AFRL-OSR-VA-TR-2013-0550

SEQUENTIAL PREDICTION FOR INFORMATION FUSION AND CONTROL

REBECCA WILLETT

DUKE UNIVERSITY

10/15/2013
Final Report

AIR FORCE RESEARCH LABORATORY
AF OFFICE OF SCIENTIFIC RESEARCH (AFOSR)/RSL
ARLINGTON, VIRGINIA 22203
AIR FORCE MATERIEL COMMAND

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
14-10-2013	Final	July 2010-July 2013

4. TITLE AND SUBTITLE
Sequential Prediction for Information Fusion and Control

5a. CONTRACT NUMBER
FA9550-10-1-0390

5b. GRANT NUMBER
P00002

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Willett, Rebecca, M
Raginsky, Maxim

5d. PROJECT NUMBER
F1ATA00138B001

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Duke University
2200 W. Main St. Suite 700
Durham, NC 27705-4677

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
USAF, AFRL
AF Office of Scientific Research
875 N. Randolph St. Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
ACRN, AA, AC

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

12. DISTRIBUTION / AVAILABILITY STATEMENT
unrestricted distribution

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The goal of this work was to develop sequential prediction methods for online information fusion and control, with methods designed to handle unknown, environmental dynamics, potentially stemming from an adversary who reacts to sensing actions, active sensing paradigms, and external feedback mechanisms. Online prediction and targeted collection of information is an emerging paradigm at the intersection of optimization, machine learning and control theory, which is concerned with real-time sequential planning of actions or decisions in the presence of model uncertainty, nonstationarity, and possibly adversarial disturbances. Several methods and underlying supporting theory which meet these objectives are described in detail in the following final report.

15. SUBJECT TERMS
Online learning, dynamical systems, active learning, stochastic control

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Rebecca Willett
U	U	U	UU	34	**19b. TELEPHONE NUMBER** *(include area code)* 608-316-4357

Sequential Prediction for Online Information Fusion and Control: Final Report for AFOSR grant FA9550-10-1-0390

PI Rebecca Willett and Co-PI Maxim Raginsky

October 14, 2013

Contents

1 Summary of program objectives and outcomes

The goal of this work was to develop sequential prediction methods for online information fusion and control, with methods designed to handle unknown, environmental dynamics, potentially stemming from an adversary who reacts to sensing actions, active sensing paradigms, and external feedback mechanisms. Online prediction and targeted collection of information is an emerging paradigm at the intersection of optimization, machine learning and control theory, which is concerned with real-time sequential planning of actions or decisions in the presence of model uncertainty, nonstationarity, and possibly adversarial disturbances.

Our team has developed several methods and underlying supporting theory which meet these objectives:

- Early in the course of the program, we developed methods for online learning using external feedback to efficiently guide active data collection and utilization of expensive expert system resources. In particular, we developed an online convex programming approach to sequential probability assignment of high-dimensional co-occurrence data [25]. Our approach consists of two main elements: (1) filtering, or assigning a belief or likelihood to each successive measurement based upon our ability to predict it from previous noisy observations, and (2) hedging, or flagging potential anomalies by comparing the current belief against a time-varying and data-adaptive threshold. The threshold is adjusted based on the available feedback from an end user. Our algorithms, which combine universal prediction with recent work on online convex programming, do not require computing posterior distributions given all current observations and involve simple primal-dual parameter updates. At the heart of our approach lie exponential-family models which can be used in a wide variety of contexts and applications, and which yield performance comparable to that of a batch algorithm which has access to all data at once rather than sequentially. methods that achieve sublinear per-round regret against both static and slowly varying product distributions with marginals drawn from the same exponential family. Moreover, the regret against static distributions coincides with the minimax value of the corresponding online strongly convex game. We also prove bounds on the number of mistakes made during the hedging step relative to the best offline choice of the threshold with access to all estimated beliefs and feedback signals. Furthermore, our approach is provably robust to unknown environmental dynamics and unmodeled statistical dependencies. As described in [19], our computationally efficient sequential anomaly detection using feedback can be used as an effective pre-processing step for large volumes of social network data associated with encrypted or other contextual information which can only be analyzed with resource-intensive expert systems.

- Online optimization methods are often designed to have a total accumulated loss comparable to that achievable by some comparator, such as a batch algorithm with access to all the data and infinite computational resources. In many settings, this comparator is allowed to vary with time, and the associated "tracking regret" bounds scale with the overall variation of the comparator sequence. However, in practical scenarios ranging from motion imagery to network analysis, the environment is nonstationary and comparator sequences with small variation are quite weak, resulting in large losses. Our work describes a "dynamic mirror descent" method which addresses this challenge, yielding low regrets bounds for comparators with small deviations from a given dynamical model. This approach is then used within a broader class of online learning methods to simultaneously track the best dynamical model and form predictions based on that model. This concept is demonstrated empirically in the context of sequential compressed sensing of a dynamic scene, solar flare detection from satellite data with missing elements, and tracking a dynamic social network [13, 14, 15].

- We also considered an online (real-time) control problem that involves an agent performing a discrete-time random walk over a finite state space. The agent's action at each time step is to specify the probability distribution for the next state given the current state. Following the set-up of Todorov, the state-action cost at each time step is a sum of a state cost and a control cost given by the Kullback-Leibler (KL) divergence between the agent's next-state distribution and that determined by some fixed passive dynamics. The online aspect of the problem is due to the fact that the state cost functions are generated by a dynamic environment, and the agent learns the current state cost only after selecting an action. An explicit construction of a computationally efficient strategy with small regret (i.e., expected difference between its actual total cost and the smallest cost attainable using noncausal knowledge of the state costs) under mild regularity conditions is presented, along with a demonstration of the performance of the proposed strategy on a simulated target tracking problem. A number of new results on Markov decision processes with KL control cost are also obtained [10, 11].

Most recently, we have built upon and to some extent unified the above results to develop a general procedure for developing low-regret algorithms for online Markov decision processes. This culminating work is described in detail in this report. Online learning algorithms can deal with nonstationary environments, but generally there is no notion of a dynamic *state* to model constraints on current and future actions as a function of past actions. State-based models are common in stochastic control settings, but commonly used frameworks such as Markov Decision Processes (MDPs) assume a known stationary environment. In recent years, there has been a growing interest in combining the above two frameworks and considering an MDP setting in which the cost function is allowed to change arbitrarily after each time step. However, most of the work in this area has been algorithmic: given a problem, one would develop an algorithm almost from scratch. Moreover, the presence of the state and the assumption of an arbitrarily varying environment complicate both the theoretical analysis and the development of computationally efficient methods. This report describes a broad extension of the ideas proposed by Rakhlin, Shamir, and Sridharan [27] to give a general framework for deriving algorithms in an MDP setting with arbitrarily changing costs. This framework leads to a unifying view of existing methods and provides a general procedure for constructing new ones. One such new method is presented and shown to have important advantages over a similar method developed outside the framework proposed in this report.

2 Relationship between program outcomes and previous state of the art

Markov decision processes, or MDPs for short [2, 17, 24], are a popular framework for sequential decision-making in a dynamic environment. In an MDP, we have states and actions. At each time step of the sequential decision-making process, the agent observes the current state and chooses an action, and the system transitions to the next state according to a fixed and known Markov law. The costs incurred by the agent depend both on his action and the current state. Traditional theory of MDPs deals with the case when both the transition law and the state-action cost function are known in advance. In this case, the problem of policy design reduces to dynamic programming. However, *a priori* known costs are typically unavailable in practical settings. In this report, instead of considering a fixed cost function, we study Markov decision processes with finite state and action spaces, where the cost functions are chosen arbitrarily and allowed to change with time. More specifically, we are interested in the *online MDP* problem: just as in the usual online leaning framework [7, 16, 28], the one-step cost functions form an arbitrarily varying sequence, and the cost function corresponding to each time step is revealed to the agent after an action has been taken. The objective of the agent is to minimize regret relative to the best stationary Markov policy that could have been selected with full knowledge of the cost function sequence over the horizon of interest. The assumption of arbitrary time-varying cost functions makes sense in highly un-

certain and complex environments whose temporal evolution may be difficult or costly to model, and it also accounts for collective (and possibly irrational) behavior of any other agents that may be present. The regret minimization viewpoint then ensures that the agent's *online* policy is robust against these effects.

Online MDP problems can be viewed as *online control problems*. The online aspect is due to the fact that the cost functions are generated by a dynamic environment under no distributional assumptions, and the agent learns the current state-action cost only after selecting an action. The control aspect comes from the fact that the choice of an action at each time step influences future states and costs. Taking into account the effect of past actions on future costs in a dynamic distribution-free setting makes online MDPs hard to solve. To the best of our knowledge, only a few methods have been developed in this area over the past decade [1, 3, 9, 12, 21, 23, 33]. Most research in this area has been algorithmic: given a problem, one would present a method and prove a guarantee (i.e., a regret bound) on its performance. Thus, it is desirable to provide a unifying view of existing methods and a general procedure for constructing new ones. In this report, we present such a general framework for online MDP problems, bringing two well-known existing methods under a single theoretical interpretation. This general framework not only enables us to recover known algorithms, but it also gives us a generic toolbox for deriving new algorithms from a more principled perspective rather than from scratch.

The online MDP setting we are considering was first defined and studied in the work of [9] and [33], which deals with MDPs with arbitrarily varying rewards. Like these authors, we assume a full information feedback model and known stochastic state transition dynamics. (However, it should be pointed out that these assumptions have been relaxed in some recent works — for example, [23] and [3] assume only bandit-type feedback, while [1] prove regret bounds for MDPs with arbitrarily varying transition models and cost functions. An extension of our framework to these settings is an interesting avenue for future research.)

Our general approach is motivated by recent work of [27], which gives a principled way of deriving online learning algorithms (and bounding their regret) from a minimax analysis. Of course, many online learning algorithms have been developed in various settings over the past few decades, but a comprehensive and systematic treatment was still lacking. Starting from a general formulation of online learning as a repeated game between a learner and an adversary, [27] analyze the minimax value of this online learning game. It was known before [26] that one could derive sublinear non-constructive upper bounds on the minimax value. However, algorithm design was done on a case-by-case basis, and a separate analysis was needed in each case to derive performance guarantees matching these upper bounds. The work of [27] bridges this gap between minimax value analysis and algorithm design: They have shown that, by choosing appropriate relaxations of a certain recursive decomposition of the minimax value, one can recover many known online learning algorithms and give a general recipe for developing new ones. In short, the framework proposed by [27] can be used to convert an upper bound on the value of the game into an algorithm.

Our main contribution is an extension of the framework of [27] to online MDPs. Since online learning problems are studied in a state-free setting, it is not straightforward to generalize the ideas of [27] to the case when the system has a state, and the technical nature of the arguments involved in online MDPs is significantly heavier than their state-free counterpart. We formulate the online MDP problem as a two-player repeated game and study its minimax value in the presence of state variables. We introduce the notion of an online MDP *relaxation* and show how it can be used to recover existing methods and to construct new algorithms. More specifically, we use Poisson inequalities for MDPs [22] to move from a state-dependent setting to a state-free one. As a consequence, each possible state value is associated with an individual online learning algorithm. We show that the algorithm proposed by [9] arises from a particular relaxation, and we also derive a new algorithm in the spirit of [33] which exhibits improved regret bounds.

The remainder of the report is organized as follows. We close this section with a brief summary of our results and frequently used notation. Section 3 contains precise formulation of the online MDP problem and points out the general idea and major challenges. Section 4 describes our proposed framework and contains the main result. Section 5 shows the power of our framework by recovering an existing method proposed in [9] and further derives a new algorithm using the framework. Section 6 contains discussion about future research. Proofs of all intermediate results are relegated to the Appendix.

2.1 A summary of results

We start by recasting an MDP with arbitrary costs as a one-sided *stochastic game*, where an agent who wishes to minimize his long-term average cost is facing a Markovian environment, which is also affected by arbitrary actions of an opponent. A stochastic game [31] is a repeated two-player game, where the state changes at every time step according to a transition law depending on the current state and the moves of both players. Here we are considering a special type of a stochastic game, where the agent controls the state transition alone and the opponent chooses the cost functions. By "one-sided", we mean that the utility of the opponent is left unspecified. In other words, we do not need to study the strategy and objectives of the opponent, and only assume that the changes in the environment in response to the opponent's moves occur arbitrarily. As a result, we simply model the opponent as the environment.

A popular and common objective in such settings is regret minimization. Regret is defined as the difference between the cost the agent actually incurred, and what could have been incurred if the agent knew the observed sequence of cost functions in advance. We will give the precise definition of this regret notion in Section 3. We start by studying the minimax regret, i.e., the regret the agent will suffer when both the agent and the environment play optimally. By applying the theory of dynamic programming for stochastic games [31], we can give the minimax strategy for the agent that achieves minimax regret. It can be interpreted as choosing the best action that takes into account the current cost and the worst case future. Unfortunately, this minimax strategy in general is not computationally feasible due to the fact that the number of possible futures grows exponential with time. The idea is to find a way to approximate the term that represents the "future" and derive near-optimal strategy that is easy to compute using the approximation.

Our main contribution is a construction of a general procedure for deriving algorithms in the online MDP setting. More specifically:

1. Just as in the state-free setting considered by [27], we argue that algorithms can be constructed systematically by first deriving a sequence of upper bounds (*relaxations*) on a quantity called *sequential Rademacher complexity*, and then plugging these upper bounds into a recursively defined system of inequalities (called the *admissibility conditions)*.

2. Once a relaxation and an algorithm are derived in this way, we give a general regret bound of that algorithm as follows:

$$\text{Expected regret} \leq \text{Relaxation} + \text{Stationarization error}.$$

   The first term on the right-hand side of the above inequality is related to the derived relaxation, while the second term is an approximation error that results from approximating the Markovian evolution of the underlying process by a simpler steady-state using a procedure we refer to as *stationarization*. The first term can be analyzed using essentially the same techniques as the ones employed by [27], with some modifications; by contrast, the second term can be handled using only Markov chain methods. This approach significantly alleviates the technical burden of proving a regret bound as in the literature before our work.

3. Using the above procedure, we recover an existing method proposed in [9], which achieves $O(\sqrt{T})$ expected regret against the best stationary policy. We show that our derived relaxation gives us the same exponentially weighted average forecaster as in [9] and leads to the same regret bound.

4. We also derive a new algorithm using our proposed framework and argue that, while this new algorithm is similar in nature to the work of [33], it has several advantages — in particular, better scaling of the regret with the horizon T.

