NOTES ON FUNDAMENTALS OF OPTIMIZATION FOR ENGINEERS

by

E. Polak
(415) 642-2644

Memorandum No. UCB/ERL M89/40

18 April 1989

NOTES ON FUNDAMENTALS OF OPTIMIZATION FOR ENGINEERS

by

E. Polak
(415) 642-2644

Memorandum No. UCB/ERL M89/40

18 April 1989

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

TABLE OF CONTENTS

1. OPTIMIZATION IN ENGINEERING DESIGN

1.1. EVOLUTION OF OPTIMIZATION-BASED ENGINEERING DESIGN

Over the years, engineering design has been increasing in complexity. This constant growth in complexity is due to several factors, such as, (i) progressively increasing expectations in product performance, (ii) progressively more restrictive constraints imposed by environmental and resource cost considerations, and (iii) progressively more and more ambitious projects being launched.

For example, in structural engineering, the increase in design complexity is due to the need to ensure the earthquake survivability of sky scrapers and nuclear reactors at reasonable cost; in control engineering and electronics to the need for reliable, high performance, worst case designs; in the automotive world, to the need to conserve energy while eliminating pollution; and in the area of space exploration, to attempts to design complex shaped, highly flexible, large space structures and their control systems simultaneously, to unprecedented performance standards.

Fortunately, over the last decade, while material and labor costs have grown rapidly, computing costs have decreased dramatically and hence, not surprisingly, engineers have been turning more and more frequently to the computer for assistance in design. As a result, a new, interdisciplinary engineering specialty has emerged which is commonly referred to as computer-aided design (CAD). Most of the existing CAD methodology is based on computer-aided analysis, with the design parameter selection carried out by the designer on a trial and error basis. Since decision making in a multiparameter space is very difficult, the trial and error approach is not very effective. Therefore, there is growing hope that considerable benefits in engineering design might be obtained from the use of sophisticated optimization tools. However, the effective use of optimization algorithms in engineering design is predicated on the supposition that engineering design problems are transcribable into a suitable canonical optimization problem.

Now, as we shall shortly illustrate by example, engineering design specifications can frequently be expressed as inequalities in terms of a finite dimensional *design vector* $x \in \mathbb{R}^n$. These inequalities are either of the form

$$g(x) \leq 0 \tag{1.1.1}$$

where $g:\mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, or of the form

$$\phi(x,y) \leq 0, \ \forall\, y \in Y\,, \tag{1.1.2a}$$

or, equivalently

$$\max_{y \in Y} \phi(x,y) \leq 0\,, \tag{1.1.2b}$$

where $\phi:\mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ is locally Lipschitz continuous and $Y \subset \mathbb{R}^p$ is compact. Constraints of the

form (1.1.1) often express simple bounds on the design variable or a "static" design condition. Constraints of the form (1.1.2b) can be used to express bounds on time and frequency responses of a dynamical system as well as tolerancing or uncertainty conditions in worst case design. Consequently, a rather large number of engineering design problems are transcribable into the following *canonical optimization problem:*

$$\min\{f(x) \mid g^i(x) \leq 0,\ i \in \underline{k};\ \phi^j(x,y_j) \leq 0,\ y_j \in Y_j,\ j \in \underline{m}\} \tag{1.1.3}$$

where we use the notation $\underline{k} \triangleq \{1,2,\cdots,k\}$, for any positive integer k. At a minimum, the functions $f:\mathbb{R}^n \to \mathbb{R}$, $g^i:\mathbb{R}^n \to \mathbb{R}$, $i \in \underline{k}$ and $\phi^j:\mathbb{R}^n \times \mathbb{R}^{p_j} \to \mathbb{R}$, $j \in \underline{m}$, must be assumed to be locally Lipschitz continuous, while the sets $Y_j \subset \mathbb{R}^{p_j}$ must be assumed to be compact.

Occasionally one encounters equality constraints as well, in engineering design. These can be removed by means of exact penalty function techniques.

Problems of the form (1.1.1) are often referred to as *semi-infinite* optimization problems, or SIP for short, because the *design vector* x is finite dimensional, while the number of constraints is infinite.

