

UNCLASSIFIED

AD **400 745**

*Reproduced
by the*

ARMED SERVICES TECHNICAL INFORMATION AGENCY
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA



UNCLASSIFIED

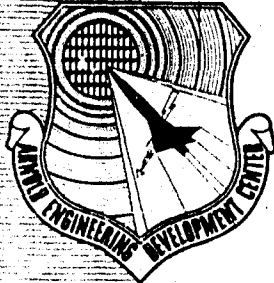
NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

AEDC-TDR-63-82

400 745

63-3-1
10

829900

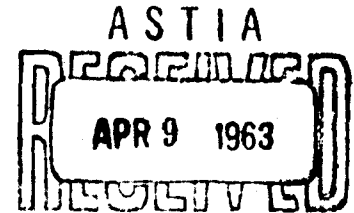


**PROBLEMS
UNDERLYING THE NUMERICAL INTEGRATION OF
THE CHEMICAL AND VIBRATIONAL RATE EQUATIONS
IN A NEAR-EQUILIBRIUM FLOW**

By

George Emanuel

Department of Aeronautics and Astronautics
Stanford University
Stanford, California



TECHNICAL DOCUMENTARY REPORT NO. AEDC-TDR-63-82

March 1963

AFSC Program Area 750A , Project 8952 , Task 895210

(Prepared under Contract No. AF 40(600)-930 by
Stanford University, Stanford, California.)

**ARNOLD ENGINEERING DEVELOPMENT CENTER
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE**

400 745
AD NO. 1
ASTIA FILE COPY

47

NOTICES

Qualified requesters may obtain copies of this report from ASTIA. Orders will be expedited if placed through the librarian or other staff member designated to request and receive documents from ASTIA.

When Government drawings, specifications or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

PROBLEMS
UNDERLYING THE NUMERICAL INTEGRATION OF
THE CHEMICAL AND VIBRATIONAL RATE EQUATIONS
IN A NEAR-EQUILIBRIUM FLOW

By

George Emanuel
Department of Aeronautics and Astronautics
Stanford University
Stanford, California

(The reproducibles used in the reproduction
of this report were supplied by the authors.)

March 1963

ABSTRACT

The work described is a continuation of that reported previously in AEDC-TN-61-85 and AEDC-TDR-62-131. In particular, the present report contains the basis for the method described in AEDC-TDR-62-131 for the precise numerical calculation of one-dimensional nonequilibrium flow of a complex gas mixture through a nozzle. ~~The contributions of the two above reports are reviewed briefly in the introduction.~~

X is discussed.

A transformation is shown to exist when the flow is close to equilibrium that transforms the chemical rate equations into a simpler system of uncoupled rate equations. Each of the transformed chemical rate equations is similar in form to a vibrational rate equation. The coefficients appearing in the transformed equations are given by solutions of an eigenvalue problem. These coefficients are shown to be real and positive. An example is given that illustrates the transformation for a simple gas model and shows that the transformation behaves correctly in the frozen and equilibrium limits.

*cut it
100%*

Because of their simple mathematical structure, the vibrational and transformed chemical rate equations can be subjected to numerical analysis. Thus, a detailed analysis of both Runge-Kutta and predictor-corrector integration procedures is performed in which truncation error, stability, and computation speed are examined. The analysis of Runge-Kutta procedures demonstrates that the greater the order of the procedure, the more stable it is. High-order Runge-Kutta procedures, unfortunately, also have reduced computation speed. On an overall basis, the commonly employed fourth-order procedure still appears to be the most suitable Runge-Kutta procedure for the integration of the chemical and vibrational rate equations. The basis for the improved technique for controlling the Runge-Kutta integration step size, given in AEDC-TDR-62-131, is also established.

*cut it
100%*

When Adams predictor-corrector procedures are similarly analyzed, the greater the order of the procedure the less stable it is. Furthermore, the overall Adams procedure is considerably less stable than is the Adams corrector formula. The common assumption that stability is determined primarily by the corrector is thus not valid. A fourth-order predictor-corrector procedure is also given that is stable for a larger integration step size than the fourth-order Adams procedure.

✓

PUBLICATION REVIEW

This report has been reviewed and publication is approved.

Donald R. Eastman, Jr.
Donald R. Eastman, Jr.
DCS/Research

Jean A. Jack
Jean A. Jack
Colonel, USAF
DCS/Test

CONTENTS

	<u>Page</u>
ABSTRACT	iii
SYMBOLS	vi
1.0 INTRODUCTION	1
2.0 PRELIMINARY REMARKS	2
3.0 CANONICAL FORM FOR THE CHEMICAL RATE EQUATIONS	6
4.0 GENERAL PROPERTIES OF STIFF EQUATIONS	15
5.0 RUNGE-KUTTA ANALYSIS	17
6.0 PREDICTOR-CORRECTOR ANALYSIS	26
7.0 CONCLUDING REMARKS	38
REFERENCES	41
APPENDIXES	
I PROOF THAT ALL OF THE EIGENVALUES κ_1 ARE REAL AND POSITIVE	43
II EXAMPLE ILLUSTRATING THE TRANSFORMATION TO CANONICAL FORM OF THE CHEMICAL RATE EQUATIONS	62

SYMBOLS

PRIMARY SYMBOLS

A_i	constants in the expansion for $\bar{K}_m^{(1)}$ (see equation (I-18))
A_i, B_i, a_i, b_i	constants in the predictor-corrector integration procedure
A_{ij}	constants in equation (27)
a_{ij}	constants in equation (4)
B	stability parameter for Runge-Kutta integration procedures
B_{ik}	coefficients in the transformation (19a)
$B^{(pr)}$	stability parameter for predictor formula
$B^{(co)}$	stability parameter for corrector formula
$B^{(pc)}$	stability parameter for overall predictor-corrector procedure
$C^{(pr)}$	characteristic polynomial for predictor formula
$C^{(co)}$	characteristic polynomial for corrector formula
$C^{(pc)}$	characteristic polynomial for overall predictor-corrector procedure
E_p	truncation error of a p^{th} - order integration procedure
e_{v_i}	vibrational energy per mole of species i
$e_{v_{i,e}}$	local equilibrium vibrational energy corresponding to static temperature T
$f(x,y)$	right-hand side of differential equation (31)
$G(x)$	arbitrary function in equation (2a)
H	defined by equation (I-16)

h	integration step size
$\bar{K}_m^{(0)}, \bar{K}_m^{(1)}$	defined by equations (I-13) and (I-15a)
K_{cr}	equilibrium constant for reaction r
k', k	see predictor-corrector formulas (48) and (49)
k_{br}	backward rate constant for reaction r
k_{fr}	forward rate constant for reaction r
L_r	see equation (6b)
N_1	total number of reacting chemical species
N_2	number of chemical reactions
N_3	number of chemical species minus total number of components and polyatomic inert species
N_4	number of chemical species plus number of catalytic bodies that consist of a linear combination of other species
n_i	mole-mass ratio of species i ; that is, number of moles of species i per unit mass of fluid
$n_{i,e}$	local equilibrium value of n_i based on the local static density and temperature
P	defined by equation (I-8a)
p	order of the truncation procedure
q_1	generalized mole-mass ratio (see equation (19a))
T	static temperature
u	flow speed
x	distance along nozzle axis
x, y	variables in equation (2a)

$z(x_0+h)$	estimate for $y(x_0+h)$ given by predictor formula (48)
$z'(x_0+h)$	estimate for $y'(x_0+h)$ used in corrector formula (49)
$\alpha_i, \beta_{ij}, \gamma_i$	constants in the Runge-Kutta integration procedure
α_r, β_r	see equation (7)
α_{ri}	forward stoichiometric coefficient for species i in reaction r
β_{ri}	backward stoichiometric coefficient for species i in reaction r
δ	see equation (41a)
δ_{kl}	Kronecker delta
ϵ	small constant (see equation (2b))
θ_r	defined by equation (6c)
κ	unknown in the characteristic polynomial given by equation (24b)
κ_i	roots of the characteristic polynomial given by equation (24b)
$\bar{\kappa}_{il}$	defined by equation (16)
κ_{il}	defined by equation (18)
v_{ri}	stoichiometric coefficients (see equation (6a))
ρ	mass density of fluid
ρ	unknown in the stability characteristic polynomial
ρ_i	roots of the stability characteristic polynomial
τ_i	vibrational relaxation time for species i

SUPERSCRIPTS AND SUBSCRIPTS

$()^*$	equilibrium reference value
$()'$	perturbation value
$()''$	denotes differentiation with respect to x
e	value at local equilibrium condition
i	value for species i
i, j, j', k, l, m, s	indices
in	initial value
o	value at the beginning of an integration step
r	value for reaction r
v	vibrational quantity

1.0 INTRODUCTION

This is the third in a series of reports concerning the calculation of the reacting flow of a complex gas in the nozzle of a hypersonic wind tunnel at high stagnation enthalpy. The first of these reports (Vincenti (1961), hereafter referred to as Part I) describes a five-species model for air, governed by eight chemical kinetic reactions. A method is described, in particular, for the numerical calculation of the one-dimensional nonequilibrium flow of this gas through a hypersonic nozzle. A specific calculation carried out on an IBM 709 computer, however, revealed that the method required too much computer time to be practical for engineering purposes.

The second report (Emanuel and Vincenti (1962), hereafter referred to as Part II) describes an improved method for calculating nonequilibrium nozzle flows. This method is relatively simple, numerically accurate, and does not require an excessive amount of machine time. It can be summarized as follows:

- (a) An equilibrium solution for the nozzle flow is first obtained. To facilitate this, a new, simple procedure for computing an inexact but numerically accurate equilibrium solution was given.
- (b) Equilibrium initial conditions are used to start the nonequilibrium calculation. This simple method for starting the nonequilibrium calculation replaces the complicated perturbation schemes frequently employed.
- (c) Forward integration of the nonequilibrium equations then proceeds from the equilibrium initial point. This point is chosen somewhat upstream of the nozzle location at which the chemistry first departs appreciably from the equilibrium-flow solution.

With this method, the numerical solution is insensitive to the location of the equilibrium initial point, provided that it is chosen as described in (c) above.

The above method requires that the nonequilibrium calculation includes a portion of the nozzle in which the flow is close to equilibrium. Although the extent of this region can be minimized, a very small step size is nevertheless necessary for the integration therein. If too large a step size is used in this near-equilibrium region, then the integration procedure becomes unstable and the numerical solution diverges from the correct solution. Thus, to avoid an excessive amount of computer time, the integration procedure must allow the step size to vary in a manner that maximizes its size but still keeps the procedure stable. Part II gave, for a Runge-Kutta integration procedure, a simple method of controlling step size that approximately satisfies the foregoing criterion.

The present report contains the analytical basis for the numerical method described in Part II and outlined above. Therefore, some acquaintance by the reader with the system of equations of Part II is desirable. We shall first show that a system of chemical and vibrational rate equations is equivalent, for purposes of numerical analysis, to a simple, so-called "stiff" equation. The remainder of the report is then concerned with the numerical analysis of this equation.

The author wishes to express his sincere gratitude to Professor Walter G. Vincenti for his guidance in this work. Thanks are also due to Mrs. Lita Emanuel for editorial assistance.

2.0 PRELIMINARY REMARKS

As Curtiss and Hirschfelder (1952) have pointed out, a type of differential equation occurs in many physical problems that is extremely difficult to solve numerically. They designate this type of equation as "stiff" since a typical example is the equation that describes the motion of a simple mechanical system with a stiff spring. We shall here define a stiff equation as any ordinary or partial differential equation in which the highest-order derivative is multiplied by a small parameter. Problems governed by this type of equation are also referred to as singular perturbation problems. This designation, however, is correct only in the limit as the small parameter goes to zero. Since our interest here is in the numerical analysis of this type of equation, and not in the limiting process, the terminology of Curtiss and Hirschfelder will be used.

In a near-equilibrium flow, each chemical and vibration rate equation involves a small parameter that multiplies the derivative. Thus, in a near-equilibrium flow, each rate equation is stiff. For example, $u\tau_1$ is the small parameter in the vibrational rate equations

$$\frac{de_{v_1}}{dx} = - \frac{e_{v_1} - e_{v_1,e}}{u\tau_1}, \quad (1)$$

which applies to steady one-dimensional flow. Here e_{v_1} and $e_{v_1,e}$ are, respectively, the actual vibrational energy and the local equilibrium vibrational energy per mole of species i . The flow speed is denoted by u , vibrational relaxation time of species i by τ_1 , and x is the distance along the nozzle axis. (For additional details see either Parts I or II.)

Our object is the analysis of stiff rate equations, both chemical and vibrational. Because of their complexity, however, the chemical rate equations will not be studied directly. Instead, they will first

be transformed into a system of equations of the same form as the vibrational rate equations (1). It will therefore not be necessary to distinguish between the two types of equations in the analysis.

To derive the transformation, which is given in detail in Chapter 3.0 and illustrated by an example in Appendix II, the chemical rate equations are first linearized about an equilibrium reference state. As a result, the transformation is valid only for a near-equilibrium flow. In general, the reference state is variable, as in nozzle flow, and the transformation is then only a local one. In other words, the transformation is valid only over a limited region of the flow. A sequence of transformations may thus be required to represent the entire near-equilibrium region. Our purpose in deriving the transformation, however, is to be able to analyze numerically the behavior of the rate equations when the flow is close to equilibrium. Since this analysis is also of a local nature, the limited region of validity of the transformation is of no consequence.

Inasmuch as the difficulties of integration of the nonequilibrium equations arise solely out of the stiffness of the rate equation, certain assumptions may be introduced to simplify the analysis. These assumptions, however, in no way alter or reduce the numerical difficulties from those encountered when the original rate equations are integrated. To better understand these assumptions, we first note that the vibrational rate equations (1) are coupled only through the dependence of $e_{v_i,e}$ and u_{T_i} on the density and temperature of the flow. The same is also true of the transformed chemical rate equations (see equations (20)), which for steady one-dimensional flow have the same structure as the vibrational equations. Thus, there will be no loss of generality if the following is assumed (after the chemical rate equations have been transformed into their simplified form):

- (a) All of the rate equations are uncoupled. This is equivalent to assuming that the density, temperature and flow speed are known functions of the independent variable x , since any coupling in the rate equations occurs through these variables. More specifically, when the equilibrium-flow solution is known, these variables may be approximated by their equilibrium-flow values. This assumption in no way alters the structure of the rate equations, and only slightly modifies the magnitudes of the small parameters. Hence, as stated earlier, no loss of generality ensues from this decoupling assumption.
- (b) In place of a system of uncoupled rate equations, it is sufficient to consider only one such equation. This equation is taken to be that containing the smallest parameter. When this assumption is examined a posteriori, it is readily found to be justified.

An important consequence of the transformation referred to earlier is that the small parameters in the transformed chemical rate equations are all positive. (Appendix I contains the proof for this statement.) This fact, in conjunction with the foregoing assumptions, implies that we need examine only one equation with the same form as the vibrational rate equations (1). Under these assumptions, the stiff equations may be represented by

$$\frac{dy}{dx} = - \frac{y + G(x)}{\epsilon(x)}, \quad (2a)$$

and any numerical analysis that is valid for this equation will be equally valid for the original system of chemical and vibrational rate equations of Parts I and II. In equation (2a), $\epsilon(x)$ is a small positive function, while $G(x)$ is an arbitrary function. Equation (2a) can be further simplified, however, still without any loss of generality insofar as the numerical analysis is concerned. To this end, we now also assume the following:

(c) The functions $\epsilon(x)$ and $G(x)$ are given by

$$\left. \begin{aligned} \epsilon(x) &= \text{small positive constant,} \\ G(x) &= x. \end{aligned} \right\} \quad (2b)$$

These simplifications are justified by first expanding ϵ and G in Taylor series about any point x' . The variables x, y are then transformed so that the numerator of the right-hand side of equation (2a) has (in terms of the new variables) the form $y + x + (\text{constant}) x^2 + \dots$. Finally, only the lowest order terms are retained in the right-hand side of equation (2a), which leads to equations (2b). This last step is consistent with the local nature of the initial transformation of the chemical equations as well as with the subsequent numerical analysis.

Because of their simplicity as compared with the original rate equations, equations (2) are readily analyzed. From this analysis, two important facts become apparent. First, at the outset of the nonequilibrium calculation a small integration step size is necessary. This requirement is a direct consequence of the magnitude of the truncation error when the integration starts from an equilibrium condition. Second, the integration step size can increase appreciably after a short fictitious transient region is completed wherein the derivatives of the variables change from their frozen values to approximately their equilibrium values. After this transient region, the size that the integration step can attain depends on the specific integration procedure. For this reason, the two most commonly used procedures for numerical integration, Runge-Kutta (Chapter 5.0) and predictor-corrector (Chapter 6.0), are analyzed. This analysis, it should be noted, is applicable only to the

integration of the nonequilibrium equations when the flow is close to equilibrium. Once the integration variables have begun to diverge from their local equilibrium values, the nonequilibrium equations are easily integrated, and any standard procedure is satisfactory. The comparative evaluation of the two procedures, given in Chapter 7.0, reveals that a particular fourth-order predictor-corrector procedure appears to be the most suitable for the integration. It is not, however, appreciably superior to a fourth-order Runge-Kutta procedure. Because this latter procedure is considerably simpler to code for a computer than are the other procedures, it was utilized for all the numerical integrations given in Part II.

The author, in an extensive search of the literature, could locate only two published articles dealing directly with the numerical integration of stiff equations. The first is the previously mentioned paper by Curtiss and Hirschfelder (1952), which recommends the use of a certain, simple integration procedure. Inasmuch as their analysis is misleading, this procedure is discussed in Section 6.3.

The second article is by Certaine and is contained in the volume edited by Ralston and Wilf (1960). Hamming (1962) also presents Certaine's procedure in a simplified form. The procedure is as follows:

- (a) First, add y/ϵ to both sides of equation (2a) and then multiply by $\exp(x/\epsilon)dx$, thereby obtaining

$$\exp(x/\epsilon)dy + y \exp(x/\epsilon) \frac{dx}{\epsilon} = -\frac{1}{\epsilon} \exp(x/\epsilon)G(x)dx. \quad (3a)$$

- (b) If ϵ is assumed constant over an interval, say from x_1 to x_2 , then equation (3a) may be written as

$$d\left(y \exp(x/\epsilon)\right) = -\frac{1}{\epsilon} \exp(x/\epsilon)G(x)dx. \quad (3b)$$

- (c) Finally, integrate equation (3b) from x_1 to x_2 and then rearrange the results, thereby obtaining

$$y_2 = y_1 \exp\left(\frac{x_1 - x_2}{\epsilon}\right) - \frac{1}{\epsilon} \exp(-x_2/\epsilon) \int_{x_1}^{x_2} \exp(x/\epsilon)G(x)dx, \quad (3c)$$

where the quadrature on the right is readily obtained by standard numerical techniques.