2.2 Notation

We will denote the underlying finite state space and action space by X and U, respectively. The set of all probability distributions on X will be denoted by $\mathscr{P}(\mathsf{X})$, and the same goes for U and $\mathscr{P}(\mathsf{U})$. A matrix $P = [P(u|x)]_{x\in\mathsf{X},u\in\mathsf{U}}$ with nonnegative entries, and with the rows and the columns indexed by the elements of U and X respectively, is called *Markov* (or *stochastic*) if its rows sum to one: $\sum_{u\in\mathsf{U}} P(u|x) = 1, \forall x \in \mathsf{X}$. We will denote the set of all such Markov matrices (or randomized state feedback laws) by $\mathscr{M}(\mathsf{U}|\mathsf{X})$. Markov matrices in $\mathscr{M}(\mathsf{U}|\mathsf{X})$ transform probability distributions on X into probability distributions on U: for any $\mu \in \mathscr{P}(\mathsf{X})$ and any $P \in \mathscr{M}(\mathsf{U}|\mathsf{X})$, we have

$$\mu P(u) \triangleq \sum_{x\in\mathsf{X}} \mu(x)P(u|x), \qquad \forall u \in \mathsf{U}.$$

The same applies to Markov matrices on X and to their action on the elements of $\mathscr{P}(\mathsf{X})$.

The fixed and known stochastic transition kernel of the MDP will be denoted throughout by K – that is, $K(y|x,u)$ is the probability that the next state is y if the current state is x and the action u is taken. For any Markov matrix (randomized state feedback law) $P \in \mathscr{M}(\mathsf{U}|\mathsf{X})$, we will denote by $K(y|x,P)$ the Markov kernel

$$K(y|x,P) \triangleq \sum_{u\in\mathsf{U}} K(y|x,u)P(u|x).$$

Similarly, for any $\nu \in \mathscr{P}(\mathsf{U})$,

$$K(y|x,\nu) \triangleq \sum_{u\in\mathsf{U}} K(y|x,u)\nu(u)$$

(this can be viewed as a special case of the previous definition if we interpret ν as a state feedback law that ignores the state and draws a random action according to ν). For any $\mu \in \mathscr{P}(\mathsf{X})$ and $P \in \mathscr{M}(\mathsf{U}|\mathsf{X})$, $\mu \otimes P$ denotes the induced joint state-action distribution on $\mathsf{X} \times \mathsf{U}$:

$$\mu \otimes P(x,u) = \mu(x)P(u|x), \qquad \forall (x,u) \in \mathsf{X} \times \mathsf{U}.$$

We say that P is *unichain* [18] if the corresponding Markov chain with transition kernel $K(\cdot|\cdot,P)$ has a single recurrent class of states (plus a possibly empty transient class). This is equivalent to the induced kernel $K(\cdot|P)$ having a unique invariant distribution π_P [29].

The total variation (or L_1) distance between $\nu_1, \nu_2 \in \mathscr{P}(\mathsf{U})$ is

$$\|\nu_1 - \nu_2\|_1 \triangleq \sum_{u\in\mathsf{U}} |\nu_1(u) - \nu_2(u)|.$$

It admits the following variational representation:

$$\|\nu_1 - \nu_2\|_1 = \sup_{f:\|f\|_\infty \le 1} \big|\langle \nu_1, f\rangle - \langle \nu_2, f\rangle\big|, \tag{1}$$

where the supremum is over all functions $f : \mathsf{U} \to \mathbb{R}$ with absolute value bounded by 1, and we are using the linear functional notation for expectations:

$$\langle \nu, f \rangle = \mathbb{E}_\nu[f] = \sum_{u \in \mathsf{U}} \nu(u) f(u).$$

The *Kullback–Leibler divergence* (or *relative entropy*) between ν_1 and ν_2 [8] is

$$D(\nu_1 \| \nu_2) \triangleq \begin{cases} \sum_{u \in \mathsf{U}} \nu_1(u) \log \frac{\nu_1(u)}{\nu_2(u)} & \text{if } \mathrm{supp}(\nu_1) \subseteq \mathrm{supp}(\nu_2) \\ +\infty & \text{otherwise} \end{cases}$$

where $\mathrm{supp}(\nu) \triangleq \{u \in \mathsf{U} : \nu(u) > 0\}$ is the *support* of ν. Here and in the sequel, we work with natural logarithms. The same applies, *mutatis mutandis*, to probability distributions on X.

We will also be dealing with binary trees that arise in symmetrization arguments, as in [27]: Let $\mathscr{H}$ be an arbitrary set. An $\mathscr{H}$-valued tree $\mathbf{h}$ of depth d is defined as a sequence $(\mathbf{h}_1, \ldots, \mathbf{h}_d)$ of mappings $\mathbf{h}_t : \{\pm 1\}^{t-1} \to \mathscr{H}$ for $t = 1, 2, \ldots, d$. Given a tuple $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d) \in \{\pm 1\}^d$, we will often write $\mathbf{h}_t(\varepsilon)$ instead of $\mathbf{h}_t(\varepsilon_{1:t-1})$.

3 Problem formulation

We consider an online MDP with finite state and action spaces X and U and transition kernel $K(y|x,u)$. Let $\mathscr{F}$ be a fixed class of functions $f : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$, and let $x \in \mathsf{X}$ be a fixed initial state. Consider an agent performing a controlled random walk on X in response to signals coming from the environment. The agent is using mixed strategies to choose actions, where a mixed strategy is a probability distribution over the action space. The interaction between the agent and the environment proceeds as follows:

$X_1 = x$
for $t = 1, 2, \ldots, T$
  The agent observes the state X_t and selects a mixed strategy $P_t \in \mathscr{P}(\mathsf{U})$
  The environment selects $f_t \in \mathscr{F}$ and announces it to the agent
  The agent draws an action U_t from P_t and incurs one-step cost $f_t(X_t, U_t)$.
  The system transitions to the next state $X_{t+1} \sim K(\cdot|X_t, U_t)$
end for

Here, T is a fixed finite horizon. We assume throughout that the environment is *oblivious* (or *open-loop*), in the sense that the evolution of the sequence $\{f_t\}$ is not affected by the state and action sequences $\{X_t\}$ and $\{U_t\}$. We view the above process as a two-player repeated game between the agent and the environment. At each $t \geq 1$, the process is at state $X_t = x_t$. The agent observes the current state x_t and selects the mixed strategy P_t, where $P_t(u|x_t) = \Pr\{U_t = u | X_t = x_t\}$, based on his knowledge of all the previous states and current state x^t and the previous moves of the environment $f^{t-1} = (f_1, \ldots, f_{t-1})$. After drawing the action U_t from P_t, the agent incurs the one-step cost $f_t(X_t, U_t)$. Adopting game-theoretic terminology [4], we define the agent's closed-loop *behavioral strategy* as a tuple $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)$, where $\gamma_t : \mathsf{X}^t \times \mathscr{F}^{t-1} \to \mathscr{P}(\mathsf{U})$. Similarly, the environment's open-loop behavioral strategy is a tuple $\boldsymbol{f} = (f_1, \ldots, f_t)$. Once the initial state $X_1 = x$ and the strategy pair $(\boldsymbol{\gamma}, \boldsymbol{f})$ are specified, the joint distribution of the state-action process (X^T, U^T) is well-defined.

Let $\mathscr{M}_0 = \mathscr{M}_0(\mathsf{U}|\mathsf{X}) \subseteq \mathscr{M}(\mathsf{U}|\mathsf{X})$ denote the subset of all Markov policies P, for which the induced state transition kernel $K(\cdot|\cdot, P)$ has a unique invariant distribution $\pi_P \in \mathscr{P}(\mathsf{X})$. The goal of the agent is to minimize the expected *steady-state regret*

$$R_x^{\boldsymbol{\gamma}, \boldsymbol{f}} \triangleq \mathbb{E}_x^{\boldsymbol{\gamma}, \boldsymbol{f}} \left\{ \sum_{t=1}^T f_t(X_t, U_t) - \inf_{P \in \mathscr{M}_0} \mathbb{E} \left[\sum_{t=1}^T f_t(X, U) \right] \right\}, \tag{2}$$

where the outer expectation $\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{f}}$ is taken w.r.t. both the Markov chain induced by the agent's behavioral strategy $\boldsymbol{\gamma}$ (including randomization of the agent's actions), the environment's behavior strategy $\boldsymbol{f}$, and the initial state $X_1 = x$. The inner expectation (after the infimum) is w.r.t. the state-action distribution $\pi_P \otimes P(x,u) = \pi_P(x)P(u|x)$, where π_P denotes the unique invariant distribution of $K(\cdot|\cdot,P)$. The regret $R_x^{\boldsymbol{\gamma},\boldsymbol{f}}$ can be interpreted as the gap between the expected cumulative cost of the agent using strategy $\boldsymbol{\gamma}$ and the best steady-state cost the agent could have achieved in hindsight by using the best stationary policy $P \in \mathscr{M}_0$ (with full knowledge of $\boldsymbol{f} = f^T$). This gap arises through the agent's lack of prior knowledge on the sequence of cost functions.

Here we consider the steady-state regret, so that the expectation w.r.t. the state evolution in the comparator term $\mathbb{E}\left[\sum_{t=1}^T f_t(X,U)\right]$ is taken over the invariant distribution π_P instead of the Markov transition law $K(\cdot|\cdot,P)$ induced by P. Under the additional assumptions that the cost functions f_t are uniformly bounded and the induced Markov chains $K(\cdot|\cdot,P)$ are uniformly exponentially mixing for all $P \in \mathscr{M}(\mathsf{U}|\mathsf{X})$, the difference we introduce here by considering the steady state is bounded by a constant independent of T [9, 33], and so is negligible in the long run. In our main results, we only consider baseline policies in $\mathscr{M}_0$ that are uniformly exponentially mixing, so we restrict our attention to the steady-state regret without any loss of generality.

3.1 Minimax value

We start our analysis by studying the value of the game (the minimax regret), which we first write down in *strategic form* as

$$V(x) \triangleq \inf_{\boldsymbol{\gamma}} \sup_{\boldsymbol{f}} R_x^{\boldsymbol{\gamma},\boldsymbol{f}} = \inf_{\boldsymbol{\gamma}} \sup_{\boldsymbol{f}} \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{f}} \left[\sum_{t=1}^T f_t(X_t,U_t) - \Psi(\boldsymbol{f}) \right], \tag{3}$$

where we have introduced the shorthand Ψ for the comparator term:

$$\Psi(\boldsymbol{f}) \triangleq \inf_{P \in \mathscr{M}_0} \mathbb{E}\left[\sum_{t=1}^T f_t(X,U) \right].$$

In operational terms, $V(x)$ gives the best value of the regret the agent can secure by any closed-loop behavioral strategy against the worst-case choice of an open-loop behavioral strategy of the environment. However, the strategic form of the value hides the *timing protocol* of the game, which encodes the information available to the agent at each time step. To that end, we give the following equivalent expression of $V(x)$ in *extensive form*:

Proposition 1 *The minimax value* (3) *is given by*

$$V(x) = \inf_{P_1} \sup_{f_1} \ldots \inf_{P_T} \sup_{f_T} \mathbb{E}\left[\sum_{t=1}^T f_t(X_t,U_t) - \Psi(\boldsymbol{f}) \right]. \tag{4}$$

Proof: See Appendix A. ∎

From this minimax formulation, we can immediately get an optimal algorithm that attains the minimax value. To see this, we give an equivalent recursive form for the value of the game. For any $t \in \{0,1,\ldots,T-1\}$, any given prefix $f^t = (f_1,\ldots,f_t)$ (where we let f^0 be the empty tuple e), and any state $X_{t+1} = x$, define the conditional value

$$V_t(x,f^t) \triangleq \inf_{\nu \in \mathscr{P}(\mathsf{U})} \sup_{f} \left\{ \sum_{u \in \mathsf{U}} f(x,u)\nu(u) + \mathbb{E}\Big[V_{t+1}(Y,f_1,\ldots,f_t,f) \Big| x,\nu \Big] \right\}, \qquad t = T-1,\ldots,0 \tag{5a}$$

$$V_T(x,f^T) \triangleq -\Psi(\boldsymbol{f}). \tag{5b}$$

Remark 1 Recursive decompositions of this sort arise frequently in problems involving decision-making in the presence of uncertainty. For instance, we may view (5) as a dynamic program for a finite-horizon minimax control problem [6]. Alternatively, we can think of (5) as applying the *Shapley operator* [31] to the conditional value in a two-player stochastic game, where one player controls only the state transitions, while the other player specifies the cost function. A promising direction for future work is to derive some characteristics of the conditional value from analytical properties of the Shapley operator.

From Proposition 1, we see that $V(x) = V_0(x, \mathsf{e})$. Moreover, we can immediately write down the minimax-optimal behavioral strategy for the agent:

$$\gamma_{t+1}(x, f^t) = \operatorname*{argmin}_{\nu \in \mathscr{P}(\mathsf{U})} \sup_{f \in \mathscr{F}} \left\{ \sum_{u \in \mathsf{U}} f(x,u)\nu(u) + \mathbb{E}\Big[V_{t+1}(Y, f_1, \ldots, f_t, f) \Big| x, \nu \Big] \right\}, \qquad t = 0, \ldots, T-1.$$

Note that the expression being minimized is a supremum of affine functions of ν, so it is a lower-semicontinuous function of ν. Any lower-semicontinuous function achieves its infimum on a compact set. Since the probability simplex $\mathscr{P}(\mathsf{U})$ is compact, we are assured that a minimizing ν always exists. Using the above strategy at each time step, we can secure the minimax value in the worst-case scenario. Note also that this strategy is very intuitive: it balances the tendency to minimize the present cost against the risk of incurring high future costs. However, with all the future infimum and supremum pairs involved, computing this conditional value is intractable. As a result, the minimax optimal strategy is not computationally feasible. The idea is to give tight bounds of the conditional value, which can be minimized to form a near-optimal strategy. We address this challenge by developing computable bounds for the conditional value functions, choosing a strategy based on these bounds. In general, tighter bounds yield lower regret and looser bounds are easier to compute, and various online MDP methods occupy different points in this domain.

In the spirit of [27], we come up with approximations of the conditional value $V_t(x, f^t)$ in (5). We say that a sequence of functions $\widehat{V}_t : \mathsf{X} \times \mathscr{F}^t \to \mathbb{R}$ is an *admissible relaxation* if

$$\widehat{V}_t(x, f^t) \geq \inf_{\nu \in \mathscr{P}(\mathsf{U})} \sup_{f} \left\{ \sum_{u \in \mathsf{U}} f(x,u)\nu(u) + \mathbb{E}[\widehat{V}_{t+1}(Y, f_1, \ldots, f_t, f) | x, \nu] \right\}, \qquad t = T-1, \ldots, 0 \tag{6a}$$

$$\widehat{V}_T(x, f^T) \geq -\Psi(\boldsymbol{f}). \tag{6b}$$

We can associate a behavior strategy $\widehat{\boldsymbol{\gamma}}$ to any admissible relaxation as follows:

$$\widehat{\gamma}_t(x, f^{t-1}) = \operatorname*{argmin}_{\nu \in \mathscr{P}(\mathsf{U})} \sup_{f \in \mathscr{F}} \left\{ \sum_{u \in \mathsf{U}} f(x,u)\nu(u) + \mathbb{E}\Big[\widehat{V}_t(Y, f_1, \ldots, f_{t-1}, f) \Big| x, \nu \Big] \right\}.$$

Proposition 2 *Given an admissible relaxation* $\{\widehat{V}_t\}_{t=0}^T$ *and the associated behavioral strategy* $\widehat{\boldsymbol{\gamma}}$*, for any open-loop strategy of the environment we have the regret bound*

$$R_x^{\widehat{\boldsymbol{\gamma}}, \boldsymbol{f}} = \mathbb{E}_x^{\widehat{\boldsymbol{\gamma}}, \boldsymbol{f}} \left[\sum_{t=1}^{T} f_t(X_t, U_t) - \Psi(\boldsymbol{f}) \right] \leq \widehat{V}_0(x).$$

Proof: See Appendix B. ∎

Based on the above sequential decompositions, it suffices to restrict attention only to Markov strategies for the agent, i.e., sequences of mappings $\gamma_t : \mathsf{X} \times \mathscr{F}^{t-1} \to \mathscr{P}(\mathsf{U})$ for all t, so that U_t is conditionally independent of X^{t-1}, U^{t-1} given X_t, f^{t-1}. From now on, we will just say "behavioral strategy" and really mean "Markov behavioral strategy." In other words, given X_t, f^{t-1}, the history of past states and actions is *irrelevant*, as far as the value of the game is concerned.

