R-1036-PR
April 1974

Implicit Function Theorems for Optimization Problems and for Systems of Inequalities

James H. Bigelow and Norman Z. Shapiro

A Report prepared for

UNITED STATES AIR FORCE PROJECT RAND

Rand
SANTA MONICA, CA. 90406

The research described in this Report was sponsored by the United States Air Force under Contract No. F44620-73-C-0011 — Monitored by the Directorate of Operational Requirements and Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Reports of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

R-1036-PR
April 1974

Implicit Function Theorems for Optimization Problems and for Systems of Inequalities

James H. Bigelow and Norman Z. Shapiro

A Report prepared for

UNITED STATES AIR FORCE PROJECT RAND

Rand
SANTA MONICA, CA. 90406

Published by The Rand Corporation

PREFACE

A great many Air Force operations research and other scientific problems have the following mathematical structure: minimize a function $f(x)$ of n variables ($x = (x_1, x_2, \ldots x_n)$) subject to m inequality constraints $g_1(x) \leq 0, \ldots, g_m(x) \leq 0$. For example, in an optimal rate-of-climb calculation the function f might be the rate-of-climb; the variables x_i such quantities as current weight, speed, and rate of fuel consumption; and the functions g_i limits imposed by characteristics of the engine, structural strength, and the aerodynamic characteristics of the aircraft. For another example, in a chemical equilibrium calculation (say calculating the blood-chemistry of an astronaut in a special atmosphere) the function f would be the Gibbs free-energy function; the variables x_i would be the number of moles of the several chemical species; and the functions g_i would reflect the mass-balance laws. In any application, these functions will depend on parameters (such as engine horsepower and fuel quality in the first example; or temperature, pressure, and initial conditions in the second).

If the functions f, $g_1, \ldots g_m$ depend on a parameter λ, then the solution (that is, the value $x = (x_1, x_2, \ldots x_n)$ which minimizes the function f subject to the constraints) is dependent on the same parameter. In this report it is proved that, under rather general conditions, the solution to this optimization problem is a differentiable function of the parameters, and the derivative may often be calculated about as easily as the solution itself. The calculation of this derivative is of considerable practical significance: it identifies how the parameters

influence the optimal policy; it tells how sensitive the "optimal" solution is to variations in the underlying assumptions; and it can be used to estimate approximate solutions to the problem for different sets of parameters.

This work was done as part of supporting research for USAF Project RAND.

SUMMARY

Implicit function formulas for differentiating the solutions of mathematical programming problems satisfying the conditions of the Kuhn-Tucker theorem are motivated and rigorously demonstrated. The special case of a convex objective function with linear constraints is also treated with emphasis on computational details. An example, an application to chemical equilibrium problems, is given.

Implicit function formulas for differentiating the unique solution of a system of simultaneous inequalities are also derived.

CONTENTS

PREFACE iii

SUMMARY v

Section
I. INTRODUCTION 1

II. THE GENERAL CASE, MOTIVATION 3

III. PROOF OF RESULTS OF SECTION I 8

IV. CONVEX OBJECTIVE FUNCTION AND LINEAR CONSTRAINTS 13

V. THE CHEMICAL EQUILIBRIUM PROBLEM 18

Appendix
IMPLICIT FUNCTION THEOREMS FOR SYSTEMS OF INEQUALITIES ... 21

REFERENCES 27

I. INTRODUCTION

The general mathematical programming problem can be formulated as a constrained minimization problem; that is, minimizing a real valued function of several variables $(x_1, \ldots, x_n) \in E^n$, subject to inequality constraints $g_1(x) \leq 0, \ldots, g_m(x) \leq 0$.

In addition to the variables $x_1, \ldots, x_n$ the objective function f, as well as the constraints $g_1, \ldots, g_m$, may depend on other parameters. For example in an optimal resource allocation problem these parameters are such factors as unit costs, depreciation rates, productivity of labor, and availability of resources. In an optimal aircraft rate-of-climb calculation these parameters are such factors as initial aircraft weight, fuel consumption rates, structural damage limits, and initial aircraft altitude. In a chemical equilibrium calculation, which is an optimization problem (see Ref. 1), these parameters are such factors as temperature, pressure and initial amounts.

Although it is relatively rarely calculated, the sensitivity of the solution of a mathematical programming problem (that is, the derivatives of the minimizing $x_1, \ldots, x_n$ with respect to these parameters) is imperfectly known, and errors in determining them may have unsuspected consequences on the optimum solution.

Furthermore, many minimization techniques are based on iterative approximation techniques which are considerably more efficient when good initial approximate solutions are available. The ability to compute the partial derivatives of solutions easily with respect to parameters can then lead to considerable efficiency when the solution to each of a series of different optimization problems is calculated.

In Sec. II we informally derive the basic result for an arbitrary mathematical programming problem, assuming only that the conditions of the Kuhn-Tucker theorem are met. In Sec. III the result is rigorously demonstrated. In Sec. IV we consider the special case of a convex objective function, with linear constraints. In Sec. V we further specialize, by way of example, to chemical equilibrium problems. Sections IV and V emphasize practical computational problems.

In the Appendix, we present an implicit function theorem for systems of simultaneous inequalities. This result is of interest in itself and is heavily used by Secs. II and III.

Our basic result was previously obtained by Fiacco and McCormick (Theorem 6, p. 34 of Ref. 2), in the special case of "strict complementarity"; that is, in the special case that the Kuhn-Tucker multipliers and the corresponding constraint do not simultaneously vanish; that is, the set, K, of Sec. II is non-empty. In Refs. 3 and 4 the derivative of the objective function, but not of the solution of a linear programming problem, is treated.