The principal objection to Certaine's procedure is that it cannot be used to integrate the chemical rate equations given in Parts I or II unless their right-hand sides are first linearized and the small parameters explicitly exhibited. A second difficulty inherent in the procedure is the choice of the appropriate integration step size $x_2 - x_1$. In particular, the step size must be sufficiently small that the small parameters are nearly constant over the integration interval. Thus, the rate of change of magnitude of each small parameter must be periodically checked to see whether the step size should be altered.

3.0 CANONICAL FORM FOR THE CHEMICAL RATE EQUATIONS

In this chapter we shall show that when the flow is close to equilibrium, the chemical rate equations can be transformed into a system of stiff equations having the same form as the vibrational rate equations. While this transformation is not directly useful for solving problems, it does justify the use of equations (2), which are basic to the subsequent analysis. The particular transformation to be developed is not limited to steady one-dimensional flow, but is valid for unsteady three-dimensional flow as well.

The situation here is similar to that of a conservative mechanical system that performs small oscillations about a fixed equilibrium configuration. Many textbooks, such as Goldstein (1950), show that a linear transformation can be introduced to transform the equations of motion into a form referred to as canonical because of its simple mathematical structure. Since this procedure will be followed approximately in our derivation of the canonical form for the chemical rate equations, the general procedure for a vibrating mechanical system will first be summarized as follows:

- (a) The potential and kinetic energies are each expanded in a Taylor series about the equilibrium configuration. Only the first approximation to the potential and kinetic energies is retained. Consequently, the equations of motion are linear, second-order differential equations.
- (b) A principal-axis (or normal-coordinate) transformation is introduced that diagonalizes the coefficients of the potential and kinetic energies and thereby transforms the equations of motion into their canonical form.
- (c) An eigenvalue problem is solved for the eigenfrequencies in order to find the coefficients in the principal-axis transformation. Because the first approximation to the potential energy is positive definite, all the eigenfrequencies can be shown to be real and positive.

In parallel with the foregoing, the right-hand side of the chemical rate equations will first be linearized about an equilibrium reference state in a manner similar to that given by Vincenti (1959). The reference state for a slightly varying flow may be chosen as the uniform undisturbed (or free-stream) condition. When the flow has large variations, as in a nozzle, the reference state is taken as the nonconstant equilibrium-flow solution. In general, we shall consider the reference state to be a function of position and time. A linear transformation of the quantities that specify the chemical composition is then introduced. This transformation, when subjected to the condition that the rate equations transform into their canonical form, results in an eigenvalue problem. Each eigenvalue is the reciprocal of the small parameter in a canonical rate equation. Because of its complexity and length, the proof that all eigenvalues are real and positive is given in Appendix I.

3.1 CHEMICAL RATE EQUATIONS

The following four quantities, necessary in the subsequent analysis, are here defined as follows (cf. Part II):

N_1 = total number of reacting chemical species,

N_2 = number of chemical reactions,

N_3 = number of chemical species minus total number of components and polyatomic inert species,

N_4 = number of chemical species plus number of catalytic bodies that consist of a linear combination of other species.

As an illustration, given also in Part II, consider a model of air consisting of O, N, NO, N_2 , O_2 , and A and having eight chemical reactions to specify the chemistry. The number of species N_1 is then six, and the number of reactions N_2 is eight. Since there are three atomic species (one of which is inert), N_3 is three. The quantity N_4 is introduced in order to simplify the treatment of catalytic bodies made up of a group of species, such as $N + N_2 + NO$. Thus, a fictitious species M, which represents any atom or molecule from the species N, N_2 , or NO, can be used to represent the catalytic body. If M is the only such fictitious species, then N_4 in the present example is seven.

The composition of the fluid is specified by the number of moles of species i per unit mass of fluid n_i , and is referred to as the mole-mass ratio. When the mole-mass ratios n_i for the reacting species ($i = 1, \dots, N_1$) are known, the values for the fictitious species ($i = N_1 + 1, \dots, N_4$) are readily found from $N_4 - N_1$ algebraic equations. In addition, there exist $N_1 - N_3$ algebraic equations that express conservation of components and the constancy of n_i for any polyatomic inert species. The number of independent values of n_i , which is the minimum number that must be accounted for by rate

equations, is thus $N_4 - (N_4 - N_1) - (N_1 - N_3) = N_3$. The additional variables are given by algebraic equations of the form

$$n_i = a_{i0} + \sum_{j=1}^{N_3} a_{ij} n_j, \quad i = N_3+1, \dots, N_4, \quad (4)$$

where the a_{ij} are constants and the independent species are numbered from 1 to N_3 . The first $N_1 - N_3$ of equations (4) account for conservation of components and inert species. The remaining $N_4 - N_1$ equations provide for any nonsimple catalytic bodies.

The system of reactions is represented by

$$\sum_{i=1}^{N_4} \alpha_{ri} X_i \xrightleftharpoons[k_{br}]{k_{fr}} \sum_{i=1}^{N_4} \beta_{ri} X_i, \quad r = 1, \dots, N_2,$$

where X_i represents the chemical species and α_{ri} and β_{ri} are the stoichiometric coefficients of the reactants and products respectively. The forward and backward rate constants for reaction r are denoted by k_{fr} and k_{br} .

The chemical rate equations are now written for a general unsteady three-dimensional flow as follows:

$$\frac{Dn_i}{Dt} = \sum_{r=1}^{N_2} v_{ri} \frac{L_r}{\theta_r}, \quad i = 1, \dots, N_1, \quad (5)$$

where $D(\)/Dt$ is the substantial derivative and

$$v_{ri} \equiv \beta_{ri} - \alpha_{ri}, \quad i = 1, \dots, N_4, \quad r = 1, \dots, N_2, \quad (6a)$$

$$L_r(\rho, T, n_i) \equiv 1 - \frac{(\rho)^{\beta_r - \alpha_r}}{K_{cr}} \prod_{k=1}^{N_4} (n_k)^{v_{rk}}, \quad r = 1, \dots, N_2, \quad (6b)$$

$$[\theta_r(\rho, T, n_i)]^{-1} \equiv (\rho)^{\alpha_r - 1} k_{fr} \prod_{k=1}^{N_4} (n_k)^{\alpha_{rk}}, \quad r = 1, \dots, N_2. \quad (6c)$$

The density is denoted by ρ , the equilibrium constant for reaction r by K_{cr} , and α_r and β_r are defined as follows:

$$\alpha_r \equiv \sum_{j=1}^{N_4} \alpha_{rj}, \quad \beta_r \equiv \sum_{j=1}^{N_4} \beta_{rj}, \quad r = 1, \dots, N_2. \quad (7)$$

For convenience, the number of rate equations (5) is here taken as N_1 instead of the minimum number N_3 . The equations expressing conservation of components will be taken into account later when the number of rate equations will be reduced to N_3 . We may also note that any catalytic body or inert species actually appears in the foregoing equations only in θ_r , since $v_{ri} = 0$ for $i = N_1 + 1, \dots, N_4$. Consequently, all inert species are taken in the group numbered from $N_1 + 1$ to N_4 , formerly used only for fictitious species. The first N_1 species may then be arranged in any arbitrary manner.

3.2 LINEARIZED FORM OF THE CHEMICAL RATE EQUATIONS

For simplicity, we require that the reactions be sequenced such that the first N_3 are linearly independent. Using the notation $()^*$ to denote the equilibrium reference state, the reference composition is then given by

$$L_r(\rho^*, T^*, n_1^*) = 0, \quad r = 1, \dots, N_3, \quad (8a)$$

and

$$n_l^* = a_{l0} + \sum_{j=1}^{N_3} a_{lj} n_j^*, \quad l = N_3 + 1, \dots, N_1, \quad (8b)$$

where T denotes the temperature, and where equations (8b) express conservation of components (cf. equations (4)). In general, as noted earlier, the quantities with asterisks are functions of position and time. These quantities are thus assumed to be known. Since the local equilibrium composition is also necessary for the linearization, the pertinent equations are now given as follows:

$$L_r(\rho, T, n_1, e) = 0, \quad r = 1, \dots, N_3, \quad (9a)$$

and

$$n_{l,e} = a_{l0} + \sum_{j=1}^{N_3} a_{lj} n_{j,e}, \quad l = N_3+1, \dots, N_1. \quad (9b)$$

We now perturb about the equilibrium reference state, denoting the perturbation quantities by $()'$. Thus, we have

$$\left. \begin{aligned} \rho &= \rho^* + \rho', \\ T &= T^* + T', \\ n_i &= n_i^* + n_i', & i = 1, \dots, N_1, \\ n_{i,e} &= n_i^* + n_{i,e}', & i = 1, \dots, N_1, \\ \theta_r &= \theta_r^* + \theta_r', & r = 1, \dots, N_2, \end{aligned} \right\} \quad (10)$$

where $\theta_r^* \equiv \theta_r(\rho^*, T^*, n_1^*)$. Both $L_r(\rho, T, n_1)$ and $L_r(\rho, T, n_{1,e})$ are now expanded about the reference state thereby resulting in

$$\begin{aligned} L_r(\rho, T, n_1) &= L_r(\rho^*, T^*, n_1^*) + \left(\frac{\partial L_r}{\partial \rho} \right)^* \rho' + \left(\frac{\partial L_r}{\partial T} \right)^* T' \\ &+ \sum_{l=1}^{N_1} \left(\frac{\partial L_r}{\partial n_l} \right)^* n_l' + \dots, \quad r = 1, \dots, N_2, \end{aligned} \quad (11)$$

and

$$\begin{aligned} L_r(\rho, T, n_{1,e}) &= L_r(\rho^*, T^*, n_1^*) + \left(\frac{\partial L_r}{\partial \rho} \right)^* \rho' + \left(\frac{\partial L_r}{\partial T} \right)^* T' \\ &+ \sum_{l=1}^{N_1} \left(\frac{\partial L_r}{\partial n_l} \right)^* n_{l,e}' + \dots, \quad r = 1, \dots, N_2. \end{aligned} \quad (12)$$

According to equation (9a), the left-hand side of equation (12) is zero. Subtracting equation (12) from (11) thereby results in the following:

$$L_r = L_r(\rho, T, n_i) = \sum_{l=1}^{N_1} \left(\frac{\partial L_r}{\partial n_l} \right)^* (n'_l - n'_{l,e}) , \quad r = 1, \dots, N_2 \quad (13a)$$

Using equations (10)

$$n'_l - n'_{l,e} = n_l - n_{l,e} , \quad l = 1, \dots, N_1 , \quad (14)$$

we obtain finally the linear approximation

$$L_r = \sum_{l=1}^{N_1} \left(\frac{\partial L_r}{\partial n_l} \right)^* (n_l - n_{l,e}) , \quad r = 1, \dots, N_2 \quad (13b)$$

The right-hand side of equations (5) can thus be linearized, resulting in

$$\frac{Dn_i}{Dt} = \sum_{r=1}^{N_2} \frac{v_{ri}}{\theta_r^*} \sum_{l=1}^{N_1} \left(\frac{\partial L_r}{\partial n_l} \right)^* (n_l - n_{l,e}) , \quad i = 1, \dots, N_1 \quad (15a)$$

By interchanging the order of the r and l summations, we can write equation (15a) as follows

$$\frac{Dn_i}{Dt} = \sum_{l=1}^{N_1} \bar{\kappa}_{il} (n_{l,e} - n_l) , \quad i = 1, \dots, N_1 , \quad (15b)$$

where

$$\bar{\kappa}_{il} = \bar{\kappa}_{il}(\rho^*, T^*) = - \sum_{r=1}^{N_2} \frac{v_{ri}}{\theta_r^*} \left(\frac{\partial L_r}{\partial n_l} \right)^* , \quad i = 1, \dots, N_1 , \quad l = 1, \dots, N_1 \quad (16)$$

Conservation of components is finally accounted for by replacing both $n_{l,e}$ and n_l ($l = N_3+1, \dots, N_1$) in the right-hand side of equation (15b) by equations (4) and (9b). After some rearrangement, the following is obtained:

$$\frac{Dn_i}{Dt} = \sum_{j=1}^{N_3} \kappa_{ij}(n_{j,e} - n_j), \quad i = 1, \dots, N_3, \quad (17)$$

where

$$\kappa_{ij} = \bar{\kappa}_{ij} + \sum_{j=N_3+1}^{N_1} a_{j,i} \bar{\kappa}_{ij}, \quad i = 1, \dots, N_3, \quad j = 1, \dots, N_3. \quad (18)$$

3.3 TRANSFORMATION OF THE CHEMICAL RATE EQUATIONS

In order to transform the linearized form of the chemical rate equations (17) to their canonical form, we hypothesize the linear transformation

$$q_i = \sum_{k=1}^{N_3} B_{ik} n_k, \quad i = 1, \dots, N_3, \quad (19a)$$

and hence

$$q_{i,e} = \sum_{k=1}^{N_3} B_{ik} n_{k,e}, \quad i = 1, \dots, N_3, \quad (19b)$$

where the q_i are generalized mole-mass ratios and the B_{ik} are constants. The canonical form of equations (17) corresponding to equations (1), is then given by

$$\frac{Dq_i}{Dt} = -\kappa_i(q_i - q_{i,e}), \quad i = 1, \dots, N_3, \quad (20)$$

where the κ_i are functions of ρ^* and T^* . So far, the B_{ik} and κ_i are still undetermined. They are found by the following procedure:

- (a) Differentiate equation (19a), and replace Dn_k/Dt by means of equation (17), thereby obtaining

$$\frac{Dq_1}{Dt} = \sum_{k=1}^{N_3} B_{1k} \sum_{l=1}^{N_3} \kappa_{kl} (n_{l,e} - n_l) , \quad i = 1, \dots, N_3 . \quad (21a)$$

Then interchange the l and k summations to obtain

$$\frac{Dq_1}{Dt} = \sum_{l=1}^{N_3} \left(\sum_{k=1}^{N_3} \kappa_{kl} B_{1k} \right) (n_{l,e} - n_l) , \quad i = 1, \dots, N_3 . \quad (21b)$$

- (b) Replace $q_1 - q_{1,e}$ in equation (20) by means of equations (19), thereby obtaining

$$\frac{Dq_1}{Dt} = - \kappa_1 \sum_{k=1}^{N_3} B_{1k} (n_k - n_{k,e}) , \quad i = 1, \dots, N_3 . \quad (22a)$$

Replace the k -variable of summation by l in equation (22a) as follows:

$$\frac{Dq_1}{Dt} = \sum_{l=1}^{N_3} \kappa_1 B_{1l} (n_l - n_{l,e}) , \quad i = 1, \dots, N_3 . \quad (22b)$$

- (c) Compare equations (21b) and (22b) to obtain

$$\kappa_1 B_{1l} = \sum_{k=1}^{N_3} \kappa_{kl} B_{1k}$$

or

$$\sum_{k=1}^{N_3} \left\{ \delta_{kl} \kappa_1 - \kappa_{kl} \right\} B_{1k} = 0 , \quad i = 1, \dots, N_3, \quad l = 1, \dots, N_3 , \quad (23)$$

where δ_{kl} is the Kronecker delta.

- (d) For the moment, assume that the κ_1 and κ_{kl} have known constant values. With the i subscript in equations (23) held fixed, we then have N_3 homogeneous linear equations for the N_3 unknown B_{ik} . This system has a non-trivial solution if, and only if, the determinant of the coefficients satisfies the condition

$$|\delta_{kl}\kappa_1 - \kappa_{kl}| = 0, \quad i = 1, \dots, N_3. \quad (24a)$$

These N_3 equations are equivalent to the single characteristic equation

$$|\delta_{kl}\kappa - \kappa_{kl}| = 0, \quad (24b)$$

whose N_3 roots are the N_3 eigenvalues κ_1 . In general, the κ_{kl} are functions of ρ^* and T^* and therefore can be determined once an equilibrium-flow solution is available. For a given value of ρ^* and T^* , the numbers κ_{kl} then determine the eigenvalues by means of equation (24b). The κ_{kl} , in conjunction with a specific κ_1 , for example κ_1 , are next used to determine the coefficients B_{ik} ($k=1, \dots, N_3$) via equations (23). Thus, each eigenvalue κ_1 is associated with an eigenvector B_{ik} . The resulting transformation is valid however, only in a region of the flow that includes the specific values chosen for ρ^* and T^* . The extent of this region depends on the magnitude of the changes in the reference state. For example, if the reference state is taken to be the uniform undisturbed equilibrium condition for a slightly varying flow, then the transformation is valid for the entire flow field.

- (e) For a given root κ_1 of equation (24b), equations (23) do not uniquely determine the B_{ik} . This indeterminacy also occurs in vibration theory. One possible simple way of removing it is to require that

$$B_{i1} = 1, \quad i = 1, \dots, N_3. \quad (25)$$

Thus we have proved the existence of the linear transformation (19a) that transforms the chemical rate equations into their canonical form (20).

Two points still remain to be discussed. First, are the κ_1 real and positive? This is answered in the affirmative by the proof given in Appendix I. The physical significance of this result is discussed in the next chapter.

The second point is that the transformation (19a) can be extended to the mole-mass ratios n_i ($i = N_3+1, \dots, N_1$) by defining new generalized mole-mass ratios as follows:

$$q_i = n_i, \quad i = N_3+1, \dots, N_1. \quad (26)$$

Equations expressing conservation of components can then be written in terms of the generalized mole-mass ratios providing no two eigenvalues are equal, as is generally the case. (See Goldstein (1950) for additional details.) The transformation (19a) is then non-singular and an inverse transformation exists. When the inverses of transformation (19a) and (26) are substituted into equations (4), the desired result is obtained in the form

$$q_i = a_{i0} + \sum_{j=1}^{N_3} A_{ij} q_j, \quad i = N_3+1, \dots, N_1, \quad (27)$$

where the A_{ij} are constants.

4.0 GENERAL PROPERTIES OF STIFF EQUATIONS

In this chapter, certain general properties of stiff equations are discussed. These properties are based on equations (2) and are extensively used in the rest of the report.