Remark 2 What happens if the environment is nonoblivious? [33] gave a simple counterexample of an aperiodic and recurrent MDP to show that the regret is linear in T regardless of the agent's policy when the opponent can adapt to the agent's state trajectory. We can gain additional insight into the challenges associated with an adaptive environment from the perspective of the minimax value. In particular, an adaptive environment's *closed-loop* behavioral strategy is $\boldsymbol{\delta} = (\delta_1, \dots, \delta_T)$ with $\delta_t : \mathsf{X}^t \times \mathsf{U}^{t-1} \to \mathscr{P}(\mathscr{F})$, and the corresponding regret will be given by

$$\begin{aligned}\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^{T} f_t(X_t, U_t) - \Psi(\boldsymbol{f})\right] &\leq \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^{T} f_t(X_t, U_t) + \widehat{V}_T(X_{T+1}, f^T)\right] \\ &= \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[\sum_{t=1}^{T-1} f_t(X_t, U_t)\right] + \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[f_T(X_T, U_T) + \widehat{V}_T(X_{T+1}, f^T)\right].\end{aligned}$$

Let's analyze the last two terms:

$$\begin{aligned}&\mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{\delta}}\left[f_T(X_T, U_T) + V_T(X_{T+1}, f^T)\right] \\ &= \int_{\mathsf{X}^T, \mathscr{F}^T} \mathbb{P}(\mathrm{d}x^T, \mathrm{d}\boldsymbol{f}) \int_{\mathsf{U}} P(\mathrm{d}u_T | x_T, f^{T-1}) \left\{f_T(x_T, u_T) + \mathbb{E}\left[\widehat{V}_T\big(X_{T+1}, f^T\big)\middle| x^T, f^T\right]\right\}.\end{aligned}$$

In the above conditional expectation, $\boldsymbol{f}$ may depend on the entire x^T, so we cannot replace this conditional expectation by $\mathbb{E}[\cdot | x_T, \gamma_T(x_T)]$. This implies we cannot get similar results as in Proposition 2 in a fully adaptive environment.

3.2 Major challenges

From Proposition 2, we can see that we can bound the expected steady-state regret in terms of the chosen relaxation. Ideally, if we construct an admissible relaxation by deriving certain upper bounds of the conditional value and implement the associated behavioral strategy, we obtain an algorithm that achieves the regret bound corresponding to the relaxation. In principle, this gives us a general framework to develop low-regret algorithms for online MDPs. However, with an additional state variable involved, it is difficult to derive admissible relaxations $\widehat{V}_t(x, f^t)$ to bound the conditional value. The difficulty stems from the fact that now the current cost depends not only on the current action, but also on past actions. Our plan is to first find a way to reduce this setting to a simpler setting where there is no Markov dynamics involved, and the agent only has to choose actions. Then we will be able to incorporate the ideas of [26, 27] into our simpler setting. More specifically, using Rademacher complexity tools introduced by [26, 27], we can derive algorithms in the simpler static setting and then transfer them to the original problem. In the same vein, we will also prove a general regret bound for the derived algorithms. Thus we will have a general recipe for developing algorithms and showing performance guarantees for online MDPs.

4 The general framework for constructing algorithms in online MDPs

As mentioned in the above section, the main challenge to overcome is the dependence of the conditional value in (6) on the state variable. Our plan is to reduce the original online MDP problem to a simpler one, where there is no Markov dynamics, and the agent only has to choose actions.

We proceed with our plan in several steps. First, we introduce a stationarization technique that will allow us to reduce the online MDP setting to a simpler setting without Markov dynamics. This effectively decouples current costs from past actions. Note that this reduction is fundamentally different from just naively applying stateless online learning methods in an online MDP setting, which would amount to a

very poor stationarization strategy with larger errors and consequently large regret bounds. In contrast, our proposed stationarization performs the decoupling with minimal loss in accuracy by exploiting the transition kernel, yielding lower regret bounds. We then state a new admissibility condition for relaxations that differs from (6) in that there is no conditioning on the state variable. The advantage of working with this new type of relaxation is that the corresponding admissibility conditions are much easier to verify. The main result of this section is that we can apply any algorithm derived in the simpler static setting to the original dynamic setting and automatically bound its regret.

4.1 Stationarization

Our stationarization technique makes use of Poisson inequalities for MDPs [22] to bound the regret defined in (2) in terms of a different function as opposed to the one-step cost function f.

As before, we let K denote the fixed and known transition law of the MDP. Following [9] and [33], we assume that K is a *unichain model*, i.e., $K(\cdot|\cdot,P)$ is unichain for any choice of $P \in \mathscr{M}(\mathsf{U}|\mathsf{X})$ — see Section 2.2 for definitions. Thus, every state feedback law $P \in \mathscr{M}(\mathsf{U}|\mathsf{X})$ belongs to $\mathscr{M}_0$. For future reference, we record the following crucial consequence of the unichain assumption: There exists a finite constant $\tau > 0$ such that for all Markov policies $P \in \mathscr{M}(\mathsf{U}|\mathsf{X})$ and all distributions $\mu_1, \mu_2 \in \mathscr{P}(\mathsf{X})$,

$$\|\mu_1 K(\cdot|P) - \mu_2 K(\cdot|P)\|_1 \leq e^{-1/\tau} \|\mu_1 - \mu_2\|_1, \tag{7}$$

where $K(\cdot|P) \in \mathscr{M}(\mathsf{X}|\mathsf{X})$ is the Markov matrix on the state space induced by P. In other words, the collection of all state transition laws induced by all Markov policies P is *uniformly mixing*. Here we assume that $\tau \geq 1$.

Remark 3 The unichain assumption is rather strong, since it places significant simultaneous restrictions on an exponentially large family of Markov chains on the state space (each chain corresponds to a particular choice of a deterministic state feedback law, and there are $|\mathsf{U}|^{|\mathsf{X}|}$ such laws). It is also difficult to verify, since the problem of determining whether an MDP is unichain is NP-hard [32]. [3] relax the unichain assumption in a different way by considering deterministic state transition dynamics and weakly communicating structure, under which it is possible to move from any state to any other state under *some* policy. Although it is not clear yet if we can derive positive results with stochastic state transition dynamics and weakly communicating structure, putting weaker assumption on state connectivity is our goal in the future.

Consider now a behavioral strategy $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)$ for the agent. For a given choice $\boldsymbol{f} = (f_1, \ldots, f_T)$ of costs, the following objects are well-defined:

- $P_t^{\boldsymbol{\gamma},\boldsymbol{f}} \in \mathscr{M}(\mathsf{U}|\mathsf{X})$ — the Markov matrix that governs the conditional distribution of U_t given X_t, i.e.,

$$P_t^{\boldsymbol{\gamma},\boldsymbol{f}}(u|x) = \left[\gamma_t(x, f^{t-1})\right](u);$$

- $\mu_t^{\boldsymbol{\gamma},\boldsymbol{f}} \in \mathscr{P}(\mathsf{X})$ — the distribution of X_t;

- $K_t^{\boldsymbol{\gamma},\boldsymbol{f}} \in \mathscr{M}(\mathsf{X}|\mathsf{X})$ — the Markov matrix that describes the state transition from X_t to X_{t+1}, i.e.,

$$K_t^{\boldsymbol{\gamma},\boldsymbol{f}}(y|x) = K(y|x, P_t^{\boldsymbol{\gamma},\boldsymbol{f}}) \equiv \sum_u K(y|x,u) P_t^{\boldsymbol{\gamma},\boldsymbol{f}}(u|x);$$

- $\pi_t^{\boldsymbol{\gamma},\boldsymbol{f}} \in \mathscr{P}(\mathsf{X})$ — the unique stationary distribution of $K_t^{\boldsymbol{\gamma},\boldsymbol{f}}$, satisfying $\pi_t^{\boldsymbol{\gamma},\boldsymbol{f}} = \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}} K_t^{\boldsymbol{\gamma},\boldsymbol{f}}$, where existence and uniqueness are guaranteed by virtue of the unichain assumption;

- $\eta_t^{\boldsymbol{\gamma},\boldsymbol{f}} = \langle \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}} \otimes P_t^{\boldsymbol{\gamma},\boldsymbol{f}}, f_t \rangle$ — the steady-state cost at time t.

Moreover, for any other state feedback law $P \in \mathscr{M}(\mathsf{U}|\mathsf{X})$, we will denote by $\eta_t^{P,\boldsymbol{f}}$ the steady-state cost $\langle \pi_P \otimes P, f_t \rangle$, where π_P is the unique invariant distribution of $K(\cdot|\cdot, P)$.

It will be convenient to introduce the regret w.r.t. a fixed $P \in \mathscr{M}(\mathsf{U}|\mathsf{X})$ with initial state $X_1 = x$:

$$\begin{aligned} R_x^{\boldsymbol{\gamma},\boldsymbol{f}}(P) &\triangleq \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{f}} \left[\sum_{t=1}^T f_t(X_t, U_t) - \sum_{t=1}^T \eta_t^{P,\boldsymbol{f}} \right] \\ &= \sum_{t=1}^T \left[\langle \mu_t^{\boldsymbol{\gamma},\boldsymbol{f}} \otimes P_t^{\boldsymbol{\gamma},\boldsymbol{f}}, f_t \rangle - \langle \pi_P \otimes P, f_t \rangle \right], \end{aligned}$$

as well as the *stationarized regret*

$$\begin{aligned} \bar{R}^{\boldsymbol{\gamma},\boldsymbol{f}}(P) &\triangleq \sum_{t=1}^T \left(\eta_t^{\boldsymbol{\gamma},\boldsymbol{f}} - \eta_t^{P,\boldsymbol{f}} \right) \\ &= \sum_{t=1}^T \left[\langle \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}} \otimes P_t^{\boldsymbol{\gamma},\boldsymbol{f}}, f_t \rangle - \langle \pi_P \otimes P, f_t \rangle \right]. \end{aligned}$$

Using (1), we get the bound

$$R_x^{\boldsymbol{\gamma},\boldsymbol{f}}(P) \le \bar{R}^{\boldsymbol{\gamma},\boldsymbol{f}}(P) + \sum_{t=1}^T \|f_t\|_\infty \|\mu_t^{\boldsymbol{\gamma},\boldsymbol{f}} - \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}}\|_1. \tag{8}$$

The key observation here is that the task of analyzing the regret $R_x^{\boldsymbol{\gamma},\boldsymbol{f}}(P)$ splits into separately upper-bounding the two terms on the right-hand side of (8): the stationarized regret $\bar{R}^{\boldsymbol{\gamma},\boldsymbol{f}}(P)$ and the stationarization error $\sum_{t=1}^T \|f_t\|_\infty \|\mu_t^{\boldsymbol{\gamma},\boldsymbol{f}} - \pi_t^{\boldsymbol{\gamma},\boldsymbol{f}}\|_1$ using Markov chain techniques.

In order to analyze the stationarized regret, we introduce the *reverse Poisson inequality*. Fix a Markov matrix $P \in \mathscr{M}(\mathsf{U}|\mathsf{X})$ and let $\pi_P \in \mathscr{P}(\mathsf{X})$ be the (unique) invariant distribution of $K(\cdot|\cdot, P)$. Then we say that $\widehat{Q} : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ satisfies the reverse Poisson inequality with *forcing function* $g : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ if

$$\mathbb{E}\left[\widehat{Q}(Y, P) \middle| x, u \right] - \widehat{Q}(x, u) \ge -g(x, u) + \langle \pi_P \otimes P, g \rangle, \qquad \forall (x, u) \in \mathsf{X} \times \mathsf{U} \tag{9}$$

where

$$\widehat{Q}(y, P) \triangleq \sum_{u \in \mathsf{U}} P(u|y) \widehat{Q}(y, u)$$

and $\mathbb{E}[\cdot|x, u]$ is w.r.t. the transition law $K(y|x, u)$. We should think of this as a relaxation of the Poisson equation [22], i.e., when (9) holds with equality. The Poisson equation arises naturally in the theory of Markov chains and Markov decision processes, where it provides a way to evaluate the long-term average cost along the trajectory of a Markov process. We are using the term "reverse Poisson inequality" to distinguish (9) from the Poisson inequality, which also arises in the theory of Markov chains and is obtained by replacing $\ge$ with $\le$ in (9) [22]. Here we impose the following assumption that we use throughout the rest of the report:

Assumption 1 *For any $P \in \mathscr{M}(\mathsf{U}|\mathsf{X})$ and any $f \in \mathscr{F}$, there exists some $\widehat{Q}_{P,f} : \mathsf{X} \times \mathsf{U} \to \mathbb{R}$ that solves the reverse Poisson inequality for P with forcing function f. Moreover,*

$$L(\mathsf{X}, \mathsf{U}, \mathscr{F}) \triangleq \sup_{P \in \mathscr{M}(\mathsf{U}|\mathsf{X})} \sup_{f \in \mathscr{F}} \|\widehat{Q}_{P,f}\|_\infty < \infty.$$

Remark 4 In Section 5, we will show this assumption is automatically satisfied under an additional condition of uniform ergodicity.

The main consequence of the reverse Poisson inequality is the following:

Lemma 1 (Comparison principle) *Suppose that $\widehat{Q}$ satisfies the reverse Poisson inequality (9) with forcing function g. Then for any other Markov matrix P' we have*

$$\langle \pi_P \otimes P, g\rangle - \langle \pi_{P'} \otimes P', g\rangle \le \sum_x \pi_{P'}(x) \sum_u \left[P(u|x)\widehat{Q}(x,u) - P'(u|x)\widehat{Q}(x,u)\right]$$

Proof: See Appendix C. ■

Armed with this lemma, we can now analyze the stationarized regret $\bar{R}^{\gamma,f}(P)$: suppose that, for each t, $\widehat{Q}_t^{\gamma,f}$ satisfies reverse Poisson inequality for $P_t^{\gamma,f}$ with forcing function f_t. Then we apply the comparison principle to get

$$\eta_t^{\gamma,f} - \eta_t^{P,f} \le \sum_x \pi_P(x) \left(\sum_u P_t^{\gamma,f}(u|x)\widehat{Q}_t^{\gamma,f}(x,u) - P(u|x)\widehat{Q}_t^{\gamma,f}(x,u) \right).$$

This in turn yields

$$\begin{aligned} R_x^{\gamma,f}(P) \le & \sum_x \pi_P(x) \sum_{t=1}^{T} \left(\sum_u P_t^{\gamma,f}(u|x)\widehat{Q}_t^{\gamma,f}(x,u) - P(u|x)\widehat{Q}_t^{\gamma,f}(x,u) \right) \\ & + \sum_{t=1}^{T} \|f_t\|_\infty \|\mu_t^{\gamma,f} - \pi_t^{\gamma,f}\|_1. \end{aligned}$$

Note that $\widehat{Q}_t^{\gamma,f}$ depends functionally on $P_t^{\gamma,f}$ and on f_t, which in turn depend functionally on f^t but not on $f_{t+1},\dots,f_T$. This ensures that any algorithm using $\widehat{Q}_t^{\gamma,f}$ respects the causality constraint that any decision made at time t depends only on information available by time t.