II. THE GENERAL CASE, MOTIVATION

Consider the mathematical programming problem,

$$\begin{aligned} &\text{Min } f(x) \\ &\text{s.t. } g_i(x) \leq 0 , \qquad i = 1, \ldots, m , \end{aligned} \tag{1}$$

where $x \in E^n$. We suppose that f and the g_i satisfy the conditions of the Kuhn-Tucker theorem. Then, if x is a solution to (1), there exists $\pi \in E^m$ for which

$$\frac{\partial f}{\partial x_j}(x) + \sum_{i=1}^{m} \pi_i \frac{\partial g_i}{\partial x_j}(x) = 0 \qquad 1 \leq j \leq n , \tag{2.1}$$

$$g_i(x) \leq 0 \qquad 1 \leq i \leq m , \tag{2.2}$$

$$\pi_i \geq 0 \qquad 1 \leq i \leq m , \tag{2.3}$$

$$\pi_i \, g_i(x) = 0 \qquad 1 \leq i \leq m . \tag{2.4}$$

We shall initially derive our results in a nonrigorous, rather uninhibited fashion, to provide the reader with motivation. A correct, but less direct, proof of the results is in the next section.

We shall use the following notation: the parameters of the problem are taken to be a vector $p \in E^q$; the direction in which we vary this vector is $\dot{p}$. Similarly, the derivations of x and π in the direction $\dot{p}$ shall be $\dot{x}$ and $\dot{\pi}$ respectively.

As before, $\frac{\partial f}{\partial x}$ is a vector of partial derivatives of the function f, each of whose components is a partial derivative with respect to

an x_j. We shall take $\frac{\partial^2 f}{\partial x^2}$ to be the matrix of second partial derivatives with respect to x, and $\frac{\partial^2 f}{\partial p \partial x}$ to be the matrix of second partial derivatives, first with respect to x, second with respect to p. Similar comments apply to the g_i.

Define sets of indices I, J, K to be

$$I = \{i \mid \pi_i > 0\} ,$$

$$J = \{i \mid g_i(x) < 0\} ,$$

$$K = \{i \mid \pi_i = 0 \quad \text{and} \quad g_i(x) = 0\} .$$

Clearly, by (2), $I \cup J \cup K = \{1, 2, \ldots, m\}$.

Differentiating (2.1) yields

$$\left[\left(\frac{\partial^2 f}{\partial x^2}\right) + \sum_{i \in I} \pi_i \left(\frac{\partial^2 g_i}{\partial x^2}\right)\right] \dot{x} + \sum_i \dot{\pi}_i \frac{\partial g_i}{\partial x} = -\left[\frac{\partial^2 f}{\partial p \partial x} + \sum_i \pi_i \frac{\partial^2 g_i}{\partial p \partial x}\right] \dot{p} . \qquad (3)$$

Differentiating conditions (2.2) gives us

$$\frac{\partial g_i}{\partial x} \cdot \dot{x} \leq -\frac{\partial g_i}{\partial p} \cdot \dot{p} \qquad i \in I \cup K . \qquad (4)$$

Note that there is no differential condition (4) for $i \in J$, since $g_i(x) < 0$ in this case. (See the Appendix.)

Conditions (2.3) yield the derivatives,

$$\dot{\pi}_i \geq 0 \qquad i \in J \cup K . \qquad (5)$$

Again, note that for $i \in I$, $\pi_i > 0$, so for these i we obtain no conditions on $\dot{\pi}_i$.

Finally, from conditions (2.4) we obtain

$$\dot{\pi}_i g_i(x) + \pi_i \left(\frac{\partial g_i}{\partial x} \dot{x} + \frac{\partial g_i}{\partial p} \cdot \dot{p} \right) = 0 , \qquad 1 \leq i \leq m . \tag{6}$$

Note that for $i \in I$, we have $\pi_i > 0$ implies $g_i(x) = 0$. Then (6) reduces to

$$\frac{\partial g_i}{\partial x} \dot{x} = - \frac{\partial g_i}{\partial p} \cdot \dot{p} , \qquad i \in I . \tag{7}$$

Similarly, for $i \in J$, we have $g_i(x) < 0$ implies $\pi_i = 0$. Thus (6) becomes

$$\dot{\pi}_i = 0 , \qquad i \in J . \tag{8}$$

However, in the event that $i \in K$, Eq. (6) becomes an identity. When this occurs, we differentiate (6) once more, obtaining (after dropping terms that must vanish)

$$\dot{\pi}_i \left(\frac{\partial g_i}{\partial x} \dot{x} + \frac{\partial g_i}{\partial p} \cdot \dot{p} \right) = 0 , \qquad i \in K . \tag{9}$$

This is the point at which this derivation lacks rigor. If one knows that x and π are twice differentiable, or assumes it, then conditions (9) can properly be obtained as above. However, even if x and π are not twice differentiable, conditions (9) must hold, as will be proved in the next section.