A number of optimal control problems with state space constraints also have the formal form of (1.1.3), except that the design vector x is a control (in $L^\infty[0,1]$, say) rather than a finite dimensional vector. Although the theory that we will present will be entirely in terms of problems in which the design vector x is finite dimensional, it is very easy to extend the algorithms that we will be presenting, both formally and analytically, to the case where x is a control.

1.2. DESIGN EXAMPLES

We shall now illustrate by means of a few simple examples how SIP problems of the form (1.1.3) arise in a variety of engineering design situations.

1.2.1. Design of Earthquake Resistant Structures

One of the simplest examples of a problem of the form (1.1.3) is found in the design of braced frame buildings which are expected to withstand small earthquakes with no damage and large ones with repairable damage. A simple three story braced frame is shown in Fig 1.2.1.1. The components of the design vector x are the stiffnesses of the frame members, as indicated in Fig.1.2.1.1. Under the hypotheses of a lumped parameter model, the horizontal floors and roof are assumed to be rigid and to concentrate the mass of the structure. The relative displacements of the three floors and roof form the components of the displacement vector y. The lumped parameter model of the braced frame obeys a second order vector differential equation of the form:

$$M\ddot{y}(t,x) + D(y,\dot{y},x)\,\dot{y}(t,x) + K(y,\dot{y},x)\,y(t,x) = F(t), \tag{1.2.1.1}$$

where $F(t)$ represents the seismic forces. When F is small, i.e., when the earthquake is small D and K can be taken to be constant so that (1.2.1.1) is a linear differential equation, but when F is large, the bending of steel introduces gross nonlinearities due to its hysteretic behavior. It is common to consider

a whole family of earthquakes $(F_k)_{k \in K}$, both large and small, in carrying out a design. When an earthquake is small, a building is expected to remain elastic and no structural damage is allowed. When an earthquake is large, survival of occupants becomes a major consideration and large, energy absorbing, non eleastic deformations are accepted, short of outright failure of the structure. A simple optimal design problem consists of minimizing the weight of the structure subject to bounds on the relative floor displacements over the entire duration of the family of earthquakes considered as well as simple bounds on the stiffness of the structural members. This leads to a SIP of the form

$$\begin{aligned} &\min f(x) | 0 < \alpha \le x^i \le \beta, \ \forall\, i \in \underline{n}; \\ &\quad |y^{j+1}(t,x,F_k) - y^j(t,x,F_k)| \le d_k^j, \\ &\quad \forall\, t \in [0,T], \ \forall\, k \in K, \ j = 0,1,2\} \,. \end{aligned} \tag{1.2.1.2}$$

1.2.2. Design of a MIMO Control System

We shall now consider a simple design of a multi-input multi-output (MIMO) control system, with specifications both in time and frequency domains. Consider the feedback configuration in Fig. 1.2.2.1, where $C(x,s)$ is a compensator transfer function matrix that needs to be designed. The equations governing the behavior of this system in the time domain are of the form

$$\dot{z}_p = A_p z_p + B_p u_p \tag{1.2.2.1a}$$

$$y_p = C_p z_p \tag{1.2.2.1b}$$

$$\dot{z}_c = A_c(x) z_c + B_c(x) u_c \tag{1.2.2.2a}$$

$$y_c = C_c(x) z_c \tag{1.2.2.2b}$$

$$u_p = y_c \tag{1.2.2.3a}$$

$$u_c = r - y \tag{1.2.2.3b}$$

$$y = y_p + d \tag{1.2.2.3c}$$

where (1.2.2.1a,b) represents the plant, (1.2.2.2a,b) represents the compensator to be designed and (1.2.2.3a-c) are the interconnection relations. We assume that r, u_p, u_c, y_p, y_c are all m-dimensional vectors and that the matrices A_c, B_c, C_c are continuously differentiable in the design vector x which, most likely, consists of the "free" elements of these matrices.