Focusing on stationarized regret and upper-bounding it in terms of the $\widehat{Q}$-functions is one of the key steps that lets us consider a simpler setting without Markov dynamics. The next step is to define a new type of relaxation with an accompanying new admissibility condition for this simpler setting. That is, we will find a relaxation and admissibility condition for the stationarized regret rather than for the expected steady-state regret directly. A new admissibility condition is needed because we have decoupled current costs from past actions, which makes the previous admissibility condition (6) inapplicable. The new admissibility condition is similar to the one in [27], which was derived in a stateless setting. The difference is that we are still in a *state-dependent* setting in the sense that the new type of relaxation is indexed by the state variable. Now instead of having a Markov dynamics that depends on the state, we consider all the states in parallel and have a separate algorithm running on each state. The interaction between different states is generated by providing these algorithms with common information that comes from actual dynamical process. Thus, starting from this new admissibility condition, we further construct algorithms using relaxations and then use Lemma 1 to bound the regret of these algorithms.

4.2 A new admissibility condition and the main result

Now we are in a position to pass to a simpler setting without Markov dynamics. Instead, we associate each state with a separate game. Within each game, the agent chooses an action and observes a signal from the environment, and the current cost in each state is independent from the past actions taken in that state. The signal generated by the environment is the $\widehat{Q}$-function mentioned above in Assumption 1. Although here we don't use the one-step cost functions f_t as the signal, we know that the $\widehat{Q}$-functions

actually contain payoff-relevant information on f_t. From this perspective, the environment choose one-step cost functions f_t is equivalent to letting the environment choose the corresponding $\widehat{Q}$-functions.

To proceed, we need to introduce a new type of relaxation with a new admissibility condition. The reason is that the relaxation $\widehat{V}$ defined in (6) is a sequence of functions with a state variable, and the state is changing at every time step according to the state transition dynamics and the agent's action. If we view the interaction between the agent and the environment as a stochastic game, then the relaxation is indeed a sequence of upper bounds on the conditional values of this game. However, after stationarization, we end up with a *family* of relaxations indexed by the state variable — for each state, we have a separate online learning game, and we have a separate relaxation for each of these online learning games. Consequently, there is no Markov dynamics involved in each of the new relaxations. The new relaxation at each state $x \in \mathsf{X}$, which we will denote by $\{\widehat{W}_x\}_{t=1}^T$, is a sequence of upper bounds on the conditional value of the corresponding online learning game. We define such a relaxation as follows.

For each $x \in \mathsf{X}$, let $\mathscr{H}_x$ denote the class of all functions $h_x : \mathsf{U} \to \mathbb{R}$ for which there exist some $P \in \mathscr{M}(\mathsf{U}|\mathsf{X})$ and $f \in \mathscr{F}$, such that

$$h_x(u) = \widehat{Q}_{P,f}(x,u), \qquad \forall\, u \in \mathsf{U}.$$

We say that a sequence of functions $\widehat{W}_{x,t} : \mathscr{H}_x^t \to \mathbb{R}$, $t = 0, \ldots, T$, is an admissible relaxation at state x if the following condition holds for any $h_{x,1}, \ldots, h_{x,T} \in \mathscr{H}_x$:

$$\widehat{W}_{x,T}(h_x^T) \geq -\inf_{\nu \in \mathscr{P}(\mathsf{U})} \mathbb{E}_{U\sim\nu}\left[\sum_{t=1}^T h_{x,t}(U)\right], \tag{10a}$$

$$\widehat{W}_{x,t}(h_x^t) \geq \inf_{\nu \in \mathscr{P}(\mathsf{U})} \sup_{h_x \in \mathscr{H}_x} \left\{\mathbb{E}_{U\sim\nu}[h(U)] + \widehat{W}_{x,t+1}(h_x^t, h_x)\right\}, \qquad t = T-1, \ldots, 0. \tag{10b}$$

Given such an admissible relaxation, we can associate to it a behavioral strategy

$$\widehat{\gamma}_t(x, f^{t-1}) = P_t^{\widehat{\boldsymbol{\gamma}}, \boldsymbol{f}}(\cdot|x) = \operatorname*{arg\,min}_{\nu \in \mathscr{P}(\mathsf{U})} \sup_{h_x \in \mathscr{H}_x} \left\{\mathbb{E}_{U\sim\nu}[h_x(U)] + \widehat{W}_{x,t}(h_x^{t-1}, h_x)\right\}$$

$$h_{y,t} = \widehat{Q}_t^{\widehat{\boldsymbol{\gamma}}, \boldsymbol{f}}(y, \cdot), \quad \forall\, y \in \mathsf{X}.$$

(Even though the above notation suggests the dependence of $h_{y,t}$ on the T-tuples $\boldsymbol{\gamma}$ and $\boldsymbol{f}$, this dependence at time t is only w.r.t. γ^t and f^t, so the resulting strategy is still causal). The relaxation $\{\widehat{W}_x\}_{t=1}^T$ at state x is a sequence of upper bounds on the conditional value of the online learning game associated with that state. In this game, at time step t, the agent chooses actions $u_t \in \mathsf{U}$ and the environment chooses function $h_{x,t} \in \mathscr{H}_x$. Although this relaxation is still state-dependent, there is no Markov dynamics involved here, which means that now the state-free techniques of [27] can be brought to bear on the problem of constructing algorithms and bounding their regret. Specifically, we derive a separate relaxation $\{\widehat{W}_x\}_{t=1}^T$ and the associated behavioral strategy for each state $x \in \mathsf{X}$. Then we assemble these into an overall algorithm for the MDP as follows: if at time t the state $X_t = x$, the agent will choose actions according to the corresponding behavioral strategy $\widehat{\gamma}_t(x, \cdot)$. Note that although the agent's behavioral strategy switches between different relaxations depending on the current state, the agent still needs to update all the h-functions simultaneously for all the states. This is because the computation of the h-functions (in terms of the $\widehat{Q}$ functions) requires the knowledge of the behavioral strategy at other states. In other words, the algorithm has to keep updating all the relaxations in parallel for all states.

Under the constructed relaxation, we state our main result:

Theorem 1 *Suppose that the MDP is unichain, the environment is oblivious, and Assumption 1 holds. Then, for any family of admissible relaxations given by* (10) *and the corresponding behavioral strategy* $\widehat{\boldsymbol{\gamma}}$*,*

we have

$$R_x^{\widehat{Y},\boldsymbol{f}} = \mathbb{E}_x^{\widehat{Y},\boldsymbol{f}}\left[\sum_{t=1}^{T} f_t(X_t,U_t) - \Psi(\boldsymbol{f})\right] \leq \sup_{P\in\mathscr{M}(\mathsf{U}|\mathsf{X})} \sum_x \pi_P(x)\widehat{W}_{x,0} + C_{\mathscr{F}} \sum_{t=1}^{T} \|\mu_t^{\widehat{Y},\boldsymbol{f}} - \pi_t^{\widehat{Y},\boldsymbol{f}}\|_1 \tag{11}$$

where $C_{\mathscr{F}} = \sup_{f\in\mathscr{F}} \|f\|_\infty$.

Proof: See Appendix D. ■

This general framework gives us a recipe for deriving algorithms for online MDPs. First, we use stationarization to pass to a simpler setting without Markov dynamics. Here we need to find $\widehat{Q}_t$ functions satisfying (9) with forcing function f_t at each time t. In this simpler setting, we associate each state with a separate online learning game. Next, we derive appropriate relaxations (upper bounds on the conditional values) for each of these online learning games. Then we plug the relaxation into the admissibility condition (10) to derive the associated algorithm. This algorithm in turn gives us a behavioral strategy for the original online MDP problem, and Theorem 1 automatically gives us a regret bound for this strategy. We emphasize that, in general, multiple different relaxations are possible for a given problem, allowing for a flexible tradeoff between computational costs and regret.

We have reduced the original problem to a collection of standard online learning problems, each of which is associated with a particular state. We proceed by constructing a separate relaxation for each of these problems. Because we have removed the Markov dynamics, we may now use available techniques for constructing these relaxations. In particular, as shown by [26], a particularly versatile method for constructing relaxations relies on the notion of *sequential Rademacher complexity* (SRC).

5 Example derivations of explicit algorithms using our framework

The algorithms derived using our general framework belongs to a class of algorithms called expert advice algorithm [7]. Expert advice algorithm is a well-known method that combines the recommendation of several individual "experts" into another strategy of choosing actions. Every expert is assigned a "weight" indicating how much the agent trusts that expert, based on the previous performance of the experts. The weights direct the agent with regard to which expert to follow at the next time step. The most popular algorithm for the expert advice framework is the Randomized Weighted Majority algorithm (RWM), sometimes called Hedge. RWM maintains a set of weights over experts and updates these weights multiplicatively. It has an alternative interpretation from a regularization perspective. It has been shown that the weights chosen by this RWM algorithm minimize a combination of empirical cost and an entropic regularization term. This observation makes RWM algorithm a special case of a broader class of algorithms known as follow the regularized leader (FRL) algorithms [30]. In order to add some stability, RWM induces high entropy, which leads to more uniform weights. The algorithms we derive in this section are examples in which RWM algorithms are applied to each state.

5.1 Recovering an expert-based algorithm for online MDPs

Similar to our set-up, [9] consider an MDP with arbitrarily varying cost functions. The main idea of their work is to efficiently incorporate existing expert-based algorithms [7, 20] into the MDP setting. For an MDP with state space X and action space U, there are $|\mathsf{U}|^{|\mathsf{X}|}$ deterministic Markov policies (state feedback laws), which renders the obvious approach of associating an expert with each possible deterministic policy computationally infeasible. Instead, they propose an alternative efficient scheme that works by associating a separate expert algorithm to each *state*, where experts correspond to *actions* and the feedback to provided each expert algorithm depends on the aggregate policy determined by the action choices of all the individual algorithms. Under a unichain assumption similar to the one we have made above, they

show that the expected regret of their algorithm is sublinear in T and independent of the size of the state space. Their algorithm can be summarized as follows:

> Put in every state x an expert algorithm B_x
> for $t = 1, 2, \ldots$ do
> Let $P_t(\cdot|x_t)$ be the distribution over actions of B_{x_t}
> Use policy P_t and obtain f_t from the environment
> Feed B_x with loss function $\widehat{Q}_{P_t,f_t}(x,\cdot) = \mathbb{E}\left[\sum_{i=0}^{\infty}\left(f_t(X_i,U_i) - \eta^{P_t,f_t}\right)\right]$,
> where $\mathbb{E}$ is taken w.r.t. the Markov chain induced by P_t from the initial state x.
> $\eta_t^{P,\boldsymbol{f}}$ is the steady-state cost $\langle \pi_{P_t} \otimes P_t, f_t \rangle$, where π_{P_t} is the invariant distribution of P_t.
> end for

As we show next, the algorithm proposed by [9] arises from a particular relaxation of the kind that was introduced in the preceding section. For every possible state value $x \in \mathsf{X}$, we want to construct an admissible relaxation that satisfies (10). Here we show that the relaxation can be obtained as an upper bound of a quantity called *conditional sequential Rademacher complexity*, which is defined in [27] as follows. Let ε be a vector $(\varepsilon_1, \ldots, \varepsilon_T)$ of i.i.d. Rademacher random variables, i.e., $\Pr(\varepsilon_i = \pm 1) = 1/2$. For a given $x \in \mathsf{X}$, an $\mathscr{H}_x$-valued tree $\mathbf{h}$ of depth d is defined as a sequence $(\mathbf{h}_1, \ldots, \mathbf{h}_d)$ of mappings $\mathbf{h}_t : \{\pm 1\}^{t-1} \to \mathscr{H}_x$, where $\mathscr{H}_x$ is the function class defined in Section 4.2. Then the conditional sequential Rademacher complexity at state x is defined as

$$\mathscr{R}_{x,t}(h_x^t) = \sup_{\mathbf{h}} \mathbb{E}_{\varepsilon_{t+1:T}} \max_{u \in \mathsf{U}} \left[2 \sum_{s=t+1}^{T} \varepsilon_s \left[\mathbf{h}_{s-t}(\varepsilon_{t+1:s-1})\right](u) - \sum_{s=1}^{t} h_{x,s}(u) \right], \qquad \forall h_x^t \in \mathscr{H}_x^t.$$

Here the supremum is taken over all $\mathscr{H}_x$-valued binary trees of depth $T-t$. The term containing the tree $\mathbf{h}$ can be seen as "future", while the term being subtracted off can be seen as "past". This quantity is conditioned on the already observed h_x^t, while for the future we consider the worst possible binary tree. As shown by [27], this Rademacher complexity is itself an admissible relaxation for standard (state-free) online optimization problems; moreover, one can obtain other relaxations by further upper-bounding the Rademacher complexity. As we will now show, because the action space U is finite and the functions in $\mathscr{H}_x$ are uniformly bounded (Assumption 1), the following upper bound on $\mathscr{R}_{x,t}(\cdot)$ is an admissible relaxation, i.e., it satisfies condition (10):

$$\widehat{W}_{x,t}(h_x^t) = \rho \log\left(\sum_{u \in \mathsf{U}} \exp\left(-\frac{1}{\rho} \sum_{s=1}^{t} h_{x,s}(u) \right) \right) + \frac{2}{\rho}(T-t)L(\mathsf{X},\mathsf{U},\mathscr{F})^2, \tag{12}$$

where the learning rate $\rho > 0$ can be tuned to optimize the resulting regret bound. This relaxation leads to an algorithm that turns out to be exactly the scheme proposed by [9]:

Proposition 3 *The relaxation* (12) *is admissible and it leads to a recursive exponential weights algorithm, specified recursively as follows: for all* $x \in \mathsf{X}$, $u \in \mathsf{U}$

$$P_{t+1}(u|x) = \frac{P_t(u|x)\exp\left(-\frac{1}{\rho}h_{x,t}(u)\right)}{\left\langle P_t(\cdot|x), \exp\left(-\frac{1}{\rho}h_{x,t}\right)\right\rangle} = \frac{\nu_1(u)\exp\left(-\frac{1}{\rho}\sum_{s=1}^{t}h_{x,s}(u)\right)}{\left\langle \nu_1, \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t}h_{x,s}\right)\right\rangle}, \qquad t = 0, \ldots, T-1 \tag{13}$$

where ν_1 *is the uniform distribution on* U.