Combining (3) through (9), we obtain the differential conditions on $\dot{x}$, $\dot{\pi}$.

$$\left[\frac{\partial^2 f}{\partial x^2} + \sum_i \pi_i \frac{\partial^2 g_i}{\partial x^2}\right]\dot{x} + \sum_i \frac{\partial g_i}{\partial x}\dot{\pi}_i = -\left[\frac{\partial^2 f}{\partial p \partial x} + \sum_i \pi_i \frac{\partial^2 g_i}{\partial p \partial x}\right]\dot{p} ,$$

$$\begin{aligned}
\frac{\partial g_i}{\partial x}\dot{x} &= -\frac{\partial g_i}{\partial p}\cdot\dot{p} && i \in I , \\
\frac{\partial g_i}{\partial x}\dot{x} &\le -\frac{\partial g_i}{\partial p}\cdot\dot{p} && i \in K , \\
\dot{\pi}_i &= 0 && i \in J , \\
\dot{\pi}_i &\ge 0 && i \in K , \\
\dot{\pi}_i\left(\frac{\partial g_i}{\partial x}\dot{x} + \frac{\partial g_i}{\partial p}\cdot\dot{p}\right) &= 0 && i \in K .
\end{aligned} \qquad (10)$$

The theorems of Sec. III and the Appendix ensure that when (10) possesses a unique solution $(\dot{x}, \dot{\pi})$, then x and π are differentiable (from the right) in the direction $\dot{p}$, and that the derivatives are $\dot{x}$ and $\dot{\pi}$.

It can easily be shown that conditions (10) are the Kuhn-Tucker conditions for the following quadratic programming problem:

$$\text{Min } \tfrac{1}{2}\dot{x}^T\left[\frac{\partial^2}{\partial x^2}\left(f + \sum_i \pi_i g_i(x)\right)\right]\dot{x} + \dot{x}^T\left[\frac{\alpha^2 f}{\alpha p \alpha x} + \sum_i \pi_i \frac{\partial^2 g_i}{\alpha p \alpha x}\right]\dot{p} ,$$

$$\text{s.t.} \qquad \begin{aligned}
\frac{\partial g_i}{\partial x}\dot{x} &= -\frac{\partial g_i}{\partial p}\cdot\dot{p} && i \in I , \\
\frac{\partial g_i}{\partial x}\dot{x} &\le -\frac{\partial g_i}{\partial p}\cdot\dot{p} && i \in K ,
\end{aligned} \qquad (11)$$

where the multiplier on constraint i is $\dot{\pi}_i$.

Finally, note that if K is empty, then (10) becomes a system of linear *equations*. Furthermore even if K is non-empty, (10) can be solved by solving several systems of simultaneous linear inequalities.

III. PROOF OF RESULTS OF SECTION I

Without loss of generality, we may assume our problem (1) to have only one parameter, t. Then we rewrite condition (2) as

$$\frac{\partial f(x,t)}{\partial x} + \sum_i \pi_i \frac{\partial g_i(x,t)}{\partial x} = 0 \; ; \tag{12.1}$$

$$g_i(x,t) \leq 0 \; ; \tag{12.2}$$

$$\pi_i \geq 0 \; ; \tag{12.3}$$

$$\pi_i \cdot g_i(x,t) = 0 \; . \tag{12.4}$$

Conditions (12) are assumed to possess a unique solution $(x(t), \pi(t))$ for each $t \in N = \{t \mid 0 \leq t < s\}$, where $s > 0$.

As before, we let

$$K = \{i \mid \pi_i(0) = 0 \quad \text{and} \quad g_i[x(0), 0] = 0\} \; ,$$

and we will suppose there are q indices in K.

There are 3^q different ways to partition K into three sets, S_1, S_2 and S_3. For each such partition,* we define a set T_{S_1,S_2,S_3} by

$$T_{S_1,S_2,S_3} = \{t \in N \mid \pi_i(t) > 0 \; , \quad (\forall i) \in S_1 \; ,$$
$$g_i(x(t),t) < 0 \; , \quad (\forall i) \in S_2 \; , \text{ and}$$
$$\pi_i(t) = g_i(x(t),t) = 0 \; , \quad (\forall i) \in S_3\} \; .$$

*By a "partition," we mean a separation of K into three disjoint sets whose union is K.

Then clearly, $\bigcup_{P} T_{S_1,S_2,S_3} = N$, where P is the set of all partitions of K.

Further, for any particular partition (S_1, S_2, S_3), we may rewrite (12) as:

$$\frac{\partial f}{\partial x}(x,t) + \sum \pi_i \cdot \frac{\partial g_i}{\partial x}(x,t) = 0 \; ; \tag{13.1}$$

$$\left.\begin{array}{lll} \pi_i = 0 & (\forall i) \in S_2 \cup S_3 \; , \\ g_i(x,t) = 0 & (\forall i) \in S_1 \cup S_3 \; , \\ \pi_i \geq 0 & (\forall i) \in S_1 \; , \\ g_i(x,t) \leq 0 & (\forall i) \in S_2 \; ; \end{array}\right\} \tag{13.2}$$

$$\left.\begin{array}{lll} \pi_i \geq 0 & (\forall i) \notin K \; , \\ g_i(x,t) \leq 0 & (\forall i) \notin K \; , \\ \pi_i \cdot g_i(x,t) = 0 & (\forall i) \notin K \; . \end{array}\right\} \tag{13.3}$$

And (13) must hold for every $t \in T_{S_1,S_2,S_3}$. For exactly these t, (13) is equivalent to (12).