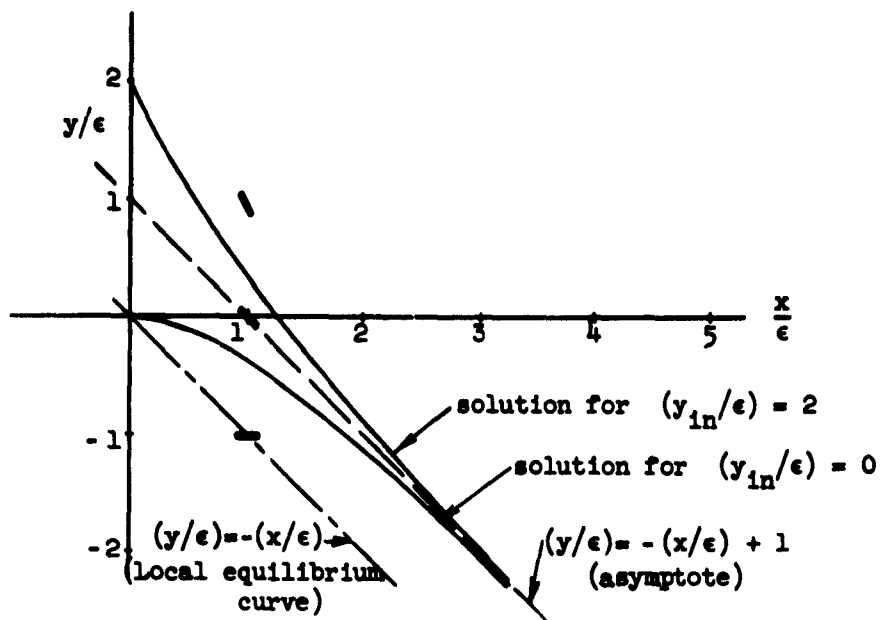
For convenience, we now combine equations (2a) and (2b) to give

$$\frac{dy}{dx} = -\frac{y+x}{\epsilon}, \quad (28)$$

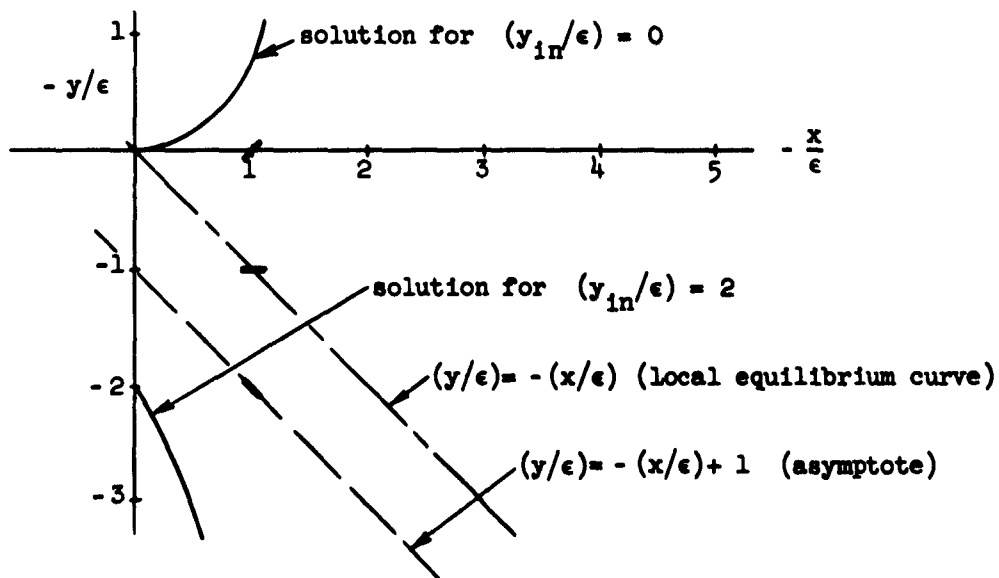
where ϵ is a small constant. The parameter ϵ thus corresponds to $u\tau_1$ of a vibrational rate equation or to u/κ_1 of a transformed chemical rate equation. The general solution of equation (28), in non-dimensional form, is readily seen to be

$$\frac{y}{\epsilon} = -\frac{x}{\epsilon} + 1 + \left(\frac{y_{in}}{\epsilon} - 1 \right) \exp \left(-\frac{x}{\epsilon} \right), \quad (29)$$

where $y(0) = y_{in}$ is the initial condition. Sketch 1 shows y/ϵ versus x/ϵ for positive ϵ for two different values of y_{in}/ϵ . The line $(y/\epsilon) = -(x/\epsilon)$, referred to as the local equilibrium curve, corresponds to $q_i = q_{i,e}$. Furthermore, the local equilibrium curve corresponds to



Sketch 1. Two Solutions of Equation (28) for Positive ϵ



Sketch 2. Two Solutions of Equation (28) for Negative ϵ

the equilibrium-flow solution when $\epsilon \rightarrow 0$. The other dashed curve, $(y/\epsilon) = -(x/\epsilon) + 1$, referred to as the asymptote, is a particular solution of equation (28) for $(y_{in}/\epsilon) = 1$. This curve corresponds to the true nonequilibrium solution. Also shown in Sketch 1, by heavy line segments, is the direction field of equation (28) for three different values of y/ϵ at $(x/\epsilon) = 1$. Because of the converging nature in the positive x -direction of the direction field, all solutions tend toward the asymptote exponentially fast, regardless of the initial value y_{in} . As a result, the two solutions shown differ only slightly from the asymptote for any x/ϵ greater than about 3. Any solution thus consists of two portions: a transient region in which the solution decays rapidly towards the asymptote, and a subsequent region in which the solution differs only negligibly from the asymptote. When equation (29) is plotted in dimensional coordinates (i.e., x and y) and ϵ is very small (e.g., $\epsilon = 10^{-4}$), then the local equilibrium curve $y = -x$ lies very close to the asymptote. Consequently, the direction field, when ϵ is small, changes extremely rapidly in the vicinity of the local equilibrium curve.

Sketch 2 is similar to Sketch 1 except ϵ is now negative. The coordinates in this sketch are chosen as $-(y/\epsilon)$ and $-(x/\epsilon)$ in order that the curves $(y/\epsilon) = -(x/\epsilon)$ and $(y/\epsilon) = -(x/\epsilon) + 1$ have similar orientation in the two figures. The direction field now diverges in the positive x -direction, as shown at $-(x/\epsilon) = 1$. Thus, all solutions tend to diverge exponentially fast from $(y/\epsilon) = -(x/\epsilon) + 1$, which, for convenience, is still referred to as the asymptote.

From a physical point of view, negative values for ϵ (or κ_1) imply that the nonequilibrium-flow solution cannot remain close to the equilibrium-flow solution. Any nonequilibrium nozzle calculation, even with equilibrium initial conditions, would in this case, diverge immediately from the equilibrium-flow solution. This situation is, however, physically unrealistic. We have, therefore, established the fact that negative ϵ (or κ_1) cannot occur by physical reasoning. Appendix I also establishes this by purely mathematical reasoning. Consequently, positive ϵ will be assumed in the rest of this report.

5.0 RUNGE-KUTTA ANALYSIS

The Runge-Kutta procedure is one of the oldest and most commonly used methods for the numerical integration of ordinary differential equations. This chapter considers the behavior of this procedure as applied to the integration of the stiff equation (28). Our main result will be the determination of the truncation error and of a stability criterion. To clarify the discussion, a brief preliminary description of the Runge-Kutta procedure will be helpful. For additional details, the interested reader is referred to Ince (1926), Hildebrand (1956) or Ralston and Wilf (1960).

The Runge-Kutta procedure is designed to approximate the Taylor-series solution

$$y(x_0+h) = y(x_0) + hy'(x_0) + \frac{h^2}{2!} y''(x_0) + \dots, \quad (30)$$

to the equation

$$y'(x) \equiv \frac{dy}{dx} = f(x,y), \quad (31)$$

at the point $x_0 + h$. The quantity h denotes the integration step size. In contrast to the formal Taylor-series solution (30), the Runge-Kutta procedure does not require the explicit evaluation of the second or higher derivatives. Rather, an approximation to expansion (30) is obtained at the expense of several evaluations of the first derivative. The approximation to the first $p+1$ terms on the right-hand side of (30), referred to as a p th-order Runge-Kutta procedure, requires the evaluation of p first derivatives. These derivatives are evaluated at various points in the x,y plane as follows:

$$\left. \begin{aligned} f_1 &= f(x_0, y_0), \\ f_2 &= f(x_0 + \alpha_2 h, y_0 + \beta_{21} h f_1), \\ f_3 &= f(x_0 + \alpha_3 h, y_0 + \beta_{31} h f_1 + \beta_{32} h f_2), \\ &\vdots \\ f_p &= f(x_0 + \alpha_p h, y_0 + h \sum_{j=1}^{p-1} \beta_{pj} f_j), \end{aligned} \right\} \quad (32)$$

where $y_0 \equiv y(x_0)$ and the α_i and β_{ij} are constants whose values are discussed shortly. The desired approximation is then obtained from

$$y(x_0 + h) = y_0 + h \sum_{i=1}^p \gamma_i f_i. \quad (33)$$

The constants α_i , γ_i , and β_{ij} are determined by the condition that equations (30) and (33) match to the p th order at the point (x_0, y_0) . In other words, the coefficients of like powers of h from $(h)^0$ to $(h)^p$ are equated with each other. This matching condition does not determine all the constants. For example, for a fourth-order procedure,

the matching condition leaves two of the constants unspecified. Values for these two constants may therefore be chosen arbitrarily.

At the start of a nonequilibrium calculation, as Section 5.1 will demonstrate, the step size h must be small. It is therefore imperative that provision be included in the integration procedure for increasing or, if necessary, decreasing the step size as the calculation proceeds in the x -direction. Any technique for varying h requires an estimate of the error that is introduced into the solution when a single step is integrated. This error is actually an estimate of the combined round-off and truncation errors. To estimate this error for a step of size h , the step is recomputed by means of two steps, each of length $h/2$, and the results compared. (See Ince (1926) for further details.) Although this is customarily referred to as a single step in the integration, the numerical solution, in fact, is primarily given by the two half-step calculations, which are more accurate than the full-step result. (This terminology is also used in Part II.) Each step of length h thus requires $3p$ complete evaluations of the right-hand side of the differential equations if the error estimation is performed. In comparison with two half steps without error estimation, which would require $2p$ evaluations, we see that error estimation increases overall computation time approximately 50%.

Runge-Kutta procedures will be compared with predictor-corrector procedures in Chapter 7.0. These latter procedures, however, do not estimate the error by recomputing with two half steps. This comparison therefore requires that each half step, henceforth denoted by h , be considered as the Runge-Kutta step size*. The number of evaluations of the right-hand side of the differential equations, in this instance, is taken as $1.5p$ per integration step.

In practice, the most widely used Runge-Kutta procedure is the fourth-order one. The results obtained from the subsequent analysis will be valid, however, for any order Runge-Kutta procedure from second-order ($p = 2$) on. Because of the considerable complexity of this procedure, such generality is rare. For an example of how extremely complicated the results usually are, the reader is referred to a paper by Lotkin (1951). The error associated with the integration of equation (31) by a specific fourth-order Runge-Kutta procedure is given in that paper. The simplicity and generality of the results that will be obtained here are directly due to the simple mathematical structure of equation (28).

* A single step of length h' in Part II will now be considered as two steps, each of length h , where $h' = 2h$.

5.1 TRUNCATION ERROR

We now show that the values chosen for the free constants in the Runge-Kutta procedure are immaterial when f is given by*

$$f = -\frac{y+x}{\epsilon} \quad (34)$$

The first step in the proof is to show that equation (33) does not contain higher powers of h than $(h)^p$. To do this, substitute equation (34) into (32) as follows:

$$\left. \begin{aligned} f_1 &= -\epsilon^{-1}(y_0 + x_0) \\ f_2 &= -\epsilon^{-1}(y_0 + \beta_{21}hf_1 + x_0 + \alpha_2h) = -\epsilon^{-1}(y_0 + x_0 + h[\beta_{21}f_1 + \alpha_2]) \\ &\vdots \end{aligned} \right\} \quad (35)$$

Thus, f_1 does not involve h , f_2 is linear in h , f_3 is quadratic in h , etc. The highest power of h in f_p is therefore $(h)^{p-1}$. Since the summation term in equation (33) is multiplied by an h , the highest power of h in this equation is $(h)^p$. Consequently, the right-hand side of equation (33) is precisely equal to the sum of the first $p+1$ terms of the Taylor-series expansion (30). Thus, no matter what values are chosen for the free constants, equation (33) will compute the same number, thereby proving the above assertion. The underlying reason for this result is that f , as given by equation (34), is linear in both x and y .

As a consequence of the foregoing, a simple expression for the truncation error** can now be obtained. The truncation error is defined as the difference between the $(p+2)^{\text{th}}$ term in the Taylor-series expansion (30) and the corresponding term computed by equation (33). The $(p+2)^{\text{th}}$ term in the series expansion is given by

* The quantity $-(\epsilon)^{-1}$, as used in this report, is equivalent to $A \equiv (\partial f / \partial y)$, which is frequently used in the literature of numerical analysis.

** This truncation error should not be confused with the estimate for the combined round-off and truncation errors mentioned in Section 5.0.

$$\frac{(h)^{p+1}}{(p+1)!} \left(\frac{d^{p+1}y}{dx^{p+1}} \right)_{x=x_0}, \quad (36)$$

whereas the corresponding term of equation (33), i.e., the term containing $(h)^{p+1}$, is zero. Denoting the truncation error of a p^{th} -order Runge-Kutta procedure by E_p , we therefore obtain the result

$$\begin{aligned} E_p &= \frac{(h)^{p+1}}{(p+1)!} \left[\left(-\frac{1}{\epsilon} \right)^{p+2} (y_0 + x_0 - \epsilon) \right], \\ &= -\frac{1}{\epsilon(p+1)!} [y_0 + x_0 - \epsilon] \left(-\frac{h}{\epsilon} \right)^{p+1}, \end{aligned} \quad (37)$$

where equation (34) is used to evaluate the derivative in expression (36).

Three important conclusions stem from this equation. First, the term $y_0 + x_0 - \epsilon$ represents the vertical distance from a point $(x_0, -x_0 + \epsilon)$ on the asymptote to the integration point (x_0, y_0) . Thus, when the numerical solution is close to the asymptote this term is small. In this case, it is possible to have $h > \epsilon$ and still have a small truncation error. A necessary condition then for a large integration step size is that the numerical solution be close to the asymptote. In terms of the integration of the nonequilibrium equations of Part II, this means that the numerical solution must be close to the physical solution in order that the step size be large.

Second, when $x_0 = y_0 = 0$ the truncation error is given by

$$E_p = \frac{1}{(p+1)!} \left(-\frac{h}{\epsilon} \right)^{p+1}. \quad (38)$$

The condition $x_0 = y_0 = 0$ represents the equilibrium initial condition for the rate equations. This topic was discussed in detail in Part II. We here note that for small E_p , in this condition, it is necessary that h be less than ϵ . For example, with $p = 4$ and $|E_p| = 10^{-5}$, equation (38) requires that $(h/\epsilon) = .2605$.

The third result to stem from equation (37) is that the condition

$$|E_{p+1}| \leq |E_p|, \quad (39)$$

requires that

$$h/\epsilon \leq p + 2 . \quad (40)$$

In other words, for the same integration step size, higher-order procedures will have smaller absolute truncation errors provided condition (40) is satisfied. This condition must be fulfilled, as will be shown, in order that the integration procedure remain stable.

5.2 STABILITY

This chapter concludes with a discussion of stability of Runge-Kutta procedures applicable to the region in which the numerical solution differs only negligibly from the asymptote.

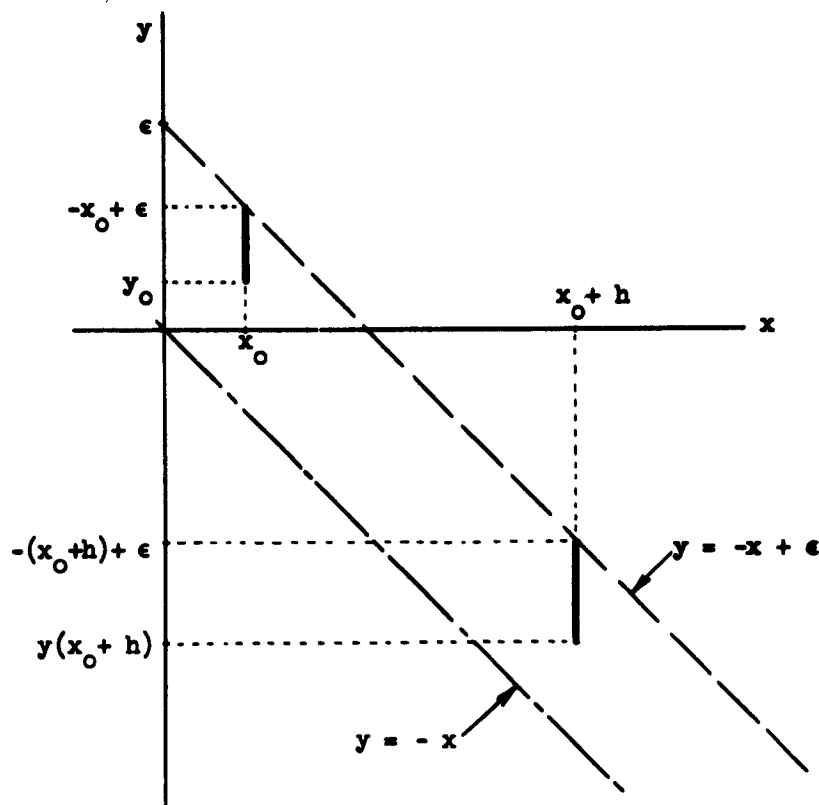
The accuracy of a numerical integration depends both on the error incurred at each step of the integration and on how the error grows. If the error decreases as the succeeding steps are performed, the integration procedure is referred to as stable. In other words, if the numerical solution converges to the exact solution as the succeeding steps are performed, the integration procedure is considered to be stable. If the error remains constant, the procedure is termed neutrally stable, and if the error increases, the procedure is termed unstable.