The most elementary requirement is that of closed loop stability. With

$$G_p(s) = C_p(sI - A_c)^{-1} B_p, \tag{1.2.2.4a}$$

$$G_c(x,s) = C_c(x)(sI - A_c(x))^{-1} B_c(x), \tag{1.2.2.4b}$$

it can be shown that the eigenvalues of the closed loop system are the zeros of the polynomial in s

$$\chi(x,s) \triangleq \det(sI-A_p)\det(sI-A_c(x))\det(I+G_p(s)G_c(x,s)) . \tag{1.2.2.5}$$

To ensure that the zeros of $\chi(x,s)$ are all in the open left half plane, we make use of the modified Nyquist stability test. For this purpose, let $d(s)$ be a monic polynomial of the same degree as $\chi(s)$, such that all zeros of $d(s)$ are in the open left half plane. Let $T(x,s) \triangleq \chi(x,s)/d(s)$. The closed loop system is stable if the locus of $T(x,j\omega)$, traced out in the complex plane for $\omega \in (-\infty,\infty)$, does not pass through or encircle the origin. A sufficient condition for ensuring this consists of keeping the locus of $T(x,j\omega)$ out of a parabolic region containing the origin (see Fig. 1.2.2.1) by imposing the semi-infinite inequality:

$$-dRe[T(x,j\omega)]^2 + Im[T(x,j\omega)] + c \le 0 \ \forall \ \omega \ge 0 . \tag{1.2.2.6}$$

where $c,d > 0$.

Next, for a set of specified inputs $\{r_k(\cdot)\}_{k \in K}$, the designer may require that the zero initial conditions response error be limited as follows (see Fig. 1.2.2.2):

$$\underline{b}^i_k(t) \le y^i_p(t;x,r_k) - r^i_k(t) \le \overline{b}^i_k(t) \tag{1.2.2.7}$$

for all $k \in K$ and $i = 1,2,\cdots,m$, with the $\underline{b}^i_k$, $\overline{b}^i_k$ piecewise continuous functions.

Finally, for the purpose of expressing insensitivity to the disturbance d, we set $r = 0$, which leads to the Laplace transform equation

$$\begin{aligned}\hat{y}(s) &= [I + P(s)\ C(x,s)]^{-1}\ \hat{d}(s) \\ &\triangleq Q(x,s)\ \hat{d}(s)\end{aligned} \tag{1.2.2.8a}$$

$$\begin{aligned}\hat{u}_p(s) &= -C(x,s)\ Q(x,s)\ \hat{d}(s) \\ &\triangleq R(x,s)\ \hat{d}(s) \ ,\end{aligned} \tag{1.2.2.8b}$$

where $\hat{u}_p(s)$, $\hat{d}(s)$, $\hat{y}(s)$ denote the Laplace transforms of $u_p(t)$, $d(t)$, $y(t)$, respectively.

Let $\bar{\sigma}H$ denote the largest singular value of a complex $m\times m$ matrix H. Since the largest singular value of a matrix is its induced L_2 norm, to make the response y of the system small for a large class of disturbances d, without unduly saturating the system as a result of u becoming too large, control system designers strive to keep $\bar{\sigma}[Q(x,j\omega)]$ small and $\bar{\sigma}[R(s,j\omega)]$ bounded over the frequency range $[\omega', \omega'']$ in which the energy of the disturbances is known to be concentrated. This leads to the following formulation of the MIMO control system design problem:

minimize $f(x)$,
where

$$f(x) \triangleq \max\{\bar{\sigma}[Q(x,j\omega)] | \omega \in [\omega',\omega']\} \tag{1.2.2.9}$$

subject to (1.2.2.6), (1.2.2.7) and

$$\bar{\sigma}[R(x,j\omega)] \le b(\omega), \ \forall \ \omega \in [\omega',\omega''], \tag{1.2.2.10}$$

$$\underline{x}^i \le x^i \le \bar{x}^i, \tag{1.2.2.11}$$

where $b(\omega)$ is a continuous, real valued function.

In addition, there could be constraints expressing *decoupling* i.e., the requirement that when only a single component of the input vector is a nonzero function, only the corresponding component of the output vector is nonzero, as well as stability robustness requirements, all of which are semi-infinite in form. We note that from an algorithmic point of view, since singular values are non-differentiable, the optimization problem corresponding to MIMO control system design is considerably more difficult than the one corresponding to structural design.