For all partitions S_1,S_2,S_3 of K, the conditions (13.1) and (13.3) are the same, and hence the conditions on $\dot{x}$ and $\dot{\pi}$ arising from (13.1) and (13.3) are the same; they are shown below:

$$\left[\frac{\partial^2 f}{\partial x^2} + \sum_i \pi_i \frac{\partial^2 g_i}{\partial x^2}\right]\dot{x} + \sum_i \frac{\partial g_i}{\partial x} \cdot \dot{\pi}_i = -\left[\frac{\partial^2 f}{\partial t \partial x} + \sum_i \pi_i \frac{\partial^2 g_i}{\partial t \partial x}\right],$$

$$\frac{\partial g_i}{\partial x} \cdot \dot{x} = \frac{\partial g_i}{\partial t} \qquad i \in I \; , \tag{14}$$

$$\dot{\pi}_i = 0 \qquad i \in J \; .$$

The sets I and J are as in the previous section.

For a particular partition (S_1,S_2,S_3) of K, conditions (13.2) yield the following constraints on $\dot{x}$ and $\dot{\pi}$:

$$\begin{aligned} &\dot{\pi}_i = 0 , && i \in S_2 \cup S_3 , \\ &\frac{\partial g_i}{\partial x} \cdot \dot{x} = - \frac{\partial g_i}{\partial t} , && i \in S_1 \cup S_3 , \\ &\dot{\pi}_i \geq 0 , && i \in S_1 , \\ &\frac{\partial g_i}{\partial x} \cdot \dot{x} \leq - \frac{\partial g_i}{\partial t} , && i \in S_2 . \end{aligned} \tag{15}$$

Theorem 3 (ii) of the Appendix then states:

THEOREM 1. If conditions (14) and (15) together possess a unique solution $(\dot{x},\dot{\pi})$, and $0 \in c\ell(T_{S_1,S_2,S_3})$, then:

$$\lim_{\substack{t \to 0^+ \\ t \in T_{S_1,S_2,S_3}}} \frac{x(t) - x(0)}{t} = \dot{x} ,$$

and

$$\lim_{\substack{t \to 0^+ \\ t \in T_{S_1,S_2,S_3}}} \frac{\pi(t) - \pi(0)}{t} = \dot{\pi} .$$

Theorem 3 (i) of the Appendix shows that:

LEMMA 2. If $0 \in c\ell(T_{S_1,S_2,S_3})$, then conditions (14) and (15) together possess at least one solution $(\dot{x},\dot{\pi})$.

Next, we introduce the conditions

$$\dot{\pi}_i \geq 0 \qquad i \in K ;$$

$$\frac{\partial g_i}{\partial x} \cdot \dot{x} \leq - \frac{\partial g_i}{\partial t} \qquad i \in K ; \tag{16}$$

$$\dot{\pi}_i \left(\frac{\partial g_i}{\partial x} \cdot \dot{x} + \frac{\partial g_i}{\partial t} \right) = 0 \qquad i \in K ;$$

and we notice that for any partition (S_1,S_2,S_3) of K, every solution to the corresponding conditions (14) and (15) is also a solution to the conditions (14) and (16), and that therefore:

THEOREM 3. If conditions (14) and (16) together possess a unique solution $(\dot{x},\dot{\pi})$, then $x(t)$ and $\pi(t)$ are differentiable from the right at $t = 0$, and:

$$\frac{dx}{dt} = \lim_{t \to 0^+} \frac{x(t) - x(0)}{t} = \dot{x} ,$$

$$\frac{d\pi}{dt} = \lim_{t \to 0^+} \frac{\pi(t) - \pi(0)}{t} = \dot{\pi} .$$

Proof. It is obvious that for at least one partition (S_1,S_2,S_3) of K, $(\dot{x},\dot{\pi})$ is a solution to the associated set of conditions (14) and (15). Let Q be the set of all partitions (S_1,S_2,S_3) for which $(\dot{x},\dot{\pi})$ solves (14) and (15). Clearly, by the remark above, $(\dot{x},\dot{\pi})$ is the unique solution to (14) and (15) for each $(S_1,S_2,S_3) \in Q$.

If we only show that, for some $t_0 > 0$,

$$\bigcup_{(S_1,S_2,S_3) \in Q} T_{S_1S_2S_3} \supset \{t \mid 0 < t < t_0\}$$

then by Theorem 1 we will be done.

But from Lemma 2, if $(S_1,S_2,S_3) \notin Q$, then $0 \notin c\ell(T_{S_1,S_2,S_3})$, so for each such partition, there exists $t_{S_1,S_2,S_3} > 0$, a lower bound for $t \in T_{S_1,S_2,S_3}$, and $t_{S_1,S_2,S_3} \notin T_{S_1,S_2,S_3}$.

Now let $t_0 = \min\{t_{S_1,S_2,S_3} \mid (S_1S_2S_3) \notin Q\}$. Since there are a finite number of partitions, this minimum exists and $t_0 > 0$.

Clearly, this t_0 has the property required; Q.E.D.

Since conditions (14) and (16) together are identical to conditions (10), Theorem 3 establishes the results of the previous section.

IV. CONVEX OBJECTIVE FUNCTION AND LINEAR CONSTRAINTS

We wish to consider:

$$\begin{aligned} &\text{Min} \quad f(x) \ , \\ &\text{s.t.} \quad Ax = b \ , \\ &\qquad x \geq 0 \ , \end{aligned} \tag{17}$$

where f is a convex function, A an $m \times n$ matrix with $m < n$, $b \in E^m$, $x \in E^n$.