We now turn to the problem of applying this definition of stability to our consideration of equation (28). As Sketch 1 shows, all analytical solutions tend toward the asymptote. Consequently, any stable numerical solution must also tend toward it. In other words, the distance from the point $(x_0 + h, y(x_0 + h))$ to the asymptote must be less than the distance from (x_0, y_0) to the asymptote, in order for the numerical procedure to be stable. For simplicity of formulation, we shall use vertical distance rather than actual distance; the final result, of course, is the same. Sketch 3 presents a sketch of the pertinent quantities. The two heavy vertical lines represent the two distances. Their ratio, denoted by δ , is given by

$$\delta \equiv \frac{-(x_0 + h) + \epsilon - y(x_0 + h)}{-x_0 + \epsilon - y_0} , \quad (41a)$$

where the numerator is the vertical distance at $x_0 + h$. After a certain amount of algebraic manipulation δ can be written as

$$\delta = 1 - \frac{\frac{y(x_0 + h) - y_0}{\epsilon} + \frac{h}{\epsilon}}{1 - \frac{y_0 + x_0}{\epsilon}} . \quad (41b)$$



Sketch 3. Stability Diagram

The quantity $y(x_0 + h) - y_0$ in equation (41b) is to be computed by a p^{th} -order Runge-Kutta procedure. As shown previously, the same value can also be obtained from the Taylor-series expansion (30) truncated after the $(h)^p$ term. When this expansion is written out, we obtain the equation

$$\begin{aligned}
 y(x_0 + h) - y_0 = & h \left\{ -\frac{y_0 + x_0}{\epsilon} \right\} + \frac{h^2}{2!} \left\{ -\frac{-\frac{y_0 + x_0}{\epsilon} + 1}{\epsilon} \right\} \\
 & + \frac{h^3}{3!} \left\{ \frac{-\frac{y_0 + x_0}{\epsilon} + 1}{\epsilon^2} \right\} + \dots + \frac{(h)^p}{p!} \left\{ \frac{-\frac{y_0 + x_0}{\epsilon} + 1}{(-\epsilon)^{p-1}} \right\}, \quad (42a)
 \end{aligned}$$

where the curly brackets contain the derivatives $y'(x_0)$, $y''(x_0), \dots, y^{(p)}(x_0)$, evaluated in accordance with equation (28). Again after some algebraic manipulation, equation (42a) can be rewritten in the more convenient form

$$\frac{y(x_0 + h) - y_0}{\epsilon} + \frac{h}{\epsilon} = - \left(1 - \frac{y_0 + x_0}{\epsilon} \right) \sum_{i=1}^p \frac{1}{i!} \left(- \frac{h}{\epsilon} \right)^i. \quad (42b)$$

After substituting equation (42b) into equation (41b), we obtain the final result

$$\begin{aligned} \delta\left(\frac{h}{\epsilon}; p\right) &= \sum_{i=0}^p \frac{1}{i!} \left(- \frac{h}{\epsilon} \right)^i \\ &= 1 - \frac{h}{\epsilon} + \frac{1}{2!} \left(\frac{h}{\epsilon} \right)^2 - \dots + \frac{1}{p!} \left(- \frac{h}{\epsilon} \right)^p. \end{aligned} \quad (41c)$$

According to the foregoing considerations, a Runge-Kutta procedure of order p is stable in the integration of equation (28) if

$$|\delta| < 1. \quad (43)$$

For each integer value of p there will exist two non-negative values of h/ϵ such that $|\delta| = 1$. One of these values is $(h/\epsilon) = 0$. The other value, denoted by the stability parameter $B(p)$, represents the upper limit for values of h/ϵ for which the integration is stable. Thus, condition (43) is equivalent to the following stability condition

$$h/\epsilon < B(p), \quad (44)$$

where the quantity $B(p)$ is determined by the relation $|\delta(B;p)| = 1$. The following table lists $B(p)$ for different values of p .

STABILITY PARAMETER $B(p)$

p	$\delta(B;p)$	$B(p)$
2	1	2.00
3	-1	2.52
4	1	2.80
5	-1	3.20

An important conclusion of the above analysis is that stability considerations impose a limit on the Runge-Kutta step size when ϵ is small. Stable integration with $(h/\epsilon) > 1$, however, is possible. When $(h/\epsilon) > 1$, the truncation error will be inordinately large unless $y_0 + x_0 - \epsilon$ is small, i.e., unless the numerical solution is quite close to the asymptote. Thus, only after the transient region is completed are large integration steps, i.e., $B > h/\epsilon > 1$, possible.

Negative values for $\delta(B;p)$ in the above table signify that $y(x_0 + h)$ and y_0 are on opposite sides of the asymptote. For example, for $p = 4$ the integration procedure is stable for $0 < h/\epsilon < 2.80$, and y_0 and $y(x_0 + h)$ are both on the same side of the asymptote. The above table shows that condition (40), as already mentioned, is always satisfied when the integration procedure is stable. Therefore, if h/ϵ is chosen to achieve stability, then the higher-order the procedure, the smaller the truncation error. This table also demonstrates that higher-order Runge-Kutta procedures are more stable than lower-order procedures. Each increase in order, however, requires additional evaluations of the derivative $f(x,y)$ (see equation (31)). Since the evaluations are frequently very time consuming on a digital computer, a compromise between stability and computation time becomes necessary. When the truncation error, stability, and computation time are all considered, the fourth-order procedure appears to be a reasonable choice.

The value of B , given in Part II for a fourth-order Runge-Kutta procedure, is based on a double step of length $2h$. Its value is therefore $2 \times 2.8 = 5.6$. (The value actually given in Part II is 5.7, and is slightly in error.)

6.0 PREDICTOR-CORRECTOR ANALYSIS

This chapter analyzes predictor-corrector procedures in a fashion similar to that of the Runge-Kutta analysis. As in the preceding chapter, a brief description of predictor-corrector procedures will help clarify the analysis. For additional details, the reader is referred to Ralston and Wilf (1960) or Hamming (1962). The latter reference, in particular, contains an extensive discussion of predictor-corrector procedures.

Any predictor-corrector procedure is a finite-difference method that uses previously computed data to assist in obtaining $y(x_0 + h)$. In other words, the calculation of $y(x_0 + h)$ generally requires the use of $y(x_0)$, $y(x_0 - h)$, $y(x_0 - 2h)$, ... and of the derivatives $y'(x_0 + h)$, $y'(x_0)$, $y'(x_0 - h)$, ... The more back data used, the higher is the order of the procedure. This use of back data represents the chief difference between a Runge-Kutta procedure, which is not a finite-difference method, and a predictor-corrector procedure. Nevertheless, both types of procedure attempt to estimate the solution of a differential equation by means of a Taylor-series expansion.

A typical corrector (or predictor) formula is given by (see Hamming (1962))

$$y(x_0 + h) = a_0 y(x_0) + a_1 y(x_0 - h) + a_2 y(x_0 - 2h) + h \left[b_{-1} y'(x_0 + h) + b_0 y'(x_0) + b_1 y'(x_0 - h) + b_2 y'(x_0 - 2h) \right], \quad (45)$$

where a_1 and b_1 are constants. Of the seven constants here, five are usually chosen such that formula (45) matches the Taylor-series expansion (30) to fourth-order. The remaining two constants may then be chosen to improve stability. Corrector formulas require the use of $y'(x_0 + h)$, i.e., $b_{-1} \neq 0$, whereas predictor formulas do not, i.e., $b_{-1} = 0$. With $b_{-1} = 0$, equation (45) is first used to predict a value for $y(x_0 + h)$. Based on this predicted value, a first estimate for the derivative $y'(x_0 + h)$ is computed by means of the differential equation. This derivative is next used to calculate a new, more precise value for $y(x_0 + h)$ by means of a corrector formula. A second, and final, value for $y'(x_0 + h)$ is now computed using this new value for $y(x_0 + h)$ in conjunction with the differential equation. The last values for $y(x_0 + h)$ and $y'(x_0 + h)$ are the ones used for the calculation of the next step.

As already noted in Chapter 5.0, any feasible integration procedure must be able to vary the integration step size. Therefore, a method for estimating the local truncation error is necessary. An error estimate with the above predictor-corrector procedure is possible if the order of the truncation errors for the predictor and the corrector are equal.

We shall always assume this to be the case. The truncation error of the overall procedure is then of the same order as that of the predictor or corrector. This error is obtained by properly comparing the results of the predictor and the corrector calculations. One often quoted advantage of predictor-corrector as compared to Runge-Kutta procedures is that no additional evaluations of the derivative $f(x,y)$ (see equation (31)) are necessary to estimate the error.

The analysis contained in this chapter is restricted to the foregoing predictor-corrector procedure. Unlike Runge-Kutta procedures, there are many different types of predictor-corrector procedures. An exhaustive analysis of the numerous types is beyond the scope of this work. In view of this, the results of this chapter are of a preliminary nature only.

A further comparison between Runge-Kutta and predictor-corrector procedures will be useful before discussing truncation error and stability. First, predictor-corrector procedures are not self-starting. Thus, a Runge-Kutta procedure, which is self-starting, is generally used to compute the first few steps. As a consequence, predictor-corrector procedures are more involved than Runge-Kutta procedures to program for a computer.

Second, varying the integration step size is more troublesome with predictor-corrector procedures, where the step size is either halved or doubled, as compared to Runge-Kutta procedures, where the step size can be changed arbitrarily. According to criterion (44), which is also applicable here, a predictor-corrector procedure will be stable after the step size is doubled only if $(h/\epsilon) < \frac{1}{2}B$, where h is the original step size. Thus, the step size must be less than half its permissible size if the procedure is to be stable after the step size is doubled. The stability parameters that will be derived in Section 6.2 therefore overestimate the average step size that is possible. An efficient technique for controlling the step size would double it whenever $(h/\epsilon) < \frac{1}{2}B$. In this circumstance, the step size would on average be $(3/4)$ that allowed by criterion (44). Consequently, a more realistic stability criterion for predictor-corrector procedures is given by

$$\frac{h}{\epsilon} < \frac{3}{4} B, \quad (46)$$

as compared to criterion (44).

Third, the nature of instability is different for the two procedures. As we have seen, Runge-Kutta instability is due to the numerical solution diverging from the exact one. Any predictor-corrector procedure, being a finite-difference method, may converge toward one of perhaps several spurious solutions instead of toward the exact solution. When this occurs the procedure is also said to be unstable. Both types of

instability, however, stem from too large a value for the parameter h/ϵ . Thus, stability considerations impose a limit on the step size for both types of integration procedures.

Finally, predictor-corrector procedures require fewer evaluations of the derivative $f(x,y)$ and, therefore, may require less computation time than Runge-Kutta procedures. In particular, the foregoing predictor-corrector procedure requires only two evaluations of the derivative per integration step. With any Runge-Kutta procedure, the corresponding number of evaluations is $1.5p$.

6.1 TRUNCATION ERROR

The truncation error for a predictor-corrector procedure of order p is given by

$$E_p = c_p \frac{(h)^{p+1}}{(p+1)!} y^{(p+1)}(\theta), \quad (47)$$

where c_p is a constant that depends on the specific procedure. The derivative $y^{(p+1)}(\theta)$ is evaluated at a point θ in the interval from $x_0 - rh$, the farthest back-data point used in the integration formula, to $x_0 + h$.^{*} This derivative is essentially the same as that in expression (36) with the exception that here $x = \theta$. Thus, when equation (28) is integrated, the truncation error is still given by equation (37), except for the constant c_p , which generally is within an order of magnitude of unity, and the replacement of $y_0 + x_0$ by $y(\theta) + \theta$. Consequently, the first two conclusions in the previous section following equation (37) also apply here. In particular, it is again possible for h to be greater than ϵ and for the truncation error nevertheless to be small provided the numerical solution is sufficiently close to the asymptote.

6.2 STABILITY

At least two different definitions of stability for predictor-corrector procedures are currently in use. The first of these, referred to as relative stability, requires that spurious solutions be small relative to the exact solution. Relative stability is important only when h/ϵ is negative and therefore is not considered here. This section deals with the more common form of stability, which is not as restrictive as relative stability, but is applicable only when h/ϵ is positive. A finite-difference scheme is considered stable when all the roots of its characteristic polynomial are interior to the unit circle. In this circumstance, all spurious solutions of the finite-

* Equation (47) assumes that the influence function is of constant sign in the interval from $x_0 - rh$ to $x_0 + h$. For additional details see Hamming (1962).

difference scheme decay toward zero. Thus, once the transient region (see Chapter 4.0) is completed, the numerical solution must converge to the exact solution when the finite-difference scheme is stable.

There are many stability studies of the type of predictor-corrector procedure that we are here concerned with. For instance, the reader is referred to Dahlquist (1956) and (1959), Hamming (1959) and (1962), Crane and Lambert (1962), Henrici (1962), and Chase (1962). With the exception of Chase, all of the referenced authors have assumed that stability of the overall predictor-corrector procedure is determined primarily by the stability of the corrector alone. Consequently, these authors derive and examine the characteristic polynomial relevant only to the corrector. The foregoing assumption, however, is usually correct only in the limit as h tends to zero. By investigating a few special predictor-corrector procedures, Chase (1962) has shown that this assumption may not be valid as h/ϵ increases. Therefore, to determine stability parameters, it is necessary to consider the overall integration procedure, not just the corrector. This is accomplished by deriving a characteristic polynomial for the overall procedure.

We now define a generalized predictor-corrector procedure by the following formulas:

$$z(x_0 + h) = \sum_{i=0}^{k'} A_i y(x_0 - ih) + h \sum_{i=0}^{k'} B_i y'(x_0 - ih), \quad (48)$$

$$y(x_0 + h) = \sum_{i=0}^k a_i y(x_0 - ih) + h b_{-1} z'(x_0 + h) + h \sum_{i=0}^k b_i y'(x_0 - ih), \quad (49)$$

where A_i , B_i , a_i and b_i are constants and k' and k are positive integers. Equation (48) is the predictor and is used to determine a first estimate $z(x_0 + h)$ for $y(x_0 + h)$. The quantity $z'(x_0 + h)$ in the corrector, formula (49), is given by $f(x_0 + h, z(x_0 + h))$ (see equation (31)).

As mentioned earlier, we shall require that each of the above formulas have a truncation error of order $(h)^{p+1}$. This condition will then determine some of the constants A_i , B_i , a_i , and b_i . The formulation of this condition also requires that the estimates $z(x_0 + h)$ and $z'(x_0 + h)$ be replaced by $y(x_0 + h)$ and $y'(x_0 + h)$ respectively. When this is done, formulas (48) and (49) are of the same form except that $B_{-1} = 0$ in formula (48). To satisfy the condition that the corrector be of order p , we require that formula (49) be exact for

the functions $y = 1, x, x^2, \dots, x^p$. In other words, the $p + 1$ functions

$$\left. \begin{array}{l} y = 1, \\ y = x, \\ \vdots \\ y = x^p, \end{array} \right\} \quad (50)$$

are substituted successively into formula (49), thereby resulting in the system of equations

$$(x_0 + h)^m = \sum_{i=0}^k a_i (x_0 - ih)^m + mh \sum_{i=-1}^k b_i (x_0 - ih)^{m-1}, \quad m=0,1,\dots,p. \quad (51)$$

Since the a_i 's and b_i 's ought to be independent of the arbitrary parameters x_0 and h , it should be possible to replace equations (51) by a simpler, alternate system not containing these two parameters. This system, given without proof by Dahlquist (1956), is derived in the following paragraph. Equation (51), and hence equation (59), can also be used to determine the constants in the predictor providing a_i , b_i , and k are replaced by A_i , B_i , and k' respectively and $B_{-1} = 0$.

We proceed as follows: Apply the binomial theorem successively to $(x_0 + h)^m$, $(x_0 - ih)^m$, and $(x_0 - ih)^{m-1}$, thereby resulting in the following equations:

$$(x_0 + h)^m = \sum_{j=0}^m \binom{m}{j} x_0^{m-j} h^j, \quad (52)$$

where

$$\binom{m}{j} \equiv \text{binomial coefficient} = \frac{m(m-1)\dots(m-j+1)}{j!}, \quad (53)$$

$$\begin{aligned}
 (x_0 - ih)^m &= \sum_{j=0}^m \binom{m}{j} x_0^{m-j} (-ih)^j, \\
 &= \sum_{j=0}^m (-1)^j (i)^j \binom{m}{j} x_0^{m-j} h^j, \quad (54)
 \end{aligned}$$

and

$$(x_0 - ih)^{m-1} = \sum_{j'=0}^{m-1} (-1)^{j'} (i)^{j'} \binom{m-1}{j'} x_0^{m-1-j'} h^{j'}. \quad (55)$$

Replace j' by $j-1$, and use the identity

$$\binom{m-1}{j-1} \equiv \left(\frac{1}{m}\right) \binom{m}{j} \quad (56)$$

to transform equation (55) into

$$(x_0 - ih)^{m-1} = -\frac{1}{m} \sum_{j=1}^m (-1)^j (i)^{j-1} j \binom{m}{j} x_0^{m-j} h^{j-1}. \quad (57)$$

Substitute equations (52), (54) and (57) into equation (51) and then interchange the j and i summations, thereby resulting in

$$\begin{aligned}
 \sum_{j=0}^m \binom{m}{j} x_0^{m-j} h^j &= \sum_{j=0}^m (-1)^j \binom{m}{j} x_0^{m-j} h^j \sum_{i=0}^k (i)^j a_i \\
 &\quad - \sum_{j=0}^m (-1)^j \binom{m}{j} x_0^{m-j} h^j j \sum_{i=-1}^k (i)^{j-1} b_i, \\
 &\quad m=0,1,\dots,p. \quad (58)
 \end{aligned}$$

When the coefficients of like powers of h , or x_0 , are equated, the desired result is obtained:

$$\sum_{i=0}^k (i)^j a_i - j \sum_{i=-1}^k (i)^{j-1} b_i = (-1)^j, \quad j = 0, 1, \dots, p. \quad (59)$$

Equations (59) have the following simple form when explicitly written out

$$\left. \begin{aligned} a_0 + a_1 + a_2 + \dots + a_k &= 1, \\ a_1 + 2a_2 + \dots + ka_k - (b_{-1} + b_0 + b_1 + b_2 + \dots + b_k) &= -1, \\ a_1 + 4a_2 + \dots + k^2 a_k - 2(-b_{-1} + b_1 + 2b_2 + \dots + kb_k) &= 1, \\ &\vdots \\ a_1 + 2^p a_2 + \dots + k^p a_k - p \left((-1)^{p-1} b_{-1} + b_1 + 2^{p-1} b_2 + \dots + (k)^{p-1} b_k \right) &= (-1)^p. \end{aligned} \right\} (60)$$

To further illustrate the importance of Chase's assertion, we shall analyze the Adams method in detail. This will be accomplished by examining the stability of the predictor alone, of the corrector alone, and of the overall procedure. Furthermore, in order to parallel the Runge-Kutta analysis given in Chapter 5.0, this analysis will also consider all orders from $p = 2$ to $p = 5$.

We now give without proof (see Chase (1962)) the characteristic polynomials needed for the investigation. All of the polynomials are based directly on formulas (48) and (49). The various constants that appear in them will then be specialized for the Adams method and for equations (60). The polynomials are denoted by $C^{(pr)}$, $C^{(co)}$, and $C^{(pc)}$ where the superscripts denote predictor, corrector, and predictor-corrector respectively. They are as follows:

$$C^{(pr)} \equiv \rho^{k'+1} + \sum_{i=0}^{k'} (-A_i + B_i \frac{h}{\epsilon}) \rho^{k'-i} = 0, \quad (61)$$

$$C^{(co)} = (1 + b_{-1} \frac{h}{\epsilon}) \rho^{k+1} + \sum_{i=0}^k (-a_i + b_i \frac{h}{\epsilon}) \rho^{k-i} = 0, \quad (62)$$

$$C^{(pc)} = \begin{cases} \rho^{k+1} + \sum_{i=0}^k (-a_i + b_i \frac{h}{\epsilon}) \rho^{k-i} + b_{-1} \frac{h}{\epsilon} \sum_{i=0}^{k'} (A_i - B_i \frac{h}{\epsilon}) \rho^{k-i} = 0, & \text{if } k \geq k', \quad (63) \\ \rho^{k'+1} + \sum_{i=0}^k (-a_i + b_i \frac{h}{\epsilon}) \rho^{k'-i} + b_{-1} \frac{h}{\epsilon} \sum_{i=0}^{k'} (A_i - B_i \frac{h}{\epsilon}) \rho^{k'-i} = 0, & \text{if } k \leq k'. \quad (64) \end{cases}$$

Following Nordsieck (1962), we shall require any Adams formula to have a characteristic polynomial of the form

$$(\rho-1)\rho^{k'} = 0, \quad (65a)$$

or

$$(\rho-1)\rho^k = 0, \quad (65b)$$

when $(h/\epsilon) = 0$. The root $\rho = 1$ of equations (65) corresponds to the exact solution, whereas the remaining roots $\rho = 0$ are the spurious solutions. For $(h/\epsilon) = 0$, the spurious solutions for any Adams formula do not, in fact, exist. Thus, the Adams formulas are optimum in terms of damping out the spurious solutions when $(h/\epsilon) = 0$. Because of this important property, the Adams method is probably the most widely used of all predictor-corrector procedures. By comparing equation (61) with (65a), the following values for the constants A_i are obtained:

$$A_0 = 1, \quad A_1 = A_2 = \dots = A_{k'} = 0. \quad (66a)$$

By comparison of equations (62) and (63) with (65b), and equation (64) with (65a), the following values for the constants a_i are obtained for all three cases:

$$a_0 = 1, \quad a_1 = a_2 = \dots = a_k = 0. \quad (66b)$$

The preceding result means that both of the summations farthest to the left in formulas (48) and (49) can now be replaced by $y(x_0)$. Since conditions (66b) also satisfy the first of equations (60), the latter is no longer needed.