Proof: See Appendix E. ∎

The above algorithm works with any collection of $\widehat{Q}$ functions satisfying the reverse Poisson inequalities determined by the f_t's (recall Assumption 1). Here is one particular example of such a function — the usual Q-function that arises in reinforcement learning and that was used by [9]. Recall our assumption that every stochastic state feedback law $P \in \mathscr{M}(\mathsf{U}|\mathsf{X})$ has a unique stationary distribution π_P. For given choices of $P \in \mathscr{M}(\mathsf{U}|\mathsf{X})$ and $f \in \mathscr{F}$, consider the function

$$\widehat{Q}_{P,f}(x,u) = \lim_{T\to\infty} \mathbb{E}_P \left[\sum_{t=1}^{T} f(X_t, U_t) - \langle \pi_P \otimes P, f \rangle \middle| X_1 = x, U_1 = u \right],$$

where X_t and U_t are the state and action at time step t after starting from the initial state $X_1 = x$, applying the immediate action $U_1 = u$, and following P onwards. It is easy to check that $\widehat{Q}_{P,f}(x,u)$ satisfies the reverse Poisson inequality for P with forcing function f. In fact, it satisfies (9) with equality. We can also derive a bound on the Q-function as a function of the mixing time τ. Let us first bound $\widehat{Q}_{P,f}(x,P)$ where P is used on the first step instead of u. For all t, let $\mu_{x,t}^{P,f}$ be the state distribution at time t starting from x and following P onwards. So we have

$$\begin{aligned}
\widehat{Q}_{P,f}(x,P) = \lim_{T\to\infty} \sum_{t=1}^{T} \left[\langle \mu_{x,t}^{P,f} \otimes P, f \rangle - \langle \pi_P \otimes P, f \rangle \right] &\le \|f\|_\infty \sum_{t=1}^{T} \|\delta_x P^t - \pi_P P^t\|_1 \\
&\le 2\|f\|_\infty \sum_{t=1}^{T} e^{-t/\tau} \\
&\le 2\tau \|f\|_\infty,
\end{aligned}$$

where the first inequality results from repeated application of the uniform mixing bound (7). Due to the fact that the one-step cost is bounded by $C_{\mathscr{F}} = \sup_{f\in\mathscr{F}} \|f\|_\infty$, we have

$$\widehat{Q}_{P,f}(x,u) \le \widehat{Q}_{P,f}(x,P) + f(x,u) - \langle \mu_{x,1}^{P,f} \otimes P, f \rangle \le 2\tau C_F + C_F \le 3\tau C_F.$$

We can now establish the following regret bound for the exponential weights strategy (13):

Theorem 2 *Let $L \triangleq L(\mathsf{X}, \mathsf{U}, \mathscr{F})$. Assume the state transition dynamics have a unichain structure. Then for the relaxation (12) and the corresponding behavioral strategy $\widehat{\boldsymbol{\gamma}}$ given by (13) with an appropriate choice of ρ, we have*

$$\mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} \left[\sum_{t=1}^{T} f_t(X_t, U_t) - \Psi(\boldsymbol{f}) \right] \le 2L\sqrt{2T\log|\mathsf{U}|} + C_{\mathscr{F}}\tau^2 \sqrt{\frac{\log|\mathsf{U}|\,T}{2}} + 2\tau C_{\mathscr{F}}.$$

Proof: See Appendix F. ∎

As we can see, this regret bound is consistent with the bound derived in [9]. Therefore, we have shown that our framework, with a specific choice of relaxation, can recover their algorithm. The advantage of our general framework is that we can analyze the part of the corresponding regret bound simply by instantiating our analysis on specific relaxations, without the need of ad-hoc proof techniques applied in [9].

5.2 A novel lazy RWM algorithm for online MDPs

In the preceding section, we used our framework to recover a particular policy for an online MDP that relies on exponential weight updates. In this section, we derive a "lazy" version of that policy, which can be interpreted as a lazy RWM algorithm. The starting point is to divide time into phases of increasing

length, so that during each phase the agent applies a fixed state feedback law. The main advantage of lazy strategies is their computational efficiency, which is the result of a looser relaxation and hence suboptimal scaling of the regret with the time horizon.

We partition the set of time indices $1,2,\ldots$ into nonoverlapping contiguous phases of (possibly) increasing duration. The phases are indexed by $m \in \mathbb{N}$, where we denote the mth phase by $\mathscr{T}_m$ and its duration by τ_m. We also define $\mathscr{T}_{1:m} \triangleq \mathscr{T}_1 \cup \ldots \cup \mathscr{T}_m$ (the union of phases 1 through m) and denote its duration by $\tau_{1:m}$. Let $M \leq T$ denote the number of complete phases concluded before time T. Here we need a describe a generic algorithm that works in phases:

> Initialize at $t=0$ and phases $\mathscr{T}_1,\ldots,\mathscr{T}_M$ s.t. $\tau_{1:M}=T$
> For $t \in \mathscr{T}_1$, choose u_t uniformly at random over U
> for $m=2,3,\ldots$
> for $t \in \mathscr{T}_m$ do
> if the process is at state x_t, choose action u_t randomly according to $P_m(u|x)$
> where $P_m(u|x)$ is the state feedback law only using information from phase 1 to $m-1$
> end for
> end for

Because in this section we work in phases instead of time steps, we need to provide an alternative definition of relaxations and admissibility condition. For every state $x \in \mathsf{X}$, we denote by h_x^m the τ_m-tuple $(h_{x,s} : s \in \mathscr{T}_m)$, and by $h_{x,1:m}$ the $\tau_{1:m}$-tuple $(h_{x,1}, h_{x,2}, \ldots, h_{x,\tau_{1:m}})$. For each $x \in \mathsf{X}$, we will say that a sequence of functions $\widehat{W}_{x,m} : \mathscr{H}_x^{\tau_{1:m}} \to \mathbb{R}$, $m=1,\ldots,M$, is an admissible relaxation if

$$\widehat{W}_{x,M}(h_{x,1:M}) \geq -\inf_{\nu \in \mathscr{P}(\mathsf{U})} \mathbb{E}_{U\sim\nu}\left[\sum_{t=1}^{T} h_{x,t}(U)\right]$$

$$\widehat{W}_{x,m}(h_{x,1:m}) \geq \inf_{\nu \in \mathscr{P}(\mathsf{U})} \sup_{h_x^m \in \mathscr{H}_x^{\tau_m}} \left\{ \mathbb{E}_{U\sim\nu}\left[\sum_{s\in\mathscr{T}_m} h_{x,s}(U)\right] + \widehat{W}_{x,m+1}(h_{x,1:m}, h_x^{m+1}) \right\}, \quad t = M-1,\ldots,1$$

For a given state x, we also define the conditional sequential Rademacher complexity in terms of phases:

$$\mathscr{R}_{x,m}(h_{x,1:m}) = \sup_{\mathbf{h}} \mathbb{E}_{\varepsilon_{m+1:M}} \max_{u\in\mathsf{U}} \left[2\sum_{j=m+1}^{M} \varepsilon_j \sum_{t\in\mathscr{T}_j} \left[\mathbf{h}_{x,t}(\varepsilon)\right](u) - \sum_{i=1}^{m}\sum_{s\in\mathscr{T}_i} h_{x,s}(u) \right].$$

Here the supremum is taken over all $\mathscr{H}_x$-valued binary trees of depth $M-m$. In the preceding section, we replaced the actual future induced by the infimum and supremum pairs in the conditional value by the "worst future" binary tree, which involves expectation over a sequence of coin flips in every time step. By contrast, in the above quantity we replace the real future by the "worst future" binary tree that branches only once per phase. Now we can construct the following relaxation:

$$\widehat{W}_{x,m}(h_{x,1:m}) = \rho \log\left(\sum_{u\in\mathsf{U}} \exp\left(-\frac{1}{\rho} \sum_{i=1}^{m}\sum_{s\in\mathscr{T}_i} h_{x,s}(u) \right) \right) + \frac{2L(\mathsf{X},\mathsf{U},\mathscr{F})^2}{\rho} \sum_{j=m+1}^{M} \tau_j^2. \tag{14}$$

The corresponding algorithm, specified in (15) below, uses a fixed state feedback law throughout each phase:

Proposition 4 *The relaxation* (14) *is admissible and it leads to the following Markov policy for phase m:*

$$P_{t,m}(u|x) = \frac{\nu_1(u) \exp\left(-\frac{1}{\rho}\sum_{i=1}^{m-1}\sum_{s\in\mathscr{T}_i} h_{x,s}(u)\right)}{\left\langle \nu_1, \exp\left(-\frac{1}{\rho}\sum_{i=1}^{m-1}\sum_{s\in\mathscr{T}_i} h_{x,s}\right)\right\rangle}, \tag{15}$$

where ν_1 is the uniform distribution on U.

Proof: See Appendix G. ■

Now we derive the regret bound for (15):

Theorem 3 *Let $L \triangleq L(\mathsf{X},\mathsf{U},\mathscr{F})$. Under the same assumptions as before, the behavioral strategy $\widehat{\boldsymbol{\gamma}}$ corresponding to (15) enjoys the following regret bound:*

$$\mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T} f_t(X_t,U_t)-\Psi(\boldsymbol{f})\right] \leq 2L\sqrt{2\log|\mathsf{U}|\sum_{i=1}^{M}\tau_i^2}+\frac{2C_{\mathscr{F}}M}{1-e^{-1/\tau}}. \tag{16}$$

Proof: See Appendix H. ■

Our behavioral strategy (13) is a mixed strategy for choosing actions, and is essentially a randomized weighted majority (RWM) algorithm developed in online learning. As a result, our recursive algorithm (13) can be seen as a FRL algorithm. We add randomness automatically when choosing actions by using a mixed strategy; effectively, we are using a FPL algorithm to choose actions. By presenting a "lazy" version of that recursive algorithm, we derive a novel lazy FPL algorithm which is similar in spirit to the algorithm in [33].

[33] also consider a similar model where the decision-maker has full knowledge of the transition kernel and the costs are chosen by an adversary. They propose an algorithm for MDPs with arbitrarily changing costs that achieves sublinear regret based on the oblivious opponent assumption. Their algorithm computes and changes policy periodically according to a perturbed version of the empirically observed cost functions, and follows the computed stationary policy for increasingly long time intervals. As a result, their algorithm has diminishing computational effort per time step and is computationally more efficient than that of [9]. [33] call their algorithm the lazy FPL algorithm.

Although our new algorithm is similar in nature to the algorithm presented in [33], our method has several advantages. First, in the algorithm of [33], the policy computation at the beginning of each phase requires solving a linear program and then adding a carefully tuned random perturbation to the solution. As a result, the performance analysis in [33] is rather lengthy and technical (in particular, it invokes several advanced results from perturbation theory for linear programs). By contrast, we can analyze the part of the corresponding regret bound simply by instantiating our analysis on specific relaxations, and we don't need to add additional randomization, which renders our proof much less technical. Second, the regret bound of Theorem 3 shows that we can control the scaling of the regret with T by choosing the duration of each phase. By contrast, the algorithm of [33] relies on a specific choice of phase durations in order to guarantee that the regret is sublinear in T and scales as $O(T^{3/4})$. We show that if the horizon T is known in advance, then it is possible to choose the phase durations to secure $O(T^{2/3})$ regret, which is better than the $O(T^{3/4})$ bound derived by [33].

Corollary 1 *For a given horizon T, the optimal choice of phase lengths is $T^{1/3}$, which gives the regret of $O(T^{2/3})$.*

Proof: See Appendix I. ■

6 Conclusions

We provide a unified viewpoint on the design and the analysis of online MDPs algorithms which is an extension of a general relaxation-based approach of [26] to a certain class of stochastic game models. We showed that an algorithm previously proposed by [9] naturally arises from our framework via a specific relaxation. Moreover, we showed that one can obtain lazy strategies (where time is split into phases, and a different stationary policy is followed in each phase) by means of relaxations as well. In particular, we

have obtained a new strategy, which is similar in spirit to the one previously proposed by [33], but with several advantages, including better scaling of the regret.

A Proof of Proposition 1

The agent's closed-loop behavioral strategy $\boldsymbol{\gamma}$ is a tuple of mappings $\gamma_t : \mathscr{F}^{t-1} \to \mathscr{P}(\mathsf{U}), 1 \le t \le T$; the environment's open-loop behavior strategy $\boldsymbol{f}$ is a tuple of functions $(f_1, \dots, f_T)$ in $\mathscr{F}$. Thus,

$$\begin{aligned} V(x) &= \inf_{\boldsymbol{\gamma}} \sup_{\boldsymbol{f}} \mathbb{E}_x^{\boldsymbol{\gamma},\boldsymbol{f}} \left[\sum_{t=1}^{T} f_t(X_t, U_t) - \Psi(\boldsymbol{f}) \right] \\ &= \inf_{\gamma_1} \dots \inf_{\gamma_T} \sup_{f_1} \dots \sup_{f_T} \mathbb{E}_x^{\gamma_1,\dots,\gamma_T,f_1,\dots,f_T} \left[\sum_{t=1}^{T} f_t(X_t, U_t) - \Psi(\boldsymbol{f}) \right]. \end{aligned}$$

We start from the final step T and proceed by backward induction. Assuming $\gamma_1, \dots, \gamma_{T-1}$ were already chosen, we have

$$\begin{aligned} &\inf_{\gamma_T} \sup_{f_1,\dots,f_T} \mathbb{E}_x^{\gamma^{T-1},\gamma_T,f^{T-1},f_T} \left\{ \sum_{t=1}^{T-1} \left[f_t(X_t, U_t) \right] + f_T(X_T, U_T) - \Psi(f^T) \right\} \\ &= \inf_{\gamma_T} \sup_{f_1,\dots,f_{T-1}} \sup_{f_T} \left\{ \mathbb{E}_x^{\gamma^{T-1},f^{T-1}} \left(\sum_{t=1}^{T-1} \left[f_t(X_t, U_t) \right] \right) + \mathbb{E}_x^{\gamma^{T-1},\gamma_T,f^{T-1},f_T} \left[f_T(X_T, U_T) - \Psi(f^T) \right] \right\} \\ &= \inf_{\gamma_T} \sup_{f_1,\dots,f_{T-1}} \left\{ \mathbb{E}_x^{\gamma^{T-1},f^{T-1}} \left(\sum_{t=1}^{T-1} \left[f_t(X_t, U_t) \right] \right) + \sup_{f_T} \mathbb{E}_x^{\gamma^{T-1},\gamma_T,f^{T-1},f_T} \left[f_T(X_T, U_T) - \Psi(f^T) \right] \right\} \\ &= \sup_{f_1,\dots,f_{T-1}} \left\{ \mathbb{E}_x^{\gamma^{T-1},f^{T-1}} \left(\sum_{t=1}^{T-1} \left[f_t(X_t, U_t) \right] \right) + \inf_{P_T(U_T|X_T)} \sup_{f_T} \mathbb{E}_x^{\gamma^{T-1},\gamma_T,f^{T-1},f_T} \left[f_T(X_T, U_T) - \Psi(f^T) \right] \right\}. \end{aligned}$$

The last step is due to the easily proved fact that, for any two sets A, B and bounded functions $g_1 : A \to \mathbb{R}$, $g_2 : A \times B \to \mathbb{R}$,

$$\inf_{\gamma : A \to B} \sup_{a} \left\{ g_1(a) + g_2(a, \gamma(a)) \right\} = \sup_{a} \left[g_1(a) + \inf_{b \in B} g_2(a, b) \right]$$

(see, e.g., Lemma 1.6.1 in [5]). Proceeding inductively in this way, we get (4).