The Kuhn-Tucker optimality conditions for (17) may be written as

$$\begin{aligned} \frac{\partial f}{\partial x} - A^T\pi - u &= 0 \ , \\ Ax &= b \ , \\ x &\geq 0 \ , \\ u &\geq 0 \ , \\ (x \cdot u) &= 0 \ . \end{aligned} \tag{18}$$

For this problem, the set K is $\{j \mid x_j = 0 \text{ and } u_j = 0\}$, and we shall assume K is empty. Then it is easy to see that, by eliminating those x_j which are zero from Problem (17), we may construct a new problem whose (unique) solution x is strictly positive. Conditions (18) then become

$$\begin{aligned} \frac{\partial f}{\partial x} - A^T\pi &= 0 \ , \\ Ax &= b \ . \end{aligned} \tag{19}$$

The differential conditions arising from (19) are just

$$\frac{\partial^2 f}{\partial x^2} \cdot \dot{x} - A^T \dot{\pi} = - \frac{\partial^2 f}{\partial p \partial x} \cdot \dot{p} + \dot{A}^T \pi \ ,$$

$$A\dot{x} = - \dot{A}x + \dot{b} \ . \tag{20}$$

Those conditions (20) will have a unique solution $(\dot{x},\dot{\pi})$ if, and only if, (1) the m rows of the matrix A are linearly independent, and (2) there is no vector $\theta \in E^n$ for which both $A\theta = 0$ and $\left(\frac{\partial^2 f}{\partial x^2}\right)\theta = 0$.

To solve (20), one might simply invert the coefficient matrix.

$$\begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & - A^T \\ A & 0 \end{pmatrix}.$$

However, there is a way to save computation time.

Let Θ be an $r \times n$ matrix whose r rows form a basis for the null space of the linear operator $\frac{\partial^2 f}{\partial x^2}: E^n \to E^n$. [Of course, $\left(\frac{\alpha^2 f}{\alpha x^2}\right)$ is a function of x; but we have evaluated it at that point x which solves (17), and hence now consider it to be a fixed matrix.]

Define r new variables, $\bar{x} \in E^r$, and let $G(\bar{x})$ be a strictly *concave*, twice continuously differentiable function of $\bar{x}$.

Now define a new function $F: E^n \to E^1$ so that whenever $\bar{x} = \Theta \cdot x$, we have

$$f(x) = F(x) + G(\bar{x}) \ .$$

Then, we may replace (17) with the equivalent problem:

$$\begin{aligned} \text{Min} \quad & F(x) + G(\bar{x}) \ , \\ \text{s.t.} \quad & Ax = b \ , \\ & \Theta x - I\bar{x} = 0 \ , \\ & x \geq 0 \ . \end{aligned}$$

The Kuhn-Tucker optimality conditions for this expanded problem must be (recalling that the solution x of (17) is strictly positive)

$$\begin{aligned} \frac{\partial F}{\partial x} \qquad & - A^T\pi - \Theta^T\bar{\pi} = 0 , \\ \frac{\partial G}{\partial \bar{x}} \qquad & + I\,\bar{\pi} = 0 , \\ Ax \qquad & = b , \\ \Theta x - I\bar{x} \qquad & = 0 . \end{aligned} \tag{22}$$

In addition, the differential conditions for (21) become

$$\frac{\partial^2 F}{\partial x^2}\dot{x} - A^T\dot{\pi} - \dot{\Theta}^T\dot{\bar{\pi}} = -\frac{\partial^2 f}{\partial p \partial x}\dot{p} + \dot{A}^T\pi \tag{23.1}$$

$$\frac{\partial^2 G}{\partial \bar{x}^2}\dot{\bar{x}} + I\dot{\bar{\pi}} = 0 \tag{23.2}$$

$$A\dot{x} = -\dot{A}x + \dot{b} \tag{23.3}$$

$$\Theta\dot{x} - I\dot{\bar{x}} = 0 \tag{23.4}$$

We have gained something more real than apparent, for while the coefficient matrix of (23) is larger than that of (20), (23) requires less computation to solve than does (20). To see this, note that

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial^2 F}{\partial x^2} + \Theta^T \frac{\partial^2 G}{\partial \bar{x}^2}\Theta ,$$

so, by the convexity of f, the strict concavity of G, and the definition of Θ, the matrix $\frac{\partial^2 F}{\partial x^2}$ is strictly positive, definite, and hence, nonsingular. Also, since G was strictly concave, $\frac{\partial^2 G}{\partial \bar{x}^2}$ is nonsingular.

And since G is arbitrary, we may choose it to make the matrix $\frac{\partial^2 G}{\partial \bar{x}^2}$ easy to invert; for example, by making it a diagonal matrix.

We may solve (23) as follows: first, invert $\frac{\partial^2 F}{\partial x^2}$ (an $n \times n$ matrix) and $\frac{\partial^2 G}{\partial \bar{x}^2}$ (a diagonal $r \times r$ matrix). Multiply (23.1) and (23.2) by $\left(\frac{\partial^2 F}{\partial x^2}\right)^{-1}$ and $\left(\frac{\partial^2 G}{\partial \bar{x}^2}\right)^{-1}$ respectively. Then eliminate the variables $\dot{x}$, $\dot{\bar{x}}$ from the remaining two equations. The result to this point is

$$\begin{aligned} I\dot{x} - \left(\frac{\partial^2 F}{\partial x^2}\right)^{-1} A^T\dot{\pi} - \left(\frac{\partial^2 F}{\partial x^2}\right)^{-1} \Theta^T\dot{\bar{\pi}} &= \left(\frac{\partial^2 F}{\partial x^2}\right)^{-1}\left(-\frac{\partial^2 f}{\partial p \partial x}\dot{p} + \dot{A}^T\pi\right), \\ I\dot{\bar{x}} + \left(\frac{\partial^2 G}{\partial \bar{x}^2}\right)^{-1}\dot{\bar{\pi}} &= 0, \end{aligned} \tag{24}$$