To determine uniquely the remaining $k' + 1$ constants B_i in formula (48) by the p linear equations (60), k' must equal $p - 1$. It is, of course, possible for k' to be greater than $p - 1$, thereby providing $k' - p + 1$ free parameters, which might then be chosen to improve stability when $(h/\epsilon) > 0$. Such an extension is, however, beyond the scope of this work, and is not given. To determine uniquely the remaining $k + 2$ constants b_i in formula (49) by the p linear equations (60), k must equal $p - 2$. The same remark as above concerning free-parameters also applies to the corrector (49). The results of these calculations are summarized for $p = 2, 3, 4$, and 5 in the following table:

CONSTANTS B_i AND b_i FOR THE ADAMS METHOD

p	B_0	B_1	B_2	B_3	B_4	b_{-1}	b_0	b_1	b_2	b_3
2	$\frac{3}{2}$	$-\frac{1}{2}$	-	-	-	$\frac{1}{2}$	$\frac{1}{2}$	-	-	-
3	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$	-	.	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$	-	-
4	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$	-	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$	-
5	$\frac{1901}{720}$	$-\frac{2774}{720}$	$\frac{2616}{720}$	$-\frac{1274}{720}$	$\frac{251}{720}$	$\frac{251}{720}$	$\frac{646}{720}$	$-\frac{264}{720}$	$\frac{106}{720}$	$-\frac{19}{720}$

Since $k' = p - 1$ and $k = p - 2$, the three relevant characteristic polynomials (61), (62), and (64) now take the form

$$c^{(pr)} = \rho^p - \rho^{p-1} + \frac{h}{\epsilon} \sum_{i=0}^{p-1} B_i \rho^{p-1-i} = 0, \quad (67)$$

$$c^{(co)} = (1 + b_{-1} \frac{h}{\epsilon}) \rho^{p-1} - \rho^{p-2} + \frac{h}{\epsilon} \sum_{i=0}^{p-2} b_i \rho^{p-2-i} = 0, \quad (68)$$

$$c^{(pc)} = \rho^p + (-1 + b_{-1} \frac{h}{\epsilon}) \rho^{p-1} + \frac{h}{\epsilon} \sum_{i=0}^{p-2} b_i \rho^{p-1-i} - b_{-1} (\frac{h}{\epsilon})^2 \sum_{i=0}^{p-1} B_i \rho^{p-1-i} = 0, \quad (69)$$

where B_i and b_i are given by the preceding table. For a given value of p , each root of each of the characteristic polynomials (67), (68), and (69) depends on the value of the parameter h/ϵ . For a sufficiently small positive value of h/ϵ all of the roots are located in the interior of the unit circle in the complex plane. In this circumstance, as mentioned earlier, the particular formula is stable. Thus, the predictor is referred to as stable when all of the roots of equation (67) are in the interior of the unit circle. The Adams method itself, is stable, of course, only when the roots of equation (69), the characteristic equation for the overall procedure, are in the interior of the unit circle.

For some positive value of h/ϵ , equation (67) will have a root with unit magnitude, i.e., $|\rho| = 1$. When this is the smallest positive value of h/ϵ for which this occurs, then this value is denoted by the stability parameter $B^{(pr)}(p)$. Hence, the predictor is stable for all values of h/ϵ that satisfy $0 < (h/\epsilon) < B^{(pr)}$. In a similar fashion, the stability parameters $B^{(co)}(p)$ and $B^{(pc)}(p)$ are defined.

By means of Wilf's (1959) criterion, the stability parameters were found for various values of p . The following table summarizes the results of these lengthy calculations:

STABILITY PARAMETERS RELEVANT TO THE ADAMS METHOD

p	$B^{(pr)}(p)$	$B^{(co)}(p)$	$B^{(pc)}(p)$
2	1.00	∞	2.00
3	.545	6.00	1.72
4	.300	3.00	1.28
5	.163	1.84	.947

The preceding table shows that all three stability parameters decrease with increasing p . This result is directly opposite to that found for the Runge-Kutta procedure. As might be anticipated, for a given value of p , the overall-procedure stability parameter $B^{(pc)}$ has a value between $B^{(pr)}$ and $B^{(co)}$. Nevertheless, it differs appreciably from $B^{(co)}$. Thus, Chase's assertion is valid here. Since $B^{(pc)}$ does not vary much, there is little reason to prefer the low-order ($p = 2$ or 3) Adams procedures to the higher-order ones.

The foregoing analysis indicates that stability is the central problem when stiff equations are integrated by a predictor-corrector procedure. For a given value of p , however, the Adams method does not have an optimum, i.e., a maximum, stability parameter. We shall now consider briefly the question of a procedure with an optimum stability parameter. The discussion is, however, restricted to $p = 4$.

Set $k' = 2$ and $k = 2$ in the generalized predictor-corrector procedure given by formulas (48) and (49). These formulas then become

$$z(x_0 + h) = A_0 y(x_0) + A_1 y(x_0 - h) + A_2 y(x_0 - 2h) + h \left[B_0 y'(x_0) + B_1 y'(x_0 - h) + B_2 y'(x_0 - 2h) \right], \quad (70)$$

$$y(x_0 + h) = a_0 y(x_0) + a_1 y(x_0 - h) + a_2 y(x_0 - 2h) + h \left[b_{-1} z'(x_0 + h) + b_0 y'(x_0) + b_1 y'(x_0 - h) + b_2 y'(x_0 - 2h) \right], \quad (71)$$

where, as a result of equations (60), the various constants satisfy the relations

$$\left. \begin{aligned} a_0 &= 1 - a_1 - a_2 \\ b_{-1} &= \frac{1}{24}(9 - a_1) \\ b_0 &= \frac{1}{24}(19 + 13a_1 + 8a_2) \\ b_1 &= \frac{1}{24}(-5 + 13a_1 + 32a_2), \\ b_2 &= \frac{1}{24}(1 - a_1 + 8a_2), \end{aligned} \right\} \begin{aligned} A_0 &= -8 - A_2, \\ A_1 &= 9, \\ B_0 &= \frac{1}{3}(17 + A_2), \\ B_1 &= \frac{2}{3}(7 + 2A_2), \\ B_2 &= \frac{1}{3}(-1 + A_2). \end{aligned} \quad (72)$$

Formulas (70) and (71) involve three free parameters a_1 , a_2 , and A_2 , which will be chosen later. The corrector, formula (71), has been examined in detail for relative stability by Hamming (1962). Crane and Lambert (1962) have also examined this corrector for stability. They find, for example, that formula (71) is stable for $0 < (h/\epsilon) < 10^4$ when $a_1 = -1.6$ and $a_2 = .925$.

Stability of the overall procedure is determined by the roots of the characteristic polynomial

$$\rho^3 - \left[a_0 - (b_0 + b_{-1}A_0)\left(\frac{h}{\epsilon}\right) + b_{-1}B_0\left(\frac{h}{\epsilon}\right)^2 \right] \rho^2 - \left[a_1 - (b_1 + b_{-1}A_1)\left(\frac{h}{\epsilon}\right) + b_{-1}B_1\left(\frac{h}{\epsilon}\right)^2 \right] \rho - \left[a_2 - (b_2 + b_{-1}A_2)\left(\frac{h}{\epsilon}\right) + b_{-1}B_2\left(\frac{h}{\epsilon}\right)^2 \right] = 0, \quad (73)$$

which is readily obtained from equation (64). Each set of values for a_1 , a_2 , and A_2 results in a specific polynomial (73) and thus a specific stability parameter. It has not been possible, however, to determine analytically the value of a_1 , a_2 , and A_2 such that the overall stability parameter is an optimum. Consequently, stability parameters for a wide choice of values for a_1 , a_2 , and A_2 were computed. From among the different values chosen, $a_1 = .8$, $a_2 = 0$, and $A_2 = 0$ resulted in the largest value of the stability parameter. With this choice, the overall procedure is stable for $0 < (h/\epsilon) < 1.82$. While this choice probably does not correspond to the optimum one, the author believes that it does not differ from it appreciably. Compared to the $p = 4$ Adams method, this choice results in a 40% increase in the value of the stability parameter. Consequently, this procedure rather than any of the Adams procedures, will be used in the comparative evaluation given in Chapter 7.0.

6.3 CURTISS-HIRSCHFELDER PROCEDURE

This chapter concludes with an analysis of the Curtiss-Hirschfelder procedure. Their procedure utilizes the following first-order corrector:

$$y(x_0 + h) = y(x_0) + hy'(x_0 + h). \quad (74)$$

They avoid the use of a predictor by noting that G and ϵ in equation (2a) are functions only of x , rather than of x and y . In this case, after equation (2a) is used to replace the unknown derivative $y'(x_0 + h)$, equation (74) can be solved explicitly for $y(x_0 + h)$. While this is possible with equation (2a), it is generally not possible with the chemical or vibrational rate equations.

Curtiss and Hirschfelder are concerned with both positive and negative values for h/ϵ in their article, even though the exact solution diverges from the asymptote (see Sketch 2) when the value is negative. They assert that the numerical solution of equation (2a) as obtained by their procedure will always converge toward the asymptote for both positive and negative values of h/ϵ . This assertion is based on an intuitive, but incorrect, geometric argument. They then verify this assertion by integrating equation (2a) first with $(h/\epsilon) = 5$ and then with $(h/\epsilon) = -5$. In both cases, the numerical solution does indeed converge to the asymptote. To further check their assertion, we now examine the stability of formula (74) by means of its characteristic polynomial, which is given by

$$\rho = \frac{1}{1 + h/\epsilon} \quad (75)$$

Equation (75) represents, in a $\rho, (h/\epsilon)$ plane, a hyperbola with one branch crossing $\rho = 1$ and the other crossing $\rho = -1$. Since $|\rho| > 1$ for $-2 < h/\epsilon < 0$, formula (74) is unstable in this region. This instability is readily verified by attempting to integrate equation (2a) with a value of h/ϵ in this range. The numerical solution is found to oscillate wildly in this case, thereby contradicting the Curtiss-Hirschfelder assertion. For $h/\epsilon > 0$, the procedure is always stable and the numerical solution does converge to the exact solution, which, after the transient region, is essentially the asymptote. When $h/\epsilon < -2$, the characteristic root is given by the branch of the hyperbola crossing $\rho = -1$. In this situation, the numerical solution converges toward a spurious solution. This spurious solution happens to be the asymptote; the exact solution actually diverges exponentially fast from the asymptote, as noted in Chapter 4.0. The reason that the usual stability condition is not applicable here is that the characteristic polynomial (75) has only one root but two branches. Of the two branches, only the one crossing $\rho = 1$ corresponds to the exact solution.

7.0 CONCLUDING REMARKS

7.1 COMPARATIVE EVALUATION OF THE DIFFERENT INTEGRATION PROCEDURES

Many numerical procedures are available for the integration of systems of ordinary differential equations. Basically, three factors should be considered as criteria for deciding which method is best suited to a particular problem - namely, truncation error, stability, and computation time. The last factor is of special importance when the integration step size is very small in comparison with the total distance to be covered. With reference to these three factors, two different integration procedures can be compared as follows: For a given step size, the higher-order procedure is more precise. When two different procedures are of the same order their truncation errors are

here considered to be approximately equal. Stability is compared on the basis of B , except that $(3/4)B$ is used for a predictor-corrector procedure (see equation (46)). Computation speed is compared by the number of derivative evaluations per integration step and the maximum step size as dictated by stability requirements.

As noted in Chapter 5.0, the fourth-order procedure appears to be the most suitable of the Runge-Kutta procedures for the solution of the nonequilibrium equations. This conclusion is based on the observation that the third-order Runge-Kutta procedure is not as stable or as accurate as the fourth-order procedure, whereas the fifth-order procedure requires too much computation time.

All of the predictor-corrector procedures discussed in Chapter 6.0 require the evaluation of two derivatives per integration step. Thus, these methods differ primarily in the order of their truncation error and in their stability. When these two factors are considered, the method given at the conclusion of Section 6.2 appears to be somewhat superior to the Adams method.

A comparison of the fourth-order Runge-Kutta procedure with the fourth-order predictor-corrector procedure given at the end of Section 6.2, reveals that the Runge-Kutta procedure is more stable ($B = 2.8$ vs. $(3/4)B = 1.36$), but that the predictor-corrector procedure is slightly faster. The latter requires two evaluations of the derivatives per integration step, whereas the fourth-order Runge-Kutta procedure requires six. On an overall basis then, the fourth-order predictor-corrector procedure appears to be slightly superior to the Runge-Kutta procedure.

Despite the foregoing theoretical advantage, the Runge-Kutta procedure was chosen for the numerical calculations given in Part II because its simplicity makes it easy to program for a digital computer. Had the computation time using the Runge-Kutta procedure proven to be unreasonably long, then a predictor-corrector procedure would have been tried. The already programmed Runge-Kutta procedure, however, would still be needed to start this procedure and to compute intervening points if the step size is halved. Since the computation time, when the Runge-Kutta procedure was used, turned out to be acceptably short, no other procedures were attempted.

7.2 CONTROL OF THE INTEGRATION STEP SIZE

It is the purpose of any method that controls the step size h to maximize this quantity while maintaining stability and keeping the truncation error small. Since an estimate of the truncation error for a given step is available, the error condition is readily met. The stability condition requires that h/ϵ be slightly less than the appropriate stability parameter. The value of the quantity ϵ , although considered a constant in the foregoing analysis, actually changes continuously; and as demonstrated in Chapter 3.0, it is not

easily found. It would therefore be desirable to develop a method for controlling h that does not require the evaluation of h/ϵ .

The only readily available quantity for controlling h is the estimated truncation error per step. For example, Runge-Kutta integration generally controls the step size by a technique that attempts to hold constant the estimated error per step. As discussed in Part II, this method results in an erratic variation of step size when the rate equations of Part II are integrated. In terms of maximizing the step size and still maintaining stability, this method is poor.

A method that is based solely on the estimated truncation error for a single step is not feasible, since there is no direct connection between the magnitude of this error and the stability of the integration procedure. It is possible, for example, for a stable procedure to have a large error in a given step or for an unstable procedure to have a small one. The stability of the procedure, however, determines the rate of growth of the error over successive steps. When a procedure is unstable, the error will increase with each succeeding step, whereas the reverse is true when the procedure is stable. The improved technique, given in detail in Part II for controlling the step size when the Runge-Kutta procedure is used is based on this concept.

In this method, three constants R_1 , R_2 , and R_5 are defined by

R_1 = lower reference value for the estimated truncation error,

R_2 = upper reference value for the estimated truncation error,

R_5 = number of integration steps between possible increases in h .

These constants approximately measure the rate of growth of the truncation error. When the estimated truncation error is less than R_1 for R_5 consecutive steps, the procedure is then considered sufficiently stable to allow for a small increase in step size. On the other hand, if the estimated truncation error grows with each succeeding step such that it ultimately exceeds R_2 , where $0 < R_1 \leq R_2$, then the procedure is considered unstable and the step size is reduced by a small amount. The percent increase or decrease in step size is governed by two constants. To avoid going from a stable to an unstable integration condition, the percent increase in h should be small, e.g., 30%. Similarly, to avoid going from a slightly unstable condition to an overly stable condition that would require excessive computation time, the percent decrease in h should also be small, e.g., 30%. As verified by Figure 1 of Part II, the above method approximately maximizes h while at the same time stability is maintained.

REFERENCES

- Chase, P. E., 1962, "Stability Properties of Predictor-Corrector Methods for Ordinary Differential Equations". J.ACM, Vol.9, No.4, p.457.
- Crane, R. L., and Lambert, R. J., 1962, "Stability of a Generalized Corrector Formula". J.ACM, Vol.9, No.1, p.104.
- Curtiss, C. F., and Hirschfelder, J. O., 1952, "Integration of Stiff Equations". Proc. of the Nat. Acad. of Sci., Vol.38, No.3, p.235.
- Dahlquist, G., 1956, "Convergence and Stability in the Numerical Integration of Ordinary Differential Equations". Math. Scand., Vol.4, p.33.
- Dahlquist, G., 1959, "Stability and Error Bounds in the Numerical Integration of Ordinary Differential Equations". Trans. of the Royal Swedish Inst. of Tech., No.130.
- Emanuel, G., and Vincenti, W. G., 1962, "Method for Calculation of the One-Dimensional Nonequilibrium Flow of a General Gas Mixture through a Hypersonic Nozzle". AEDC-TDR-62-131.
- Frazer, R. A., Duncan, W. J., and Collar, A. R., 1947, Elementary Matrices. The Macmillan Co., New York.
- Goldstein, H., 1950, Classical Mechanics. Addison-Wesley Pub.Co., Cambridge, Mass.
- Hamming, R. W., 1962, Numerical Methods for Scientists and Engineers. McGraw-Hill Book Co., New York.
- Hamming, R. W., 1959, "Stable Predictor-Corrector Methods for Ordinary Differential Equations". J.ACM, Vol.6, p.37.
- Henrici, P., 1962, Discrete Variable Methods in Ordinary Differential Equations. John Wiley and Sons, New York.
- Hildebrand, F. B., 1956, Introduction to Numerical Analysis. McGraw-Hill Book Co., New York.
- Ince, E. L., 1926, Ordinary Differential Equations. Dover Pub. Inc., New York.
- Lotkin, M., 1951, "On the Accuracy of Runge-Kutta's Method". Math. Tables and other Aids to Computation, Vol.5, p.128.
- Marcus, M., 1960, "Basic Theorems in Matrix Theory". Nat. Bur. of Stds., Appl. Math. Ser. No.57.

Nordsieck, A., 1962, "On Numerical Integration of Ordinary Differential Equations". Math. of Computation, Vol.16, No.77, p.22.

