}^{\mathrm{T}} \end{bmatrix}.$$

By Propositions 1 and 2, (X, Y) has at most one generalized eigenvalue in $[0, 1]$. If there is none in $[0, 1]$, then $\lambda^* = 0$ or 1. Experience shows that computing the matrices X and Y and their generalized eigenvalues requires much less time than solving the LMI problem (14).

Fig. 1 illustrates a two dimensional level set which is the convex hull of three ellipsoids. Fig. 2 plots the values of $\gamma^* = (\gamma_1^*, \gamma_2^*, \gamma_3^*)$ as x varies along the boundary of $L_{V_c}(1)$ in counterclockwise direction, where the abscissa is the angle of x (from 0 to π). From Fig. 1, we see that parts of $\partial L_{V_c}(1)$ overlap with segments of $\mathcal{E}(P_i), i = 1, 2, 3$. The overlapped segments correspond to the intervals in Fig. 2 where $\gamma_i^* = 1$. Fig. 3 illustrates a three dimensional level set. It is also the convex hull of three ellipsoids.

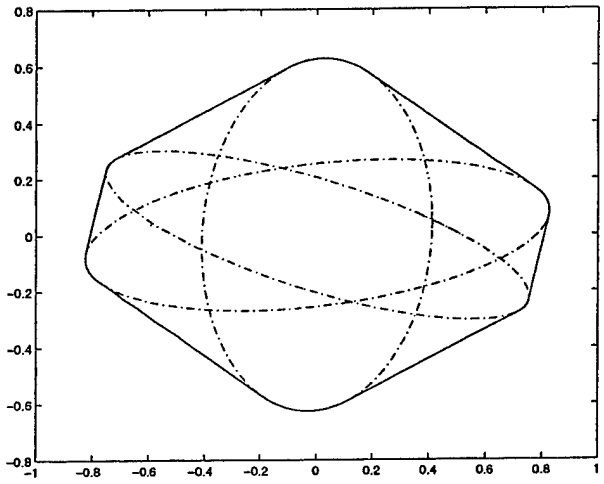


Figure 1: A two dimensional level set $L_{V_c}(1)$.

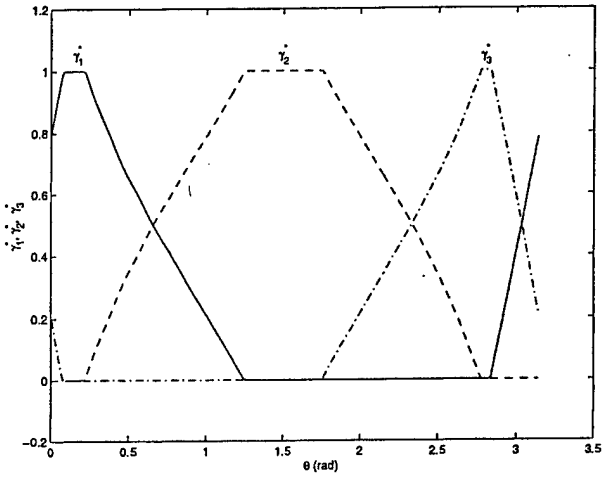


Figure 2: γ^* along the boundary of $L_{V_c}(1)$.

3 Invariant sets under a given saturated linear feedback

This section will also include some results on a class of saturation-like nonlinearities.

Consider the open-loop system

$$\dot{x} = Ax + Bu, \quad x \in \mathbf{R}^n, \ u \in \mathbf{R}^m, \tag{16}$$

where u is the output of saturating actuators and is assumed to satisfy the bound $|u|_\infty \leq 1$. The state constraint is represented by a convex set Ω_0, which contains the origin in its interior. It is required that the system operate in Ω_0 for all $t \geq 0$. Suppose that we have a stabilizing

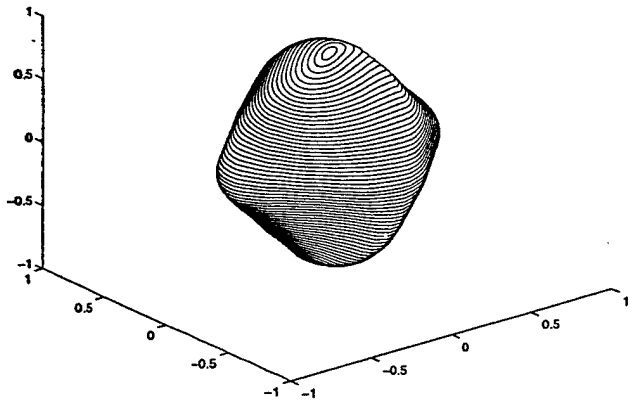


Figure 3: A three dimensional level set $L_{V_c}(1)$.

feedback law $u = \text{sat}(Fx)$, under which the closed-loop system is

$$\dot{x} = Ax + B\text{sat}(Fx). \tag{17}$$

Since Ω_0 is generally not an invariant set, we would like to determine a maximal subset of Ω_0 such that, for any initial state x_0 in this subset, the state trajectory of (17) will stay in it and converge to the origin. Because of the intrinsic difficulty involved in determining the maximal invariant set inside Ω_0, alternative problems have been formulated such as determining the invariance of ellipsoids and searching for the largest invariant ellipsoid inside Ω_0.

In [11], we derived a sufficient condition for checking the invariance of a given ellipsoid. This condition turns out to be also necessary for single input systems [9]. We need some notation to state the set invariance condition of [11]. Let $\mathcal{D}$ be the set of $m \times m$ diagonal matrices whose diagonal elements are either 1 or 0. There are 2^m elements in $\mathcal{D}$. Suppose that each element of $\mathcal{D}$ is labeled as D_i, $i = 1, 2, \cdots, 2^m$. Then, $\mathcal{D} = \left\{D_i : i \in [1, 2^m]\right\}$. Denote $D_i^- = I - D_i$. Clearly, D_i^- is also an element of $\mathcal{D}$ if $D_i \in \mathcal{D}$. Given two vectors, $u, v \in \mathbf{R}^m$,

$$\left\{D_i u + D_i^- v : i \in [1, 2^m]\right\}$$

is the set of vectors formed by choosing some elements from u and the rest from v. Given two matrices $F, H \in \mathbf{R}^{m \times n}$,

$$\left\{D_i F + D_i^- H : i \in [1, 2^m]\right\}$$

is the set of matrices formed by choosing some rows from F and the rest from H.

Given a positive definite matrix P, let $V(x) = x^{\mathrm{T}}Px$. The ellipsoid $\mathcal{E}(P,\rho)$ is said to be contractively invariant if

$$\dot{V}(x) = 2x^{\mathrm{T}}P(Ax + B\mathrm{sat}(Fx)) < 0 \tag{18}$$

for all $x \in \mathcal{E}(P,\rho) \setminus \{0\}$. The invariance of $\mathcal{E}(P,\rho)$ can be defined by replacing "$<$" in (18) with "$\leq$". Clearly, if $\mathcal{E}(P,\rho)$ is contractively invariant, then for every initial state $x_0 \in \mathcal{E}(P,\rho)$, the state trajectory will converge to the origin and $\mathcal{E}(P,\rho)$ is in the domain of attraction.

Proposition 3 [11, 8] *Given an ellipsoid $\mathcal{E}(P,\rho)$, if there exists an $H \in \mathbf{R}^{m\times n}$ such that*

$$(A + B(D_iF + D_i^- H))^{\mathrm{T}}P + P(A + B(D_iF + D_i^- H)) \leq (<)0 \quad \forall i \in [1, 2^m], \tag{19}$$

and $\mathcal{E}(P,\rho) \subset \mathcal{L}(H)$, then $\mathcal{E}(P,\rho)$ is a (contractively) invariant set.

The condition in Proposition 3 is easy to check with LMI method. To impose the state constraint, we only need to require that $\mathcal{E}(P,\rho) \subset \Omega_0$. In the case that Ω_0 is a symmetric polytope, there exists a matrix $G_0 \in \mathbf{R}^{\ell\times n}$ for some integer ℓ such that $\Omega_0 = \mathcal{L}(G_0)$. In light of Fact 1, the requirement that $\mathcal{E}(P,\rho) \subset \Omega_0$ can be easily transformed into LMIs. In [11, 8], we also developed LMI methods for choosing the largest invariant ellipsoid with respect to some shape reference set, where the matrix P was taken as an optimizing parameter. The shape reference set could be a polygon or a fixed ellipsoid. It could also be a single point $x_0 \in \mathbf{R}^n$. In this case, the largest invariant ellipsoid inside Ω_0 is the one that includes αx_0 with the maximal $\alpha > 0$. By choosing different x_0, say, $x_{0,j}, j \in [1, N]$, we can obtain N optimized invariant ellipsoids $\mathcal{E}(P_j,\rho_j) \subset \Omega_0, j \in [1, N]$. It is easy to see that the union of these ellipsoids, $\bigcup_{j=1}^N \mathcal{E}(P_j,\rho_j)$, is also an invariant set inside Ω_0. But this union does not necessarily include the convex hull of $x_{0,j}, j \in [1, N]$. What is desired here is that the convex hull of the ellipsoids, $\mathrm{co}\{\mathcal{E}(P_j,\rho_j),\ j \in [1, N]\}$, is also an invariant set.

For simplicity and without loss of generality, we will consider a group of invariant ellipsoids $\mathcal{E}(P_j,\rho_j), j \in [1, N]$, with $\rho_j = 1$. The following theorem says that if each $\mathcal{E}(P_j)$ satisfies the condition of Proposition 3, then their convex hull, $\mathrm{co}\{\mathcal{E}(P_j), j \in [1, N]\}$, is also invariant.

Theorem 3 *Given a group of ellipsoids $\mathcal{E}(P_j), j \in [1, N]$. If there exist matrices $H_j, j \in [1, N]$ such that*

$$(A + B(D_iF + D_i^- H_j))^{\mathrm{T}}P_j + P_j(A + B(D_iF + D_i^- H_j)) \leq 0 \quad \forall i \in [1, 2^m],\ j \in [1, N], \tag{20}$$

and $\mathcal{E}(P_j) \subset \mathcal{L}(H_j)$, $j \in [1, N]$, *then* $\mathrm{co}\{\mathcal{E}(P_j), j \in [1, N]\}$ *is an invariant set. If "<" holds for each of the above inequalities, then for every initial state* $x_0 \in \mathrm{co}\{\mathcal{E}(P_j), j \in [1, N]\}$, *the state trajectory will converge to the origin.*

Proof. Let $Q_j = P_j^{-1}$ and $Z_j = H_j Q_j$. The inequalities in (20) are equivalent to

$$Q_j(A + BD_iF)^{\mathrm{T}} + (A + BD_iF)Q_j + Z_j^{\mathrm{T}} D_i^{-} B^{\mathrm{T}} + BD_i^{-} Z_j \leq 0 \quad \forall i \in [1, 2^m], \ \ j \in [1, N]. \quad (21)$$

The condition $\mathcal{E}(P_j) \subset \mathcal{L}(H_j)$, $j \in [1, N]$, can be written as

$$\begin{bmatrix} 1 & z_{jk} \\ z_{jk}^{\mathrm{T}} & Q_j \end{bmatrix} \geq 0, \quad j \in [1, N], \ \ k \in [1, m], \quad (22)$$

where z_{jk} is the kth row of the matrix Z_j. Consider $x_0 \in \mathrm{co}\{\mathcal{E}(P_j), j \in [1, N]\}$. There exists $x_j \in \mathcal{E}(P_j)$ and $\gamma_j \geq 0, j \in [1, N]$, such that $\gamma_1 + \gamma_2 + \cdots + \gamma_N = 1$ and $x_0 = \gamma_1 x_1 + \gamma_2 x_2 + \cdots + \gamma_N x_N$. Let $Q = \gamma_1 Q_1 + \gamma_2 Q_2 + \cdots + \gamma_N Q_N$ and $P = Q^{-1}$. Then by Theorem 1, $\mathcal{E}(P) \subset \mathrm{co}\{\mathcal{E}(P_j), j \in [1, N]\}$. From $x_j \in \mathcal{E}(P_j)$, we have $x_j^{\mathrm{T}} P_j x_j \leq 1$, which is equivalent to

$$\begin{bmatrix} 1 & x_j^{\mathrm{T}} \\ x_j & Q_j \end{bmatrix} \geq 0, \quad j \in [1, N].$$

By the convexity, we have

$$\begin{bmatrix} 1 & x_0^{\mathrm{T}} \\ x_0 & Q \end{bmatrix} \geq 0,$$

which implies that $x_0^{\mathrm{T}} P x_0 \leq 1$ and $x_0 \in \mathcal{E}(P)$.

Let $Z = \gamma_1 Z_1 + \gamma_2 Z_2 + \cdots + \gamma_N Z_N$, and z_k be the kth row of Z, then by (21), (22) and the convexity, we have

$$Q(A + BD_iF)^{\mathrm{T}} + (A + BD_iF)Q + Z^{\mathrm{T}} D_i^{-} B^{\mathrm{T}} + BD_i^{-} Z \leq 0 \quad \forall i \in [1, 2^m], \quad (23)$$

and

$$\begin{bmatrix} 1 & z_k \\ z_k^{\mathrm{T}} & Q \end{bmatrix} \geq 0, \ \ k \in [1, m]. \quad (24)$$

Let $H = ZQ^{-1} = ZP$. The inequalities in (23) and (24) can be rewritten as

$$(A + B(D_iF + D_i^{-} H))^{\mathrm{T}} P + P(A + B(D_iF + D_i^{-} H)) \leq 0, \quad i \in [1, 2^m], \quad (25)$$

and

$$\begin{bmatrix} 1 & h_k P^{-1} \\ P^{-1} h_k^{\mathrm{T}} & P^{-1} \end{bmatrix} \geq 0, \ \ k \in [1, m] \iff \mathcal{E}(P) \subset \mathcal{L}(H). \quad (26)$$

The inequalities in (25) and the condition (26) jointly show that $\mathcal{E}(P)$ is an invariant set by Proposition 3. Hence a trajectory starting from x_0 will stay inside of $\mathcal{E}(P)$, which is a subset of $\text{co}\{\mathcal{E}(P_j): \ j \in [1,N]\}$. Since x_0 is an arbitrary point inside $\text{co}\{\mathcal{E}(P_j): \ j \in [1,N]\}$, it follows that this convex hull is an invariant set. If "<" holds for all the inequalities in (20), then we also have "<" in (25), which guarantees that the trajectory starting form x_0 will converge to the origin. □

For single input systems, it was shown in [9] that the set invariance condition in Proposition 3 is also necessary. Hence we have

Corollary 1 *Suppose that (17) is a single input system. Given a group of ellipsoids $\mathcal{E}(P_j), j \in [1,N]$. If each ellipsoid $\mathcal{E}(P_j)$ is (contractively) invariant, then $\text{co}\{\mathcal{E}(P_j), j \in [1,N]\}$ is a (contractively) invariant set.*

Actually, Corollary 1 can be extended to a more general class of saturation-like functions.

Theorem 4 *Let $\phi(s): \mathbf{R} \mapsto \mathbf{R}$ be a continuous function such that $\phi(0) = 0$ and $|\phi(s)| \leq |s|$. Assume that for $s \geq 0$, $\phi(s) \geq 0$ is concave and for $s \leq 0$, $\phi(s) \leq 0$ is convex. Consider a single input system (16) with $u = \phi(Fx)$. The closed-loop system is*

$$\dot{x} = Ax + B\phi(Fx). \tag{27}$$

Given a group of ellipsoids $\mathcal{E}(P_j), j \in [1,N]$. Suppose that

$$(A+BF)^{\text{T}}P_j + P_j(A+BF) \leq 0 (< 0), \quad j \in [1,N]. \tag{28}$$

If each ellipsoid $\mathcal{E}(P_j)$ is (contractively) invariant for (27), then $\text{co}\{\mathcal{E}(P_j), j \in [1,N]\}$ is a (contractively) invariant set.

Note that when $\phi(s) = \text{sat}(s)$, the condition (28) is necessary for the (contractive) invariance of each $\mathcal{E}(P_j)$. Hence it is not an additional condition as compared with Corollary 1. Fig. 4 illustrates a few functions that satisfy the constraint for $\phi(s)$ in Theorem 4. The restriction $|\phi(s)| \leq |s|$ can be extended to $|\phi(s)| \leq \alpha|s|$ for any $\alpha > 0$. If we have $\alpha \neq 1$, then we can replace B with αB and $\phi(s)$ with $\phi(s)/\alpha$.

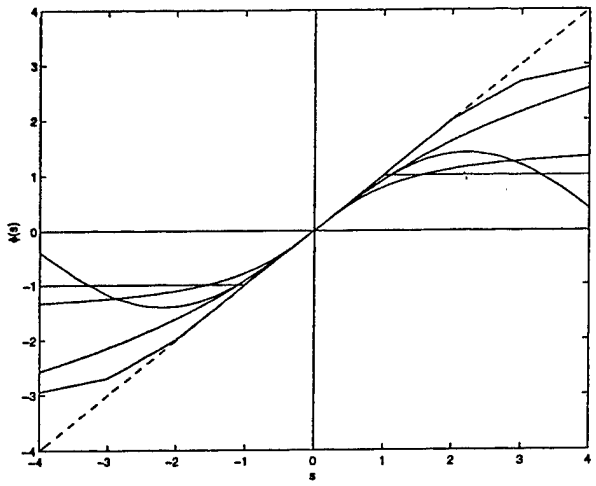


Figure 4: A class of saturation-like functions.

Proof of Theorem 4. Recall from Theorem 1 that $\mathrm{co}\{\mathcal{E}(P_j), j \in [1,N]\} = L_{V_c}(1)$. We prove the invariance of $L_{V_c}(1)$ instead. It suffices to show that

$$\dot{V}_c(x) = \left(\frac{\partial V_c}{\partial x}\right)^{\mathrm{T}} (Ax + B\phi(Fx)) \leq 0$$

for all $x \in \partial L_{V_c}(1)$. For contractive invariance, we can show that $\dot{V}_c(x) < 0$ for all $x \in \partial L_{V_c}(\rho), \rho \in (0,1]$, following a same procedure.

Now we consider an arbitrary $x_0 \in \partial L_{V_c}(1)$. For simplicity, assume that $Fx_0 \geq 0$ (the proof for $Fx_0 \leq 0$ is similar). If $x_0 \in \partial\mathcal{E}(P_j)$ for some $j \in [1,N]$, then $\frac{\partial V_c}{\partial x}\Big|_{x=x_0} = 2P_jx_0$ and $\dot{V}_c(x) = \dot{V}_j(x) \leq 0$ follows from the invariance of $\mathcal{E}(P_j)$. Hence we assume that $x_0 \notin \partial\mathcal{E}(P_j)$ for any j. Then there exist an integer $N_0 \leq N$, some numbers $\alpha_j \in (0,1)$ and vectors $x_j \in \mathcal{E}(P_j), j \in [1,N_0]$, such that

$$\sum_{j=1}^{N_0} \alpha_j = 1, \quad x_0 = \sum_{j=1}^{N_0} \alpha_j x_j.$$

(Here we have assumed for simplicity that x_0 is only related to the first N_0 ellipsoids. Otherwise, the ellipsoids can be reordered to meet this assumption.) Let $h_0 = \frac{1}{2}\left(\frac{\partial V_c}{\partial x}\Big|_{x=x_0}\right)^{\mathrm{T}}$, then by Theorem 2, $h_0x_0 = x_0^{\mathrm{T}}P(\gamma^*)x_0 = 1$. It follows that the hyperplane $h_0x = 1$ is tangential to the convex set $L_{V_c}(1)$ at $x = x_0$. Hence $L_{V_c}(1)$ lies between $h_0x = 1$ and $h_0x = -1$, i.e., $L_{V_c}(1) \subset \mathcal{L}(h_0)$. Therefore,

$$\mathcal{E}(P_j) \subset \mathcal{L}(h_0) \quad \forall j \in [1,N_0], \tag{29}$$

and

$$1 = h_0 x_0 \geq h_0 x_j \quad \forall j \in [1, N_0].$$

We claim that $h_0 x_j = 1$ for all $j \in [1, N_0]$. Suppose on the contrary that $h_0 x_j < 1$ for some j, say, $h_0 x_1 < 1$, then

$$1 = h_0 x_0 = \alpha_1 h_0 x_1 + \sum_{j=2}^{N_0} \alpha_j h_0 x_j \leq \alpha_1 h_0 x_1 + \sum_{j=2}^{N_0} \alpha_j < \sum_{j=1}^{N_0} \alpha_j = 1,$$

which is a contradiction. Because of (29) and $x_j \in \mathcal{E}(P_j)$, the equality $h_0 x_j = 1$ implies that $\mathcal{E}(P_j)$ touches the hyperplane $h_0 x = 1$ at $x = x_j$. Hence the hyperplane $h_0 x = 1$ is tangential to $\mathcal{E}(P_j)$ at x_j for every $j \in [1, N_0]$. It follows from Fact 1 that

$$h_0^{\mathrm{T}} = P_j x_j \quad \forall j \in [1, N_0].$$

Let $V_j(x) = x^{\mathrm{T}} P_j x$. Since each ellipsoid $\mathcal{E}(P_j)$ is invariant, we have

$$\dot{V}_j(x_j) = 2x_j^{\mathrm{T}} P_j (Ax_j + B\phi(Fx_j)) = 2h_0(Ax_j + B\phi(Fx_j)) \leq 0, \quad j \in [1, N_0]. \tag{30}$$

We need to show that

$$\dot{V}_c(x_0) = 2h_0(Ax_0 + B\phi(Fx_0)) \leq 0. \tag{31}$$

Because of (28), we have

$$2x_j^{\mathrm{T}} P_j (Ax_j + BFx_j) = 2h_0(Ax_j + BFx_j) \leq 0 \quad \forall j \in [1, N_0].$$

Hence, for all $x \in \mathrm{co}\{x_1, x_2, \cdots, x_{N_0}\}$,

$$2h_0(Ax + BFx) \leq 0. \tag{32}$$

We first assume that $Fx_j \geq 0$ for all $j \in [1, N_0]$. In this case, $Fx \geq 0$ for all $x \in \mathrm{co}\{x_1, x_2, \cdots, x_{N_0}\}$. If $h_0 B \geq 0$, then

$$\dot{V}_c(x_0) = 2h_0(Ax_0 + B\phi(Fx_0)) \leq 2h_0(Ax_0 + BFx_0) \leq 0,$$

noticing that, by the assumption that $Fx_0 \geq 0$ and by the property of the function $\phi(s)$, we have $\phi(Fx_0) \leq Fx_0$. If $h_0 B \leq 0$, then by the property of $\phi(s)$, $h_0 Ax + h_0 B\phi(Fx)$ is a convex function for $x \in \mathrm{co}\{x_1, x_2, \cdots, x_{N_0}\}$. Hence we also have (31) by (30).

If $Fx_j \geq 0$ does not hold for all $j \in [1, N_0]$, then we can get an intersection of the set $\text{co}\{x_1, x_2, \cdots, x_{N_0}\}$ with the half space $Fx \geq 0$. This intersection is also a polygon and can be denoted as $\text{co}\{y_1, y_2, \cdots, y_{N_1}\}$. Since $Fx_0 \geq 0$, we have $x_0 \in \text{co}\{y_1, y_2, \cdots, y_{N_1}\}$. Some y_j's belong to $\{x_1, x_2, \cdots, x_{N_0}\}$, others are not. For those $y_j \notin \{x_i : i \in [1, N_0]\}$, we must have $Fy_j = 0$ and $y_j \in \text{co}\{x_i : i \in [1, N_0]\}$. It follows from (32) that $h_0(Ay_j + BFy_j) \leq 0$. Since $\phi(0) = 0$, for those y_j's such that $Fy_j = 0$, we have

$$\dot{V}_c(y_j) = 2h_0(Ay_j + B\phi(Fy_j)) = 2h_0(Ay_j + BFy_j) \leq 0.$$

In summary, we have

$$\dot{V}_c(y_j) = 2h_0(Ay_j + B\phi(Fy_j)) \leq 0 \quad \forall j \in [1, N_1]. \tag{33}$$

Because of this, we can work on the set $\text{co}\{y_1, y_2, \cdots, y_{N_1}\}$ instead of $\text{co}\{x_1, x_2, \cdots, x_{N_0}\}$. Since $Fy_j \geq 0$ for all $j \in [1, N_1]$, same arguments can be used to prove (31) by using (33) instead of (30). □

4 Controlled invariant sets

In this section, we investigate the possibility that a level set can be made invariant with controls delivered by the saturating actuators. Given a positive definite function $V(x)$, suppose that the level set $L_V(1)$ is bounded and $V(kx) = k^2V(x)$. A level set $L_V(\rho)$ is said to be controlled contractively invariant if for every $x \in L_V(\rho) \setminus \{0\}$, there exists a $u \in \mathbf{R}^m$, $|u|_\infty \leq 1$, such that

$$\dot{V}(x, u) = \left(\frac{\partial V}{\partial x}\right)^{\mathrm{T}} (Ax + Bu) < 0.$$

The controlled invariance can be defined by replacing "$<$" with "$\leq$". Since $V(kx) = k^2V(x)$, we have $\frac{\partial V}{\partial x}\Big|_{x=kx_0} = k\,\frac{\partial V}{\partial x}\Big|_{x=x_0}$. Hence if $L_V(\rho)$ is controlled (contractively) invariant, then $L_V(\rho_1)$ is for all $\rho_1 \leq \rho$. Therefore, to determine the controlled (contractive) invariance of $L_V(\rho)$, it suffices to check all the points in $\partial L_V(\rho)$. For the composite quadratic Lyapunov function $V_c(x)$ defined in (1), we have

Theorem 5 *Suppose that each of the ellipsoids $\mathcal{E}(P_j), j \in [1, N]$, is controlled (contractively) invariant, then $L_{V_c}(1)$ is controlled (contractively) invariant.*

Proof. We only prove controlled invariance. The controlled contractively invariance can be shown similarly.

Denote $V_j(x) = x^{\mathrm{T}} P_j x$. The condition implies that for all $x \in \partial\mathcal{E}(P_j)$, there exists a $u \in \mathbf{R}^m$, $|u|_\infty \le 1$, such that

$$\dot{V}_j(x,u) = 2x^{\mathrm{T}} P_j (Ax + Bu) \le 0. \tag{34}$$

Now we consider an arbitrary $x_0 \in \partial L_{V_c}(1)$. If $x_0 \in \partial\mathcal{E}(P_j)$ for some $j \in [1,N]$, then $\frac{\partial V_c}{\partial x}\Big|_{x=x_0} = 2P_j x_0$ and $\dot{V}_c(x,u) = \dot{V}_j(x,u) \le 0$ follows from (34). Hence we assume that $x_0 \notin \partial\mathcal{E}(P_j)$ for any j. Then, similar to the proof of Theorem 4, there exists an integer $N_0 \le N$, some numbers $\alpha_j \in (0,1)$ and vectors $x_j \in \mathcal{E}(P_j), j \in [1,N_0]$, such that

$$\sum_{j=1}^{N_0} \alpha_j = 1, \quad x_0 = \sum_{j=1}^{N_0} \alpha_j x_j.$$

Letting $h_0 = \frac{1}{2}\left(\frac{\partial V_c}{\partial x}\Big|_{x=x_0}\right)^{\mathrm{T}}$, we have

$$h_0^{\mathrm{T}} = P_j x_j \quad \forall\, j \in [1,N_0],$$

and the hyperplane $h_0 x = 1$ is tangential to each $\mathcal{E}(P_j)$ at x_j. By assumption, there exists a $u_j \in \mathbf{R}^m, |u_j|_\infty \le 1$, such that

$$\dot{V}_j(x_j,u_j) = 2x_j^{\mathrm{T}} P_j (Ax_j + Bu_j) \le 0,$$

i.e., $2h_0(Ax_j + Bu_j) \le 0$, for all $j \in [1,N_0]$. Let $u_0 = \sum_{j=1}^{N_0} \alpha_j u_j$. Then $|u_0|_\infty \le 1$ and by the convexity, we have

$$\dot{V}_c(x_0,u_0) = 2h_0(Ax_0 + Bu_0) \le 0.$$

Since x_0 is an arbitrary point in $\partial L_{V_c}(1)$, this implies that the level set $L_{V_c}(1)$ is controlled invariant. □

If a level set $L_{V_c}(1)$ is controlled contractively invariant, a simple feedback law to make it contractively invariant is

$$u_i = -\mathrm{sign}\left(b_i^{\mathrm{T}} \frac{\partial V_c}{\partial x}\right), \quad i \in [1,m], \tag{35}$$

where b_i is the ith column of B. However, due to the discontinuity of the sign function, the closed-loop system under this control may be not well behaved. For instance, the closed-loop differential equation may have no solution. It can be shown with methods in [8], Chapter 11, that

there exists a positive number k such that $L_{V_c}(1)$ is contractively invariant under the saturated linear feedback law

$$u = -\text{sat}\left(kB^{\mathrm{T}}\frac{\partial V_c}{\partial x}\right).$$

This control is continuous in x since both $\text{sat}(\cdot)$ and $\frac{\partial V_c}{\partial x}$ are continuous. The value of k may be however difficult to determine. In the next section, we will provide a method for constructing a controller from a group of saturated linear feedback laws.

5 Construction of continuous feedback laws

Suppose that we have a group of ellipsoids $\mathcal{E}(P_j), j \in [1, N]$, each of them (contractively) invariant under a corresponding saturated linear feedback $u = \text{sat}(F_j x)$. It was shown in [13, 8] that a switching feedback law can be constructed such that the union $\cup_{j=1}^{N}\mathcal{E}(P_j)$ is invariant. In this section, we would like to construct a continuous feedback law from these F_j's such that the convex hull of the ellipsoids, $\text{co}\{\mathcal{E}(P_j) : j \in [1, N]\} = L_{V_c}(1)$, is invariant.

Theorem 6 *Given ellipsoids $\mathcal{E}(P_j)$ and feedback matrices $F_j \in \mathbf{R}^{m\times n}$, $j \in [1, N]$. Suppose that there exist $H_j \in \mathbf{R}^{m\times n}$ such that $\mathcal{E}(P_j) \subset \mathcal{L}(H_j)$ and*

$$(A + B(D_i F_j + D_i^- H_j))^{\mathrm{T}} P_j + P_j(A + B(D_i F_j + D_i^- H_j)) \leq 0 (< 0) \tag{36}$$

for all $i \in [1, 2^m]$ and $j \in [1, N]$. Let $Q_j = P_j^{-1}$, $Y_j = F_j Q_j$. Let $\gamma^*(x)$ be such that $x^{\mathrm{T}} P(\gamma^*) x = V_c(x)$. Suppose that the vector function $\gamma^*(x)$ is continuous in x. Define $F(\gamma) := Y(\gamma)Q^{-1}(\gamma)$, where

$$Y(\gamma) = \sum_{j=1}^{N} \gamma_j Y_j, \quad Q(\gamma) = \sum_{j=1}^{N} \gamma_j Q_j.$$

Then $u = \text{sat}(F(\gamma^*(x))x)$ is a continuous feedback law and $L_{V_c}(1)$ is (contractively) invariant under the feedback $u = \text{sat}(F(\gamma^*(x))x)$.

Proof. We only prove the invariance of $L_{V_c}(1)$. The contractive invariance follows from similar arguments. Let $Z_j = H_j Q_j$. Denote $Z(\gamma) = \sum_{j=1}^{N} \gamma_j Z_j$ and $H(\gamma) = Z(\gamma)Q^{-1}(\gamma)$. We see that (36) can be rewritten as

$$Q_j A^{\mathrm{T}} + A Q_j + (D_i Y_j + D_i^- Z_j))^{\mathrm{T}} B^{\mathrm{T}} + B(D_i Y_j + D_i^- Z_j) \leq 0$$

for all $i \in [1, 2^m]$ and $j \in [1, N]$. It follows from the convexity that $\mathcal{E}(P(\gamma)) \subset \mathcal{L}(H(\gamma))$ (see the proof of Theorem 3, equation (26), where the dependence on γ is suppressed) and

$$Q(\gamma)A^{\mathrm{T}} + Q(\gamma)A^{\mathrm{T}} + (D_iY(\gamma) + D_i^-Z(\gamma)))^{\mathrm{T}}B^{\mathrm{T}} + B(D_iY(\gamma) + D_i^-Z(\gamma))) \leq 0$$

for all $i \in [1, 2^m]$ and $\gamma \in \Gamma$. The above inequality is equivalent to

$$(A + B(D_iF(\gamma) + D_i^-H(\gamma)))^{\mathrm{T}}P(\gamma) + P(\gamma)(A + B(D_iF(\gamma) + D_i^-H(\gamma))) \leq 0 \quad \forall i \in [1, 2^m].$$

By Proposition 3, this inequality along with $\mathcal{E}(P(\gamma)) \in \mathcal{L}(H(\gamma))$ shows that $\mathcal{E}(P(\gamma))$ is invariant under the control of $u = \mathrm{sat}(F(\gamma)x)$, i.e.,

$$2x^{\mathrm{T}}P(\gamma)(Ax + B\mathrm{sat}(F(\gamma)x)) \leq 0 \quad \forall x \in \mathcal{E}(P(\gamma)). \tag{37}$$

For an arbitrary $x_0 \in \partial L_{V_c}(1)$, let $\gamma_0 = \gamma^*(x_0)$. Then $x_0^{\mathrm{T}}P(\gamma_0)x_0 = 1$, i.e., $x_0 \in \partial\mathcal{E}(P(\gamma_0))$, and from Theorem 2,

$$\left.\frac{\partial V_c}{\partial x}\right|_{x=x_0} = 2P(\gamma_0)x_0,$$

It follows from (37) that

$$\dot{V}_c(x_0) = 2x_0^{\mathrm{T}}P(\gamma_0)(Ax_0 + B\mathrm{sat}(F(\gamma_0)x_0))) \leq 0.$$

This shows that $L_{V_c}(1)$ is invariant under the control $u = \mathrm{sat}(F(\gamma^*(x))x)$. □

For the case where $N = 2$ and $Q_1 - Q_2$ is nonsingular, $\gamma^*(x)$ is continuous by Proposition 1 and can be computed with Proposition 2.

Example. Consider system (16) with

$$A = \begin{bmatrix} 0 & -0.5 \\ 1 & 1.5 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

We have designed two feedback matrices

$$F_1 = \begin{bmatrix} 0.9471 & 1.6000 \end{bmatrix}, \quad F_2 = \begin{bmatrix} -0.1600 & 1.6000 \end{bmatrix},$$

along with two ellipsoids, $\mathcal{E}(P_1)$ and $\mathcal{E}(P_2)$, where

$$P_1 = \begin{bmatrix} 1.6245 & -1.5364 \\ -1.5364 & 15.3639 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 5.9393 & -0.2561 \\ -0.2561 & 2.5601 \end{bmatrix}.$$

The matrices P_1 and F_1 are designed such that the value α_1 is maximized, where $\alpha_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \mathcal{E}(P_1)$ and $V_1(x) = x^{\mathrm{T}} P_1 x$ has a guaranteed convergence rate inside $\mathcal{E}(P_1)$ under $u = \mathrm{sat}(F_1 x)$. The matrices P_2 and F_2 are designed such that the value α_2 is maximized, where $\alpha_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} \in \mathcal{E}(P_2)$ and $V_2(x) = x^{\mathrm{T}} P_2 x$ has a guaranteed convergence rate inside $\mathcal{E}(P_2)$ under $u = \mathrm{sat}(F_2 x)$ (see [10, 8] for the detailed design method). In Fig. 5, the boundaries of the two ellipsoids are plotted

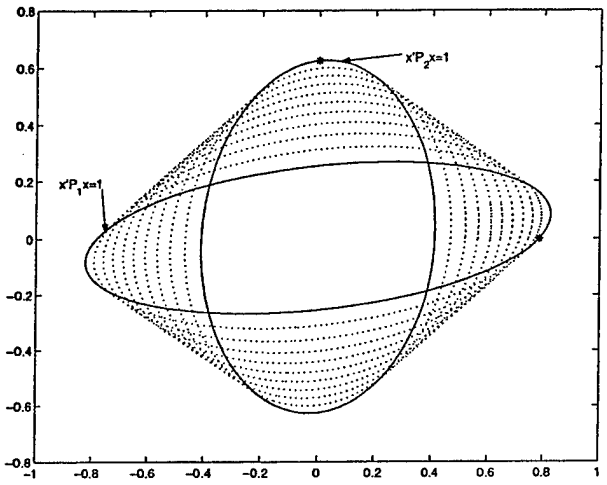


Figure 5: The convex hull of two ellipsoids.

in solid curves. The dotted curves are the boundaries of $\mathcal{E}(P(\gamma))$ as γ varies in the set Γ. The shape of $L_{V_c}(1) = \bigcup_{\gamma \in \Gamma} \mathcal{E}(P(\gamma))$ can be seen from these dotted curves. It can be verified that $Q_1 - Q_2 = P_1^{-1} - P_2^{-1}$ is nonsingular. So we can use the method in Proposition 2 to compute the function $\gamma^*(x)$, which is guaranteed to be continuous by Proposition 1. Simulation is carried out under the feedback law $u = \mathrm{sat}(F(\gamma^*(x))x)$. In Fig. 6, a trajectory starting from $\partial L_{V_c}(1)$ is plotted. Fig. 7 plots the control signal $u(t)$ (in solid curve) and the composite quadratic function $V_c(x(t))$ (in dashed curve).

6 Conclusions

We introduced a composite quadratic function for analysis and design of linear systems with input and state constraint. We have shown that this function is continuously differentiable and its level set is the convex hull of a group of ellipsoids. Using these results, we studied some set invariance properties of linear systems with input saturation or saturation-like nonlinearities. In particular, we showed that if every ellipsoid in a group is invariant under a saturated (or

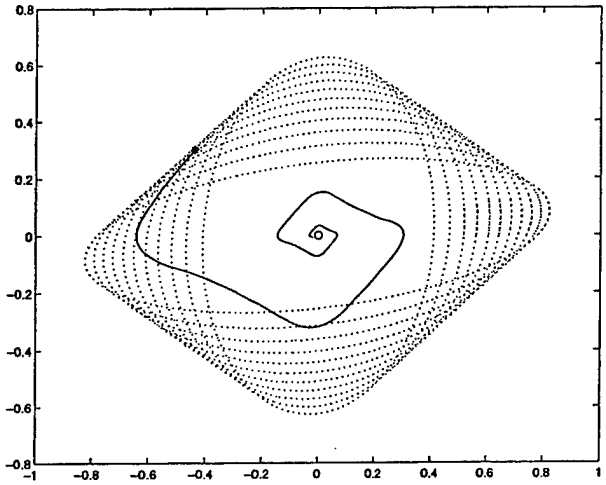


Figure 6: A trajectory and the invariant set $L_{V_c}(1)$.

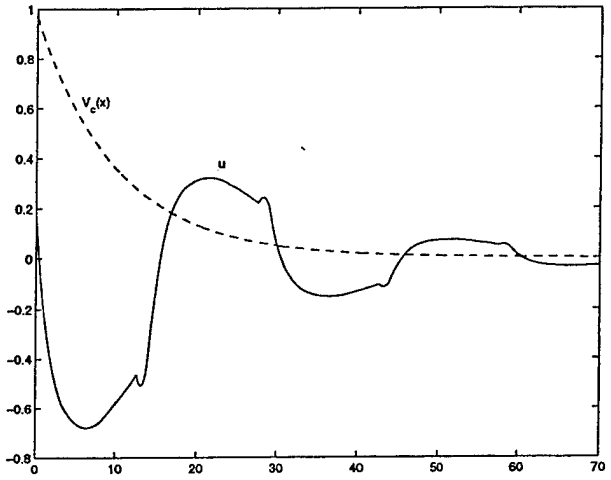


Figure 7: The control signal $u(t)$ and the Lyapunov function $V_c(x(t))$.

saturated-like) feedback, then they convex hull is also invariant. Similar results on controlled invariance have also been established. We also proposed a method to construct a continuous feedback law based on a group of saturated linear feedback laws to make the convex hull of a group of ellipsoids invariant. The composite quadratic function is relatively easier to handle than a general nonlinear Lyapunov function and we expect to use it to study more general nonlinear systems.

A Proof of Proposition 2

In the proof, we will use the following algebraic fact. Suppose that X_1, X_4 and $\begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \end{bmatrix}$ are square matrices. If X_1 is nonsingular, then

$$\det \begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \end{bmatrix} = \det(X_1) \det(X_4 - X_3 X_1^{-1} X_2), \tag{38}$$

and if X_4 is nonsingular, then

$$\det \begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \end{bmatrix} = \det(X_4) \det(X_1 - X_2 X_4^{-1} X_3). \tag{39}$$

Let $Q(\lambda) \in \mathbf{R}^{n \times n}$ be a matrix function. Suppose that $Q(\lambda)$ is nonsingular, then

$$\frac{dQ^{-1}(\lambda)}{d\lambda} = -Q^{-1}(\lambda) \frac{dQ(\lambda)}{d\lambda} Q^{-1}(\lambda). \tag{40}$$

In what follows, we use I to denote the $n \times n$ unit matrix and 0 to denote a zero matrix of appropriate dimension. For simplicity, we use $Q(\lambda)$ to denote $\lambda Q_1 + (1-\lambda)Q_2$.

If $\alpha(\lambda^*, x) = V_c(x)$ and $\lambda^* \in (0,1)$, then we must have $\frac{\partial \alpha}{\partial \lambda}\Big|_{\lambda=\lambda^*} = 0$. From (40), we have

$$x^{\mathrm{T}} Q^{-1}(\lambda^*)(Q_1 - Q_2) Q^{-1}(\lambda^*) x = 0,$$

which can be written as

$$\det\left(1 - x^{\mathrm{T}} Q^{-1}(\lambda^*)(Q_1 - Q_2) Q^{-1}(\lambda^*) x\right) = 1.$$

By applying (38) and (39), we obtain a sequence of equivalent relations:

$$\det \begin{bmatrix} 1 & x^{\mathrm{T}} Q^{-1}(\lambda^*) \\ Q^{-1}(\lambda^*) x & (Q_1 - Q_2)^{-1} \end{bmatrix} = \det(Q_1 - Q_2)^{-1}$$

$$\Updownarrow$$

$$\det\left(\begin{bmatrix} 1 & 0 \\ 0 & (Q_1 - Q_2)^{-1} \end{bmatrix} - \begin{bmatrix} x^{\mathrm{T}} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} Q^{-1}(\lambda^*) & 0 \\ 0 & Q^{-1}(\lambda^*) \end{bmatrix} \begin{bmatrix} 0 & I \\ x & 0 \end{bmatrix}\right) = \det(Q_1 - Q_2)^{-1}$$

$$\Updownarrow$$

$$\det\left(\begin{bmatrix} Q(\lambda^*) & 0 \\ 0 & Q(\lambda^*) \end{bmatrix} - \begin{bmatrix} 0 & I \\ x & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (Q_1 - Q_2) \end{bmatrix} \begin{bmatrix} x^{\mathrm{T}} & 0 \\ 0 & I \end{bmatrix}\right) = \det \begin{bmatrix} Q(\lambda^*) & 0 \\ 0 & Q(\lambda^*) \end{bmatrix}$$

$$\Updownarrow$$

$$\det \begin{bmatrix} Q(\lambda^*) & Q_1 - Q_2 \\ x x^{\mathrm{T}} & Q(\lambda^*) \end{bmatrix} = \det \begin{bmatrix} Q(\lambda^*) & 0 \\ 0 & Q(\lambda^*) \end{bmatrix}$$

Multiplying the matrices on both sides from left with $\begin{bmatrix} U^{\mathrm{T}} & 0 \\ 0 & U^{\mathrm{T}} \end{bmatrix}$ and from right with $\begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}$, we obtain

$$\det \begin{bmatrix} \hat{Q}_2 + \lambda^*(\hat{Q}_1 - \hat{Q}_2) & \hat{Q}_1 - \hat{Q}_2 \\ \begin{bmatrix} c & 0 \\ 0 & 0_{n-1} \end{bmatrix} & \hat{Q}_2 + \lambda^*(\hat{Q}_1 - \hat{Q}_2) \end{bmatrix} = \det \begin{bmatrix} \hat{Q}_2 + \lambda^*(\hat{Q}_1 - \hat{Q}_2) & \hat{Q}_1 - \hat{Q}_2 \\ 0 & \hat{Q}_2 + \lambda^*(\hat{Q}_1 - \hat{Q}_2) \end{bmatrix} \tag{41}$$

Here we notice that $\hat{Q}_1 - \hat{Q}_2$ has been added to the upper-right block of the right-hand side matrix. This does not change the value of the determinant.

The two determinants in (41) are different only at the first column. By using the property of determinant on column addition and the partition of $\hat{Q}_1$ and $\hat{Q}_2$, we have

$$\det \begin{bmatrix} 0 & \lambda^*(\hat{Q}_{12} - \hat{Q}_{22}) + \hat{Q}_{22} & \hat{Q}_1 - \hat{Q}_2 \\ c & 0 & \lambda^*(\hat{q}_1 - \hat{q}_2)^{\mathrm{T}} + \hat{q}_2^{\mathrm{T}} \\ 0 & 0 & \lambda^*(\hat{Q}_{12} - \hat{Q}_{22})^{\mathrm{T}} + \hat{Q}_{22}^{\mathrm{T}} \end{bmatrix} = 0$$

$$\Updownarrow$$

$$\det \begin{bmatrix} \lambda^*(\hat{Q}_{12} - \hat{Q}_{22}) + \hat{Q}_{22} & \hat{Q}_1 - \hat{Q}_2 \\ 0 & \lambda^*(\hat{Q}_{12} - \hat{Q}_{22})^{\mathrm{T}} + \hat{Q}_{22}^{\mathrm{T}} \end{bmatrix} = 0.$$

The last equality is (15). □

References

[1] D. S. Bernstein and A. N. Michel, editors, Special issue on "Saturating Actuators," *Int. J. Robust and Nonlinear Control*, Vol. 5, 1995.

[2] F. Blanchini, "Set invariance in control – a survey", *Automatica*, Vol. 35, No.11, pp. 1747-1767, 1999.

[3] S. Boyd, L. El Ghaoui, E. Feron and V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory*, SIAM Studies in Appl. Mathematics, Philadelphia, 1994.

[4] E. J. Davison and E. M. Kurak, "A computational method for determining quadratic Lyapunov functions for non-linear systems," *Automatica*, Vol. 7, pp. 627-636, 1971.

[5] E. G. Gilbert and K. T. Tan, "Linear systems with state and control constraints: the theory and application of maximal output admissible sets," *IEEE Trans. Automat. Contr.* Vol. 36, pp. 1008-1020, 1991.

[6] P. O. Gutman, and M.Cwikel, "Admissible sets and feedback control for discrete-time linear dynamical systems with bounded control and dynamics", *IEEE Trans. Auto. Contr.*, Vol. 31, pp. 373-376, 1986.

[7] H. Hindi & S. Boyd, "Analysis of linear systems with saturating using convex optimization," *Proc of the 37th IEEE CDC*, pp903-908, Florida, 1998.

[8] T. Hu and Z. Lin. *Control Systems with Actuator Saturation: Analysis and Design*, Birkhäuser, Boston, 2001.

[9] T. Hu and Z. Lin, "Exact characterization of invariant ellipsoids for linear systems with saturating actuators," *IEEE Trans. Automat. Contr.*, scheduled to appear in February 2002.

[10] T. Hu and Z. Lin, "On enlarging the basin of attraction for linear systems under saturated linear feedback," *Sys. & Contr. Lett.*, Vol. 40, No. 1, pp. 59-69, May, 2000.

[11] T. Hu, Z. Lin and B. M. Chen, "An analysis and design method for linear systems subject to actuator saturation and disturbance," *Automatica*, Vol. 38, No. 2, pp. 351-359, 2002.

[12] T. Hu, Z. Lin and L. Qiu, "Stabilization of exponentially unstable linear systems with saturating actuators," *IEEE Transactions on Automatic Control*, Vol. 46, No. 6, pp. 973-979, 2001.

[13] T. Hu, Z. Lin and Y. Shamash, "Semi-global stabilization with guaranteed regional performance of linear systems subject to actuator saturation," *Systems & Control Letters*, Vol. 43, No. 3, pp. 203-210, 2001.

[14] H. Khalil, *Nonlinear Systems*, Prentice-Hall, Upper Saddle River, New Jersey, 1996.

[15] Z. Lin, *Low Gain Feedback*, Lecture Notes in Control and Information Sciences, Vol. 240, Springer-Verlag, London, 1998.

[16] Z. Lin and A. Saberi, "Semi-global exponential stabilization of linear systems subject to 'input saturation' via linear feedbacks," *Systems and Control Letters*, Vol. 21, pp. 225-239, 1993.

[17] K. A. Loparo and G. L. Blankenship, "Estimating the domain of attraction of of nonlinear feedback systems," *IEEE Trans. Automat. Contr.*, Vol. 23, No.4, pp. 602-607, 1978.

[18] D. Liu and A. N. Michel, *Dynamical Systems with Saturation Nonlinearities*, Lecture Notes in Control and Information Sciences, Vol. 195, Springer-Verlag, London, 1994.

[19] H. J. Sussmann, E. D. Sontag and Y. Yang, "A general result on the stabilization of linear systems using bounded controls", *IEEE Trans. Automat. Contr.* Vol. 39, pp. 2411-2425, 1994.

[20] H. J. Sussmann and Y. Yang, "On the stabilizability of multiple integrators by means of bounded feedback controls", *Proc. 30th IEEE Conf. Decision and Control*, pp. 70-72, 1991.

[21] R. Suarez, J. Alvarez-Ramirez and J. Solis-Daun, "Linear systems with bounded inputs: global stabilization with eigenvalue placement," *Int. J. Robust Nonlin. Contr.*, Vol. 7, pp. 835-845, 1997.

[22] A. R. Teel, "Global stabilization and restricted tracking for multiple integrators with bounded controls", *System and Control Letters*, Vol. 18, pp. 165-171, 1992.

[23] A. R. Teel, "Linear systems with input nonlinearities: global stabilization by scheduling a family of H_∞-type controllers," *Int. J. of Robust and Nonlinear Control*, Vol. 5, pp. 399-441, 1995.

[24] S. Weissenberger, "Application of results from the absolute stability to the computation of finite stability domains," *IEEE Trans. Automatic Control*, AC-13, pp124-125, 1968.

[25] G. F. Wredenhagen and P. R. Belanger, "Piecewise-linear LQ control for systems with input constraints," *Automatica*, Vol. 30, pp. 403-416, 1994.

Publication 24

Submitted to *IEEE Transactions on Automatic Control*

The Equivalence of Several Set Invariance Conditions under Saturation

Tingshu Hu† Zongli Lin†

† Department of Electrical and Computer Engineering
University of Virginia
Charlottesville, VA 22903, U.S.A.
Email: th7f, zl5y@virginia.edu

Abstract

Several equivalent conditions or statements for set invariance were obtained for systems with one saturating actuator in a recent paper. In particular, it was shown that the invariance of an ellipsoid under a saturated linear feedback is equivalent to its controlled invariance and also to the existence of a feedback linear inside the ellipsoid that makes it invariant. In this paper, we attempt to extend the results to systems with multiple saturating actuators. Our analysis reveals that the equivalence holds conditionally for some pairs of the statements and does not hold for some other pairs.

Keywords: invariant ellipsoid, contractive invariance, controlled invariance, actuator saturation

[1]This work was supported in part by the US Office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

1 Introduction

The notion of invariant set has played a very important role in the study of systems with state and control constraint (see [1]-[13], [15]-[17] and the references therein). Invariant sets can be used to estimate the domain of attraction and to study disturbance rejection properties. We have been particularly interested in invariant ellipsoids since quadratic Lyapunov functions are the most popular and the results can often be put into simple and compact forms, which makes analysis and design easily implementable. Recently, we obtained a series of encouraging results on invariant ellipsoids (see, e.g., [7, 8, 10, 11]). In [10] and [11], we derived simple conditions for set invariance in the form of LMIs, for both continuous-time and discrete-time linear systems under actuator saturation. These conditions are then used for designing controllers to enlarge the domain of attraction and to improve the performance of disturbance rejection.

In a recent paper [7], we gave a complete characterization of invariant ellipsoids for single input systems with actuator saturation. Four equivalent conditions or statements on set invariance are presented. In particular, it was shown that if an ellipsoid is controlled invariant (i.e., it can be made invariant with a nonlinear feedback bounded by the saturation level), then there exists a feedback which is linear inside the ellipsoid that makes it invariant. Also, suppose that an ellipsoid is invariant under a linear feedback $u = Fx$, then it is invariant under the saturated linear feedback $u = \text{sat}(Fx)$ if and only if the ellipsoid is controlled invariant. Since one of the equivalent conditions can be put in the form of LMI, these properties make it very easy to analyze and construct invariant ellipsoids.

In this paper, we attempt to extend the results in [7] to multi-input systems. Our investigation shows that some pairs of the statements are conditionally equivalent and some other pairs are not equivalent.

Notation: In this paper, we use $\text{sat} : \mathbf{R}^m \to \mathbf{R}^m$ to denote the standard saturation function of appropriate dimensions. For $u \in \mathbf{R}^m$, the ith component of $\text{sat}(u)$ is $\text{sign}(u_i)\min\{1, |u_i|\}$. The infinity norm of u is denoted as $|u|_\infty$. For an $m \times n$ matrix H, we use h_i to denote its ith row and for an $n \times m$ matrix B, we use b_i to denote its ith column.

2 Problem statement and preliminaries

Consider a linear system with input saturation

$$\dot{x} = Ax + Bu, \quad x \in \mathbf{R}^n, \quad u \in \mathbf{R}^m, \quad |u|_\infty \le 1. \tag{1}$$

Let $P \in \mathbf{R}^{n\times n}$ be a positive-definite matrix. For a positive number ρ, denote

$$\mathcal{E}(P,\rho) = \left\{x \in \mathbf{R}^n : x^{\mathrm{T}} P x \le \rho\right\}.$$

We are interested in the controlled invariance of an ellipsoid $\mathcal{E}(P,\rho)$ and its invariance under a given feedback $u = f(x)$, especially, $u = \text{sat}(Fx)$. Let us first give formal definitions of these notions.

Denote $V(x) = x^{\mathrm{T}}Px$. Under the control u, the derivative of V along the trajectory of the system (1) is

$$\dot{V}(x,u) = 2x^{\mathrm{T}}P(Ax + Bu).$$

The ellipsoid $\mathcal{E}(P,\rho)$ is invariant under a given feedback control $u = f(x)$ if

$$\dot{V}(x,f(x)) = 2x^{\mathrm{T}}P(Ax + Bf(x)) \leq 0 \quad \forall x \in \mathcal{E}(P,\rho).$$

It is contractively invariant if

$$\dot{V}(x,f(x)) = 2x^{\mathrm{T}}P(Ax + Bf(x)) < 0 \quad \forall x \in \mathcal{E}(P,\rho) \setminus \{0\}.$$

The ellipsoid $\mathcal{E}(P,\rho)$ is controlled (contractively) invariant if there exists a vector function $f(x)$: $\mathbf{R}^n \mapsto \mathbf{R}^m$, $|f(x)|_\infty \leq 1$, that makes it (contractively) invariant.

As said before, we are particularly interested in the invariance of an ellipsoid under a saturated linear feedback $u = \text{sat}(Fx)$. This problem has been studied in [4]-[12] and [17]. For a matrix $F \in \mathbf{R}^{m\times n}$, define

$$\mathcal{L}(F) := \left\{x \in \mathbf{R}^n : \ |Fx|_\infty \leq 1\right\}.$$

If F is a feedback gain matrix, then $\mathcal{L}(F)$ is the region where the feedback control $u = \text{sat}(Fx)$ is linear in x. We call $\mathcal{L}(F)$ the linear region of the saturated feedback $\text{sat}(Fx)$, or simply, the linear region of saturation.

For single input systems ($m = 1$), we obtained four equivalent conditions or statements:

Proposition 1 *[7] Given a $P > 0$, assume that $\mathcal{E}(P,\rho)$ is controlled contractively invariant for some $\rho > 0$. Let $\rho > 0$ be given. The following statements are equivalent:*

a) $\mathcal{E}(P,\rho)$ is controlled contractively invariant;

b) $\mathcal{E}(P,\rho)$ is contractively invariant under the control $u = -\text{sign}(B^{\mathrm{T}}Px)$;

c) $\mathcal{E}(P,\rho)$ is contractively invariant under $u = \text{sat}(Fx)$, where F satisfies

$$(A + BF)^{\mathrm{T}}P + P(A + BF) < 0;$$

d) There exists an $H \in \mathbf{R}^{1\times n}$ satisfying

$$(A + BH)^{\mathrm{T}}P + P(A + BH) < 0$$

and $\mathcal{E}(P,\rho) \subset \mathcal{L}(H)$.

The equivalence of a) and c) implies the equivalence of controlled invariance and the invariance under a saturated linear feedback. Item d) implies that there exists a feedback which is linear inside $\mathcal{E}(P,\rho)$ that makes it contractively invariant.