$$R\begin{pmatrix}\dot{\pi}\\ \dot{\bar{\pi}}\end{pmatrix} = s,$$

where R is the $(m + r) \times (m + r)$ matrix:

$$R = \left(\begin{array}{c|c} A\left(\frac{\partial^2 F}{\partial x^2}\right)^{-1} A^T & A\left(\frac{\partial^2 F}{\partial x^2}\right)^{-1}\Theta^T \\ \hline \Theta\left(\frac{\partial^2 F}{\partial^2 x}\right)^{-1} A^T & \Theta\left(\frac{\partial^2 F}{\partial x^2}\right)^{-1}\Theta^T + \left(\frac{\partial^2 G}{\partial \bar{x}^2}\right)^{-1} \end{array}\right),$$

and

$$s = \begin{pmatrix} -\dot{A}x + b - A\left(\frac{\partial^2 f}{\partial x^2}\right)^{-1}\left(-\frac{\partial^2 f}{\partial p \partial x}\cdot\dot{p} + \dot{A}^T\pi\right) \\ -\Theta\left(\frac{\partial^2 F}{\partial x^2}\right)^{-1}\left(-\frac{\partial^2 f}{\partial p \partial x}\cdot\dot{p} + \dot{A}^T\pi\right) \end{pmatrix}.$$

Finally, we invert the $(m + r) \times (m + r)$ matrix R, to find $\dot{\pi}$, $\dot{\bar{\pi}}$ as $R^{-1}s$, and substitute into (24) to obtain $\dot{x}$, $\dot{\bar{x}}$.

Note that this procedure is worthwhile only if r is relatively small, for it requires the inversion of one $n \times n$ matrix and one $(m + r) \times (m + r)$ matrix. For sufficiently large r (so that $m + r$ is near n, its maximum possible value), and sufficiently small m, it is probably more efficient to work directly with (20).

Note also that if $\frac{\partial^2 f}{\partial x^2}$ is nonsingular, as it will be if f is strictly convex, then $r = 0$, Θ is vacuous, R becomes $A\left(\frac{\partial^2 f}{\partial x^2}\right)^{-1} A^T$, $s = (-\dot{A}x + \dot{b}) - A\left(\frac{\partial^2 f}{\partial x^2}\right)^{-1}\left(\frac{\partial^2 f}{\partial p \partial x} \cdot \dot{p} + \dot{A}^T\pi\right)$. This procedure is then simply the obvious method for solving (20).

V. THE CHEMICAL EQUILIBRIUM PROBLEM

A chemical equilibrium problem may be expressed mathematically in the form of (17). The variables $x = (x_1, \ldots, x_n)$ correspond to the number of moles of each of the chemical species in the system. The constraints $Ax = b$ are called the mass-balance constraints. For a more complete description, see Ref. 1.

The species are partitioned into non-empty subsets called phases. We denote the phase containing the j*th* species by $\langle j \rangle$.

For each phase $\langle j \rangle$ we define a function,

$$\bar{x}_{\langle j \rangle} = \sum_{k \in \langle j \rangle} x_k \ . \tag{25}$$

Then the objective function f, called the Gibbs free-energy function, is defined by

$$f(x) = \sum_{j=1}^{n} x_j \left(c_j + \log \frac{x_j}{\bar{x}_{\langle j \rangle}}\right) ,$$

where the c_j are constants. Here, f is defined as above for all $x > 0$ and can be extended continuously to the boundary of $\{x \mid x \geq 0\}$ by defining $t \log t$ to be zero when $t = 0$.

The chemical equilibrium problem then becomes:

$$\begin{aligned} \text{Min} \quad & f(x) = \sum x_j \left(c_j + \log \frac{x_j}{\bar{x}_{\langle j \rangle}}\right) , \\ \text{s.t.} \quad & Ax = b , \\ & x \geq 0 . \end{aligned} \tag{26}$$

If we assume that (26) has a unique, strictly positive solution, the Kuhn-Tucker optimality conditions become

$$\begin{aligned} c_j + \log \frac{x_j}{\bar{x}_{<j>}} - A_j^T \pi &= 0 \ , \\ Ax &= b \ , \end{aligned} \tag{27}$$

For this problem, the method of adding new variables to compute derivatives is extremely useful. We simply add a new variable $\bar{x}_{<j>}$ for each phase <j> and introduce Eqs. (25) as constraints, instead of as definitions. The function $G(\bar{x})$ becomes

$$G(\bar{x}) = - \sum_{<j>} \bar{x}_{<j>} \log \bar{x}_{<j>} \ .$$

Then:

$$\begin{aligned} F(x) &= f(x) - G(\bar{x}) \\ &= \sum x_j (c_j + \log x_j) \ . \end{aligned}$$

The matrix $\frac{\partial^2 F}{\partial x^2}$ is then a diagonal matrix whose jj^{th} element is $1/x_j$; $\frac{\partial^2 G}{\partial \bar{x}^2}$ is also a diagonal matrix with its $<j><j>^{th}$ element equal to $-1/\bar{x}_{<j>}$.

The matrix Θ (see previous section) has as many rows as there are phases; and in its $<j>^{th}$ row has a "1" in each column k for which $k \in <j>$, and zeros elsewhere.