1.2.3. Design of a Wide Band Amplifier

The design of a wide band amplifier usually involves three transfer functions: the input impedance $Z_{in}(x,s)$, the output impedance, $Z_{out}(x,s)$ and the gain, $A(x,s)$, which are all proper rational functions in the complex variable s. The design vector $x \in \mathbf{R}^n$ determines certain critical component values (e.g., resistor, capacitor values) in the circuit, which affect the impedances and the gain. Thus, the coefficients of the rational functions Z_{in}, Z_{out} and A are functions of the design vector x.

The simplest formulation of a wide band amplifier design has the form

$$\begin{aligned}\max_{(x,\omega_f)} \{\omega_f \mid & \underline{b}_{in} \le |Z_{in}(x,j\omega)|^2 \le \bar{b}_{in}, \ \forall\ \omega \in [\omega_0,\omega_f] \ ; \\ & \underline{b}_{out} \le |Z_{out}(x,j\omega)|^2 \le \bar{b}_{out}, \ \forall\ \omega \in [\omega_0,\omega_f] \ ; \\ & \underline{A} \le |A(s,j\omega)|^2 \le \bar{A}, \ \forall\ \omega \in [\omega_0,\omega_f]; \\ & \underline{x}^i \le x^i \le \bar{x}^i, \ i = 1,2,\cdots,n\} \ .\end{aligned} \tag{1.2.3.1a}$$

As stated, this problem is not quite of the form (1.1.2.3). To bring it in line with the canonical form (1.1.2.3), we augment the design variable by one component, x^0, to $\bar{x} = (x^0,x) \in \mathbf{R}^{n+1}$. Problem (1.2.3.1a) can then seen to be equivalent to the problem:

$$\begin{aligned}\min_{\bar{x}} \{-x^0 \mid & \underline{b}_{in} \le |Z_{in}(x,j(\omega_0 + yx^0))|^2 \le \bar{b}_{in}, \ \forall\ y \in [0,1]; \\ & \underline{b}_{out} \le |Z_{out}(x,j(\omega_0 + yx^0))|^2 \le \bar{b}_{out}, \ \forall\ y \in [0,1]; \\ & \underline{A} \le |A(x,j(\omega_0+yx^0))|^2 \le \bar{A}, \ \forall\ y \in [0,1]; \\ & \underline{x}^i \le x^i \le \bar{x}^i, \ i = 1,2,\ldots,n\} \ .\end{aligned} \tag{1.2.3.1b}$$

1.2.4. Robot Arm Path Planning

In designing a sequence of moves to be carried out by a robot manipulator in a manufacturing situation, it is necessary to find a number of paths which take the robot arm from one location to another without collision with the workpiece. We shall describe a simple problem involving a two link robot manipulator and a circular workpiece obstacle in $\mathbf{R}^2$. Our transcription of this problem into

optimization form is rather simplistic, we refer the reader to for a more sophisticated formulation. Let $\theta^1(t)$, $\theta^2(t)$ be the angles at time t between reference rays and the robot links (see Fig. 1.2.4.1), and let $\theta(t) \triangleq (\theta^1(t), \theta^2(t))$. Then the dynamics of the robot have the form

$$M(\theta(t))\ddot{\theta}(t) = \tau(t) - C(\theta(t),\theta(t))\dot{\theta}(t) + G(\theta(t)) \tag{1.2.4.1}$$

where $M(\cdot)$ and $C(\cdot,\cdot)$ are 2×2 continuously differentiable matrices, and $G:\mathbb{R}^2 \to \mathbb{R}^2$ is continuously differentiable and $\tau(t) \in \mathbb{R}^2$ is a torque vector, with $\tau^1(t)$ the torque applied at the first joint and $\tau^2(t)$ the torque applied at the second joint. The circular workpiece is described by an inequality of the form

$$h(x) \le 0 \ , \tag{1.2.4.2}$$

where $f:\mathbb{R}^2 \to \mathbb{R}$ is defined by

$$h(x) = 1 - (x^1 - a)^2 - (x^2 - b)^2 \ . \tag{1.2.4.3}$$

for some $a,b \in \mathbb{R}$.