For a multiple input system ($m > 1$), the four statements in Proposition 1 are respectively

a1) $\mathcal{E}(P,\rho)$ is controlled contractively invariant;

b1) $\mathcal{E}(P,\rho)$ is contractively invariant under the control $u_i = -\text{sign}(b_i^{\mathrm{T}}Px), i \in [1,m]$;

c1) $\mathcal{E}(P,\rho)$ is contractively invariant under $u = \text{sat}(Fx)$, where F satisfies

$$(A+BF)^{\mathrm{T}}P + P(A+BF) < 0;$$

d1) There exists an $H \in \mathbf{R}^{m\times n}$ satisfying

$$(A+BH)^{\mathrm{T}}P + P(A+BH) < 0$$

and $\mathcal{E}(P,\rho) \subset \mathcal{L}(H)$.

In this paper, we will investigate the equivalence of a1)–d1). It is clear that a1) is the weakest, i.e., it is a necessary condition for the other three. From Chapter 11 of [6], we know that a1) and b1) are always equivalent. Hence we need to study the pairs (b1,c1), (b1,d1) and (c1,d1). We will establish that the equivalence of b1) and d1) is conditional and the condition under which the equivalency holds will be identified. Numerical examples show that this condition is usually satisfied. In particular, d1) could be stronger than b1) in some special situations. However, the pairs (c1,d1) and (b1,c1) are not equivalent.

3 Condition for the equivalence of b1) and d1)

In [6], it was shown that $\mathcal{E}(P,\rho)$ is controlled contractively invariant for some $\rho > 0$ if and only if there exists an F such that

$$(A+BF)^{\mathrm{T}}P + P(A+BF) < 0. \tag{2}$$

Since controlled invariance is the weakest among the four conditions, we will assume the existence of F satisfying (2) throughout this paper.

We have also shown in [6] that if $\mathcal{E}(P,\rho_0)$ satisfies b1) (or d1)), then $\mathcal{E}(P,\rho)$ satisfies b1) (or d1)) for all $\rho < \rho_0$. In view of this, we only need to compare the largest ellipsoid satisfying b1) and the largest ellipsoid satisfying d1). Let ρ_b^* be the supreme of ρ such that b1) is satisfied and let ρ_d^* be the supreme of ρ such that d1) is satisfied. Then we have

$$\rho_b^* = \sup\left\{\rho > 0 : \ x^{\mathrm{T}}P\left(Ax - \sum_{i=1}^{m} b_i \text{sign}(b_i^{\mathrm{T}}Px)\right) < 0 \ \ \forall x \in \mathcal{E}(P,\rho)\setminus\{0\}\right\} \tag{3}$$

and

$$\begin{aligned} \rho_d^* = \quad & \sup_H \quad \rho \\ & \text{s.t} \quad (A+BH)^{\mathrm{T}}P + P(A+BH) < 0, \\ & \qquad \rho h_i P^{-1} h_i^{\mathrm{T}} \le 1, \ \ i \in [1,m]. \end{aligned} \tag{4}$$

Here we note that, the condition $\rho h_i P^{-1} h_i^{\mathrm{T}} \le 1, \in [1,m]$, is equivalent to the set inclusion condition $\mathcal{E}(P,\rho) \subset \mathcal{L}(H)$.

Lemma 1 *Assume that there exists an F satisfying (2). We have*

$$\rho_b^* = \min\left\{ x^{\mathrm{T}}Px : \ x^{\mathrm{T}}P\left(Ax - \sum_{i=1}^{m} b_i \mathrm{sign}(b_i^{\mathrm{T}}Px)\right) = 0, \ x \neq 0 \right\}, \tag{5}$$

and

$$\begin{aligned} \rho_d^* = \quad & \max_H \quad \rho \\ & \text{s.t} \quad (A+BH)^{\mathrm{T}}P + P(A+BH) \le 0, \\ & \qquad \rho h_i P^{-1} h_i^{\mathrm{T}} \le 1, \ \ i \in [1,m]. \end{aligned} \tag{6}$$

Proof. Let

$$\rho_1 := \min\left\{ x^{\mathrm{T}}Px : \ x^{\mathrm{T}}P\left(Ax - \sum_{i=1}^{m} b_i \mathrm{sign}(b_i^{\mathrm{T}}Px)\right) = 0, \ x \neq 0 \right\}.$$

From the definition of ρ_b^*, we know that $\rho_b^* \le \rho_1$. To prove that $\rho_b^* = \rho_1$, it suffices to show that $\mathcal{E}(P,\rho)$ is controlled contractively invariant for all $\rho < \rho_1$.

Under the assumption of the existence of F satisfying (2), there exists a $\rho_0 > 0$ such that $\mathcal{E}(P,\rho)$ is controlled contractively invariant. Hence

$$g(k,x) := (kx)^{\mathrm{T}}P\left(kAx - \sum_{i=1}^{m} b_i \mathrm{sign}(b_i^{\mathrm{T}}Pkx)\right) < 0 \quad \forall\, k \in (0,1], x^{\mathrm{T}}Px = \rho_0.$$

Given any $x \neq 0$, it is clear that $g(k,x)$ is a quadratic function of k with $g(0,x) = 0$. If $g(k_0,x) = 0$, then we must have $g(k,x) < 0$ for all $k \in (0,k_0)$. By the definition of ρ_1, we have $g(k,x) < 0$ for all $k \in (0,1), x \in \partial\mathcal{E}(P,\rho_1)$, which means the controlled contractive invariance of $\mathcal{E}(P,k^2\rho_1)$ for all $k \in (0,1)$. Therefore, $\rho_1 = \rho_b^*$.

We next prove (6). Suppose that the right hand side of (6) equals to some number ρ_2 instead of ρ_d^*. We need to prove that $\rho_2 = \rho_d^*$. It is clear that $\rho_2 \ge \rho_d^*$. Let H_* be an optimal solution to (6) and let F satisfy (2). Denote $H(\alpha) = \alpha F + (1-\alpha)H_*$, then

$$(A+BH(\alpha))^{\mathrm{T}}P + P(A+BH(\alpha)) < 0,$$

for all $\alpha \in (0,1]$. Let $\rho = \rho_2 - \varepsilon$. Then for any small number $\varepsilon > 0$, there exists a sufficiently small $\alpha > 0$ such that $\rho h_i(\alpha) P^{-1} h_i^{\mathrm{T}}(\alpha) \leq 1$ for all $i \in [1,m]$. This means that $\mathcal{E}(P,\rho)$ satisfies condition d1) and hence $\rho_d^* \geq \rho = \rho_2 - \varepsilon$. By taking ε arbitrarily small, we obtain $\rho_d^* = \rho_2$ and (6) follows. □

Lemma 2 *Define*

$$\begin{aligned} \lambda^*(\gamma) = \quad & \min_H \quad \lambda_{\max}((A+BH)^{\mathrm{T}}P + P(A+BH)), \\ & \text{s.t} \quad h_i P^{-1} h_i^{\mathrm{T}} \leq \gamma, \ \ i \in [1,m]. \end{aligned} \tag{7}$$

Then $\lambda^*(\gamma)$ is a convex function of γ and it is monotonically decreasing. Moreover, $\lambda^*(1/\rho_d^*) = 0$.

Proof. The monotonically decreasing property is obvious and the convexity of $\lambda^*(\gamma)$ follows from the fact that both $\lambda_{\max}((A+BH)^{\mathrm{T}}P + P(A+BH))$ and $h_i P^{-1} h_i^{\mathrm{T}}$ are convex in H.

We now prove that $\lambda^*(1/\rho_d^*) = 0$. Let H_* be an optimal solution to (6), then we must have $\lambda_{\max}((A+BH_*)^{\mathrm{T}}P + P(A+BH_*)) = 0$ and $h_{*i} P^{-1} h_{*i}^{\mathrm{T}} \leq 1/\rho_d^*$. It follows that $\lambda^*(1/\rho_d^*) \leq 0$. Suppose on the contrary that $\lambda^*(1/\rho_d^*) < 0$, then there exists an H_1 such that $h_{1i} P^{-1} (h_{1i})^{\mathrm{T}} \leq 1/\rho_d^*$, $i \in [1,m]$, and

$$\lambda_{\max}((A+BH_1)^{\mathrm{T}}P + P(A+BH_1)) < 0.$$

Hence there exists a $k \in (0,1)$ such that

$$(A+BkH_1)^{\mathrm{T}}P + P(A+BkH_1) \leq 0.$$

Since $(\rho_d^*/k^2)(kh_{1i}) P^{-1} (kh_{1i})^{\mathrm{T}} \leq 1$, we have

$$\begin{aligned} \rho_d^*/k^2 \leq \quad & \max_H \quad \rho \\ & \text{s.t} \quad (A+BH)^{\mathrm{T}}P + P(A+BH) \leq 0, \\ & \qquad \rho h_i P^{-1} h_i^{\mathrm{T}} \leq 1, \ \ i \in [1,m], \end{aligned}$$

which implies that $\rho_d^*/k^2 \leq \rho_d^*$. This is a contradiction since $k \in (0,1)$.

Therefore, we must have $\lambda^*(1/\rho_d^*) = 0$. □

From Lemma 2, we also know that H_* is an optimal solution to (6) if and only if it is an optimal solutuion to (7) for $\gamma = 1/\rho_d^*$.

We next proceed to investigate the possibility that $\rho_b^* = \rho_d^*$.

Theorem 1 *Let H_* be an optimal solution to (6) and (7) for $\gamma = 1/\rho_d^*$. Denote*

$$W_* = (A+BH_*)^{\mathrm{T}}P + P(A+BH_*).$$

If W_* has only one eigenvalue at 0, then $\rho_b^* = \rho_d^*$.

Proof. It is clear that $\rho_b^* \geq \rho_d^*$. In view of (5), to prove that $\rho_b^* = \rho_d^*$, it suffices to find an $x_0 \in \partial\mathcal{E}(P, \rho_d^*)$ such that

$$\dot{V}(x_0) = 2x_0^{\mathrm{T}} P \left(Ax_0 - \sum_{i=1}^{m} b_i \mathrm{sign}(b_i^{\mathrm{T}} P x_0) \right) = 0. \tag{8}$$

The proof is lengthy. The main idea is to construct such an x_0 from the eigenvector of W_* corresponding to the 0 eigenvalue. To show (8), we need to carry out an intricate analysis on the relation between this eigenvector and H_* by exploring the properties of the optimal solution to (7).

We assume for simplicity and without loss of generality that $P = I$ and $\rho_d^* = 1$. If not so, a state transformation of the form $\hat{x} = (P/\rho_d^*)^{\frac{1}{2}} x$ can be used to satisfy this assumption. Under this assumption, we have

$$\begin{aligned} 1 = \rho_d^* = \quad & \max_H \quad \rho \\ & \text{s.t} \quad (A + BH)^{\mathrm{T}} + (A + BH) \leq 0, \\ & \qquad h_i h_i^{\mathrm{T}} \leq 1/\rho, \quad i \in [1, m]. \end{aligned} \tag{9}$$

By Lemma 2,

$$\begin{aligned} 0 = \lambda^*(1) = \quad & \min_H \quad \lambda_{\max}((A + BH)^{\mathrm{T}} + (A + BH)), \\ & \text{s.t} \quad h_i h_i^{\mathrm{T}} \leq 1, \quad i \in [1, m]. \end{aligned} \tag{10}$$

Let H_* be an optimal solution. There must exist an i such that $h_{*i} h_{*i}^{\mathrm{T}} = 1$. Otherwise, there would be a neighborhood of H_*, $\mathcal{N}(H_*)$, such that $h_i h_i^{\mathrm{T}} \leq 1$ for all $H \in \mathcal{N}(H_*)$. Since there exists an F such that $(A + BF)^{\mathrm{T}} + A + BF < 0$, there exists an $H = \alpha F + (1 - \alpha) H_* \in \mathcal{N}(H_*)$ such that $(A + BH)^{\mathrm{T}} + (A + BH) < 0$, which is in contradiction with that $\lambda^*(1) = 0$. Let the number of i such that $h_{*i} h_{*i}^{\mathrm{T}} = 1$ be m_1. Without loss of generality, assume that

$$h_{*i} h_{*i}^{\mathrm{T}} = 1, \quad i \in [1, m_1],$$

and

$$h_{*i} h_{*i}^{\mathrm{T}} < 1, \quad i \in [m_1 + 1, m].$$

Otherwise, the columns of B and the rows of H_* can be permuted to make it so.

The optimality of the solution H_* to (10) means that $\lambda_{\max}((A + BH)^{\mathrm{T}} + (A + BH))$ cannot be decreased by varying H in a neighborhood of H_* along any direction which keeps it within the constraint. Let $H = H_* + k\Delta H$, where ΔH represents the varying direction, then $\lambda_{\max}((A + BH)^{\mathrm{T}} + (A + BH))$ cannot be decreased along the following direction ΔH,

1) For $i \in [1, m_1]$, Δh_i is tangential to the sphere surface $h_i h_i^{\mathrm{T}} = 1$ at h_{*i}, or, Δh_i points inward of the sphere $h_i h_i^{\mathrm{T}} \leq 1$ from h_{*i};

2) For $i \in [m_1 + 1, m]$, $\Delta h_i^{\mathrm{T}} \in \mathbf{R}^n$.

Let us get more exact implication of the optimality by using the eigenvalue perturbation theory (see, e.g., [14]). Let $v \in \mathbf{R}^n$ be a unit eigenvector of W_* corresponding to the single eigenvalue at 0, i.e.,

$$v^{\mathrm{T}} v = 1, \quad W_* v = 0.$$

For the perturbed matrix $H = H_* + k\Delta H$, denote

$$\lambda_{\max}(k, \Delta H) = \lambda_{\max}((A + B(H_* + k\Delta H))^{\mathrm{T}} + A + B(H_* + k\Delta H)).$$

Then

$$\partial \lambda_{\max} / \partial k|_{k=0} = v^{\mathrm{T}}((B\Delta H)^{\mathrm{T}} + B\Delta H)v.$$

The optimality of the solution H_* implies that for $i \in [1, m_1]$,

$$v^{\mathrm{T}}((b_i \Delta h_i)^{\mathrm{T}} + b_i \Delta h_i)v = 0 \ \ \forall \Delta h_i \ \text{ such that } \Delta h_i h_{*i}^{\mathrm{T}} = 0, \tag{11}$$

where $\Delta h_i h_{*i}^{\mathrm{T}} = 0$ means that Δh_i is tangential to the sphere surface $h_i h_i^{\mathrm{T}} = 1$, and

$$v^{\mathrm{T}}((b_i h_{*i})^{\mathrm{T}} + b_i h_{*i})v \leq 0, \tag{12}$$

and for $i \in [m_1 + 1, m]$,

$$v^{\mathrm{T}}((b_i \Delta h_i)^{\mathrm{T}} + b_i \Delta h_i)v = 0 \ \ \forall \Delta h_i^{\mathrm{T}} \in \mathbf{R}^n. \tag{13}$$

Since there exists an F such that $(A + BF)^{\mathrm{T}} + A + BF < 0$ and $\lambda_{\max}((A + BH)^{\mathrm{T}} + A + BH)$ is convex in H, there is a direction $\Delta H = F - H_*$ such that

$$v^{\mathrm{T}}((B\Delta H)^{\mathrm{T}} + B\Delta H)v < 0.$$

This implies that there is at least one $i \in [1, m_1]$ such that

$$v^{\mathrm{T}}((b_i h_{*i})^{\mathrm{T}} + b_i h_{*i})v < 0. \tag{14}$$

Let the number of i satisfying (14) be m_0. Without loss of generality, we assume that

$$v^{\mathrm{T}}((b_i h_{*i})^{\mathrm{T}} + b_i h_{*i})v < 0 \ \ \forall i \in [1, m_0]. \tag{15}$$

Then we have, for all $i \in [m_0 + 1, m]$,

$$v^{\mathrm{T}}((b_i \Delta h_i)^{\mathrm{T}} + b_i \Delta h_i)v = 0 \ \ \forall \Delta h_i^{\mathrm{T}} \in \mathbf{R}^n. \tag{16}$$

Let $e_i \in \mathbf{R}^n$ be the ith unit vector whose only nonzero element is the ith one and it equals 1. Let $U \in \mathbf{R}^{n\times n}$ be a unitary matrix such that

$$U^{\mathrm{T}}U = UU^{\mathrm{T}} = I, \;\; e_1^{\mathrm{T}}U = h_{*1}.$$

To utilize (11), consider $\Delta h_1 = e_j^{\mathrm{T}}U, j \in [2,n]$. Then

$$\Delta h_1 h_{*1}^{\mathrm{T}} = e_j^{\mathrm{T}}Uh_{*1}^{\mathrm{T}} = e_j^{\mathrm{T}}UU^{\mathrm{T}}e_1 = e_j^{\mathrm{T}}e_1 = 0.$$

From (11), we obtain

$$v^{\mathrm{T}}((b_1e_j^{\mathrm{T}}U)^{\mathrm{T}} + b_1e_j^{\mathrm{T}}U)v = 0 \;\; \forall\, j \in [2,n].$$

Let $\hat{v} = Uv$, $\hat{b}_1 = Ub_1$, then

$$\hat{v}^{\mathrm{T}}((\hat{b}_1e_j^{\mathrm{T}})^{\mathrm{T}} + \hat{b}_1e_j^{\mathrm{T}})\hat{v} = 2\hat{v}_j\hat{v}^{\mathrm{T}}\hat{b}_1 = 0 \;\; \forall\, j \in [2,n], \tag{17}$$

and (15) implies that

$$\hat{v}^{\mathrm{T}}((\hat{b}_1e_1^{\mathrm{T}})^{\mathrm{T}} + \hat{b}_1e_1^{\mathrm{T}})\hat{v} = 2\hat{v}_1\hat{v}^{\mathrm{T}}\hat{b}_1 < 0. \tag{18}$$

Equations (17) and (18) jointly show that $\hat{v}_j = 0$ for $j \in [2,n]$ and $\hat{v}_1 \neq 0$. Since $\hat{v}$ is a unit vector, we must have $\hat{v} = \pm e_1^{\mathrm{T}}$. Recalling that $e_1^{\mathrm{T}}U = h_{*1}$ and $\hat{v} = Uv$, we obtain

$$v^{\mathrm{T}} = \hat{v}^{\mathrm{T}}U = \pm e_1^{\mathrm{T}}U = \pm h_{*1}. \tag{19}$$

This means that h_{*1}^{T} is a unit eigenvector of W_*. For simplicity, we let $v = h_{*1}^{\mathrm{T}}$. It follows from same arguments that $v^{\mathrm{T}} = \pm h_{*i}$, $i \in [2, m_0]$. Therefore, $h_{*i} = \pm h_{*1}$ for $i \in [2, m_0]$ and

$$W_*h_{*i}^{\mathrm{T}} = 0 \quad \forall\, i \in [1, m_0]. \tag{20}$$

With $v = h_{*1}^{\mathrm{T}}$, condition (15) can be rewritten as

$$h_{*1}((b_ih_{*i})^{\mathrm{T}} + b_ih_{*i})h_{*1}^{\mathrm{T}} = 2h_{*1}h_{*i}^{\mathrm{T}}b_i^{\mathrm{T}}h_{*1}^{\mathrm{T}} = 2h_{*1}h_{*1}^{\mathrm{T}}b_i^{\mathrm{T}}h_{*i}^{\mathrm{T}} < 0.$$

It follows that

$$b_i^{\mathrm{T}}h_{*i}^{\mathrm{T}} < 0, \quad i \in [1, m_0]. \tag{21}$$

For $i \in [m_0+1, m]$, it follows from (16) that $v^{\mathrm{T}}((b_iv^{\mathrm{T}})^{\mathrm{T}} + b_iv^{\mathrm{T}})v = 0$. Thus, $v^{\mathrm{T}}b_i = 0$, i.e.,

$$b_i^{\mathrm{T}}h_{*1} = 0, \quad i \in [m_0+1, m]. \tag{22}$$

Let $x_0 = h_{*1}^{\mathrm{T}}$. Then for $i \in [1, m_0]$, we have $h_{*i}x_0 = \pm 1$. If $h_{*i}x_0 = 1$, then $h_{*i} = h_{*1}$ and from (21),

$$-\mathrm{sign}(b_i^{\mathrm{T}}x_0) = -\mathrm{sign}(b_i^{\mathrm{T}}h_{*i}) = 1 = h_{*i}x_0.$$

Similarly, if $h_{*i}x_0 = -1$, then $h_{*i} = -h_{*1}$ and from (21),

$$-\text{sign}(b_i^{\mathrm{T}}x_0) = \text{sign}(b_i^{\mathrm{T}}h_{*i}) = -1 = h_{*i}x_0.$$

In summary, we have

$$-\text{sign}(b_i^{\mathrm{T}}x_0) = h_{*i}x_0, \quad i \in [1, m_0]. \tag{23}$$

As for $i \in [m_0+1, m]$, since $x_0^{\mathrm{T}}b_i = h_1^{\mathrm{T}}b_i = 0$,

$$x_0^{\mathrm{T}}b_i h_{*i}x_0 = -x_0^{\mathrm{T}}b_i\,\text{sign}(b_i^{\mathrm{T}}x_0) = 0, \quad i \in [m_0+1, m]. \tag{24}$$

Combining (23) and (24), the derivative $\dot{V}(x_0)$ under the control of $u_i = -\text{sign}(b_i^{\mathrm{T}}Px)$ is

$$\begin{aligned}
\dot{V}(x_0) &= x_0^{\mathrm{T}}(A^{\mathrm{T}} + A)x_0 - 2\sum_{i=1}^{m} x_0^{\mathrm{T}}b_i\,\text{sign}(b_i^{\mathrm{T}}x_0) \\
&= x_0^{\mathrm{T}}(A^{\mathrm{T}} + A)x_0 + 2\sum_{i=1}^{m} x_0^{\mathrm{T}}b_i h_{*i}x_0 \\
&= x_0^{\mathrm{T}}(A^{\mathrm{T}} + A + (BH_*)^{\mathrm{T}} + BH_*)x_0 \\
&= x_0^{\mathrm{T}}W_*x_0 \\
&= 0.
\end{aligned}$$

Since $x_0 \in \mathcal{E}(P, \rho_d^*)$, it follows from (5) that $\rho_b^* \le \rho_d^*$. □

From Theorem 1, we see that conditions b1) and d1) are equivalent if W_* has a single eigenvalue at 0. If W_* has two identical eigenvalues at 0, then it is possible that $\rho_d^* < \rho_b^*$.

Example 1 We consider a special case where

$$A = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \quad B = A, \quad P = I.$$

The optimal solution to (6) is $H_* = -I$ with $\rho_d^* = 1$ and $W_* = 0$. Since W_* has two identical eigenvalues at 0, Theorem 1 is not applicable. For second-order systems, ρ_b^* can be obtained by computing $\dot{V}(x)$ under the control $u_i = -\text{sign}(b_i^{\mathrm{T}}Px)$ for all $x \in \partial\mathcal{E}(P, \rho)$ and iterating on ρ with bisection method. For this system, we have $\rho_b^* = 1.2504 > 1 = \rho_d^*$.

In the next example, we illustrate the possibility that W_* has multiple eigenvalues at 0.

Example 2 There are three parameters, A, B and P, that determine the optimal W_*. To ensure that $\mathcal{E}(P, \rho)$ can be made invariant for some $\rho > 0$, we designed a fixed algorithm to determine P from A and B. For instance, a feedback matrix F is chosen to place the eigenvalues of $A+BF$ to some fixed location and P is solved from $(A+BF)^{\mathrm{T}}P + P(A+BF) = -I$.

In each test, we fix the matrix A and let B be generated by "randn(2)". We have 1000 samples of B. Because of numerical error, we set a threshold $\varepsilon = 10e^{-5}$. If $\lambda_{\min}(W_*) \geq -\varepsilon$, then we say that W_* has two identical eigenvalues at 0, otherwise it has two distinct eigenvalues. It turns out that the outcome depends on the matrix A. Here are two cases:

1) $A = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, the number of B's that result in two identical zero eigenvalues of W_* is 230;

2) $A = \begin{bmatrix} 1 & 0.5 \\ -0.5 & 1 \end{bmatrix}$, the number of B's that result in two identical zero eigenvalues of W_* is 103.

Of course, these numbers are changing every time we run the test but they are close to these values.

These computational results show that the condition in Theorem 1 is usually satisfied. This means that the controlled invariance of an ellipsoid is usually equivalent to the existence of a feedback linear inside the ellipsoid that makes it invariant.

4 The equivalence of (c1,d1) and (b1,c1)

We will actually show in this section that the pairs (c1,d1) and (b1,c1) are not equivalent for multiple input systems. Define

$$\rho_c^* := \max\left\{\rho > 0 : \ x^{\mathrm{T}} P(Ax + B\mathrm{sat}(Fx)) \leq 0 \ \ \forall x \in \mathcal{E}(P,\rho)\right\}.$$

Then ρ_c^* and ρ_d^* are generally not equal. Usually, we have $\rho_c^* < \rho_d^* \leq \rho_b^*$, but it is also possible that $\rho_c^* > \rho_d^*$.

In [10], we established a sufficient condition for an ellipsoid $\mathcal{E}(P,\rho)$ to be contractively invariant under $u = \mathrm{sat}(Fx)$: there exists an $H \in \mathbf{R}^{m\times n}$ such that $\mathcal{E}(P,\rho) \subset \mathcal{L}(H)$ and

$$(A + B(DF + D^-H))^{\mathrm{T}} P + P(A + B(DF + D^-H)) < 0 \quad \forall D \in \mathcal{D},$$

where $\mathcal{D}$ is the set of $m \times m$ diagonal matrices with only 0 or 1 diagonal elements and $D^- = I - D$. The largest ρ satisfying the sufficient condition can be computed as follows,

$$\begin{aligned} \rho^* = \ & \max_H \ \ \rho \\ & \text{s.t} \quad (A + B(DF + D^-H))^{\mathrm{T}} P + P(A + B(DF + D^-H)) \leq 0 \quad \forall D \in \mathcal{D}, \\ & \qquad \rho h_i P^{-1} h_i^{\mathrm{T}} \leq 1, \ \ i \in [1, m]. \end{aligned} \tag{25}$$

In [9], we showed that we usually have $\rho^* = \rho_c^*$. Since the constraint in (25) is more restrictive than that of (6) (noting that if $D = 0$, then $DF + D^-H = H$), we must have $\rho^* \leq \rho_d^*$. Therefore, we usually have $\rho_c^* \leq \rho_d^*$.

The possibility of the situation that $\rho_c^* > \rho_d^*$ can be observed from the possibility that $\rho_b^* > \rho_d^*$, which we have shown in Section 3. In [6], we showed that if an ellipsoid is controlled contractively invariant, then there exists a saturated high gain linear feedback of the form $u = \text{sat}(kB^{\mathrm{T}}Px)$ to make it contractively invariant. This implies that F can be constructed such that ρ_c^* is arbitrarily close to ρ_b^*. In the case that $\rho_b^* > \rho_d^*$, we can always find an F such that $\rho_c^* > \rho_d^*$.

Example 3 In this example, we show the possibility that $\rho_c^* \leq \rho_d^*$. As in Example 2, we fix the matrix $A = \begin{bmatrix} 1 & 0.5 \\ -0.5 & 1 \end{bmatrix}$ and let B be generated by "randn(2)". The feedback matrix F is chosen to place the eigenvalues of $A + BF$ to -1 ± 0.5 and P is solved from $(A + BF)^{\mathrm{T}}P + P(A + BF) = -I$.

In the test, we generated 1000 matrices B. Among these B's, 894 of which result in $\rho^* = \rho_c^*$ and in all of these cases, $\rho_c^* < \rho_d^*$. In 665 of these cases, we have $\rho_d^* > 1.01\rho_c^*$.

Example 4 In this example, we would like to illustrate the possibility that $\rho_c^* > \rho_d^*$. We use the same system as in Example 1, where $A = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, B = A$ and $P = I$. We know that $\rho_b^* = 1.2504$ and $\rho_d^* = 1$. By choosing $F = -\text{sat}(20B^{\mathrm{T}}Px)$, we obtain $\rho_c^* = 1.248$, which is greater than $\rho_d^* = 1$.

Let us give some remark on the pair (b1,c1). The relation between ρ_b^* and ρ_c^* can be derived from the relation between other pairs. We know that $\rho_b^* \geq \rho_c^*$. Since we usually have $\rho_b^* = \rho_d^*$ (see Example 2), and $\rho_d^* > \rho_c^*$ (see Example 3), it is usually true that $\rho_b^* > \rho_c^*$.

5 Conclusions

We attempted to extend our previous results on the equivalence of several set invariance conditions for single input systems to multi-input systems. It turns out that the equivalence hold conditionally for some pairs and does not hold for some other pairs. Computational examples were worked out to illustrate the possibility of the equivalence of each pair of conditions.

References

[1] F. Blanchini, "Set invariance in control – a survey", *Automatica*, Vol. 35, No.11, pp. 1747-1767, 1999.

[2] S. Boyd, L. El Ghaoui, E. Feron and V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory*, SIAM Studies in Appl. Mathematics, Philadelphia, 1994.

[3] E. J. Davison and E. M. Kurak, "A computational method for determining quadratic Lyapunov functions for non-linear systems," *Automatica*, Vol. 7, pp. 627-636, 1971.

[4] E. G. Gilbert and K. T. Tan, "Linear systems with state and control constraints: the theory and application of maximal output admissible sets," *IEEE Trans. Automat. Contr.* Vol. 36, pp. 1008-1020, 1991.

[5] H. Hindi & S. Boyd, "Analysis of linear systems with saturating using convex optimization," *Proc of the 37th IEEE CDC*, pp903-908, Florida, 1998.

[6] T. Hu and Z. Lin. *Control Systems with Actuator Saturation: Analysis and Design*, Birkhäuser, Boston, 2001.

[7] T. Hu and Z. Lin, "Exact characterization of invariant ellipsoids for single input linear systems subject to actuator saturation," *IEEE Trans. Automat. Contr.*, scheduled to appear in February 2002.

[8] T. Hu and Z. Lin, "On enlarging the basin of attraction for linear systems under saturated linear feedback," *Sys. & Contr. Lett.*, Vol. 40, No. 1, pp. 59-69, May, 2000.

[9] T. Hu and Z. Lin, "On the necessity of a recent set invariance condition under actuator saturation," to be presented at *American Control Conference*, Anchorage, Alaska, 2002.

[10] T. Hu, Z. Lin and B. M. Chen, "An analysis and design method for linear systems subject to actuator saturation and disturbance," *Automatica*, Vol. 38, No. 2, pp. 351-359, 2002.

[11] T. Hu, Z. Lin and B. M. Chen, "Analysis and design for linear discrete-time systems subject to actuator saturation," *Sys. & Contr. Lett.*, Vol. 45, No. 2, pp. 97-112, 2002.

[12] H. Khalil, *Nonlinear Systems*, Prentice-Hall, Upper Saddle River, New Jersey, 1996.

[13] K. A. Loparo and G. L. Blankenship, "Estimating the domain of attraction of of nonlinear feedback systems," *IEEE Trans. Automat. Contr.*, Vol. 23, No.4, pp. 602-607, 1978.

[14] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

[15] A. Vanelli and M. Vidyasagar, "Maximal Lyapunov functions and domain of attraction for autonomous nonlinear systems," *Automatica*, 21(1), pp. 69-80, 1985.

[16] S. Weissenberger, "Application of results from the absolute stability to the computation of finite stability domains," *IEEE Trans. Automatic Control*, AC-13, pp124-125, 1968.

[17] G. F. Wredenhagen and P. R. Belanger, "Piecewise-linear LQ control for systems with input constraints," *Automatica*, Vol. 30, pp. 403-416, 1994.

Publication 25

Submitted to *Systems & Control Letters*

On the Necessity of a Recent Set Invariance Condition under Actuator Saturation

Tingshu Hu† Zongli Lin†

† Department of Electrical and Computer Engineering
University of Virginia
Charlottesville, VA 22903, U.S.A.
Email: th7f, zl5y@virginia.edu

Abstract

A sufficient condition for an ellipsoid to be invariant was obtained recently and an LMI approach was developed to find the largest ellipsoid satisfying the condition. This condition was later shown to be necessary for the single input case. This paper is dedicated to the multi-input case. We will examine when this condition is also necessary for multi-input systems. Our investigation is based on studying the optimal solution to a related LMI problem. A criterion is presented to determine when the condition is necessary and when the largest invariant ellipsoid has been obtained by using the LMI method.

Keywords: invariant ellipsoid, contractive invariance, actuator saturation

[1]This work was supported in part by the US Office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

1 Introduction

In this paper, we will continue to study the set invariance property for a linear system under saturated feedback,

$$\dot{x} = Ax + B\text{sat}(Fx). \tag{1}$$

This problem has been studied in our recent works [7, 9]. We have restricted our attention to invariant ellipsoids since quadratic Lyapunov functions are the most popular and the results can be put into simple and compact forms, which makes analysis and design easily implementable. Moreover, our method can be applied to different ellipsoids. The union of multiple invariant ellipsoids forms a new invariant set. In the literature, invariant ellipsoids have been used to estimate the domain of attraction for nonlinear systems (see *e.g.*, [1, 2, 3, 4, 5, 10, 11, 13] and the references therein). The problem of estimating the domain of attraction for (1) has been a focus of study in recent years.

For a matrix $F \in \mathbf{R}^{m\times n}$, denote the ith row of F as f_i and define

$$\mathcal{L}(F) := \left\{x \in \mathbf{R}^n : \ |f_i x| \le 1, \ i = 1, 2, \cdots, m\right\}.$$

If F is a feedback gain matrix, then $\mathcal{L}(F)$ is the region where the feedback control $u = \text{sat}(Fx)$ is linear in x. We call $\mathcal{L}(F)$ the linear region of the saturated feedback $\text{sat}(Fx)$, or simply, the linear region of saturation.

Let $P \in \mathbf{R}^{n\times n}$ be a positive-definite matrix. For a positive number ρ, denote

$$\mathcal{E}(P, \rho) = \left\{x \in \mathbf{R}^n : x^{\mathrm{T}} P x \le \rho\right\}.$$

If

$$(A + BF)^{\mathrm{T}} P + P(A + BF) < 0$$

and $\mathcal{E}(P, \rho) \subset \mathcal{L}(F)$, then $\mathcal{E}(P, \rho)$ is an invariant ellipsoid inside the domain of attraction. The largest of these $\mathcal{E}(P, \rho)$'s was used as an estimate of the domain of attraction in the earlier literature (see *e.g.*, [4, 14]). This saturation avoidance estimation method, though simple, could be very conservative. Recent efforts have been made to extend the ellipsoid beyond the linear region $\mathcal{L}(F)$ (see, *e.g.*, [5, 10]). In particular, simple and general methods have been derived by applying the absolute stability analysis tools, such as the circle and Popov criteria, where the saturation is treated as a locally sector bounded nonlinearity.

More recently, we developed a new sufficient condition for an ellipsoid to be invariant in [9] (see also [6]). It was shown that this condition is less conservative than the existing conditions resulting from the circle criterion or the vertex analysis. The most important feature of this new condition is that it can be expressed as LMIs in terms of all varying parameters and hence

can easily be used for controller synthesis. A recent discovery makes this condition even more attractive. In [7], we showed that for the single input case, this condition is also necessary, thus the largest ellipsoid obtained with the LMI approach is actually the largest one. With this new finding, we are tempted to try to understand if this condition is also necessary for the multi-input case. Our investigation identifies cases where this condition is also necessary for multiple input systems.

Notation: In this paper, we use $\text{sat} : \mathbf{R}^m \to \mathbf{R}^m$ to denote the standard saturation function of appropriate dimensions. For $u \in \mathbf{R}^m$, the ith component of $\text{sat}(u)$ is $\text{sign}(u_i)\min\{1, |u_i|\}$. The infinity norm of u is denoted as $|u|_\infty$. For an $m \times n$ matrix H, we use h_i to denote its ith row and for an $n \times m$ matrix B, we use b_i to denote its ith column.

2 A Sufficient Condition for Set Invariance

Consider the linear system subject to input saturation,

$$\dot{x} = Ax + Bu, \quad x \in \mathbf{R}^n, \quad u \in \mathbf{R}^m, \quad |u|_\infty \le 1. \tag{2}$$

Under a saturated linear feedback $u = \text{sat}(Fx)$, the closed-loop system is,

$$\dot{x} = Ax + B\text{sat}(Fx). \tag{3}$$

Given a positive definite matrix P, let $V(x) = x^{\mathrm{T}}Px$. The ellipsoid $\mathcal{E}(P,\rho)$ is said to be (contractively) invariant if

$$\dot{V}(x) = 2x^{\mathrm{T}}P(Ax + B\text{sat}(Fx)) \le (<)0$$

for all $x \in \mathcal{E}(P,\rho) \setminus \{0\}$. Clearly, if $\mathcal{E}(P,\rho)$ is contractively invariant, then it is inside the domain of attraction. We need more notation to present the sufficient condition for $\mathcal{E}(P,\rho)$ to be contractively invariant.

Let $\mathcal{D}$ be the set of $m \times m$ diagonal matrices whose diagonal elements are either 1 or 0. There are 2^m elements in $\mathcal{D}$. Suppose that each element of $\mathcal{D}$ is labeled as D_i, $i = 1, 2, \cdots, 2^m$. Then, $\mathcal{D} = \left\{D_i : i \in [1, 2^m]\right\}$. Denote $D_i^- = I - D_i$. Clearly, D_i^- is also an element of $\mathcal{D}$ if $D_i \in \mathcal{D}$. Given two matrices $F, H \in \mathbf{R}^{m\times n}$,

$$\left\{D_iF + D_i^-H : i \in [1, 2^m]\right\}$$

is the set of matrices formed by choosing some rows from F and the rest from H.

Theorem 1 ([9, 6]). *Given an ellipsoid $\mathcal{E}(P,\rho)$, if there exists an $H \in \mathbf{R}^{m\times n}$ such that*

$$(A + B(D_iF + D_i^-H))^{\mathrm{T}}P + P(A + B(D_iF + D_i^-H)) < 0, \quad \forall i \in [1, 2^m], \tag{4}$$

and $\mathcal{E}(P,\rho) \subset \mathcal{L}(H)$, then $\mathcal{E}(P,\rho)$ is contractively invariant under the feedback $u = \text{sat}(Fx)$.

It is clear that the condition,

$$(A+BF)^{\mathrm{T}}P+P(A+BF)<0, \tag{5}$$

which corresponds to (4) for $D_i = I$, is necessary for the contractive invariance of $\mathcal{E}(P,\rho)$ for any $\rho > 0$. Hence we assume throughout the paper that (5) is satisfied.

For the single input case ($m = 1$), we have shown in [7] that the condition in Theorem 1 is also necessary.

Theorem 2 *Assume that $m = 1$. Given an ellipsoid $\mathcal{E}(P,\rho)$, suppose that (5) is satisfied. Then $\mathcal{E}(P,\rho)$ is contractively invariant under the feedback $u = \mathrm{sat}(Fx)$ if and only if there exists an $H \in \mathbf{R}^{1\times n}$ such that*

$$(A+BH)^{\mathrm{T}}P+P(A+BH)<0, \tag{6}$$

and $\mathcal{E}(P,\rho) \subset \mathcal{L}(H)$.

Here we note that when $m = 1$, there are only two inequalities in (4), namely, (5) and (6).

3 Necessity of the Set Invariance Condition

In this section, we will study the necessity of the condition in Theorem 1 for multi-input case. First, let us consider the largest ellipsoid that satisfies the condition. Let $P > 0$ be given, define

$$\begin{aligned} \rho^* := \ & \sup_H \rho \\ \text{s.t.} \quad & \text{a)} \ (A+B(D_iF+D_i^-H))^{\mathrm{T}}P+P(A+B(D_iF+D_i^-H))<0, \quad i\in[1,2^m], \\ & \text{b)} \ \rho h_jP^{-1}h_j^{\mathrm{T}}\le 1, \quad j\in[1,m]. \end{aligned} \tag{7}$$

Recall from [8, 6] that constraint b) is equivalent to $\mathcal{E}(P,\rho) \subset \mathcal{L}(H)$. Consider a closely related optimization problem

$$\begin{aligned} \rho_1^* := \ & \sup_H \rho \\ \text{s.t.} \quad & \text{a)} \ (A+B(D_iF+D_i^-H))^{\mathrm{T}}P+P(A+B(D_iF+D_i^-H))\le 0, \quad i\in[1,2^m], \\ & \text{b)} \ \rho h_jP^{-1}h_j^{\mathrm{T}}\le 1, \quad j\in[1,m]. \end{aligned} \tag{8}$$

The only difference between (7) and (8) is the "$<$" in (7a) and "$\le$" in (8a). We claim that $\rho^* = \rho_1^*$. It is easy to see that $\rho^* \le \rho_1^*$ since (7a) is more restrictive than (8a). Let $(\rho_1^*, \overline{H})$ be an optimal solution to (8). To prove that $\rho^* \ge \rho_1^*$, it suffices to show that given any $\varepsilon > 0$, there exist H and ρ, with $\rho \ge \rho_1^* - \varepsilon$, such that (7a) and (7b) are satisfied. Let $H = (1-\delta)\overline{H} + \delta F$, where $\delta \in (0,1)$ is to be determined later. Recalling that $D_i + D_i^- = I$, we have

$$D_iF + D_i^-H = \delta F + (1-\delta)(D_iF + D_i^-\overline{H}).$$

By assumption, we have (5). It follows that

$$\begin{aligned}
&(A+B(D_iF+D_i^-H))^{\mathrm{T}}P+P(A+B(D_iF+D_i^-H))\\
&=\delta\left((A+BF)^{\mathrm{T}}P+P(A+BF)\right)\\
&\quad+(1-\delta)\left((A+B(D_iF+D_i^-\overline{H}))^{\mathrm{T}}P+P(A+B(D_iF+D_i^-\overline{H}))\right)\\
&<0,
\end{aligned}$$

for all $\delta\in(0,1)$. Hence (7a) is satisfied for all these H's. Since $H-\overline{H}=\delta(F-\overline{H})$ can be made arbitrarily small and $\rho_1^*\bar{h}_jP^{-1}\bar{h}_j^{\mathrm{T}}\le 1$, given any $\varepsilon>0$, there exists a $\delta>0$ such that

$$(\rho_1^*-\varepsilon)h_jP^{-1}h_j^{\mathrm{T}}=(\rho_1^*-\varepsilon)(\bar{h}_j+\delta(f_j-\bar{h}_j))P^{-1}(\bar{h}_j+\delta(f_j-\bar{h}_j))^{\mathrm{T}}\le 1.$$

This means that $\rho=\rho_1^*-\varepsilon$ and $H=(1-\delta)\overline{H}+\delta F$ satisfies (7a) and (7b) for some $\delta>0$. Thus we have $\rho^*=\rho_1^*$. Here we note that condition (5) is critical in establishing the equality $\rho^*=\rho_1^*$.

From Theorem 1, we know that if $\rho<\rho^*$, then $\mathcal{E}(P,\rho)$ is contractively invariant. If we can conclude that $\rho\ge\rho^*$ implies that $\mathcal{E}(P,\rho)$ is not contractively invariant, then the condition in Theorem 1 is also necessary. However, this is not always the case.

Define

$$\rho_c:=\sup\left\{\rho>0:\ \ \mathcal{E}(P,\rho)\ \text{ is contractively invariant}\right\}.$$

It is clear that $\rho_c\ge\rho^*$. We will show that $\rho_c=\rho^*$ is conditional.

Also, let $(\rho^*,\overline{H})$ be an optimal solution to (8). There must be a j such that $\rho^*\bar{h}_jP^{-1}\bar{h}_j^{\mathrm{T}}=1$ and there must be an i such that

$$\lambda_{\max}\left((A+B(D_iF+D_i^-\overline{H}))^{\mathrm{T}}P+P(A+B(D_iF+D_i^-\overline{H}))\right)=0. \tag{9}$$

Theorem 3 *Let $(\rho^*,\overline{H})$ be an optimal solution to (8). Suppose that*

1) *there is only one j such that $\rho^*\bar{h}_jP^{-1}\bar{h}_j^{\mathrm{T}}=1$ (i.e., the boundary of $\mathcal{E}(P,\rho^*)$ only touches one pair of the planes $\bar{h}_jx=\pm1$);*

2) *there is only one i satisfying (9), the matrix in (9) has a single eigenvalue at 0 and the only nonzero element in D_i^- is the jth diagonal one ($D_i^-\overline{H}$ chooses only $\bar{h}_j$).*

Let $x_0=\rho^*P^{-1}\bar{h}_j^{\mathrm{T}}$, then x_0 is the unique intersection of $\mathcal{E}(P,\rho^*)$ with $\bar{h}_jx=1$. If

3) *$|f_kx_0|\le 1$ for all $k\ne j$,*

then $\rho^*=\rho_c$.

Proof. To simplify the proof, we would like to make some special assumptions. First, we assume that $\rho^* = 1$. Otherwise we can scale the matrix P to make it so. We also assume that $j = 1$. Otherwise we can permute the columns of the B matrix.

Next, we assume some special forms of the matrices P and H. Suppose that we have a state transformation, $x \to z = Tx$. Then the invariance of $\mathcal{E}(P, \rho)$ for the x system is equivalent to the invariance of $\mathcal{E}(\bar{P}, \rho)$ for the z system

$$\dot{z} = \hat{A}z + \hat{B}\text{sat}(\hat{F}z),$$

with

$$\hat{P} = (T^{-1})^{\mathrm{T}} P T^{-1}, \quad \hat{A} = TAT^{-1}, \quad \hat{B} = TB, \quad \hat{F} = FT^{-1}.$$

Also let $\hat{\overline{H}} = \overline{H}T^{-1}$. With the above transformation, the three conditions 1)-3) in Theorem 3 remain unchanged. In view of the above arguments, we can assume that $P = I$ and $\overline{h}_1 = [1 \;\; 0 \;\; \cdots \;\; 0]$. Otherwise, we can use a unitary transformation ($T^{\mathrm{T}}T = I$) to make it so, noting that $\rho^* \overline{h}_1 P^{-1} \overline{h}_1^{\mathrm{T}} = 1$ and $\rho^* = 1$.

In summary, we assume that $\rho^* = 1, P = I$ and

$$\overline{h}_1 = [1 \;\; 0 \;\; \cdots \;\; 0].$$

In this case, we have $x_0 = [1 \;\; 0 \;\; \cdots \;\; 0]^{\mathrm{T}}$, $x_0^{\mathrm{T}} P x_0 = 1 = \rho^*$ and $\overline{h}_1 x_0 = 1$, *i.e.*, x_0 is the unique intersection of the ellipsoid $\mathcal{E}(P, \rho^*)$ with the hyperplane $\overline{h}_1 x = 1$.

Denote

$$\begin{aligned} Q(h_1) &= \left(A + B \begin{bmatrix} h_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} \right)^{\mathrm{T}} P + P \left(A + B \begin{bmatrix} h_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} \right) \\ &= \left(A + \sum_{j=2}^{m} b_j f_j \right)^{\mathrm{T}} P + P \left(A + \sum_{j=2}^{m} b_j f_j \right) + (b_1 h_1)^{\mathrm{T}} P + P b_1 h_1. \end{aligned}$$

Under condition 2), we have $\lambda_{\max}\left(Q(\overline{h}_1)\right) = 0$. Let the unit eigenvector of $Q(\overline{h}_1)$ corresponding to the zero eigenvalue be v, *i.e.*, $v^{\mathrm{T}}v = 1$ and $Q(\overline{h}_1)v = 0$. In the remaining part of the proof, we will first show that $Q(\overline{h}_1)x_0 = 0$, and under condition 3), we will further have $\dot{V}(x_0) = 0$. This leads to $\rho_c \leq \rho^*$ and hence $\rho_c = \rho^*$.

Step 1: $Q(\overline{h}_1)x_0 = 0$.

Under condition 1), the only equality in (8b) is

$$\rho^* \overline{h}_1 P^{-1} \overline{h}_1^{\mathrm{T}} = 1,$$

and all the others have strict " $<$ ". Under condition 2), we have

$$(A + B(D_i F + D_i^- \overline{H}))^{\mathrm{T}} P + P(A + B(D_i F + D_i^- \overline{H})) < 0$$

for all $i \in [1, 2^m]$, except

$$Q(\overline{h}_1) \leq 0.$$

Hence, there is a neighborhood of $\overline{H}$, $\mathcal{N}(\overline{H})$, where all the conditions in (8a) and (8b) are satisfied except $\rho^* h_1 P^{-1} h_1^{\mathrm{T}} \leq 1$ and $Q(h_1) \leq 0$. Let us restrict $H \in \mathcal{N}(\overline{H})$, we must also have

$$\begin{aligned} \rho^* \quad &= \sup_{H \in \mathcal{N}(\overline{H})} \rho \\ \text{s.t.} \quad &(8a), \ (8b), \end{aligned} \tag{10}$$

with an optimizer $\overline{H}$. Since for any $H \in \mathcal{N}(\overline{H})$, all the conditions in (8a) and (8b) are satisfied except $\rho^* h_1 P^{-1} h_1^{\mathrm{T}} \leq 1$ and $Q(h_1) \leq 0$, the pair $(\rho^*, \overline{H})$ must also be the optimal solution to

$$\sup_{H \in \mathcal{N}(\overline{H})} \rho \tag{11}$$

$$\text{s.t.} \quad \rho h_1 P^{-1} h_1 \leq 1, \tag{12}$$

$$Q(h_1) \leq 0. \tag{13}$$

The optimality of the solution means that if we scale down h_1 from $\overline{h}_1$ to $k\overline{h}_1$, $k < 1$, condition (13) must be violated, otherwise a ρ greater than ρ^* would be allowed for condition (12). Observing the special form of $\overline{h}_1$, we have

$$v^{\mathrm{T}} \frac{\partial Q(h_1)}{\partial h_{11}} v \Big|_{h_1 = \overline{h}_1} < 0, \tag{14}$$

which means that a decrease of h_{11} from 1 would increase the largest eigenvalue of $Q(h_1)$. This relation is obtained by using eigenvalue perturbation theory (*e.g.*, see [12]). Recalling that $P = I$, we can rewrite (14) as

$$v^{\mathrm{T}} \begin{bmatrix} 2b_{11} & b_{12} & \cdots & b_{1n} \\ b_{12} & 0 & \cdots & 0 \\ \vdots & \vdots & & \\ b_{1n} & 0 & \cdots & 0 \end{bmatrix} v = 2v_1 \sum_{i=1}^{n} b_{1i} v_i < 0. \tag{15}$$

Also, the optimality of ρ^* means that if we change other elements of h_1 over the sphere surface $\rho^* h_1 P^{-1} h_1^{\mathrm{T}} = h_1 h_1^{\mathrm{T}} = 1$, we will have $\lambda_{\max}(Q(h_1)) \geq 0$, otherwise a ρ greater than ρ^* would be allowed. All the h_1 in the surface $h_1 h_1^{\mathrm{T}} = 1$ and in a neighborhood of $\overline{h}_1$ can be expressed as

$$h_1 = [\sqrt{1 - (d_2^2 + d_3^2 + \cdots + d_n^2)} \ \ d_2 \ \ \cdots \ \ d_n], \ \ d_2^2 + d_3^2 + \cdots + d_n^2 < 1.$$

Since $\lambda_{\max}(Q(h_1))$ has a local minimum at $h_1 = \bar{h}_1$, by eigenvalue perturbation theory, we must have

$$v^{\mathrm{T}} \frac{\partial Q(h_1)}{\partial d_j} v \Big|_{h_1=\bar{h}_1} = 0, \quad j \in [2,n]. \tag{16}$$

With the special form of P, we can rewrite (16) as

$$v^{\mathrm{T}} \begin{bmatrix} 0 & \cdots & b_{11} & \cdots & 0 \\ \vdots & & \vdots & & \\ b_{11} & \cdots & 2b_{1j} & \cdots & b_{1n} \\ \vdots & & \vdots & & \\ 0 & \cdots & b_{1n} & \cdots & 0 \end{bmatrix} v = 2v_j \sum_{i=1}^{n} b_{1i} v_i = 2v_j \sum_{i=1}^{n} b_{1i} v_i = 0. \tag{17}$$

The relations (15) and (17) jointly show that $v_1 \neq 0$ and $v_j = 0, j \in [2,n]$, and hence v is aligned with x_0. Therefore,

$$Q(\bar{h}_1) x_0 = 0.$$

Step 2: $\dot{V}(x_0) = 0$.

From (15) and $v_j = 0, j \in [2,n]$, we have $b_{11} < 0$. From condition 3) of the theorem, $|f_k x_0| \leq 1$ for all $k \in [2,m]$. It follows that

$$\begin{aligned} \dot{V}(x_0) &= x_0^{\mathrm{T}} \left(A^{\mathrm{T}}P + PA + \sum_{k=2}^{m} (f_k^{\mathrm{T}} b_k^{\mathrm{T}} P + P b_k f_k) \right) x_0 + 2x_0^{\mathrm{T}} P b_1 \mathrm{sat}(f_1 x_0) \\ &= x_0^{\mathrm{T}} \left(A^{\mathrm{T}}P + PA + \sum_{k=2}^{m} (f_k^{\mathrm{T}} b_k^{\mathrm{T}} P + P b_k f_k) \right) x_0 + 2b_{11} \mathrm{sat}(f_1 x_0). \end{aligned}$$

Since $Q(\bar{h}_1) x_0 = 0$, we have

$$\begin{aligned} 0 &= x_0^{\mathrm{T}} Q(\bar{h}_1) x_0 \\ &= x_0^{\mathrm{T}} \left(A^{\mathrm{T}}P + PA + \sum_{k=2}^{m} (f_k^{\mathrm{T}} b_k^{\mathrm{T}} P + P b_k f_k) + \bar{h}_1^{\mathrm{T}} b_1^{\mathrm{T}} P + P b_1 \bar{h}_1 \right) x_0 \\ &= x_0^{\mathrm{T}} \left(A^{\mathrm{T}}P + PA + \sum_{k=2}^{m} (f_k^{\mathrm{T}} b_k^{\mathrm{T}} P + P b_k f_k) \right) x_0 + 2b_{11}, \end{aligned} \tag{18}$$

noting that $\bar{h}_1 x_0 = 1$.

On the other hand, from (5),

$$\begin{aligned} & x_0^{\mathrm{T}} ((A+BF)^{\mathrm{T}}P + P(A+BF)) x_0 \\ &= x_0^{\mathrm{T}} \left(A^{\mathrm{T}}P + PA + \sum_{k=2}^{m} (f_k^{\mathrm{T}} b_k^{\mathrm{T}} P + P b_k f_k) \right) x_0 + 2b_{11} f_1 x_0 \\ &< 0. \end{aligned} \tag{19}$$

By comparing (18) with (19), we know that

$$2b_{11}f_1x_0 < 2b_{11}.$$

Recalling that $b_{11} < 0$, we obtain $f_1x_0 > 1$. Thus $\mathrm{sat}(f_1x_0) = 1$ and

$$\dot{V}(x_0) = x_0^{\mathrm{T}}Q(\overline{h}_1)x_0 = 0.$$

Since $x_0 \in \mathcal{E}(P, \rho^*)$, this implies that $\rho_c \leq \rho^*$. Observing that $\rho_c \geq \rho^*$, we finally have $\rho_c = \rho^*$. □

Corollary 1 *If the system has only one input, i.e., $m = 1$, then $\rho_c = \rho^*$.*

Proof. In this case, (8b) has only one equality and for an optimal solution, we must have $\rho^* HP^{-1}H^{\mathrm{T}} = 1$. Hence condition 1) in Theorem 3 is satisfied. As to condition 2), there are two inequalities involved,

$$(A + BF)^{\mathrm{T}}P + P(A + BF) \leq 0,$$

and

$$(A + BH)^{\mathrm{T}}P + P(A + BH) \leq 0.$$

For the first one, we have the strict " $<$ " by assumption and for the second one, we must have

$$\lambda_{\max}\Big((A + BH)^{\mathrm{T}}P + P(A + BH)\Big) = 0.$$

Hence condition 2) is also satisfied. Since $m = 1$, condition 3) vanishes (or is satisfied automatically). □

For systems with multiple inputs, computational experience shows that conditions 1) and 2) of Theorem 3 are generally true. This can be explained as follows. It is easy to see that 1) is generally true. Assume that it is $\rho^*\overline{h}_1P^{-1}\overline{h}_1^{\mathrm{T}} = 1$. For condition 2), there is also generally only one D_i such that

$$\lambda_{\max}\Big((A + B(D_iF + D_i^-\overline{H})^{\mathrm{T}}P + P(A + B(D_iF + D_i^-\overline{H})\Big) = 0. \tag{20}$$

We would like to show that this D_i^- should be the matrix whose only nonzero element is at $(1,1)$. First, D_i^- must chooses $\overline{h}_1$. Otherwise, (8a) would be true for all h_1 in a neighborhood of $\overline{h}_1$, allowing a greater ρ^*. Suppose that D_i^- also chooses some other h_j, say h_2, then we would have the term

$$(b_2\overline{h}_2)^{\mathrm{T}}P + Pb_2\overline{h}_2$$

in the matrix in (20). Since $\rho^* \overline{h}_2 P^{-1} \overline{h}_2^{\mathrm{T}} < 1$, we can let h_2 vary in a neighborhood of $\overline{h}_2$ without violating other conditions except (20). Generally, there would be certain direction Δh_2 such that the additional term

$$(b_2 \Delta h_2)^{\mathrm{T}} P + P b_2 \Delta h_2$$

will cause

$$\lambda_{\max} \Big((A + B(D_i F + D_i^- H)^{\mathrm{T}} P + P(A + B(D_i F + D_i^- H)) \Big) < 0.$$

This would also allow a greater ρ^*.