B Proof of Proposition 2

The proof is by backward induction. Starting at time T and using the admissibility condition (6), we write

$$
\begin{aligned}
&\mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T} f_t(X_t,U_t)-\Psi(\boldsymbol{f})\right]\\
&\le \mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T} f_t(X_t,U_t)+\widehat{V}_T(X_{T+1},f^T)\right]\\
&= \mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T-1} f_t(X_t,U_t)\right]+\mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[f_T(X_T,U_T)+\widehat{V}_T(X_{T+1},f^T)\right]\\
&= \mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T-1} f_t(X_t,U_t)\right]\\
&\qquad +\sum_{x_T}\mu_T(x_T)\left\{\sum_{u\in\mathsf{U}} f_T(x_T,u)\left[\widehat{\gamma}_T\left(x_T,f^{T-1}\right)\right](u)+\mathbb{E}\left[\widehat{V}_T(X_{T+1},f^T)\middle| x_T,\widehat{\gamma}_T\left(x_T,f^{T-1}\right)\right]\right\}\\
&\le \mathbb{E}_x^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\left[\sum_{t=1}^{T-1} f_t(X_t,U_t)+\widehat{V}_{T-1}(X_T,f^{T-1})\right],
\end{aligned}
$$

where $\mu_T\in\mathscr{P}(\mathsf{X})$ denotes the probability distribution of X_T. The last inequality is due to the fact that $\widehat{\boldsymbol{\gamma}}$ is the behavioral strategy associated to the admissible relaxation $\{\widehat{V}_t\}_{t=0}^T$. Continuing in this manner, we complete the proof.

C Proof of Lemma 1

Let us take expectations of both sides of (9) w.r.t. $\pi_{P'}\otimes P'$:

$$
\begin{aligned}
\langle \pi_P\otimes P,g\rangle-\langle \pi_{P'}\otimes P',g\rangle &\le \mathbb{E}_{\pi_{P'}\otimes P'}\left\{\mathbb{E}[\widehat{Q}(Y,P)|X,U]-\widehat{Q}(X,U)\right\}\\
&=\sum_{x,u}\pi_{P'}(x)P'(u|x)\left\{\mathbb{E}[\widehat{Q}(Y,P)|x,u]-\widehat{Q}(x,u)\right\}\\
&=\sum_{x,u}\pi_{P'}(x)P'(u|x)\mathbb{E}[\widehat{Q}(Y,P)|x,u]-\sum_{x,u}\left(\sum_y \pi_{P'}(y)K(x|y,P')\right)P'(u|x)\widehat{Q}(x,u)
\end{aligned}
$$

where in the third step we have used the fact that $\pi_{P'}$ is invariant w.r.t. $K(\cdot|\cdot,P')$. Then we have

$$
\begin{aligned}
&\sum_{x,u}\pi_{P'}(x)P'(u|x)\mathbb{E}[\widehat{Q}(Y,P)|x,u]-\sum_{x,u}\left(\sum_y \pi_{P'}(y)K(x|y,P')\right)P'(u|x)\widehat{Q}(x,u)\\
&\qquad=\sum_x \pi_{P'}(x)\left\{\sum_u P'(u|x)\mathbb{E}[\widehat{Q}(Y,P)|x,u]-\sum_{u,y}K(y|x,P')P'(u|y)\widehat{Q}(y,u)\right\}\\
&\qquad=\sum_x \pi_{P'}(x)\left\{\sum_u P'(u|x)\mathbb{E}[\widehat{Q}(Y,P)|x,u]-\sum_{y}K(y|x,P')\widehat{Q}(y,P')\right\},
\end{aligned}
$$

where the second step is by definition of $\widehat{Q}(y,P')$. Then we can write

$$\begin{aligned}
&\sum_x \pi_{P'}(x)\left\{\sum_u P'(u|x)\mathbb{E}[\widehat{Q}(Y,P)|x,u]-\sum_y K(y|x,P')\widehat{Q}(y,P')\right\}\\
&\quad\overset{(a)}{=}\sum_x \pi_{P'}(x)\left\{\sum_u P'(u|x)\left(\mathbb{E}[\widehat{Q}(Y,P)|x,u]-\sum_y K(y|x,u)\widehat{Q}(y,P')\right)\right\}\\
&\quad=\sum_{x,u}\pi_{P'}(x)P'(u|x)\Big\{\mathbb{E}[\widehat{Q}(Y,P)|x,u]-\mathbb{E}[\widehat{Q}(Y,P')|x,u]\Big\}\\
&\quad\overset{(b)}{=}\sum_{x,u,y}\pi_{P'}(x)P'(u|x)K(y|x,u)\left\{\sum_{u'}P(u'|y)\widehat{Q}(y,u')-\sum_{u'}P'(u'|y)\widehat{Q}(y,u')\right\}\\
&\quad\overset{(c)}{=}\sum_{x,y}\pi_{P'}(x)K(y|x,P')\left\{\sum_{u'}P(u'|y)\widehat{Q}(y,u')-\sum_{u'}P'(u'|y)\widehat{Q}(y,u')\right\}\\
&\quad\overset{(d)}{=}\sum_x\pi_{P'}(x)\sum_u\left[P(u|x)\widehat{Q}(x,u)-P'(u|x)\widehat{Q}(x,u)\right],
\end{aligned}$$

where (a) and (c) are by definition of $K(\cdot|\cdot,P')$; (b) is by definition of $\widehat{Q}(y,P')$; and in (d) we use the fact that $\pi_{P'}$ is invariant w.r.t. $K(\cdot|\cdot,P')$.

D Proof of Proposition 1

We have

$$\begin{aligned}
&\mathbb{E}_x^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\left[\sum_{t=1}^T f_t(X_t,U_t)-\Psi(\boldsymbol{f})\right]\\
&\le\sup_{P\in\mathscr{M}(\mathsf{U}|\mathsf{X})}\sum_{t=1}^T\left[\langle\pi_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\otimes P_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}},f_t\rangle-\langle\pi_P\otimes P,f_t\rangle\right]+\sum_{t=1}^T\|f_t\|_\infty\|\mu_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}-\pi_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\|_1\\
&\le\sup_{P\in\mathscr{M}(\mathsf{U}|\mathsf{X})}\sum_x\pi_P(x)\sum_{t=1}^T\left(\sum_u P_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u)-P(u|x)\widehat{Q}_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u)\right)+\sum_{t=1}^T\|f_t\|_\infty\|\mu_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}-\pi_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\|_1,
\end{aligned}$$

where in the first equality we have used (8), while the second inequality is by Lemma 1. Then we write the last term out and get

$$\begin{aligned}
&\sup_{P\in\mathscr{M}(\mathsf{U}|\mathsf{X})}\sum_x\pi_P(x)\left[\sum_{t=1}^{T-1}\left(\sum_u P_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u)\right)+\sum_u P_T^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(u|x)\widehat{Q}_T^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u)-\sum_{t=1}^T P(u|x)\widehat{Q}_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u)\right]\\
&\quad+\sum_{t=1}^T\|f_t\|_\infty\|\mu_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}-\pi_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\|_1\\
&\le\sup_{P\in\mathscr{M}(\mathsf{U}|\mathsf{X})}\sum_x\pi_P(x)\left[\sum_{t=1}^{T-1}\left(\sum_u P_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u)\right)+\sum_u P_T^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(u|x)\widehat{Q}_T^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u)+\widehat{W}_{x,T}(h_x^T)\right]\\
&\quad+\sum_{t=1}^T\|f_t\|_\infty\|\mu_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}-\pi_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\|_1\\
&\le\sup_{P\in\mathscr{M}(\mathsf{U}|\mathsf{X})}\sum_x\pi_P(x)\left[\sum_{t=1}^{T-1}\left(\sum_u P_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u)\right)+\widehat{W}_{x,T-1}(h_x^{T-1})\right]+\sum_{t=1}^T\|f_t\|_\infty\|\mu_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}-\pi_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\|_1,
\end{aligned}$$

where the two inequalities are by the admissibility condition (10). Continuing this induction backward, and noting that $\sum_{t=1}^T\|f_t\|_\infty\|\mu_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}-\pi_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\|_1\le C_{\mathscr{F}}\sum_{t=1}^T\|\mu_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}-\pi_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\|_1$, we arrive at (11).

E Proof of Proposition 3

First we show that the relaxation (12) arises as an upper bound on the conditional sequential Rademacher complexity. The proof of this is similar to the one given by [26], except that they also optimize over the choice of the learning rate ρ. For any $\rho > 0$,

$$\begin{aligned}
&\mathbb{E}_\varepsilon\left[\max_{u\in\mathsf{U}}\left\{2\sum_{i=1}^{T-t}\varepsilon_i\left[\mathbf{h}_i(\varepsilon)\right](u)-\sum_{s=1}^{t}h_{x,s}(u)\right\}\right]\\
&\le\rho\log\left(\mathbb{E}_\varepsilon\left[\max_{u\in\mathsf{U}}\exp\left(\frac{2}{\rho}\sum_{i=1}^{T-t}\varepsilon_i\left[\mathbf{h}_i(\varepsilon)\right](u)-\frac{1}{\rho}\sum_{s=1}^{t}h_{x,s}(u)\right)\right]\right)\\
&\le\rho\log\left(\mathbb{E}_\varepsilon\left[\sum_{u\in\mathsf{U}}\exp\left(\frac{2}{\rho}\sum_{i=1}^{T-t}\varepsilon_i\left[\mathbf{h}_i(\varepsilon)\right](u)-\frac{1}{\rho}\sum_{s=1}^{t}h_{x,s}(u)\right)\right]\right),
\end{aligned}$$

where the first inequality is by Jensen's inequality, while the second inequality is due to the non-negativity of exponential function. Then we pull out the second term inside the expectation $\mathbb{E}_\varepsilon$ and get

$$\begin{aligned}
&\rho\log\left(\sum_{u\in\mathsf{U}}\exp\left(-\frac{1}{\rho}\sum_{s=1}^{t}h_{x,s}(u)\right)\mathbb{E}_\varepsilon\left[\prod_{i=1}^{T-t}\exp\left(\frac{2}{\rho}\varepsilon_i\left[\mathbf{h}_i(\varepsilon)\right](u)\right)\right]\right)\\
&\le\rho\log\left(\sum_{u\in\mathsf{U}}\exp\left(-\frac{1}{\rho}\sum_{s=1}^{t}h_{x,s}(u)\right)\times\exp\left(\frac{2}{\rho^2}\max_{\varepsilon_1,\dots,\varepsilon_{T-t}\in\{\pm1\}}\sum_{i=1}^{T-t}\left(\left[\mathbf{h}_i(\varepsilon)\right](u)\right)^2\right)\right)\\
&\le\rho\log\left(\sum_{u\in\mathsf{U}}\exp\left(-\frac{1}{\rho}\sum_{s=1}^{t}h_{x,s}(u)\right)\max_{u}\exp\left(\frac{2}{\rho^2}\max_{\varepsilon_1,\dots,\varepsilon_{T-t}\in\{\pm1\}}\sum_{i=1}^{T-t}\left(\left[\mathbf{h}_i(\varepsilon)\right](u)\right)^2\right)\right)\\
&\le\rho\log\left(\sum_{u\in\mathsf{U}}\exp\left(-\frac{1}{\rho}\sum_{s=1}^{t}h_{x,s}(u)\right)\right)+\frac{2}{\rho}\sup_{\mathbf{h}}\max_{u\in\mathsf{U}}\max_{\varepsilon_1,\dots,\varepsilon_{T-t}\in\{\pm1\}}\sum_{i=1}^{T-t}\left(\left[\mathbf{h}_i(\varepsilon)\right](u)\right)^2,
\end{aligned}$$

where the first inequality is due to Hoeffding's lemma (see, e.g., Lemma A.1 in [7]) applied to the expectation w.r.t. ε. The last term, representing the worst-case future, is upper bounded by $\frac{2}{\rho}(T-t)L(\mathsf{X},\mathsf{U},\mathscr{F})^2$. We thus obtain our exponential weight relaxation from (12).

Next we prove that the relaxation (12) is admissible and leads to the recursive algorithm (13). To keep the notation simple, we drop the subscript x in the following. In particular, we use h_t for $h_{x,t}$, $\widehat{W}_t$ for $\widehat{W}_{x,t}$, ν_t for $P_t(\cdot|x)$, etc. The admissibility condition to be proved is

$$\sup_{h_t\in\mathscr{H}_x}\left\{\mathbb{E}_{U\sim\nu_t}\left[h_t(U)\right]+\widehat{W}_t(h^t)\right\}\le\widehat{W}_{t-1}(h^{t-1}).$$

Note that

$$\left\langle\nu_t,\exp\left(-\frac{1}{\rho}h_t\right)\right\rangle=\sum_{u\in\mathsf{U}}\frac{\nu_1(u)\exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-1}h_s(u)\right)}{\left\langle\nu_1,\exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-1}h_s\right)\right\rangle}\exp\left(-\frac{1}{\rho}h_t(u)\right)=\frac{\left\langle\nu_1,\exp\left(-\frac{1}{\rho}\sum_{s=1}^{t}h_s\right)\right\rangle}{\left\langle\nu_1,\exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-1}h_s\right)\right\rangle}.$$

We have

$$\begin{aligned}
&\rho \log\left(\sum_{u\in\mathsf{U}} \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t} h_s(u)\right)\right)\\
&= \rho \log\left\langle \nu_1, \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t} h_s\right)\right\rangle + \rho\log|\mathsf{U}|\\
&= \rho \log\left\langle \nu_t, \exp\left(-\frac{1}{\rho} h_t\right)\right\rangle + \rho \log\left\langle \nu_1, \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-1} h_s\right)\right\rangle + \rho\log|\mathsf{U}|\\
&\le -\mathbb{E}_{U\sim\nu_t} h_t(U) + \frac{L(\mathsf{X},\mathsf{U},\mathscr{F})^2}{2\rho} + \rho \log\left(\sum_{u\in\mathsf{U}} \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-1} h_s(u)\right)\right),
\end{aligned}$$

where the first equality is due to the fact that ν_1 is the uniform distribution on U, while the inequality is due to Hoeffding's lemma. Plugging the resulting bound into the admissibility condition, we get

$$\begin{aligned}
&\sup_{h_t\in\mathscr{H}_x} \left\{\mathbb{E}_{U\sim\nu_t}[h_t(U)] + \widehat{W}_{x,t}(h^t)\right\}\\
&\le \rho \log\left(\sum_{u\in\mathsf{U}} \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-1} h_s(u)\right)\right) + 2\frac{1}{\rho}(T-t+1)L(\mathsf{X},\mathsf{U},\mathscr{F})^2\\
&= \widehat{W}_{x,t-1}(h^{t-1}).
\end{aligned}$$

Thus, the recursive algorithm (13) is admissible for the relaxation (12).