For the chemical equilibrium problem, the R matrix of the last section can be computed as

$$R = \left(\begin{array}{c|c} r_{ik} & \bar{r}_{i,<j>} \\ \hline \bar{r}_{<j>,k} & 0 \end{array} \right),$$

where:

$$r_{ik} = r_{ki} = \sum_{j=1}^{n} a_{ij}\, a_{kj}\, x_j \,,$$

$$\bar{r}_{<j>,i} = \bar{r}_{i,<j>} = \sum_{k \varepsilon <j>} a_{ik}\, x_k \,.$$

APPENDIX

IMPLICIT FUNCTION THEOREMS FOR SYSTEMS OF INEQUALITIES

Let $F(y,p) = (F_1(y,p), F_2(y,p), \ldots, F_r(y,p)) \in E^r$ be a function of $y = (y_1, y_2, \ldots, y_n) \in E^n$ and $p = (p_1, \ldots, p_m) \in E^m$; then the system,

$$(1) \quad F(y,p) \geq 0 ,$$

may be thought of as a system of simultaneous inequalities in the unknowns $y_1, \ldots, y_n$ which depends on the parameters $p_1, \ldots, p_m$. This system may therefore define $y_1, \ldots, y_m$ which depend on $p_1, \ldots, p_m$. We wish to obtain results concerning this dependence. Is it continuous? Is it differentiable? What are the relations defining its derivatives? How may the derivatives be computed?

Theorem 1

Let $F:UxV \to E^r$ be continuous where U is an open, bounded subset of E^n and $V \subseteq E^m$. For $p \in V$, define $Y(p) = \{y \in U \mid F(y,p) \geq 0\}$. Suppose that for some $p_0 \in V$, $Y(p_0)$ contains exactly one element y_0 and that the closure of $\bigcup_{p \in V} Y(p)$ is contained in U. Then for any sequence $\{p^\ell\}$ of elements of V which converges to p_0 and any sequence $\{y^\ell\}$ such that $y^\ell \in Y(p^\ell)$, we have $\{y^\ell\}$ convergent to y_0.

Proof

Since F is continuous, every limit point in U of $\{y^\ell\}$ is y_0; and $\{y^\ell\}$, being a sequence of elements of the bounded set U, is bounded and has all its limit points in U, Q.E.D.

Corollary 2

Let $F:UxV \to E^r$ be continuous where U is an open bounded subset of E^n and $V \subseteq E^m$. Suppose that for every $p \in V$ there is exactly one element, $y(p)$ of U for which $F(y(p),p) \geq 0$ and that the closure of $\bigcup_{p \in V} \{y(p)\}$ is contained in U. Then y is continuous.

Proof

Theorem 1 Q.E.D.

Theorem 3

Let $F = (f_1, \ldots, f_r):UxV \to E^r$ where U is an open subset of E^n, V is an open subset of E^m and each f_i, $i = 1, \ldots, r$ has continuous partial derivatives of UxV. Let $\{p^\ell\}$ be a sequence of elements of V converging to $p^0 \in V$ for which

$$\dot{p} = \lim_{\ell \to \infty} \frac{p^\ell - p^0}{|p^\ell - p^0|}$$

exists. Let $\{y^\ell\}$ be a sequence of elements of U convergent to (see Theorem 1 for conditions which guarantee this) $y^0 \in U$ so that $F(y^\ell,p^\ell) \geq 0$ for $\ell = 1, \ldots$.

Let I be the index set defined by (note that $F(y^0,p^0) \geq 0$, because F is continuous):

$$I = \{i = 1, \ldots, r \mid f_i(y^0,p^0) = 0\}$$

Let F_I be a vector-valued function obtained from F by deleting all components with index $i \notin I$; let $\frac{\partial F_I}{\partial y}$ be the matrix whose i,k*th*

element for $i \in I$ and $k = 1, \ldots, n$ is $\frac{\partial f_i}{\partial y_k}$; let $\frac{\partial F_I}{\partial p}$ be the matrix whose i,k*th* element for $i \in I$ and $k = 1, \ldots, m$ is $\frac{\partial f_i}{\partial p_k}$.

Then

(i) If $\dot{y} = \lim_{\ell \to \infty} \frac{y^\ell - y^0}{|p^\ell - p^0|}$ exists, it satisfies

$$\frac{\partial F_I}{\partial y}(y^0,p^0)\ \dot{y} + \frac{\partial F_I}{\partial p}(y^0,p^0)\ \dot{p} \geq 0 \ . \tag{3.1}$$

(ii) If the system (3.1) of simultaneous inequalities in $\dot{y}$ possesses exactly one solution, $\dot{y}$, then $\lim_{\ell \to \infty} \frac{y^\ell - y^0}{|p^\ell - p^0|}$ exists and is $\dot{y}$.