However, condition 3) in Theorem 3 is not always satisfied. In that case, we may have $\rho_c > \rho^*$. This will be illustrated in an example.

4 An Example

Consider a two-input system with

$$A = \begin{bmatrix} 0.6 & -0.8 \\ 0.8 & 0.6 \end{bmatrix}.$$

The input matrix B is generated randomly with normal distribution. F is a feedback matrix such that $A + BF$ has eigenvalues $-1 \pm j0.6$, and P is the solution to

$$(A + BF)^{\mathrm{T}} P + P(A + BF) = -I.$$

Computational results show that conditions 1) and 2) are generally satisfied (97 out of 100). condition 3) is often satisfied but not always (88 out of 100) .

The following are two sets of parameters B generated randomly and the corresponding optimization results.

Case 1.

$$B = \begin{bmatrix} 0.8030 & 0.9455 \\ 0.0839 & 0.9159 \end{bmatrix}.$$

The pole assignment feedback matrix F and the P matrix are

$$F = \begin{bmatrix} -1.2031 & 1.7926 \\ -0.4441 & -2.1447 \end{bmatrix}, \quad P = \begin{bmatrix} 0.5366 & -0.2676 \\ -0.2676 & 0.7179 \end{bmatrix}.$$

The optimal solution to (8) is $\rho^* = 0.4050$ and

$$\overline{H} = \begin{bmatrix} -0.6633 & 1.1973 \\ -0.0359 & -1.1828 \end{bmatrix}.$$

We also have $x_0 = \begin{bmatrix} -0.4420 \\ -0.8320 \end{bmatrix}$. All the conditions in Theorem 3 are satisfied. This is illustrated in Fig. 1, where the four solid lines are $h_1 x = \pm 1$ and $h_2 x = \pm 1$, and the four dotted lines are

$f_1 x = \pm 1$ and $f_2 x = \pm 1$. The ellipsoid only intersects $h_2 x = \pm 1$. This shows that condition 1) is satisfied. condition 2) is verified by checking the eigenvalues of the matrices in (8a). We also see that x_0 is between the two lines $f_1 x = \pm 1$. This means that condition 3) is satisfied. According to Theorem 3, $\mathcal{E}(P, \rho^*)$ is the largest invariant ellipsoid. This is verified in Fig. 2, where $\dot{V}(x)$ along the boundary of the ellipsoid $\mathcal{E}(P, \rho^*)$ is plotted. We see that the maximal value reaches 0.

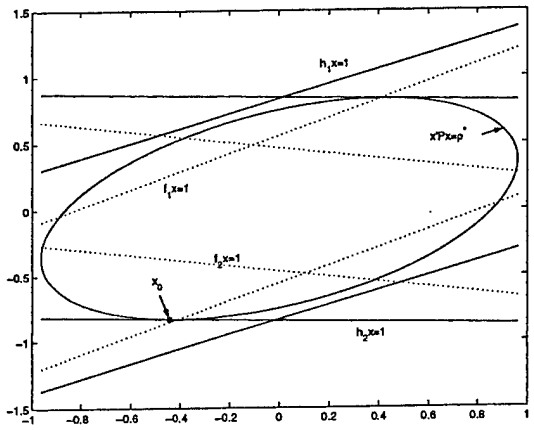


Figure 1: Illustration of the conditions for case 1.

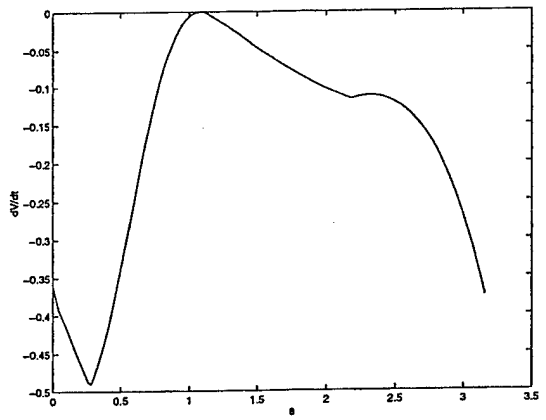


Figure 2: The derivative $\dot{V}(x)$ along $\partial\mathcal{E}(P, \rho)$: case 1.

Case 2. We have

$$B = \begin{bmatrix} 0.8828 & -0.1455 \\ 0.2842 & -0.0896 \end{bmatrix},$$

and

$$F = \begin{bmatrix} -2.6921 & -9.1511 \\ -0.9778 & -18.2487 \end{bmatrix}, \quad P = \begin{bmatrix} 0.2773 & -0.3815 \\ -0.3815 & 7.8606 \end{bmatrix}.$$

The optimal solution to (8) is $\rho^* = 0.0342$, and

$$\overline{H} = \begin{bmatrix} -1.2666 & -11.3745 \\ 0.0110 & 5.9181 \end{bmatrix}.$$

Here we have $x_0 = \begin{bmatrix} -0.2403 \\ -0.0612 \end{bmatrix}$. It is verified that conditions 1) and 2) are satisfied, but condition 3) is not, as we can see in Fig. 3, where $|f_i x_0| > 1, i = 1, 2$. In this case, it is likely that $\mathcal{E}(P, \rho^*)$ is not the largest invariant ellipsoid. As can be seen from Fig. 4, the maximal value of $\dot{V}(x)$ along the ellipsoid is strictly less than 0. This means the the largest invariant ellipsoid is strictly larger than $\mathcal{E}(P, \rho^*)$.

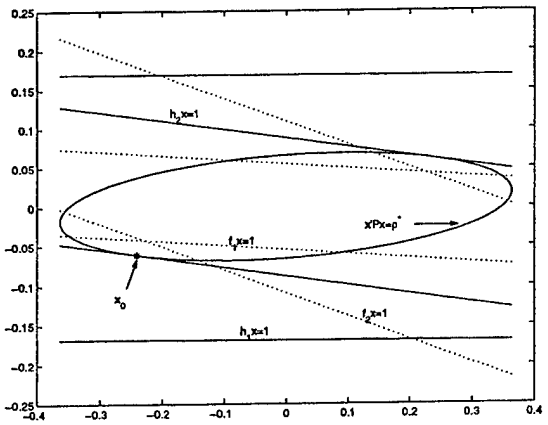


Figure 3: Illustration of the conditions for case 2.

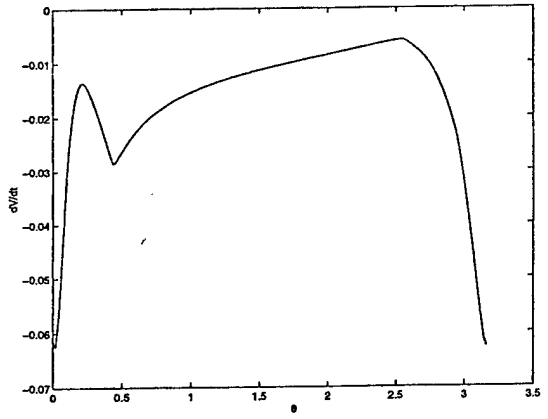


Figure 4: The derivative $\dot{V}(x)$ along $\partial\mathcal{E}(P, \rho^*)$: case 2.

5 Conclusions

We investigated the necessity of a recent condition for set invariance by studying the optimal solution of a related LMI problem. We developed criterion for checking if the largest invariant

ellipsoid has been obtained by solving the LMI problem. Examples show that the condition may not be necessary under certain circumstances.

References

[1] F. Blanchini, "Set invariance in control – a survey", *Automatica*, Vol. 35, No.11, pp. 1747-1767, 1999.

[2] S. Boyd, L. El Ghaoui, E. Feron and V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory*, SIAM Studies in Appl. Mathematics, Philadelphia, 1994.

[3] E. J. Davison and E. M. Kurak, "A computational method for determining quadratic Lyapunov functions for non-linear systems," *Automatica*, Vol. 7, pp. 627-636, 1971.

[4] E. G. Gilbert and K. T. Tan, "Linear systems with state and control constraints: the theory and application of maximal output admissible sets," *IEEE Trans. Automat. Contr.* Vol. 36, pp. 1008-1020, 1991.

[5] H. Hindi & S. Boyd, "Analysis of linear systems with saturating using convex optimization," *Proc of the 37th IEEE CDC*, pp903-908, Florida, 1998.

[6] T. Hu and Z. Lin. *Control Systems with Actuator Saturation: Analysis and Design*, Birkhäuser, Boston, 2001.

[7] T. Hu and Z. Lin, "Exact characterization of invariant ellipsoids for linear systems with saturating actuators," *IEEE Trans. Automat. Contr.*, scheduled to appear in February, 2002.

[8] T. Hu and Z. Lin, "On enlarging the basin of attraction for linear systems under saturated linear feedback," *Sys. & Contr. Lett.*, Vol. 40, No. 1, pp. 59-69, May, 2000.

[9] T. Hu, Z. Lin and B. M. Chen, "An analysis and design method for linear systems subject to actuator saturation and disturbance," *Automatica*, Vol. 38, No. 2, pp. 351-359, 2002.

[10] H. Khalil, *Nonlinear Systems*, Prentice-Hall, Upper Saddle River, New Jersey, 1996.

[11] K. A. Loparo and G. L. Blankenship, "Estimating the domain of attraction of of nonlinear feedback systems," *IEEE Trans. Automat. Contr.*, Vol. 23, No.4, pp. 602-607, 1978.

[12] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

[13] S. Weissenberger, "Application of results from the absolute stability to the computation of finite stability domains," *IEEE Trans. Automatic Control*, AC-13, pp124-125, 1968.

[14] G.F. Wredenhagen and P.R. Belanger, "Piecewise-linear LQ control for systems with input constraints," *Automatica*, Vol. 30, pp. 403-416, 1994.

Publication 26

SUBMITTED TO *IEEE Transactions on Automatic Control*

On Maximizing the Convergence Rate for Linear Systems with Input Saturation

Tingshu Hu† Zongli Lin† Yacov Shamash*

† Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903, U.S.A.
Email: th7f, zl5y@virginia.edu

*Department of Electrical Engineering, State University of New York, Stony Brook, NY 11794
yshamash@notes.sunysb.edu

Abstract

In this paper, we consider the problem of maximizing the convergence rate inside a given level set for both continuous-time and discrete-time systems with input saturation. We also provide simple methods for finding the largest ellipsoid of a given shape that can be made invariant with a saturated control. For the continuous-time case, the maximal convergence rate is achieved by a bang-bang type control with a simple switching scheme. Sub-optimal convergence rate can be achieved with saturated high-gain linear feedback. We also study the problem of maximizing the convergence rate in the presence of disturbances. For the discrete-time case, the maximal convergence rate is achieved by a coupled saturated linear feedback.

Keywords: Convergence rate, stability, invariant set, saturation.

[1]This work was supported in part by the US Office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

1 Introduction

Fast response is always a desired property for control systems. The time optimal control problem was formulated for this purpose (see, e.g., [3, 6, 8, 14, 15]. Although it is well known that the time optimal control is a bang-bang control, this control strategy is rarely implemented in real systems. The main reason is that it is generally impossible to characterize the switching surface. For discrete-time systems, online computation was proposed but the computation burden is very heavy since linear programming has to be solved recursively with increasing time horizon. Also, as the time horizon is extended, numerical problems become more severe. Another reason is that even if the optimal control can be obtained exactly and efficiently, it results in open-loop controls.

A notion directly related to fast response is the convergence rate of the state trajectories. To motivate our problem formulation, we consider the following linear system

$$\dot{x} = Ax, \quad x \in \mathbf{R}^n. \tag{1}$$

Assume that the system is asymptotically stable. For this system, the overall convergence rate is measured by the maximal real part of the eigenvalues of the A matrix. Let

$$\alpha = -\max\left\{\mathrm{Re}(\lambda_i(A)) : \ i = 1, 2, \cdots, n\right\},$$

where $\mathrm{Re}(\lambda_i(A))$ is the real part of the ith eigenvalue of A, then $\alpha > 0$. For simplicity, assume that $A + \alpha I$ is neutrally stable, *i.e.*, there exists a positive definite matrix P such that

$$A^{\mathrm{T}}P + PA \leq -2\alpha P.$$

Then, there exists a nonzero $x \in \mathbf{R}^n$ such that

$$x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x = -2\alpha x^{\mathrm{T}}Px. \tag{2}$$

Let $V(x) = x^{\mathrm{T}}Px$. Along the trajectory of the system (1),

$$\dot{V}(x) = x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x \leq -2\alpha x^{\mathrm{T}}Px = -2\alpha V(x).$$

Hence,

$$\frac{\dot{V}(x)}{V(x)} \leq -2\alpha, \quad \forall x \in \mathbf{R}^n \setminus \{0\}. \tag{3}$$

Furthermore, because of (2), we have

$$\alpha = \frac{1}{2}\min\left\{-\frac{\dot{V}(x)}{V(x)} : \ x \in \mathbf{R}^n \setminus \{0\}\right\}. \tag{4}$$

From (3), we obtain

$$V(x(t)) \leq e^{-2\alpha t}V(x_0).$$

Therefore, we call α the overall convergence rate of the system (1). For a general nonlinear system, the convergence rate can be defined similarly as in (4). Since local stability is a more

general property than global stability, we would like to define the term convergence rate on some subset of the domain of attraction of the origin. Consider a nonlinear system

$$\dot{x} = f(x).$$

Assume that the system is asymptotically stable. Given a Lyapunov function $V(x)$, let $L_V(\rho)$ be a level set

$$L_V(\rho) = \left\{ x \in \mathbf{R}^n : \ V(x) \leq \rho \right\}.$$

Suppose that $\dot{V}(x) < 0$ for all $x \in L_V(\rho) \setminus \{0\}$. Then, the overall convergence rate of the system on $L_V(\rho)$ can be defined as

$$\alpha := \frac{1}{2} \inf \left\{ -\frac{\dot{V}(x)}{V(x)} : \ x \in L_V(\rho) \setminus \{0\} \right\}. \tag{5}$$

For a discrete-time system

$$x(k+1) = f(x(k)),$$

we can define the overall convergence rate on $L_V(\rho)$ as

$$\alpha := \frac{1}{2} \inf \left\{ -\frac{\Delta V(x)}{V(x)} : \ x \in L_V(\rho) \setminus \{0\} \right\},$$

where $\Delta V(x)$ is the increment of $V(x)$ along the trajectory of the system. We can also define the convergence rate at each point x in the state space as

$$-\frac{\dot{V}(x)}{V(x)} \quad \left(\text{or} \ -\frac{\Delta V(x)}{V(x)} \right).$$

For linear systems subject to actuator saturation, efforts have been made to increase the convergence rate in various heuristic ways. For example, the Q matrix in LQR design can be increased piecewisely[18] or continuously[16, 17] as the state trajectory converges to the origin. The objective of this paper is, for a given $V(x)$, to find a feedback law constrained by the actuator saturation such that $-\dot{V}(x)$ is maximized at each x. It turns out that the optimal control law is a bang-bang type control with a simple switching scheme. Since the discontinuity might be undesirable, for example, causing chattering around the switching surface, we will also derive a continuous control law which results in a convergence rate that is arbitrarily close to the optimal one. The proposed continuous control law is a saturated high gain feedback. As the gain goes to infinity, the saturated high gain feedback approaches the optimal bang-bang control law.

For a discrete-time system, the control law that maximizes the convergence rate is a coupled saturated linear feedback. If the system has one or two inputs, the control law can be put into a simple formula. If the system has more than two inputs, the controller is more complicated. It is linear inside some polyhedron. Outside of this polyhedron, we need to solve a simple convex optimization problem.

A very important consequence of the maximal convergence control is that it produces the maximal invariant level set of a given shape (in the absence or in the presence of disturbances). It is easy to see that a level set can be made invariant if and only if the maximal $\dot{V}(x)$ on the boundary of the level set under the maximal convergence control is less than 0. As is pointed out in [1], set invariance is a very important notion and a powerful tool in studying the stability and other performances of systems. In [2, 4, 5, 9, 10, 11], invariant ellipsoids are used to estimate the domain of attraction and to study disturbance rejection capability of the closed-loop system. In this paper, we will first study the quadratic type Lyupnov functions and then extend the results to a general Lyapunov function. We will present a simple method for checking if an ellipsoid can be made invariant and for determining the largest ellipsoid that can be made invariant.

In [11], we presented methods for enlarging the invariant ellipsoid with respect to some shape reference set. As a result, the closed-loop system behaves linearly in the ellipsoid. In other words, the feedback control does not saturate in the ellipsoid. In this paper, the control that achieves maximal convergence rate and maximal invariant ellipsoid saturates almost everywhere in the ellipsoid (for continuous-time systems). This seemingly contradiction can be explained with the result in [12], where it was shown that for a single input system, the largest invariant ellipsoid is somehow independent of a particular stabilizing controller. Although both methods in [10, 11] and in this paper produce large invariant ellipsoids, the focuses of these papers are different. [10, 11] put the optimization problems into LMI framework and makes it very easy to choose the shape of the ellipsoid. This paper assumes that the shape of the ellipsoid is given, say, produced by the method of [10, 11] and tries to maximize the convergence rate and to find the maximal invariant set of the given shape.

This paper is organized as follows. Section 2 and Section 3 study the maximal convergence rate control problems for continuous-time and discrete-time systems, respectively. Sections 2.1 – 2.4 deal with quadratic Lyapunov functions and Section 2.5 extends the results to a general Lypunov function. In particular, Section 2.1 shows that the maximal convergence control is a bang-bang type control with a simple switching scheme and that it produces the maximal invariant ellipsoid of a given shape. A method for determining the largest ellipsoid that can be made invariant with a bounded control is also given in this section. Section 2.2 presents a saturated high gain feedback strategy to avoid the discontinuity of the bang-bang control. An example is included in Section 2.2 to illustrate the effectiveness of high gain feedback control. Section 2.3 reveals some properties and limitations about the overall convergence rate and provides methods to deal with these limitations. Section 2.4 shows that the maximal convergence control also achieves both the maximal and the minimal invariant ellipsoids in the presence of disturbances. A brief concluding remark is made in Section 4.

Throughout the paper, we will use standard notation. For a vector $u \in \mathbf{R}^m$, we use $|u|_\infty$ to denote the ∞-norm. We use $\mathrm{sat}(\cdot)$ to denote the standard saturation function $\mathrm{sat}(s) = \mathrm{sign}(s)\min\{1, |s|\}$. With a slight abuse of notation and for simplicity, for a vector $u \in \mathbf{R}^m$, we

also use the same $\text{sat}(u)$ to denote the vector saturation function, i.e.,

$$\text{sat}(u) = [\text{sat}(u_1)\ \ \text{sat}(u_2)\ \cdots\ \text{sat}(u_m)]^{\mathrm{T}}.$$

We use $\text{sign}(\cdot)$ to denote the sign function which takes value $+1$ or -1.

2 Continuous-time Systems

2.1 Maximal Convergence Rate Control and Maximal Invariant Ellipsoid

Consider a linear system subject to actuator saturation,

$$\dot{x} = Ax + Bu, \quad x \in \mathbf{R}^n,\ u \in \mathbf{R}^m,\ |u|_\infty \le 1. \tag{6}$$

Assume that the system is stabilizable and that B has full column rank. Let

$$B = \begin{bmatrix} b_1 & b_2 & \cdots & b_m \end{bmatrix}.$$

In [11, 10], we developed a design method for enlarging the domain of attraction under a saturated linear state feedback. The approach was to maximize the invariant ellipsoid with respect to some shape reference set. The optimized controller, however, results in very slow convergence rate of the closed-loop system. The objective of this paper is to design a controller that maximizes the convergence rate inside a given ellipsoid. Given a positive definite matrix P, let

$$V(x) = x^{\mathrm{T}}Px.$$

For a positive number ρ, the level set associated with $V(x)$ is the ellipsoid,

$$\mathcal{E}(P,\rho) = \left\{x \in \mathbf{R}^n :\ x^{\mathrm{T}}Px \le \rho\right\}.$$

Along the trajectory of the system (6),

$$\begin{aligned} \dot{V}(x,u) &= 2x^{\mathrm{T}}P(Ax+Bu) \\ &= x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x + 2\sum_{i=1}^{m} x^{\mathrm{T}}Pb_iu_i. \end{aligned}$$

Under the constraint that $|u|_\infty \le 1$, the control that maximizes the convergence rate, or minimizes $\dot{V}(x,u)$, is simply

$$u_i = -\text{sign}(b_i^{\mathrm{T}}Px), \quad i = 1,2,\cdots,m, \tag{7}$$

where $\text{sign} : \mathbf{R} \to \mathbf{R}$ is the sign function. Under this bang-bang control, we have

$$\dot{V}(x) = x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2\sum_{i=1}^{m} x^{\mathrm{T}}Pb_i\,\text{sign}(b_i^{\mathrm{T}}Px).$$

Now consider the closed-loop system

$$\dot{x} = Ax - \sum_{i=1}^{m} b_i\text{sign}(b_i^{\mathrm{T}}Px). \tag{8}$$

Because of the discontinuity of the sign function, equation (8) may have no solution for some $x(0)$ or have solution only in a finite time interval. For example, for the single input case $m = 1$, equation (8) will have no solution if $B^{\mathrm{T}}Px(0) = 0$ and $\dot{x}$ at each side of the switching plane $B^{\mathrm{T}}Px = 0$ points to the other side. We will use a continuous feedback law,

$$u_i = -\mathrm{sat}(kb_i^{\mathrm{T}}Px),$$

where sat : $\mathbf{R} \to \mathbf{R}$ is the standard saturation function, to approximate the bang-bang control (7) in the next subsection. In what follows, we use the bang-bang control law to investigate the possibility that an ellipsoid can be made invariant with a bounded control $|u|_\infty \leq 1$.

Recall that an ellipsoid $\mathcal{E}(P, \rho)$ is invariant for a system $\dot{x} = f(x)$ if all the trajectories starting from it will stay inside of it. It is contractive invariant if

$$\dot{V}(x) = 2x^{\mathrm{T}}Pf(x) < 0, \quad \forall x \in \mathcal{E}(P, \rho) \setminus \{0\}.$$

Since the bang-bang control (7) minimizes $\dot{V}(x, u)$ at each x, we have the following obvious fact.

Fact 1 *The following two statements are equivalent:*

a) The ellipsoid $\mathcal{E}(P, \rho)$ can be made contractive invariant for (6) with a bounded control $|u|_\infty \leq 1$;

b) The ellipsoid $\mathcal{E}(P, \rho)$ is contractive invariant for (8), i.e., the following condition is satisfied,

$$\dot{V}(x) = x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2\sum_{i=1}^{m} x^{\mathrm{T}}Pb_i\mathrm{sign}(b_i^{\mathrm{T}}Px) < 0, \quad \forall x \in \mathcal{E}(P, \rho) \setminus \{0\}. \tag{9}$$

It is clear from Fact 1 that the maximal convergence rate control produces the maximal invariant ellipsoid of a given shape. We will see in Section 2.2 that if (9) is satisfied, there also exists a continuous feedback law such that $\mathcal{E}(P, \rho)$ is contractive invariant. In this case, all the trajectories starting from $\mathcal{E}(P, \rho)$ will converge to the origin asymptotically.

For an arbitrary positive definite matrix P, there may exists no ρ such that $\mathcal{E}(P, \rho)$ can be made invariant. In what follows we give condition on P such that $\mathcal{E}(P, \rho)$ can be made invariant for some ρ and provide a method for finding the largest ρ.

Proposition 1 *For a given positive definite matrix P, the following three statements are equivalent:*

a) There exists a $\rho > 0$ such that (9) is satisfied;

b) There exists an $F \in \mathbf{R}^{m \times n}$ such that

$$(A + BF)^{\mathrm{T}}P + P(A + BF) < 0; \tag{10}$$

c) There exists a $k > 0$ such that

$$(A - kBB^{\mathrm{T}}P)^{\mathrm{T}}P + P(A - kBB^{\mathrm{T}}P) < 0. \tag{11}$$

Proof. b)→ a). If (10) is satisfied, then there exists a $\rho > 0$ such that

$$\mathcal{E}(P,\rho) \subset \left\{x \in \mathbf{R}^n : \ |Fx|_\infty \leq 1\right\}.$$

If $x_0 \in \mathcal{E}(P,\rho)$, then under the control $u = Fx$, $x(t)$ will stay in $\mathcal{E}(P,\rho)$ and we also have $|u|_\infty \leq 1$ for all $t \geq 0$. This means that $\mathcal{E}(P,\rho)$ can be made contractive invariant with a bounded control. Hence by the equivalence of the statements a) and b) in Fact 1, we have (9).

c) → b). It is obvious.

a) → c). Let us assume that

$$PB = \begin{bmatrix} 0 \\ R \end{bmatrix},$$

where R is an $m \times m$ nonsingular matrix. If not so, we can use a state transformation, $\bar{x} = Tx$, with T nonsingular such that

$$P \to \bar{P} = (T^{-1})^{\mathrm{T}}PT^{-1},$$

$$B \to \bar{B} = TB$$

and

$$PB \to \bar{P}\bar{B} = (T^{-1})^{\mathrm{T}}PB = \begin{bmatrix} 0 \\ R \end{bmatrix}.$$

Recall that we have assumed that B has full column rank. Also, let us accordingly partition x as $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $A^{\mathrm{T}}P + PA$ and P as

$$A^{\mathrm{T}}P + PA = \begin{bmatrix} Q_1 & Q_{12} \\ Q_{12}^{\mathrm{T}} & Q_2 \end{bmatrix}, \quad P = \begin{bmatrix} P_1 & P_{12} \\ P_{12}^{\mathrm{T}} & P_2 \end{bmatrix}.$$

For all

$$\begin{bmatrix} x_1 \\ 0 \end{bmatrix} \in \partial\mathcal{E}(P,\rho),$$

we have $x^{\mathrm{T}}PB = 0$. So, if a) is true, then (9) holds for some $\rho > 0$, which implies that

$$x_1^{\mathrm{T}}Q_1x_1 < 0,$$

for all x_1 such that $x_1^{\mathrm{T}}P_1x_1 = \rho$. It follows that $Q_1 < 0$. Hence there exists a $k > 0$ such that

$$(A - kBB^{\mathrm{T}}P)^{\mathrm{T}}P + P(A - kBB^{\mathrm{T}}P) = \begin{bmatrix} Q_1 & Q_{12} \\ Q_{12}^{\mathrm{T}} & Q_2 - kRR^{\mathrm{T}} \end{bmatrix} < 0.$$

This shows that c) is true. □

Suppose that we are given a shape of ellipsoid, characterized by $P_0 > 0$, and that the maximal convergence rate is desired with respect to $V(x) = x^{\mathrm{T}}P_0x$, but this shape cannot be

made invariant for any size, then we can take $\mathcal{E}(P_0, 1)$ as a shape reference set and find another invariant set $\mathcal{E}(P, \rho)$ such that $\alpha\mathcal{E}(P_0, 1) \subset \mathcal{E}(P, \rho)$ with α maximized by the method in [11, 10]. The shape of the resulting invariant ellipsoid $\mathcal{E}(P, \rho)$ will be the closest to that of $\mathcal{E}(P_0, 1)$.

Now assume that we have a $P > 0$ such that the conditions in Proposition 1 are satisfied. Given $\rho > 0$, we would like to determine if $\mathcal{E}(P, \rho)$ is contractive invariant for the closed-loop system (8). Let's start with the single input case. In this case, condition (9) simplifies to

$$\dot{V}(x) = x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2x^{\mathrm{T}}PB\mathrm{sign}(B^{\mathrm{T}}Px) < 0, \quad \forall x \in \mathcal{E}(P, \rho) \setminus \{0\}. \tag{12}$$

We claim that (12) is equivalent to

$$x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2x^{\mathrm{T}}PB\mathrm{sign}(B^{\mathrm{T}}Px) < 0, \quad \forall x \in \partial\mathcal{E}(P, \rho). \tag{13}$$

To see this, we consider kx for $k \in (0, 1]$ and $x \in \partial\mathcal{E}(P, \rho)$. Suppose that

$$x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2x^{\mathrm{T}}PB\mathrm{sign}(B^{\mathrm{T}}Px) < 0.$$

Since

$$-2x^{\mathrm{T}}PB\mathrm{sign}(B^{\mathrm{T}}Px) \leq 0,$$

we have

$$x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - \frac{2x^{\mathrm{T}}PB}{k}\mathrm{sign}(B^{\mathrm{T}}Px) < 0, \quad \forall k \in (0, 1].$$

Therefore,

$$\begin{aligned}
&(kx)^{\mathrm{T}}(A^{\mathrm{T}}P + PA)(kx) - 2(kx)^{\mathrm{T}}PB\mathrm{sign}(B^{\mathrm{T}}Pkx) \\
&\quad = k^2\left(x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - \frac{2x^{\mathrm{T}}PB}{k}\mathrm{sign}(B^{\mathrm{T}}Px)\right) \\
&\quad < 0,
\end{aligned}$$

for all $k \in (0, 1]$. This shows that condition (12) is equivalent to (13). Based on this equivalence property, we have the following necessary and sufficient condition for the contractive invariance of a given ellipsoid.

Theorem 1 *Assume that $m = 1$. Suppose that $\mathcal{E}(P, \rho)$ can be made contractive invariant for some $\rho > 0$. Let $\lambda_1, \lambda_2, \cdots, \lambda_{\mathcal{J}} > 0$ be real numbers such that*

$$\det\begin{bmatrix} \lambda_j P - A^{\mathrm{T}}P - PA & P \\ \rho^{-1}PBB^{\mathrm{T}}P & \lambda_j P - A^{\mathrm{T}}P - PA \end{bmatrix} = 0 \tag{14}$$

and

$$B^{\mathrm{T}}P(A^{\mathrm{T}}P + PA - \lambda_j P)^{-1}PB > 0. \tag{15}$$

Then, $\mathcal{E}(P, \rho)$ is contractive invariant for the system (8) if and only if

$$\lambda_j\rho - B^{\mathrm{T}}P(A^{\mathrm{T}}P + PA - \lambda_j P)^{-1}PB < 0, \quad \forall j = 1, 2, \cdots, \mathcal{J}.$$

If there exists no $\lambda_j > 0$ satisfying (14) and (15), then $\mathcal{E}(P, \rho)$ is contractive invariant.

In the proof of Theorem 1, we will use the following algebraic fact. Suppose that X_1, X_4 and $\begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \end{bmatrix}$ are square matrices. If X_1 is nonsingular, then

$$\det \begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \end{bmatrix} = \det(X_1) \det(X_4 - X_3 X_1^{-1} X_2), \tag{16}$$

and if X_4 is nonsingular, then

$$\det \begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \end{bmatrix} = \det(X_4) \det(X_1 - X_2 X_1^{-1} X_3). \tag{17}$$

Proof of Theorem 1. Denote

$$g(x) = x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2x^{\mathrm{T}}PB.$$

By the equivalence of (12) and (13), the contractive invariance of $\mathcal{E}(P, \rho)$ is equivalent to

$$\max \left\{ g(x) : \ B^{\mathrm{T}}Px \geq 0, \ x^{\mathrm{T}}Px = \rho \right\} < 0. \tag{18}$$

Since $\mathcal{E}(P, \rho)$ can be made contractive invariant for some $\rho > 0$, we must have $g(x) < 0$ for all $B^{\mathrm{T}}Px = 0$. In this case, the contractive invariance of $\mathcal{E}(P, \rho)$ is equivalent to that all the extrema of $g(x)$ in the surface $x^{\mathrm{T}}Px = \rho$, $B^{\mathrm{T}}Px > 0$, if any, are less than zero.

By the Lagrange multiplier method, an extremum of $g(x)$ in the surface $x^{\mathrm{T}}Px = \rho$, $B^{\mathrm{T}}Px > 0$, must satisfy

$$(A^{\mathrm{T}}P + PA - \lambda P)x = PB, \quad x^{\mathrm{T}}Px = \rho, \quad x^{\mathrm{T}}PB > 0, \tag{19}$$

for some real number λ. And at the extremum, we have

$$g(x) = \lambda\rho - x^{\mathrm{T}}PB.$$

If $\lambda \leq 0$, then $g(x) < 0$ since $x^{\mathrm{T}}PB > 0$. So we only need to consider $\lambda > 0$.

Now suppose that $\lambda > 0$. From

$$(A^{\mathrm{T}}P + PA - \lambda P)x = PB,$$

we conclude that

$$\det(A^{\mathrm{T}}P + PA - \lambda P) \neq 0.$$

To show this, we assume, without loss of generality, that

$$A^{\mathrm{T}}P + PA = \begin{bmatrix} Q_1 & Q_{12} \\ Q_{12}^{\mathrm{T}} & q_2 \end{bmatrix}, \quad P = \begin{bmatrix} P_1 & P_{12} \\ P_{12}^{\mathrm{T}} & p_2 \end{bmatrix}, \quad PB = \begin{bmatrix} 0 \\ r \end{bmatrix},$$

as in the proof of Proposition 1, it follows then that $Q_1 < 0$. Since $\lambda > 0$, $Q_1 < 0$ and $P_1 > 0$, $Q_1 - \lambda P_1$ is nonsingular. Let

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad x_2 \in \mathbf{R}$$

and suppose that $x \neq 0$ satisfies

$$(A^{\mathrm{T}}P + PA - \lambda P)x = PB,$$

then,

$$x_1 = -(Q_1 - \lambda P_1)^{-1}(Q_{12} - \lambda P_{12})x_2,$$

and

$$\Big(- (Q_{12}^{\mathrm{T}} - \lambda P_{12}^{\mathrm{T}})(Q_1 - \lambda P_1)^{-1}(Q_{12} - \lambda P_{12}) + q_2 - \lambda p_2 \Big) x_2 = r.$$

Multiplying both sides with $\det(Q_1 - \lambda P_1)$ and applying (16), we obtain

$$\det(A^{\mathrm{T}}P + PA - \lambda P)x_2 = \det(Q_1 - \lambda P_1)r.$$

Since $r \neq 0$ and $\det(Q_1 - \lambda P_1) \neq 0$, we must have

$$\det(A^{\mathrm{T}}P + PA - \lambda P) \neq 0.$$

So for all $\lambda > 0$ and x satisfying (19), we have

$$x = (A^{\mathrm{T}}P + PA - \lambda P)^{-1}PB,$$

and hence from $x^{\mathrm{T}}Px = \rho$,

$$B^{\mathrm{T}}P(A^{\mathrm{T}}P + PA - \lambda P)^{-1}P(A^{\mathrm{T}}P + PA - \lambda P)^{-1}PB = \rho. \tag{20}$$

Denote

$$\Phi = \lambda P - A^{\mathrm{T}}P - PA,$$

then the equation (20) can be written as,

$$B^{\mathrm{T}}P\Phi^{-1}P\Phi^{-1}PB = \rho.$$

By invoking (16) and (17), we obtain

$$\det \begin{bmatrix} \rho & -B^{\mathrm{T}}P\Phi^{-1} \\ -\Phi^{-1}PB & P^{-1} \end{bmatrix} = 0$$

$$\Updownarrow$$

$$\det \left(\begin{bmatrix} \rho & 0 \\ 0 & P^{-1} \end{bmatrix} - \begin{bmatrix} B^{\mathrm{T}}P & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \Phi^{-1} & 0 \\ 0 & \Phi^{-1} \end{bmatrix} \begin{bmatrix} 0 & I \\ PB & 0 \end{bmatrix} \right) = 0$$

$$\Updownarrow$$

$$\det \left(\begin{bmatrix} \Phi & 0 \\ 0 & \Phi \end{bmatrix} - \begin{bmatrix} 0 & I \\ PB & 0 \end{bmatrix} \begin{bmatrix} \rho^{-1} & 0 \\ 0 & P \end{bmatrix} \begin{bmatrix} B^{\mathrm{T}}P & 0 \\ 0 & I \end{bmatrix} \right) = 0$$

$$\Updownarrow$$

$$\det \begin{bmatrix} \lambda P - A^{\mathrm{T}}P - PA & P \\ \rho^{-1}PBB^{\mathrm{T}}P & \lambda P - A^{\mathrm{T}}P - PA \end{bmatrix} = 0.$$

This last equation is (14).

Also, at the extremum, we have $x^{\mathrm{T}}PB > 0$. This is equivalent to (15),

$$B^{\mathrm{T}}P(A^{\mathrm{T}}P + PA - \lambda P)^{-1}PB > 0.$$

Finally, at the extremum

$$\begin{aligned} g(x) &= x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2x^{\mathrm{T}}PB \\ &= \lambda\rho - B^{\mathrm{T}}P(A^{\mathrm{T}}P + PA - \lambda P)^{-1}PB. \end{aligned}$$

Hence the result of the theorem follows. □

Here we note that all the λ_j's satisfying (14) are the eigenvalues of the matrix

$$\begin{bmatrix} P^{-\frac{1}{2}}A^{\mathrm{T}}P^{\frac{1}{2}} + P^{\frac{1}{2}}AP^{-\frac{1}{2}} & -I \\ -\rho^{-1}P^{\frac{1}{2}}BB^{\mathrm{T}}P^{\frac{1}{2}} & P^{-\frac{1}{2}}A^{\mathrm{T}}P^{\frac{1}{2}} + P^{\frac{1}{2}}AP^{-\frac{1}{2}} \end{bmatrix}.$$

Hence the condition of Theorem 1 can be easily checked.

Recall that condition (12) is equivalent to (13). This implies that there is a $\rho^* > 0$ such that $\mathcal{E}(P,\rho)$ is contractive invariant if and only if $\rho < \rho^*$. Therefore, the maximum value ρ^* can be obtained by checking the condition of Theorem 1 bisectionally.

For systems with multiple inputs, we may divide the state space into cones: $x^{\mathrm{T}}Pb_i < 0$ (> 0, or $= 0$), and check the maximum value of

$$\dot{V}(x) = x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2\sum_{i=1}^{m} x^{\mathrm{T}}Pb_i \mathrm{sign}(b_i^{\mathrm{T}}Px)$$

within each cone. For example, consider $m = 2$, the surface of the ellipsoid $\mathcal{E}(P,\rho)$ can be divided into the following subsets:

$$\begin{aligned} S_1 &= \left\{x \in \mathbf{R}^n : b_1^{\mathrm{T}}Px = 0,\ b_2^{\mathrm{T}}Px \geq 0,\ x^{\mathrm{T}}Px = \rho\right\}, \quad -S_1, \\ S_2 &= \left\{x \in \mathbf{R}^n : b_1^{\mathrm{T}}Px \geq 0,\ b_2^{\mathrm{T}}Px = 0,\ x^{\mathrm{T}}Px = \rho\right\}, \quad -S_2, \\ S_3 &= \left\{x \in \mathbf{R}^n : b_1^{\mathrm{T}}Px > 0,\ b_2^{\mathrm{T}}Px > 0,\ x^{\mathrm{T}}Px = \rho\right\}, \quad -S_3, \\ S_4 &= \left\{x \in \mathbf{R}^n : b_1^{\mathrm{T}}Px > 0,\ b_2^{\mathrm{T}}Px < 0,\ x^{\mathrm{T}}Px = \rho\right\}, \quad -S_4. \end{aligned}$$

With this partition, $\mathcal{E}(P,\rho)$ is contractive invariant if and only if

$$\max_{x \in S_1} \dot{V}(x) < 0, \quad \max_{x \in S_2} \dot{V}(x) < 0, \tag{21}$$

and all the local extrema of $\dot{V}(x)$ in S_3 and S_4 are negative.

In S_1,

$$\dot{V}(x) = x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2x^{\mathrm{T}}Pb_2.$$

Let $N \in \mathbf{R}^{n\times(n-1)}$ be a matrix of rank $n-1$ such that $b_1^{\mathrm{T}}PN = 0$, *i.e.*, $\{Ny : y \in \mathbf{R}^{n-1}\}$ is the kernel of $b_1^{\mathrm{T}}P$. The constraint $b_1^{\mathrm{T}}Px = 0$ can be replaced by $x = Ny$, $y \in \mathbf{R}^{n-1}$. Thus,

$$\max_{x\in S_1}\dot{V}(x) = \max\Big\{y^{\mathrm{T}}N^{\mathrm{T}}(A^{\mathrm{T}}P+PA)Ny - 2y^{\mathrm{T}}N^{\mathrm{T}}Pb_2 : b_2^{\mathrm{T}}PNy \geq 0,\ y^{\mathrm{T}}N^{\mathrm{T}}PNy = \rho\Big\}.$$

This is similar to the optimization problem (18) in the proof of Theorem 1 except with a reduced order. The second optimization problem in (21) can be handled in the same way.

In S_3,

$$\dot{V}(x) = x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x - 2x^{\mathrm{T}}P(b_1+b_2).$$

All the local extrema of $\dot{V}(x)$ in S_3 (and in S_4) can be obtained like those of $g(x)$ in the proof of Theorem 1.

2.2 Saturated High Gain Feedback

With the bang-bang type control law (7), $V(x)$ of the closed-loop system will decrease with a maximal convergence rate. As we have noted in Section 2.1, the discontinuity of the bang-bang control law may cause the state equation to have no solution. When the control law is implemented on a sampled-data system, it may cause frequent switching (chattering) around the switching surface. In this section, we will replace the bang-bang control law with a saturated high gain linear feedback at the cost of a slight reduction in convergence rate. We also start with the single input case.

Theorem 2 *Assume that $m = 1$. Suppose that $\mathcal{E}(P,\rho)$ can be made contractive invariant with a bounded control. If*

$$(A - k_0BB^{\mathrm{T}}P)^{\mathrm{T}}P + P(A - k_0BB^{\mathrm{T}}P) < 0, \tag{22}$$

then $\mathcal{E}(P,\rho)$ is contractive invariant under the saturated linear control

$$u = -\mathrm{sat}(kB^{\mathrm{T}}Px),$$

for all $k \geq k_0$.

This theorem says that the saturated linear control $u = -\mathrm{sat}(kB^{\mathrm{T}}Px)$ can produce the same invariant set as the bang-bang control law as long as (22) is satisfied and $k > k_0$. As has been pointed out in Proposition 1, there always exists a $k > 0$ satisfying (22) under the condition that $\mathcal{E}(P,\rho)$ can be made contractive invariant. So the invariance of $\mathcal{E}(P,\rho)$ does not require high gain feedback. But high gain is necessary for achieving fast convergence rate.

Proof of Theorem 2. Suppose that $\mathcal{E}(P,\rho)$ can be made contractive invariant, then by the equivalence of statements a) and b) in Fact 1,

$$x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x - 2x^{\mathrm{T}}PB\mathrm{sign}(B^{\mathrm{T}}Px) < 0, \quad \forall\, x \in \mathcal{E}(P,\rho)\setminus\{0\}. \tag{23}$$

If $|B^{\mathrm{T}}Px| \geq \frac{1}{k}$, then

$$\mathrm{sat}(kB^{\mathrm{T}}Px) = \mathrm{sign}(B^{\mathrm{T}}Px).$$

It follows from (23) that

$$x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2x^{\mathrm{T}}PB\mathrm{sat}(kB^{\mathrm{T}}Px) < 0 \tag{24}$$

for $x \in \mathcal{E}(P,\rho) \setminus \{0\}$ such that $|B^{\mathrm{T}}Px| \geq \frac{1}{k}$.

If $|B^{\mathrm{T}}Px| < \frac{1}{k}$, then

$$\mathrm{sat}(kB^{\mathrm{T}}Px) = kB^{\mathrm{T}}Px.$$

Since $k \geq k_0$, it follows from (22) that

$$(A - kBB^{\mathrm{T}}P)^{\mathrm{T}}P + P(A - kBB^{\mathrm{T}}P) < 0.$$

Hence we also obtain (24). This shows that $\mathcal{E}(P,\rho)$ is contractive invariant under the control

$$u = -\mathrm{sat}(kB^{\mathrm{T}}Px).$$

□

Let us now compare the convergence rates under the two controls

$$u = -\mathrm{sat}(kB^{\mathrm{T}}Px)$$

and

$$u = -\mathrm{sign}(B^{\mathrm{T}}Px).$$

The difference between the two $\dot{V}(x)$'s under these two controls is

$$\left|2x^{\mathrm{T}}PB\left(\mathrm{sign}(B^{\mathrm{T}}Px) - \mathrm{sat}(kB^{\mathrm{T}}Px)\right)\right| \begin{cases} = 0, & \text{if } |B^{\mathrm{T}}Px| > \dfrac{1}{k}, \\ \leq \dfrac{2}{k}, & \text{if } |B^{\mathrm{T}}Px| \leq \dfrac{1}{k}. \end{cases}$$

Note that $|\mathrm{sign}(B^{\mathrm{T}}Px) - \mathrm{sat}(kB^{\mathrm{T}}Px)| \leq 1$. By letting $k \to \infty$, the difference between the two $\dot{V}(x)$'s will go to zero uniformly on $\mathcal{E}(P,\rho)$. Thus we can say that saturated high gain linear feedback will produce sub-optimal convergence rate.

We now consider the multiple input case. For $v \in \mathbf{R}^m$, denote

$$\mathrm{sat}(v) = \left[\begin{array}{cccc} \mathrm{sat}(v_1) & \mathrm{sat}(v_2) & \cdots & \mathrm{sat}(v_m) \end{array}\right]^{\mathrm{T}}.$$

Here, as usual, we have slightly abused the notation by using sat to denote both the scalar and the vector saturation functions. We have the following result.

Theorem 3 *For a multiple input system, suppose that $\mathcal{E}(P,\rho)$ can be made contractive invariant with a bounded control, then there exists a $k > 0$ such that $\mathcal{E}(P,\rho)$ is contractive invariant under the control*

$$u = -\mathrm{sat}(kB^{\mathrm{T}}Px).$$

Proof. The condition that $\mathcal{E}(P,\rho)$ can be made contractive invariant implies that

$$x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x-2\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\mathrm{sign}(b_i^{\mathrm{T}}Px)<0,\quad\forall\,x\in\mathcal{E}(P,\rho)\setminus\{0\}. \tag{25}$$

We need to show that there exists a $k>0$ such that

$$x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x-2\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\mathrm{sat}(kb_i^{\mathrm{T}}Px)<0,\quad\forall\,x\in\mathcal{E}(P,\rho)\setminus\{0\}. \tag{26}$$

First, we will show that (26) is equivalent to

$$x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x-2\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\mathrm{sat}(kb_i^{\mathrm{T}}Px)<0,\quad\forall\,x\in\partial\mathcal{E}(P,\rho). \tag{27}$$

If this is true, then we only need to consider $x\in\partial\mathcal{E}(P,\rho)$. Obviously, (27) is implied by (26). We need to prove the converse.

Suppose that $x\in\partial\mathcal{E}(P,\rho)$ satisfies (27). For $\gamma>0$, define

$$g(\gamma)=x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x-2\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\frac{\mathrm{sat}(kb_i^{\mathrm{T}}P\gamma x)}{\gamma},$$

then $g(1)<0$. Since

$$x^{\mathrm{T}}Pb_i\,\mathrm{sat}(kb_i^{\mathrm{T}}Px)\geq 0$$

and $\left|\frac{\mathrm{sat}(kb_i^{\mathrm{T}}P\gamma x)}{\gamma}\right|$ decreases as γ increases, it follows that $g(\gamma)$ increases as γ increases. Hence,

$$g(\gamma)\leq g(1)<0,\quad\forall\gamma\in(0,1].$$

Therefore, for all $\gamma\in(0,1]$,

$$(\gamma x)^{\mathrm{T}}(A^{\mathrm{T}}P+PA)(\gamma x)-2\sum_{i=1}^{m}(\gamma x)^{\mathrm{T}}Pb_i\mathrm{sat}(kb_i^{\mathrm{T}}P\gamma x)=\gamma^2g(\gamma)<0.$$

This shows that (26) is implied by (27).

Let

$$\varepsilon=-\max_{x\in\partial\mathcal{E}(P,\rho)}\left(x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x-2\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\mathrm{sign}(b_i^{\mathrm{T}}Px)\right).$$

Then, from (25), we have $\varepsilon>0$. For all $x\in\partial\mathcal{E}(P,\rho)$,

$$\begin{aligned}
&x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x-2\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\,\mathrm{sat}(kb_i^{\mathrm{T}}Px)\\
&=x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x-2\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\mathrm{sign}(b_i^{\mathrm{T}}Px)+2\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\left(\mathrm{sign}(b_i^{\mathrm{T}}Px)-\mathrm{sat}(kb_i^{\mathrm{T}}Px)\right)\\
&\leq-\varepsilon+2\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\left(\mathrm{sign}(b_i^{\mathrm{T}}Px)-\mathrm{sat}(kb_i^{\mathrm{T}}Px)\right).
\end{aligned}\tag{28}$$

Since

$$\left|x^{\mathrm{T}}Pb_i\left(\mathrm{sign}(b_i^{\mathrm{T}}Px)-\mathrm{sat}(kb_i^{\mathrm{T}}Px)\right)\right|\begin{cases}=0, & \text{if } |b_i^{\mathrm{T}}Px|>\dfrac{1}{k},\\ \le\dfrac{1}{k}, & \text{if } |b_i^{\mathrm{T}}Px|\le\dfrac{1}{k},\end{cases}$$

we have

$$\left|2\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\left(\mathrm{sign}(b_i^{\mathrm{T}}Px)-\mathrm{sat}(kb_i^{\mathrm{T}}Px)\right)\right|\le\frac{2m}{k}.$$

If we choose $k>\frac{2m}{\varepsilon}$, then from (28),

$$x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x-2\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\mathrm{sat}(kb_i^{\mathrm{T}}Px)<0,\quad\forall x\in\partial\mathcal{E}(P,\rho),$$

which is (27). It follows from the equivalence of (26) and (27) that $\mathcal{E}(P,\rho)$ is contractive invariant under the control

$$u=-\mathrm{sat}(kB^{\mathrm{T}}Px).$$

□

Example 1 Consider the system (6) with

$$A=\begin{bmatrix}0.6 & -0.8\\ 0.8 & 0.6\end{bmatrix},\quad B=\begin{bmatrix}2\\ 4\end{bmatrix}.$$

Let

$$P=\begin{bmatrix}0.0836 & -0.0639\\ -0.0639 & 0.1460\end{bmatrix}.$$

By checking the condition of Theorem 1 bisectionally, the largest ellipsoid that can be made contractive invariant with a bounded control is $\mathcal{E}(P,\rho^*)$ with $\rho^*=1.059$. By using the design method in [11], a feedback

$$u=\mathrm{sat}(F_0x),\quad F_0=\begin{bmatrix}0.0036 & -0.3057\end{bmatrix}$$

is found such that

$$(A+BF_0)^{\mathrm{T}}P+P(A+BF_0)<0,$$

with $\mathcal{E}(P,\rho)$, $\rho=1.058$ inside the linear region of the saturation

$$\mathcal{L}(F_0)=\left\{x\in R^2:\ |F_0x|_\infty\le 1\right\}$$

(see Fig. 1). The eigenvalues of $A+BF_0$ are $-0.0078\pm j0.8782$. It can be expected that the convergence rate under this feedback is very slow. The convergence rate is accelerated by using a saturated high gain feedback

$$u=-\mathrm{sat}(5B^{\mathrm{T}}Px).$$

This is illustrated in Figs. 1 and 2. In Fig. 1, "$*$" represents the initial state, the solid trajectory is under the control of $u=\mathrm{sat}(F_0x)$ and the dash-dotted one is under the control of $u=-\mathrm{sat}(5B^{\mathrm{T}}Px)$. Fig. 2 shows $V(x)=x^{\mathrm{T}}Px$ as a function of time. Also, the solid one is under the control of $u=\mathrm{sat}(F_0x)$ and the dash-dotted one is under the control of $u=-\mathrm{sat}(5B^{\mathrm{T}}Px)$.

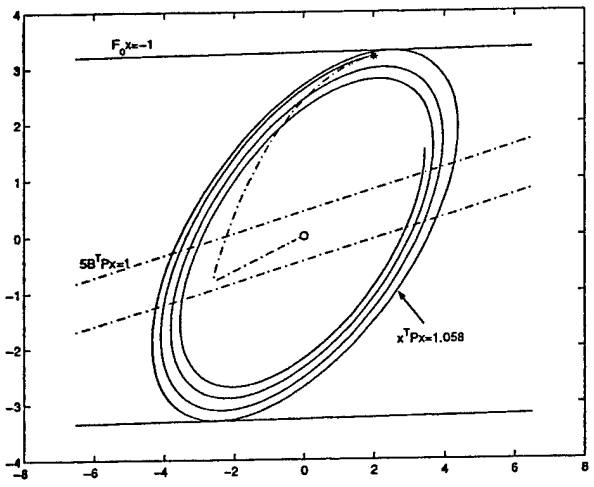


Figure 1: Comparison of the trajectory convergence rates.

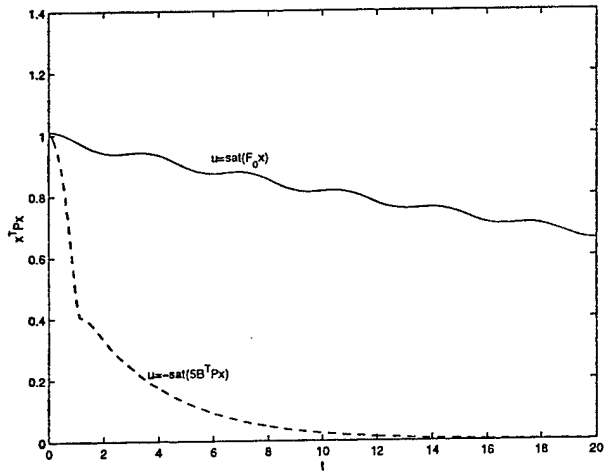


Figure 2: Comparison of the convergence rates of $x^{\mathrm{T}}Px$.

From Fig. 2, we see that the decrease of $x^{\mathrm{T}}Px$ becomes slower and slower under the control

$$u = -\mathrm{sat}(5B^{\mathrm{T}}Px).$$

This will be the case even if we increase the gain k in

$$u = -\mathrm{sat}(kB^{\mathrm{T}}Px)$$

to infinity. Actually, the overall convergence rate under the bang-bang control

$$u_i = -\mathrm{sign}(b_i^{\mathrm{T}}Px), \quad i = 1, 2, \cdots, m,$$

is limited by the shape of the ellipsoid or the matrix P. This problem will be discussed in the next section.

2.3 Overall Convergence Rate

We now consider the system (8) under the maximal convergence control,

$$\dot{x} = Ax - \sum_{i=1}^{m} b_i \text{sign}(b_i^{\mathrm{T}} P x). \tag{29}$$

Assume that $\mathcal{E}(P,\rho)$ is contractive invariant for (29). We would like to know the overall convergence rate in $\mathcal{E}(P,\rho)$. We will see later that as ρ decreases (note that a trajectory goes into smaller $\mathcal{E}(P,\rho)$ as time goes by), the overall convergence rate increases but is limited by the shape of $\mathcal{E}(P,\rho)$. This limit can be raised by choosing P properly.

The overall convergence rate, denoted by α, is defined by (5) in Section 1. Here we would like to examine its dependence on ρ, so we write,

$$\alpha(\rho) := \frac{1}{2} \min \left\{ -\frac{\dot{V}(x)}{V(x)} : \; x \in \mathcal{E}(P,\rho) \setminus \{0\} \right\}.$$

The main results of this subsection are contained in the following theorem.