F Proof of Theorem 2

Again, we drop the subscript x and write ν_t for $P_t(\cdot|x)$, etc. We have

$$\mathbb{E}_x^{\widehat{\Upsilon},\boldsymbol{f}}\left[\sum_{t=1}^{T} f_t(X_t,U_t) - \Psi(\boldsymbol{f})\right] \le \sup_{P\in\mathscr{M}(\mathsf{U}|\mathsf{X})} \sum_x \pi_P(x)\widehat{W}_{x,0} + C_{\mathscr{F}}\sum_{t=1}^{T}\|\mu_t^{\widehat{\Upsilon},\boldsymbol{f}} - \pi_t^{\widehat{\Upsilon},\boldsymbol{f}}\|_1. \tag{17}$$

From the relaxation (12), it is easy to see $\widehat{W}_{x,0} \le 2L\sqrt{2T\log|\mathsf{U}|}$ for all states x (in fact, the bound is met with equality with the optimal choice of $\rho = \sqrt{\frac{2TL^2}{\log|\mathsf{U}|}}$). Since we have bounded the first term, now we focus on bounding the second term of the regret bound.

The relative entropy between ν_t and ν_{t-1} is given by

$$\begin{aligned}
D(\nu_t\|\nu_{t-1}) &= \left\langle \nu_t, \log\frac{\exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-1} h_s\right)}{\exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-2} h_s\right)}\right\rangle + \log\frac{\left\langle \nu_1, \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-2} h_s\right)\right\rangle}{\left\langle \nu_1, \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-1} h_s\right)\right\rangle}\\
&= -\frac{1}{\rho}\langle \nu_t, h_{t-1}\rangle + \log\frac{\left\langle \nu_1, \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-2} h_s\right)\right\rangle}{\left\langle \nu_1, \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-1} h_s\right)\right\rangle},
\end{aligned} \tag{18}$$

where

$$\begin{aligned}
\frac{\left\langle \nu_1, \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-2} h_s\right)\right\rangle}{\left\langle \nu_1, \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-1} h_s\right)\right\rangle} &= \frac{\sum_{u\in\mathsf{U}} \nu_1(u)\exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-1} h_s(u)\right)\exp\left(\frac{1}{\rho}h_{t-1}(u)\right)}{\left\langle \nu_1, \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-1} h_s\right)\right\rangle}\\
&= \left\langle \nu_t, \exp\left(\frac{1}{\rho}h_{t-1}\right)\right\rangle.
\end{aligned}$$

Using Hoeffding's lemma, we can write

$$\log \frac{\left\langle \nu_1, \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-2} h_s\right)\right\rangle}{\left\langle \nu_1, \exp\left(-\frac{1}{\rho}\sum_{s=1}^{t-1} h_s\right)\right\rangle} \leq \frac{1}{\rho}\langle \nu_t, h_{t-1}\rangle + \frac{L^2}{2\rho^2}$$

Substituting this bound into (18), we see that the terms involving the expectation of h_{t-1} w.r.t. ν_t cancel, and we are left with

$$D(\nu_t \| \nu_{t-1}) \leq \frac{L^2}{2\rho^2}.$$

Plugging in the optimal value of ρ and using Pinsker's inequality [8], we find

$$\|\nu_t - \nu_{t-1}\|_1 \leq \sqrt{\frac{\log|\mathsf{U}|}{2T}}.$$

So far, we have been working with a fixed state $x \in \mathsf{X}$, so we had $\nu_t = P_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}(\cdot|x)$, where $\widehat{\boldsymbol{\gamma}}$ is the agent's behavioral strategy induced by the relaxation (12). Since x was arbitrary, we get the uniform bound

$$\max_{x\in\mathsf{X}} \left\| P_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}(\cdot|x) - P_{t-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}(\cdot|x)\right\|_1 \leq \sqrt{\frac{\log|\mathsf{U}|}{2T}}. \tag{19}$$

Armed with this estimate, we now bound the total variation distance between the actual state distribution at time t and the unique invariant distribution of $K_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}$. For any time $k \leq t$, we have

$$\begin{aligned}
\left\|\mu_k^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} - \pi_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\right\|_1 &= \left\|\mu_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} K_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} - \mu_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} K_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} + \mu_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} K_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} - \pi_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\right\|_1 \\
&\overset{\text{(a)}}{\leq} \left\|\mu_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} K_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} - \pi_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\right\|_1 + \left\|\mu_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} K_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} - \mu_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} K_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\right\|_1 \\
&\overset{\text{(b)}}{=} \left\|\mu_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} K_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} - \pi_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} K_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\right\|_1 + \left\|\mu_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} K_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} - \mu_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} K_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\right\|_1 \\
&\overset{\text{(c)}}{\leq} e^{-1/\tau}\left\|\mu_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} - \pi_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\right\|_1 + \max_{x\in\mathsf{X}}\left\|P_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}(\cdot|x) - P_t^{\widehat{\boldsymbol{\gamma}}}(\cdot|x)\right\|_1 \\
&\overset{\text{(d)}}{\leq} e^{-1/\tau}\left\|\mu_{k-1}^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}} - \pi_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}\right\|_1 + \sum_{i=k-1}^{t-1}\sqrt{\frac{\log|\mathsf{U}|}{2T}},
\end{aligned} \tag{20}$$

where (a) is by triangle inequality; (b) is by invariance of $\pi_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}$ w.r.t. $K_t^{\widehat{\boldsymbol{\gamma}},\boldsymbol{f}}$; (c) is by the uniform mixing bound (7); and (d) follows from repeatedly using (19) together with triangle inequality and the easily proved fact that, for any state distribution $\mu \in \mathscr{P}(\mathsf{X})$ and any two Markov kernels $P, P' \in \mathscr{M}(\mathsf{U}|\mathsf{X})$,

$$\left\|\mu K(\cdot|P) - \mu K(\cdot|P')\right\|_1 \leq \max_{x\in\mathsf{X}}\left\|P(\cdot|x) - P'(\cdot|x)\right\|_1.$$

Letting now the initial state distribution be μ_1, we can apply the bound (20) recursively to obtain

$$\begin{aligned}
\left\| \mu_t^{\widehat{\boldsymbol{Y}},\boldsymbol{f}} - \pi_t^{\widehat{\boldsymbol{Y}},\boldsymbol{f}} \right\|_1 &\leq e^{-t/\tau} \left\| \mu_1 - \pi_t^{\widehat{\boldsymbol{Y}},\boldsymbol{f}} \right\|_1 + \sum_{k=2}^{t} e^{-\frac{t-k}{\tau}} \sum_{i=k-1}^{t-1} \sqrt{\frac{\log|\mathsf{U}|}{2T}} \\
&\leq 2e^{-t/\tau} + \sum_{k=2}^{t} e^{-\frac{t-k}{\tau}} (t-k) \sqrt{\frac{\log|\mathsf{U}|}{2T}} \\
&\leq 2e^{-t/\tau} + \sqrt{\frac{\log|\mathsf{U}|}{2T}} \sum_{k=1}^{\infty} k e^{-\frac{k}{\tau}} \\
&\leq 2e^{-t/\tau} + \sqrt{\frac{\log|\mathsf{U}|}{2T}} \int_0^{\infty} k e^{-\frac{k}{\tau}} dk \\
&\leq 2e^{-t/\tau} + \tau^2 \sqrt{\frac{\log|\mathsf{U}|}{2T}}.
\end{aligned}$$

So, the second term on the right-hand side of (17) can bounded by

$$C_{\mathscr{F}} \sum_{t=1}^{T} \| \mu_t^{\widehat{\boldsymbol{Y}},\boldsymbol{f}} - \pi_t^{\widehat{\boldsymbol{Y}},\boldsymbol{f}} \|_1 \leq C_{\mathscr{F}} \tau^2 \sqrt{\frac{\log|\mathsf{U}|T}{2}} + 2\tau C_{\mathscr{F}},$$

which completes the proof.

G Proof of Proposition 4

First we show that the relaxation (14) arises as an upper bound on the conditional sequential Rademacher complexity. Once again, we omit the subscript x from $h_{x,t}$ etc. to keep the notation light. Following the same steps as in Appendix E, we have, for any $\rho > 0$,

$$\begin{aligned}
&\mathbb{E}_{\varepsilon} \left[\max_{u \in \mathsf{U}} \left\{ 2 \sum_{j=m+1}^{M} \varepsilon_j \sum_{t \in \mathscr{T}_j} [\mathbf{h}_t(\varepsilon)](u) - \sum_{i=1}^{m} \sum_{s \in \mathscr{T}_i} h_s(u) \right\} \right] \\
&\leq \rho \log \left(\mathbb{E}_{\varepsilon} \left[\max_{u \in \mathsf{U}} \exp \left(\frac{2}{\rho} \sum_{j=m+1}^{M} \varepsilon_j \sum_{t \in \mathscr{T}_j} [\mathbf{h}_t(\varepsilon)](u) - \frac{1}{\rho} \sum_{i=1}^{m} \sum_{s \in \mathscr{T}_i} h_s(u) \right) \right] \right) \\
&\leq \rho \log \left(\mathbb{E}_{\varepsilon} \left[\sum_{u \in \mathsf{U}} \exp \left(\frac{2}{\rho} \sum_{j=m+1}^{M} \varepsilon_j \sum_{t \in \mathscr{T}_j} [\mathbf{h}_t(\varepsilon)](u) - \frac{1}{\rho} \sum_{i=1}^{m} \sum_{s \in \mathscr{T}_i} h_s(u) \right) \right] \right).
\end{aligned}$$

In the same vein,

$$
\begin{aligned}
&\rho\log\left(\sum_{u\in\mathsf{U}}\exp\left(-\frac{1}{\rho}\sum_{i=1}^{m}\sum_{s\in\mathscr{T}_i}h_s(u)\right)\mathbb{E}_{\varepsilon}\left[\prod_{j=m+1}^{M}\exp\left(\frac{2}{\rho}\varepsilon_j\sum_{t\in\mathscr{T}_j}[\mathbf{h}_t(\varepsilon)](u)\right)\right]\right)\\
&\le\rho\log\left(\sum_{u\in\mathsf{U}}\exp\left(-\frac{1}{\rho}\sum_{i=1}^{m}\sum_{s\in\mathscr{T}_i}h_s(u)\right)\times\exp\left(\frac{2}{\rho^2}\max_{\varepsilon_{m+1},\ldots,\varepsilon_M\in\{\pm1\}}\sum_{j=m+1}^{M}\left(\tau_j[\mathbf{h}(\varepsilon)](u)\right)^2\right)\right)\\
&\le\rho\log\left(\sum_{u\in\mathsf{U}}\exp\left(-\frac{1}{\rho}\sum_{i=1}^{m}\sum_{s\in\mathscr{T}_i}h_s(u)\right)\max_{u}\exp\left(\frac{2}{\rho^2}\max_{\varepsilon_{m+1},\ldots,\varepsilon_M\in\{\pm1\}}\sum_{j=m+1}^{M}\left(\tau_j[\mathbf{h}(\varepsilon)](u)\right)^2\right)\right)\\
&\le\rho\log\left(\sum_{u\in\mathsf{U}}\exp\left(-\frac{1}{\rho}\sum_{i=1}^{m}\sum_{s\in\mathscr{T}_i}h_s(u)\right)\right)+\frac{2}{\rho}\sup_{\mathbf{h}}\max_{u\in\mathsf{U}}\max_{\varepsilon_{m+1},\ldots,\varepsilon_M\in\{\pm1\}}\sum_{j=m+1}^{M}\left(\tau_j[\mathbf{h}(\varepsilon)](u)\right)^2\\
&\le\rho\log\left(\sum_{u\in\mathsf{U}}\exp\left(-\frac{1}{\rho}\sum_{i=1}^{m}\sum_{s\in\mathscr{T}_i}h_s(u)\right)\right)+\frac{2}{\rho}\sum_{j=m+1}^{M}\tau_j^2L(\mathsf{X},\mathsf{U},\mathscr{F})^2,
\end{aligned}
$$

where the first inequality is due to Hoeffding's lemma, while the last inequality is by Assumption 1. We thus derive the relaxation in (14).

Now we prove that this relaxation is admissible, and leads to the lazy algorithm (15) The admissibility condition to be proved is

$$
\sup_{h_m\in\mathscr{H}_x^{\tau_m}}\left\{\mathbb{E}_{U\sim\nu_m}\left[\sum_{s\in\mathscr{T}_m}h_s(U)\right]+\widehat{W}_{x,m}(h_{1:m})\right\}\le\widehat{W}_{x,m-1}(h_{1:m-1}),
$$

where $\nu_m=P_m(\cdot|x)$ is the Markov policy used in phase m. We have

$$
\begin{aligned}
&\rho\log\left(\sum_{u\in\mathsf{U}}\exp\left(-\frac{1}{\rho}\sum_{i=1}^{m}\sum_{s\in\mathscr{T}_i}h_s(u)\right)\right)\\
&=\rho\log\left\langle\nu_1,\exp\left(-\frac{1}{\rho}\sum_{i=1}^{m}\sum_{s\in\mathscr{T}_i}h_s\right)\right\rangle+\rho\log|\mathsf{U}|\\
&=\rho\log\left\langle\nu_m,\exp\left(-\frac{1}{\rho}\sum_{s\in\mathscr{T}_m}h_s\right)\right\rangle+\rho\log\left\langle\nu_1,\exp\left(-\frac{1}{\rho}\sum_{i=1}^{m-1}\sum_{s\in\mathscr{T}_i}h_s\right)\right\rangle+\rho\log|\mathsf{U}|\\
&\le-\mathbb{E}_{U\sim\nu_m}\left[\sum_{s\in\mathscr{T}_m}h_s(U)\right]+\frac{\tau_m^2L(\mathsf{X},\mathsf{U},\mathscr{F})^2}{2\rho}+\rho\log\left(\sum_{u\in\mathsf{U}}\exp\left(-\frac{1}{\rho}\sum_{i=1}^{m-1}\sum_{s\in\mathscr{T}_i}h_s(u)\right)\right),
\end{aligned}
$$

Plugging this into the admissibility condition, we have

$$
\begin{aligned}
&\sup_{h_m\in\mathscr{H}_x^{\tau_m}}\left\{\mathbb{E}_{U\sim\nu_m}\left[\sum_{s\in\mathscr{T}_m}h_s(U)\right]+\widehat{W}_{x,m}(h_{1:m})\right\}\\
&\le\rho\log\left(\sum_{u\in\mathsf{U}}\exp\left(-\frac{1}{\rho}\sum_{i=1}^{m-1}\sum_{s\in\mathscr{T}_i}h_s(u)\right)\right)+\frac{2}{\rho}\sum_{j=m+1}^{M}\tau_j^2L(\mathsf{X},\mathsf{U},\mathscr{F})^2+\frac{\tau_m^2L(\mathsf{X},\mathsf{U},\mathscr{F})^2}{2\rho}\\
&\le\rho\log\left(\sum_{u\in\mathsf{U}}\exp\left(-\frac{1}{\rho}\sum_{i=1}^{m-1}\sum_{s\in\mathscr{T}_i}h_s(u)\right)\right)+\frac{2}{\rho}\sum_{j=m}^{M}\tau_j^2L(\mathsf{X},\mathsf{U},\mathscr{F})^2\\
&=\widehat{W}_{x,m-1}(h^{m-1}).
\end{aligned}
$$

So the lazy algorithm (15) is an admissible strategy for the relaxation (14).