Proof. Apply the mean value theorem to the function,

$$g(\lambda) = f_i(\lambda y^0 + (1 - \lambda)y^\ell,\ \lambda p^0 + (1 - \lambda)p^\ell) \ ,$$

to see that for all $i = 1, \ldots, r$ and all ℓ there is a $Z^{i,\ell} \in U \times V$ on the line segment joining (y^0,p^0) to (y^ℓ,p^ℓ) for which

$$f_i(y^\ell,p^\ell) - f_i(y^0,p^0) = \frac{\partial f_i}{\partial y}(Z^{i\ell})(y^\ell - y^0) + \frac{\partial f_i}{\partial p}(Z^{i\ell})(p^\ell - p^0) \ .$$

Hence for $i \in I$ (note $f_i(y^0,p^0) = 0$), we have

$$\frac{\partial f_i}{\partial y}(Z^{i\ell})\ \frac{(y^\ell - y^0)}{|p^\ell - p^0|} + \frac{\partial f_i}{\partial p}(Z^{i\ell})\ \frac{p^\ell - p^0}{|p^\ell - p^0|} \geq 0 \ .$$

But $\frac{\partial F_i}{\partial y}$ and $\frac{\partial f_i}{\partial p}$ are continuous and $\lim_{\ell \to \infty} Z^{i\ell} = (y^0,p^0)$. Thus (i) follows at once, and (ii) follows* from Ref. 5, Theorem 2.2, p. 539.

*It is obvious that every convergent subsequence of $\left\{\frac{(y^\ell - y^0)}{|p^\ell - p^0|}\right\}$ converges to a solution of (3.1) and hence to $\dot{y}$, but it is not obvious that it is bounded.

Corollary 4

Let $F = (f_1, \ldots, f_r): U \times V \to E^r$ where U is an open subset of E^m and each f_i, $i = 1, \ldots, r$ has continuous partial derivatives on UxV. Let $y: V \to U$ satisfy $F(y(p),p) \geq 0$ for all $p \in V$ and define the index set I by

$$I = \{i = 1, \ldots, r \mid f_i(y(p_0), p_0) = 0\} .$$

Let F_I be a vector-valued function obtained from F by deleting all components with index $i \notin I$. Let $\frac{\partial F_I}{\partial y}$ be the matrix whose i,k*th* element for $i \in I$ and $k = 1, \ldots, n$ is $\frac{\partial f_i}{\partial y_k}$; let $\frac{\partial F_I}{\partial p}$ be the matrix whose i,k*th* element for $i \in I$ and $k = 1, \ldots, m$ is $\frac{\partial f_i}{\partial p_k}$. Let $\dot{p} \in E^m$.

(i) If the right-hand derivative at $t = 0$ of the function $y(p_0 + t\dot{p})$ exists and is equal to $\dot{y}$; that is, if

$$\dot{y} = \lim_{t \to 0+} \frac{y(p_0 + t\dot{p} - y(p_0))}{t}$$

then (3.1) is satisfied.

And conversely,

(ii) If there is exactly one element $\dot{y}$ of E^n satisfying (3.1), then the right-hand derivative at $t = 0$ of $y(p_0 + t\dot{p})$ exists and is equal to $\dot{y}$.

Proof. Theorem 3, Q.E.D.

Corollary 5

Let F, U, V, p_0, I, $\frac{\partial F_I}{\partial y}$, $\frac{\partial F_I}{\partial p}$ be as in Corollary 4.

(i) Suppose that each y_i, i = 1, ..., n has a gradient at p_0; that is, suppose that $\frac{\partial y}{\partial p}(p_0)$, the matrix whose i,j*th* element, i = 1, ..., n; j = 1, ..., m is $\frac{\partial y_i}{\partial p_j}(p_0)$ exists; and suppose that for all $\dot{p} \in E^m$,

$$\lim_{t \to 0} \frac{y(p_0 + t\dot{p}) - y(p_0)}{t}$$

exists and equals $\frac{\partial y}{\partial p}(p_0)\dot{p}$. Note that if y has continuous partial derivatives in a neighborhood of p_0, it has a gradient at p_0, but not necessarily conversely.

Then

$$\frac{\partial F_I}{\partial y}(y(p_0),p_0)\ \frac{\partial y}{\partial p}(p^0) = -\frac{\partial F_I}{\partial p}(y(p_0),p_0)\ , \tag{5.1}$$

and conversely,

(ii) If the system of simultaneous linear equations

$$\frac{\partial F_I}{\partial y}(y(p_0),p_0)\ X = -\frac{\partial F_I}{\partial p}(y(p_0),p_0) \tag{5.2}$$

has exactly one solution, X, then y has a gradient at p_0 and $\frac{\partial y}{\partial p}(p_0) = X$.

Proof of (i)

By Theorem 3 (i), for all $\dot{p} \in E^n$,

$$\frac{\partial F_I}{\partial y}(y(p_0),p_0)\ \frac{\partial y}{\partial p}(p_0)\dot{p} + \frac{\partial F_I}{\partial p}(y(p_0,p_0))\dot{p} \geq 0 \ ,$$

from which (5.1) follows.

Part (ii) follows even more directly from Theorem 3 (ii). Q.E.D.

REFERENCES

1. Shapiro, N. Z., and L. S. Shapley, "Mass Action Laws and the Gibbs Free Energy Function," *J. Soc. Indust. Appl. Math.*, Vol. 13, No. 2, June 1965, pp. 353-375.

2. Fiacco, A. V., and G. P. McCormack, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley & Sons, Inc., New York, 1968.

3. Mills, H. D., "Marginal Values of Matrix Games and Linear Programs," *Linear Inequalities and Related Systems*, Princeton University Press, New Jersey, 1956, pp. 183-193.

4. Williams, A. C., "Marginal Values in Linear Programming," *SIAM J. Appl. Math.*, Vol. 11, No. 1, 1963, pp. 82-94.

5. Dantzig, G. B., J. H. Folkman, and N. Z. Shapiro, "On the Continuity of the Minimum Set of a Continuous Function," *J. of Mathematical Analysis and Applications*, Vol. 17, No. 3, March 1967.