Theorem 4

a)

$$\alpha(\rho) = \frac{1}{2} \min \left\{ -\frac{\dot{V}(x)}{\rho} : \; x^{\mathrm{T}} P x = \rho \right\}; \tag{30}$$

b) $\alpha(\rho)$ increases as ρ decreases;

c) Let

$$\beta_0 = \min \left\{ -x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x : \; x^{\mathrm{T}} P x = 1, \; x^{\mathrm{T}} P B = 0 \right\},$$

then,

$$\lim_{\rho \to 0} \alpha(\rho) = \frac{\beta_0}{2}. \tag{31}$$

Proof.

a) Consider $x \in \partial\mathcal{E}(P,\rho)$ and $k \in (0,1]$,

$$\begin{aligned} -\frac{\dot{V}(kx)}{V(kx)} &= -\frac{k^2 x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2k \sum_{i=1}^{m} x^{\mathrm{T}} P b_i \text{sign}(b_i^{\mathrm{T}} P k x)}{k^2 x^{\mathrm{T}} P x} \\ &= \frac{-x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x + \frac{2}{k} \sum_{i=1}^{m} x^{\mathrm{T}} P b_i \text{sign}(b_i^{\mathrm{T}} P x)}{x^{\mathrm{T}} P x}. \end{aligned} \tag{32}$$

Since

$$\sum_{i=1}^{m} x^{\mathrm{T}} P b_i \text{sign}(b_i^{\mathrm{T}} P x) \geq 0,$$

$-\frac{\dot{V}(kx)}{V(kx)}$ increases as k decreases. It follows that the minimal value of $-\frac{\dot{V}(x)}{V(x)}$ is obtained on the boundary of $\mathcal{E}(P,\rho)$, which implies (30).

b) This follows from the proof of a).

c) From a), we see that

$$\begin{aligned}
2\alpha(\rho) &= \min\left\{\frac{-x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x+2\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\,\mathrm{sign}(b_i^{\mathrm{T}}Px)}{\rho}:\ x^{\mathrm{T}}Px=\rho\right\}\\
&= \min\left\{-x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x+\frac{2}{\sqrt{\rho}}\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\,\mathrm{sign}(b_i^{\mathrm{T}}Px):\ x^{\mathrm{T}}Px=1\right\}\\
&\le \min\left\{-x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x+\frac{2}{\sqrt{\rho}}\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\,\mathrm{sign}(b_i^{\mathrm{T}}Px):\ x^{\mathrm{T}}Px=1,\ x^{\mathrm{T}}PB=0\right\}\\
&= \min\left\{-x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x:\ x^{\mathrm{T}}Px=1,\ x^{\mathrm{T}}PB=0\right\}\\
&= \beta_0.
\end{aligned}$$

It follows then $2\alpha(\rho)\le\beta_0$ for all $\rho>0$. To prove (31), it suffices to show that given any $\varepsilon>0$, there exists a $\rho>0$ such that

$$2\alpha(\rho)\ge\beta_0-\varepsilon.$$

Denote

$$\mathcal{X}_0=\left\{x\in\mathbf{R}^n:\ x^{\mathrm{T}}Px=1,\ x^{\mathrm{T}}PB=0\right\},$$

and

$$\mathcal{X}(\delta)=\left\{x\in\mathbf{R}^n:\ x^{\mathrm{T}}Px=1,\ |x^{\mathrm{T}}PB|_\infty\le\delta\right\}.$$

It is clear that

$$\lim_{\delta\to0}\mathrm{dist}(\mathcal{X}(\delta),\mathcal{X}_0)=0,$$

where $\mathrm{dist}(\cdot,\cdot)$ is the Hausdorff distance[1]. By the uniform continuity of $x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x$ on the surface $\{x\in\mathbf{R}^n:\ x^{\mathrm{T}}Px=1\}$, we have that, given any ε, there exists a $\delta>0$ such that

$$\min\left\{-x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x:\ x^{\mathrm{T}}Px=1,\ |x^{\mathrm{T}}PB|_\infty\le\delta\right\}\ge\beta_0-\varepsilon. \tag{33}$$

Since

$$\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\,\mathrm{sign}(b_i^{\mathrm{T}}Px)\ge0,$$

we have,

$$\min\left\{-x^{\mathrm{T}}(A^{\mathrm{T}}P+PA)x+\frac{2}{\sqrt{\rho}}\sum_{i=1}^{m}x^{\mathrm{T}}Pb_i\,\mathrm{sign}(b_i^{\mathrm{T}}Px):x^{\mathrm{T}}Px=1,\ |x^{\mathrm{T}}PB|_\infty\le\delta\right\}\ge\beta_0-\varepsilon, \tag{34}$$

for all $\rho>0$.

[1]Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two bounded subsets of $\mathbf{R}^n$. Their Hausdorff distance is defined as

$$\mathrm{dist}(\mathcal{X}_1,\mathcal{X}_2):=\max\left\{\vec{d}(\mathcal{X}_1,\mathcal{X}_2),\vec{d}(\mathcal{X}_2,\mathcal{X}_1)\right\},$$

where

$$\vec{d}(\mathcal{X}_1,\mathcal{X}_2)=\sup_{x_1\in\mathcal{X}_1}\inf_{x_2\in\mathcal{X}_2}|x_1-x_2|.$$

Here the vector norm used is arbitrary.

Let

$$\beta_1 = \min\Big\{ -x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x : \ x^{\mathrm{T}}Px = 1, \ |x^{\mathrm{T}}PB|_\infty \geq \delta \Big\}.$$

If $\beta_1 \geq \beta_0 - \varepsilon$, then for all $\rho > 0$,

$$\begin{aligned} &\min\left\{ -x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x + \frac{2}{\sqrt{\rho}} \sum_{i=1}^{m} x^{\mathrm{T}}Pb_i \operatorname{sign}(b_i^{\mathrm{T}}Px) : x^{\mathrm{T}}Px = 1, \ |x^{\mathrm{T}}PB|_\infty \geq \delta \right\} \\ &\quad \geq \beta_1 \geq \beta_0 - \varepsilon. \end{aligned}$$

Combining the above with (34), we have

$$\begin{aligned} 2\alpha(\rho) &= \min\left\{ -x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x + \frac{2}{\sqrt{\rho}} \sum_{i=1}^{m} x^{\mathrm{T}}Pb_i \operatorname{sign}(b_i^{\mathrm{T}}Px) : x^{\mathrm{T}}Px = 1 \right\} \\ &\quad \geq \beta_0 - \varepsilon. \end{aligned} \tag{35}$$

for all $\rho > 0$. This shows that

$$2\alpha(\rho) \geq \beta_0 - \varepsilon, \quad \forall \rho > 0.$$

If $\beta_1 < \beta_0 - \varepsilon$, then for

$$\rho < \left(\frac{2\delta}{-\beta_1 + \beta_0 - \varepsilon} \right)^2,$$

we have

$$\begin{aligned} &\min\left\{ -x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x + \frac{2}{\sqrt{\rho}} \sum_{i=1}^{m} x^{\mathrm{T}}Pb_i \operatorname{sign}(b_i^{\mathrm{T}}Px) : x^{\mathrm{T}}Px = 1, \ |x^{\mathrm{T}}PB|_\infty \geq \delta \right\} \\ &\quad \geq \beta_1 + \frac{2\delta}{\sqrt{\rho}} \\ &\quad > \beta_0 - \varepsilon. \end{aligned}$$

Combining the above with (34), we also obtain (35) and

$$2\alpha(\rho) \geq \beta_0 - \varepsilon$$

for

$$\rho < \left(\frac{2\delta}{-\beta_1 + \beta_0 - \varepsilon} \right)^2.$$

This completes the proof. □

Theorem 4 says that $\alpha(\rho)$ can be obtained by computing the maximum of $\dot{V}(x)$ over $\partial\mathcal{E}(P,\rho)$. For the single input case, Theorem 1 provides a method for determining if this maximum is negative. The exact value of $\alpha(\rho)$ can be computed with a procedure similar to the proof of Theorem 1. To avoid too much technical detail, we assume for simplicity that, for any real eigenvalue λ_j of $A + P^{-1}A^{\mathrm{T}}P$, there exists no x satisfying

$$(PA + A^{\mathrm{T}}P - \lambda_j P)x = PB.$$

This is the case if $(A + P^{-1}A^{\mathrm{T}}P, B)$ is controllable, thus the assumption is generally true. Let $\lambda_1, \lambda_2, \cdots, \lambda_{\mathcal{J}}$ be real numbers satisfying (14) and (15). Denote

$$\beta_j = -\lambda_j + \frac{1}{\rho} B^{\mathrm{T}}P(A^{\mathrm{T}}P + PA - \lambda_j P)^{-1}PB,$$

then,

$$\alpha(\rho) = \frac{1}{2}\min\left\{\beta_j : \ j = 0, 1, \cdots, \mathcal{J}\right\},$$

where β_0 is as defined in Theorem 4.

Since the overall convergence rate is limited by $\beta_0/2$ and it approaches this limit as ρ goes to 0, we would like β_0 not to be too small. For a fixed P, a formula for computing β_0 can be derived directly from the definition. Let the kernel of $B^{\mathrm{T}}P$ be $\{Ny : \ y \in \mathbf{R}^{n-m}\}$. Then

$$\beta_0 = -\lambda_{\max}\left((N^{\mathrm{T}}PN)^{-1}N^{\mathrm{T}}(A^{\mathrm{T}}P + PA)N\right). \tag{36}$$

Since N depends on P, the above formula gives us no hint for choosing P to obtain a satisfactory β_0. The following proposition will lead to an LMI approach to choosing P.

Proposition 2

$$\begin{aligned} \beta_0 = \quad & \sup_F \lambda \\ & s.t. \ (A + BF)^{\mathrm{T}}P + P(A + BF) \le -\lambda P. \end{aligned} \tag{37}$$

Proof. Notice that, for any F, we have

$$x^{\mathrm{T}}((A + BF)^{\mathrm{T}}P + P(A + BF))x = x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x, \quad \forall x^{\mathrm{T}}PB = 0.$$

It follows that

$$\begin{aligned} \beta_0 &= \min\left\{ - x^{\mathrm{T}}((A + BF)^{\mathrm{T}}P + P(A + BF))x : \ x^{\mathrm{T}}Px = 1, \ x^{\mathrm{T}}PB = 0\right\} \\ &\ge \min\left\{ - x^{\mathrm{T}}((A + BF)^{\mathrm{T}}P + P(A + BF))x : \ x^{\mathrm{T}}Px = 1\right\}, \end{aligned}$$

and hence,

$$\beta_0 \ge \sup_F \min\left\{ - x^{\mathrm{T}}((A + BF)^{\mathrm{T}}P + P(A + BF))x : \ x^{\mathrm{T}}Px = 1\right\}. \tag{38}$$

In what follows, we will prove that

$$\beta_0 = \sup_F \min\left\{ - x^{\mathrm{T}}((A + BF)^{\mathrm{T}}P + P(A + BF))x : \ x^{\mathrm{T}}Px = 1\right\}. \tag{39}$$

In view of (38), it suffices to show that for any $\varepsilon > 0$, there exists an $F = -kB^{\mathrm{T}}P$, with $k > 0$, such that

$$\min\left\{ - x^{\mathrm{T}}((A + BF)^{\mathrm{T}}P + P(A + BF))x : \ x^{\mathrm{T}}Px = 1\right\} \ge \beta_0 - \varepsilon. \tag{40}$$

From the definition of β_0, we see that there exists a $\delta > 0$ such that

$$\min\Big\{ -x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x :\ x^{\mathrm{T}}Px = 1,\ |x^{\mathrm{T}}PB|_2 \le \delta \Big\} \ge \beta_0 - \varepsilon$$

(refer to (33). Since $x^{\mathrm{T}}PBB^{\mathrm{T}}Px \ge 0$,

$$\begin{aligned} &\min\Big\{ -x^{\mathrm{T}}((A - kBB^{\mathrm{T}}P)^{\mathrm{T}}P + P(A - kBB^{\mathrm{T}}P))x : x^{\mathrm{T}}Px = 1,\ |x^{\mathrm{T}}PB| \le \delta \Big\} \\ &\quad \ge \beta_0 - \varepsilon, \end{aligned} \tag{41}$$

for all $k > 0$.

For every $x \in \mathbf{R}^n$,

$$-x^{\mathrm{T}}((A - kBB^{\mathrm{T}}P)^{\mathrm{T}}P + (A - kBB^{\mathrm{T}}P))x$$

is an increasing function of k. Similar to the proof of Theorem 4, it can be shown that there exists a $k > 0$, such that

$$\begin{aligned} &\min\Big\{ -x^{\mathrm{T}}((A - kBB^{\mathrm{T}}P)^{\mathrm{T}}P + P(A - kBB^{\mathrm{T}}P))x : x^{\mathrm{T}}Px = 1,\ |x^{\mathrm{T}}PB| \ge \delta \Big\} \\ &\quad \ge \beta_0 - \varepsilon. \end{aligned} \tag{42}$$

Combining (41) and (42), we have

$$\begin{aligned} &\min\Big\{ -x^{\mathrm{T}}((A - kBB^{\mathrm{T}}P)^{\mathrm{T}}P + P(A - kBB^{\mathrm{T}}P))x :\ x^{\mathrm{T}}Px = 1 \Big\} \\ &\quad \ge \beta_0 - \varepsilon. \end{aligned}$$

This proves (40) and hence (39).

Denote

$$\beta(F) = \min\Big\{ -x^{\mathrm{T}}((A + BF)^{\mathrm{T}}P + P(A + BF))x :\ x^{\mathrm{T}}Px = 1 \Big\}.$$

From (39), we have

$$\beta_0 = \sup_F \beta(F).$$

It can be shown that

$$\beta(F) = \max\Big\{ \lambda :\ (A + BF)^{\mathrm{T}}P + P(A + BF) \le -\lambda P \Big\}.$$

This brings us to (37). □

For a fixed P, β_0 is a finite value given by (36). Assume that (A, B) is controllable, then the eigenvalues of $(A + BF)$ can be arbitrarily assigned. If we also take P as a variable, then $-\beta_0/2$ can be made equal to the largest real part of the eigenvalues of $A + BF$ (see the definition, as given in Section 1, of the overall convergence rate for a linear system). This means that β_0 can be made arbitrarily large. But generally, as β_0 becomes very large, the matrix P will be badly conditioned, *i.e.*, the ellipsoid $\mathcal{E}(P, \rho)$ will become very thin in certain direction, and hence very "small", with respect to a fixed shape reference set. On the other hand, as mentioned in [11, 10], if our only objective is to enlarge the domain of attraction with respect to a reference

set, some eigenvalues of $A+BF$ will be very close to the imaginary axis, resulting in very small β_0. These two conflicting objectives can be balanced, for example, by pre-specifying a lower bound on β_0 and then maximizing the invariant ellipsoid with respect to some shape reference set. This mixed problem can be described as follows:

$$\begin{aligned} & \sup_{P>0,\rho,F,H} \quad \alpha \\ \text{s.t.} \quad & \text{a)} \ \alpha\mathcal{X}_{\mathrm{R}} \subset \mathcal{E}(P,\rho), \\ & \text{b)} \ (A+BF)^{\mathrm{T}}P + P(A+BF) < 0, \\ & \text{c)} \ \mathcal{E}(P,\rho) \subset \mathcal{L}(F), \\ & \text{d)} \ (A+BH)^{\mathrm{T}}P + P(A+BH) \leq -\underline{\beta}P, \end{aligned} \tag{43}$$

where $\mathcal{X}_{\mathrm{R}}$ is a shape reference set (see [9, 10, 11]). The constraint a) means that $\mathcal{E}(P,\rho)$ contains $\alpha\mathcal{X}_{\mathrm{R}}$. By maximizing α, $\mathcal{E}(P,\rho)$ will be maximized with respect to $\mathcal{X}_{\mathrm{R}}$. The constraints b) and c) guarantee that $\mathcal{E}(P,\rho)$ can be made contractive invariant and the constraint d) guarantees a lower bound $\underline{\beta}$ on the convergence rate. This optimization problem can be transformed into one with LMI constraints as those in [9, 10, 11]. By solving (43), we obtain the optimal ellipsoid $\mathcal{E}(P,\rho)$ along with two feedback matrices F and H. We may actually discard both F and H but instead use the bang-bang control law

$$u_i = -\mathrm{sign}(b_i^{\mathrm{T}}Px), \quad i = 1, 2, \cdots, m, \tag{44}$$

or the high gain controller

$$u_i = -\mathrm{sat}(kb_i^{\mathrm{T}}Px), \quad i = 1, 2, \cdots, m. \tag{45}$$

The final outcome is that under the control of (44) or (45), the closed-loop system will have a contractive invariant set $\mathcal{E}(P,\rho)$ and a guaranteed limit of the convergence rate

$$\frac{\beta_0}{2} \geq \frac{\underline{\beta}}{2}.$$

2.4 Maximal Convergence Control in the Presence of Disturbances

Consider the open-loop system

$$\dot{x} = Ax + Bu + Ew, \quad x \in \mathbf{R}^n, \ u \in \mathbf{R}^m, \ w \in \mathbf{R}^q, \tag{46}$$

where u is the control bounded by $|u|_\infty \leq 1$ and w is the disturbance bounded by $|w|_\infty \leq 1$. We also consider the quadratic Lyapunov function $V(x) = x^{\mathrm{T}}Px$.

Along the trajectory of the system (46),

$$\begin{aligned} \dot{V}(x,u,w) &= 2x^{\mathrm{T}}P(Ax + Bu + Ew) \\ &= x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x + 2\sum_{i=1}^{m} x^{\mathrm{T}}Pb_iu_i + 2x^{\mathrm{T}}PEw. \end{aligned}$$

No matter what w is, the control that maximizes the convergence rate, or minimizes $\dot{V}(x,u,w)$, is also

$$u_i = -\text{sign}(b_i^{\mathrm{T}} P x), \quad i = 1, 2, \cdots, m. \tag{47}$$

Under this control, we have

$$\dot{V}(x,w) = x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2\sum_{i=1}^{m} x^{\mathrm{T}} P b_i \,\text{sign}(b_i^{\mathrm{T}} P x) + 2x^{\mathrm{T}} P E w. \tag{48}$$

Now consider the closed-loop system

$$\dot{x} = Ax - \sum_{i=1}^{m} b_i \text{sign}(b_i^{\mathrm{T}} P x) + Ew. \tag{49}$$

An ellipsoid $\mathcal{E}(P,\rho)$ is invariant for the system (49) if $\dot{V}(x,w) \leq 0$ for all possible w bounded by $|w|_\infty \leq 1$. Unlike the system in the absence of disturbance, this system does not possess the property of contractive invariance. We may define strict invariance as in [11]. Since the invariance property can be easily extended to the strict invariance property (*e.g.*, by changing "$\leq$" to "$<$"), we will only investigate the former.

Denote the ith column of E as e_i. From (48), we see that, for $x \in \mathbf{R}^n$, the worst w that maximizes $\dot{V}(x,w)$ is

$$w_i = \text{sign}(e_i^{\mathrm{T}} P x), \quad i = 1, 2, \cdots, q.$$

Thus, we have the following obvious fact.

Fact 2 *The following three statements are equivalent:*

a) The ellipsoid $\mathcal{E}(P,\rho)$ can be made invariant for (46) with a bounded control $|u|_\infty \leq 1$;

b) The ellipsoid $\mathcal{E}(P,\rho)$ is invariant for (49);

c) The following condition is satisfied,

$$x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2\sum_{i=1}^{m} x^{\mathrm{T}} P b_i \text{sign}(b_i^{\mathrm{T}} P x) + \sum_{i=1}^{q} x^{\mathrm{T}} P e_i \text{sign}(e_i^{\mathrm{T}} P x) \leq 0, \quad \forall x \in \partial\mathcal{E}(P,\rho).$$

Denote

$$\dot{V}(x) = x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2\sum_{i=1}^{m} x^{\mathrm{T}} P b_i \text{sign}(b_i^{\mathrm{T}} P x) + 2\sum_{i=1}^{q} x^{\mathrm{T}} P e_i \text{sign}(e_i^{\mathrm{T}} P x).$$

Fact 2 says that $\mathcal{E}(P,\rho)$ is invariant for (49) if and only if $\dot{V}(x) \leq 0$ for all $x \in \partial\mathcal{E}(P,\rho)$. In the following, when we say that $\mathcal{E}(P,\rho)$ is invariant, we mean it is invariant for the system (49), or for the system (46) under the control

$$u_i = -\text{sign}(b_i^{\mathrm{T}} P x).$$

Proposition 3 *Suppose that $\mathcal{E}(P,\rho_1)$ and $\mathcal{E}(P,\rho_2)$ are both invariant and $\rho_1 < \rho_2$. Then $\mathcal{E}(P,\rho)$ is invariant for all $\rho \in [\rho_1, \rho_2]$.*

Proof. For a fixed $x \in \mathbf{R}^n$, consider $\dot{V}(kx), k \geq 0$, as a function of k. Clearly, this is a quadratic function and we also have $\dot{V}(0x) = 0$. Hence, if both $\dot{V}(k_1 x)$ and $\dot{V}(k_2 x)$ are negative, we must have $\dot{V}(kx)$ negative for all k between k_1 and k_2. The result of the proposition follows readily from this argument. □

From Proposition 3, we see that all the ρ such that $\mathcal{E}(P,\rho)$ is invariant form an interval. By Fact 2, this is the largest interval we can achieve with any feedback control constrained by the input saturation. In other words, the maximal convergence control produces both the maximal invariant ellipsoid and the minimal invariant ellipsoid. As we have mentioned in [11], a large invariant ellipsoid is desired for a large domain of attraction, and a small invariant ellipsoid indicates a good capability of disturbance rejection. In summary, the maximal convergence control is ideal for dealing with disturbances.

Similar to the system in the absence of disturbance, it can be shown that the saturated high gain feedback

$$u = -\text{sat}(kB^{\mathrm{T}}Px)$$

can be used to approximate the bang-bang control

$$u_i = -\text{sign}(b_i^{\mathrm{T}}Px), \quad i = 1, 2, \cdots, m,$$

and to produce a sub-optimal disturbance rejection performance.

The condition in Fact 2 c) can be verified by solving a set of optimization problems. Consider the single input and single disturbance case, $m = q = 1$. In this case, we have

$$\dot{V}(x) = x^{\mathrm{T}}(A^{\mathrm{T}}P + PA)x - 2x^{\mathrm{T}}PB\text{sign}(B^{\mathrm{T}}Px) + 2x^{\mathrm{T}}PE\text{sign}(E^{\mathrm{T}}Px).$$

Similar to the two-input case in Section 2.1, we can divide the surface $\partial\mathcal{E}(P,\rho)$ into 8 subsets:

$$\begin{aligned}
S_1 &= \left\{x \in \mathbf{R}^n :\ B^{\mathrm{T}}Px = 0,\ E^{\mathrm{T}}Px \geq 0,\ x^{\mathrm{T}}Px = \rho\right\}, \quad -S_1,\\
S_2 &= \left\{x \in \mathbf{R}^n :\ B^{\mathrm{T}}Px \geq 0,\ E^{\mathrm{T}}Px = 0,\ x^{\mathrm{T}}Px = \rho\right\}, \quad -S_2,\\
S_3 &= \left\{x \in \mathbf{R}^n :\ B^{\mathrm{T}}Px > 0,\ E^{\mathrm{T}}Px > 0,\ x^{\mathrm{T}}Px = \rho\right\}, \quad -S_3,\\
S_4 &= \left\{x \in \mathbf{R}^n :\ B^{\mathrm{T}}Px > 0,\ E^{\mathrm{T}}Px < 0,\ x^{\mathrm{T}}Px = \rho\right\}, \quad -S_4.
\end{aligned}$$

With this partition, $\mathcal{E}(P,\rho)$ is invariant if and only if

$$\max_{x \in S_1} \dot{V}(x) \leq 0, \quad \max_{x \in S_2} \dot{V}(x) \leq 0, \tag{50}$$

and all the local extrema of $\dot{V}(x)$ in S_3 and S_4 are non-positive. These optimization problems can be handled similarly as those in Section 2.1.

2.5 Extension to General Lyapunov Functions

In (5), the convergence rate is defined for a general Lyapunov function. It is known that the quadratic type function is the most popular and the simplest to deal with. In some special situations, one may prefer to use other types of Lyapunov functions. However, it should be expected that not all of the previous results can be extended to a general Lyapunov function.

Given a positive definite function $V(x)$, assume that $\frac{\partial V}{\partial x}$ exists and is continuous in $\mathbf{R}^n$. For $\rho > 0$, the level set is denoted as $L_V(\rho)$. Also consider the system (6). Along the trajectory of the system (6),

$$\begin{aligned}\dot{V}(x,u) &= \left(\frac{\partial V}{\partial x}\right)^{\mathrm{T}}(Ax+Bu)\\ &= \left(\frac{\partial V}{\partial x}\right)^{\mathrm{T}}Ax+\sum_{i=1}^{m}\left(\frac{\partial V}{\partial x}\right)^{\mathrm{T}}b_iu_i.\end{aligned}$$

Under the constraint that $|u|_\infty \le 1$, the control that maximizes the convergence rate, or minimizes $\dot{V}(x,u)$, is

$$u_i = -\mathrm{sign}\left(b_i^{\mathrm{T}}\frac{\partial V}{\partial x}\right), \quad i=1,2,\cdots,m. \tag{51}$$

Under this control, the closed-loop system is

$$\dot{x} = Ax - \sum_{i=1}^{m} b_i \mathrm{sign}\left(b_i^{\mathrm{T}}\frac{\partial V}{\partial x}\right). \tag{52}$$

It is clear that Fact 1 is true if we replace the ellipsoid $\mathcal{E}(P,\rho)$ with the level set $L_V(\rho)$ and (9) with

$$\dot{V}(x) = \left(\frac{\partial V}{\partial x}\right)^{\mathrm{T}}Ax - \sum_{i=1}^{m}\left(\frac{\partial V}{\partial x}\right)^{\mathrm{T}}b_i\,\mathrm{sign}\left(b_i\frac{\partial V}{\partial x}\right) < 0, \quad \forall x \in L_V(\rho)\setminus\{0\}. \tag{53}$$

Also, the bang-bang control (51) can be replaced with a continuous one, the saturated high gain feedback,

$$u = -\mathrm{sat}\left(kB^{\mathrm{T}}\frac{\partial V}{\partial x}\right)$$

to achieve a sub-optimal convergence rate.

For the system (46) subject to disturbance, it can be shown with similar arguments that the bang-bang control (51) will also maximize the convergence rate and achieve the largest invariant level set.

However, for those results that involve exact characterization of the maximal invariant ellipsoid (Theorem 1) and the maximal convergence rate (Theorem 4), they cannot be extended to a general Lyapunov function.

3 Discrete-time Systems

Consider the discrete-time system

$$x(k+1) = Ax(k)+Bu(k), \quad x\in\mathbf{R}^n,\ u\in\mathbf{R}^m,\ |u|_\infty \le 1. \tag{54}$$

Assume that the system is stabilizable and that B has full column rank. Given $V(x) = x^{\mathrm{T}}Px$, we would like to maximize the convergence rate within $\mathcal{E}(P,\rho)$.

Define

$$\Delta V(x,u) = (Ax+Bu)^{\mathrm{T}}P(Ax+Bu) - x^{\mathrm{T}}Px.$$

Our objective is to minimize $\Delta V(x,u)$ under the constraint $|u|_\infty \le 1$ for every $x \in \mathcal{E}(P,\rho)$.

Since B has full column rank, $B^{\mathrm{T}}PB$ is nonsingular and we have

$$\begin{aligned}\Delta V(x,u) &= u^{\mathrm{T}}B^{\mathrm{T}}PBu + 2u^{\mathrm{T}}B^{\mathrm{T}}PAx + x^{\mathrm{T}}A^{\mathrm{T}}PAx - x^{\mathrm{T}}Px \\ &= \left(u+(B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PAx\right)^{\mathrm{T}} B^{\mathrm{T}}PB\left(u+(B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PAx\right) \\ &\quad -x^{\mathrm{T}}A^{\mathrm{T}}PB(B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PAx + x^{\mathrm{T}}A^{\mathrm{T}}PAx - x^{\mathrm{T}}Px.\end{aligned}$$

Let

$$F_0 = -(B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PA.$$

It is clear that the convergence maximization problem is equivalent to

$$\min_{|u|_\infty \le 1} (u-F_0x)^{\mathrm{T}}B^{\mathrm{T}}PB(u-F_0x). \tag{55}$$

Let us first consider the single input case. In this case, the control that maximizes the convergence rate is simply,

$$u = \mathrm{sat}(F_0x). \tag{56}$$

This control is continuous in x, so it is better behaved than its continuous-time counterpart.

Consider the closed-loop system

$$x(k+1) = Ax(k) + B\mathrm{sat}(F_0x). \tag{57}$$

The system is asymptotically stable at the origin if

$$(A+BF_0)^{\mathrm{T}}P(A+BF_0) - P < 0,$$

i.e.,

$$\left(A-B(B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PA\right)^{\mathrm{T}} P\left(A-B(B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PA\right) - P < 0. \tag{58}$$

As in the continuous-time case, we also have

Fact 3 *For the single input case, the following two statements are equivalent:*

a) The ellipsoid $\mathcal{E}(P,\rho)$ can be made contractive invariant for the system (54) with a bounded control;

b) The ellipsoid $\mathcal{E}(P,\rho)$ is contractive invariant for (57), i.e., the following condition is satisfied,

$$(Ax+B\mathrm{sat}(F_0x))^{\mathrm{T}}P(Ax+B\mathrm{sat}(F_0x)) - x^{\mathrm{T}}Px < 0, \quad x \in \mathcal{E}(P,\rho)\setminus\{0\}. \tag{59}$$

Proposition 4 *For the single input case, the ellipsoid $\mathcal{E}(P,\rho)$ can be made contractive invariant for some $\rho > 0$ if and only if (58) is satisfied.*

Proof. There exists a $\rho_0 > 0$ such that

$$\mathcal{E}(P,\rho_0) \subset \left\{x \in \mathbf{R}^n : \ |F_0 x|_\infty \le 1\right\}.$$

Suppose that (58) is satisfied. Let $u = \mathrm{sat}(F_0 x)$, then $u = F_0 x$ for $x \in \mathcal{E}(P,\rho_0)$. Hence $\mathcal{E}(P,\rho_0)$ is contractive invariant under $u = \mathrm{sat}(F_0 x)$.

On the other hand, suppose that $\mathcal{E}(P,\rho)$ can be made contractive invariant. By Fact 3, we have (59). Let $\rho_1 = \min\{\rho_0, \rho\}$, then $\mathrm{sat}(F_0 x) = F_0 x$ for $x \in \mathcal{E}(P,\rho_1)$. It follows from (59) that (58) is true. □

Suppose that (58) is satisfied, we would like to obtain the condition for $\mathcal{E}(P,\rho)$ to be invariant for (57) and to find the largest ρ such that $\mathcal{E}(P,\rho)$ is invariant.

Because of (58), we have

$$(Ax + B\mathrm{sat}(F_0 x))^{\mathrm{T}} P (Ax + B\mathrm{sat}(F_0 x)) - x^{\mathrm{T}} P x < 0$$

for all

$$x \in \left\{x \in \mathbf{R}^n : \ |F_0 x| \le 1, \ x \ne 0\right\} = \left\{x \in \mathbf{R}^n : \ |B^{\mathrm{T}} P A x| \le B^{\mathrm{T}} P B, \ x \ne 0\right\}.$$

Hence, $\mathcal{E}(P,\rho)$ is invariant if and only if

$$\begin{aligned}(Ax - B)^{\mathrm{T}} P (Ax - B) - x^{\mathrm{T}} P x < 0,& \\ \forall x \in \mathcal{E}(P,\rho) \text{ such that } B^{\mathrm{T}} P A x \ge B^{\mathrm{T}} P B.&\end{aligned} \tag{60}$$

Like the equivalence of (12) and (13), it can be shown that (60) is equivalent to

$$\begin{aligned}(Ax - B)^{\mathrm{T}} P (Ax - B) - x^{\mathrm{T}} P x < 0,& \\ \forall x \in \partial\mathcal{E}(P,\rho) \text{ such that } B^{\mathrm{T}} P A x \ge B^{\mathrm{T}} P B.&\end{aligned} \tag{61}$$

And we have

Theorem 5 *For the single input case, assume that (58) is satisfied. Let $\lambda_1, \lambda_2, \cdots, \lambda_{\mathcal{J}} > 1$ be real numbers such that*

$$\det \begin{bmatrix} \lambda_j P - A^{\mathrm{T}} P A & P \\ \rho^{-1} A^{\mathrm{T}} P B B^{\mathrm{T}} P A & \lambda_j P - A^{\mathrm{T}} P A \end{bmatrix} = 0 \tag{62}$$

and

$$B^{\mathrm{T}} P A (A^{\mathrm{T}} P A - \lambda_j P)^{-1} A^{\mathrm{T}} P B \ge B^{\mathrm{T}} P B. \tag{63}$$

Then, $\mathcal{E}(P,\rho)$ is contractive invariant for the system (57) if and only if

$$\begin{aligned}(\lambda_j - 1)\rho - B^{\mathrm{T}} P A (A^{\mathrm{T}} P A - \lambda_j P)^{-1} A^{\mathrm{T}} P B + B^{\mathrm{T}} P B < 0,& \\ \forall j = 1, 2, \cdots \mathcal{J}.&\end{aligned} \tag{64}$$

If there is no $\lambda_j > 1$ satisfying (62) and (63), then $\mathcal{E}(P,\rho)$ is contractive invariant.

Proof. Denote

$$g(x) = (Ax - B)^{\mathrm{T}} P(Ax - B) - x^{\mathrm{T}} P x.$$

On the plane $B^{\mathrm{T}} PAx = B^{\mathrm{T}} PB$, we have

$$Ax - B(B^{\mathrm{T}} PB)^{-1} B^{\mathrm{T}} PAx = Ax - B.$$

Since P satisfies (58), we have $g(x) < 0$ for all x on the plane $B^{\mathrm{T}} PAx = B^{\mathrm{T}} PB$. So the invariance of $\mathcal{E}(P, \rho)$ is equivalent to that all the extrema in the surface $x^{\mathrm{T}} Px = \rho, B^{\mathrm{T}} PAx > B^{\mathrm{T}} PB$, if any, are less than zero.

Suppose that x is an extremum in the surface, then by Lagrange multiplier method, there exists a real number λ such that

$$(A^{\mathrm{T}} PA - \lambda P)x = A^{\mathrm{T}} PB, \quad x^{\mathrm{T}} Px = \rho, \quad B^{\mathrm{T}} PAx > B^{\mathrm{T}} PB. \tag{65}$$

At the extremum, we have

$$g(x) = (\lambda - 1)\rho - B^{\mathrm{T}} PAx + B^{\mathrm{T}} PB.$$

Since $B^{\mathrm{T}} PAx > B^{\mathrm{T}} PB$, we have $g(x) < 0$ for $\lambda \leq 1$. So we only need to consider $\lambda > 1$. Assume that

$$A^{\mathrm{T}} PB = \begin{bmatrix} 0 \\ r \end{bmatrix}$$

and partition $A^{\mathrm{T}} PA$ and P accordingly as

$$A^{\mathrm{T}} PA = \begin{bmatrix} Q_1 & Q_{12} \\ Q_{12}^{\mathrm{T}} & q_2 \end{bmatrix}, \quad P = \begin{bmatrix} P_1 & P_{12} \\ P_{12}^{\mathrm{T}} & p_2 \end{bmatrix}.$$

It follows from (58) that $Q_1 - P_1 < 0$ and $Q_1 - \lambda P_1 < 0$ for all $\lambda > 1$. It can be shown as in the proof of Theorem 1 that $(A^{\mathrm{T}} PA - \lambda P)x = A^{\mathrm{T}} PB$ and $\lambda > 1$ imply that $A^{\mathrm{T}} PA - \lambda P$ is nonsingular. Hence

$$x = (A^{\mathrm{T}} PA - \lambda P)^{-1} A^{\mathrm{T}} PB,$$

and from $x^{\mathrm{T}} Px = \rho$,

$$B^{\mathrm{T}} PA(A^{\mathrm{T}} PA - \lambda P)^{-1} P(A^{\mathrm{T}} PA - \lambda P)^{-1} A^{\mathrm{T}} PB = \rho.$$

The rest of the proof is similar to that of Theorem 1. □

The supremum of ρ such that $\mathcal{E}(P, \rho)$ is contractive invariant can be obtained by checking the condition in Theorem 5 bisectionally.

Let us now consider the multiple input case and examine the optimization problem (55). If $B^{\mathrm{T}} PB$ is diagonal, then we also have

$$u = \mathrm{sat}(F_0 x),$$

as the optimal control. But for the general case, the optimal u cannot be put into this simple form. Actually, (55) is a minimal distance problem. Let

$$v = F_0 x = -(B^{\mathrm{T}} P B)^{-1} B^{\mathrm{T}} P A x,$$

then the optimization problem is to find a point in the box $|u|_\infty \leq 1$ that is closest to v, with the distance induced by the weighted 2-norm, *i.e.*,

$$|u - v|_{B^{\mathrm{T}} P B} = ((u - v)^{\mathrm{T}} B^{\mathrm{T}} P B (u - v))^{\frac{1}{2}}.$$

This is illustrated in Fig. 3. In Fig. 3, v is marked with "o" and the optimal u is marked with "*". Suppose that $v \in \mathbf{R}^2$ is outside of the unit square, then there exists a smallest ellipsoid

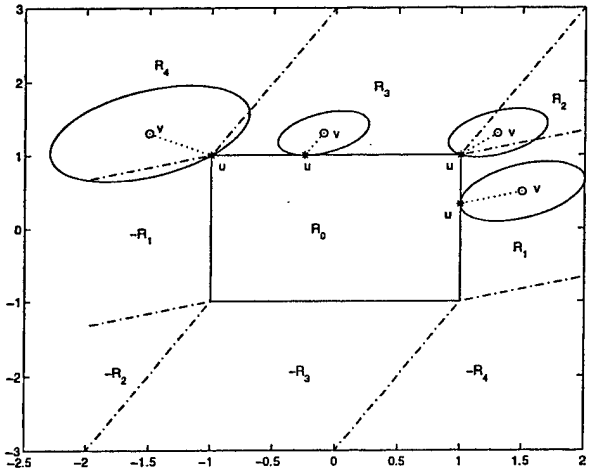


Figure 3: Illustration for the relation between v and the optimal u.

centered at v,

$$\mathcal{E}_0 = \left\{ u \in \mathbf{R}^2 : \ (u - v)^{\mathrm{T}} B^{\mathrm{T}} P B (u - v) \leq r^2 \right\},$$

which includes a point of the unit square. Since the unit square is convex and $\mathcal{E}_0$ is strictly convex, they have a unique intersection, which is the optimal u.

For the case $m = 2$, the solution can be put into an explicit formula. For simplicity, let

$$B^{\mathrm{T}} P B = \begin{bmatrix} a & -c \\ -c & b \end{bmatrix}$$

and assume that $c > 0$. We partition the plane into 9 regions (see Fig. 3, where the regions are divided by the dash-dotted lines):

$$\begin{aligned} R_0 &= \left\{ v \in \mathbf{R}^2 : \ |v|_\infty \leq 1 \right\}, \\ R_1 &= \left\{ v \in \mathbf{R}^2 : \ v_1 > 1, \ -1 + \frac{c}{b}(v_1 - 1) < v_2 \leq 1 + \frac{c}{b}(v_1 - 1) \right\}, \\ R_2 &= \left\{ v \in \mathbf{R}^2 : \ 1 + \frac{c}{b}(v_1 - 1) < v_2 \leq 1 + \frac{a}{c}(v_1 - 1) \right\}, \end{aligned}$$

$$R_3 = \left\{v \in \mathbf{R}^2 :\ v_2 > 1,\ -1 + \frac{c}{a}(v_2 - 1) \le v_1 < 1 + \frac{c}{a}(v_2 - 1)\right\},$$
$$R_4 = \left\{v \in \mathbf{R}^2 :\ v_1 < -1 + \frac{c}{a}(v_2 - 1),\ v_2 \ge 1 + \frac{c}{b}(v_1 + 1)\right\},$$

$-R_1, -R_2, -R_3$, and $-R_4$. With this partition, the maximal convergence control that solves (55) can be written as,

$$u = \begin{cases} v, & \text{if } v \in R_0, \\ \begin{bmatrix} 1 \\ v_2 - c(v_1 - 1)/b \end{bmatrix}, & \text{if } v \in R_1, \\ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, & \text{if } v \in R_2, \\ \begin{bmatrix} v_1 - c(v_2 - 1)/a \\ 1 \end{bmatrix}, & \text{if } v \in R_3, \\ \begin{bmatrix} -1 \\ 1 \end{bmatrix}, & \text{if } v \in R_4. \end{cases} \tag{66}$$

and for $v \in -R_1, -R_2, -R_3$ or $-R_4$, the optimal control u is symmetric. Since u_i does not solely depend on v_i, we can view u as a coupled saturation function of v. If $m > 2$, it is hard to put the optimal solution into an explicit formula. However, the optimization problem (55) can be transformed into a very simple LMI problem:

$$\begin{aligned} &\min_{u,\gamma} \gamma \\ \text{s.t.} \quad &\begin{bmatrix} \gamma & (u - v)^{\mathrm{T}} \\ u - v & (B^{\mathrm{T}}PB)^{-1} \end{bmatrix} \ge 0, \quad |u|_\infty \le 1. \end{aligned} \tag{67}$$

This optimization problem can be efficiently solved so that the control can be computed and implemented on line. It can be shown that the optimal control u resulting from (67) is continuous in v and hence continuous in x.

We have some remarks for the case that B has full row rank. It is known that there exists a linear dead-beat control for this case. Let B^+ be the right inverse of B, then

$$A^{\mathrm{T}}PB(B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PA = A^{\mathrm{T}}PB(B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PBB^+A = A^{\mathrm{T}}PA.$$

It follows that

$$\begin{aligned} V(Ax + Bu) &= \left(u + (B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PAx\right)^{\mathrm{T}} B^{\mathrm{T}}PB \\ &\qquad \times \left(u + (B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PAx\right). \end{aligned}$$

It is easy to see that the optimal convergence control is also a dead-beat control for x in the region,

$$\left\{x \in \mathbf{R}^n :\ |(B^{\mathrm{T}}PB)^{-1}B^{\mathrm{T}}PAx|_\infty \le 1\right\}.$$

Outside of this region, the optimal control can be obtained by solving a simple LMI problem.

Similar to the continuous-time case, it can be shown that the overall convergence rate increases as ρ decreases. And if $\mathcal{E}(P,\rho)$ is in the linear region, the overall convergence rate is a constant which equals to that of a linear system.

The problem of disturbance rejection for a discrete-time system is much more complicated than its continuous-time counterpart. This is because the expression of $\Delta V(x,u,w)$ contains some cross terms between u and w and the maximal convergence control will be dependent on w.

4 Conclusions

We have shown in this paper that, for a continuous-time system subject to input saturation and persistent disturbance, the maximal convergence rate control is a bang-bang type control with a simple switching strategy. A saturated high gain feedback is developed to avoid the discontinuity of the bang-bang control. For a discrete-time system, the maximal convergence control is a coupled saturated linear feedback. For both continuous-time and discrete-time systems, we also provided methods for determining the largest ellipsoid that can be made invariant with a bounded control.

References

[1] F. Blanchini, "Set invariance in control – a survey", *Automatica*, Vol. 35, No.11, pp. 1747-1767, 1999.

[2] S. Boyd, L. El Ghaoui, E. Feron and V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory*, SIAM Studies in Appl. Mathematics, Philadelphia, 1994.

[3] P. Bruck, "A new result in time optimal control of discrete systems," *IEEE Transactions on Automatic Control*, Vol. 19, pp. 597-598, 1974.

[4] E. J. Davison and E. M. Kurak, "A computational method for determining quadratic Lyapunov functions for non-linear systems," *Automatica*, Vol. 7, pp. 627-636, 1971.

[5] E. G. Gilbert, and K. T. Tan, "Linear systems with state and control constraints: the theory and application of maximal output admissible sets", *IEEE Trans. Automat. Contr.* Vol. 36, pp. 1008-1020, 1991.

[6] C. A. Desoer and J. Wing, "An optimal strategy for a saturating sampled data system," *IRE Transactions on Automatic Control*, Vol. 6, pp. 5-15, 1961.

[7] P. O. Gutman, and M.Cwikel, "Admissible sets and feedback control for discrete-time linear dynamical systems with bounded control and dynamics", *IEEE Trans. Auto. Contr.*, Vol. 31, pp. 373-376, 1986.

[8] L. M. Hocking, *Optimal Control, An Introduction to the Theory and Applications*, Oxford University Press, Oxford, 1991.

[9] T. Hu and Z. Lin. *Control Systems with Actuator Saturation: Analysis and Design*, Birkhäuser, Boston, 2001.

[10] T. Hu and Z. Lin, "On enlarging the basin of attraction for linear systems under saturated linear feedback," *Sys. & Contr. Lett.*, Vol. 40, No. 1, pp. 59-69, May, 2000.

[11] T. Hu, Z. Lin and B. M. Chen, "An analysis and design method for linear systems subject to actuator saturation and disturbance," *American Control Conferences*, Chicago, pp. 725-729, 2000, also under revision for *Automatica.*

[12] T. Hu, Z. Lin and B. M. Chen, " Analysis and design for discrete-time systems subject to actuator saturation and disturbance," submitted for publication.

[13] H. Khalil, *Nonlinear Systems*, Prentice-Hall, Upper Saddle River, New Jersey, 1996.

[14] E. B. Lee and L. Markus, *Foundations of Optimal Control*, John Wiley and Sons Inc., New York, 1967.

[15] J. Macki and M. Strauss. *Introduction to Optimal Control*, Springer-Verlag, 1982.

[16] A. Megretski, "Output feedback stabilization with saturated control: making the input-output map L_2-bounded," preprint. See also "L_2 BIBO output feedback stabilization with saturated control," *Proceedings of the 13th Triennial World Congress of IFAC*, Vol. D, pp. 435-440, 1996.

[17] R. Suarez, J. Alvarez-Ramirez and J. Solis-Daun, "Linear systems with bounded inputs: global stabilization with eigenvalue placement," *International Journal of Robust and Nonlinear Control*, Vol. 7, pp. 835-845, 1997.

[18] G.F. Wredenhagen and P.R. Belanger, "Piecewise-linear LQ control for systems with input constraints," *Automatica*, Vol. 30, pp. 403-416, 1994.

Publication 27

Revised for *Automatica*

Output Regulation of General Linear Systems with Saturating Actuators

TINGSHU HU[†] ZONGLI LIN[†]

[†]Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903
th7f, zl5y@virginia.edu

Abstract

This paper studies the classical problem of output regulation for linear systems subject to actuator saturation. The asymptotically regulatable region, the set of all initial conditions of the plant and the exosystem for which output regulation is possible, is characterized in terms of the null controllable region of the anti-stable subsystem of the plant. Output regulation laws are constructed from a given stabilizing state feedback law. It is shown that a stabilizing feedback law that achieves a larger domain of attraction leads to a feedback law that achieves output regulation on a larger subset of the asymptotically regulatable region. A feedback law that achieves global stabilization on the asymptotically null controllable region leads to a feedback law that achieves output regulation on the entire asymptotically regulatable region.

Keywords: actuator saturation, output regulation, regulatable region.

[1]Work supported in part by the US office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

1 Introduction

There has been considerable research on the problem of stabilization and output regulation of linear systems subject to actuator saturation. The problem of stabilization involves the issues of the characterization of null controllable region (or, asymptotically null controllable region), the set of all initial conditions that can be driven to the origin by the saturating actuators in some finite time (or, asymptotically), and the construction of feedback laws that achieve stabilization on the entire or a large portion of the asymptotically null controllable region. Recent years have witnessed extensive research that addresses these issues. In particular, for an open loop system that is stabilizable and has all its poles in the closed left-half plane, it was established in Macki and Strauss (1982) and Sontag (1984) that the asymptotically null controllable region is the entire state space. For this reason, a linear system that is stabilizable in the usual linear sense and has all its poles in the closed left-half plane is said to be asymptotically null controllable with bounded controls, or ANCBC. For ANCBC systems subject to actuator saturation, various feedback laws that achieve global or semi-global stabilization on the null controllable region have been constructed (see, for example, Teel, 1992; Sussmann, Sontag and Yang, 1994; Suarez, Alvarez-Ramirez and Solis-Daun, 1997; Lin and Saberi, 1993; and Lin, 1998). Here by semi-global stabilization on the null controllable region we mean the construction of a feedback law that achieves a domain of attraction large enough to include any *a priori* given (arbitrarily large) bounded set in the null controllable region.

For exponentially unstable open-loop systems subject to actuator saturation, it was shown by Teel (1996) that the problem of stabilization can be reduced to one of stabilizing its anti-stable subsystem, whose null controllable region is a bounded convex open set. A complete characterization of the null controllable region for a general linear system was developed in Hu and Lin (2001a), where simple feedback laws were also constructed that achieve semi-global stabilization on the null controllable region for linear systems with only two exponentially unstable poles. More recently, feedback laws have been constructed that achieve semi-global stabilization on the null controllable region (Hu, Lin and Shamash, 2001) for general linear systems subject to actuator saturation.

In comparison with the problem of stabilization, the problem of output regulation for linear systems subject to actuator saturation, however, has received relatively less attention. The few works that have motivated our current research are De Santis (2000), De Santis and Isidori (2001), Teel (1992) and Lin, Stoorvogel and Saberi (1996). In Teel (1992) and Lin et al (1996), the problem of output regulation was studied for ANCBC systems subject to actuator saturation. In Lin et al (1996), necessary and sufficient conditions on the plant/exosystem and their initial conditions were derived under which output regulation can be achieved, and feedback laws that achieve semi-global output regulation were constructed. De Santis (2000) and De Santis and Isidori (2001) made an attempt to address the problem of output regulation for exponentially unstable linear systems subject to actuator saturation. The attempt was to enlarge the set of initial conditions of the plant and the exosystem under which output regulation can be achieved. In particular, for plants

with only one positive pole and exosystems that contain only one frequency component, feedback laws were constructed that achieve output regulation on what we will characterize in this paper as the regulatable region.

Apart from the above mentioned literature, there also exist a variety of results on the problem of output regulation/output tracking with saturating actuators. For example, Saberi, Stoorvogel and Sannuti (1999) considered the problem for ANCBC systems subject to actuator and rate saturation. Blanchini and Miani (2000) studied the relation between stabilization and tracking a reference signal. Tarbouriech, Pittet and Burgat (2000) made an attempt to synthesize nonlinear control laws for local output regulation. A gain scheduling design was proposed in Lin, Pachter and Banda (1998) to improve the transient tracking performance for second order systems and was generalized by Turner, Postlethwaite and Walker (2000). Casavola and Mosca (2001) developed a switching logic to achieve global output regulation for ANCBC systems.

The objective of this paper is to systematically study the problem of output regulation for general linear systems subject to actuator saturation. Unlike in the case for a linear system without actuator saturation, where output regulation can be achieved for all initial states, here in the presence of saturation, the set of initial states for which output regulation is possible may not be the whole state space. For instance, suppose that we have an exponentially unstable plant, then it is known that the set of initial states of the plant that can be kept bounded with saturating actuators is not the whole state space. The set of initial states of the plant where output regulation is possible is even more restricted. For this reason, we will start our investigation by characterizing the set of initial states of the plant and the exosystem where output regulation is possible and we will call this set the regulatable region. It turns out that the regulatable region can be characterized in terms of the null controllable region of the anti-stable subsystem of the plant.

We then proceed to construct output regulation laws from a given stabilizing state feedback law. We show that a stabilizing feedback law that achieves a larger domain of attraction leads to a feedback law that achieves output regulation on a larger subset of the regulatable region and, a stabilizing feedback law on the entire null controllable region leads to a feedback law that achieves output regulation on the entire regulatable region. Our feedback law employs a switching scheme which generalizes the safe switching idea of Wredenhagen and Belanger (1994), where the switching was based on a group of nested ellipsoids. In this paper, the switching is based on a group of nested invariant sets, which may not even be convex.

The remainder of this paper is organized as follows. In Section 2, we state the problem of output regulation for linear systems with saturating actuators. Section 3 characterizes the regulatable region. Sections 4 and 5 respectively construct state feedback and error feedback laws that achieve output regulation on the regulatable region. In Section 6, we illustrate our results on an aircraft model. Finally, Section 7 gives a brief concluding remark to our current work.

Throughout the paper, we will use standard notation. For a vector $u \in \mathbf{R}^m$, we use $|u|_\infty$ and $|u|_2$ to denote the vector ∞-norm and the 2-norm. For a measurable function $u : [0, \infty) \to \mathbf{R}^m$, we define $\|u\|_\infty = \sup_{t\in[0,\infty]} |u(t)|_\infty$. We use $\mathrm{sat}(\cdot)$ to denote the standard saturation function, i.e., the ith component of $\mathrm{sat}(u)$ is $\mathrm{sgn}(u_i)\min\{1, |u_i|\}$.

2 Preliminaries and Problem Statement

In this section, we first recall from Francis (1975) and Isidori and Byrnes (1990) the classical formulation and results on the problem of output regulation for linear systems. This brief review will motivate our formulation as well as the solution to the problem of output regulation for linear systems subject to actuator saturation.

Consider a linear system

$$\begin{aligned} \dot{x} &= Ax + Bu + Pw, \\ \dot{w} &= Sw, \\ e &= Cx + Qw. \end{aligned} \tag{1}$$

The first equation of this system describes a plant, with state $x \in \mathbf{R}^n$ and input $u \in \mathbf{R}^m$, subject to the effect of a disturbance represented by Pw. The third equation defines the error $e \in \mathbf{R}^q$ between the actual plant output Cx and a reference signal $-Qw$ that the plant output is required to track. The second equation describes an autonomous system, often called the exosystem, with state $w \in \mathbf{R}^r$. The exosystem models the class of disturbances and references taken into consideration.

The control action to the plant, u, can be provided either by state feedback or by error feedback. The objective is to achieve internal stability and output regulation. *Internal stability* means that if we disconnect the exosystem and set w equal to zero then the closed-loop system is asymptotically stable. *Output regulation* means that, for any initial conditions of the plant and the exosystem, the state of the plant is bounded and $e(t) \to 0$ as $t \to \infty$.

The solution to the output regulation problem was first obtained by Francis (1975). It is now well known that under some mild necessary assumptions, the output regulation problem is solvable if and only if there exist matrices Π and Γ that solve the linear matrix equations

$$\begin{aligned} &\Pi S = A\Pi + B\Gamma + P, \\ &C\Pi + Q = 0. \end{aligned} \tag{2}$$

For more detail about the assumptions and the solution, see Francis (1975) and Isidori and Byrnes (1990).

In this paper, we study the problem of output regulation for the linear system (1) subject to actuator saturation, where the control signal u is the output of saturating actuators and can be assumed to be measurable and satisfy the bound $\|u\|_\infty \leq 1$. A control u that satisfies these assumptions will be referred to as an admissible control.

Following Francis (1975), Isidori and Byrnes (1990) and Lin et al (1996), we make the following necessary assumptions on the plant and the exosystem:

A1. The matrix equations (2) have solution (Π, Γ);

A2. The matrix S has all its eigenvalues on the imaginary axis and is neutrally stable;

A3. The pair (A, B) is stabilizable;

A4. The initial state w_0 of the exosystem is in the following set

$$\mathcal{W}_0 = \left\{ w_0 \in \mathbf{R}^r : |\Gamma w(t)|_\infty = |\Gamma e^{St} w_0|_\infty \leq \rho, \forall t \geq 0 \right\}, \tag{3}$$

for some $\rho \in [0, 1)$ and $\mathcal{W}_0$ is compact. For later use, we denote $\delta = 1 - \rho$.

We note that the compactness of $\mathcal{W}_0$ can be guaranteed by the observability of (Γ, S). Indeed, if (Γ, S) is not observable, then the exosystem can be reduced to make it so. As will be seen shortly, the exosystem affects the output regulation property through the signal $\Gamma w(t)$.

3 The Regulatable Region

To begin with, we define a new state $z = x - \Pi w$ and rewrite the system equations as

$$\begin{aligned} \dot{z} &= Az + Bu - B\Gamma w, \\ \dot{w} &= Sw, \\ e &= Cz. \end{aligned} \tag{4}$$

This particular state transformation has been traditionally used in the output regulation literature to transform the output regulation problem into a stabilization problem. The remaining part of the paper will be focused on system (4). All the results can be easily restated in terms of the original state of the plant x by replacing z with $x - \Pi w$. Here we note that, when $w = 0$, the internal stability in terms of z is the same as that in terms of x. As to output regulation, it is clear that $e(t)$ goes to zero asymptotically if $z(t)$ does. To combine the objectives of achieving internal stability and achieving output regulation, we will define the notion of regulatable region in terms of driving $z(t)$ to zero instead of driving $e(t)$ to zero. As will be explained in detail in Remark 1, this will result in essentially the same description of the regulatable region and will avoid some tedious technical discussions.

Definition 1

1) *Given a $T > 0$, a pair $(z_0, w_0) \in \mathbf{R}^n \times \mathbf{R}^r$ is regulatable in time T if there exists an admissible control $u(\cdot)$, such that the response of (4) satisfies $z(t) = 0$ for all $t \geq T$;*

2) *A pair (z_0, w_0) is regulatable if there exist a finite $T > 0$ and an admissible control $u(\cdot)$ such that $z(t) = 0$ for all $t \geq T$;*

3) *The set of all (z_0, w_0) regulatable in time T is denoted as $\mathcal{R}_g(T)$ and the set of all regulatable (z_0, w_0) is referred to as the regulatable region and is denoted as $\mathcal{R}_g$;*

4) *The set of all (z_0, w_0) for which there exists an admissible control $u(\cdot)$ such that the response of (4) satisfies $\lim_{t\to\infty} z(t) = 0$ is referred to as the asymptotically regulatable region and is denoted as $\mathcal{R}_g^a$.*

It is clear that $\mathcal{R}_g(T_1) \subset \mathcal{R}_g(T_2)$ if $T_2 > T_1$. Because of Assumption A4, the requirement in Defintion 1 that $z(t) = 0$ for all $t \geq T$ can be replaced with $z(T) = 0$.

We will describe $\mathcal{R}_g(T)$, $\mathcal{R}_g$ and $\mathcal{R}_g^a$ in terms of the null controllable region of the plant $\dot{v} = Av + Bu$, $\|u\|_\infty \leq 1$, which is defined as follows.