Ralston, A., and Wilf, H. S., (editors), 1960, Mathematical Methods for Digital Computers. John Wiley and Sons, New York.

Vincenti, W. G., 1959, "Non-equilibrium Flow Over a Wavy Wall". J.Fluid Mech., Vol.6, p.481.

Vincenti, W. G., 1961, "Calculations of the One-Dimensional Nonequilibrium Flow of Air Through a Hypersonic Nozzle - Interim Report". AEDC-TN-61-65.

Wilf, H. S., 1959, "A Stability Criterion for Numerical Integration". J.ACM, Vol.6, p.363.

APPENDIX I

PROOF THAT ALL OF THE EIGENVALUES κ_i ARE REAL AND POSITIVE

This appendix proves the following fundamental theorem referred to in Chapter 3.0:

Theorem: All eigenvalues κ_i ($i=1, \dots, N_3$), given by the chemical characteristic equation

$$|\delta_{kl}\kappa - \kappa_{kl}| = 0, \quad (24b)$$

are real and positive.

A sequence of five lemmas is used to establish this theorem. A few preliminary remarks concerning the nature of the proof will be helpful.

Only two physical concepts are required in the proof. First, the equations that express conservation of components

$$n_i = a_{i0} + \sum_{j=1}^{N_3} a_{ij}n_j, \quad i = N_3+1, \dots, N_1, \quad (I-1)$$

are necessary. Second, the structure of the chemical rate equations (as given in Chapter 3.0), when the flow is close to equilibrium, is also involved. Thus, the proof does not require conservation of momentum or of energy. The theorem is therefore valid for inviscid or viscous flows as well as non-radiating or radiating flows. Likewise, the second law of thermodynamics is not required in the proof. The theorem as given here, is not applicable when certain physical processes that alter the form of the chemical rate equations, such as diffusion, are to be considered.

Two matrices, $((\bar{\kappa}_{kl}))$ and $((\kappa_{kl}))$, whose elements are $\bar{\kappa}_{kl}$ and κ_{kl} (see equations (16) and (18)), are used throughout the proof. The first is an N_3 by N_3 matrix, while the second is an N_1 by N_1 matrix. Although the theorem is directly concerned only with the N_3 eigenvalues of $((\kappa_{kl}))^*$ as determined by equation (24b), most of the proof actually deals with $((\bar{\kappa}_{kl}))$. The basic reason for this is the greater simplicity of the elements of $((\bar{\kappa}_{kl}))$ as compared with those of $((\kappa_{kl}))$.

* We here follow standard matrix theory terminology by referring to the eigenvalues determined by equation (24b) as belonging to the matrix $((\kappa_{kl}))$.

The five lemmas can be briefly summarized as follows: Lemma 1 establishes two useful relations necessary in the subsequent development. Lemma 2 establishes the connection between eigenvalues of $((\kappa_{kl}))$ and those of $((\bar{\kappa}_{kl}))$. Lemma 3 demonstrates that $((\kappa_{kl}))$ has N_3 non-zero eigenvalues. Lemmas 4 and 5 are used to prove that $((\kappa_{kl}))$ is positive semidefinite.* Once these lemmas are established, the actual proof of the theorem then follows readily. Following this, the appendix concludes with the demonstration of two corollaries that help clarify the preceding proof. The first of these demonstrates that it is immaterial which side of a reaction is taken to refer to the reactants and which to the products. The second shows that the eigenvalues of $((\kappa_{kl}))$ do not depend on the specific choice of the mole-mass ratios n_i ($i=N_3+1, \dots, N_1$) that are taken to be given by the conservation-of-components equation (I-1). Both of these corollaries could have been anticipated from purely physical reasoning.

Lemma 1:

(a) The quantities $\bar{\kappa}_{il}$ are given by

$$\bar{\kappa}_{il} = \frac{1}{n_l^*} \sum_{r=1}^{N_2} \frac{v_{ri} v_{rl}}{\theta_r^*}, \quad i = 1, \dots, N_1, \quad l = 1, \dots, N_1. \quad (\text{I-2})$$

(b) The relations

$$\bar{\kappa}_{il} = \sum_{j=1}^{N_3} a_{ij} \bar{\kappa}_{jl}, \quad i = N_3+1, \dots, N_1, \quad l = 1, \dots, N_1 \quad (\text{I-3})$$

are a consequence of the equations expressing conservation of components (I-1).

* A matrix is referred to as positive semidefinite if all of its eigenvalues are real and non-negative. If all the eigenvalues of a matrix are real and positive, i.e., none are zero valued, the matrix is called positive definite.

Proof: Differentiating equations (6b) gives

$$\begin{aligned} \frac{\partial L_r}{\partial n_l} &= - \frac{(\rho) \beta_r - \alpha_r}{K_{cr}(T)} \prod_{k=1}^{N_1} (n_k)^{v_{rk}} \frac{v_{rl}}{n_l} , \\ &= (L_r - 1) \frac{v_{rl}}{n_l} , \quad r = 1, \dots, N_2, \quad l = 1, \dots, N_1 , \quad (I-4) \end{aligned}$$

where ρ, T , and all n_k , except n_l , are held fixed in accordance with equations (11). Since $(L_r)^* = 0$ (see equation (8a)), the reference value for $\partial L_r / \partial n_l$ is therefore given by

$$\left(\frac{\partial L_r}{\partial n_l} \right)^* = - \frac{v_{rl}}{n_l^*} , \quad r = 1, \dots, N_2, \quad l = 1, \dots, N_1 . \quad (I-5)$$

Substituting equation (I-5) into equations (16) results in equations (I-2), thereby establishing part (a) of the lemma. It is important to note that equations (I-5) are valid only if the reference state is given by the equilibrium-flow solution or the local equilibrium value.

Part (b) is proved by differentiating equation (I-1) and then replacing Dn_j/Dt by equations (15b) as follows:

$$\begin{aligned} \frac{Dn_i}{Dt} &= \sum_{j=1}^{N_3} a_{ij} \frac{Dn_j}{Dt} , \\ &= \sum_{j=1}^{N_3} a_{ij} \sum_{l=1}^{N_1} \bar{\kappa}_{jl} (n_l, e^{-n_l}) , \\ &= \sum_{l=1}^{N_1} \left\{ \sum_{j=1}^{N_3} a_{ij} \bar{\kappa}_{jl} \right\} (n_l, e^{-n_l}) , \quad i = N_3+1, \dots, N_1 . \quad (I-6) \end{aligned}$$

The i and j summations have been interchanged in the last step. By comparing this result with equations (15b), we obtain equations (I-3). Thus, the lemma is proved.

Lemma 2 will prove that any eigenvalue of $((\kappa_{k\ell}))$ is also an eigenvalue of $((\bar{\kappa}_{k\ell}))$. In addition, the lemma will show that any eigenvalue of $((\bar{\kappa}_{k\ell}))$ that is not an eigenvalue of $((\kappa_{k\ell}))$, is zero. These two results are established by demonstrating that $((\kappa_{k\ell}))$ and $((\bar{\kappa}_{k\ell}))$, except for the factor $(\kappa)^{N_1-N_3}$, have identical characteristic equations.

Lemma 2: The relation

$$|\delta_{k\ell}\kappa - \bar{\kappa}_{k\ell}| = (\kappa)^{N_1-N_3} |\delta_{k\ell}\kappa - \kappa_{k\ell}| \quad (I-7)$$

is valid.

Proof: The left-hand side of equation (I-7) is a polynomial in κ of degree N_1 . This polynomial is denoted by $P = P(\kappa)$, and when explicitly written out is given by

$$P(\kappa) = \begin{vmatrix} \kappa - \bar{\kappa}_{11} & -\bar{\kappa}_{12} & -\bar{\kappa}_{13} & \dots & -\bar{\kappa}_{1N_1} \\ -\bar{\kappa}_{21} & \kappa - \bar{\kappa}_{22} & -\bar{\kappa}_{23} & \dots & -\bar{\kappa}_{2N_1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\bar{\kappa}_{N_1 1} & -\bar{\kappa}_{N_1 2} & \dots & \dots & \kappa - \bar{\kappa}_{N_1 N_1} \end{vmatrix} \quad (I-8a)$$

First, column N_3+1 of P is multiplied by $a_{N_3+1,1}$ and the result is added to column 1. Commas are now used between certain double subscripts such as N_3+1 and 1, solely for clarity. Next, column N_3+2 is multiplied by $a_{N_3+2,1}$ and again the result is added to column 1. By repeating this process, we finally obtain for the first column of P

$$\left. \begin{aligned}
 \kappa &= \kappa_{11} \\
 &= \kappa_{21} \\
 &\vdots \\
 &= \kappa_{N_3,1} \\
 &= \bar{\kappa}_{N_3+1,1} + a_{N_3+1,1} \kappa - \sum_{j=N_3+1}^{N_1} a_{j,1} \bar{\kappa}_{N_3+1,j} \\
 &= \bar{\kappa}_{N_3+2,1} + a_{N_3+2,1} \kappa - \sum_{j=N_3+1}^{N_1} a_{j,1} \bar{\kappa}_{N_3+2,j} \\
 &\vdots \\
 &= \bar{\kappa}_{N_1,1} + a_{N_1,1} \kappa - \sum_{j=N_3+1}^{N_1} a_{j,1} \bar{\kappa}_{N_1,j}
 \end{aligned} \right\} \quad (I-9)$$

where equations (18) are used to obtain the top N_3 elements. Column 2 is next altered in a similar manner, i.e., by multiplying column N_3+1 of P by $a_{N_3+1,2}$ and adding the result to column 2 of P , etc. After column 2 the process is repeated for columns 3 through N_3 of P , thereby resulting in

P =

$$\begin{array}{c}
 \begin{array}{c}
 \kappa - \kappa_{11} \\
 - \kappa_{21} \\
 \vdots \\
 - \kappa_{N_3,1}
 \end{array}
 \begin{array}{c}
 \kappa - \kappa_{12} \\
 \kappa - \kappa_{22} \\
 \vdots \\
 - \kappa_{N_3,2}
 \end{array}
 \begin{array}{c}
 \cdot \\
 \cdot \\
 \cdot \\
 \cdot
 \end{array}
 \begin{array}{c}
 - \bar{\kappa}_{1,N_1} \\
 \cdot \\
 \cdot \\
 \cdot
 \end{array}
 \end{array}
 \begin{array}{c}
 - \bar{\kappa}_{N_3+1,1} + a_{N_3+1,1} \kappa - \sum_{j=N_3+1}^{N_1} a_{j,1} \bar{\kappa}_{N_3+1,j} \\
 \cdot \\
 \cdot \\
 \cdot
 \end{array}
 \begin{array}{c}
 - \bar{\kappa}_{N_3+1,2} + a_{N_3+1,2} \kappa - \sum_{j=N_3+1}^{N_1} a_{j,2} \bar{\kappa}_{N_3+1,j} \dots - \bar{\kappa}_{N_3+1,N_1} \\
 \cdot \\
 \cdot \\
 \cdot
 \end{array}
 \end{array}
 \cdot (I-8b)$$

$$\begin{array}{c}
 - \bar{\kappa}_{N_1,1} + a_{N_1,1} \kappa - \sum_{j=N_3+1}^{N_1} a_{j,1} \bar{\kappa}_{N_1,j} \\
 \cdot \\
 \cdot \\
 \cdot
 \end{array}
 \begin{array}{c}
 - \bar{\kappa}_{N_1,2} + a_{N_1,2} \kappa - \sum_{j=N_3+1}^{N_1} a_{j,2} \bar{\kappa}_{N_1,j} \dots \kappa - \bar{\kappa}_{N_1,N_1}
 \end{array}$$

Columns $N_3 + 1$ through N_1 of (I-8b), of course, are identical with the corresponding columns of (I-8a). The top N_3 rows of (I-8b) are now multiplied by $-a_{N_3+1,k}$ ($k=1, \dots, N_3$) and the result added to the $N_3 + 1$ row of (I-8b). The first element in this row then becomes

$$\begin{aligned}
 & -\bar{\kappa}_{N_3+1,1} + \cancel{a_{N_3+1,1} \bar{\kappa}_{N_3+1,1}} - \sum_{j=N_3+1}^{N_1} a_{j,1} \bar{\kappa}_{N_3+1,j} \cancel{a_{N_3+1,1} \bar{\kappa}_{N_3+1,1}} + a_{N_3+1,1} \kappa_{11} \\
 & + a_{N_3+1,2} \kappa_{21} + \dots + a_{N_3+1,N_3} \kappa_{N_3,1} \\
 & = -\bar{\kappa}_{N_3+1,1} - \sum_{j=N_3+1}^{N_1} a_{j,1} \bar{\kappa}_{N_3+1,j} + \sum_{k=1}^{N_3} a_{N_3+1,k} \kappa_{k1} \quad (I-10a)
 \end{aligned}$$

When κ_{k1} is replaced by equations (18), and the result is regrouped, we obtain

$$- \left(\bar{\kappa}_{N_3+1,1} - \sum_{k=1}^{N_3} a_{N_3+1,k} \bar{\kappa}_{k1} \right) - \sum_{j=N_3+1}^{N_1} a_{j,1} \left(\bar{\kappa}_{N_3+1,j} - \sum_{k=1}^{N_3} a_{N_3+1,k} \bar{\kappa}_{kj} \right) \quad (I-10b)$$

According to equations (I-3), each term enclosed by parentheses in (I-10b) is zero. Consequently, the first element of row $N_3 + 1$ of (I-8b) is zero. In a similar fashion, each element from the first to the $(N_3)^{\text{th}}$ of this row of (I-8b) is shown to be zero. The $N_3 + 1$ element of this row becomes

$$\kappa = \bar{\kappa}_{N_3+1,N_3+1} + \sum_{k=1}^{N_3} a_{N_3+1,k} \bar{\kappa}_{k,N_3+1}$$

which, by virtue of equations (I-3), is equal to κ . In a similar fashion, the remaining elements of the N_3+1 row of (I-8b) are shown to be zero. Hence, this row contains only zero elements, except for the diagonal element, which is κ . The argument for the other rows of (I-8b) from row N_3+2 through row N_1 is the same. Thus, determinant (I-8b) can be written as follows:

$$P = \begin{vmatrix} \kappa - \kappa_{11} & -\kappa_{12} & \dots & -\kappa_{1,N_3} & -\bar{\kappa}_{1,N_3+1} & \dots & -\bar{\kappa}_{1,N_1} \\ -\kappa_{21} & \kappa - \kappa_{22} & & & & & \\ \vdots & \vdots & & & & & \\ -\kappa_{N_3,1} & -\kappa_{N_3,2} & \dots & \kappa - \kappa_{N_3,N_3} & -\bar{\kappa}_{N_3,N_3+1} & \dots & -\bar{\kappa}_{N_3,N_1} \\ 0 & 0 & \dots & 0 & \kappa & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots & 0 & \kappa & & \vdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & \kappa \end{vmatrix} .$$

(I-8c)

This determinant, however, is precisely equal to

$$(\kappa)^{N_1-N_3} |\delta_{kl}\kappa - \kappa_{kl}| ,$$

which proves the lemma.

The purpose of the next lemma is to show that $((\kappa_{kl}))$ has N_3 non-zero eigenvalues. Since $((\kappa_{kl}))$ is an N_3 by N_3 matrix, it is sufficient to prove that the rank of $((\kappa_{kl}))$ is N_3 .^{*} Implicit in the proof of this lemma is the assumption that $N_2 \geq N_3$. As noted in Part II, an immediate consequence of this assumption is that the minimum number of chemical rate equations is N_3 .

^{*} If the rank of $((\kappa_{kl}))$ is N_3 , then the determinant of $((\kappa_{kl}))$ is not zero and the matrix has no zero eigenvalues. If, however, the rank of $((\kappa_{kl}))$ is less than N_3 , then the determinant of $((\kappa_{kl}))$ is zero and the matrix has at least one zero eigenvalue.

Lemma 3: The rank of $((\kappa_{kl}))$ is N_3 .

Proof: If the rank of $((\kappa_{kl}))$ were less than N_3 , say $N_3 - 1$, then there would exist a non-zero constant vector $\{c_i\}$ ($i=1, \dots, N_3$) such that

$$\sum_{i=1}^{N_3} c_i \kappa_{il} = 0, \quad l = 1, \dots, N_3. \quad (I-11)$$

Equations (17) could thus be transformed as follows:

$$\sum_{i=1}^{N_3} c_i \frac{Dn_i}{Dt} = \sum_{l=1}^{N_3} (n_l e^{-n_l}) \sum_{i=1}^{N_3} c_i \kappa_{il}. \quad (I-12a)$$

The terms $\sum_{i=1}^{N_3} c_i \kappa_{il}$ are equal to zero as a result of equations (I-11).

Equation (I-12a) could therefore be integrated to yield

$$\sum_{i=1}^{N_3} c_i n_i = \text{constant}. \quad (I-12b)$$

This result is not possible, however, since equations (17) represent the minimum possible number of rate equations. Consequently, the rank of $((\kappa_{kl}))$ must be N_3 and the lemma is established.

The remaining two lemmas are used to show that $((\bar{\kappa}_{kl}))$ is positive semidefinite. A sequence of determinants, given by

$$\bar{K}_m^{(o)} = \begin{vmatrix} \bar{\kappa}_{11} & \bar{\kappa}_{12} & \dots & \bar{\kappa}_{1m} \\ \bar{\kappa}_{21} & & & \vdots \\ \vdots & & & \vdots \\ \bar{\kappa}_{m1} & \dots & \dots & \bar{\kappa}_{mm} \end{vmatrix}, \quad m = 1, \dots, N_1, \quad (I-13)$$

forms the starting point for the proof. Basic to the proof is the following theorem from matrix theory: An N_1 by N_1 matrix $((\bar{\kappa}_{kl}))$ is positive semidefinite if, and only if,

$$\bar{K}_m^{(0)} \geq 0, \quad m = 1, \dots, N_1. \quad (I-14)$$

The above theorem is usually proved for positive definite matrices. In this case, the symbol $(\geq)$ in conditions (I-14) should be changed to the symbol $(>)$. A proof of this theorem can be found in Frazer, Duncan, and Collar (1947).

Our ultimate purpose is to show that $((\kappa_{kl}))$ is positive definite. This goal will be achieved, in part, by proving $((\bar{\kappa}_{kl}))$ positive semidefinite. This round-about approach is undertaken because a direct proof that $((\kappa_{kl}))$ is positive definite would be considerably more difficult owing to the complicated structure of the elements of $((\kappa_{kl}))$.

The actual structure of the chemical rate equations, when the flow is close to equilibrium, is embodied in the quantities $\bar{\kappa}_{kl}$, as given by equations (I-2). This structure enters in an essential way into the next two lemmas. To put it differently, $((\bar{\kappa}_{kl}))$ is positive semidefinite because of this structure.