Now suppose that we are given that at $t=0$, the angles are $\theta^1(0) = \theta^1_0$, $\theta_2(0)=\theta^2_0$, and that we are supposed to find a torque vector $\tau(t)$, $t \in [0,1]$, which results in a collision free path that takes the robot manipulator from these initial angles to the angles $\theta_1(1) = \theta^1_f$, $\theta^2(1) = \theta^2_f$ at time $t=1$, with $|\tau^j(t)| \le c$, $j = 1,2$, for $t \in [0,1]$. We assume that $\tau(\cdot)$ is an $L^2_\infty[0,1]$ function.

Let us denote the solution of (1.2.4.1), which satisfies the initial condition $\theta(0) = \theta_0$, and which corresponds to the torque $\tau(\cdot)$ by $\theta^\tau(\cdot)$. We can now express our problem in the form

$$\min\{f(\tau) \mid g^j(\tau) \le 0 \ , j = 1,2;\ \phi^k(\tau,y) \le 0,\ k = 1,2,\ \forall\ y \in Y\} \tag{1.2.4.4a}$$

where $f:L^2_\infty[0,1] \to \mathbb{R}$ is defined by

$$f(\tau) \triangleq |\theta^\tau(1) - \theta_f|^2 \ ; \tag{1.2.4.4b}$$

the $g^j:L^2_\infty[0,1] \to \mathbb{R}$, $j = 1,2$ are defined by

$$g^1(\tau) \triangleq \max_{t \in [0,1]} |\tau^1(t)| - c \ ; \tag{1.2.4.4c}$$

$$g^2(\tau) \triangleq \max_{t \in [0,1]} |\tau^2(t)| - c \ ; \tag{1.2.4.4d}$$

$Y = [0,1] \times [0,1] \subset \mathbb{R}^2$ and, for $k = 1,2$, and $y \triangleq (s,t)$, $\phi^k:L^2_\infty[0,1]\times\mathbb{R}^2 \to \mathbb{R}$ are defined, by

$$\phi^1(\tau,y) \triangleq h(sl_1\cos\theta^{1\tau}(t), sl_1\sin\theta^{2\tau}(t)) \tag{1.2.4.4e}$$

$$\begin{aligned}\phi^2(\tau,y) \triangleq h((&l_1\cos\theta^{1\tau}(t) + sl_2\cos(\theta^{1\tau}(t) + \pi - \theta^{2\tau}(t)) \ , \\ &l_1\sin\theta^{1\tau}(t) + sl_2\sin(\theta^{1\tau}(t) + \pi - \theta^{2\tau}(t))\end{aligned} \tag{1.2.4.4f}$$

where l_1 is the length of the first link and l_2 is the length of the second link. The function $\phi^1(\cdot,\cdot)$ is used to ensure that the *entire first link* will avoid collision with the workpiece, while the funcion $\phi^2(\cdot,\cdot)$ is used to insure that the entire second link will avoid collision with the workpiece. As stated, the design vector $\tau(\cdot)$ is a function. The problem can be made finite dimensional by representing $\tau(\cdot)$ in terms of splines, say, over a fixed set of nodes.

2. MATHEMATICAL PRELIMINARIES

2.1. NORMS AND SETS IN $\mathbb{R}^n$

Definition 2.1.1 : A norm in $\mathbb{R}^n$ is a function $|\cdot|:\mathbb{R}^n \to \mathbb{R}_+$ such that

(*i*) $|x| = 0 \Leftrightarrow x = 0$; (2.1.1a)

(*ii*) $|\lambda x| = |\lambda||x|, \forall\ \lambda \in \mathbb{R}, x \in \mathbb{R}^n$; (2.1.1b)

(*iii*) $|x + y| \leq |x| + |y|, \forall\ x, y \in \mathbb{R}^n$. (2.1.1c)