Definition 2 *The null controllable region in time T, denoted as $\mathcal{C}(T)$, is the set of $v_0 \in \mathbf{R}^n$ that can be driven to the origin in time T by admissible controls; The null controllable region, denoted*

as $\mathcal{C}$, is the set of $v_0 \in \mathbf{R}^n$ that can be driven to the origin in a finite time by admissible controls; The asymptotically null controllable region, denoted as $\mathcal{C}^a$, is the set of all v_0 that can be driven to the origin asymptotically by admissible controls.

Clearly, $\mathcal{C} = \bigcup_{T\in[0,\infty)} \mathcal{C}(T)$ and

$$\mathcal{C}(T) = \left\{ v \in \mathbf{R}^n : \ v = \int_0^T e^{-A\tau} B u(\tau) d\tau, \ \|u\|_\infty \le 1 \right\}. \tag{5}$$

It is also clear that the null controllable region and the asymptotically null controllable region are identical if the pair (A, B) is controllable. Some simple methods to describe $\mathcal{C}$ and $\mathcal{C}^a$ were recently developed in Hu and Lin (2001a).

To simplify the characterization of $\mathcal{R}_g$ and $\mathcal{R}_g^a$, let us assume, without loss of generality, that $z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, z_1 \in \mathbf{R}^{n_1}, z_2 \in \mathbf{R}^{n_2}$ and

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \tag{6}$$

where $A_1 \in \mathbf{R}^{n_1 \times n_1}$ is semi-stable (i.e., all its eigenvalues are in the closed left-half plane) and $A_2 \in \mathbf{R}^{n_2 \times n_2}$ is anti-stable (i.e., all its eigenvalues are in the open right-half plane). The system including the anti-stable subplant and the exosystem

$$\begin{aligned} \dot{z}_2 &= A_2 z_2 + B_2 u - B_2 \Gamma w, \\ \dot{w} &= S w, \end{aligned} \tag{7}$$

is of crucial importance. Denote its regulatable regions as $\mathcal{R}_{g2}(T)$, $\mathcal{R}_{g2}$ and $\mathcal{R}_{g2}^a$, and the null controllable regions for the system $\dot{v}_2 = A_2 v_2 + B_2 u$ as $\mathcal{C}_2(T)$ and $\mathcal{C}_2$. Then, the asymptotically null controllable region for the system $\dot{v} = Av + Bu$ is given by $\mathcal{C}^a = \mathbf{R}^{n_1} \times \mathcal{C}_2$ (Hájek, 1991), where $\mathcal{C}_2$ is a bounded convex open set. Denote the closure of $\mathcal{C}_2$ as $\bar{\mathcal{C}}_2$, then

$$\bar{\mathcal{C}}_2 = \left\{ v_2 \in \mathbf{R}^{n_2} : \ v_2 = \int_0^\infty e^{-A_2\tau} B_2 u(\tau) d\tau, \ \|u\|_\infty \le 1 \right\}.$$

Also, if D is a closed subset of $\mathcal{C}_2$, then there is a finite $T > 0$ such that $D \subset \mathcal{C}_2(T)$.

Theorem 1 *Let $V_2 \in \mathbf{R}^{n_2 \times r}$ be the unique solution to the matrix equation*

$$-A_2 V_2 + V_2 S = -B_2 \Gamma$$

and let $V(T) = V_2 - e^{-A_2 T} V_2 e^{ST}$. Then

$$i) \quad \mathcal{R}_{g2}(T) = \{(z_2, w) \in \mathbf{R}^{n_2} \times \mathcal{W}_0 : z_2 - V(T) w \in \mathcal{C}_2(T)\}; \tag{8}$$

$$ii) \quad \mathcal{R}_{g2} = \{(z_2, w) \in \mathbf{R}^{n_2} \times \mathcal{W}_0 : z_2 - V_2 w \in \mathcal{C}_2\}; \tag{9}$$

$$iii) \quad \mathcal{R}_g^a = R^{n_1} \times \mathcal{R}_{g2}. \tag{10}$$

Proof. i) Given $(z_{20}, w_0) \in \mathbf{R}^{n_2} \times \mathcal{W}_0$ and an admissible control u, the solution of (7) at $t = T$ is,

$$z_2(T) = e^{A_2 T}\left(z_{20} + \int_0^T e^{-A_2\tau} B_2 u(\tau) d\tau - \int_0^T e^{-A_2\tau} B_2 \Gamma e^{S\tau} w_0 d\tau\right). \tag{11}$$

Since $-A_2V_2 + V_2S = -B_2\Gamma$, we have

$$\frac{de^{-A_2\tau}V_2e^{S\tau}}{d\tau} = e^{-A_2\tau}(-A_2V_2)e^{S\tau} + e^{-A_2\tau}V_2Se^{S\tau} = e^{-A_2\tau}(-A_2V_2 + V_2S)e^{S\tau} = -e^{-A_2\tau}B_2\Gamma e^{S\tau},$$

noting that A_2 and $e^{-A_2\tau}$ commute, and S and $e^{S\tau}$ commute. Hence,

$$-\int_0^T e^{-A_2\tau}B_2\Gamma e^{S\tau}d\tau = \int_0^T de^{-A_2\tau}V_2e^{S\tau} = e^{-A_2T}V_2e^{ST} - V_2 = -V(T). \tag{12}$$

Therefore,

$$e^{-A_2T}z_2(T) = z_{20} - V(T)w_0 + \int_0^T e^{-A_2\tau}B_2u(\tau)d\tau. \tag{13}$$

To prove (8), it suffices to show that $z_{20} - V(T)w_0 \in \mathcal{C}_2(T) \iff (z_{20}, w_0) \in \mathcal{R}_{g_2}(T)$.

If $z_{20} - V(T)w_0 \in \mathcal{C}_2(T)$, then by (5), there exists an admissible control u_1 such that

$$z_{20} - V(T)w_0 = \int_0^T e^{-A_2\tau}B_2u_1(\tau)d\tau.$$

Let $u(t) = -u_1(t)$ for $t \in [0, T]$ and $u(t) = \Gamma w(t)$ for $t > T$, then u is admissible by Assumption A4 and it follows from (13) that $z_2(T) = 0$ and $z_2(t) = 0$ for all $t > T$. Therefore, $(z_{20}, w_0) \in \mathcal{R}_{g_2}(T)$. On the other hand, if $(z_{20}, w_0) \in \mathcal{R}_{g_2}(T)$, then there exists an admissible u such that $z_2(T) = 0$. Also by (13), we have

$$z_{20} - V(T)w_0 + \int_0^T e^{-A_2\tau}B_2u(\tau)d\tau = 0,$$

which implies that $z_{20} - V(T)w_0 \in \mathcal{C}_2(T)$.

ii) Since A_2 is anti-stable and S is stable, we have that $\lim_{T\to\infty} V(T) = V_2$. It follows from (12) that

$$V_2 = \int_0^\infty e^{-A_2\tau}B_2\Gamma e^{S\tau}d\tau. \tag{14}$$

First we show that $z_{20} - V_2w_0 \in \mathcal{C}_2 \implies (z_{20}, w_0) \in \mathcal{R}_{g_2}$. Since $\mathcal{C}_2$ is open, there exists an $\varepsilon > 0$ such that $\{z_{20} - V_2w_0 + z_2 : |z_2|_\infty \le \varepsilon\} \subset \mathcal{C}_2$. Also, there exists a $T_1 > 0$ such that $\{z_{20} - V_2w_0 + z_2 : |z_2|_\infty \le \varepsilon\} \subset \mathcal{C}_2(T_1)$. Since $\lim_{T\to\infty} V(T) = V_2$, there is a $T_2 > T_1$ such that $z_{20} - V(T_2)w_0 \in \mathcal{C}_2(T_1) \subset \mathcal{C}_2(T_2)$. It follows from i) that $(z_{20}, w_0) \in \mathcal{R}_{g_2}(T_2) \subset \mathcal{R}_{g_2}$.

Next we show that $(z_{20}, w_0) \in \mathcal{R}_{g_2} \implies z_{20} - V_2w_0 \in \mathcal{C}_2$. If $(z_{20}, w_0) \in \mathcal{R}_{g_2}$, then $(z_{20}, w_0) \in \mathcal{R}_{g_2}(T_1)$ for some $T_1 > 0$. It follows from the definition of $\mathcal{R}_g(T)$ that there is an admissible control u_1 such that

$$z_{20} - \int_0^{T_1} e^{-A_2\tau}B_2\Gamma e^{S\tau}w_0 d\tau + \int_0^{T_1} e^{-A_2\tau}B_2u_1(\tau)d\tau = 0. \tag{15}$$

Denote $Z_2 = \left\{\delta e^{-A_2T_1}v_2 : v_2 \in \tilde{\mathcal{C}}_2\right\}$. For each $z_2 \in Z_2$, there is an admissible control u_2 such that

$$z_2 = \delta e^{-A_2T_1}\int_0^\infty e^{-A_2\tau}B_2u_2(\tau)d\tau = \int_{T_1}^\infty e^{-A_2\tau}B_2\delta u_2(\tau - T_1)d\tau.$$

It follows from (14) and (15) that

$$\begin{aligned} z_{20} - V_2 w_0 + z_2 &= z_{20} - \int_0^\infty e^{-A_2\tau} B_2 \Gamma e^{S\tau} w_0 d\tau + \int_{T_1}^\infty e^{-A_2\tau} B_2 \delta u_2(\tau - T_1) d\tau \\ &= z_{20} - \int_0^{T_1} e^{-A_2\tau} B_2 \Gamma e^{S\tau} w_0 d\tau + \int_{T_1}^\infty e^{-A_2\tau} B_2 \left(\delta u_2(\tau - T_1) - \Gamma e^{S\tau} w_0 \right) d\tau \\ &= -\int_0^{T_1} e^{-A_2\tau} B_2 u_1(\tau) d\tau + \int_{T_1}^\infty e^{-A_2\tau} B_2 \left(\delta u_2(\tau - T_1) - \Gamma e^{S\tau} w_0 \right) d\tau \\ &\in \bar{\mathcal{C}}_2. \end{aligned}$$

The last step follows from the fact that $|\Gamma e^{S\tau} w_0|_\infty \leq \rho = 1 - \delta$ for all $\tau > 0$ and that the input $u(t) = u_1(t)$ for $t \in [0, T_1], u(t) = \delta u_2(t - T_1) - \Gamma e^{St} w_0$ for $t > T_1$ is admissible. This implies that

$$\{z_{20} - V_2 w_0 + z_2 : \ z_2 \in Z_2\} \subset \bar{\mathcal{C}}_2.$$

Since $e^{-A_2 T_1}$ is nonsingular, the set Z_2 contains the origin in its interior. It follows that $z_{20} - V_2 w_0 \in \mathcal{C}_2$.

iii) We first show that $\mathcal{R}^a_{g2} = \mathcal{R}_{g2}$. It is easy to see that $\mathcal{R}_{g2} \subset \mathcal{R}^a_{g2}$. To show $\mathcal{R}^a_{g2} \subset \mathcal{R}_{g2}$, suppose we are given $z_{20} \in \mathcal{R}^a_{g2}$. Then there exits a finite time T and an admissible control such that $z_2(T) \in \delta\mathcal{C}_2$ since $\delta\mathcal{C}_2$ is an open set containing the origin. For $t > T$, let $u = \Gamma w + u_\delta$ with $\|u_\delta\|_\infty \leq \delta$. Since $w_0 \in \mathcal{W}_0$, we have $\|\Gamma w\|_\infty \leq \rho$ and hence, $\|u\|_\infty \leq 1$ and

$$\dot{z}_2 = A_2 z_2 + B_2 u - B_2 \Gamma w = A_2 z_2 + B_2 u_\delta.$$

Since $z_2(T) \in \delta\mathcal{C}_2$, we have a control u_δ and a finite $T_1 > T$ such that $z_2(T_1) = 0$. So we have $z_{20} \in \mathcal{R}_{g2}$, and hence $\mathcal{R}^a_{g2} \subset \mathcal{R}_{g2}$. Therefore, $\mathcal{R}^a_{g2} = \mathcal{R}_{g2}$.

Now we show that $\mathcal{R}^a_g \subset \mathbf{R}^{n_1} \times \mathcal{R}_{g2}$. Suppose $(z_0, w_0) = (z_{10}, z_{20}, w_0) \in \mathcal{R}^a_g$, then there exists an admissible control such that $\lim_{t\to\infty} z_2(t) = 0$. This implies that $(z_{20}, w_0) \subset \mathcal{R}^a_{g2} = \mathcal{R}_{g2}$.

We next show that $\mathbf{R}^{n_1} \times \mathcal{R}_{g2} \subset \mathcal{R}^a_g$. Suppose $(z_{10}, z_{20}, w_0) \in \mathbf{R}^{n_1} \times \mathcal{R}_{g2}$, then there exist a $T \geq 0$ and an admissible control such that $z_2(T) = 0$. Since $\begin{bmatrix} z_1(T) \\ 0 \end{bmatrix}$ is inside the asymptotically null controllable region for the system $\dot{z} = Az + B u_\delta$ under the constraint $\|u_\delta\|_\infty \leq \delta$ (Hájek, 1991), there exists a $u = \Gamma w + u_\delta$ for $t > T$ such that $\lim_{t\to\infty} z(t) = 0$. Hence $(z_0, w_0) \in \mathcal{R}^a_g$. This establishes that $\mathbf{R}^{n_1} \times \mathcal{R}_{g2} \subset \mathcal{R}^a_g$ and hence $\mathbf{R}^{n_1} \times \mathcal{R}_{g2} = \mathcal{R}^a_g$. □

Remark 1 *Here we justify the requirement of driving $z(t)$, instead of $e(t)$, to zero in Defintion 1. From the above theorem, we see that (z_0, w_0) is regulatable if and only if $z_{20} - V_2 w_0 \in \mathcal{C}_2$. This is essentially the necessary condition to keep $z_2(t)$ bounded. We explain this as follows: Suppose that $z_{20} - V_2 w_0 \notin \bar{\mathcal{C}}_2$, then there exists an $\varepsilon > 0$ such that $|z_{20} - V_2 w_0 - z_2|_2 > \varepsilon \ \forall \ z_2 \in \bar{\mathcal{C}}_2$. Since*

$$\lim_{T\to\infty} \int_0^T e^{-A_2\tau} B_2 \Gamma e^{S\tau} w_0 d\tau = V_2 w_0,$$

and

$$-\int_0^T e^{-A_2\tau} B_2 u(\tau) d\tau \in \mathcal{C}_2, \quad \forall\, T > 0, \ \ |u|_\infty \leq 1,$$

there exists a $T_0 > 0$ *such that*

$$\left| z_{20} - \int_0^T e^{-A_2\tau} B_2 \Gamma e^{S\tau} w_0 d\tau + \int_0^T e^{-A_2\tau} B_2 u(\tau) d\tau \right|_2 > \frac{\varepsilon}{2}, \quad \forall\, T > T_0.$$

Since the smallest singular value of $e^{A_2 T}$ *increases exponentially with* T*, it follows from (11) that* $z_2(T)$ *will grow unbounded whatever admissible control is applied. Hence even if the requirement* $\lim_{t\to\infty} z(t) = 0$ *is replaced with* $\lim_{t\to\infty} e(t) = 0$*, we still require* $z_{20} - V_2 w_0 \in \bar{C}_2$ *to achieve output regulation. The gap only arises from the boundary of* C_2*. It is unclear whether it is possible to achieve* $\lim_{t\to\infty} e(t) = 0$ *with* $z(t)$ *bounded for* $z_{20} - V_2 w_0 \in \partial C_2$*. Since this problem involves too much technical detail and is of little practical importance (we will not take the risk to allow* $z_{20} - V_2 w_0 \in \partial C_2$*, otherwise a small perturbation can cause the state to grow unbounded), we will not address this subtle technical point here.*

4 State Feedback Controller

In view of what has been discussed in the previous section, the set of initial conditions for which output regulation can be achieved with any feedback law must be a subset of $\mathcal{R}_g^a$. In this section, we would like to search for a state feedback law $u = g(z, w)$ such that this subset is as close as possible to $\mathcal{R}_g^a$.

With a state feedback $u = g(z, w), |g(z, w)|_\infty \leq 1$ for all $(z, w) \in \mathbf{R}^n \times \mathcal{W}_0$, we have the closed-loop system

$$\begin{aligned} \dot{z} &= Az + B\, g(z, w) - B\Gamma w, \\ \dot{w} &= Sw. \end{aligned} \tag{16}$$

Suppose that the system $\dot{z} = Az + Bg(z, 0)$ is asymptotically stable at $z = 0$. Denote the time response $z(t)$ of (16) to the initial state (z_0, w_0) as $z(t, z_0, w_0)$ and define

$$\mathcal{S}_{zw} := \left\{ (z_0, w_0) \in \mathbf{R}^n \times \mathcal{W}_0 : \lim_{t\to\infty} z(t, z_0, w_0) = 0 \right\}.$$

We see that $\mathcal{S}_{zw}$ is the set of initial states for which output regulation is achieved by $u = g(z, w)$. Also, the intersection of $\mathcal{S}_{zw}$ with the z space serves as the domain of attraction for the system $\dot{z} = Az + Bg(z, 0)$, where internal stability is achieved.

It is clear that $\mathcal{S}_{zw} \subset \mathcal{R}_g^a$. Our objective is to search for a function $g(\cdot, \cdot)$ such that $\dot{z} = Az + Bg(z, 0)$ is asymptotically stable and $\mathcal{S}_{zw}$ is as large as possible, or as close as possible to $\mathcal{R}_g^a$.

We assume that a stabilizing state feedback law $u = f(v), |f(v)|_\infty \leq 1$ for all $v \in \mathbf{R}^n$, has been designed and the equilibrium $v = 0$ of the closed-loop system

$$\dot{v} = Av + Bf(v) \tag{17}$$

has a domain of attraction $\mathcal{S} \subset \mathcal{C}^a$. We will construct an output regulation law from this stabilizing feedback law.

Furthermore, we make the assumption, that will be removed later, that there exists a matrix $V \in \mathbf{R}^{n\times r}$ such that

$$-AV + VS = -B\Gamma. \tag{18}$$

This will be the case if A and S have no common eigenvalues. With the decomposition in (6), if we partition V accordingly as $\begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$, then V_2 satisfies $-A_2V_2 + V_2S = -B_2\Gamma$.

Denote

$$D_{zw} := \{(z,w) \in \mathbf{R}^n \times \mathcal{W}_0 : z - Vw \in \mathcal{S}\}. \tag{19}$$

From Theorem 1 and $\mathcal{C}^a = \mathbf{R}^{n_1} \times \mathcal{C}_2$, we see that

a) The set D_{zw} increases as $\mathcal{S}$ increases, and if $\mathcal{S} = \mathcal{C}^a$, then $D_{zw} = \mathcal{R}_g^a$;

b) In the absence of w, $x_0 \in \mathcal{S} \Rightarrow (z_0, 0) \in D_{zw}$.

We will construct an output regulation law from the stabilizing feedback law $u = f(v)$ in such a way that $D_{zw} \subset \mathcal{S}_{zw}$. It is then clear that if f is chosen such that $\mathcal{S} = \mathcal{C}^a$, then $D_{zw} = \mathcal{R}_g^a$ and $\mathcal{S}_{zw} = \mathcal{R}_g^a$. Our controller construction is based on a series of technical lemmas. In the sequel, a set D is said to be *invariant* for a system if every trajectory of the system starting from D will stay in D.

Lemma 1 *Let $u = f(z - Vw)$. Consider the closed-loop system*

$$\begin{aligned} \dot{z} &= Az + Bf(z - Vw) - B\Gamma w, \\ \dot{w} &= Sw. \end{aligned} \tag{20}$$

For this system, D_{zw} is an invariant set and for all $(z_0, w_0) \in D_{zw}$, $\lim_{t\to\infty}(z(t) - Vw(t)) = 0$.

Proof. Because of (18), we can replace $-B\Gamma$ in (20) with $-AV + VS$, i.e.,

$$\begin{aligned} \dot{z} &= Az + Bf(z - Vw) - AVw + VSw \\ &= A(z - Vw) + Bf(z - Vw) + V\dot{w}. \end{aligned}$$

Define the new state $v := z - Vw$, we have

$$\dot{v} = Av + Bf(v),$$

which has a domain of attraction $\mathcal{S}$. This also implies that $\mathcal{S}$ is an invariant set for the v-system.

If $(z_0, w_0) \in D_{zw}$, then $v_0 = z_0 - Vw_0 \in \mathcal{S}$. It follows that $v(t) = z(t) - Vw(t) \in \mathcal{S}$ for all $t \geq 0$, which means that D_{zw} is invariant, and $\lim_{t\to\infty}(z(t) - Vw(t)) = \lim_{t\to\infty} v(t) = 0$. □

Lemma 2 *Suppose that $D \subset \mathbf{R}^n$ is an invariant set in the domain of attraction for the system*

$$\dot{v} = Av + B\phi(v), \tag{21}$$

then for any $\alpha > 0$, αD is an invariant set in the domain of attraction for the system

$$\dot{v} = Av + \alpha B\phi\left(\frac{v}{\alpha}\right). \tag{22}$$

Proof. Write (22) as

$$\frac{\dot{v}}{\alpha} = A\frac{v}{\alpha} + B\phi\left(\frac{v}{\alpha}\right),$$

and replace $\frac{v}{\alpha}$ with v, then we get (21). If $v_0 \in \alpha D$, i.e., $\frac{v_0}{\alpha} \in D$, then $\frac{v(t)}{\alpha} \in D$ for all $t > 0$ and $\lim_{t\to\infty} v(t) = 0$. □

Lemma 1 says that, in the presence of w, the feedback $u = f(z - Vw)$ will cause $z(t) - Vw(t)$ to approach zero and $z(t)$ to approach $Vw(t)$, which is bounded. Our next step is to construct a finite sequence of controllers $u = f_k(z, w, \alpha)$, $k = 0, 1, 2, \cdots, N$, all parameterized in $\alpha \in (0, 1)$. By judiciously switching between these controllers, we can cause $z(t)$ to approach $\alpha^k Vw(t)$ for any k. By choosing N large enough, $z(t)$ will become arbitrarily small in a finite time. Once $z(t)$ becomes small enough, we will use the controller $u = \Gamma w + \delta\,\mathrm{sat}(\frac{Fz}{\delta})$ (F to be specified below) to make $z(t)$ converge to the origin.

Let $F \in \mathbf{R}^{m\times n}$ be such that

$$\dot{v} = Av + B\,\mathrm{sat}(Fv) \tag{23}$$

is asymptotically stable. Let $X > 0$ be such that $(A + BF)^{\mathrm{T}}X + X(A + BF) < 0$ and

$$\mathcal{E} := \{v \in \mathbf{R}^n : v^{\mathrm{T}}Xv \le 1\} \subset \{v \in \mathbf{R}^n : \ |Fv|_\infty \le 1\}.$$

Then $\mathcal{E}$ is an invariant set in the domain of attraction for the closed-loop system (23).

For any $\alpha \in (0, 1)$, there exists a positive integer N such that

$$\alpha^N |X^{\frac{1}{2}}Vw|_2 < \delta, \quad \forall\, w \in \mathcal{W}_0, \tag{24}$$

i.e., $v = \alpha^N Vw \in \delta\mathcal{E}$, for all $w \in \mathcal{W}_0$. Define a sequence of subsets in $\mathbf{R}^n \times \mathcal{W}_0$ as,

$$\begin{aligned} D_{zw}^k &= \left\{(z, w) \in \mathbf{R}^n \times \mathcal{W}_0 : z - \alpha^k Vw \in \alpha^k\mathcal{E}\right\}, \ \ k = 0, 1, \cdots, N, \\ D_{zw}^{N+1} &= \{(z, w) \in \mathbf{R}^n \times \mathcal{W}_0 : z \in \delta\mathcal{E}\}, \end{aligned}$$

and, on each of these sets, define a state feedback law as follows,

$$\begin{aligned} f_k(z, w, \alpha) &= (1 - \alpha^k)\Gamma w + \alpha^k\,\mathrm{sat}\left(\frac{F(z - \alpha^k Vw)}{\alpha^k}\right), \ \ k = 0, 1, \cdots, N, \\ f_{N+1}(z, w) &= \Gamma w + \delta\,\mathrm{sat}\left(\frac{Fz}{\delta}\right). \end{aligned}$$

It can be verified that, for any $k = 0$ to $N + 1$, $|f_k(z, w, \alpha)|_\infty \le 1$ for all $(z, w) \in \mathbf{R}^n \times \mathcal{W}_0$.

Lemma 3 *Let $u = f_k(z, w, \alpha)$. Consider the closed-loop system*

$$\begin{aligned} \dot{z} &= Az + Bf_k(z, w, \alpha) - B\Gamma w, \\ \dot{w} &= Sw. \end{aligned} \tag{25}$$

For this system, D_{zw}^k is an invariant set. Moreover, if $k = 0, 1, \cdots, N$, then for all $(z_0, w_0) \in D_{zw}^k$, $\lim_{t\to\infty}(z(t) - \alpha^k Vw(t)) = 0$; if $k = N + 1$, then, for all $(z_0, w_0) \in D_{zw}^{N+1}$, $\lim_{t\to\infty} z(t) = 0$.

Proof. With $u = f_k(z, w, \alpha), k = 0, 1, \cdots, N$, we have

$$\begin{aligned} \dot{z} &= Az + (1-\alpha^k)B\Gamma w + \alpha^k B\text{sat}\left(\frac{F(z-\alpha^k Vw)}{\alpha^k}\right) - B\Gamma w \\ &= Az + \alpha^k B\text{sat}\left(\frac{F(z-\alpha^k Vw)}{\alpha^k}\right) - \alpha^k B\Gamma w. \end{aligned} \tag{26}$$

Let $v_k = z - \alpha^k Vw$, then by (18),

$$\dot{v}_k = Av_k + \alpha^k B\text{sat}\left(\frac{Fv_k}{\alpha^k}\right). \tag{27}$$

It follows from Lemma 2 that $\alpha^k\mathcal{E}$ is an invariant set in the domain of attraction for the v_k-system. Hence D_{zw}^k is invariant for the system (25) and if $(z_0, w_0) \in D_{zw}^k$, i.e., $v_{k0} = z_0 - \alpha^k Vw_0 \in \alpha^k\mathcal{E}$, then $\lim_{t\to\infty}(z(t) - \alpha^k Vw(t)) = \lim_{t\to\infty} v_k(t) = 0$.

With $u = f_{N+1}(z, w)$, we have $\dot{z} = Az + \delta B\text{sat}\left(\frac{Fz}{\delta}\right)$ and the same argument applies. □

For $k = 0, 1, \cdots, N+1$, denote $\Omega^k = \cup_{j=k}^{N+1} D_{zw}^j$. We clearly have $\Omega^0 \supset \Omega^1 \cdots \supset \Omega^N \supset \Omega^{N+1}$. Based on the technical lemmas established above, we construct our state feedback law as follows,

$$u = g(z, w, \alpha, N) = \begin{cases} f_{N+1}(z, w), & \text{if } (z, w) \in \Omega^{N+1}, \\ f_k(z, w, \alpha), & \text{if } (z, w) \in \Omega^k \setminus \Omega^{k+1}, \ k = 0, 1, \cdots, N, \\ f(z - Vw), & \text{if } (z, w) \in \mathbf{R}^n \times \mathcal{W}_0 \setminus \Omega^0. \end{cases}$$

Since Ω^k, $k = 0, 1, \cdots, N+1$ are nested, the controller is well defined on $\mathbf{R}^n \times \mathcal{W}_0$. What remains to be shown is that this controller will accomplish our objective if the parameter α is properly chosen. Let

$$\alpha_0 = \max_{w\in\mathcal{W}_0} \frac{|X^{\frac{1}{2}}Vw|_2}{|X^{\frac{1}{2}}Vw|_2 + 1}.$$

It is obvious that $\alpha_0 \in (0, 1)$.

Theorem 2 *Choose any $\alpha \in (\alpha_0, 1)$ and let N be chosen such that (24) is satisfied. Then for all $(z_0, w_0) \in D_{zw}$, the time response of the closed-loop system*

$$\begin{aligned} \dot{z} &= Az + B\,g(z, w, \alpha, N) - B\Gamma w, \\ \dot{w} &= Sw \end{aligned} \tag{28}$$

satisfies $\lim_{t\to\infty} z(t) = 0$, i.e., $D_{zw} \subset S_{zw}$.

Here we note that the control $u = g(z, w, \alpha, N)$ is executed by choosing one from $f_k(z, w, \alpha), k = 0, 1, \cdots, N+1$, and $f(z - Vw)$. Since the control is discontinuous in z and w, it is necessary to ensure that the trajectories are well behaved and no chattering will ocur. We will generalize the safe switching idea of Wredenhagen and Belanger (1994) to guarantee this. Actually, the crucial point behind the safe switching in Wredenhagen and Belanger (1994) is that each of the nested ellipsoids, on which the switching is based, is invariant. Here under the control of $u = g(z, w, \alpha, N)$, we also

have a group of nested invariant sets, Ω^k, $k = 0, 1, \cdots, N+1$. But these sets are not ellipsoids, nor necessarily convex.

Proof of Theorem 2. We first claim that for each $k = 0, 1, \cdots, N+1$, the set Ω^k is an invariant set under the control of $u = g(z, w, \alpha, N)$, and every trajectory starting from Ω^k will enter a subset Ω^{k+1} at a finite time.

We prove the claim by induction. By Lemma 3, we know that $\Omega^{N+1} = D_{zw}^{N+1}$ is an invariant set. Hence, the claim is true for $k = N+1$. We assume that it is true for $k+1$ and need to show that it is also true for k.

Here we are given $(z_0, w_0) \in \Omega^k$. If $(z_0, w_0) \in \Omega^{k+1}$, then by assumption, $(z(t), w(t)) \in \Omega^{k+1} \subset \Omega^k$ for all $t > 0$. If $(z_0, w_0) \in \Omega^k \setminus \Omega^{k+1}$, then $(z_0, w_0) \in D_{zw}^k$ and $u = f_k(z, w, \alpha)$. By Lemma 3, $(z(t), w(t)) \in D_{zw}^k$ for all $t > 0$ and $\lim_{t\to\infty}(z(t) - \alpha^k V w(t)) = 0$. Since $\alpha \in (\alpha_0, 1)$, we have $(1-\alpha)|X^{\frac{1}{2}} V w|_2 < \alpha$ for all $w \in \mathcal{W}_0$. Therefore, for $k < N$,

$$\begin{aligned} |X^{\frac{1}{2}}(z - \alpha^{k+1} V w)|_2 &\leq |X^{\frac{1}{2}}(z - \alpha^k V w)|_2 + \alpha^k(1-\alpha)|X^{\frac{1}{2}} V w|_2 \\ &< |X^{\frac{1}{2}}(z - \alpha^k V w)|_2 + \alpha^{k+1}. \end{aligned} \tag{29}$$

Since the first term goes to zero asymptotically, there exists a finite time $t_1 > 0$ such that

$$|X^{\frac{1}{2}}(z(t_1) - \alpha^{k+1} V w(t_1))|_2 \leq \alpha^{k+1}.$$

This implies that $z(t_1) - \alpha^{k+1} V w(t_1) \in \alpha^{k+1}\mathcal{E}$, i.e., $(z(t_1), w(t_1)) \in D_{zw}^{k+1}$. If $k = N$, then, by (24),

$$\begin{aligned} |X^{\frac{1}{2}} z|_2 &\leq |X^{\frac{1}{2}}(z - \alpha^N V w)|_2 + \alpha^N |X^{\frac{1}{2}} V w|_2 \\ &< |X^{\frac{1}{2}}(z - \alpha^N V w)|_2 + \delta. \end{aligned}$$

Also, the first term goes to zero asymptotically, so there exists a finite time t_1 such that $|X^{\frac{1}{2}} z(t_1)|_2 \leq \delta$, i.e., $(z(t_1), w(t_1)) \in D_{zw}^{N+1}$. The state $(z(t), w(t))$ might have entered some other set $D_{zw}^{k+l}, l > 1$, at an earlier time than t_1. But in any case, it will enter Ω^{k+1} at a finite time under the control of $u = f_k(z, w, \alpha)$. Thus we have completed the proof of the claim.

Let us now consider $(z_0, w_0) \in D_{zw}$. If $(z_0, w_0) \in \Omega^0$, then by the foregoing claim, we have $(z(t), w(t)) \in \Omega^0$ for all $t > 0$. If $(z_0, w_0) \in D_{zw} \setminus \Omega^0$, then $u = f(z - Vw)$ and by Lemma 1, $(z(t), w(t)) \in D_{zw}$ and $\lim_{t\to\infty}(z(t) - Vw(t)) = 0$. Hence there is a finite time such that $z(t) - Vw(t) \in \mathcal{E}$, i.e., $(z(t), w(t)) \in D_{zw}^0$. Similarly, $(z(t), w(t))$ might have entered some other sets $D_{zw}^l, l > 0$, but in any case, it will enter the set Ω^0 at a finite time.

We know that the sets $\Omega^k, k = 0, 1, 2, \cdots, N+1$ are nested and each of them is an invariant set under the control $u = g(z, w, \alpha, N)$. Moreover, every trajectory starting from D_{zw} will enter Ω^0 and subsequently into smaller and smaller Ω^k and finally enter Ω^{N+1} at a finite time. After that we have $u = f_{N+1}(z, w)$ and by Lemma 3, $(z(t), w(t))$ will remain there and $\lim_{t\to\infty} z(t) = 0$.

Here we have a final remark on the behavior of the trajectories of (28). Since $\Omega^k, k = 0, 1, \cdots, N+1$ are nested and each is an invariant set, a trajectory starting from the boundary of Ω^k will stay inside Ω^k and will never go back to Ω^{k-1}. This guarantees that no chattering will occur and there are at most $N+2$ switches. And between two switches, the trajectory is

uniquely determined by the place where the first switch takes place. Hence for all initial states in D_{zw}, there is a unique trajectory of (28). □

Remark 2 *Note that our controller $u = g(z, w, \alpha, N)$ is constructed from a nonlinear controller $u = f(v)$ and a linear controller $u = Fx$. The nonlinear controller is used to keep $z - Vw$ bounded and it determines the region of output regulation. The linear controller is used to drive z closer and closer to zero. We may also replace the linear controller with a nonlinear one such as $f(v)$ and replace $\mathcal{E}$ with a general level set. The reason we have used a linear controller is for simplicity of both analysis and the switching scheme.*

From the proof of Theorem 2, we see that for all $(z_0, w_0) \in D_{zw}$, the number of switches is at most $N + 2$. To reduce the complexity of the controller, we would like to choose N as small as possible. As we can see from (24), choosing smaller $\alpha < 1$ will enable smaller N.

For better understanding, we illustrate the proof of Theorem 2 with Fig. 1. The sets D_{zw}^k for the simplest case where both z and w are one dimensional are plotted. Here we have $X = V = 1$, $\alpha = 0.6$, $\delta = 0.2$ and $N = 3$. The parallelogram bounded by the straight lines L_k (along with the two vertical lines $w = \pm 1$) is $D_{zw}^k = \left\{(z, w) : |z - \alpha^k w| \leq \alpha^k, |w| \leq 1\right\}, k = 0, 1, 3, 4$. The dotted line passing through the origin is in parallel to the line L_0. Suppose that $(z_0, w_0) \in \Omega^0 \subset D_{zw}^0$, then under the control $u = f_0(z, w, \alpha)$, (z, w) will converge to the dotted line. Since α is chosen such that this dotted line is inside D_{zw}^1 (ref. (29)), (z, w) will enter D_{zw}^1 in a finite time. After that, the controller will be switched to $f_1(z, w, \alpha)$ and so on. Finally (z, w) will enter D_{zw}^4 and $z(t)$ will go to zero. If z is two dimensional, we will have cylinders as D_{zw}^k instead of parallelograms.

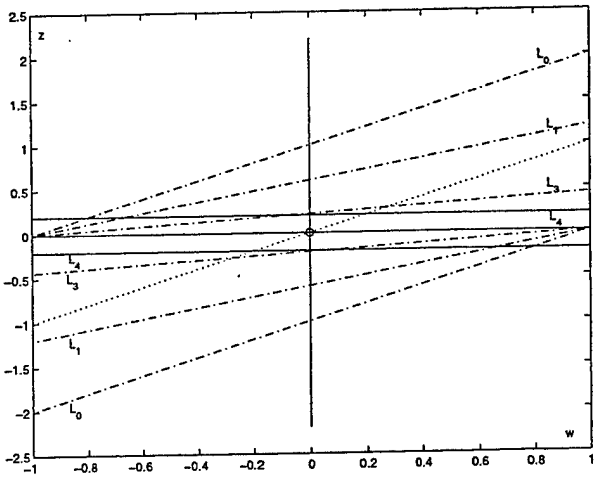


Figure 1: Illustration for the proof of Theorem 2

In what follows, we will deal with the case that there is no V satisfying $-AV + VS = -B\Gamma$. This will occur if A and S have some same eigenvalues on the imaginary axis. The following method is also useful in the case that some eigenvalues of A and S on the imaginary axis are very close. This will result in large elements of V, so α could be very close to 1 and N could be very large.

Suppose that the z-system (4) has the following form,

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u - \begin{bmatrix} B_1\Gamma \\ B_2\Gamma \end{bmatrix} w \tag{30}$$

with $A_1 \in \mathbf{R}^{n_1 \times n_1}$ semi-stable and $A_2 \in \mathbf{R}^{n_2 \times n_2}$ anti-stable. Also suppose that there is a known function $f(v_2), |f(v_2)|_\infty \leq 1$ for all $v_2 \in \mathbf{R}^{n_2}$ such that the origin of the following system

$$\dot{v}_2 = A_2 v_2 + B_2 f(v_2)$$

has a domain of attraction $\mathcal{S}_2$, which is a bounded set. Then by Lemma 2, the system

$$\dot{v}_2 = A_2 v_2 + \delta B_2 f\left(\frac{v_2}{\delta}\right)$$

has a domain of attraction $\delta\mathcal{S}_2$; and by Teel (1996) there exists a control $u = \delta\,\mathrm{sat}(h(v))$ such that the origin of

$$\dot{v} = Av + \delta B \mathrm{sat}(h(v))$$

has a domain of attraction $\mathcal{S}_\delta = \mathbf{R}^{n_1} \times \delta\mathcal{S}_2$.

Since there is a V_2 satisfying $-A_2V_2 + V_2S = -B_2\Gamma$, by the foregoing development of Theorem 2, there exists a controller $u = g(z_2, w, \alpha, N)$ such that for any (z_0, w_0) satisfying $z_{20} - V_2 w_0 \in \mathcal{S}_2$, $w_0 \in \mathcal{W}_0$, the response of the closed-loop system satisfies $\lim_{t\to\infty} z_2(t) = 0$. Hence there is a finite time $t_1 > 0$ such that $z(t_1) \in \mathbf{R}^{n_1} \times \delta\mathcal{S}_2$. After that if we switch to the control $u = \Gamma w + \delta\,\mathrm{sat}(h(z)$, we will have

$$\dot{z} = Az + \delta B \mathrm{sat}(h(z))$$

and z will stay in $\mathcal{S}_\delta = \mathbf{R}^{n_1} \times \delta\mathcal{S}_2$. To avoid the possibility that the state z_2 might be trapped on the boundary of $\delta\mathcal{S}_2$, we need to modify the controller further such that it will not switch before z_2 has entered $\gamma\delta\mathcal{S}_2$ for some $\gamma \in (0,1)$. For this purpose, we introduce a switching state variable $s(z_2(t))$:

$$s(z_2(t)) = \begin{cases} 0, & \text{if } z_2 \in \mathbf{R}^{n_2} \setminus \delta\mathcal{S}_2, \\ 1, & \text{if } z_2 \in \gamma\delta\mathcal{S}_2, \\ s(z_2(t^-)), & \text{if } z_2 \in \delta\mathcal{S}_2 \setminus \gamma\delta\mathcal{S}_2, \end{cases}$$

where $s(z_2(t)) = s(z_2(t^-))$ means that s does not change value at t. We may simply assume that $s(0) = 0$. Now the modified controller is,

$$u = \begin{cases} g(z_2, w, \alpha, N), & \text{if } z_2 \in \mathbf{R}^{n_2} \setminus \gamma\delta\mathcal{S}_2 \text{ and } s = 0, \\ \Gamma w + \delta\,\mathrm{sat}(h(z)), & \text{if } z_2 \in \delta\mathcal{S}_2 \text{ and } s = 1. \end{cases}$$

Since $g(z_2, w, \alpha, N)$ has $N+2$ switches, the modified controller has $N+3$ switches. We conclude that under this control, the following set

$$D_{zw} = \{(z, w) \in \mathbf{R}^n \times \mathcal{W}_0 : z_2 - V_2 w \in \mathcal{S}_2\}$$

will be a subset of $\mathcal{S}_{zw}$. The argument goes as follows. Given $(z_0, w_0) \in D_{zw}$, without loss of generality assume that $z_{20} \in \mathbf{R}^{n_2} \setminus \delta\mathcal{S}_2$, then $s(0) = 0$ and $u = g(z_2, w, \alpha, N)$ will be in effect and

by Theorem 2 we have $\lim_{t\to\infty} z_2(t) = 0$. So there is a finite time $t_1 > 0$ such that $z_2(t_1) \in \delta\mathcal{S}_2$. After t_1 we still have $s(z_2(t)) = 0$ so the controller will not switch until $z_2(t_2) \in \gamma\delta\mathcal{S}_2$ for some $t_2 > t_1$. After t_2, we have $s(z_2(t)) = 1$ and the controller will switch to $\Gamma w + \delta\,\mathrm{sat}(h(z))$. Under this control $\mathbf{R}^{n_1} \times \delta\mathcal{S}_2$ is an invariant set and a domain of attraction. So $z_2(t)$ will stay in $\delta\mathcal{S}_2$, $s(z_2(t))$ stays at 1 and we have $\lim_{t\to\infty} z(t) = 0$. Here we also have, if $\mathcal{S}_2 = \mathcal{C}_2$, then $D_{zw} = \mathcal{R}_g^a$.

5 Error Feedback

Consider again the open loop system (4). Here in this section, we assume that only the error $e = Cz$ is available for feedback. Also, without loss of generality, assume that the pair

$$(\overline{C}, \overline{A}) = \left([C\ \ 0], \begin{bmatrix} A & -B\Gamma \\ 0 & S \end{bmatrix}\right)$$

is observable. If it is detectable, but not observable, then the unobservable modes must be the asymptotically stable eigenvalues of A, which do not affect the output regulation (see (4)) and hence can be left out.

Our controller consists of an observer and a given state feedback law $u = f(v)$. Because of the observer error, we need an additional assumption on $f(v)$ so that some class of disturbances can be tolerated. Consider the system

$$\dot{v} = Av + Bf(v + \eta), \tag{31}$$

where η stands for the disturbance arising from, for example, the observer error. Assume that $|f(v)|_\infty \le 1$ for all $v \in \mathbf{R}^n$ and that there exist a set $D_0 \subset \mathbf{R}^n$ and positive numbers γ and d_0 such that the solution of the system satisfies

$$\|v\|_\infty \le \gamma \max\left(|v_0|_\infty, \|\eta\|_\infty\right), \quad \|v\|_a \le \gamma\|\eta\|_a, \quad \forall\, v_0 \in D_0,\ \|\eta\|_\infty \le d_0,$$

where $\|v\|_a = \limsup_{t\to\infty} |v(t)|_\infty$. This system is said to satisfy an asymptotic bound from D_0 with gain γ and restriction d_0 (Teel, 1996). In Hu and Lin (2001b), a saturated linear feedback $u = f(v) = \mathrm{sat}(F_0 v)$ with such property is constructed for second order anti-stable systems. Moreover, the set D_0 can be made arbitrarily close to the null controllable region $\mathcal{C}$.

Let $D \in \mathbf{R}^n$ be in the interior of D_0, i.e., the distance from any point in D to the boundary of D_0 is greater than a fixed positive number. Given a number M, denote

$$D_M = \left\{(z, w, \tilde{z}, \tilde{w}) \in \mathbf{R}^n \times \mathcal{W}_0 \times \mathbf{R}^{n+r} :\ z - Vw \in D, \left\|\begin{bmatrix} \tilde{z} \\ \tilde{w} \end{bmatrix}\right\| \le M\right\},$$

where $(\tilde{z}, \tilde{w})$ denotes the observer error. It is clear that the set D_M increases as D increases. From Theorem 1, we see that if $D = \mathcal{C}^a$, then the projection of D_M to the $(z, w)-$subspace equals to $\mathcal{R}_g^a$. To enlarge the set of initial conditions where output regulation is achieved, it suffices to construct a state feedback law $u = f(v)$ to enlarge D_0, choose a set D very close to D_0 and design an observer such that for all $(z_0, w_0, \tilde{z}_0, \tilde{w}_0) \in D_M$, $\lim_{t\to\infty} z(t) = 0$. The objective in this section is to construct such an observer given $u = f(v)$, D_0 and D in the interior of D_0.

We use the following observer to reconstruct the states z and w,

$$\begin{aligned} \dot{\bar{z}} &= A\bar{z} + Bu - B\Gamma\bar{w} - L_1(e - C\bar{z}), \\ \dot{\bar{w}} &= S\bar{w} - L_2(e - C\bar{z}). \end{aligned} \tag{32}$$

Letting $\tilde{z} = z - \bar{z}, \tilde{w} = w - \bar{w}$, we can write the composite system as

$$\begin{aligned} \dot{z} &= Az + Bu - B\Gamma w, \\ \dot{w} &= Sw, \\ \begin{bmatrix} \dot{\tilde{z}} \\ \dot{\tilde{w}} \end{bmatrix} &= \begin{bmatrix} A + L_1C & -B\Gamma \\ L_2C & S \end{bmatrix} \begin{bmatrix} \tilde{z} \\ \tilde{w} \end{bmatrix}. \end{aligned} \tag{33}$$

Now we have to use $(\bar{z}, \bar{w})$ instead of (z, w) to construct a feedback controller. Since $(\overline{C}, \overline{A})$ is observable, we can choose $L = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix}$ appropriately such that the estimation error $(\tilde{z}, \tilde{w})$ decays arbitrarily fast. Moreover, the following fact is easy to establish,

Lemma 4 *Denote*

$$\tilde{A} = \begin{bmatrix} A + L_1C & -B\Gamma \\ L_2C & S \end{bmatrix}.$$

Given any (arbitrarily small) positive numbers T and ε, there exists an L such that

$$\max\left\{\left|e^{\tilde{A}t}\right|, \ |L| \cdot \left|e^{\tilde{A}t}\right|\right\} \leq \varepsilon, \quad \forall\, t \geq T,$$

where $|\cdot|$ can be any matrix norm.

Because of this lemma, it is expected that the controller based on the observer can achieve almost the same performance as the state feedback controller.

Consider system (33). For simplicity, we also assume that (18) is satisfied. Letting $v = z - Vw$, we obtain $\dot{v} = Av + Bu$. Suppose that $(z_0, w_0, \tilde{z}_0, \tilde{w}_0) \in D_M$, then $v_0 = z_0 - Vw_0 \in D$. Since D is in the interior of D_0, there exists a $T_0 > 0$ such that, with any admissible control u, we have

$$v(t) = z(t) - Vw(t) \in D_0, \quad \forall\, t \leq T_0. \tag{34}$$

What we are going to do is to choose L such that the estimation error is sufficiently small after T_0, and to design switching feedback laws to make $\bar{z}(t) \to \alpha^k V\bar{w}(k)$ with increasing k and finally drive $\bar{z}(t)$ and $z(t)$ to the origin.

Lemma 5 *There exists an $L \in \mathbf{R}^{(n+r)\times q}$ such that, under the control*

$$u = \begin{cases} u_1(t), & t < T_0, \\ f(\bar{z} - V\bar{w}), & t \geq T_0, \end{cases}$$

where u_1 is any admissible control, the solution of system (33) satisfies

$$\lim_{t\to\infty}(z(t) - Vw(t)) = 0, \quad \forall\, (z_0, w_0, \tilde{z}_0, \tilde{w}_0) \in D_M.$$

Proof. Let $v = z - Vw$. Since u_1 is an admissible control, we have (34). In particular,

$$v(T_0) = z(T_0) - Vw(T_0) \in D_0. \tag{35}$$

Let $\tilde{v} = \tilde{z} - V\tilde{w}$, by Lemma 4, there exists an $L \in \mathbf{R}^{(n+r)\times q}$ such that for all $(z_0, w_0, \tilde{z}_0, \tilde{w}_0) \in D_M$,

$$|\tilde{v}(t)|_\infty = |\tilde{z}(t) - V\tilde{w}(t)|_\infty \le d_0, \quad \forall\, t \ge T_0. \tag{36}$$

We now consider the system after T_0. For $t \ge T_0$, we have $u = f(\bar{z} - V\bar{w})$ and the closed-loop system is

$$\dot{v} = Av + Bf(\bar{z} - V\bar{w}) = Av + Bf(v - \tilde{v}), \quad v(T_0) \in D_0.$$

By assumption, this system satisfies an asymptotic bound from D_0 with a finite gain and restriction d_0. It follows from (35), (36) and $\lim_{t\to\infty} \tilde{v}(t) = 0$, that $\lim_{t\to\infty}(z(t) - Vw(t)) = \lim_{t\to\infty} v(t) = 0$. □

Lemma 5 means that we can keep $z(t)$ bounded if $(z_0, w_0, \tilde{z}_0, \tilde{w}_0) \in D_M$. Just as the state feedback case, we want to move $z(t)$ to the origin gradually by making $z(t) - \alpha^k Vw(t)$ small with increased k. Due to the switching nature of the controller and that the feedback has to be based on $(\bar{z}, \bar{w})$, we need to construct a sequence of sets which are invariant with respect to $(\bar{z}, \bar{w})$ rather than (z, w) under the corresponding controllers.

Using linear system theory, it is easy to design an $F \in \mathbf{R}^{m\times n}$, along with a matrix $X > 0$ such that $A + BF$ is Hurwitz, that $\mathcal{E} = \{v \in \mathbf{R}^n : v^{\mathrm{T}} X v \le 1\} \subset \{v \in \mathbf{R}^n : \ |Fv|_\infty \le 1\}$ and that for some positive number d_1, $\mathcal{E}$ is invariant for the system

$$\dot{v} = Av + B\mathrm{sat}(Fv) - \eta, \quad |\eta|_\infty \le d_1. \tag{37}$$

Let α and N be determined from X in the same way as with the state feedback controller. With $F \in \mathbf{R}^{m\times n}$, we form a sequence of controllers,

$$u = f_k(\bar{z}, \bar{w}, \alpha) = (1 - \alpha^k)\Gamma\bar{w} + \alpha^k \mathrm{sat}\left(\frac{F(\bar{z} - \alpha^k V\bar{w})}{\alpha^k}\right), \quad k = 0, 1, \cdots, N,$$

and

$$u = f_{N+1}(\bar{z}, \bar{w}) = \Gamma\bar{w} + \delta\, \mathrm{sat}\left(\frac{F\bar{z}}{\delta}\right).$$

Under the control $u = f_k(\bar{z}, \bar{w}, \alpha)$, consider $v_k = \bar{z} - \alpha^k V\bar{w}$, then we get

$$\dot{v}_k = Av_k + \alpha^k B\, \mathrm{sat}\left(\frac{Fv_k}{\alpha^k}\right) - (L_1 - \alpha^k V L_2)C\tilde{z}. \tag{38}$$

Note the difference between this equation and the corresponding (27) in the state feedback case. Here we need to take into account the extra term $(L_1 - \alpha^k V L_2)C\tilde{z}$. For clarity, we split the discussion into three cases.

Case 1. $k = 0$. Let $v = \bar{z} - V\bar{w}$. Then the system

$$\dot{v} = Av + B\mathrm{sat}(Fv) - (L_1 - VL_2)C\tilde{z},$$

has an invariant set $\mathcal{E}$ for all $\tilde{z}$ satisfying

$$|(L_1 - VL_2)C\tilde{z}|_\infty \leq d_1. \tag{39}$$

Since $\lim_{t\to\infty} \tilde{z}(t) = 0$, we also have $\lim_{t\to\infty} v(t) = 0 \ \ \forall v_0 \in \mathcal{E}$.

Case 2. $0 < k \leq N$. Similar to Lemma 3, we have an invariant set $\alpha^k\mathcal{E}$ for the system (38) under the restriction

$$\left|(L_1 - \alpha^k VL_2)C\tilde{z}\right|_\infty \leq \alpha^k d_1. \tag{40}$$

Also, because $\lim_{t\to\infty} \tilde{z}(t) = 0$, we have $\lim_{t\to\infty} v_k(t) = 0$ for all $v_0 \in \alpha^k\mathcal{E}$.

Case 3. $k = N + 1$. We have an invariant set $\delta\mathcal{E}$ for the system

$$\dot{\tilde{z}} = A\tilde{z} + \delta B \,\mathrm{sat}\left(\frac{F\tilde{z}}{\delta}\right) - L_1 C\tilde{z}$$

under the restriction

$$|L_1 C\tilde{z}|_\infty \leq \delta d_1. \tag{41}$$

Now we define a sequence of sets in $\mathbf{R}^n \times \mathbf{R}^r$:

$$D^k_{\tilde{z}\bar{w}} = \left\{(\bar{z},\bar{w}) \in \mathbf{R}^n \times \mathbf{R}^r : \bar{z} - \alpha^k V\bar{w} \in \alpha^k\mathcal{E}\right\}, \quad k = 0,1,2,\cdots,N,$$

and

$$D^{N+1}_{\tilde{z}\bar{w}} = \left\{(\bar{z},\bar{w}) \in \mathbf{R}^n \times \mathbf{R}^r : \bar{z} \in \delta\mathcal{E}\right\}.$$

A counterpart of Lemma 3 is,

Lemma 6 *Suppose that F is chosen such that $\mathcal{E}$ is invariant for the system (37) and that L and $\tilde{z}_0$ are such that $\tilde{z}$ satisfies all the conditions (39), (40) and (41). Then, under the control $u = f_k(\bar{z},\bar{w},\alpha)$, the set $D^k_{\tilde{z}\bar{w}}$ is invariant for the system (32) and* $\lim_{t\to\infty}\left(\bar{z}(t) - \alpha^k V\bar{w}(t)\right) = 0$.

With the preliminaries we developed above, we can now construct an error feedback law as follows:

$$u = g(\bar{z},\bar{w},\alpha,N) = \begin{cases} f_{N+1}(\bar{z},\bar{w}), & \text{if } (\bar{z},\bar{w}) \in \Omega^{N+1}, \\ f_k(\bar{z},\bar{w},\alpha), & \text{if } (\bar{z},\bar{w}) \in \Omega^k \setminus \Omega^{k+1}, \ \ k = 0,1,\ldots,N, \\ f(\bar{z} - V\bar{w}), & \text{if } (\bar{z},\bar{w}) \in \mathbf{R}^n \times \mathbf{R}^r \setminus \Omega^0. \end{cases} \tag{42}$$

where $\Omega^k = \cup_{j=k}^{N+1} D^j_{\tilde{z}\bar{w}}$, $k = 0,1,\ldots,N+1$.

Suppose that the controller (42) is connected to system (33). At $t = T_0$, we have $z(T_0) - Vw(T_0) \in D_0$ (cf. (34)). To apply Lemma 6 for $t \geq T_0$, we have to ensure that all the conditions (39), (40) and (41) are satisfied. Moreover, we also require that

$$|f_k(\bar{z},\bar{w},\alpha)|_\infty \leq 1, \quad |f_{N+1}(\bar{z},\bar{w})|_\infty \leq 1. \tag{43}$$

Note that

$$|f_k(\bar{z},\bar{w},\alpha)|_\infty = \left|(1-\alpha^k)\Gamma\bar{w} + \alpha^k \mathrm{sat}\left(\frac{F(\bar{z} - \alpha^k V\bar{w})}{\alpha^k}\right)\right|_\infty \leq 1$$

can be satisfied if

$$|\Gamma\bar{w}|_\infty \le |\Gamma w|_\infty + |\Gamma\tilde{w}|_\infty \le 1. \tag{44}$$

Let d_0 and d_1 be given, then by Lemma 4, all the conditions (36), (39), (40), (41) and (44) can be satisfied for $t \ge T_0$ by suitably choosing L. Note that, in (44), $|\Gamma w|_\infty \le \rho < 1$.

Theorem 3 *Consider the feedback system (33) with (42). Let $f(\cdot)$ be chosen such that the system (31) satisfy an asymptotic bound from D_0 with finite gain and restriction d_0. Let D be a set in the interior of D_0 and T_0 be chosen such that (34) is satisfied. Let F be chosen such that $\mathcal{E}$ is invariant for the system (37) and let L be chosen such that the conditions (36), (39), (40), (41) and (44) are satisfied for $t \ge T_0$. Then for all $(z_0, w_0, \tilde{z}_0, \tilde{w}_0) \in D_M$, $\lim_{t\to\infty} z(t) = 0$.*

Proof. *Case 1.* $(\bar{z}(T_0), \bar{w}(T_0)) \in \mathbf{R}^n \times \mathbf{R}^r \setminus \Omega^0$. Then $u = f(\bar{z} - V\bar{w})$ will be in effect. This is just the situation described in Lemma 5. So we have $\lim_{t\to\infty}(\bar{z}(t) - V\bar{w}(t)) = 0$ and similar to the proof of Theorem 2, it can be shown that $(\bar{z}, \bar{w})$ will enter Ω^0, or some other Ω^k at a finite time. *Case 2.* $(\bar{z}(T_0), \bar{w}(T_0)) \in \Omega^k \setminus \Omega^{k+1}$. Since $\tilde{z}(t), t \ge T_0$ satisfies the conditions (39), (40) and (41), each $D^k_{\bar{z}\bar{w}}$ is invariant under the control $u = f_k(\bar{z}, \bar{w}, \alpha)$. Since $\lim_{t\to\infty} \tilde{z}(t) = 0$, we have (see 38) $\lim_{t\to\infty}\left(\bar{z}(t) - \alpha^k V\bar{w}(t)\right) = \lim_{t\to\infty} v_k(t) = 0$. Similar to the proof of Theorem 2, it can be shown that $(\bar{z}, \bar{w})$ will enter Ω^{k_1} for some $k_1 > k$.