Before starting the next lemma, we will define certain quantities that are necessary for its formulation. When the $\bar{\kappa}_{kl}$, given by equation (I-2), are substituted into equation (I-13), the elements of each column are seen to contain the factor $1/n_l^*$. Consequently, a new sequence of determinants is defined as follows:

$$\bar{K}_m^{(1)} = \left(\prod_{l=1}^m n_l^* \right) \bar{K}_m^{(0)}, \quad (I-15a)$$

$$= \begin{vmatrix} \sum_{r=1}^{N_2} \frac{v_{r1} v_{r1}}{\theta_r^*} & \dots & \sum_{r=1}^{N_2} \frac{v_{r1} v_{rm}}{\theta_r^*} \\ \vdots & & \vdots \\ \sum_{r=1}^{N_2} \frac{v_{rm} v_{r1}}{\theta_r^*} & \dots & \sum_{r=1}^{N_2} \frac{v_{rm} v_{rm}}{\theta_r^*} \end{vmatrix}, \quad m=1, \dots, N_1. \quad (I-15b)$$

When the various elements of determinant (I-15b) are multiplied against each other, in accord with the usual rules for computing the value of a determinant, certain quantities, such as (in this illustration $m=2$)

$$\frac{1}{(\theta_1^*)^2}, \frac{1}{\theta_1^* \theta_2^*}, \frac{1}{\theta_2^* \theta_1^*}, \frac{1}{\theta_1^* \theta_3^*}, \dots,$$

will regularly appear. These quantities are here considered to be the variables upon which $\bar{K}_m^{(1)}$ depends. The coefficients of these variables are readily seen to be functions only of the constants v_{r1} . In order to introduce a notation for these variables, a subscripted r, r_s , is now used to denote any value of r from 1 to N_2 . For a fixed m , the above variables are defined by

$$H(m; r_1, r_2, \dots, r_m) = \frac{1}{\prod_{s=1}^m (\theta_{r_s}^*)^*}, \quad (\text{I-16})$$

where the dependence of H on the parameters m and $r_1, \dots, r_m$ is explicitly indicated. For example, with $m=2$, $r_1=5$, and $r_2=1$, H is given by

$$H(2; 5, 1) = H(2; 1, 5) = \frac{1}{\theta_1^* \theta_5^*}. \quad (\text{I-17})$$

For reasons that will become apparent in the proof of Lemma 4, the definition of H given above is restricted to prevent any θ_r^* from appearing more than once. Thus, in addition to definition (I-16), we also require that if $s \neq t$ then

$$r_s \neq r_t.$$

Different H 's will be distinguished by subscripts. Therefore, $H_1(m; r_1^{(1)}, \dots, r_m^{(1)})$ must have at least one $r_s^{(1)}$ value that is different from all the $r_s^{(2)}$ values in $H_2(m; r_1^{(2)}, \dots, r_m^{(2)})$.

Lemma 4: Each m by m determinant $\bar{K}_m^{(1)}$ can be written as

$$\bar{K}_m^{(1)} = \sum_{i=1}^{\binom{N_2}{m}} A_i H_i(m; r_1^{(1)}, \dots, r_m^{(1)}) , \quad m = 1, \dots, N_1 . \quad (I-18)$$

where

$$\binom{N_2}{m} \equiv \frac{N_2!}{(N_2 - m)! m!} , \quad (I-19)$$

and the constants $A_i = A_i(m; r_1^{(1)}, \dots, r_m^{(1)})$ depend only on the parameters v_{rj} .

A simple illustration of Lemma 4 is provided when $m = 1$. In this case, we have

$$\bar{K}_1^{(1)} = n_1^* \bar{K}_1^{(0)} = \sum_{r=1}^{N_2} \frac{v_{r1}^2}{\theta_r^*} , \quad (I-20)$$

where $\binom{N_2}{1} = N_2, H_1(1; r_1^{(1)}) = (\theta_1^*)^{-1}$, and $A_1 = v_{11}^2$.

Proof: Each element of $\bar{K}_m^{(1)}$ contains N_2 terms, according to determinant (I-15b). Therefore, using the determinant rule

$$\begin{vmatrix} a_1 + a_2 & b \\ c_1 + c_2 & d \end{vmatrix} = \begin{vmatrix} a_1 & b \\ c_1 & d \end{vmatrix} + \begin{vmatrix} a_2 & b \\ c_2 & d \end{vmatrix} ,$$

we can write $\bar{K}_m^{(1)}$ as the sum of $(N_2)^m$ determinants. Each of these is an m by m determinant, whose elements have the form

$$\frac{v_{r1} v_{rj}}{\theta_r^*} . \quad (I-21)$$

Since every element in a particular column contains v_{rj}/θ_r^* , this factor may be taken outside of the determinant. Thus, each of the $(N_2)^m$ determinants is of the form

$$\left(\prod_{s=1}^m \frac{\theta_{r_s}^*}{r_s} \right) \begin{vmatrix} v_{r_1,1} & v_{r_2,1} & \cdots & v_{r_m,1} \\ \vdots & \vdots & & \vdots \\ v_{r_1,m} & v_{r_2,m} & \cdots & v_{r_m,m} \end{vmatrix} \quad (I-22)$$

Note that if $r_s = r_t$, $s \neq t$, the determinant in expression (I-22) is zero since two columns are equal. Hence, only those determinants for which $r_s \neq r_t$ when $s \neq t$ need be considered. In addition, if $m > N_2$ then at least two columns must be identical in each of the $(N_2)^m$ determinants. Consequently, we have

$$\bar{K}_m^{(1)} = 0, \quad (I-23)$$

for all $m > N_2$.

According to the foregoing, $\bar{K}_m^{(1)}$ is a linear function of the variables

$$\frac{1}{\prod_{s=1}^m \theta_{r_s}^*} \quad (I-24)$$

where $r_s \neq r_t$ when $s \neq t$. Each of the products in expression (I-24) therefore contains m distinct θ_r^* selected from the N_2 distinct θ_r^* . Thus the number of distinct variables of the form (I-24) is given by the binomial coefficient $\binom{N_2}{m}$. Each of these was defined earlier as an $H_1(m; r_1^{(1)}, \dots, r_m^{(1)})$. Therefore, $\bar{K}_m^{(1)}$ can be written as equation (I-18), thereby proving the lemma.

Lemma 5: The constants A_i are given by

$$A_i(m; r_1^{(1)}, \dots, r_m^{(1)}) = \left\{ \begin{vmatrix} (v)_{r_1^{(1)}, 1} & \dots & (v)_{r_m^{(1)}, 1} \\ \vdots & & \vdots \\ (v)_{r_1^{(1)}, m} & \dots & (v)_{r_m^{(1)}, m} \end{vmatrix} \right\}^2, \quad i = 1, \dots, \binom{N_2}{m}. \quad (\text{I-25})$$

Proof: The number of permutations of the m distinct numbers $r_1^{(1)}, \dots, r_m^{(1)}$ is $m!$. Consequently, the variable H_1 will occur $m!$ times in the original expansion of $\bar{K}_m^{(1)}$ into $(N_2)^m$ determinants. For example, with $N_2 = 3$ and $m = 2$, each of the $(N_2)^m = (3)^2 = 9$ determinants in the $\bar{K}_2^{(1)}$ expansion is multiplied by one of the following variables:

$$\frac{1}{(\theta_1^*)^2}, \frac{1}{(\theta_2^*)^2}, \frac{1}{(\theta_3^*)^2}, \frac{1}{\theta_1^* \theta_2^*}, \frac{1}{\theta_2^* \theta_1^*}, \frac{1}{\theta_1^* \theta_3^*}, \frac{1}{\theta_3^* \theta_1^*}, \frac{1}{\theta_2^* \theta_3^*}, \frac{1}{\theta_3^* \theta_2^*}.$$

As we have already seen, the coefficients of the $(\theta_1^*)^{-2}$ variables are zero. There are then $m! = 2! = 2$ variables, such as $(\theta_1^* \theta_2^*)^{-1}$ and $(\theta_2^* \theta_1^*)^{-1}$, each of which is designated by the same H_1 .

According to expression (I-22), each of the $m!$ terms in the coefficient of a particular H_1 is of the general form

$$\left(\prod_{s=1}^m (v)_{r_s^{(1)}, s} \right) \begin{vmatrix} (v)_{r_1^{(1)}, 1} & \dots & (v)_{r_m^{(1)}, 1} \\ \vdots & & \vdots \\ (v)_{r_1^{(1)}, m} & \dots & (v)_{r_m^{(1)}, m} \end{vmatrix}. \quad (\text{I-26})$$

With H_1 fixed, we now examine the $m!$ determinants appearing in the coefficient of this H_1 . Each of these determinants contains the same columns but in a different sequence. In fact, no two of the $m!$ determinants have their columns in precisely the same sequence. Since interchanging columns in a determinant simply changes its sign, all $m!$ determinants are equal to

$$(\text{sgn } \sigma) \begin{vmatrix} (v)_{r_1^{(1)},1} & \dots & (v)_{r_m^{(1)},1} \\ \vdots & & \vdots \\ (v)_{r_1^{(1)},m} & \dots & (v)_{r_m^{(1)},m} \end{vmatrix}, \quad (\text{I-27})$$

where $\text{sgn } \sigma$ is 1 or -1 depending on whether the permutation σ is even or odd. One column interchange represents an odd permutation, two an even permutation, etc.

The constant A_1 is then obtained by summing the $m!$ terms in the coefficient of H_1 as follows:

$$A_1 = \sum_{m!} \left(\prod_{s=1}^m (v)_{r_s^{(1)},s} \right) (\text{sgn } \sigma) \begin{vmatrix} (v)_{r_1^{(1)},1} & \dots & (v)_{r_m^{(1)},1} \\ \vdots & & \vdots \\ (v)_{r_1^{(1)},m} & \dots & (v)_{r_m^{(1)},m} \end{vmatrix}, \quad i = 1, \dots, \binom{N_2}{m}, \quad (\text{I-28a})$$

where the summation is over all $m!$ permutations of $r_1^{(1)}, \dots, r_m^{(1)}$. After the determinant, which is a common factor, is taken outside the summation, A_1 becomes

$$A_1 = \left| \begin{array}{cccc} (v)_{r_1^{(1)},1}^{(1)} & \cdots & (v)_{r_m^{(1)},1}^{(1)} \\ \vdots & & \vdots \\ (v)_{r_1^{(1)},m}^{(1)} & \cdots & (v)_{r_m^{(1)},m}^{(1)} \end{array} \right| \sum_{m!} (\text{sgn } \sigma) \prod_{s=1}^m (v)_{r_s^{(1)},s}^{(1)}$$

$$i = 1, \dots, \binom{N_2}{m} \quad (\text{I-28b})$$

According to the usual definition of a determinant (see Marcus (1960)), the summation is given by

$$\sum_{m!} (\text{sgn } \sigma) \prod_{s=1}^m (v)_{r_s^{(1)},s}^{(1)} = \left| \begin{array}{cccc} (v)_{r_1^{(1)},1}^{(1)} & \cdots & (v)_{r_m^{(1)},1}^{(1)} \\ \vdots & & \vdots \\ (v)_{r_1^{(1)},m}^{(1)} & \cdots & (v)_{r_m^{(1)},m}^{(1)} \end{array} \right| \quad (\text{I-29})$$

Consequently, A_1 is given by equation (I-25), thus proving the lemma.

A simple illustration of Lemma 5 is provided by the earlier example when $N_2 = 3$ and $m = 2$. The six non-vanishing determinants in the expansion of $\tilde{K}_2^{(1)}$ are then treated as follows:

$$\begin{aligned} \bar{\kappa}_2^{(1)} &= \begin{vmatrix} \frac{v_{11}^2}{\theta_1^*} & \frac{v_{21}v_{22}}{\theta_2^*} \\ \frac{v_{11}v_{12}}{\theta_1^*} & \frac{v_{22}^2}{\theta_2^*} \end{vmatrix} + \begin{vmatrix} \frac{v_{21}^2}{\theta_2^*} & \frac{v_{11}v_{12}}{\theta_1^*} \\ \frac{v_{21}v_{22}}{\theta_2^*} & \frac{v_{12}^2}{\theta_1^*} \end{vmatrix} + \begin{vmatrix} \frac{v_{11}^2}{\theta_1^*} & \frac{v_{31}v_{32}}{\theta_3^*} \\ \frac{v_{11}v_{12}}{\theta_1^*} & \frac{v_{32}^2}{\theta_3^*} \end{vmatrix} + \dots \\ \bar{\kappa}_2^{(1)} &= \frac{v_{11}v_{22}}{\theta_1^*\theta_2^*} \begin{vmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \end{vmatrix} + \frac{v_{21}v_{12}}{\theta_2^*\theta_1^*} \begin{vmatrix} v_{21} & v_{11} \\ v_{22} & v_{12} \end{vmatrix} + \frac{v_{11}v_{32}}{\theta_1^*\theta_3^*} \begin{vmatrix} v_{11} & v_{31} \\ v_{12} & v_{32} \end{vmatrix} + \dots \\ \bar{\kappa}_2^{(1)} &= \frac{(v_{11}v_{22} - v_{21}v_{12})^2}{\theta_1^*\theta_2^*} + \frac{(v_{11}v_{32} - v_{12}v_{31})^2}{\theta_1^*\theta_3^*} + \frac{(v_{21}v_{32} - v_{22}v_{31})^2}{\theta_2^*\theta_3^*} . \end{aligned} \quad (\text{I-30})$$

A second illustration of the lemma is provided by $\bar{\kappa}_1^{(1)}$, given by equation (I-20).

As a result of Lemmas 4 and 5, the proof that $((\bar{\kappa}_{kl}))$ is positive semidefinite is simple. Since all H_i and A_i are non-negative, all $\bar{\kappa}_m^{(1)}$ are non-negative. Equations (I-15a) then imply that all $\bar{\kappa}_m^{(0)}$ are non-negative. Conditions (I-14) then state that $((\bar{\kappa}_{kl}))$ is positive semidefinite, i.e., all its eigenvalues are real and non-negative.

The proof of the theorem stated at the outset of this Appendix is now also simple. The preceding paragraph demonstrated that all eigenvalues of $((\bar{\kappa}_{kl}))$ are real and non-negative. According to Lemma 2, however, all eigenvalues of $((\kappa_{kl}))$ are eigenvalues of $((\bar{\kappa}_{kl}))$. Thus, all eigenvalues of $((\kappa_{kl}))$ are real and non-negative. Lemma 3, however, proved that all N_3 eigenvalues of $((\kappa_{kl}))$ are non-zero. Consequently, all eigenvalues of $((\kappa_{kl}))$ are real and positive, thereby proving the theorem.

Two corollaries of the theorem are now given. Corollary 1 is primarily a consequence of Lemma 1.

Corollary 1: The specific values for the N_3 eigenvalues κ_1 are unchanged when the forward and backward directions of any reaction are interchanged.

Proof: All that is necessary is to show that the $\bar{\kappa}_{k\ell}$ are unaffected when the two sides of a reaction are interchanged. For convenience, assume that reaction 1 is so altered, while the remaining reactions are left unchanged. All v_{11} must then be replaced by $-v_{11}$. Also, according to equations (6c),

$$(\theta_1^*)^{-1} = (\rho^*)^{\alpha_1-1} k_{f1}(T^*) \prod_{k=1}^{N_4} (n_k^*)^{\alpha_{1k}} \quad (I-31)$$

Noting that $K_{c1}(T^*) = (k_{f1}(T^*)/k_{b1}(T^*))$ and $L_1(\rho^*, T^*, n_i^*) = 0$, (see equations (8a)), we can write equations (6b) for $r = 1$ as follows:

$$1 = \frac{(\rho^*)^{\beta_1-\alpha_1}}{K_{c1}(T^*)} \prod_{k=1}^{N_1} (n_k^*)^{v_{1k}} \quad (I-32a)$$

Multiply both sides of equation (I-32a) by

$$\prod_{k=N_1+1}^{N_4} (n_k^*)^{v_{1k}} = 1, \quad (I-33)$$

since $v_{1k} = 0$ for $k = N_1+1, \dots, N_4$, and then rearrange the results to obtain

$$(\rho^*)^{\alpha_1-1} k_{f1}(T^*) \prod_{k=1}^{N_4} (n_k^*)^{\alpha_{1k}} = (\rho^*)^{\beta_1-1} k_{b1}(T^*) \prod_{k=1}^{N_4} (n_k^*)^{\beta_{1k}} \quad (I-32b)$$

By comparing equations (I-31) and (I-32b), we observe that $(\theta_1^*)^{-1}$ is unaltered when the two sides of reaction 1 are interchanged. When these changes are introduced into the $r = 1$ term of $\kappa_{k\ell}$, as given by equations (I-2), we see that $\bar{\kappa}_{k\ell}$ remains unchanged.

Corollary 2: The specific values for the N_3 eigenvalues κ_1 do not depend on the particular choice of mole-mass ratios n_i ($i=N_3+1, \dots, N_1$) taken to be given by the conservation-of-component equations (I-1).

As an illustration of Corollary 2, consider the model for air given in Part II, where the mole-mass ratios n_i represent O, N, NO, N_2 , and O_2 , respectively. From the two equations,

$$\left. \begin{aligned} n_1 + n_3 + 2n_5 &= \text{constant} \\ n_2 + n_3 + 2n_4 &= \text{constant} \end{aligned} \right\}, \quad (\text{I-34})$$

that express conservation of atomic oxygen and atomic nitrogen, respectively, eight different sets of equations, each set containing two equations of the form (I-1), are equally possible. The corollary then asserts that the eigenvalues are the same no matter which of the eight sets of equations (I-1) is chosen. While the eigenvalues are invariant with respect to this choice, the coefficients B_{ik} in the linear transformation (19a) do depend on the choice.

Proof: Corollary 2 is established by noting that all eigenvalues of $((\kappa_{kl}))$ are also eigenvalues of $((\bar{\kappa}_{kl}))$, as proven by Lemma 2. The eigenvalues of $((\bar{\kappa}_{kl}))$, however, do not depend on the conservation of component equations (I-1). Consequently, the corollary is proven.

The eigenvalues can be computed either by solving the N_3 by N_3 determinant equation (24b), or by solving the N_1 by N_1 determinant equation

$$|\delta_{kl}\kappa - \bar{\kappa}_{kl}| = 0. \quad (\text{I-35})$$

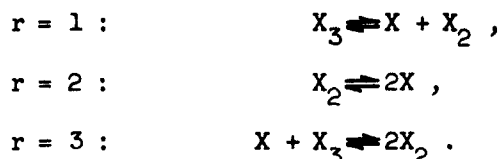
This latter equation yields the correct N_3 non-zero eigenvalues plus $N_1 - N_3$ zero eigenvalues. While the elements of equation (I-35) are considerably simpler than those of equation (24b), the higher order of equation (I-35) more than offsets this advantage. Thus, for computational purposes equation (24b) is superior. This is especially true when the eigenvalues are to be computed with a digital computer using a standard eigenvalue routine. These routines sometimes have convergence difficulties when there are multiple zero eigenvalues.