■

Exercise 2.1.1 : Show that the following three functions are all norms:

$$|x|_2 \triangleq \left[\sum_{i=1}^{n} (x^i)^2\right]^{1/2} \tag{2.1.2a}$$

$$|x|_\infty = \max_{i \in \underline{n}} |x^i| \tag{2.1.2b}$$

$$|x|_1 = \sum_{i=1}^{n} |x^i|, \tag{2.1.2c}$$

where we used the notation

$$\underline{n} \triangleq \{1,2,....,n\}. \tag{2.1.2d}$$

■

Exercise 2.1.2 : Show that there are finite constants $K_{\infty,2}, K_{2,\infty}, K_{\infty,1}, K_{1,\infty}, K_{2,1}, K_{1,2}$, such that

$$|x|_\infty \leq K_{\infty,2}|x|_2, \quad |x|_2 \leq K_{2,\infty}|x|_\infty, \tag{2.1.3a}$$

$$|x|_\infty \leq K_{\infty,1}|x|_1, \quad |x|_1 \leq K_{1,\infty}|x|_\infty, \tag{2.1.3b}$$

$$|x|_2 \leq K_{2,1}|x|_1, \quad |x|_1 \leq K_{1,2}|x|_2. \tag{2.1.3c}$$

■

Definition 2.1.2 : For any $x \in \mathbb{R}^n$ and $\rho > 0$, we denote by

$$\overset{\circ}{B}(x,\rho) \triangleq \{x' \in \mathbb{R}^n \mid |x' - x| < \rho\}, \tag{2.1.4a}$$

the *open* ball of radius ρ about x, and we denote by

$$B(x,\rho) \triangleq \{x' \in \mathbb{R}^n \mid |x' - x| \leq \rho\}, \tag{2.1.4b}$$

the *closed* ball of radius ρ about x. ■

Definition 2.1.3 : A set $X \subset \mathbb{R}^n$ is said to be *open*, if for every $x \in X$, there exists a $\rho > 0$ such that $B(x,\rho) \subset X$. A set $X \subset \mathbb{R}^n$ is said to be *closed* if X^c, its complement in $\mathbb{R}^n$, is open. ■

Exercise 2.1.3 : Let $x \in \mathbb{R}^n$ and $\rho > 0$ be given. Show that $\overset{\circ}{B}(x,\rho)$ is open and that $B(x,\rho)$ is closed. ■

Exercise 2.1.4 : Show that $X \subset \mathbb{R}^n$ is closed $\Leftrightarrow$ for every $x \in \mathbb{R}^n$, if $B(x,\rho) \cap X \neq \emptyset$ for all $\rho > 0$, then $x \in X$. ■

Definition 2.1.4 : A set $X \subset \mathbb{R}^n$ is said to be *compact* if X is closed and bounded, i.e., there exists an $M < \infty$ such that $|x| \leq M$ for all $x \in X$. ■

2.2. SEQUENCES

Let the set of nonnegative integers be denoted by $\mathbb{N}$, i.e.,

$$\mathbb{N} \triangleq \{0,1,2,.....\}. \tag{2.2.1}$$

Definition 2.2.1 : A *sequence* in $\mathbb{R}^n$ is a function from $\mathbb{N}$ into $\mathbb{R}^n$. We denote a sequence by the set of its values, i.e., $\{x_i\}_{i=0}^{\infty}$ or $\{x_i\}_{i \in \mathbb{N}}$. A *subsequence* of $\{x_i\}_{i \in \mathbb{N}}$ is a sequence of the form $\{x_i\}_{i \in K}$, where K is an *infinite* subset of $\mathbb{N}$. ■

Definition 2.2.2 : A sequence $\{x_i\}_{i \in \mathbb{N}}$ in $\mathbb{R}^n$ is said to *converge* to a point $\hat{x}$ ($x_i \to \hat{x}$ as $i \to \infty$) if $\lim_{i \to \infty} |x_i - \hat{x}| = 0$. The point $\hat{x}$ is called a *limit* point of $\{x_i\}_{i \in \mathbb{N}}$. A point x^* is said to be an *accumulation* point of a sequence $\{x_i\}_{i \in \mathbb{N}}$ in $\mathbb{R}^n$, if there exists an infinite subset $K \subset \mathbb{N}$ such that $\lim_{\substack{i \to \infty \\ i \in K}} |x_i - x^*| = 0$ ($x_i \overset{K}{\to} x^*$). ■

Exercise 2.2.1 : Suppose that $x_i \to \hat{x}$ as $i \to \infty$, show that for every $\rho > 0$ there exists an $i_\rho \in \mathbb{N}$ such that $x_i \in B(\hat{x},\rho)$ for all $i \geq i_\rho$. ■