By repeating this procedure, we will have $(\bar{z}, \bar{w}) \in \Omega^{N+1}$ at some finite time, and hence $\lim_{t\to\infty} \bar{z}(t) = 0$. Also, since $\lim_{t\to\infty} \tilde{z}(t) = 0$, we have $\lim_{t\to\infty} z(t) = 0$. □

6 Example

In this section, we will apply the results developed above to the control of an aircraft model. Consider the longitudinal dynamics of the TRANS3 aircraft under certain flight condition (Junkins, Valasek and Ward, 1999),

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx, \end{aligned} \tag{45}$$

with

$$A = \begin{bmatrix} 0 & 14.3877 & 0 & -31.5311 \\ -0.0012 & -0.4217 & 1.0000 & -0.0284 \\ 0.0002 & -0.3816 & -0.4658 & 0 \\ 0 & 0 & 1.0000 & 0 \end{bmatrix}, \quad B = 0.1745 \times \begin{bmatrix} 4.526 \\ -0.0337 \\ -1.4566 \\ 0 \end{bmatrix},$$

and $C = [1\ \ 0\ \ 0\ \ 0]$, where the state consists of the velocity x_1 (feet/s, relative to the nominal flight condition), the angle of attack x_2 (degree), the pitch rate x_3 (degree/s) and the Euler angle rotation of aircraft about the inertial y-axis x_4 (degree), the control u (degree) is the elevator input, whose value is scaled to between ± 1 (corresponding to $\pm 10^o$). The design objective is to reject disturbances Pw, where

$$P = \begin{bmatrix} -0.6526 & -0.3350 & 0.4637 & 0.9185 \\ 0.0049 & 0.0025 & -0.0035 & -0.0068 \\ 0.2100 & 0.1078 & -0.1492 & -0.2956 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

and w contains the frequencies of 0.1 rad/s and 0.3 rad/s. Clearly, this problem can be cast into an output regulation problem for the system (1) with

$$S = \text{diag}\left\{\begin{bmatrix} 0 & -0.1 \\ 0.1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -0.3 \\ 0.3 & 0 \end{bmatrix}\right\}$$

and $Q = 0$. A solution to the linear matrix equations (2) are

$$\Pi = 0, \quad \Gamma = \begin{bmatrix} 0.8263 & 0.4242 & -0.5871 & -1.1630 \end{bmatrix}.$$

Assume that the disturbances are bounded by $\|\Gamma w\|_\infty \leq \rho = 0.9$. Thus $\delta = 0.1$. Since $\Pi = 0$, we have $z = x - \Pi w = x$.

The matrix A has two stable eigenvalues $-0.4650 \pm 0.6247i$ and two anti-stable ones, $0.0212 \pm 0.1670i$. With state transformation, we get the matrices for the anti-stable subsystem:

$$A_2 = \begin{bmatrix} 0.0212 & 0.1670 \\ -0.1670 & 0.0212 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 8.2856 \\ -2.4303 \end{bmatrix}.$$

We don't need to worry about the exponentially stable z_1-subsystem since its state is bounded under any bounded input $u - \Gamma w$ and will converge to the origin as the combined input goes to zero. We now investigate two types of controllers.

Case 1. The state feedback.

With the design method in Hu and Lin (2001a), we obtain a semi-global stabilizer $u = f(z_2) = \text{sat}(\beta F_0 z_2)$, $F_0 = \begin{bmatrix} -0.0093 & 0.0041 \end{bmatrix}$ for the system in the absence of disturbance. With $\beta = 1$, the achieved domain of attraction is very close to the null controllable region (see Fig. 2, where the outermost solid close curve is the boundary of $\mathcal{C}_2$ and the dotted closed curve is the boundary of the domain of attraction $\mathcal{S}_2$).

With the method proposed in Section 3, we take $F = F_0$ and the switching parameters be $\alpha = 0.4$ and $N = 3$. So there are at most 5 switches in the controller $u = g(z_2, w, \alpha, 3)$. The closed-loop system is simulated with $z_{20} - V_2 w_0$ very close to the boundary of $\mathcal{S}_2$ (see the point marked with "o" in Fig. 2). The trajectory shown in the figure is that of $z_2(t)$ with the initial state marked with "*". We note here that the initial state z_{20} is not always inside $\mathcal{S}_2$, not even inside the null controllable region. But $z_{20} - V_2 w_0$ has to be in the null controllable region, i.e., (z_{20}, w_0) has to be in the regulatable region $\mathcal{R}_{g2}$.

The output $e = Cz$ is plotted in Fig. 3. The control (solid curve) and the switching history (dash-dotted curve) are plotted in Fig. 4, where for the dash-dotted curve, the number -0.2 indicates the control $u = f(z_2 - V_2 w) = \text{sat}(F(z_2 - V_2 w))$ is applied and the number $0.2k$, indicates that the controller $u = f_k(z_2, w, \alpha)$ is applied, $k = 0, 1, \cdots, 4$, respectively. We see that the controller $u = f_3(z_2, w, \alpha)$ is skipped. This is because that the set Ω^{k+1} as a subset of Ω^k may share some bounary points with Ω^k (see Fig. 1).

Case 2. The error feedback

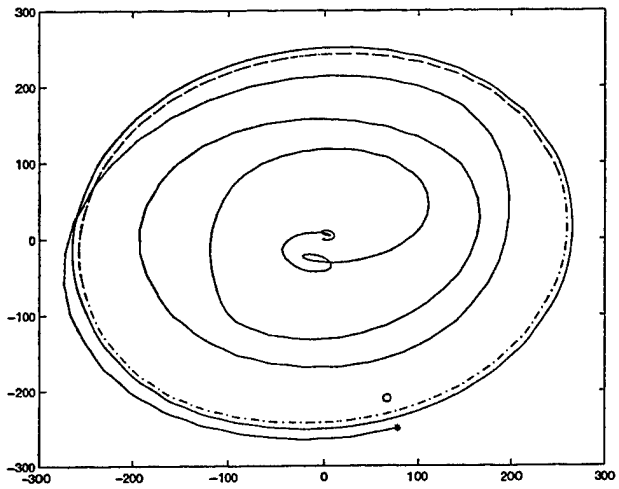


Figure 2: A trajectory of z_2 – Case 1

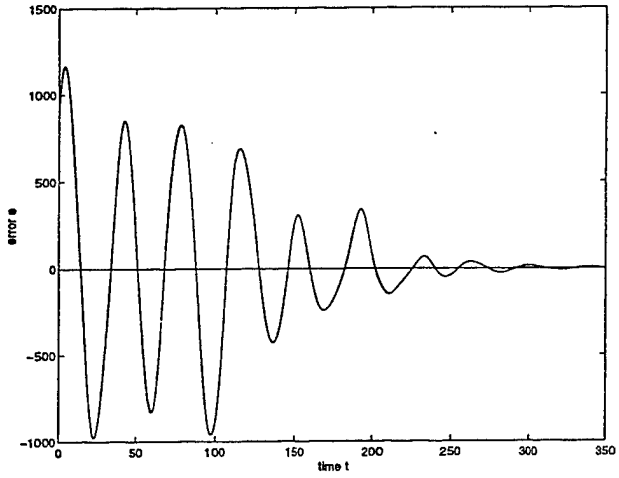


Figure 3: The time response of the error – Case 1

Here we assume that only the error signal is available. So we need to build an observer to reconstruct the plant and exosystem states. For this system, $\left(\begin{bmatrix} C & 0 \end{bmatrix}, \begin{bmatrix} A & -B\Gamma \\ 0 & S \end{bmatrix} \right)$ is observable. With the design method in Hu and Lin (2001b), we obtain a controller $u = f(v_2)$=sat$(F_0 v_2)$, $F_0 = \begin{bmatrix} -0.0175 & 0.0103 \end{bmatrix}$ such that the system $\dot{v}_2 = A_2 v_2 + B_2 f(v_2 + \eta)$ satisfies an asymptotic bound from D_0 with a finite gain and nonzero restriction, where D_0 is also very close to the null controllable region (see Fig. 5, where the outermost solid close curve is the boundary of $\mathcal{C}_2$ and the dotted closed curve is the boundary of D_0).

With the error feedback design method proposed in this paper, we obtain for the anti-stable sub-system $F = [-0.0378 \quad 0.0357]$ and the switching parameters $\alpha = 0.5$ and $N = 3$. Also, to achieve fast recovery of the state, we choose

$$L = [24.11 \quad -263.61 \quad 68.59 \quad -115.96 \quad 161070.21 \quad -231551.67 \quad 65778.10 \quad -3347.56]^{\mathrm{T}},$$

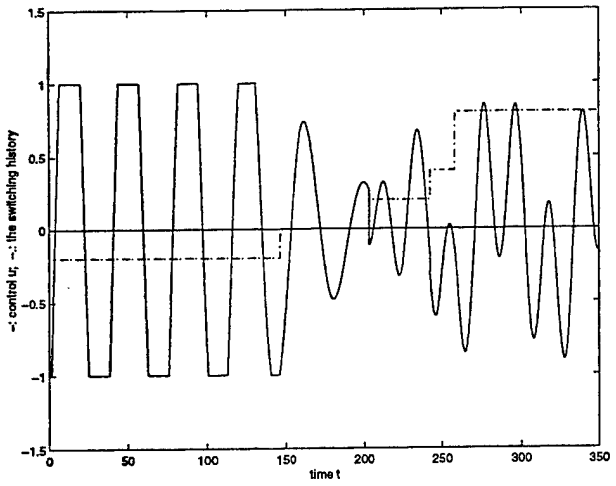


Figure 4: The control and the switching history — Case 1

for the observer.

The closed-loop system is simulated with $z_{20} - Vw_0$ very close to the boundary of D_0, see the point marked with "o" in Fig. 5. We simply set the initial states of the observer be 0. The trajectory shown in the figure is that of $z_2(t)$, with the initial point marked with '*'.

The error output $e = Cz$ is plotted in Fig. 6 and the observer error is plotted in Fig. 7. The control and the switching history is plotted in Fig. 8. The simulation results verify the effectiveness of the design method proposed in this paper.

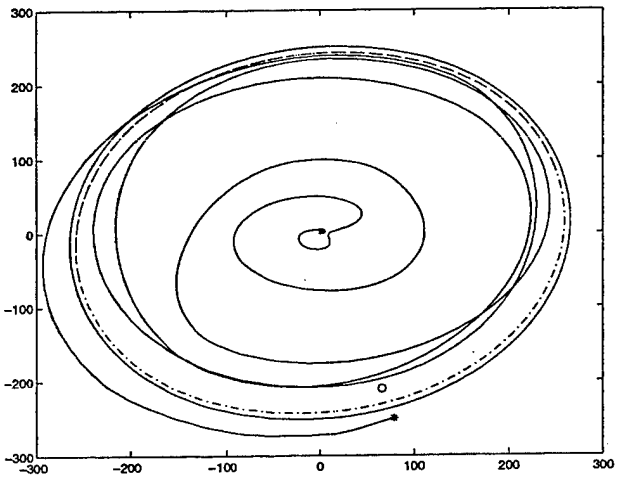


Figure 5: The trajectory of z_2 — Case 2

7 Conclusions

In this paper, we have systematically studied the problem of output regulation for linear systems subject to actuator saturation. The plants considered here are general and can be exponentially

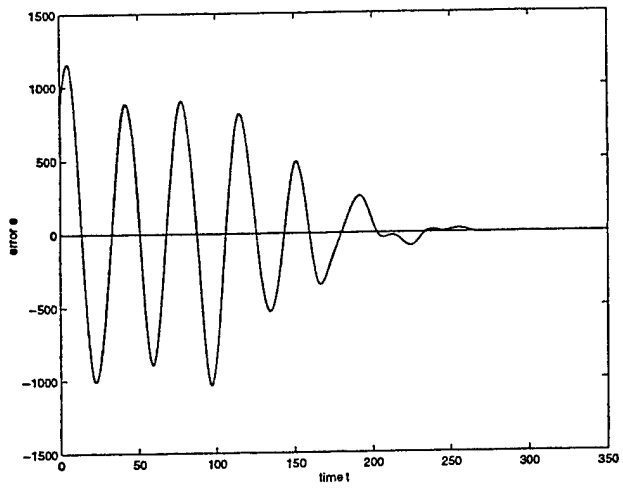


Figure 6: The error output $e = Cz$ — Case 2

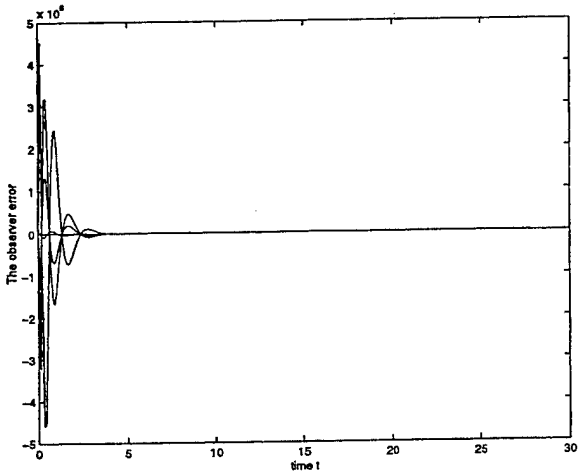


Figure 7: The observer error $\tilde{z}$ and $\tilde{w}$ — Case 2

unstable. We first characterized the regulatable region, the set of initial conditions of the plant and the exosystem for which output regulation can be achieved. Based on a given stabilizing state feedback law, we then constructed a state feedback law and an error feedback law, that achieve output regulation on a subset of the regulatable region. The size of this subset depends on the domain of attraction under the given stabilizing state feedback law.

References

[1] [Casavola and Mosca, 2001] Casavola A. and Mosca E. (2001). Global switching regulation of input-saturated discrete-time linear systems with arbitrary l(2) disturbances. *IEEE Transaction on Automatic Control*, **AC-46**, 915-919.

[2] [Blanchini and Miani, 2000] Blanchini, F. and Miani, S. (2000). Any domain of attraction

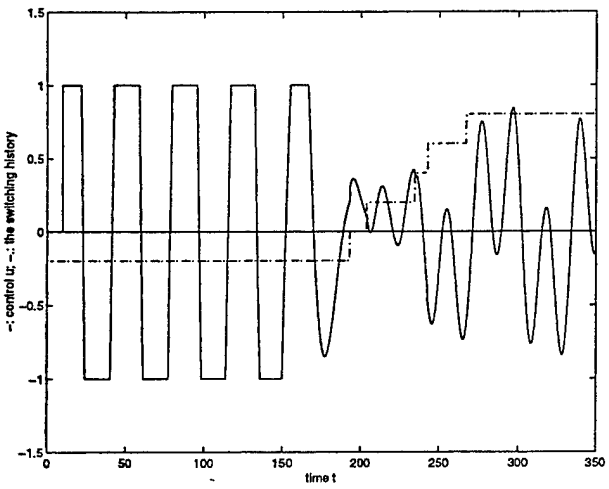


Figure 8: The control and the switching history — Case 2

for a linear constrained system is a tracking domain of attarction. *SIAM J. Control and Optimization*, **38**, pp971-994.

[3] [De Santis, 2000] De Santis, (2000). Output regulation for linear systems with anti-stable eigenvalues in the presence of input saturation. *Int. J. Robust and Nonlinear Control*, **10**, 423-438.

[4] [De Santis and Isidori, 2001] De Santis, R. and Isidori, A. (2001). On the output regulation for linear systems in the presence of input saturation. *IEEE Transaction on Automatic Control*, **AC-46**, 156-160.

[5] [Francis, 1975] Francis, B. A. (1975). The linear multivariable regulator problem. *SIAM J. Cont. and Opt.*. **15**, 486-505.

[6] [Hájek, 1991] Hájek, O. (1991). *Control Theory in the Plane*, Springer-Verlag.

[7] [Hu and Lin, 2001a] Hu, T. and Lin, Z. (2001a). *Control Systems with Actuator Saturation: Analysis and Design*. Birkhäuser, Boston.

[8] [Hu and Lin, 2001b] Hu, T. and Lin, Z. (2001b). Practical stabilization of exponentially unstable linear systems subject to actuator saturation nonlinearity and disturbance. *Int. J. of Robust and Nonlinear Control.* **11**, 555-588.

[9] [Hu, Lin and Shamash, 2001] Hu, T., Lin, Z. and Shamash, Y. (2001). Semi-global stabilization with guaranteed regional performance of linear systems subject to actuator saturation. *Systems & Control Letters*, **43**, 203-210.

[10] [Isidori and Byrnes, 1990] Isidori, A. and Byrnes, C. I. (1990). Output regulation for nonlinear systems. *IEEE Trans. Automatic Control.* **35**, 131-140.

[11] [Junkins, Valasek and Ward, 1999] Junkins, J., Valasek, J. and Ward, D. (1999) *Report of ONR UCAV Modeling Effort*, Dept. of Aerospace Engineering, Texas A&M University.

[12] [Lin, 1998] Lin, Z. (1998). *Low Gain Feedback.* Springer-Verlag, London.

[13] [Lin, Pachter and Banda, 1998] Lin, Z., Pachter, M. and Banda, S. (1998) Toward improvement on tracking performance – nonlinear feedback for linear systems. *Int. J. of Control*, **70**, 1-11.

[14] [Lin and Saberi, 1993] Lin, Z. and Saberi, A. (1993). Semi-global exponential stabilization of linear systems subject to 'input saturation' via linear feedbacks. *Systems & Control Letters*, **21**, 225-239.

[15] [Lin, Stoorvogel and Saberi, 1996] Lin, Z., Stoorvogel, A. A. and Saberi, A. (1996). Output regulation for linear systems subject to input saturation. *Automatica.* **32**, 29-47.

[16] [Macki and Strauss, 1982] Macki, J. and Strauss, M. (1982). *Introduction to Optimal Control*, Springer-Verlag.

[17] [Saberi, Stoorvogel and Sannuti, 1999] Saberi, A., Stoorvogel, A. and Sannuti, P. (1999). Output regulation and control problems with regulation constraints. Springer-Verlag.

[18] [Sontag, 1984] Sontag, E. D. (1984). An algebraic approach to bounded controllability of linear systems. *Int. J. Control*, **39**, 181-188.

[19] [Suarez, Alvarez-Ramirez and Solis-Daun, 1997] Suarez, R., Alvarez-Ramirez, and J. Solis-Daun, J. (1997). Linear systems with bounded inputs: global stabilization with eigenvalue placement. *Int. J. Robust and Nonlinear Control*, **7**, 835-845.

[20] [Sussmann, Sontag, and Yang, 1994] Sussmann, H. J., Sontag, E. D. and Yang, Y. (1994). A general result on the stabilization of linear systems using bounded controls. *IEEE Trans. Automatic Control.* **AC-39**, 2411-2425.

[21] [Tarbouriech, Pittet and Burgat, 2000] Tarbouriech, S., Pittet, C. and Burgat, C. (2000). Output tracking problem for systems with input saturations via nonlinear integrating actions. *Int. J. of Robust and Nonlinear Control.* **10**, 489-512.

[22] [Teel, 1992] Teel, A. R. (1992). *Feedback Stabilization: Nonlinear Solutions to Inherently Nonlinear Problems*, Ph.D dissertation, Berkeley, CA.

[23] [Teel, 1996] Teel, A. R. (1996). A nonlinear small gain theorem for the analysis of control systems, *IEEE Trans. Automatic Control.* **AC-42**, 1256-1270.

[24] [Turner, Postlethwaite and Walker, 2000] Turner, M. C., Postlethwaite I. and Walker D. J. (2000). Non-linear tracking control for multivariable constrained input linear systems, *Int. J. of Control.* **73**, 1160-1172.

[25] [Wredenhagen and Belanger, 1994] Wredenhagen, G. F. and P.R. Belanger, P. R. (1994). Piecewise-linear LQ control for systems with input constraints. *Automatica.* **30**, 403-416.

Publication 28

Submitted to *IEEE Transactions on Signal Processing*

Robust Filtering for Discrete-time Systems with Saturation and Its Application to Transmultiplexers

Yufei Xiao Yong-Yan Cao Zongli Lin

Department of Electrical & Computer Engineering
University of Virginia
Charlottesville, VA 22903

March 26, 2002

Abstract

This paper considers the problem of robust filtering for discrete-time linear systems subject to saturation. A generalized dynamic filter architecture is proposed and a filter design method is developed. Our approach incorporates the conventional linear H_2 and H_∞ filtering as well as a regional l_2 gain filtering feature developed specially for the saturation nonlinearity, and is applicable to the digital transmultiplexer systems for the purpose of separating filter bank design. It turns out that our filter design can be carried out by solving a constrained optimization problem with LMI constraints. Simulation shows that the resultant separating filters possess satisfactory reconstruction performance while working in the linear range, and less degraded reconstruction performance in the presence of saturation.

EDICS: 2-FILB

[1]This work was supported in part by the US Office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

1 Introduction

Digital filtering systems are inherently subject to nonlinear effects such as adder overflow and amplitude saturation [7, 19]. These nonlinearities may introduce limit cycles to the system. As seen in the literature, efforts have been dedicated to the stability issue where the generalized overflow nonlinearity as described in Fig. 1 is considered (see, *e.g.*, [3, 18]). The saturation nonlinearity brought about by magnitude truncation falls into this category and can be regarded as a special case in which $L = 1$. In [18] and [20], conditions for global asymptotic stability are given in regard to systems with generalized overflow characteristics and systems with partial state saturation, respectively.

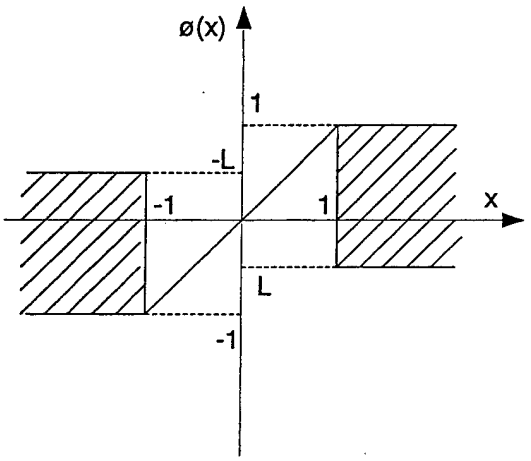


Figure 1: The generalized overflow nonlinearity.

Beyond stability, the second important issue to address is the performance. In particular, when both noise and saturation are present in the digital system, we would like to design a filter such that a certain cost function from the noise to the filtering error is minimized. As far as linear systems are concerned, besides the traditional Kalman filter, the H_2 and H_∞ filtering has also gained popularity [2, 8, 11, 15, 22]. It is noteworthy that some work has been done with general nonlinear systems [9, 23]. Earlier, results on systems represented by state-space models are usually given in the form of Riccati equations (*e.g.*, [9, 23]). In recent years, authors have been casting linear H_2 and H_∞ filter design into constrained optimization problem with linear matrix inequality (LMI) constraints [10, 21, 24], which leads to increased tractability and greater convenience.

In our work, we will treat discrete-time systems with saturation and additive noise in state-space model, propose a dynamic filter structure, and develop a framework in which the filter

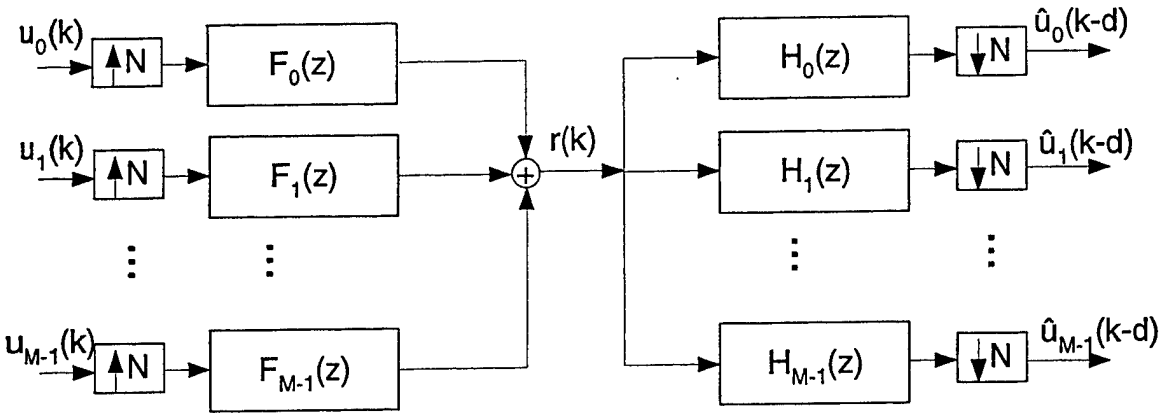


Figure 2: Noise-free TDM-FDM transmultiplexer.

design could be carried out by solving an LMI constrained optimization problem. We will also apply the proposed method to a digital transmultiplexer system to design a separating filter bank.

In communication systems, it is often necessary to transmit several data signals through one common physical medium, as with cellular phone systems, telephone systems, etc. Such necessity gives rise to the transmultiplexer systems. Fig. 2 shows a conventional noise-free multirate transmultiplexer with M bands [16]. M time division multiplexed (TDM) data signals are expanded by a factor of N and processed by a combining filter bank and summed up to become a single frequency division multiplexed (FDM) signal for transmission. The received signal passes through the separating filter bank to become M data signals again, which are decimated, and converted back to M TDM signals. Here the channel is assumed to be ideal in that its transfer function is identity up to some constant gain and fixed time-delay. In such a system, crosstalk may occur due to the decimation and imperfect filters which result in not perfectly reconstructed data. In the noise-free case with ideal channel, crosstalk cancellation is achievable with proper design of the combining/separating filter banks [1, 16]. However, these works did not consider the effect of channel distortion (linear/nonlinear filtering, fading, channel nonlinearities, etc), which is often encountered in communications systems including wireless communication and fiber optic communication. Also, in digital transmultiplexer systems, the quantization error in coding and external disturbances to the channel (both could be viewed as noise) are ubiquitous. Therefore in practice, perfect reconstruction of data is not achievable. To deal with the problem in the presence of linear time-invariant channel distortion and additive noise, in [6] a mixed H_2/H_∞ filtering design scheme is provided in the state space framework

which yields a satisfactory reconstruction performance and is shown to be more robust than the Kalman filter [17]. However, previous works assumed the transmultiplexer system to be working under the linear condition such that the effect of possible amplitude saturation (which could take place either in the power amplifier, within the transmission channel or at the receiver) is ignored. In this work, we will take amplitude saturation into account and design a filter that, in the absence of saturation, minimizes the H_2- and H_∞-norms of the transfer function from the noise to the reconstruction error, and in the presence of actuator saturation, minimizes the regional l_2 gain from the noise to the reconstruction error. We thus refer to our filter design as an hybrid design method which is expected to result in less degraded reconstruction in the presence of both saturation nonlinearity and additive noise.

This paper is organized as follows. Section 2 states the problem as well as reviews some preliminary results on linear filtering. Section 3 presents our hybrid design method with consideration of amplitude saturation. Section 4 includes an application of our proposed hybrid filter design method to the separating filter bank design. Section 5 concludes the paper.

Throughout the paper, we will use the following notation. For a vector $x \in \mathbf{R}^n$, $\|x\|$ denotes its Euclidean norm. Let l_2 denote the set of discrete-time signals such that

$$\sum_{k=0}^{\infty} \|x(k)\|^2 < \infty.$$

For any $x \in l_2$, the l_2-norm is defined as

$$\|x\|_2 = \left(\sum_{k=0}^{\infty} \|x(k)\|^2 \right)^{\frac{1}{2}}.$$

For an asymptotically stable transfer function $G(z)$, we use $\|G\|_2$ and $\|G\|_\infty$ to denote its H_2-norm and H_∞-norm, respectively.

2 Problem Statement and Preliminaries

In this section, we will first precisely state the problem we are to address in this paper and then recall some preliminaries on the linear H_2 and H_∞ filtering.

2.1 Problem Statement

Consider a discrete-time linear time-invariant system in the presence of input noise v, measurement noise w and saturation

$$\begin{aligned} x(k+1) &= Ax(k) + Bv(k), \quad x \in \mathbf{R}^n, v \in \mathbf{R}^m, \\ r(k) &= \text{sat}(Gx(k)) + Ew(k), \quad r \in \mathbf{R}^p, w \in \mathbf{R}^l, \\ u(k) &= Dx(k), \quad u \in \mathbf{R}^q, \end{aligned} \tag{1}$$

where x is the state variable, r is the measured output and u is the signal to be reconstructed by some digital filter. The saturation function sat : $\mathbf{R}^p \to \mathbf{R}^p$ is defined as

$$\text{sat}(u) = [\ \text{sat}(u_1), \text{sat}(u_2), \cdots, \text{sat}(u_m)\ ]^{\text{T}},$$

with $\text{sat}(u_i) = \text{sgn}(u_i)\min\{1, |u_i|\}$. Here we have assumed, without loss of generality, that the saturation level is unity. We have also slightly abused the notation by using sat to denote both the scalar valued and the vector valued saturation functions.

Instead of using an observer based filter structure as in [6],

$$\begin{aligned} \hat{x}(k+1) &= A\hat{x}(k) + K[r(k) - G\hat{x}(k)], \quad \hat{x} \in \mathbf{R}^n, \\ \hat{u}(k) &= D\hat{x}(k), \quad \hat{u} \in \mathbf{R}^q, \end{aligned}$$

we adopt the following more general dynamic filter, to be denoted as $\mathcal{F}(A_f, B_f, D_f)$,

$$\begin{aligned} \hat{x}(k+1) &= A_f\hat{x}(k) + B_f r(k), \quad \hat{x} \in \mathbf{R}^n, \\ \hat{u}(k) &= D_f\hat{x}(k), \quad \hat{u} \in \mathbf{R}^q, \end{aligned} \tag{2}$$

where $\hat{u}(k)$ is the reconstruction of $u(k)$. Define the filtering error (or reconstruction error) as

$$\tilde{u}(k) = u(k) - \hat{u}(k).$$

Introducing a new variable $\bar{x}$ which is composed of x and $\hat{x}$, and a new variable $\bar{w}$ consisting of v and w, the overall system can be written as

$$\begin{aligned} \bar{x}(k+1) &= \bar{A}\bar{x}(k) + \bar{B}\sigma\left(\bar{F}\bar{x}(k)\right) + \bar{E}\bar{w}(k), \\ \tilde{u}(k) &= \bar{C}\bar{x}(k), \end{aligned} \tag{3}$$

where

$$\bar{A} = \begin{bmatrix} A & 0 \\ 0 & A_f \end{bmatrix}, \bar{B} = \begin{bmatrix} 0 \\ B_f \end{bmatrix}, \bar{F} = [\ G \quad 0\ ], \bar{E} = \begin{bmatrix} B & 0 \\ 0 & B_f E \end{bmatrix}, \bar{C} = [\ D \quad -D_f\ ].$$

Note that in the absence of saturation, the system (3) simplifies to

$$\begin{aligned} \bar{x}(k+1) &= \bar{A}_L\bar{x}(k) + \bar{E}\bar{w}(k), \\ \tilde{u}(k) &= \bar{C}\bar{x}(k), \end{aligned} \tag{4}$$

where

$$\bar{A}_L = \begin{bmatrix} A & 0 \\ B_f G & A_f \end{bmatrix}.$$

Our objective is to reconstruct $u(k)$ from the received signal $r(k)$ which is corrupted by additive noise and amplitude saturation, and minimize the reconstruction error $\tilde{u}(k)$ according to certain filtering criterion.

2.2 Linear H_2 and H_∞ Filtering

Consider the discrete-time linear system (1) and the dynamic filter $\mathcal{F}(A_f, B_f, D_f)$ that are both free of saturation. The overall state-space representation is given by (4). The transfer function from the noise $\bar{w} = [v^{\mathrm{T}}, w^{\mathrm{T}}]^{\mathrm{T}}$ to filtering error $\tilde{u}$ is then

$$G(z) = \bar{C}(zI - \bar{A}_L)^{-1}\bar{E}.$$

The objective of filtering design is to make this transfer function as "small" as possible. Two popular measure of the size of a transfer function are its H_2-norm and H_∞-norm.

The H_2-norm of an asymptotically stable and proper discrete-time transfer function $G(z)$ is defined as [5]

$$\|G\|_2 := \left(\frac{1}{2\pi}\mathrm{trace}\left[\int_{-\pi}^{\pi} G(e^{j\omega})^{\mathrm{H}}\, G(e^{j\omega})d\omega\right]\right)^{\frac{1}{2}},$$

where $G^{\mathrm{H}}\,(e^{j\omega})$ is the complex conjugate transpose of $G(e^{j\omega})$. By Parseval Theorem,

$$\|G\|_2 = \left(\mathrm{trace}\left[\sum_{k=0}^{\infty} g(k)g^{\mathrm{T}}(k)\right]\right)^{\frac{1}{2}},$$

where $g(k)$ is the unit impulse response of $G(z)$. Thus, the H_2 norm of a transfer function measures the energy of its unit impulse response.

The H_∞-norm of an asymptotically stable and proper discrete-time transfer function $G(z)$ is defined as

$$\|G\|_\infty := \sup_{\omega\in[0,2\pi]} \sigma_{\max}\left[G(e^{j\omega})\right] = \sup_{\|\omega\|_2=1} \frac{\|h\|_2}{\|w\|_2},$$

where $\sigma_{\max}(G(e^{j\omega})$ is the maximum singular value of $G(e^{j\omega})$, h is the response of $G(z)$ to the input w. Clearly, H_∞ norm of a transfer function is the maximum energy gain from the input signal to the output signal of the system the transfer function represents. In other words, if the transfer function $G(z)$ has an H_∞-norm of γ, then the output h of the system to an input $w \in l_2$ satisfies,

$$\|h\|_2 \leq \gamma\|w\|_2.$$

The H_2 and H_∞ filtering problem can then me formulated as follows. For an *a priori* given $\gamma > 0$, construct a filter $\mathcal{F}(A_f, B_f, D_f)$ such that the resulting transfer function $G(z)$ from the noise $\bar{w}$ to the filtering error $\tilde{u}$ satisfies

$$\|G\|_2 \leq \gamma \quad \text{or} \quad \|G\|_\infty \leq \gamma.$$

We recall the following results on the H_2 and H_∞ filtering problems, respectively.

Lemma 1 [6, 21] *For the discrete-time linear system (4), the H_2-norm of its transfer function $G(z)$ is less than or equal to $\gamma > 0$ if and only if there exists a matrix $R = R^{\mathrm{T}} > 0$ such that*

$$\bar{A}_L R \bar{A}_L^{\mathrm{T}} - R + \bar{E}\bar{E}^{\mathrm{T}} < 0, \tag{5}$$

and

$$\mathrm{trace}(\bar{C}R\bar{C}^{\mathrm{T}}) \leq \gamma^2. \tag{6}$$

Lemma 2 [21] *For the discrete-time linear system (4), the H_∞-norm of its transfer function $G(z)$ is less than or equal to $\gamma > 0$ if and only if there exists a matrix $P = P^{\mathrm{T}} > 0$ such that*

$$\begin{bmatrix} P & 0 & \bar{A}_L^{\mathrm{T}} P & \bar{C}^{\mathrm{T}} \\ 0 & \gamma I & \bar{E}^{\mathrm{T}} P & 0 \\ P\bar{A}_L & P\bar{E} & P & 0 \\ \bar{C} & 0 & 0 & \gamma I \end{bmatrix} > 0. \tag{7}$$

3 A Hybrid Approach to Robust Filtering

3.1 Stability Analysis

The H_2 and H_∞ filtering as recalled in Section 2.2 applies only to linear systems. However when saturation takes place we have to seek other filtering criterion as the performance measure. In what follows, we will introduce the concept of regional l_2 gain filtering and develop the relevant analysis and design procedures.

Consider the following discrete-time system in the presence of amplitude saturation and unknown additive disturbance

$$\begin{array}{rcl} \bar{x}(k+1) & = & \bar{A}\bar{x}(k) + \bar{B}y(k) + \bar{E}\bar{w}(k), \quad \bar{x} \in \mathbf{R}^{2n}, \bar{w} \in \mathbf{R}^{m+l}, \\ y(k) & = & \mathrm{sat}(\bar{F}\bar{x}(k)), \quad y \in \mathbf{R}^p, \\ \tilde{u}(k) & = & \bar{C}\bar{x}(k), \quad \tilde{u} \in \mathbf{R}^q, \end{array} \tag{8}$$

where $\bar{x}$ is the state variable, $\bar{w}$ is an unknown disturbance satisfying $\bar{w}^{\mathrm{T}}\bar{w} \leq 1$, and $\tilde{u}$ is the filtering error. To facilitate our development, we need to recall the following definition.

Definition 1 [14] *A memoryless nonlinearity $\varphi : [0, \infty) \times \mathbf{R}^p \to \mathbf{R}^p$ is said to satisfy a sector condition if*

$$[\varphi(t,y) - K_{\min} y]^{\mathrm{T}} [\varphi(t,y) - K_{\max} y] \leq 0, \quad \forall t \geq 0, \forall y \in \Gamma \subset \mathbf{R}^p, \tag{9}$$

for some real matrices $K_{\min}$ and $K_{\max}$, where $K = K_{\max} - K_{\min}$ is a positive definite symmetric matrix and the interior of Γ is connected and contains the origin.

Assuming that Λ_1 and Λ_2 are diagonal matrices such that $0 \leq \Lambda_1 \leq I \leq \Lambda_2$, and that $(\bar{A} + \Lambda_1 \bar{F})$ is stable, one can write

$$\text{sat}\left(\bar{F}\bar{x}\right) = \Lambda_1 \bar{F}\bar{x} + \Psi_{\bar{F}}(\bar{x}).$$

Define

$$\mathcal{L}(\Lambda_1 \bar{F}) := \{\bar{x} : |(\Lambda_1 \bar{F})_i \bar{x}| \leq 1, \forall i\},$$

in which $(\Lambda_1 \bar{F})_i$ is the ith row of $\Lambda_1 \bar{F}$. Denote $\Lambda := \Lambda_2 - \Lambda_1$, then $\Psi_{\bar{F}}(\bar{x})$ satisfies a sector condition with $K_{\min} = 0$, $K_{\max} = \Lambda \bar{F}$ and $\Gamma = \mathcal{L}(\Lambda_1 \bar{F})$.

Our analysis goal here is to establish a condition under which an ellipsoid $\Omega(P, \rho) = \{\bar{x} \in \mathbf{R}^n : \bar{x}^{\mathrm{T}} P \bar{x} \leq \rho\}$ is invariant. A set is said to be invariant if any trajectory starting from within it will remain inside.

Let P be a positive definite matrix, define the Lyapunov function candidate

$$V(\bar{x}) = \bar{x}^{\mathrm{T}} P \bar{x}.$$

Denote

$$\Delta V(\bar{x}(k)) = V(\bar{x}(k+1)) - V(\bar{x}(k)).$$

In order for the set $\Omega(P, \rho)$ to be invariant, we need to show that

$$\Delta V(\bar{x}(k) \leq 0, \quad \forall \bar{x}(k) \in \partial\Omega(P, \rho),$$

where $\partial\Omega(P, \rho)$ denotes the boundary of the set $\Omega(P, \rho)$. Recalling that $\Psi_{\bar{F}}(\bar{x})$ satisfies a sector condition and

$$\Psi_{\bar{F}}^{\mathrm{T}}(\bar{x})[\Psi_{\bar{F}}(\bar{x}) - \Lambda \bar{F}\bar{x}] \leq 0, \quad \forall \bar{x} \in \mathcal{L}(\Lambda_1 \bar{F}).$$

Assuming that $\Omega(P, \rho) \subset \mathcal{L}(\Lambda_1 \bar{F})$, we obtain the following inequality

$$\begin{aligned}
& V\left(\bar{x}(k+1)\right) \\
\leq\ & \bar{x}^{\mathrm{T}}(k+1) P \bar{x}(k+1) - 2\Psi_{\bar{F}}^{\mathrm{T}}\left(\bar{x}\left(k\right)\right)\left[\Psi_{\bar{F}}\left(\bar{x}\left(k\right)\right) - \Lambda \bar{F}\bar{x}\left(k\right)\right] \\
=\ & \left[\left(\bar{A} + \bar{B}\Lambda_1 \bar{F}\right)\bar{x}(k) + \bar{B}\Psi_{\bar{F}}\left(\bar{x}(k)\right) + \bar{E}\bar{w}(k)\right]^{\mathrm{T}} P \left[\left(\bar{A} + \bar{B}\Lambda_1 \bar{F}\right)\bar{x}(k) + \bar{B}\Psi_{\bar{F}}\left(\bar{x}(k)\right) + \bar{E}\bar{w}(k)\right] \\
& -2\Psi_{\bar{F}}^{\mathrm{T}}\left(\bar{x}(k)\right)\Psi_{\bar{F}}\left(\bar{x}(k)\right) + 2\Psi_{\bar{F}}^{\mathrm{T}}\left(\bar{x}(k)\right)\Lambda \bar{F}\bar{x}(k), \quad \bar{x}(k) \in \Omega(P, \rho).
\end{aligned}$$

Recalling that for two vectors a, b and any positive number η,

$$(a+b)^{\mathrm{T}}(a+b) \leq (1+\eta)a^{\mathrm{T}}a + (1+1/\eta)b^{\mathrm{T}}b,$$

we have

$$
\begin{aligned}
& V\left(\bar{x}(k+1)\right) \\
\leq\ & (1+\eta)\left[\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right)\bar{x}(k)+\bar{B}\Psi_{\bar{F}}\left(\bar{x}(k)\right)\right]^{\mathrm{T}} P\left[\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right)\bar{x}(k)+\bar{B}\Psi_{\bar{F}}\left(\bar{x}(k)\right)\right] \\
& -2\Psi_{\bar{F}}^{\mathrm{T}}\left(\bar{x}(k)\right)\Psi_{\bar{F}}\left(\bar{x}(k)\right)+2\Psi_{\bar{F}}^{\mathrm{T}}\left(\bar{x}(k)\right)\Lambda\bar{F}\bar{x}(k)+(1+1/\eta)\bar{w}(k)^{\mathrm{T}}\bar{E}^{\mathrm{T}}P\bar{E}\bar{w}(k) \\
=\ & \begin{bmatrix} \bar{x}(k) \\ \Psi_{\bar{F}}\left(\bar{x}(k)\right) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} (1+\eta)\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right)^{\mathrm{T}} P\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right) & * \\ (1+\eta)\bar{B}^{\mathrm{T}}P\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right)+\Lambda\bar{F} & (1+\eta)\bar{B}^{\mathrm{T}}P\bar{B}-2I \end{bmatrix} \begin{bmatrix} \bar{x}(k) \\ \Psi_{\bar{F}}\left(\bar{x}(k)\right) \end{bmatrix} \\
& +(1+1/\eta)\bar{w}^{\mathrm{T}}(k)\bar{E}^{\mathrm{T}}P\bar{E}\bar{w}(k).
\end{aligned}
$$

where and elsewhere, $*$ denotes a block in a matrix that can be inferred by symmetry.

By assumption, $\bar{w}^{\mathrm{T}}(k)\bar{w}(k)\leq 1$, hence,

$$
\bar{w}^{\mathrm{T}}(k)\bar{E}^{\mathrm{T}}P\bar{E}\bar{w}(k)\leq\lambda_{\max}\left(\bar{E}^{\mathrm{T}}P\bar{E}\right):=\lambda.
$$

On $\partial\Omega(P,\rho)$, $\bar{x}^{\mathrm{T}}P\bar{x}=\rho$ or $\bar{x}^{\mathrm{T}}P\bar{x}/\rho=1$, we can write

$$
\lambda_{\max}\left(\bar{E}^{\mathrm{T}}P\bar{E}\right)=\lambda\bar{x}^{\mathrm{T}}(k)P\bar{x}(k)/\rho,\quad \bar{E}^{\mathrm{T}}P\bar{E}\leq\lambda I.
$$

It then follows that

$$
\begin{aligned}
& \Delta V(\bar{x}(k)) \\
=\ & V\left(\bar{x}(k+1)\right)-V\left(\bar{x}(k)\right) \\
\leq\ & \begin{bmatrix} \bar{x}(k) \\ \Psi_{\bar{F}}\left(\bar{x}(k)\right) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} (1+\eta)\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right)^{\mathrm{T}} P\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right) & * \\ (1+\eta)\bar{B}^{\mathrm{T}}P\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right)+\Lambda\bar{F} & (1+\eta)\bar{B}^{\mathrm{T}}P\bar{B}-2I \end{bmatrix} \begin{bmatrix} \bar{x}(k) \\ \Psi_{\bar{F}}\left(\bar{x}(k)\right) \end{bmatrix} \\
& +(1+1/\eta)\bar{w}^{\mathrm{T}}(k)\bar{E}^{\mathrm{T}}P\bar{E}\bar{w}(k)-\bar{x}^{\mathrm{T}}(k)P\bar{x}(k) \\
\leq\ & \begin{bmatrix} \bar{x}(k) \\ \Psi_{\bar{F}}\left(\bar{x}(k)\right) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} (1+\eta)\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right)^{\mathrm{T}} P\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right) & * \\ (1+\eta)\bar{B}^{\mathrm{T}}P\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right)+\Lambda\bar{F} & (1+\eta)\bar{B}^{\mathrm{T}}P\bar{B}-2I \end{bmatrix} \begin{bmatrix} \bar{x}(k) \\ \Psi_{\bar{F}}\left(\bar{x}(k)\right) \end{bmatrix} \\
& +(1+1/\eta)\lambda\bar{x}^{\mathrm{T}}(k)P\bar{x}(k)/\rho-\bar{x}^{\mathrm{T}}(k)P\bar{x}(k) \\
=\ & \begin{bmatrix} \bar{x}(k) \\ \Psi_{\bar{F}}\left(\bar{x}(k)\right) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} (1+\eta)\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right)^{\mathrm{T}} P\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right) & * \\ (1+\eta)\bar{B}^{\mathrm{T}}P\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right)+\Lambda\bar{F} & (1+\eta)\bar{B}^{\mathrm{T}}P\bar{B}-2I \end{bmatrix} \begin{bmatrix} \bar{x}(k) \\ \Psi_{\bar{F}}(k) \end{bmatrix} \\
& +\begin{bmatrix} \bar{x}(k) \\ \Psi_{\bar{F}}\left(\bar{x}(k)\right) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \left(\frac{1+\eta}{\eta}\frac{\lambda}{\rho}-1\right)P & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{x}(k) \\ \Psi_{\bar{F}}\left(\bar{x}(k)\right) \end{bmatrix},\quad \bar{x}(k)\in\partial\Omega(P,\rho).
\end{aligned}
$$

Therefore, $\Delta V(\bar{x}(k))\leq 0$, $\forall\bar{x}(k)\in\partial\Omega(P,\rho)$ if there exists a $\lambda\in\left[0,\frac{\eta\rho}{1+\eta}\right]$ such that

$$
\begin{bmatrix} \left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right)^{\mathrm{T}} P\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right)+\left(\frac{\lambda}{\eta\rho}-\frac{1}{1+\eta}\right)P & \left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right)^{\mathrm{T}} P\bar{B}+\frac{1}{1+\eta}\bar{F}^{\mathrm{T}}\Lambda \\ \bar{B}^{\mathrm{T}}P\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right)+\frac{1}{1+\eta}\Lambda\bar{F} & \bar{B}^{\mathrm{T}}P\bar{B}-\frac{2}{1+\eta}I \end{bmatrix}\leq 0, \quad (10)
$$

and

$$
\bar{E}^{\mathrm{T}}P\bar{E}\leq\lambda I. \quad (11)
$$

If we introduce a new variable $g \in (0,1)$, let $g = \frac{1}{1+\eta} - \frac{\lambda}{\eta\rho}$. For any fixed g and ρ, it can be verified that the maximum value of λ is $\lambda^* = \rho(1-\sqrt{g})^2$, achieved at $\eta = \frac{1}{\sqrt{g}} - 1$. We can vary g from 0 to 1 and have the following lemma.

Lemma 3 (Local stability) *Consider the discrete-time linear system with saturation (3), if there exist a positive definite matrix P and a diagonal matrix Λ_1 with $0 < \Lambda_1 < I$ and $\Lambda := I - \Lambda_1$ such that $\bar{A} + \bar{B}\Lambda_1\bar{F}$ is stable, and there exists scalar $g \in (0,1)$ such that*

$$\begin{bmatrix} (\bar{A}+\bar{B}\Lambda_1\bar{F})^{\mathrm{T}} P (\bar{A}+\bar{B}\Lambda_1\bar{F}) - gP & (\bar{A}+\bar{B}\Lambda_1\bar{F})^{\mathrm{T}} P\bar{B} + \sqrt{g}\bar{F}^{\mathrm{T}}\Lambda \\ \bar{B}^{\mathrm{T}}P(\bar{A}+\bar{B}\Lambda_1\bar{F}) + \sqrt{g}\Lambda\bar{F} & \bar{B}^{\mathrm{T}}P\bar{B} - 2\sqrt{g}I \end{bmatrix} \le 0, \tag{12}$$

$$\bar{E}^{\mathrm{T}}P\bar{E} \le \rho(1-\sqrt{g})^2 I, \tag{13}$$

$$\Omega(P,\rho) \subset \mathcal{L}(\Lambda_1\bar{F}), \tag{14}$$

then, any ellipsoid $\Omega(P,\rho) = \{\bar{x} : \bar{x}^{\mathrm{T}}P\bar{x} \le \rho\}$ inside the region $\mathcal{L}(\Lambda_1\bar{F})$ is an invariant set.

3.2 Regional l_2 Gain Analysis

We start with a definition.

Definition 2 (Regional l_2 gain [4]) *The system (8) is said to have a regional l_2 gain less than or equal to $\gamma > 0$ in $\Omega(P,\rho)$ if, for all $\bar{w} \in l_2$, the set $\Omega(P,\rho)$ is invariant and,*

$$\Delta V(\bar{x}(k)) \le \gamma^2 ||\bar{w}(k)||^2 - ||\tilde{u}(k)||^2, \quad \forall \bar{x}(k) \in \Omega(P,\rho). \tag{15}$$

If the system (8) has a regional l_2 gain in $\Omega(P,\rho)$, then, in the absence of initial condition, (15) implies that

$$\sum_{k=0}^{\infty} \left(\gamma^2 \bar{w}^{\mathrm{T}}(k)\bar{w}(k) - \tilde{u}^{\mathrm{T}}(k)\tilde{u}(k)\right) \ge 0,$$

or

$$\|\tilde{u}\|_2 \le \gamma \|\bar{w}\|_2.$$

Hence, by making γ small, we can reduce the energy of the reconstruction error signal.

The following lemma gives a sufficient condition for (15) to be true.

Lemma 4 *Consider the system (8). The regional l_2 gain from $\bar{w}$ to $\tilde{u}$ is less than or equal to $\gamma > 0$ if there exists a $P = P^{\mathrm{T}} > 0$ such that, in addition to (12),(13) and (14), the following inequality holds,*

$$\begin{bmatrix} (\bar{A}+\bar{B}\Lambda_1\bar{F})^{\mathrm{T}}P(\bar{A}+\bar{B}\Lambda_1\bar{F}) - P + \bar{C}^{\mathrm{T}}\bar{C} & * & * \\ \bar{B}^{\mathrm{T}}P(\bar{A}+\bar{B}\Lambda_1\bar{F}) + \Lambda\bar{F} & \bar{B}^{\mathrm{T}}P\bar{B} - 2I & * \\ \bar{E}^{\mathrm{T}}P(\bar{A}+\bar{B}\Lambda_1\bar{F}) & \bar{E}^{\mathrm{T}}P\bar{B} & \bar{E}^{\mathrm{T}}P\bar{E} - \gamma^2 I \end{bmatrix} \le 0. \tag{16}$$

Proof. First, we note that, by Lemma 3, inequalities (12), (13) and (14) imply that $\Omega(P,\rho)$ is an invariant set. We next show that (15) holds. To this end, consider the Lyapunov function candidate

$$V(\bar{x}) = \bar{x}^{\mathrm{T}} P \bar{x}.$$

Recalling that

$$\Psi_{\bar{F}}^{\mathrm{T}}(\bar{x})[\Psi_{\bar{F}}(\bar{x}) - \Lambda\bar{F}\bar{x}] \leq 0, \quad \forall \bar{x} \in \Omega(P,\rho),$$

we have

$$\begin{aligned}
& \Delta V(\bar{x}(k)) - \left(\gamma^2\|\bar{w}(k)\|^2 - \|\tilde{u}(k)\|^2\right) \\
< \; & \Delta V(\bar{x}(k)) - 2\Psi_{\bar{F}}^{\mathrm{T}}(\bar{x}(k))[\Psi_{\bar{F}}(\bar{x}(k)) - \Lambda\bar{F}\bar{x}(k)] - \left(\gamma^2\|\bar{w}(k)\|^2 - \|\tilde{u}(k)\|^2\right) \\
= \; & \bar{x}(k)^{\mathrm{T}}\left[(\bar{A}+\bar{B}\Lambda_1\bar{F})^{\mathrm{T}}P(\bar{A}+\bar{B}\Lambda_1\bar{F}) - P\right]\bar{x}(k) + 2\Psi_{\bar{F}}^{\mathrm{T}}(\bar{x}(k))\left[\bar{B}^{\mathrm{T}}P(\bar{A}+\bar{B}\Lambda_1\bar{F}) + \Lambda\bar{F}\right]\bar{x}(k) \\
& + \Psi_{\bar{F}}^{\mathrm{T}}(\bar{x}(k))\left(\bar{B}^{\mathrm{T}}P\bar{B} - 2I\right)\Psi_{\bar{F}}(\bar{x}(k)) + 2\bar{w}^{\mathrm{T}}(k)\bar{E}^{\mathrm{T}}P(\bar{A}+\bar{B}\Lambda_1\bar{F})\bar{x}(k) \\
& + 2\bar{w}^{\mathrm{T}}(k)\bar{E}^{\mathrm{T}}P\bar{B}\Psi_{\bar{F}}(\bar{x}(k)) + \bar{w}^{\mathrm{T}}(k)\bar{E}^{\mathrm{T}}P\bar{E}\bar{w}(k) + \bar{x}^{\mathrm{T}}\bar{C}^{\mathrm{T}}\bar{C}\bar{x} - \gamma^2\bar{w}^{\mathrm{T}}\bar{w} \\
= \; & \bar{x}^{\mathrm{T}}(k)\left[(\bar{A}+\bar{B}\Lambda_1\bar{F})^{\mathrm{T}}P(\bar{A}+\bar{B}\Lambda_1\bar{F}) - P + \bar{C}^{\mathrm{T}}\bar{C}\right]\bar{x}(k) \\
& + 2\Psi_{\bar{F}}^{\mathrm{T}}(\bar{x}(k))\left[\bar{B}^{\mathrm{T}}P(\bar{A}+\bar{B}\Lambda_1\bar{F}) + \Lambda\bar{F}\right]\bar{x}(k) + \Psi_{\bar{F}}^{\mathrm{T}}(\bar{x}(k))\left(\bar{B}^{\mathrm{T}}P\bar{B} - 2I\right)\Psi_{\bar{F}}(\bar{x}(k)) \\
& + 2\bar{w}^{\mathrm{T}}(k)\bar{E}^{\mathrm{T}}P(\bar{A}+\bar{B}\Lambda_1\bar{F})\bar{x}(k) + 2\bar{w}^{\mathrm{T}}(k)\bar{E}^{\mathrm{T}}P\bar{B}\Psi_{\bar{F}}(\bar{x}(k)) + \bar{w}^{\mathrm{T}}(k)\left(\bar{E}^{\mathrm{T}}P\bar{E} - \gamma^2 I\right)\bar{w}(k) \\
= \; & \begin{bmatrix} \bar{x}(k)^{\mathrm{T}} & \Psi_{\bar{F}}^{\mathrm{T}}(\bar{x}(k)) & \bar{w}^{\mathrm{T}}(k) \end{bmatrix} M \begin{bmatrix} \bar{x}(k) \\ \Psi_{\bar{F}}(\bar{x}(k)) \\ \bar{w}(k) \end{bmatrix}, \quad \bar{x}(k) \in \Omega(P,\rho),
\end{aligned}$$

in which

$$M := \begin{bmatrix} (\bar{A}+\bar{B}\Lambda_1\bar{F})^{\mathrm{T}}P(\bar{A}+\bar{B}\Lambda_1\bar{F}) - P + \bar{C}^{\mathrm{T}}\bar{C} & * & * \\ \bar{B}^{\mathrm{T}}P(\bar{A}+\bar{B}\Lambda_1\bar{F}) + \Lambda\bar{F} & \bar{B}^{\mathrm{T}}P\bar{B} - 2I & * \\ \bar{E}^{\mathrm{T}}P(\bar{A}+\bar{B}\Lambda_1\bar{F}) & \bar{E}^{\mathrm{T}}P\bar{B} & \bar{E}^{\mathrm{T}}P\bar{E} - \gamma^2 I \end{bmatrix}.$$

Clearly, $M \leq 0$ implies that

$$\Delta V(\bar{x}(k)) - \left(\gamma^2\|\bar{w}(k)\|^2 - \|\tilde{u}(k)\|^2\right) \leq 0, \quad \forall \bar{x}(k) \in \Omega(P,\rho).$$

■

3.3 Robust hybrid filtering design

We now proceed to develop a robust hybrid filtering algorithm. The goal of such a robust hybrid filtering algorithm is not only to meet the H_2 and H_∞ criterion under linear operation environment, but also to secure a low regional l_2 gain under amplitude saturation. By accommodating both linear/saturated working scenarios, we will be able to ensure that the saturation will not

severely deteriorate the reconstruction performance, while the response in saturation-free phase is also acceptable.