APPENDIX II

EXAMPLE ILLUSTRATING THE TRANSFORMATION TO CANONICAL FORM
OF THE CHEMICAL RATE EQUATIONS

By means of a simple example, this appendix illustrates the theory given in Chapter 3.0. The gas model, including the reactions, is not presumed to be physically realistic; it was chosen purely on the basis of its simplicity. In addition to the foregoing, the appendix also demonstrates the behavior of the transformation given in Chapter 3.0 in the frozen and equilibrium limits. This is accomplished by applying the limit processes to the illustrative example.

We assume a gas composed of the three reacting species X , X_2 and X_3 . Their mole-mass ratios are designated by n_1 , n_2 , and n_3 , respectively. Three reactions are assumed to govern the chemical kinetics as follows:



The gas model thus contains 3 species, 1 component, 3 reactions, and no inert or catalytic bodies. Hence, the parameters N_1 are given by

$$N_1 = N_2 = N_4 = 3 \quad , \quad N_3 = 2 . \quad (\text{II-1})$$

Conservation of component X is given by

$$n_1 + 2n_2 + 3n_3 = 3a_{30} = \text{constant} . \quad (\text{II-2a})$$

By solving for n_3 , we obtain the following equation for conservation of components:

$$\left. \begin{aligned} n_3 &= a_{30} + a_{31}n_1 + a_{32}n_2 , \\ n_3 &= a_{30} - \frac{1}{3}n_1 - \frac{2}{3}n_2 . \end{aligned} \right\} \quad (\text{II-2b})$$

The quantities L_r and θ_r^{-1} defined by equations (6b) and (6c) are then

$$\left. \begin{aligned} L_1 &= 1 - \frac{\rho}{K_{c1}} \frac{n_1 n_2}{n_3}, & \theta_1^{-1} &= k_{f1} n_3, \\ L_2 &= 1 - \frac{\rho}{K_{c2}} \frac{n_1^2}{n_2}, & \theta_2^{-1} &= k_{f2} n_2, \\ L_3 &= 1 - \frac{1}{K_{c3}} \frac{n_2^2}{n_1 n_3}, & \theta_3^{-1} &= \rho k_{f3} n_1 n_3, \end{aligned} \right\} \quad (\text{II-3})$$

and the rate equations (5) are given by

$$\left. \begin{aligned} \frac{Dn_1}{Dt} &= \frac{L_1}{\theta_1} + 2 \frac{L_2}{\theta_2} - \frac{L_3}{\theta_3}, \\ \frac{Dn_2}{Dt} &= \frac{L_1}{\theta_1} - \frac{L_2}{\theta_2} + 2 \frac{L_3}{\theta_3}. \end{aligned} \right\} \quad (\text{II-4})$$

No rate equation for n_3 is necessary since this quantity is determined by equation (II-2b).

The coefficients $\bar{\kappa}_{ij}$ in the linearized form (15b) of the chemical rate equations are as follows:

$$\left. \begin{aligned}
 n_1^* \bar{\kappa}_{11} &= \frac{1}{\theta_1^*} + \frac{4}{\theta_2^*} + \frac{1}{\theta_3^*}, \\
 n_2^* \bar{\kappa}_{22} &= \frac{1}{\theta_1^*} + \frac{1}{\theta_2^*} + \frac{4}{\theta_3^*}, \\
 n_3^* \bar{\kappa}_{33} &= \frac{1}{\theta_1^*} + \frac{1}{\theta_3^*}, \\
 n_2^* \bar{\kappa}_{12} &= n_1^* \bar{\kappa}_{21} = \frac{1}{\theta_1^*} - \frac{2}{\theta_2^*} - \frac{2}{\theta_3^*}, \\
 n_3^* \bar{\kappa}_{13} &= n_1^* \bar{\kappa}_{31} = -\frac{1}{\theta_1^*} + \frac{1}{\theta_3^*}, \\
 n_3^* \bar{\kappa}_{23} &= n_2^* \bar{\kappa}_{32} = -\frac{1}{\theta_1^*} - \frac{2}{\theta_3^*}.
 \end{aligned} \right\} \quad (\text{II-5})$$

The coefficients κ_{ij} that account for conservation of components in the linearized form (17) of the chemical rate equations are as follows:

$$\left. \begin{aligned}
 \kappa_{11} &= \bar{\kappa}_{11} + a_{31} \bar{\kappa}_{13} = \frac{1}{\theta_1^*} \left(\frac{1}{n_1^*} + \frac{1}{3} \frac{1}{n_3^*} \right) + \frac{4}{\theta_2^* n_1^*} + \frac{1}{\theta_3^*} \left(\frac{1}{n_1^*} - \frac{1}{3} \frac{1}{n_3^*} \right), \\
 \kappa_{22} &= \bar{\kappa}_{22} + a_{32} \bar{\kappa}_{23} = \frac{1}{\theta_1^*} \left(\frac{1}{n_2^*} + \frac{2}{3} \frac{1}{n_3^*} \right) + \frac{1}{\theta_2^* n_2^*} + \frac{4}{\theta_3^*} \left(\frac{1}{n_2^*} + \frac{1}{3} \frac{1}{n_3^*} \right), \\
 \kappa_{12} &= \bar{\kappa}_{12} + a_{32} \bar{\kappa}_{13} = \frac{1}{\theta_1^*} \left(\frac{1}{n_2^*} + \frac{2}{3} \frac{1}{n_3^*} \right) - \frac{2}{\theta_2^* n_2^*} - \frac{2}{\theta_3^*} \left(\frac{1}{n_2^*} + \frac{1}{3} \frac{1}{n_3^*} \right), \\
 \kappa_{21} &= \bar{\kappa}_{21} + a_{31} \bar{\kappa}_{23} = \frac{1}{\theta_1^*} \left(\frac{1}{n_1^*} + \frac{1}{3} \frac{1}{n_3^*} \right) - \frac{2}{\theta_2^* n_1^*} - \frac{2}{\theta_3^*} \left(\frac{1}{n_1^*} - \frac{1}{3} \frac{1}{n_3^*} \right).
 \end{aligned} \right\} \quad (\text{II-6})$$

The transformation equations (19a) and (26) are given by:

$$\left. \begin{aligned}
 q_1 &= n_1 + B_{12} n_2, \\
 q_2 &= B_{21} n_1 + n_2, \\
 q_3 &= n_3,
 \end{aligned} \right\} \quad (\text{II-7})$$

where $B_{11} = 1$ is arbitrarily assumed. Before the constants B_{12} and B_{21} can be found, the characteristic equation (24b) must be solved. This equation is:

$$\begin{vmatrix} \kappa - \kappa_{11} & \kappa_{12} \\ \kappa_{21} & \kappa - \kappa_{22} \end{vmatrix} = 0, \quad (\text{II-8a})$$

or

$$\kappa^2 - (\kappa_{11} + \kappa_{22}) \kappa + (\kappa_{11}\kappa_{22} - \kappa_{21}\kappa_{12}) = 0. \quad (\text{II-8b})$$

The solution for the two eigenvalues is then given by

$$\left. \begin{aligned} \kappa_1 &= \frac{1}{2} \left\{ \kappa_{11} + \kappa_{22} - \sqrt{(\kappa_{11} - \kappa_{22})^2 + 4\kappa_{21}\kappa_{12}} \right\}, \\ \kappa_2 &= \frac{1}{2} \left\{ \kappa_{11} + \kappa_{22} + \sqrt{(\kappa_{11} - \kappa_{22})^2 + 4\kappa_{21}\kappa_{12}} \right\}, \end{aligned} \right\} \quad (\text{II-9})$$

where κ_1 is arbitrarily chosen to go with the solution containing the minus sign.

The constants B_{12} and B_{21} are determined by equations (23). For example, B_{12} , with $i = 1$ and $l = 1$ is given by

$$B_{12} = \frac{\kappa_1 - \kappa_{11}}{\kappa_{21}}, \quad (\text{II-10a})$$

or, with $i = 1$ and $l = 2$, by

$$B_{12} = \frac{\kappa_{12}}{\kappa_1 - \kappa_{22}}. \quad (\text{II-10b})$$

These two values for B_{12} , by virtue of equation (II-8b), are equal. With $i = 2$ and $l = 2$, we obtain B_{21} as

$$B_{21} = \frac{\kappa_2 - \kappa_{22}}{\kappa_{12}}. \quad (\text{II-11})$$

Thus, the transformation (II-7) is uniquely determined.

The two transformed rate equations that replace equations (II-4) are given by equations (20). They need not be repeated here.

Although the theory of Chapter 3.0 is valid only for a near-equilibrium flow, the frozen limit nevertheless makes physical sense. For simplicity, we shall apply this limiting process to one reaction; the other reactions will, of course, behave in a similar fashion in this limit. Thus, we imagine a situation where one reaction rapidly changes from a near-equilibrium value to a frozen value. To be more precise, we let reaction 1 freeze while reactions 2 and 3 are unaltered. In mathematical terms, we require that

$$k_{f1} \rightarrow 0 \quad \text{and} \quad k_{b1} \rightarrow 0, \quad (\text{II-12})$$

which implies that (see equations (II-3))

$$\theta_1 \rightarrow \infty \quad \text{and} \quad \frac{v_{1f} v_{1b}}{\theta_1} \rightarrow 0. \quad (\text{II-13})$$

Thus, the θ_1 -term vanishes from all κ_{1l} and reaction 1 no longer affects the eigenvalues or the coefficients B_{1l} . Freezing reaction 1 is therefore tantamount to eliminating it from the original set of reactions.

We next investigate the equilibrium limit, in a similar fashion, by requiring reaction 1, for example, to tend toward equilibrium while reactions 2 and 3 are unaltered. In other words, we require that

$$k_{f1} \rightarrow \infty \quad \text{and} \quad k_{b1} \rightarrow \infty, \quad (\text{II-14})$$

which implies that

$$\theta_1 \rightarrow 0, \quad (\text{and } \theta_1^* \rightarrow 0). \quad (\text{II-15})$$

Substituting equations (II-6) into (II-9) and then simplifying, we obtain

$$\kappa_1 = \left(\frac{n_1^* + 4n_2^* + 9n_3^*}{n_1^* n_2^* + n_1^* n_3^* + n_2^* n_3^*} \right) \left(\frac{1}{\theta_2^*} + \frac{1}{\theta_3^*} \right) + O(\theta_1^*), \quad (\text{II-16a})$$

$$\kappa_2 = \left(\frac{1}{n_1^*} + \frac{1}{n_2^*} + \frac{1}{n_3^*} \right) \frac{1}{\theta_1^*} + O(1) . \quad (\text{II-16b})$$

In the limit as θ_1^* tends to zero, κ_1 tends toward a finite positive value while κ_2 tends toward $+\infty$. Equation (II-16b) in conjunction with the rate equation (20) for q_2 , then implies that $q_2 - q_{2,e}$ tend to zero, or

$$q_2 - q_{2,e} = B_{21}(n_1 - n_{1,e}) + (n_2 - n_{2,e}) \rightarrow 0 . \quad (\text{II-17})$$

To better understand the above expression, we first compute B_{21} and B_{12} via equations (II-10a) and (II-11). When this is done, the following result is obtained:

$$\left. \begin{aligned} B_{12} &\rightarrow -1 \\ B_{21} &\rightarrow \left(\frac{n_2^*}{n_1^*} \right) \left(\frac{n_1^* + 3n_3^*}{2n_2^* + 3n_3^*} \right) \end{aligned} \right\} \text{ as } \theta_1^* \rightarrow 0 . \quad (\text{II-18})$$

Next, substitute equations (I-5) into equation (13b) thereby resulting in

$$L_r = - \sum_{\ell=1}^{N_1} \frac{v_{r\ell}}{n_\ell^*} (n_\ell - n_{\ell,e}) , \quad r = 1, \dots, N_2 . \quad (\text{II-19a})$$

With $r = 1$, with $(n_3 - n_{3,e})$ eliminated via equation (II-2b), and by means of expressions (II-18) we finally obtain

$$L_1 = - \left(\frac{1}{n_2^*} + \frac{2}{3} \frac{1}{n_3^*} \right) \left\{ B_{21}(n_1 - n_{1,e}) + (n_2 - n_{2,e}) \right\} , \quad (\text{II-19b})$$

in the limit as θ_1^* tends to zero. Equation (II-19b) is the linearized form of L_1 (see equations (II-3)) with conservation of components accounted for. By comparing expressions (II-17) with equation (II-19b), we observe that the requirement that θ_1^* tend to zero is equivalent to the requirement that the linearized form of L_1 tend to zero. The equilibrium limit for reaction 1 thus reduces the system of two rate equations to a system consisting of one rate equation for q_1 , discussed below, and one algebraic equation that represents the linearized form of the law of mass action for reaction 1.

When B_{12} , given by (II-18), is substituted into the transformation (II-7), we obtain the equation

$$q_1 = n_1 - n_2. \quad (\text{II-20})$$

To better understand this equation, note that L_1/θ_1 , which is indeterminate in the equilibrium limit for reaction 1, disappears when the difference of the two rate equations (II-4) is taken as follows:

$$\frac{Dn_1}{Dt} - \frac{Dn_2}{Dt} = \frac{Dq_1}{Dt} = 3 \left(\frac{L_2}{\theta_2} - \frac{L_3}{\theta_3} \right). \quad (\text{II-21a})$$

After the right-hand side of equation (II-21a) is linearized, this equation can be shown to reduce to

$$\frac{Dq_1}{Dt} = -\kappa_1(q_1 - q_{1,e}), \quad (\text{II-21b})$$

where κ_1 is given by equation (II-16a) with $\theta_1^* = 0$, and equation (II-20) is used to obtain q_1 and $q_{1,e}$. Thus, the one applicable rate equation (II-21b), in the equilibrium limit for reaction 1, is formed by eliminating the indeterminate form L_1/θ_1 from the original rate equations (II-4).

The foregoing illustrates that the theory of Chapter 3.0 has the correct behavior in the frozen and equilibrium limits.

Arnold Engineering Development Center
Arnold Air Force Station, Tennessee

Rpt. No. AEDC-TDR-63-82. PROBLEMS UNDERLYING
THE NUMERICAL INTEGRATION OF THE CHEMICAL
AND VIBRATIONAL RATE EQUATIONS IN A NEAR-
EQUILIBRIUM FLOW. March 1963, 78 p. incl 20 refs.
Unclassified Report

The work described is a continuation of that reported previously in AEDC-TN-61-65 and AEDC-TDR-62-131. In particular, the present report contains the basis for the method described in AEDC-TDR-62-131 for the precise numerical calculation of one-dimensional nonequilibrium flow of a complex gas mixture through a nozzle. The contributions of the two above reports are reviewed briefly in the introduction. A transformation is shown to exist when the flow is close to equilibrium that transforms the chemical rate equations into a simpler system of uncoupled rate equations. Each of the transformed chemical rate equations

1. Gas flow
2. Numerical analysis
3. Hypersonic nozzles
4. Equations
1. AFSC Program Area 750A, Project 8952, Task 895210
- II. Contract AF 40(600)-930
- III. Stanford University, Stanford, Calif.
- IV. George Emanuel
- V. Available from OTS
- VI. In ASTIA Collection

Arnold Engineering Development Center
Arnold Air Force Station, Tennessee

Rpt. No. AEDC-TDR-63-82. PROBLEMS UNDERLYING
THE NUMERICAL INTEGRATION OF THE CHEMICAL
AND VIBRATIONAL RATE EQUATIONS IN A NEAR-
EQUILIBRIUM FLOW. March 1963, 78 p. incl 20 refs.
Unclassified Report

The work described is a continuation of that reported previously in AEDC-TN-61-65 and AEDC-TDR-62-131. In particular, the present report contains the basis for the method described in AEDC-TDR-62-131 for the precise numerical calculation of one-dimensional nonequilibrium flow of a complex gas mixture through a nozzle. The contributions of the two above reports are reviewed briefly in the introduction. A transformation is shown to exist when the flow is close to equilibrium that transforms the chemical rate equations into a simpler system of uncoupled rate equations. Each of the transformed chemical rate equations

1. Gas flow
2. Numerical analysis
3. Hypersonic nozzles
4. Equations
1. AFSC Program Area 750A, Project 8952, Task 895210
- II. Contract AF 40(600)-930
- III. Stanford University, Stanford, Calif.
- IV. George Emanuel
- V. Available from OTS
- VI. In ASTIA Collection

is similar in form to a vibrational rate equation. The coefficients appearing in the transformed equations are given by solutions of an eigenvalue problem. These coefficients are shown to be real and positive. An example is given that illustrates the transformation for a simple gas model and shows that the transformation behaves correctly in the frozen and equilibrium limits. Because of their simple mathematical structure, the vibrational and transformed chemical rate equations can be subjected to numerical analysis. Thus, a detailed analysis of both Runge-Kutta and predictor-corrector integration procedures is performed in which truncation error, stability, and computation speed are examined. The analysis of Runge-Kutta procedures demonstrates that the greater the order of the procedure, the more stable it is. High-order Runge-Kutta procedures, unfortunately, also have reduced computation speed. On an overall basis, the commonly employed fourth-order procedure still appears to be the most suitable Runge-Kutta procedure for the integration of the chemical and vibrational rate equations. The basis for the improved technique for controlling the Runge-Kutta integration step size, given in AEDC-TDR-62-131, is also established. When Adams predictor-corrector procedures are similarly analyzed, the greater the order of the procedure the less stable it is. Furthermore, the overall Adams procedure is considerably less stable than is the Adams corrector formula. The common assumption that stability is determined primarily by the corrector is thus not valid. A fourth-order predictor-corrector procedure is also given that is stable for a larger integration step size than the fourth-order Adams procedure.

is similar in form to a vibrational rate equation. The coefficients appearing in the transformed equations are given by solutions of an eigenvalue problem. These coefficients are shown to be real and positive. An example is given that illustrates the transformation for a simple gas model and shows that the transformation behaves correctly in the frozen and equilibrium limits. Because of their simple mathematical structure, the vibrational and transformed chemical rate equations can be subjected to numerical analysis. Thus, a detailed analysis of both Runge-Kutta and predictor-corrector integration procedures is performed in which truncation error, stability, and computation speed are examined. The analysis of Runge-Kutta procedures demonstrates that the greater the order of the procedure, the more stable it is. High-order Runge-Kutta procedures, unfortunately, also have reduced computation speed. On an overall basis, the commonly employed fourth-order procedure still appears to be the most suitable Runge-Kutta procedure for the integration of the chemical and vibrational rate equations. The basis for the improved technique for controlling the Runge-Kutta integration step size, given in AEDC-TDR-62-131, is also established. When Adams predictor-corrector procedures are similarly analyzed, the greater the order of the procedure the less stable it is. Furthermore, the overall Adams procedure is considerably less stable than is the Adams corrector formula. The common assumption that stability is determined primarily by the corrector is thus not valid. A fourth-order predictor-corrector procedure is also given that is stable for a larger integration step size than the fourth-order Adams procedure.