H Proof of Theorem 3

The state feedback law $P_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(\cdot|x)$ that the agent applies within phase m is the same for all $t \in \mathcal{T}_m$, and we denote it by $P_m^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(\cdot|x)$. Let $K_m^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}$ denote the Markov matrix that describes the state transition from X_t to X_{t+1} if $t \in \mathcal{T}_m$. Thus, we can write

$$K_m^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(y|x) = \sum_u K(y|x,u) P_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(u|x), \qquad \forall\, x, y \in \mathsf{X}.$$

First, we show that

$$\mathbb{E}_x^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\left[\sum_{t=1}^T f_t(X_t,U_t) - \Psi(\boldsymbol{f})\right] \le \sup_{P\in\mathscr{M}(\mathsf{U}|\mathsf{X})} \sum_x \pi_P(x)\widehat{W}_{x,0} + C_{\mathscr{F}} \sum_{m=1}^M \sum_{t\in\mathcal{T}_m} \|\mu_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}} - \pi_m^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\|_1, \tag{21}$$

where $\pi_m^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}$ is the invariant distribution of $K_m^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}$.

To prove (21), we write

$$\begin{aligned}
&\mathbb{E}_x^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\left[\sum_{t=1}^T f_t(X_t,U_t) - \Psi(\boldsymbol{f})\right] \\
&\le \sup_{P\in\mathscr{M}(\mathsf{U}|\mathsf{X})} \sum_{t=1}^T \left[\langle \pi_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}} \otimes P_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}, f_t\rangle - \langle \pi_P \otimes P, f_t\rangle\right] + \sum_{t=1}^T \|f_t\|_\infty \|\mu_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}} - \pi_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\|_1 \\
&\le \sup_{P\in\mathscr{M}(\mathsf{U}|\mathsf{X})} \sum_{m=1}^M \sum_{t\in\mathcal{T}_m} \left[\langle \pi_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}} \otimes P_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}, f_t\rangle - \langle \pi_P \otimes P, f_t\rangle\right] + C_{\mathscr{F}} \sum_{m=1}^M \sum_{t\in\mathcal{T}_m} \|\mu_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}} - \pi_m^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\|_1 \\
&\le \sup_{P\in\mathscr{M}(\mathsf{U}|\mathsf{X})} \sum_x \pi_P(x) \sum_{m=1}^M \sum_{t\in\mathcal{T}_m} \left(\sum_u P_m^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u) - P(u|x)\widehat{Q}_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u)\right) + C_{\mathscr{F}} \sum_{m=1}^M \sum_{t\in\mathcal{T}_m} \|\mu_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}} - \pi_m^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\|_1,
\end{aligned}$$

where the last inequality is by Lemma 1. By writing out the first term in the right hand side, we get

$$\begin{aligned}
&\sup_{P\in\mathscr{M}(\mathsf{U}|\mathsf{X})} \sum_x \pi_P(x)\left[\sum_{m=1}^{M-1} \sum_{t\in\mathcal{T}_m} \left(\sum_u P_m^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u)\right) + \sum_u \nu_M(u|x) \sum_{t\in\mathcal{T}_M} \widehat{Q}_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u)\right. \\
&\quad \left. - \sum_{t=1}^T P(u|x)\widehat{Q}_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u)\right] + C_{\mathscr{F}} \sum_{m=1}^M \sum_{t\in\mathcal{T}_m} \|\mu_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}} - \pi_m^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\|_1 \\
&\le \sup_{P\in\mathscr{M}(\mathsf{U}|\mathsf{X})} \sum_x \pi_P(x)\left[\sum_{m=1}^{M-1} \sum_{t\in\mathcal{T}_m} \left(\sum_u P_m^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u)\right) + \sum_u \nu_M(u|x) \sum_{t\in\mathcal{T}_M} \widehat{Q}_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u) + \widehat{W}_{x,M}(h^M)\right] \\
&\quad + C_{\mathscr{F}} \sum_{m=1}^M \sum_{t\in\mathcal{T}_m} \|\mu_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}} - \pi_m^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\|_1 \\
&\le \sup_{P\in\mathscr{M}(\mathsf{U}|\mathsf{X})} \sum_x \pi_P(x)\left[\sum_{m=1}^{M-1} \sum_{t\in\mathcal{T}_m} \left(\sum_u P_m^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(u|x)\widehat{Q}_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}(x,u)\right) + \widehat{W}_{x,M-1}(h^{M-1})\right] \\
&\qquad + \sum_{t=1}^T \|f_t\|_\infty \left\|\mu_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}} - \pi_t^{\widehat{\boldsymbol{\Upsilon}},\boldsymbol{f}}\right\|_1.
\end{aligned}$$

The last inequality is due to the fact that $\widehat{\boldsymbol{\gamma}}$ is the behavioral strategy associated to the admissible relaxation $\{\widehat{W}_{x,m}\}_{m=1}^M$. Continuing this induction backwards, we arrive at (21).

Next, we bound the two terms on the right-hand side of (21). From the form of the relaxation (14), it is easy to see $\widehat{W}_{x,0} \le 2L\sqrt{2\log|\mathsf{U}|\sum_{i=1}^M \tau_i^2}$ for all states x; in fact, this bound is attained with equality if we

use the optimal choice $\rho = \sqrt{\frac{2\sum_{i=1}^{M}\tau_i^2 L^2}{\log|\mathsf{U}|}}$. Since we have bounded the first term, now we focus on bounding the second term of (21).

From the contraction inequality (7) it follows that, for every $k \in \{0,1,\dots,\tau_m - 1\}$, we have

$$\begin{aligned}\left\|\mu^{\hat{\boldsymbol{\Upsilon}},\boldsymbol{f}}_{\tau_{1:m-1}+k+1} - \pi^{\hat{\boldsymbol{\Upsilon}},\boldsymbol{f}}_{m}\right\|_1 &= \left\|\mu^{\hat{\boldsymbol{\Upsilon}},\boldsymbol{f}}_{\tau_{1:m-1}+1}(K^{\hat{\boldsymbol{\Upsilon}},\boldsymbol{f}}_{m})^k - \pi^{\hat{\boldsymbol{\Upsilon}},\boldsymbol{f}}_{m}(K^{\hat{\boldsymbol{\Upsilon}},\boldsymbol{f}}_{m})^k\right\|_1 \\ &\le e^{-k/\tau}\left\|\mu^{\hat{\boldsymbol{\Upsilon}},\boldsymbol{f}}_{\tau_{1:m-1}+1} - \pi^{\hat{\boldsymbol{\Upsilon}},\boldsymbol{f}}_{m}\right\|_1 \\ &\le 2e^{-k/\tau}.\end{aligned}$$

Hence,

$$\sum_{t\in\mathscr{T}_m}\left\|\mu^{\hat{\boldsymbol{\Upsilon}},\boldsymbol{f}}_{t} - \pi^{\hat{\boldsymbol{\Upsilon}},\boldsymbol{f}}_{m}\right\|_1 \le 2\sum_{k=0}^{\tau_m-1} e^{-k/\tau} \le \frac{2}{1-e^{-1/\tau}}.$$

Plugging it in (21), we have shown that

$$\mathbb{E}^{\hat{\boldsymbol{\Upsilon}},\boldsymbol{f}}_{x}\left[\sum_{t=1}^{T} f_t(X_t,U_t) - \Psi(\boldsymbol{f})\right] \le 2L\sqrt{2\log|\mathsf{U}|\sum_{i=1}^{M}\tau_i^2} + \frac{2C_{\mathscr{F}}M}{1-e^{-1/\tau}}.$$

I Proof of Corollary 1

Let us inspect the right-hand side of (16). We see that both $\sqrt{\sum_{j=1}^{M}\tau_j^2}$ and M have to be sublinear in T. Since $\sum_{i=1}^{M}\tau_i = T$ and $\sqrt{\sum_{i=1}^{M}\tau_i^2} < \sqrt{(\sum_{i=1}^{M}\tau_i)^2}$, at least the first of these terms can be made sublinear, e.g., by having $\tau_j = 1$ for all j. Of course, this means that $M = T$, so we need longer phases. For example, if we follow [33] and let $\tau_m = \lceil m^{1/3-\varepsilon}\rceil$ for some $\varepsilon \in (0,1/3)$, then a straightforward if tedious algebraic calculation shows that $M = O(T^{3/4})$ and $\sqrt{\sum_{j=1}^{M}\tau_j^2} = O(T^{5/8})$, which yields the regret of $O(T^{3/4})$.

However, if T is known in advance, then we can do better: ignoring the rounding issues, for any constants $A_1, A_2 > 0$,

$$\min_{1\le M\le T}\min\left\{A_1\sqrt{\sum_{j=1}^{M}\tau_j^2} + A_2 M : \sum_{j=1}^{M}\tau_j = T\right\} = O(T^{2/3}), \tag{22}$$

To see this, let us first fix M and optimize the choice of the τ_j's:

$$\min\sum_{j=1}^{M}\tau_j^2 \quad \text{subject to } \sum_{j=1}^{M}\tau_j = T.$$

By the Cauchy–Schwarz inequality, we have

$$\sum_{j=1}^{M}\tau_j \le \sqrt{M\sum_{j=1}^{M}\tau_j^2}.$$

Thus, $\sum_{j=1}^{M}\tau_j^2$ achieves its minimum when the above bound is met with equality. This will happen only if all the τ_j's are equal, i.e., $\tau_j = \frac{T}{M}$ for every j (for simplicity, we assume that M divides T — otherwise,

the remainder term will be strictly smaller than M, and the bound in (22) will still hold, but with a larger multiplicative constant). Therefore,

$$\min_{1\le M\le T} \min\left\{ A_1\sqrt{\sum_{j=1}^{M}\tau_j^2} + A_2 M : \sum_{j=1}^{M}\tau_j = T \right\} = \min_{1\le M\le T}\left(\frac{A_1 T}{\sqrt{M}} + A_2 M\right) = O(T^{2/3}),$$

where the minimum on the right-hand side (again, ignoring rounding issues) is achieved by $M = T^{2/3}$ and $\tau_j = T^{1/3}$ for all j. This shows that, for a given horizon T, the optimal choice of phase lengths is $T^{1/3}$, which gives the regret of $O(T^{2/3})$, better than the $O(T^{3/4})$ bound derived by [33].

References

[1] Y. Abbasi-Yadkori, P. L. Bartlett, and C. Szepesvári. Online learning in Markov decision processes with adversarially chosen transition probability distributions. *http://arxiv.org/1303.3055*, 2013.

[2] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus. Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM J. Control Optim.*, 31 (2):282–344, 1993.

[3] R. Arora, O. Dekel, and A. Tewari. Deterministic MDPs with adversarial rewards and bandit feedback. In *Proceedings of the 28th Annual Conference on Uncertainty in Artificial Intelligence*, pages 93–101, AUAI Press, 2012.

[4] T. Başar and G. J. Olsder. *Dynamic Noncooperative Game Theory*. SIAM, Philadelphia, PA, 2nd edition, 1999.

[5] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, Nashua, NH, 3rd edition, 2005.

[6] D. P. Bertsekas and I. B. Rhodes. Sufficiently informative functions and the minimax feedback control of uncertain dynamic systems. *IEEE Trans. Automat. Control*, 18(2):117–124, April 1973.

[7] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge Univ. Press, 2006.

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

[9] E. Even-Dar, S. M. Kakade, and Y. Mansour. Online Markov decision processes. *Math. Oper. Res.*, 34 (3):726–736, 2009.

[10] P. Guan, M. Raginsky, and R. Willett. Online Markov decision processes with Kullback-Leibler control cost. In *Proc. American Control Conference*, 2012.

[11] P. Guan, M. Raginsky, and R. Willett. Online markov decision processes with Kullback-Leibler control cost. submitted to *IEEE Transactions on Automatic Control*, 2012.

[12] P. Guan, M. Raginsky, and R. Willett. Online Markov decision processes with Kullback–Leibler control cost. *IEEE Trans. Automat. Control*, 2013. conditonally accepted as a regular paper.

[13] E. Hall and R. Willett. Dynamical models and tracking regret in online convex programming. *arXiv:1301.1254*, 2013. To appear in *Proc. ICML*.

[14] E. Hall and R. Willett. Online optimization in dynamic environments. arXiv:1307:5944, submitted to *Journal of Machine Learning Research*, 2013.

[15] E. Hall and R. Willett. Online optimization in parametric dynamic environments. In *Proc. Allerton Conference on Communication, Control and Computing*, 2013.

[16] J. Hannan. Approximation to Bayes risk in repeated play. In *Contributions to the Theory of Games*, volume 3, pages 97–139. Princeton Univ. Press, 1957.

[17] O. Hernández-Lerma and J. B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, 1996.

[18] O. Hernández-Lerma and J. B. Lasserre. *Markov Chains and Invariant Probabilities*. Birkhäuser, 2003.

[19] C. Horn and R. Willett. Online anomaly detection with expert system feedback in social networks. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2011.

[20] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Inform. Comput.*, 108:212–261, 1994.

[21] H. McMahan. Planning in the presence of cost functions controlled by an adversary. *The 20th International Conference on Machine Learning*, pages 536–543, 2003.

[22] S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambrdige Univ. Press, 2nd edition, 2009.

[23] G. Neu, A. György, C. Szepesvári, and A. Antos. Online Markov decision processes under bandit feedback. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1804–1812, 2010.

[24] M. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

[25] M. Raginsky, R. Willett, C. Horn, J. Silva, and R. Marcia. Sequential anomaly detection in the presence of noise and limited feedback. *IEEE Transactions on Information Theory*, 58(8):5544–5562, 2012. doi:10.1109/TIT.2012.2201375.

[26] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: random averages, combinatorial parameters, and learnability. *Adv. Neural Inform. Processing Systems*, 2010.

[27] A. Rakhlin, O. Shamir, and K. Sridharan. Relax and randomize: from value to algorithms. *Adv. Neural Inform. Processing Systems*, 2012.

[28] H. Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probability 1950*, pages 131–148. University of California Press, Berkeley, CA, 1951.

[29] E. Seneta. *Nonnegative Matrices and Markov Chains*. Springer, 2006.

[30] S. Shalev-Shwartz and Y. Singer. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69:115–142, 2007.

[31] S. Sorin. *A first course on zero-sum repeated games*. Springer, 2002.

[32] J. N. Tsitisklis. NP-hardness of checking the unichain condition in average cost MDPs. *Oper. Res. Lett.*, 35(3):319–323, 2007.

[33] J. Y. Yu, S. Mannor, and N. Shimkin. Markov decision processes with arbitrary reward processes. *Math. Oper. Res.*, 34(3):737–757, 2009.