The robust hybrid filtering problem can be stated as follows.

Problem 1 *Given a discrete-time system, where the measured output r is corrupted by saturation and noise, and u is the signal to be reconstructed,*

$$\begin{aligned} x(k+1) &= Ax(k) + Bv(k), \\ r(k) &= \mathrm{sat}(Gx(k)) + Ew(k), \\ u(k) &= Dx(k). \end{aligned}$$

For some prescribed scalars $\tau > 0$ and $\mu > 0$, we would like to design a filter

$$\begin{aligned} \hat{x}(k+1) &= A_f\hat{x}(k) + B_f r(k), \\ \hat{u}(k) &= C_f\hat{x}(k), \end{aligned}$$

such that the regional l_2 gain from $\begin{bmatrix} v^{\mathrm{T}}(k) & w^{\mathrm{T}}(k) \end{bmatrix}^{\mathrm{T}}$ to the filtering error $\tilde{u}(k) = u(k) - \hat{u}(k)$ is minimized, while the H_2-norm and H_∞-norm of transfer function $T(z) = \bar{C}(zI - \bar{A}_L)^{-1}\bar{E}$ are less than or equal to τ and μ, respectively.

Summarizing Lemma 1 through Lemma 4, and letting $P = R^{-1}$, we can formulate the synthesis problem as follows.

$$\begin{aligned} & \min_{P=P^{\mathrm{T}}>0,\rho,g,A_f,B_f,D_f} \gamma^2 \\ \text{s.t.} \quad & (5),(6),(7),(12),(13),(14),(16), \\ & \rho > 0, \\ & 0 < g < 1. \end{aligned}$$

It is observed that the above constraints are not LMIs with respect to P, A_f, B_f and D_f. In an effort to transform the synthesis problem into an LMI constrained optimization problem, we need examine these constraints further.

First consider (5). Recalling that congruence transformation does not change positive (negative) definiteness, we apply congruence transformation R^{-1} and obtain

$$(R^{-1}\bar{A}_L)R(R^{-1}\bar{A}_L)^{\mathrm{T}} - R^{-1} + (R^{-1}\bar{E})(R^{-1}\bar{E})^{\mathrm{T}} \leq 0.$$

Applying Schur complement twice, we obtain

$$\begin{bmatrix} R^{-1} & R^{-1}\bar{A}_L & R^{-1}\bar{E} \\ \bar{A}_L^{\mathrm{T}}R^{-1} & R^{-1} & 0 \\ \bar{E}^{\mathrm{T}}R^{-1} & 0 & I \end{bmatrix} \geq 0,$$

or

$$\begin{bmatrix} P & P\bar{A}_L & P\bar{E} \\ \bar{A}_L^{\mathrm{T}}P & P & 0 \\ \bar{E}^{\mathrm{T}}P & 0 & I \end{bmatrix} > 0. \tag{17}$$

Next consider (6). We introduce an auxiliary matrix $N := \tau^2 I/q$, where q is the number of rows of $\bar{C}$. A sufficient condition for

$$\mathrm{trace}(\bar{C}R\bar{C}^{\mathrm{T}}) \leq \tau^2$$

is

$$\bar{C}R\bar{C}^{\mathrm{T}} \leq N,$$

which by Schur complement is equivalent to

$$\begin{bmatrix} N & \bar{C} \\ \bar{C}^{\mathrm{T}} & P \end{bmatrix} \geq 0. \tag{18}$$

We temporarily skip (7) and go to (12). The left hand side of inequality (12) equals

$$\begin{bmatrix} -gP & \sqrt{g}\bar{F}^{\mathrm{T}}\Lambda \\ \sqrt{g}\Lambda\bar{F} & -2\sqrt{g}I \end{bmatrix} + \begin{bmatrix} (\bar{A}+\bar{B}\Lambda_1\bar{F})^{\mathrm{T}} \\ \bar{B}^{\mathrm{T}} \end{bmatrix} P \begin{bmatrix} \bar{A}+\bar{B}\Lambda_1\bar{F} & \bar{B} \end{bmatrix},$$

or

$$\begin{bmatrix} -gP & \sqrt{g}\bar{F}^{\mathrm{T}}\Lambda \\ \sqrt{g}\Lambda\bar{F} & -2\sqrt{g}I \end{bmatrix} + \begin{bmatrix} (\bar{A}+\bar{B}\Lambda_1\bar{F})^{\mathrm{T}}P \\ \bar{B}^{\mathrm{T}}P \end{bmatrix} P^{-1} \begin{bmatrix} P\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right) & P\bar{B} \end{bmatrix}.$$

By Schur complement, we arrive at

$$\begin{bmatrix} -gP & \sqrt{g}\bar{F}^{\mathrm{T}}\Lambda & (\bar{A}+\bar{B}\Lambda_1\bar{F})^{\mathrm{T}}P \\ \sqrt{g}\Lambda\bar{F} & -2\sqrt{g}I & \bar{B}^{\mathrm{T}}P \\ P\left(\bar{A}+\bar{B}\Lambda_1\bar{F}\right) & P\bar{B} & -P \end{bmatrix} \leq 0, \quad 0 < g < 1. \tag{19}$$

As to inequality (13), it is equivalent to

$$\rho(1-\sqrt{g})^2 I - \left(\bar{E}^{\mathrm{T}}P\right) P^{-1} \left(P\bar{E}\right) \geq 0,$$

which by Schur complement gives

$$\begin{bmatrix} \rho\left(1-\sqrt{g}\right)^2 I & \bar{E}^{\mathrm{T}}P \\ P\bar{E} & P \end{bmatrix} \geq 0, \quad \rho > 0, 0 < g < 1. \tag{20}$$

Now consider (14). This constraint can be translated to

$$P \geq \rho(\Lambda_1\bar{F})_i^{\mathrm{T}}(\Lambda_1\bar{F})_i, \quad \rho > 0, \forall i, \tag{21}$$

where $(\Lambda_1\bar{F})_i$ is the ith row of $\Lambda_1\bar{F}$.

Finally, the left hand side of constraint (16) can be written as

$$\begin{bmatrix} (\bar{A}+\bar{B}\Lambda_1\bar{F})^{\mathrm{T}} \\ \bar{E}^{\mathrm{T}} \\ \bar{B}^{\mathrm{T}} \end{bmatrix} P \begin{bmatrix} \bar{A}+\bar{B}\Lambda_1\bar{F} & \bar{E} & \bar{B} \end{bmatrix} + \begin{bmatrix} \bar{C}^{\mathrm{T}}\bar{C}-P & 0 & \bar{F}^{\mathrm{T}}\Lambda \\ 0 & -\gamma^2 I & 0 \\ \Lambda\bar{F} & 0 & -2I \end{bmatrix}.$$

By Schur complement, constraint (16) is equivalent to

$$\begin{bmatrix} \bar{C}^{\mathrm{T}}\bar{C}-P & 0 & \bar{F}^{\mathrm{T}}\Lambda & (\bar{A}+\bar{B}\Lambda_1\bar{F})^{\mathrm{T}}P \\ 0 & -\gamma^2 I & 0 & \bar{E}^{\mathrm{T}}P \\ \Lambda\bar{F} & 0 & -2I & \bar{B}^{\mathrm{T}}P \\ P(\bar{A}+\bar{B}\Lambda_1\bar{F}) & P\bar{E} & P\bar{B} & -P \end{bmatrix} \leq 0. \tag{22}$$

Now we are able to reformulate the synthesis problem as follows,

$$\begin{aligned} & \min_{P=P^{\mathrm{T}}>0,\rho,g,A_f,B_f,D_f} \gamma^2 \\ \text{s.t.} \quad & (17),(18),(7),(19),(20),(21),(22), \\ & \rho > 0, \\ & 0 < g < 1. \end{aligned}$$

At this stage, we can easily turn this problem into an LMIs constrained optimization problem by the following technique. Assume P is of the form

$$P = \begin{bmatrix} X & Y \\ Y & Y \end{bmatrix},$$

where X and Y are positive definite and satisfy $X > Y$. Then define new variables

$$Q_a = YA_f, Q_b = YB_f.$$

By substituting the parameters of $A, B, D, G, X, Y, Q_a, Q_b$ and D_f into the matrix inequalities (17), (18), (7), (19)-(22), and denoting $\gamma_2 := \gamma^2$, we arrive at the following LMIs in $\rho, \gamma_2, X, Y, Q_a, Q_b$ and D_f correspondingly,

$$\begin{bmatrix} X & Y & XA+Q_bG & Q_a & XB & Q_bE \\ Y & Y & YA+Q_bG & Q_a & YB & Q_bE \\ * & * & X & Y & 0 & 0 \\ * & * & * & Y & 0 & 0 \\ * & * & * & * & I & 0 \\ * & * & * & * & * & I \end{bmatrix} \geq 0, \tag{23}$$

$$\begin{bmatrix} N & D & -D_f \\ * & X & Y \\ * & * & Y \end{bmatrix} \geq 0, \tag{24}$$

$$\begin{bmatrix} X & Y & * & * & * & * & * \\ Y & Y & * & * & * & * & * \\ 0 & 0 & \mu I & * & * & * & * \\ 0 & 0 & 0 & \mu I & * & * & * \\ XA+Q_bG & Q_a & XB & Q_bE & X & Y & * \\ YA+Q_bG & Q_a & YB & Q_bE & Y & Y & * \\ D & -D_f & 0 & 0 & 0 & 0 & \mu I \end{bmatrix} \geq 0. \tag{25}$$

$$\begin{bmatrix} -gX & -gY & * & * & * \\ -gY & -gY & * & * & * \\ \sqrt{g}\Lambda G & 0 & -2\sqrt{g}I & * & * \\ XA+Q_b\Lambda_1 G & Q_a & Q_b & -X & -Y \\ YA+Q_b\Lambda_1 G & Q_a & Q_b & -Y & -Y \end{bmatrix} \leq 0, \tag{26}$$

$$\begin{bmatrix} \rho\left(1-\sqrt{g}\right)^2 I & * & * & * \\ 0 & \rho\left(1-\sqrt{g}\right)^2 I & * & * \\ XB & Q_bE & X & Y \\ YB & Q_bE & Y & Y \end{bmatrix} \geq 0, \tag{27}$$

$$X \geq \rho\left(\Lambda_1 G\right)_i^{\mathrm{T}}\left(\Lambda_1 G\right)_i, \quad \forall i, \tag{28}$$

$$\begin{bmatrix} -X & -Y & * & * & * & * & * & * \\ -Y & -Y & * & * & * & * & * & * \\ 0 & 0 & -\gamma_2 I & * & * & * & * & * \\ 0 & 0 & 0 & -\gamma_2 I & * & * & * & * \\ \Lambda G & 0 & 0 & 0 & -2I & * & * & * \\ XA+Q_b\Lambda_1 G & Q_a & XB & Q_bE & Q_b & -X & -Y & * \\ YA+Q_b\Lambda_1 G & Q_a & YB & Q_bE & Q_b & -Y & -Y & * \\ D & -D_f & 0 & 0 & 0 & 0 & 0 & -I \end{bmatrix} \leq 0. \tag{29}$$

Therefore, the robust hybrid filter can be formulated into the following LMIs constrained optimization problem that is readily solvable by the Matlab LMI toolbox.

$$\begin{aligned} & \min_{X,Y,\rho,g,Q_a,Q_b,D_f} \gamma_2 \\ \text{s.t.} \quad & (23),(24),(25),(26),(27),(28),(29), \\ & X = X^{\mathrm{T}} > Y = Y^{\mathrm{T}} > 0, \\ & \rho > 0, \\ & \gamma_2 > 0, \\ & 0 < g < 1. \end{aligned}$$

Note that constraints (26) and (27) are not linear with respect to parameter g, and we must fix g each time until minimal γ_2 is found. The parameters A_f and B_f of filter $\mathcal{F}$ can be recovered by $A_f = Y^{-1}Q_a$ and $B_f = Y^{-1}Q_b$, respectively.

3.4 An Example

Consider the system

$$\begin{aligned} x(k+1) &= Ax(k) + Bv(k), \\ r(k) &= \mathrm{sat}(Gx(k)) + Ew(k), \\ u(k) &= Dx(k), \end{aligned}$$

with

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, B = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, G = \begin{bmatrix} -0.0007 & 0.9462 & 0.0647 \\ 0.0017 & 1.2856 & 0.0017 \\ 0.0647 & 0.9462 & -0.0007 \\ 0.3453 & 0.3453 & 0 \end{bmatrix}, D = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}, E = I.$$

We would like to design a filter of the following form,

$$\begin{aligned} \hat{x}(k+1) &= A_f\hat{x}(k) + B_f r(k), \\ \hat{u}(k) &= D_f\hat{x}(k). \end{aligned}$$

By our hybrid design method, we obtain

$$A_f = \begin{bmatrix} 0.0959 & -0.0496 & 0.0030 \\ 0.2605 & -0.1356 & 0.0085 \\ -0.0148 & 0.0527 & -0.0125 \end{bmatrix},$$

$$B_f = \begin{bmatrix} 0.0739 & 0.0960 & -0.0471 & -0.6145 \\ 0.1860 & 0.2413 & -0.1266 & -1.5899 \\ -0.2162 & -0.2945 & -0.2046 & -0.0128 \end{bmatrix},$$

and

$$D_f = \begin{bmatrix} -0.0006 & -0.0320 & -0.9895 \end{bmatrix}.$$

By the mixed H_2/H_∞ method [6], we obtain

$$K = \begin{bmatrix} -0.0000 & -0.0000 & -0.0000 & -0.0000 \\ -0.0335 & -0.0434 & 0.0251 & 0.2983 \\ 0.2034 & 0.2773 & 0.2019 & 0.0625 \end{bmatrix}.$$

To evaluate the reconstruction performance, we define the output signal to noise ratio (SNR) as a relative measure of reconstruction error,

$$SNR^o := 10\log_{10}\frac{\Sigma_k u^2(k)}{\Sigma_k (u(k)-\hat{u}(k))^2},$$

while the input signal to noise ratio is

$$SNR^i := 10\log_{10}\frac{\Sigma_k y^2(k)}{\Sigma_k w^2(k)}.$$

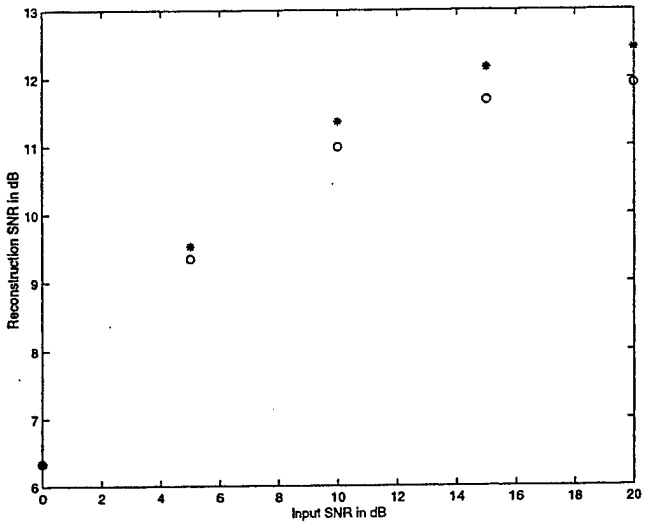


Figure 3: Reconstruction performance of the hybrid design and mixed H_2/H_∞ design: under saturation.

With amplitude saturation, when the input signal to noise ratios (SNR^i) are 0dB,5dB, 10dB, 15dB and 20dB, the reconstruction SNR of our design (denoted by SNR_H^o) and mixed H_2/H_∞ design of [6] (denoted by SNR_M^o) are listed in the the following table, and illustrated in Fig. 3.

SNR^i	0	5	10	15	20
SNR_H^o	6.3270	9.5259	11.3614	12.1525	12.4360
SNR_M^o	6.3365	9.3415	10.9891	11.6744	11.9156

The performance comparison in linear (unsaturated) case is illustrated in Fig. 4. It is noteworthy that in both cases our filter performs significantly better.

4 Digital Transmultiplexers

In this section, we will apply the hybrid filter design algorithm to the digital transmultiplexer systems. We will first recall the state space model of digital transmultiplexers established in [6, 17], with a slight modification to accommodate the amplitude saturation. To design the separating filter bank, a generalized dynamic filter in state-space form is proposed and obtained by our hybrid design method via the LMI technique in Section 3. Simulation will compare its reconstruction performance to that in [6], and show that our filter is indeed efficient to retain reconstruction performance under saturation.

4.1 State Space Model of Digital Transmultiplexer Systems

The block diagram of a multi-rate digital transmultiplexer system with linear channel distortion, amplitude saturation and measuring noise is shown in Fig. 5. The digital transmultiplexer

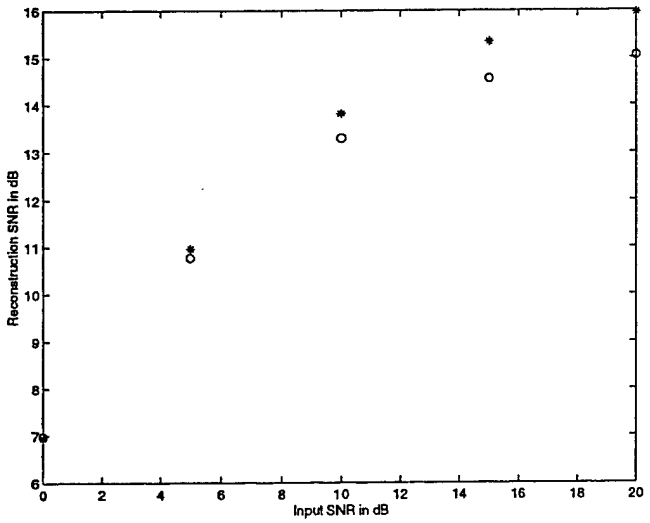


Figure 4: Reconstruction performance of the hybrid design and mixed H_2/H_∞ design: without saturation.

system without saturation can be modeled as in [6, 17]. The input signals to the M band transmultiplexer, $u_0(k), u_1(k), \cdots, u_{M-1}(k)$ are first expanded by a factor of N, then they pass through the M FIR combining filters $F_0(z), F_1(z), \cdots, F_{M-1}(z)$ (synthesis filter bank) to be combined into a single signal which then is transmitted via a common communication medium with transfer function of $C(z)$. Define $G_i(z) = F_i(z)C(z)$, the overall transfer functions for M channels are $\{G_i(z), i = 0, 1, \cdots, M-1\}$ with order L', and L' is chosen as an integer multiple of N (if not so, use zero-padding). Let $L = L'/N$, define the state vector of length ML,

$$x(k) = [x_0^{\mathrm{T}}(k), x_1^{\mathrm{T}}(k), \cdots, x_{M-1}^{\mathrm{T}}(k)]^{\mathrm{T}},$$

with

$$x_i(k) = [u_i(k), u_i(k-1), \cdots, u_i(k-L+1)]^{\mathrm{T}}, \quad i = 0, 1, \cdots, M-1,$$

and the input vector

$$v(k) = [u_0(k+1), u_1(k+1), \cdots, u_{M-1}(k+1)]^{\mathrm{T}}.$$

Let $r(k)$ denote the measured output, and $w(k)$ the additive noise including quantization error, channel disturbance and measurement noise at the receiver (for convenience we will simply call $w(k)$ the measurement noise later in this paper). Here w is assumed to have finite energy and bounded amplitude, but its statistics is not known. Now that $v(k)$ is the system driving noise with unknown statistics, according to [17] and [6], the saturation-free state space model

of transmultiplexer is

$$\begin{aligned} x(k+1) &= Ax(k) + Bv(k), \\ y(k) &= Gx(k), \\ r(k) &= y(k) + w(k). \end{aligned}$$

From the received signal $r(k)$, the following delayed version of input is to be reconstructed:

$$u(k) = \begin{bmatrix} u_0(k-d) \\ u_1(k-d) \\ \vdots \\ u_{M-1}(k-d) \end{bmatrix} = Dx(k).$$

Integer $d \in [1, L-1]$ is the time delay, and D is a constant matrix used to extract the desired signal from state variables. In the above, matrices A, B, G and D are as follows,

$$A = \mathrm{diag}\{A_0, A_1, \cdots, A_{M-1}\}, \quad A_i = \begin{bmatrix} 0 & 0 & & \cdots & 0 \\ 1 & 0 & & \cdots & 0 \\ 0 & 1 & \ddots & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} b_0 & 0_{L\times 1} & \cdots & 0_{L\times 1} \\ 0_{L\times 1} & b_1 & \cdots & 0_{L\times 1} \\ \vdots & \vdots & & \vdots \\ 0_{L\times 1} & 0_{L\times 1} & \cdots & b_{M-1} \end{bmatrix}, \quad b_i = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$G = \begin{bmatrix} G_0 & G_1 \cdots & G_{M-1} \end{bmatrix},$$

$$G_i = \begin{bmatrix} \mathbf{g}_i(0) & \mathbf{g}_i(1) & \cdots & \mathbf{g}_i(N-1) \end{bmatrix}^{\mathrm{T}},$$

$$\mathbf{g}_i(n) = \begin{bmatrix} g_i(n) & g_i(N+n) & \cdots & g_i((L-1)N+n) \end{bmatrix},$$

$$D = \begin{bmatrix} e_{d+1}^{\mathrm{T}} \\ e_{d+1+L}^{\mathrm{T}} \\ \vdots \\ e_{d+1+L(M-1)}^{\mathrm{T}} \end{bmatrix}.$$

Note that the (finite length) sequence $\{g_i(n), n = 0, 1, \cdots, L'-1\}$ denotes the impulse response of the ith channel $G_i(z)$. In matrix D, e_k stands for a column vector of length ML where the kth element is one but all other elements are zero.

If we also consider the saturation effect in the channel, then the final model of transmulti-

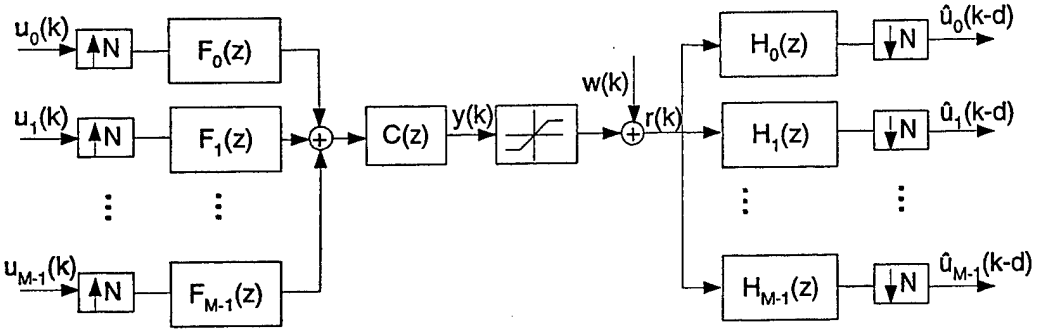


Figure 5: Digital transmultiplexer subject to saturation and additive noise.

plexer is established as

$$\begin{aligned} x(k+1) &= Ax(k)+Bv(k), \\ y(k) &= Gx(k), \\ r(k) &= \mathrm{sat}(y(k))+w(k), \\ u(k) &= Dx(k). \end{aligned} \tag{30}$$

At the receiver, what is obtained is the distorted and contaminated version of combined signal from which we must design a separating filter bank to recover the inputs up to a constant delay, i.e., $u_0(k-d)$, $u_1(k-d)$, $\cdots$, $u_{M-1}(k-d)$. In other words, we would like to estimate $u(k)$ from received signal $r(k)$. We use the following generalized dynamic filter

$$\begin{aligned} \hat{x}(k+1) &= A_f\hat{x}(k)+B_f r(k), \\ \hat{u}(k) &= D_f\hat{x}(k). \end{aligned}$$

4.2 A Numerical Example

This four band transmultiplexer example is borrowed from [6]. The input signals are generated by letting a zero-mean white Gaussian process pass through a 10-th order autoregressive (AR) model

$$\begin{aligned} A_r(z) =& 1-1.1z^{-1}-0.23z^{-2}+0.633z^{-3}-0.129z^{-4}-0.1621z^{-5} \\ &+8.8948\times 10^{-2}z^{-6}+1.6251\times 10^{-2}z^{-7}-1.7851\times 10^{-2}z^{-8} \\ &-1.3124\times 10^{-3}z^{-9}+7.5725\times 10^{-4}z^{-10}. \end{aligned}$$

The input signals are expanded by a factor of 12, then go through the length-16 combining filter bank $\{F_i(z), i=0,1,2,3\}$. The combined signal is transmitted via a communication medium whose impulse response is described by

$$C(z)=1+1.36z^{-1}+0.3948z^{-2}-0.631z^{-3}-0.7495z^{-4}-0.3055z^{-5}+0.0419z^{-6}.$$

The frenquency response of $C(z)$ is depicted in Fig. 7, and it can be seen how this channel distorted transmitted signals in magnitude and phase. The channel/measuring noise is modeled by zero-mean white Gaussian noise. By the modeling procedure developed in [6], we obtain matrix G:

$$\begin{bmatrix}
0.3637 & -2.4848 & 0.0853 & 0.4252 & 0.4290 & 0.2427 & -0.2430 & -0.0768 \\
0.6836 & -5.2951 & -0.0442 & 5.7500 & 0.4764 & 1.9490 & -0.2929 & 0.0393 \\
0.1542 & -4.8631 & 0.0247 & 3.6600 & 0.1633 & 0.1059 & -0.0938 & -0.0734 \\
-1.3389 & -2.6194 & -0.0322 & 0.7067 & -1.1246 & -1.1791 & 0.6688 & 0.1423 \\
-2.8577 & -0.8069 & -2.2945 & -0.2236 & -1.1345 & -0.4731 & -0.9064 & 0.2820 \\
-2.7536 & -0.2139 & -4.9396 & -0.3266 & 2.9017 & 0.1951 & 1.1991 & 0.2389 \\
0.1305 & -0.2202 & -0.9387 & -0.4355 & 0.8707 & 0.2461 & -0.8508 & -0.0481 \\
5.2274 & -0.1957 & 9.0781 & -0.3601 & -4.4889 & 0.1216 & 0.3316 & -0.2045 \\
10.0426 & -0.0758 & 11.6440 & -0.1355 & 0.3573 & 0.0328 & 0.2642 & -0.1032 \\
11.6369 & -0.0102 & 0.1275 & -0.0180 & 4.0435 & 0.0036 & -0.5275 & -0.0152 \\
8.8303 & 0 & -11.8774 & 0 & -0.7888 & 0 & 0.4027 & 0 \\
3.0940 & 0 & -9.9863 & 0 & -2.8276 & 0 & -0.1470 & 0
\end{bmatrix}.$$

By using the hybrid design method, we obtain the parameters A_f, B_f and D_f of the filter, and the frequency response of the separating filters are depicted in Fig. 6. Note that the orthogonality of the filter bank is ruined due to the channel distortion and amplitude saturation. In particular, looking at the frequency response of channel $C(z)$ without saturation (Fig. 7), we observe that the channel suppresses the high frequency components. As can be seen, the separating filters of the 3rd and 4th band have higher gains than the other two, which could be interpreted as the higher gains serve to compensate the suppression of the magnitude.

To evaluate the reconstruction performance, we define the output signal to noise ratio (SNR) of the ith band as a relative measure of reconstruction error,

$$SNR^o := 10 \log_{10} \frac{\Sigma_k u_i^2(k)}{\Sigma_k (u_i(k) - \hat{u}_i(k))^2},$$

while the input signal to noise ratio of the ith band is

$$SNR^i := 10 \log_{10} \frac{\Sigma_k y_i^2(k)}{\Sigma_k w_i^2(k)}.$$

When no saturation occurs, with input signal to noise ratio SNR^i at the receiver equals 0dB, 5dB, 10dB, 15dB and 20dB respectively, we test the system by obtaining the output signal to noise ratio SNR^o at the reconstruction output. We list the simulation results of the second band from the hybrid design scheme, denoted as SNR^o_H and the results from mixed H_2/H_∞

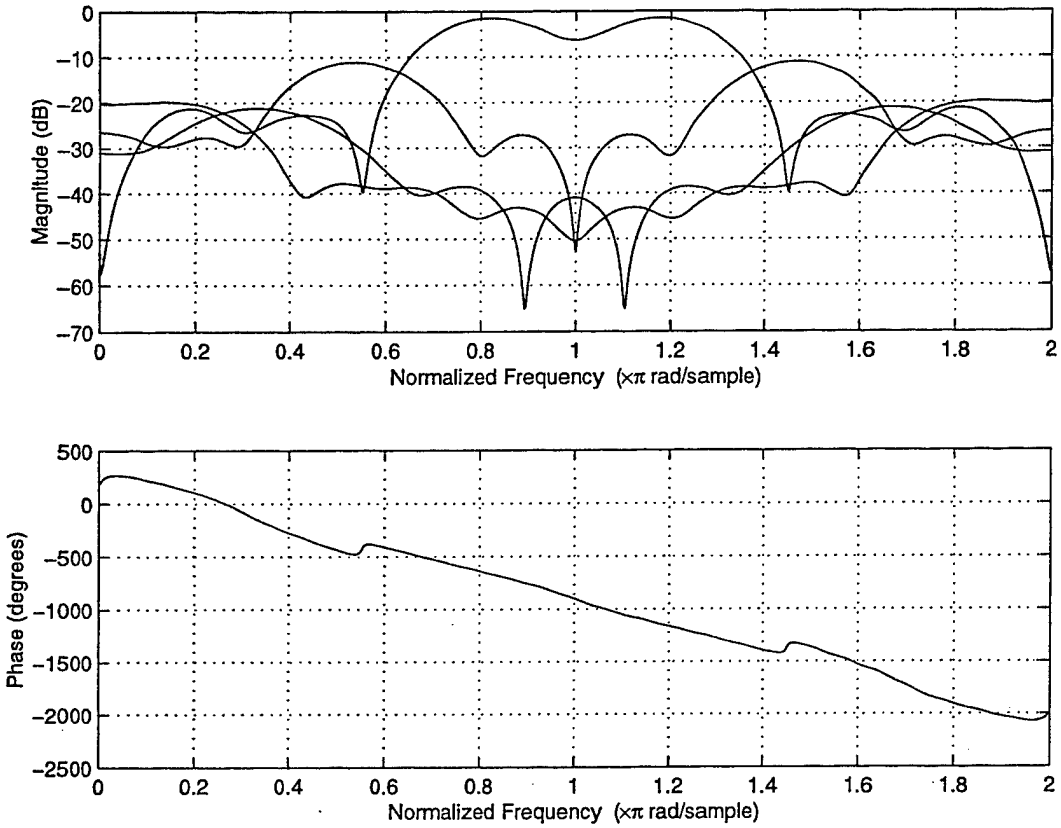


Figure 6: Frequency responses of the separating filter bank.

method denoted as SNR_M^o.

SNR^i	0	5	10	15	20
SNR_H^o	16.5880	21.5734	26.5268	31.3817	35.9507
SNR_M^o	16.6627	21.6577	26.6422	31.5940	36.4455

When (amplitude) saturation takes place in the process of transmission, i.e.,

$$r(k) = \text{sat}(y(k)) + w(k),$$

we run the system and obtain the following results.

SNR^i	0	5	10	15	20
SNR_H^o	15.9272	19.8136	22.4881	23.8654	24.4169
SNR_M^o	15.9155	19.6867	22.2012	23.4525	23.9422

From the simulation results, we notice that in linear case, our design has slightly lower output SNR (0.08-0.5dB) compared to the mixed H_2/H_∞ approach, but when amplitude saturation is present in the channel, our design outperforms the mixed H_2/H_∞ by as much as 0.47dB (11.4%). Our design method demonstrates a better reconstruction property in saturated case

at the slight cost of linear performance. However it is worth noticing that, in linear case when output SNR is already high, a degradation of 0.1-0.5dB is often acceptable, while in saturated operation, an improvement of 0.1-0.5dB will be critical to the overall system performance.

5 Conclusions

In this paper, we have discussed the design of separating filter bank of digital transmultiplexer systems with consideration of amplitude saturation and proposed a so-called hybrid filter design method. The hybrid method takes advantage of linear filtering and nonlinear regional l_2 gain filtering in order to preserve the reconstruction performance when saturation takes place. This problem is formulated into an optimization problem with LMI constraints. Simulation results show the robustness of our design in the presence of saturation.

6 Acknowledgment

The authors would like to thank Prof. Bor-Sen Chen of National Tsing Hua University, Taiwan, for generously providing us with the data used in the example in Section 4 and a preprint of [6].

References

[1] A.N. Akansu, M.V. Tazebay and R. A. Haddad, "A New Look at Digital Orthogonal Transmultiplexers for CDMA Communications", *IEEE Trans. on Signal Processing*, Vol. 45, No. 1, pp. 263-267, 1997.

[2] P. Apkarian, P.C. Pellanda and H.D. Tuan, "Mixed H_2/H_∞ multi-channel linear parameter-varying control in discrete time", *Systems & Control Letters*, Vol. 41, pp. 333-346, 2000.

[3] T. Bose and M.-Q. Chen, "Overflow oscillations in state-space digital filters", *IEEE Trans. on Circuits and Systems II*, Vol. 38, No. 7, pp. 807-810, 1991.

[4] Y.-Y. Cao, Z. Lin and D. G. Ward, "H_∞ anti-windup design for linear systems subject to input saturation", *AIAA J. of Guidance, Control & Dynamics*, to appear, 2002.

[5] B.M. Chen, *Robust and H_∞ Control*, Springer, London, 2000.

[6] B.-S. Chen, C.-L. Tsai and Y.-F. Chen, "Mixed H_2/H_∞ filtering design in multirate transmultiplexer systems: LMI approach", *IEEE Trans. Signal Processing*, Vol. 49, No. 11, pp. 2693-2701, 2001.

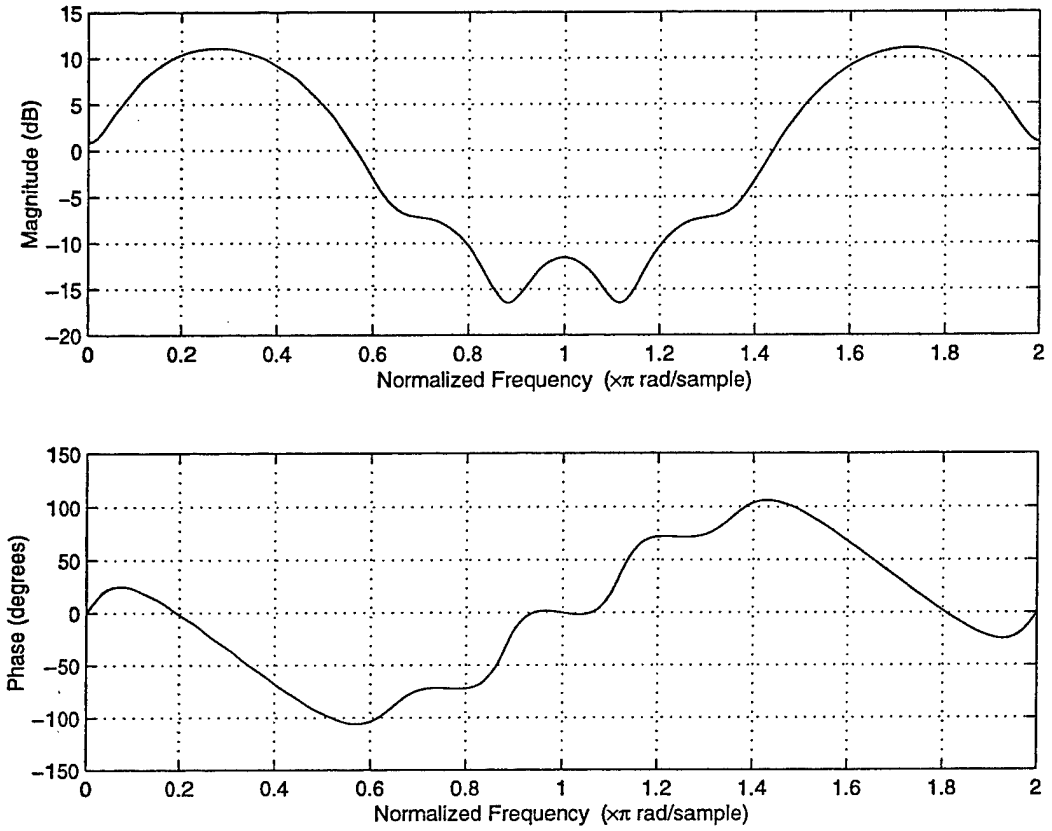


Figure 7: Frequency responses of the communication channel $C(z)$.

[7] S. Dussy, "Robust stabilization of discrete-time parameter-dependent systems: the finite precision problem", *Proceedings of the 35th Conference on Decision & Control*, pp. 3976-3981, 1996.

[8] A. T. Erdogan, B. Hassibi and T. Kailath, "On linear H^∞ equalization of communication channels", *IEEE Trans. on Signal Processing*, Vol. 48, No. 11, pp. 3227-3231, 2000.

[9] E. Fridman and U. Shaked, "Regional nonlinear H_∞ filtering", *UKACC Inter. Conf. on Control*, pp. 1102-1106, 1996.

[10] J.C. Geromel, J. Bernussou, G. Garcia and M.C. de Oliveira, "H_2 and H_∞ robust filtering for discrete-time linear systems", *Proceedings of the 37th IEEE Conf. on Decision & Control*, pp. 632-637, 1998.

[11] B. Hassibi and T. Kailath, "H^∞ bounds for least-squares estimators", *IEEE trans. on Automatic control*, Vol. 46, No. 2, pp. 309-314, 2001.

[12] H. Hindi and S. Boyd, "Analysis of linear systems with saturation using convex optimization", *Proceedings of the 37th IEEE Conf. on Decision & Control*, pp. 903-908, 1998.

[13] T. Hu, Z. Lin and B. M. Chen, "Analysis and design for linear discrete-time systems subject to actuator saturation", *Systems & Control Letters*, Vol. 45, No. 2, pp. 97-112, 2002.

[14] H.K. Khalil, *Nonlinear Systems*, Prentice Hall, 2nd Ed. 1996.

[15] P.P. Khargonekar and M.A. Rotea, "Mixed H_2/H_∞ filtering", *Proceedings of the 31st Conf. on Decision and Control*, pp. 2299-2304, 1992.

[16] R.D. Koilpillai, T.Q. Nguyen and P. P. Vaidyanathan, "Some results in the theory of crosstalk-free transmultiplexers", *IEEE Trans. on Signal Processing*, Vol. 39, No. 10, pp. 2174-2183, 1991.

[17] C.-W. Lin and B.-S. Chen, "State space model and noise filtering design in transmultiplexer systems", *Signal Processing*, Vol. 43, pp. 65-78, 1995.

[18] D. Liu and A. N. Michel, "Asymptotic stability of discrete-time systems with saturation nonlinearities with application to digital filters", *IEEE Trans. on Circuits and Systems I*, Vol. 39, No. 10, pp. 798-807, 1992.

[19] D. Liu and A. N. Michel, *Dynamical Systems with Saturation Nonlinearities: Analysis and Design*, Lecture Notes in Control and Information Sciences, Springer-Verlag, 1994.

[20] D. Liu and A. N. Michel, "Stability analysis of systems with partial state saturation nonlinearities", *IEEE Trans. on Circuits and Systems I*, Vol. 43, No. 3, pp. 230-232, 1996.

[21] R.M. Palhares and P.L.D. Peres, "LMI approach to the mixed H_2/H_∞ filtering design for discrete-time uncertain systems", *IEEE Trans. on Aerospace & Electronic Systems*, Vol. 37, No. 1, pp. 292-296, 2001.

[22] B. Sayyarrodsari, J. P. How, B. Hassibi and A. Carrier, "Estimation- based synthesis of H_∞-optimal adaptive FIR filters for filtered-LMS problems", *IEEE Trans. on Signal Processing*, Vol. 49, No. 1, pp. 164-178, 2001.

[23] U. Shaked and N. Berman, "H_∞ nonlinear filtering of discrete-time processes", *IEEE Trans. on Signal Processing*, Vol. 43, No. 9, pp. 2205-2209, 1995.

[24] Z. Tan, Y. C. Soh and L. Xie, "Envelope-constrained H_∞ filter design: an LMI optimization approach", *IEEE Trans. on Signal Processing*, Vol. 48, No. 10, pp. 2960-2963, 2000.

Publication 29

Submitted to *IEEE Transactions on Fuzzy Systems*

Robust Stability Analysis and Fuzzy-Scheduling Control for Nonlinear Systems Subject to Actuator Saturation *

Yong-Yan Cao Zongli Lin

Department of Electrical and Computer Engineering
University of Virginia
Charlottesville, VA 22903
email: {yycao, zl5y}@virginia.edu

March 8

Abstract

Takagi-Sugeno (TS) fuzzy models can provide an effective representation of complex nonlinear systems in terms of fuzzy sets and fuzzy reasoning applied to a set of linear input-output submodels. In this paper, the TS fuzzy modeling approach is utilized to carry out the stability analysis and control design for nonlinear systems with actuator saturation. The TS fuzzy representation of a nonlinear system subject to actuator saturation is presented. In our TS fuzzy representation, the modeling error is also captured by norm-bounded uncertainties. A set invariance condition for the system in the TS fuzzy representation is first established. Based on this set invariance condition, the problem of estimating the domain of attraction of a TS fuzzy system under a constant state feedback law is formulated and solved as an LMI optimization problem. By viewing the state feedback gain as an extra free parameter in the LMI optimization problem, we arrive at a method for designing state feedback gain that maximizes the domain of attraction. A fuzzy scheduling control design method is also introduced to further enlarge the domain of attraction. An inverted pendulum is used to show the effectiveness of the proposed fuzzy controller.

Key Words: nonlinear systems, Fuzzy control, uncertainty, linear matrix inequality (LMI), robust control, actuator saturation.

*This work was supported in part by the US office of Naval Research Young Investigator Program under grant N00014-99-1-0670.

1 Introduction

Fuzzy logic control [24] is an effective approach to designing nonlinear control systems, especially in the absence of complete knowledge of the plant. It has found successful applications not only in consumer products but also in industrial processes (see, *e.g.*, [6, 12, 15] and the references therein). Recently, a conceptually simple nonlocal approach to fuzzy control design was proposed for nonlinear systems [17, 18, 22, 21]. The procedure is as follows: First the nonlinear plant is represented by a so-called Takagi-Sugeno (TS) type fuzzy model. In this type of fuzzy model, local dynamics in different state space regions are represented by linear models. The overall model of the system is obtained by fuzzy "blending" of these local models. The control design is carried out based on the fuzzy model by the so-called parallel distributed compensation (PDC) scheme [18, 22]. For each local linear model, a linear feedback control is designed. The resulting overall controller, which is nonlinear in general, is again a fuzzy blending of those individual linear controllers.

Actuator saturation can severely degrade the closed-loop system performance and sometimes even make the otherwise stable closed-loop system unstable. The analysis and synthesis of control systems with actuator saturation nonlinearities have been receiving increasing attention recently (see, *e.g.*, [8, 14] and the references therein). Very often, actuator saturation is dealt with by either designing low gain control laws that, for a given bound on the initial conditions, avoid the saturation limits, or estimating the domain of attraction in the presence of actuator saturation. In this paper, we will utilize the TS fuzzy modeling approach to analyze the domain of attraction of nonlinear systems with actuator saturation. In our analysis procedure, a given nonlinear system with actuator saturation is first represented by a set of TS models with input saturation. The system dynamics is captured by a set of fuzzy implications which characterize local relations in the state space. The main feature of the TS fuzzy model is to express the local dynamics of each fuzzy rule by a linear state space system model with input saturation. The overall fuzzy model of the system is obtained by fuzzy "blending" of those individual models with input saturation. In [20], actuator saturation constraint was dealt with by designing low gain control laws that, for a given bound on the initial conditions, avoid the saturation limits. It is known that low-gain controllers that avoid saturation will often result in low levels of performance. The domain of attraction was not discussed either.

This paper takes the fuzzy control approach to dealing with stability analysis and control design of nonlinear systems with actuator saturation. A TS fuzzy model with input saturation and norm-bounded uncertainties is proposed to represent the original nonlinear systems with actuator saturation. A set invariance condition for the system in the TS fuzzy representation is first established. Based on this set invariance condition, the problem of estimating the domain of attraction of a TS fuzzy system under a constant state feedback law is formulated and solved as an LMI optimization problem. By viewing the state feedback gain as an extra free parameter in the LMI optimization problem, we arrive at a method for designing state feedback gain that maximizes the domain of attraction.

The paper is organized as follows. In Section 2, the TS fuzzy model with input saturation is first introduced and a set invariance condition is derived using Lyapunov function based approach. In Section 3, robust stability condition is given for the TS fuzzy model with input saturation and modeling uncertainties. In Section 4, a robust state feedback fuzzy scheduling control law for the uncertain fuzzy systems with actuator saturation is proposed based on the parallel distributed compensation. In

Section 5, an inverted pendulum subject to actuator saturation is used to demonstrate the effectiveness of our analysis and design method. The paper is concluded in Section 6.

Notations: The following notations will be used throughout the paper. $\mathbb{R}$ denotes the set of real numbers, $\mathbb{R}^+$ the set of nonnegative real numbers, $\mathbb{R}^m$ the m dimensional Euclidean space, and $\mathbb{R}^{n\times m}$ the set of all $n \times m$ real matrices. In the sequel, if not explicitly stated, matrices are assumed to have compatible dimensions. The notation $M > (\geq, <, \leq)0$ is used to denote a symmetric positive definite (positive semidefinite, negative definite, negative semidefinite, respectively) matrix.

2 Problem Statement and Preliminaries

2.1 Problem Statement

Consider a nonlinear system described by

$$\dot{x}(t) = f(x(t), v(t)), \tag{1}$$

where $x \in \mathbb{R}^n$, $v \in \mathbb{R}^m$ and f is sufficiently smooth in x and affine in v. The control input v is subject to actuator saturation. Our goal is to design a state feedback controller

$$v(t) = \sigma(u(t)),\ u(t) = F(x(t)), \tag{2}$$

such that the origin of the closed-loop system is asymptotically stable with a domain of attraction as large as possible, where the function $\sigma : \mathbb{R}^m \to \mathbb{R}^m$ is the standard saturation function of appropriate dimensions defined as follows

$$\sigma(u) = \left[\ \sigma(u_1)\quad \sigma(u_2)\quad \cdots\quad \sigma(u_m)\ \right]^{\mathrm{T}},$$

where $\sigma(u_i) = \mathrm{sign}(u_i)\min\{1, |u_i|\}$. Here we have slightly abused the notation by using σ to denote both the scalar valued and the vector valued saturation functions. Also note that it is without loss of generality to assume unity saturation level. The non-unity saturation level can be absorbed into the input by applying the following substitution

$$\hat{v} = vU,\ \ \hat{u} = U^{-1}u,$$

where $U = \mathrm{diag}(u_{\max,i})$, and $u_{\max,i}$ is the saturation amplitude of the i-th input.

We will develop a simplified model for which control design is easier. Such a simplified model is labeled as the design model. In many situations, there may be human experts who can provide a linguistic description of the system in terms of IF-THEN rules. For example, Takagi and Sugeno [17] proposed an approach to modeling the nonlinear process. This method is further developed by Sugeno and Kang [16]. This type of models are referred as Takagi-Sugeno (TS) or Takagi-Sugeno-Kang (TSK) fuzzy models [19, 2, 21].

The fuzzy model is described by fuzzy IF-THEN rules, which represent local linear input-output relations of a nonlinear system. The ith rules of the fuzzy models are of the following form

$$\begin{aligned}\textbf{Plant Rule } i: \quad & \text{IF } z_1(t) \text{ is } M_{i1} \text{ and } \cdots \text{ and } z_q(t) \text{ is } M_{iq} \text{ THEN} \\ & \dot{x}(t) = A_i x(t) + B_i v(t),\ \ i = 1, 2, \cdots, r,\end{aligned} \tag{3}$$

where M_{ij} is the fuzzy set and r is the number of IF-THEN rules and $z_1(t), z_2(t), \cdots, z_q(t)$ are the premise variables. It is assumed in this paper that the premise variables do not explicitly depend on the input variables v. Then, given a pair $(x(t), v(t))$, the resulting fuzzy system model is inferred as the weighted average of the local models and has the form

$$\dot{x}(t) = A(p(t))x(t) + B(p(t))\sigma(u(t)), \tag{4}$$

where

$$\begin{aligned} A(p(t)) &= \sum_{i=1}^{r} p_i(t)A_i, \quad B(p(t)) = \sum_{i=1}^{r} p_i(t)B_i, \\ p_i(z(t)) &= w_i(z(t))/\sum_{j=1}^{r} w_j(z(t)), \\ w_i(z(t)) &= \prod_{j=1}^{p} M_{ij}(z_j(t)), \end{aligned}$$

and where $M_{ij}(z_j(t))$ is the grade of membership of $z_j(t)$ in M_{ij}. In this paper, we assume that all membership functions are continuous and piecewise continuously differentiable. It is easy to find that the time-varying parameter vector $p(t)$ belongs to a convex polytope $\mathcal{P}$, where

$$\mathcal{P} := \left\{ p \in \mathbb{R}^r : \sum_{j=1}^{r} p_j(t) = 1, \ \ 0 \le p_j(t) \le 1 \right\}. \tag{5}$$

Therefore, when $p_i(t) = 1$ and $p_j(t) = 0$ for $i, j \in [1,\ r], j \neq i$, the fuzzy model (4) reduces to its i-th linear time-invariant "local" model, *i.e.*, $(A(p),\ B(p)) = (A_i,\ B_i)$. It is clear that as p varies inside the polytope $\mathcal{P}$, the system matrices of (4) vary inside a corresponding polytope Ω whose vertices consist of r local system matrices

$$[A(p(t)),\ B(p(t))] \in \Omega = \mathrm{co}\left\{(A_i,\ B_i),\ \ i \in [1, r]\right\}, \tag{6}$$

where co denotes the convex hull.

Based on the parallel distributed compensation (PDC) [18, 22], we consider the following fuzzy control law for the fuzzy model (4)

$$\begin{aligned} \textbf{Control Rule } i: \quad & \text{IF } z_1(t) \text{ is } M_{i1} \text{ and } \cdots \text{ and } z_q(t) \text{ is } M_{iq} \text{ THEN} \\ & u(t) = F_i x(t), \quad i = 1, 2, \cdots, r. \end{aligned} \tag{7}$$

The overall state feedback fuzzy control law is represented by

$$u(t) = \frac{\sum_{i=1}^{r} w_i(z(t))F_i x(t)}{\sum_{i=1}^{r} w_i(z(t))} = \sum_{i=1}^{r} p_i(z(t))F_i x(t). \tag{8}$$

Because the fuzzy models (4) are subject to input saturation, in general, global stabilizing controllers do not exist. The aim of this paper is to design r local linear state feedback law (7) or a time-varying parameter-dependent linear state feedback law (8) such that the origin of the closed-loop

system with actuator saturation is asymptotically stable in a region as large as possible. For simplicity, we will first consider the following linear feedback control law

$$u(t) = Fx(t). \tag{9}$$

This control law can also be obtained by setting $F_i = F$ in (7). Control law (9) is a constant feedback law, while (8) is a time-varying feedback law. Control law (8) is the so-called fuzzy scheduling controller.

2.2 Set Invariance Analysis for Fuzzy Systems

Let f_i be the i-th row of the matrix F. We define the symmetric polyhedron

$$\mathcal{L}(F) = \{x \in \mathbb{R}^n : |f_i x| \leq 1, \quad i = 1, 2, \cdots, m\}.$$

If the control u does not saturate for all $i = 1, 2, \cdots, m$, that is $x \in \mathcal{L}(F)$, then the system (4) under the control (9) admits the following linear representation

$$\dot{x}(t) = (A(p(t)) + B(p(t))F)x(t). \tag{10}$$

Let $P \in \mathbb{R}^{n\times n}$ be a positive-definite matrix and define $V = x^{\mathrm{T}}Px$. For a positive number ρ, denote the ellipsoid

$$\Omega(P, \rho) = \{x \in \mathbb{R}^n : x^{\mathrm{T}}Px \leq \rho\}.$$

An ellipsoid is said to be contractively invariant if

$$\dot{V} = 2x^{\mathrm{T}}Pf(x, v) < 0, \quad \forall x \in \Omega(P, \rho)\backslash\{0\}.$$

Thus, if an ellipsoid is contractively invariant, it is inside the domain of attraction. An ellipsoid $\Omega(P, \rho)$ is inside $\mathcal{L}(F)$ if and only if

$$f_i(P/\rho)^{-1}f_i^{\mathrm{T}} \leq 1, \quad i = 1, 2, \cdots m.$$

Let $\mathcal{V}$ be the set of $m \times m$ diagonal matrices whose diagonal elements are either 1 or 0. For example, if $m = 2$, then

$$\mathcal{V} = \left\{ \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\}.$$

There are 2^m elements in $\mathcal{V}$. Suppose that each element of $\mathcal{V}$ is labeled as E_i, $i = 1, 2, \cdots, 2^m$, and denote $E_i^- = I - E_i$. Clearly, E_i^- is also an element of $\mathcal{V}$ if $E_i \in \mathcal{V}$.

Lemma 1 [8] *Let $F, H \in \mathbb{R}^{m\times n}$ be given. For $x \in \mathbb{R}^n$, if $x \in \mathcal{L}(H)$, then*

$$\sigma(Fx) \in \mathrm{co}\left\{E_iFx + E_i^-Hx : \quad i \in [1, 2^m]\right\}.$$

Consequently, $\sigma(Fx)$ can be rewritten as

$$\sigma(Fx) = \sum_{i=1}^{2^m} \eta_i(E_iF + E_i^-H)x,$$

where $0 \leq \eta_i \leq 1$, $\sum_{i=1}^{2^m} \eta_i = 1$.

Lemma 2 [4] *Suppose that matrices $M_i \in \mathbb{R}^{m\times n}, i = 1, 2, \cdots, r$, and a positive semi-definite matrix $P \in \mathbb{R}^{m\times m}$ are given. If $\sum_{i=1}^{r} p_i = 1$ and $0 \le p_i \le 1$, then*

$$\left(\sum_{i=1}^{r} p_i M_i\right)^{\mathrm{T}} P \left(\sum_{i=1}^{r} p_i M_i\right) \le \sum_{i=1}^{r} p_i M_i^{\mathrm{T}} P M_i. \tag{11}$$

For the fuzzy system subject to actuator saturation (4) and a given linear control law (9), we have the following set invariance condition.

Theorem 3 *For a given fuzzy system (4) and a given state feedback control matrix F, the ellipsoid $\Omega(P,\rho)$ is a contractively invariant set of the closed-loop system under linear state feedback control law (9) if there exists a matrix $H \in \mathbb{R}^{m\times n}$ such that the following matrix inequalities hold*

$$\left(A_i + B_i(E_jF + E_j^- H)\right)^{\mathrm{T}} P + P\left(A_i + B_i(E_jF + E_j^- H)\right) < 0, \quad i \in [1, r],\ j \in [1, 2^m], \tag{12}$$

and $\Omega(P,\rho) \subset \mathcal{L}(H)$. Consequently, the closed-loop system is asymptotically stable at the origin with $\Omega(P,\rho)$ contained in the domain of attraction.

Proof. Choose a Lyapunov function

$$V(x(t)) = x^{\mathrm{T}}(t) P x(t).$$

Then,

$$\dot{V}(x) = \left[A(p)x(t) + B(p)\sigma(Fx(t))\right]^{\mathrm{T}} P x(t) + x^{\mathrm{T}}(t) P \left[A(p)x(t) + B(p)\sigma(Fx(t))\right].$$

By Lemma 1, we have

$$\begin{aligned}
\dot{V}(x) &= x^{\mathrm{T}}(t)\left[\sum_{i=1}^{r} p_i A_i + \sum_{i=1}^{r} p_i B_i \sum_{j=1}^{2^m} \eta_j (E_jF + E_j^- H)\right]^{\mathrm{T}} P x(t) + \\
&\quad x^{\mathrm{T}}(t) P \left[\sum_{i=1}^{r} p_i A_i + \sum_{i=1}^{r} p_i B_i \sum_{j=1}^{2^m} \eta_j (E_jF + E_j^- H)\right] x(t) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{2^m} p_i \eta_j x^{\mathrm{T}}(t) \left[\left(A_i + B_i(E_jF + E_j^- H)\right)^{\mathrm{T}} P + P\left(A_i + B_i(E_jF + E_j^- H)\right)\right] x(t)
\end{aligned}$$

for all $x \in \mathcal{L}(H)$. By (12), we have

$$\dot{V}(x) < 0, \quad \forall x \in \Omega(P,\gamma)\backslash\{0\} \subset \mathcal{L}(H).$$

Thus, if $x_0^{\mathrm{T}} P x_0 \le \rho$, then $x^{\mathrm{T}}(t) P x(t) \le \rho$ for $t \ge 0$, *i.e.*, $\Omega(P,\rho)$ is a contractively invariant set. This also implies that the closed-loop system (4) under state feedback (9) is asymptotically stable at the origin with $\Omega(P,\rho)$ contained in the domain of attraction. ■

Remark 1 *Theorem 3 gives a condition for the region $\Omega(P,\rho)$ to be inside the domain of attraction for the closed-loop system (4) under a linear constant feedback control law (9). For the special case of $r = 1$, Theorem 3 recovers the set invariance condition for linear time-invariant systems subject to actuator saturation [8].*

With all the ellipsoids satisfying the set invariance condition of Theorem 3, we may choose the "largest" one to obtain the least conservative estimate of the domain of attraction. As in [8], we will measure the largeness of the ellipsoids with respect to a shape reference set. Let $\mathcal{X}_R \subset \mathbb{R}^n$ be a prescribed bounded convex set containing origin. For a set $\mathcal{S} \subset \mathbb{R}^n$ which contains origin, define

$$\alpha_R(\mathcal{S}) := \sup\{\alpha > 0 : \alpha\mathcal{X}_R \subset \mathcal{S}\}.$$

Obviously, if $\alpha_R(\mathcal{S}) \geq 1$, then $\mathcal{X}_R \subset \mathcal{S}$. Two typical types of $\mathcal{X}_R$ are the ellipsoid

$$\mathcal{X}_R = \{x \in \mathbb{R}^n : x^{\mathrm{T}} R x \leq 1\}, \quad R > 0, \tag{13}$$

and the polyhedron

$$\mathcal{X}_R = \mathrm{co}\{x_0^1, x_0^2, \cdots, x_0^l\}, \tag{14}$$

where $x_0^1, x_0^2, \cdots, x_0^l$ are *a priori* given points in $\mathbb{R}^n$.

With the above reference sets, we can choose an $\Omega(P,\rho)$ from all that satisfy the condition such that the quantity $\alpha_R(\Omega(P,\rho))$ is maximized. This problem can be formulated into the following optimization problem:

$$\begin{aligned} &\min_{P>0,H} \alpha, \\ \text{s.t.} \quad &a)\ \alpha\mathcal{X}_R \subset \Omega(P,\rho), \\ &b)\ \text{inequalities (12)}, \quad \forall i \in [1,r],\ j \in [1,2^m], \\ &c)\ |h_i x| \leq 1, \quad \forall x \in \Omega(P,\rho), \quad i = [1,m], \end{aligned} \tag{15}$$

where h_i denotes the i-th row of H.

In what follows, we will show that the optimization problem (15) can be solved as an LMI optimization problem.

In the case that the shape reference set $\mathcal{X}_R$ is given by (14), Constraint a) is equivalent to

$$\alpha^2 (x_0^i)^{\mathrm{T}} P x_0^i \leq \rho \Longleftrightarrow \begin{bmatrix} \alpha^{-2} & (x_0^i)^{\mathrm{T}} \\ x_0^i & (P/\rho)^{-1} \end{bmatrix} \geq 0, \quad i = [1,l]. \tag{16}$$

In the case that $\mathcal{X}_R$ is given by (13), Constraint a) is equivalent to

$$\alpha^{-2} R \geq P/\rho \Longleftrightarrow R^{-1} \leq \alpha^{-2}(P/\rho)^{-1}. \tag{17}$$

Let

$$Q = (P/\rho)^{-1},\ Z = HQ,\ \gamma = \alpha^{-2}.$$

Then, (16) and (17) can be written as the following LMIs

$$\begin{bmatrix} \gamma & (x_0^i)^{\mathrm{T}} \\ x_0^i & Q \end{bmatrix} \geq 0, \quad i = [1,l], \tag{18}$$

and

$$R^{-1} \leq \gamma Q, \tag{19}$$

respectively. Also, Constraint b) is equivalent to

$$\left(A_i Q + B_i(E_j F Q + E_j^- Z)\right)^{\mathrm{T}} + \left(A_i Q + B_i(E_j F Q + E_j^- Z)\right) < 0, \forall i \in [1,r],\ \forall j \in [1,2^m]. \tag{20}$$

Constraint c) is equivalent to

$$h_i\left(\frac{P}{\rho}\right)^{-1}h_i^{\mathrm{T}} \le 1 \Longleftrightarrow \begin{bmatrix} 1 & h_iQ \\ Qh_i^{\mathrm{T}} & Q \end{bmatrix} \ge 0, \quad \forall i \in [1,m].$$

Also let the i-th row of Z be z_i, *i.e.*, $z_i = h_iQ$. The optimization problem (15) can then be reduced to the following one with LMI constraints,

$$\begin{aligned} & \min_{Q>0,Z} \gamma, \\ \text{s.t.} \quad & a)\ \text{LMI (18) or (19)}, \\ & b)\ \text{LMI (20)}, \\ & c)\ \begin{bmatrix} 1 & z_i \\ z_i^{\mathrm{T}} & Q \end{bmatrix} \ge 0, \quad \forall i \in [1,m]. \end{aligned} \tag{21}$$

If we would like to design a control law $u = Fx$ such that the domain of attraction of the closed-loop system is as large as possible, we only need to replace $Y = FQ$ in (20). With the solution (Q, Y, Z), the state feedback control matrix F such that the origin of the system (4) is stabilized with a domain of attraction as large as possible with respect to a given shape reference $\mathcal{X}_R$ can be solved by

$$F = YQ^{-1}.$$

In optimization problem (21), the amplitude of control law (9) is not constrained, *i.e.*, there is no control amplitude constraint on the control law. In [9], the authors proved that this controller design method is less conservative than the approaches based on circle criterion and Popov criterion [7]. On the other hand, to avoid the controller gain being too large, we may constrain it to be bounded by $\mu_0 > 1$, *i.e.*, $|f_ix| \le \mu_0$, which is equivalent to the following LMIs

$$\begin{bmatrix} \mu_0^2 & y_i \\ y_i^{\mathrm{T}} & Q \end{bmatrix} \ge 0, \quad \forall i \in [1,m],$$

where y_i denotes i-th row of Y.

If we require $Y = Z$, then we can recover the design algorithm which constrains the control law to be unsaturated [1, 20]. The unsaturated control algorithm can be described as:

$$\begin{aligned} & \min_{Q>0,Y} \gamma, \\ \text{s.t.} \quad & a)\ \text{LMI (18) or (19)}, \\ & b)\ (A_iQ + B_iY)^{\mathrm{T}} + (A_iQ + B_iY) < 0, \quad \forall i \in [1,r], \\ & c)\ \begin{bmatrix} 1 & y_i \\ y_i^{\mathrm{T}} & Q \end{bmatrix} \ge 0, \quad \forall i \in [1,m]. \end{aligned} \tag{22}$$

Note that the constraints in (22) imply that $\Omega(Q^{-1}, 1) \subset \mathcal{L}(F)$ and hence the control $u = Fx$ will never reach saturation limits. This will lead to a very conservative control law. In (21), we permit the control to be saturated and hence our algorithm will result in a larger domain of attraction. It is also known that low-gain controllers that avoid saturation will often result in low levels of performance [14].

3 Robust Stability of Fuzzy Systems with Uncertainty

In the previous section, we consider the stability of the fuzzy system (4) rather than the original nonlinear system (1). It is an obvious fact that the closed-loop stability of fuzzy system (4) cannot guarantee that of nonlinear system (1). As discussed in [3], we can present a fuzzy model with norm-bounded uncertainty to analyze the stability of the original nonlinear system (1). A TS fuzzy model with uncertainty is composed of r plant rules that can be represented as [25, 11, 13],

$$\begin{aligned} \textbf{Plant Rule } i: \quad & \text{IF } z_1(t) \text{ is } M_{i1} \text{ and } \cdots \text{ and } z_q(t) \text{ is } M_{iq} \text{ THEN} \\ & \dot{x}(t) = \hat{A}_i x(t) + \hat{B}_i v(t), \quad i = 1, 2, \cdots, r, \end{aligned} \tag{23}$$

where $\hat{A}_i$, and $\hat{B}_i$ are real-valued time-varying matrices of appropriate dimensions. Fuzzy model (23) is an extension of local fuzzy model (3). We assume that the time-varying uncertainties enter the system matrices in the following manner:

$$\hat{A}_i(t) = A_i + \Delta A_i(t), \quad \hat{B}_i(t) = B_i + \Delta B_i(t), \tag{24}$$

where A_i and B_i are some constant matrices of compatible dimensions and $\Delta A_i(t)$ and $\Delta B_i(t)$ are real-valued matrix functions of compatible dimensions representing time-varying parameter uncertainties. Such uncertainties arise in the fuzzy representation (3) of the original nonlinear system (1). The uncertainties are assumed to be norm-bounded and be given by

$$\left[\begin{array}{cc} \Delta A_i(t) & \Delta B_i(t) \end{array} \right] = M_i \Theta_i(t) \left[\begin{array}{cc} N_{1i} & N_{2i} \end{array} \right], \tag{25}$$

where M_i, N_{1i} and N_{2i} are known constant matrices with compatible dimensions and $\Theta_i(t) \in \mathbb{R}^{n_{f1i} \times n_{f2i}}$ are unknown nonlinear time-varying matrix functions satisfying

$$\Theta_i^{\mathrm{T}}(t)\Theta_i(t) \leq I. \tag{26}$$

It is assumed that the elements of $\Theta_i(t)$ are Lebesgue measurable. This type of uncertainty is an effective representation of some nonlinear uncertainties (see [5, 10, 23] and the references therein). By including these uncertainties, the fuzzy model (23) is expected to better represent the original system (1).

Now, given a pair of (x, v), the final fuzzy system is inferred as follows:

$$\dot{x}(t) = \hat{A}(p(t))x(t) + \hat{B}(p(t))v(t) \tag{27}$$

$$= (A(p(t)) + \Delta A(p,t))x(t) + (B(p(t)) + \Delta B(p,t))v(t), \tag{28}$$

where

$$\hat{A}(p) = \sum_{i=1}^{r} p_i(z(t))\hat{A}_i, \qquad \hat{B}(p) = \sum_{i=1}^{r} p_i(z(t))\hat{B}_i$$

$$\Delta A(p,t) = \sum_{i=1}^{r} p_i M_i \Theta_i(t) N_{1i}, \quad \Delta B(p,t) = \sum_{i=1}^{r} p_i M_i \Theta_i(t) N_{2i},$$

and $\Delta f := \Delta A x + \Delta B v$ represents the difference between (4) and (1). By suitably selecting M_i, N_{1i} and N_{2i}, we can make the original nonlinear system (1) completely included in the differential inclusions system (27) [1, 11, 3].

Theorem 4 *For a given uncertain fuzzy system (27) and a given state feedback control matrix F, the ellipsoid $\Omega(P,\rho)$ is a contractively invariant set of the closed-loop system under linear state feedback control law (9) if there exist matrices $H \in \mathbb{R}^{m\times n}$ and $X_i > 0$ such that*

$$\begin{bmatrix} P\left(A_i + B_i(E_jF + E_j^- H)\right) + (*)^{\mathrm{T}} + PM_iX_iM_i^{\mathrm{T}}P & (*)^{\mathrm{T}} \\ N_{1i} + N_{2i}(E_jF + E_j^- H) & X_i \end{bmatrix} < 0, \quad i \in [1,r],\ j \in [1,2^m], \tag{29}$$

and $\Omega(P,\rho) \subset \mathcal{L}(H)$, where $*$'s represent blocks that are readily inferred by symmetry.

Proof. Choose a Lyapunov function

$$V(x) = x^{\mathrm{T}}(t)Px(t).$$

Then,

$$\dot{V}(x) = \left[\hat{A}(p)x(t) + \hat{B}(p)\sigma(Fx(t))\right]^{\mathrm{T}} Px(t) + x^{\mathrm{T}}(t)P\left[\hat{A}(p)x(t) + \hat{B}(p)\sigma(Fx(t))\right].$$

By Lemma 1, we have for all $x \in \mathcal{L}(H)$,

$$\begin{aligned}
\dot{V}(x) &= x^{\mathrm{T}}(t)\left[\sum_{i=1}^{r} p_i\left(\hat{A}_i + \hat{B}_i\sum_{j=1}^{2^m}\eta_j(E_jF + E_j^- H)\right)\right]^{\mathrm{T}} Px(t) \\
&\quad + x^{\mathrm{T}}(t)P\left[\sum_{i=1}^{r} p_i\left(\hat{A}_i + \hat{B}_i\sum_{j=1}^{2^m}\eta_j(E_jF + E_j^- H)\right)\right]x(t) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{2^m} p_i\eta_j x^{\mathrm{T}}(t)\left[\left(\hat{A}_i + \hat{B}_i(E_jF + E_j^- H)\right)^{\mathrm{T}} P + P\left(\hat{A}_i + \hat{B}_i(E_jF + E_j^- H)\right)\right]x(t) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{2^m} p_i\eta_j x^{\mathrm{T}}(t)\left[\left(A_i + B_i(E_jF + E_j^- H)\right)^{\mathrm{T}} P + P\left(A_i + B_i(E_jF + E_j^- H)\right)\right. \\
&\quad \left. + \left(N_{1i} + N_{2i}(E_jF + E_j^- H)\right)^{\mathrm{T}}\Theta_i^{\mathrm{T}}(t)M_i^{\mathrm{T}}P + PM_i\Theta_i(t)\left(N_{1i} + N_{2i}(E_jF + E_j^- H)\right)\right]x(t) \\
&\le \sum_{i=1}^{r}\sum_{j=1}^{2^m} p_i\eta_j x^{\mathrm{T}}(t)\left[\left(A_i + B_i(E_jF + E_j^- H)\right)^{\mathrm{T}} P + P\left(A_i + B_i(E_jF + E_j^- H)\right)\right. \\
&\quad \left. + PM_iX_iM_i^{\mathrm{T}}P + \left(N_{1i} + N_{2i}(E_jF + E_j^- H)\right)^{\mathrm{T}} X_i^{-1}\left(N_{1i} + N_{2i}(E_jF + E_j^- H)\right)\right]x(t).
\end{aligned}$$

By (29), we have

$$\dot{V}(x) < 0, \quad \forall x \in \Omega(P,\gamma)\backslash\{0\}.$$

Thus, $\Omega(P,\rho)$ is a contractively invariant set. This also implies that the closed-loop system (27) under state feedback (9) is asymptotically stable at the origin with $\Omega(P,\rho)$ contained in the domain of attraction. ■

It is easy to see that matrix inequality (29) can be transformed to the following LMIs

$$\begin{bmatrix} \left(A_iQ + B_i(E_jY + E_j^- Z)\right) + (*)^{\mathrm{T}} + M_iX_iM_i^{\mathrm{T}} & (*)^{\mathrm{T}} \\ N_{1i}Q + N_{2i}(E_jY + E_j^- Z) & X_i \end{bmatrix} < 0, \quad i \in [1,r],\ j \in [1,2^m], \tag{30}$$

As in Section 2.2, we can formulate the following LMI optimization to maximize the domain of attraction of the closed-loop system:

$$\begin{aligned} & \min_{Q>0, X_i, Y, Z} \gamma, \\ \text{s.t.} \quad & a) \text{ LMI (18) or (19)}, \\ & b) \text{ LMI (30)}, \\ & c) \begin{bmatrix} 1 & z_i \\ z_i^{\mathrm{T}} & Q \end{bmatrix} \geq 0, \quad \forall i \in [1, m]. \end{aligned} \tag{31}$$

4 Robust Fuzzy Scheduling Control Law Design

As shown in Section 2, the approach to fuzzy scheduling involves the design of several LTI controllers for a parameterized family of LTI fuzzy models. The resulting controller (8) is the interpolation of these gains. Note that F in (9) is a constant matrix, while the control gain in (8) is a time-varying matrix function of time-varying membership function $z(t)$. It is reasonable to expect that this kind of control laws can result in a larger domain of attraction and better performance.

With control law (8), the closed-loop system can be rewritten as

$$\dot{x}(t) = \sum_{i=1}^{r} p_i(A_i + \Delta A_i)x(t) + \sum_{i=1}^{r} p_i(B_i + \Delta B_i)\sigma\left(\tilde{F}x(t)\right),$$

where

$$\tilde{F} = \sum_{j=1}^{r} p_j(z(t))F_j.$$

By Lemma 1, we have that for any matrix $\tilde{H}$ of the same dimensions of $\tilde{F}$ such that $x \in \mathcal{L}(\tilde{H})$,

$$\begin{aligned} \dot{x}(t) &= \sum_{i=1}^{r} p_i \left[(A_i + \Delta A_i)x(t) + (B_i + \Delta B_i) \sum_{s=1}^{2^m} \eta_s(t)(E_s\tilde{F} + E_s^-\tilde{H})x(t) \right] \\ &= \sum_{i=1}^{r}\sum_{s=1}^{2^m} p_i\eta_s \left[(A_i + \Delta A_i) + (B_i + \Delta B_i)(E_s\tilde{F} + E_s^-\tilde{H}) \right] x(t), \end{aligned}$$

where $0 \leq \eta_s(t) \leq 1$, $\sum_{s=1}^{2^m} \eta_s(t) = 1$, for all $s = 1, 2, \cdots, 2^m$. If we let

$$\tilde{H} = \sum_{j=1}^{r} p_j H_j,$$

then

$$\dot{x}(t) = \sum_{s=1}^{2^m}\sum_{i=1}^{r}\sum_{j=1}^{r} \eta_s p_i p_j \tilde{A}_{s,i,j} x(t), \tag{32}$$

where

$$\tilde{A}_{s,i,j} := (A_i + \Delta A_i) + (B_i + \Delta B_i)(E_sF_j + E_s^-H_j), \quad s \in [1, 2^m],\ i, j \in [1, r].$$

Remark 2 *It is easy to verify that the closed-loop system described by (32) can be further simplified if the subsystem (23) possesses a common input matrix $\hat{B}$, namely $\hat{B}_i = \hat{B}$ for all i. In this case, the closed-loop system can be simplified as*

$$\dot{x}(t) = \sum_{s=1}^{2^m} \sum_{i=1}^{r} \eta_s p_i \left((A_i + \Delta A_i) + (B + \Delta B)(E_s F_i + E_s^- H_i) \right) x(t).$$

Theorem 5 *For a given uncertain fuzzy system (27), suppose that the local state feedback control matrices F_j, $j = 1, 2, \cdots, r$, are given. The ellipsoid $\Omega(P, \rho)$ is a contractively invariant set of the closed-loop system under the fuzzy scheduling state feedback law (8) if there exist matrices $H_j \in \mathbb{R}^{m \times n}$, and $X_{ij} > 0$, $i, j = 1, 2, \cdots, r$, such that*

$$\begin{bmatrix} P\left(A_i + B_i(E_s F_j + E_s^- H_j)\right) + (*)^{\mathrm{T}} + P M_i X_{ij} M_i^{\mathrm{T}} P & (*)^{\mathrm{T}} \\ N_{1i} + N_{2i}(E_s F_j + E_s^- H_j) & X_{ij} \end{bmatrix} < 0, \quad i, j \in [1, r],\ s \in [1, 2^m], \tag{33}$$

and $\Omega(P, \rho) \subset \bigcap_{i=1}^{r} \mathcal{L}(H_i)$.

Proof. Choose a Lyapunov function $V(x(t)) = x^{\mathrm{T}}(t) P x(t)$. We note that

$$x \in \bigcap_{j=1}^{r} \mathcal{L}(H_j),$$

implies

$$x \in \mathcal{L}\left(\sum_{j=1}^{r} p_j H_j \right),$$

since $\sum_{j=1}^{r} p_j = 1$ and $0 \le p_j \le 1$. Let

$$\tilde{H} = \sum_{j=1}^{r} p_j H_j.$$

Then, by Lemma 1,

$$\begin{aligned} \dot{V}(x) &= x^{\mathrm{T}}(t) \left\{ \left[\sum_{i=1}^{r} p_i \sum_{s=1}^{2^m} \eta_s \left(\hat{A}_i + \hat{B}_i(E_s \tilde{F} + E_s^- \tilde{H}) \right) \right]^{\mathrm{T}} P \right. \\ &\quad \left. + P \left[\sum_{i=1}^{r} p_i \sum_{s=1}^{2^m} \eta_s \left(\hat{A}_i + \hat{B}_i(E_s \tilde{F} + E_s^- \tilde{H}) \right) \right] \right\} x(t) \\ &= x^{\mathrm{T}}(t) \left\{ \sum_{s=1}^{2^m} \eta_s \sum_{i=1}^{r} \sum_{j=1}^{r} p_i p_j \left(\tilde{A}_{s,i,j}^{\mathrm{T}} P + P \tilde{A}_{s,i,j} \right) \right\} x(t), \end{aligned} \tag{34}$$

for all $x \in \mathcal{L}(\tilde{H})$. It is easy to see that $\dot{V}(x) < 0$ if

$$\tilde{A}_{s,i,j}^{\mathrm{T}} P + P \tilde{A}_{s,i,j} < 0,$$

for all $i, j \in [1, r]$, and $s \in [1, 2^m]$. Note that

$$\begin{aligned} \tilde{A}_{s,i,j} &= A_i + \Delta A_i + (B_i + \Delta B_i)(E_s F_j + E_s^- H_j) \\ &= A_i + B_i(E_s F_j + E_s^- H_j) + M_i \Theta_i(t) \left(N_{1i} + N_{2i}(E_s F_j + E_s^- H_j) \right), \end{aligned}$$

and

$$\begin{aligned} & PM_i\Theta_i(t)\left(N_{1i}+N_{2i}(E_sF_j+E_s^-H_j)\right)+(*)^{\mathrm{T}} \\ \leq\ & PM_iX_{ij}M_i^{\mathrm{T}}P+\left(N_{1i}+N_{2i}(E_sF_j+E_s^-H_j)\right)^{\mathrm{T}}X_{ij}^{-1}\left(N_{1i}+N_{2i}(E_sF_j+E_s^-H_j)\right). \end{aligned}$$

This implies that if the matrix inequalities (33) hold, then we have

$$\tilde{A}_{s,i,j}^{\mathrm{T}}P+P\tilde{A}_{s,i,j}<0, \quad i,j\in[1,r],\ s\in[1,2^m],$$

and hence $\dot{V}(x)<0$ for all $x\in\Omega(P,\rho)\backslash\{0\}$. That is, $\Omega(P,\rho)$ is contractively invariant. ■

Corollary 6 *For the special case of $\hat{B}_i=\hat{B}$ for all i, the ellipsoid $\Omega(P,\rho)$ is a contractively invariant set of the closed-loop system under the fuzzy scheduling state feedback control law (8), if there exist matrices $H_i\in\mathbb{R}^{m\times n}$ and $X_i>0$ such that*

$$\left[\begin{array}{cc} P\left(A_i+B_i(E_sF_i+E_s^-H_i)\right)+(*)^{\mathrm{T}}+PM_iX_iM_i^{\mathrm{T}}P & (*)^{\mathrm{T}} \\ N_{1i}+N_{2i}(E_sF_i+E_s^-H_i) & X_i \end{array}\right]<0, \quad i\in[1,r],\ s\in[1,2^m], \tag{35}$$

and $\Omega(P,\rho)\subset\bigcap_{i=1}^r\mathcal{L}(H_i)$.

In what follows, we will present a less conservative set invariance condition.

Note that system (32) can be rewritten as

$$\dot{x}(t)=\sum_{s=1}^{2^m}\eta_s\left\{\sum_{i=1}^r p_i^2\tilde{A}_{s,i,i}+\sum_{i=1}^r\sum_{j<i}^r 2p_ip_j\left(\frac{\tilde{A}_{s,i,j}+\tilde{A}_{s,j,i}}{2}\right)\right\}x(t).$$

Let

$$\bar{p}_l:=\left\{\begin{array}{ll} p_i^2 & l=i^2 \\ 2p_ip_j & l=i\cdot j \end{array}\right., \text{ for } i<j=1,2,\cdots,r,$$

$$\bar{A}_{s,l}:=\left\{\begin{array}{ll} \tilde{A}_{s,i,i} & l=i^2 \\ (\tilde{A}_{s,i,j}+\tilde{A}_{s,j,i})/2 & l=i\cdot j \end{array}\right., \text{ for } i<j=1,2,\cdots,r.$$

We then have

$$0\leq\bar{p}_l\leq 1, \quad \sum_{l=1}^{r(r+1)/2}\bar{p}_l=1,$$

and

$$\dot{x}(t)=\sum_{s=1}^{2^m}\eta_s\sum_{l=1}^{r(r+1)/2}\bar{p}_l\bar{A}_{s,l}x(t).$$

Hence,

$$\dot{V}(x)=x^{\mathrm{T}}(t)\left[\sum_{s=1}^{2^m}\sum_{l=1}^{r(r+1)/2}\eta_s\bar{p}_l\left(\bar{A}_{s,l}^{\mathrm{T}}P+P\bar{A}_{s,l}\right)\right]x(t).$$

Theorem 7 *For a given uncertain fuzzy system (27), suppose that the local state feedback control matrices F_j, $j=1,2,\cdots,r$, are known. The ellipsoid $\Omega(P,\rho)$ is a contractively invariant set of the*

closed-loop system under the fuzzy scheduling control law (8), if there exist matrices $H_j \in \mathbb{R}^{m\times n}$ *and* $X_{ij} > 0$, $i,j = 1,2,\cdots,r$, *such that*

$$\begin{bmatrix} P\left(A_i + B_i(E_sF_i + E_s^- H_i)\right) + (*)^{\mathrm{T}} + PM_iX_{ii}M_i^{\mathrm{T}}P & (*)^{\mathrm{T}} \\ N_{1i} + N_{2i}(E_sF + E_s^- H) & X_{ii} \end{bmatrix} < 0, \quad i \in [1,r],\ s \in [1,2^m], \tag{36}$$

$$\begin{bmatrix} P\left(A_i + B_i(E_sF_j + E_s^- H_j)\right) + (*)^{\mathrm{T}} & & \\ +P\left(A_j + B_j(E_sF_i + E_s^- H_i)\right) + (*)^{\mathrm{T}} & (*)^{\mathrm{T}} & (*)^{\mathrm{T}} \\ +PM_iX_{ij}M_i^{\mathrm{T}}P + PM_jX_{ji}M_j^{\mathrm{T}}P & & \\ N_{1i} + N_{2i}(E_sF_j + E_s^- H_j) & X_{ij} & 0 \\ N_{1j} + N_{2j}(E_sF_i + E_s^- H_i) & 0 & X_{ji} \end{bmatrix} < 0, \quad i < j \in [1,r],\ s \in [1,2^m], \tag{37}$$

and $\Omega(P,\rho) \subset \bigcap_{i=1}^r \mathcal{L}(H_i)$.

Proof. Choose the Lyapunov function as

$$V(x) = x^{\mathrm{T}}(t)Px(t).$$

Then,

$$\dot{V}(x) = \sum_{s=1}^{2^m}\sum_{i=1}^{r}\sum_{j=1}^{r}\eta_s p_i p_j x^{\mathrm{T}}(t)\left[\left(\hat{A}_i + \hat{B}_i(E_sF_j + E_s^- H_j)\right)^{\mathrm{T}}P + P\left(\hat{A}_i + \hat{B}_i(E_sF_j + E_s^- H_j)\right)\right]x(t).$$

It is easy to see that $\dot{V}(x) < 0$ if

$$\left(\hat{A}_i + \hat{B}_i(E_sF_i + E_s^- H_i)\right)^{\mathrm{T}}P + P\left(\hat{A}_i + \hat{B}_i(E_sF_i + E_s^- H_i)\right) < 0, \tag{38}$$

and

$$P\Phi_{s,i,j} + \Phi_{s,i,j}^{\mathrm{T}}P < 0, \tag{39}$$

where

$$\Phi_{s,i,j} = \hat{A}_i + \hat{B}_i(E_sF_j + E_s^- H_j) + \hat{A}_j + \hat{B}_j(E_sF_i + E_s^- H_i).$$

Similar to the proof of Theorem 5, we can find that matrix inequality (38) holds if (36) holds. In what follows, we will prove that matrix inequality (39) holds if (37) holds. Note that

$$\begin{aligned} \Phi_{s,i,j} \;=\; & A_i + B_i(E_sF_j + E_s^- H_j) + M_i\Theta_i(t)\left(N_{1i} + N_{2i}(E_sF_j + E_s^- H_j)\right) \\ & + A_j + B_j(E_sF_i + E_s^- H_i) + M_j\Theta_j(t)\left(N_{1j} + N_{2j}(E_sF_i + E_s^- H_i)\right), \end{aligned}$$

and

$$\begin{aligned} & PM_i\Theta_i(t)\left(N_{1i} + N_{2i}(E_sF_j + E_s^- H_j)\right) + M_j\Theta_j(t)\left(N_{1j} + N_{2j}(E_sF_i + E_s^- H_i)\right) + (*)^{\mathrm{T}} \\ \le\; & PM_iX_{ij}M_i^{\mathrm{T}}P + \left(N_{1i} + N_{2i}(E_sF_j + E_s^- H_j)\right)^{\mathrm{T}}X_{ij}^{-1}\left(N_{1i} + N_{2i}(E_sF_j + E_s^- H_j)\right) \\ & + PM_jX_{ji}M_j^{\mathrm{T}}P + \left(N_{1j} + N_{2j}(E_sF_i + E_s^- H_i)\right)^{\mathrm{T}}X_{ji}^{-1}\left(N_{1j} + N_{2j}(E_sF_i + E_s^- H_i)\right). \end{aligned}$$

Then, $P\Phi_{s,i,j} + \Phi_{s,i,j}^T P < 0$ if (37) holds. ■

Remark 3 *In comparison with Theorem 5, the number of matrix inequalities in Theorem 7 is reduced by* $r(r-1)\cdot 2^{m-1}$. *In comparison with Corollary 6, another* $r(r-1)\cdot 2^{m-1}$ *matrix inequalities can be removed for the special case of* $B_i = B, \forall i$.

Let

$$Q = (P/\rho)^{-1},\ Y_j = F_j Q,\ Z_j = H_j Q,\ j = [1, r].$$

Denote the i-th row of the matrix Z_j as z_i^j. Then (36) and (37) are equivalent to the following LMIs

$$\begin{bmatrix} (A_iQ + B_i(E_sY_i + E_s^- Z_i)) + (*)^{\mathrm{T}} + M_i X_{ii} M_i^{\mathrm{T}} & (*)^{\mathrm{T}} \\ N_{1i}Q + N_{2i}(E_jY_i + E_s^- Z_i) & X_{ii} \end{bmatrix} < 0, \quad i \in [1, r], s \in [1, 2^m], \tag{40}$$

and

$$\begin{bmatrix} (A_iQ + B_i(E_sY_j + E_s^- Z_j)) + (*)^{\mathrm{T}} & & \\ + (A_jQ + B_j(E_sY_i + E_s^- Z_i)) + (*)^{\mathrm{T}} & (*)^{\mathrm{T}} & (*)^{\mathrm{T}} \\ + M_i X_{ij} M_i^{\mathrm{T}} + M_j X_{ji} M_j^{\mathrm{T}} & & \\ N_{1i}Q + N_{2i}(E_sY_j + E_s^- Z_j) & X_{ij} & 0 \\ N_{1j}Q + N_{2j}(E_sY_i + E_s^- Z_i) & 0 & X_{ji} \end{bmatrix} < 0, \quad i < j \in [1, r],\ s \in [1, 2^m], \tag{41}$$

respectively. Then, we have the following theorem.

Theorem 8 *For a given uncertain fuzzy system (27), the fuzzy scheduling state feedback control law (8) such that the closed-loop system is robustly stable at the origin with a domain of attraction as large as possible can be solved by*

$$F_j = Y_j Q^{-1},\ j = [1, r],$$

where $(Q > 0, Y_j)$ is a solution to the following LMI optimization problem

$$\begin{aligned} & \min_{Q>0, Y_j, Z_j} \gamma, \\ \text{s.t.} \quad & a)\ \text{LMI (18) or (19)}, \\ & b)\ \text{LMIs (40) and (41)}, \\ & c)\ \begin{bmatrix} 1 & z_i^j \\ \left(z_i^j\right)^{\mathrm{T}} & Q \end{bmatrix} \geq 0, \quad \forall i \in [1, m],\ j \in [1, r]. \end{aligned} \tag{42}$$

5 An Example

Consider the problem of balancing and swing-up of an inverted pendulum on a cart. The equations of motion for the pendulum are [22],

$$\dot{x}_1 = x_2, \tag{43}$$

$$\dot{x}_2 = \frac{g\sin(x_1) - aml x_2^2 \sin(2x_1)/2 - \mu a \cos(x_1) v}{4l/3 - aml\cos^2(x_1)}, \tag{44}$$

where x_1 denotes the angle (in radians) of the pendulum from the vertical, x_2 is the angular velocity, $g = 9.8m/s^2$ is the gravity constant, m is the mass of the pendulum, M is the mass of the cart, $2l$ is the length of the pendulum, v is the force applied to the cart (in Kilo-Newtons), $a = 1/(m + M)$ and $\mu = 1000$. We choose $m = 2.0kg$, $M = 8.0kg$, $2l = 1.0m$ in the simulations.

The control objective is to balance the inverted pendulum for the approximate range $x_1 \in [-\pi/2, \pi/2]$. The actuator is subject to saturation and the saturation level is $u_{\max} = 1$. That is, we would like to design a state feedback control law

$$v = \sigma(u(x)), \tag{45}$$

such that the inverted pendulum can be balanced in the approximate range $[-\pi/2, \pi/2]$.

First we linearize the plant (43)-(44) at the origin and design a constant linear control law $u = Fx$ based on the linearizing model. It is easy to get the linearizing model,

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ \dfrac{g}{4l/3 - aml} & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ \dfrac{-\mu a}{4l/3 - aml} \end{bmatrix} v. \tag{46}$$

By placing the closed-loop eigenvalues at $\{-2, -2\}$, we have [22]

$$F = \begin{bmatrix} 0.1207 & 0.0227 \end{bmatrix}. \tag{47}$$

First, we use Theorem 3 and optimization problem (21) to estimate the permissible balancing range of x_1 through estimating the domain of attraction of the closed-loop system. To apply the optimization method introduced in Section 2.2, we set $\mathcal{X}_R = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$. Solving optimization problem (21), we obtain

$$\alpha_{\max} = 10.2026,$$

which corresponds to an angle much larger than $\pi/2$. By applying the above controller to the true plant (43)-(44), however, we find that when $x_1 > 42°$, $x_2 = 0$, the constant controller (47) fails to balance the pendulum. This is because of the very large modeling error between linear system (46) and original nonlinear system (43)-(44).

Now, we take the fuzzy-scheduling control design approach proposed in this paper. We approximate the system by the following two-rule fuzzy model

$$\begin{array}{ll} \textbf{Rule 1:} & \text{IF } x_1 \text{ is about } 0 \\ & \text{THEN } \dot{x} = A_1 x + B_1 v, \\ \textbf{Rule 2:} & \text{IF } x_1 \text{ is about } \pm\pi/2 \ (|x_1| < \pi/2) \\ & \text{THEN } \dot{x} = A_2 x + B_2 v, \end{array} \tag{48}$$

where

$$A_1 = \begin{bmatrix} 0 & 1 \\ \dfrac{g}{4l/3 - aml} & 0 \end{bmatrix}, \ B_1 = \begin{bmatrix} 0 \\ \dfrac{-\mu a}{4l/3 - aml} \end{bmatrix},$$

$$A_2 = \begin{bmatrix} 0 & 1 \\ \dfrac{g}{\pi(4l/3 - aml\beta^2)} & 0 \end{bmatrix}, \ B_2 = \begin{bmatrix} 0 \\ \dfrac{-\mu a\beta}{4l/3 - aml\beta^2} \end{bmatrix},$$

and $\beta = \cos(86°)$. Membership functions for Rule 1 and Rule 2 are set as follows

$$p_1 = \cos(x_1), \ p_2 = 1 - p_1,$$

which are shown in Figure 1.

The following fuzzy-scheduling control law is used to balance the plant

$$\begin{array}{ll} \textbf{Rule 1:} & \text{IF } x_1 \text{ is about } 0 \\ & \text{THEN } u = F_1 x, \\ \textbf{Rule 2:} & \text{IF } x_1 \text{ is about } \pm\pi/2 \ (|x_1| < \pi/2) \\ & \text{THEN } u = F_2 x. \end{array} \tag{49}$$

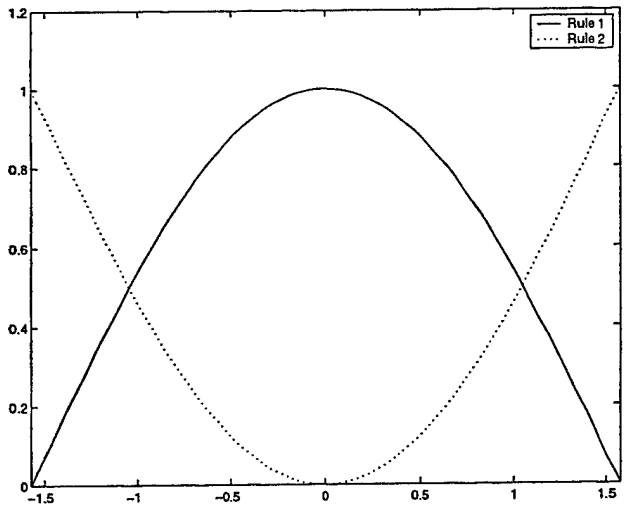


Figure 1: Membership functions of two-rule model.

By placing the eigenvalues of both $A_1 + B_1F_1$ and $A_2 + B_2F_2$ at $\{-2, -2\}$, we have

$$F_1 = [\ 0.1207 \quad 0.0227\ ],\ \ F_2 = [\ 0.8292 \quad 0.3820\ ]. \tag{50}$$

In the absence of saturation, this control law can balance the pendulum for initial conditions $x_1 \in [-88°, 88°]$ ($x_2 = 0$) as the control law proposed in [22].

Then, we apply our optimization method to estimate the balancing range. By solving optimization problem (42), we obtain

$$\alpha_{\max} = 2.2357.$$

which still corresponds to an angle larger than $\pi/2$. This implies that control law (50) should be able to balance the pendulum in the required range $[-\pi/2, \pi/2]$ even in the presence of saturation. However, simulation indicates that the fuzzy-scheduling control law (8) can balance the pendulum for initial conditions $x_1 \in [-84°, 84°]$ ($x_2 = 0$) in the presence of saturation. Figure 2 shows the response of the pendulum system using linear and fuzzy scheduling controls for initial conditions $x_1 = 20°, 40°, 45°, 84°$, and $x_2 = 0$. The solid curves indicate responses with the fuzzy scheduling controller. The dashed curves show those with the linear constant controller.

¿From the curves, we can also find that the fuzzy scheduling controller can lead to fast responses.

We can also use Theorem 7 and optimization problem (42) to design a controller such that the balancing range is as large as possible. Solving optimization problem (42), we obtain the following result

$$\begin{aligned} \alpha_{\max} &= 2.2361, \\ F_1 &= [\ 5.2016 \quad 2.4764\ ],\ F_2 = [\ 6.2736 \quad 2.9962\ ]. \end{aligned} \tag{51}$$

Note that $\alpha_{\max} = 2.2361$ corresponds to an angle of 128°. By applying the above controller to the true plant (43)-(44) with initial conditions $x_1 = 84°, 45°$ and $20°$ ($x_2 = 0$), the system responses are shown in Figure 3.

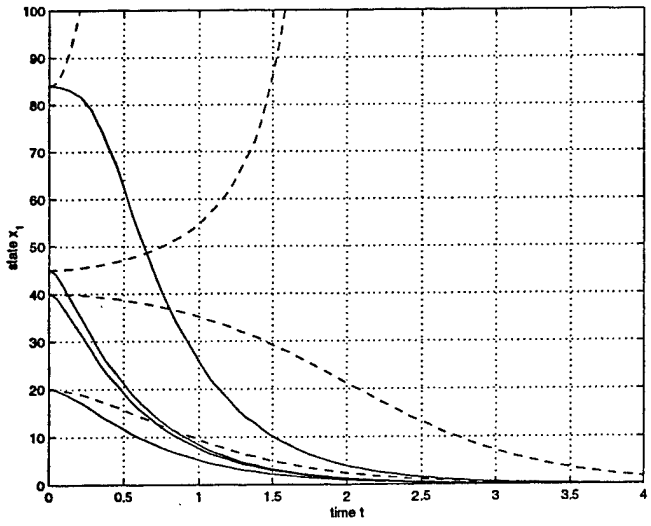


Figure 2: Angle response using linear and two-rule fuzzy control: dashed curves – linear control; solid curves – fuzzy control.

Because the modeling error is not considered, the stability region of the fuzzy system may not be that of the original nonlinear system. In what follows, we will analyze the modeling error between above fuzzy model and the actual pendulum model (43)-(44).

With the fuzzy rules shown in (48), the resulting fuzzy system is given by

$$\begin{aligned}\dot{x}_1 &= x_2,\\ \dot{x}_2 &= \left(\frac{g}{4l/3-aml}p_1+\frac{2g}{\pi(4l/3-aml\beta^2)}p_2\right)x_1+\left(\frac{-\mu a}{4l/3-aml}p_1+\frac{-\mu a\beta}{4l/3-aml\beta^2}p_2\right)v. \quad (52)\end{aligned}$$

The difference between (44) and (52) is

$$\Delta f = f_{x_1}x_1 + f_{x_2}x_2 + f_v v,$$

where

$$\begin{aligned}f_{x_1} &= \frac{g\sin(x_1)}{(4l/3-aml\cos^2(x_1))x_1}-\left(\frac{g}{4l/3-aml}p_1+\frac{2g}{\pi(4l/3-aml\beta^2)}p_2\right),\\ f_{x_2} &= \frac{-amlx_2\sin(2x_1)/2}{4l/3-aml\cos^2(x_1)},\\ f_v &= \frac{\mu a}{4l/3-aml}p_1+\frac{\mu a\beta}{4l/3-aml\beta^2}p_2-\frac{\mu a\cos(x_1)}{4l/3-aml\cos^2(x_1)}.\end{aligned}$$

Now we use the following uncertain fuzzy model to analyze the stability of the original system

$$\begin{aligned}&\textbf{Rule 1:} \text{ IF } x_1 \text{ is about } 0\\ &\qquad \text{THEN } \dot{x} = (A_1+\Delta A_1)x + (B_1+\Delta B_1)v,\\ &\textbf{Rule 2:} \text{ IF } x_1 \text{ is about } \pm\pi/2 \ (|x_1|<\pi/2)\\ &\qquad \text{THEN } \dot{x} = (A_2+\Delta A_2)x + (B_2+\Delta B_2)v.\end{aligned} \quad (53)$$

In the above uncertain fuzzy models, we assume that the uncertainties that describe the modeling errors are in the form of

$$\begin{aligned}\Delta A_1 &= M_1\Theta(t)N_{11},\ \Delta B_1 = M_1\Theta(t)N_{12},\\ \Delta A_2 &= M_2\Theta(t)N_{21},\ \Delta B_2 = M_2\Theta(t)N_{22},\end{aligned}$$

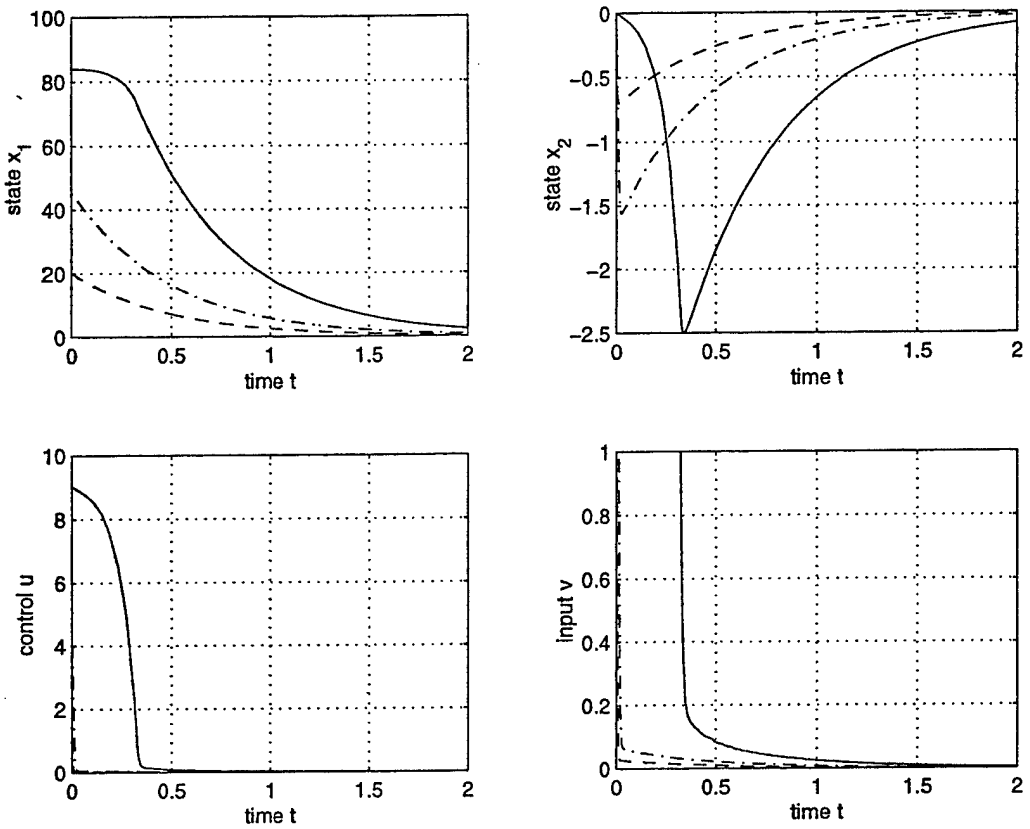


Figure 3: State and input with different initial conditions: solid curves - 84°; dotted-dashed curves – 45°; dashed curves – 20°.

where $\Theta(t) = \mathrm{diag}(\theta_1(t), \theta_2(t), \theta_3(t))$ with $||\theta_i(t)|| \leq 1$, and

$$M_1 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}^T, \; N_{11} = \begin{bmatrix} a_{11} & 0 \\ 0 & a_{12} \\ 0 & 0 \end{bmatrix}, \; N_{12} = \begin{bmatrix} 0 \\ 0 \\ b_1 \end{bmatrix},$$

$$M_2 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}^T, \; N_{21} = \begin{bmatrix} a_{21} & 0 \\ 0 & a_{22} \\ 0 & 0 \end{bmatrix}, \; N_{22} = \begin{bmatrix} 0 \\ 0 \\ b_2 \end{bmatrix}.$$

The parameters $a_{11}, a_{12}, a_{21}, a_{22}$, b_1 and b_2 are to be determined. Hence, the uncertain nonlinear fuzzy system is

$$\begin{aligned} \dot{x} &= [p_1(A_1 + \Delta A_1) + p_2(A_2 + \Delta A_2)]x \\ &\quad + [p_1(B_1 + \Delta B_1) + p_2(B_2 + \Delta B_2)]v \end{aligned} \tag{54}$$

i.e.,

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= \left(\frac{g}{4l/3 - aml} p_1 + \frac{2g}{\pi(4l/3 - aml\beta^2)} p_2\right) x_1 + \left(\frac{-\mu a}{4l/3 - aml} p_1 + \frac{-\mu a\beta}{4l/3 - aml\beta^2} p_2\right) v, \\ &\quad + (p_1 a_{11} + p_2 a_{21})\theta_1(t) x_1 + (p_1 a_{12} + p_2 a_{22})\theta_2(t) x_2 + (p_1 b_1 + p_2 b_2)\theta_3(t) v. \end{aligned} \tag{55}$$

If we set $a_{11} = a_{21} = a_1$, $a_{12} = a_{22} = a_2$ and $b_1 = b_2 = b$, then they can be chosen as

$$a_1 = \max_{x_1 \in [-\pi/2, \pi/2]} \|f_{x_1}\|, \quad a_2 = \max_{x_1 \in [-\pi/2, \pi/2], x_2} \|f_{x_2}\|, \quad b = \max_{x_1 \in [-\pi/2, \pi/2]} \|f_v\|.$$

If we constrain $\|x_2\| < 10$, then we gain

$$a_1 = 4.6757, \; a_2 = 0.813, \; b = 8.9697.$$

With the fixed controller (47), the optimization problem (31) has no solution. When the controller is not fixed, we obtain following result

$$\begin{aligned} \alpha_{\max} &= 1.2111 \; (= 69°), \\ F &= \begin{bmatrix} 0.8998 & 0.4021 \end{bmatrix}. \end{aligned}$$

Simulation results with this controller are shown in Figure 4. This simulation shows that, by optimizing the feedback gain as in Section 4, even a constant controller is able to balance the pendulum in range of $[-84°, 84°]$.

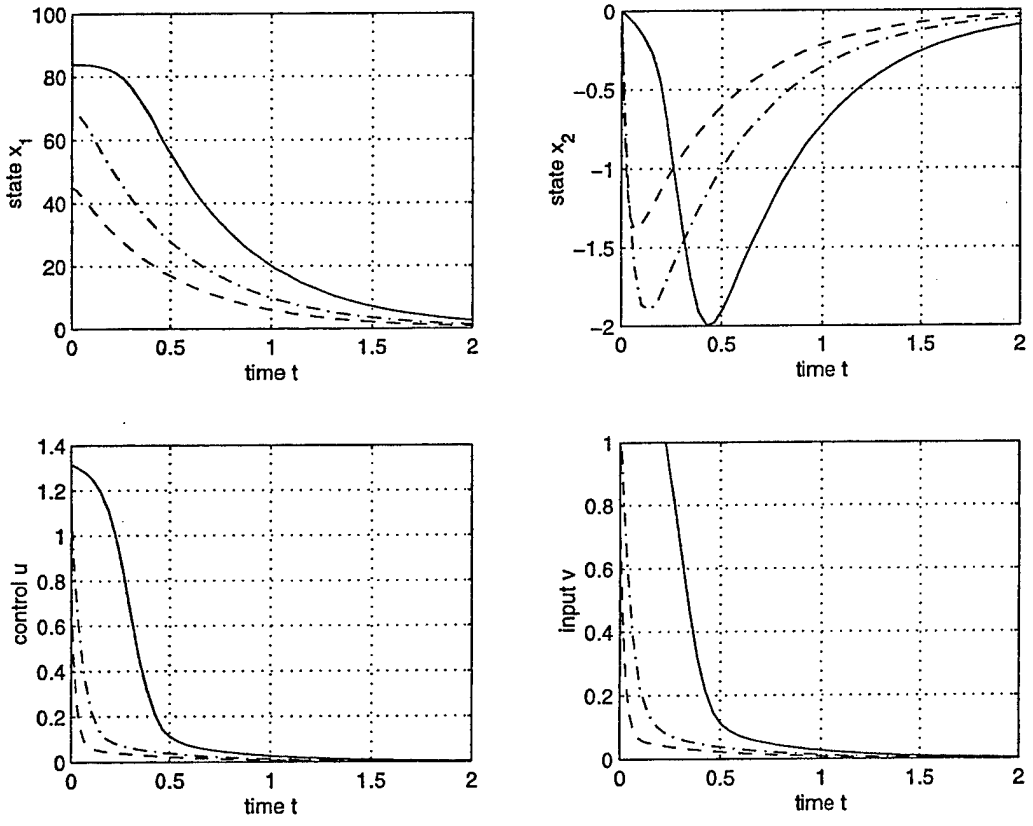


Figure 4: State and input responses with different initial conditions: solid curves - 84°; dotted-dashed curves – 69°; dashed curves – 45°.

6 Conclusions

In this paper, the TS fuzzy modeling approach was extended to represent nonlinear systems subject to actuator saturation. Based on this fuzzy representation, we developed stability analysis and design methods for nonlinear systems subject to actuator saturation. By their application to an inverted pendulum, our methods are shown to be very effective.

References

[1] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory.* SIAM, Philadelphia, 1994.

[2] S. G. Cao, N. W. Rees, and G. Feng. Analysis and design for a class of complex control systems part 2: Fuzzy controller design. *Automatica*, 33(6):1029–1039, 1997.

[3] Y.-Y. Cao and P. M. Frank. Robust H_∞ control for a class of discrete-time nonlinear systems via fuzzy switch. *IEEE Trans. Fuzzy Systems*, 8(4):406–415, 2000.

[4] Y.-Y. Cao, Z. Lin, and Y. Shamash. Set invariance analysis and gain-scheduling control for LPV systems subject to actuator saturation. *Systems Control Lett.*, page to appear, 2002.

[5] C. E. de Souza, M. Fu, and L. Xie. H_∞ analysis and synthesis of discrete-time systems with time-varying uncertainty. *IEEE Trans. Automat. Control*, 38:459–462, 1993.

[6] A. E. Gegov and P. M. Frank. Hierarchical fuzzy control of multivariable systems. *Fuzzy Sets And Systems*, 72(3):299–310, 1995.

[7] H. Hindi and S. Boyd. Analysis of linear systems with saturating using convex optimization. *Proc. 37th IEEE Conf. Decision Contr., Tampa, FL*, pages 903–908, 1998.

[8] T. Hu and Z. Lin. *Control Systems with Actuator Saturation: Analysis and Design.* Birkhäuser, Boston, 2001.

[9] T. Hu, Z. Lin, and B. M. Chen. An analysis and design method for linear systems subject to actuator saturation and disturbance. *Automatica*, 38(2):351–359, 2002.

[10] P. P. Khargonekar, I. R. Petersen, and K. Zhou. Robust stabilization of uncertain linear systems: quadratic stabilizability and H_∞ control. *IEEE Trans. Automat. Control*, 35(8):356–361, 1990.

[11] K. Kiriakidis. Fuzzy model-based control of complex plants. *IEEE Trans. Fuzzy Sys.*, 6(4):517–529, 1998.

[12] C. Lee. Fuzzy logic in control systems:fuzzy logic controller, *P*art 1, *P*art 2. *IEEE Trans. Sys., Man, Cybern.*, 20(2):404–435, 1990.

[13] F. H. F. Leung, H. K. Lam, and P. K. S. Tam. Design of fuzzy controllers for uncertain nonlinear systems using stability and robustness analyses. *Systems Control Lett.*, 35:237–243, 1998.

[14] Z. Lin. *Low Gain Feedback.* Springer, London, 1998.

[15] R. Palm, D. Driankov, and H. Hellendoorn. *Model based fuzzy control.* Springer-Verlag, Berlin, Germany, 1997.

[16] S. Sugeno and G.T. Kang. Structure identification of fuzzy model. *Fuzzy Sets And Systems*, 28(10):15–33, Oct. 1988.

[17] T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans. Sys., Man, Cybern.*, 15(1):116–132, 1985.

[18] T. Takagi and M. Sugeno. Stability analysis and design of fuzzy control systems. *Fuzzy Sets Syst.*, 45(2):135–156, 1992.

[19] K. Tanaka, T. Ikeda, and H. O. Wang. Robust stabilization of a class of uncertain nonlinear systems via fuzzy control: quadratic stabilizability, H_∞ control theory, and linear matrix inequlities. *IEEE Trans. Fuzzy Sys.*, 4(1):1–13, 1996.

[20] K. Tanaka, T. Ikeda, and H. O. Wang. Fuzzy regulators and fuzzy observers: relaxed stability conditions and LMI-based designs. *IEEE Trans. Fuzzy Sys.*, 6(2):250–265, 1998.

[21] M. C. M. Teixeira and S. H. Zak. Stabilizing controller design for uncertain nonlinear systems using fuzzy models. *IEEE Trans. Fuzzy Sys.*, 7(2):133–142, 1999.

[22] H. O. Wang, K. Tanaka, and M. F. Griffin. An approach to fuzzy control of nonlinear systems: stability and design issues. *IEEE Trans. Fuzzy Sys.*, 4(1):14–23, 1996.

[23] L. Y. Wang and W. Zhan. Robust disturbance attenuation with stability for linear systems with norm-bounded nonlinear uncertainties. *IEEE Trans. Automat. Control*, 41(6):886–888, 1996.

[24] L. Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Trans. Sys., Man, Cybern.*, 3(1):28–44, 1973.

[25] J. Zhao, V. Wertz, and R. Gorez. Linear TS models based robust stabilizing controller design. *Proc. 34th IEEE Conf. on Decision and Control*, New Orleans,USA:255–260, 1995.