Exercise 2.2.2 : Suppose that $\hat{x}, \hat{x}'$ are limit points of a sequence $\{x_i\}_{i \in \mathbb{N}}$. Show that $\hat{x} = \hat{x}'$ must hold. ■

Exercise 2.2.3 :

(a) Show that a set $X \subset \mathbb{R}^n$ is open $\Leftrightarrow$ for any $\hat{x} \in X$ and any sequence $\{x_i\}_{i \in \mathbb{N}} \subset \mathbb{R}^n$, such that $x_i \to \hat{x}$ as $i \to \infty$, there exists a $q \in \mathbb{N}$ such that $x_i \in X$ for all $i \geq q$.

(b) Show that a set $X \subset \mathbb{R}^n$ is closed $\Leftrightarrow$ for all $\{x_i\}_{i \in \mathbb{N}} \subset X$, if $x_i \to \hat{x}$ as $i \to \infty$, then $\hat{x} \in X$. ■

Theorem 2.2.1 (Bolzano-Weierstrass): Suppose $X \subset \mathbb{R}^n$ is compact and $\{x_i\}_{i \in \mathbb{N}} \subset X$. Then $\{x_i\}_{i \in \mathbb{N}}$ must have at least one accumulation point. ■
(For a proof of this result see a book on analysis.)

In proving convergence of algorithms we shall need the following special property of monotone sequences.

Proposition 2.2.1 : Suppose that $\{x_i\}_{i \in \mathbf{N}}$ is a sequence in $\mathbf{R}$ such that $x_0 \geq x_1 \geq x_2 \geq \ldots$ (i.e., it is monotone decreasing). If $\{x_i\}_{i \in \mathbf{N}}$ has an accumulation point x^*, then $x_i \to x^*$ as $i \to \infty$, i.e., x^* is a limit point.

Proof : For the sake of contradiction, suppose that $\{x_i\}_{i \in \mathbf{N}}$ does not converge to x^*. Then, for some $\rho > 0$, there exists a subsequence $\{x_i\}_{i \in K}$ such that $x_i \notin \mathbf{B}(x^*, \rho)$ for all $i \in K$, i.e., $|x_i - x^*| > \rho$ for all $i \in K$. Since x^* is an accumulation point, there exists a subsequence $\{x_i\}_{i \in K^*}$ such that $x_i \xrightarrow{K^*} x^*$ as $i \to \infty$. Hence there is an $i_1 \in K^*$ such that $|x_i - x^*| \leq \rho/2$, for all $i \geq i_1$, $i \in K^*$.

Let $i_2 \in K$ be such that $i_2 > i_1$. Then we must have that $x_{i_2} < x_{i_1}$ and $|x_{i_2} - x^*| \geq \rho$, which leads to the conclusion that $x_{i_2} < x^*$ and $x_{i_1} - x_{i_2} \geq \rho/2$. Now let $i_3 \in K^*$ be such that $i_3 > i_2$. Then we must have that $x_{i_3} < x_{i_2} < x^*$ and hence that $|x_{i_3} - x^*| > \rho$. But this contradicts the fact that $|x_{i_3} - x^*| \leq \rho/2$ by construction, and hence we conclude that $x_i \to x^*$ as $i \to \infty$. ■

Corollary 2.2.1 : Suppose that $\{x_i\}_{i \in \mathbf{N}}$ is a monotone decreasing sequence in $\mathbf{R}$. If there exists a $b \in \mathbf{R}$ such that $x_i \geq b$ for all $i \in \mathbf{N}$, then $\{x_i\}_{i \in \mathbf{N}}$ converges to some $x^* \in \mathbf{R}$. ■

Exercise 2.2.4 : Prove corollary 2.2.1. ■

2.3. CONTINUITY

We now summarize the most essential properties of continuous functions.

Definition 2.3.1 : A function $f: \mathbf{R}^n \to \mathbf{R}^m$ is said to be *continuous at a point* $x \in \mathbf{R}^n$, if for every $\delta > 0$ there exists a $\rho > 0$ such that

$$|f(x') - f(x)| < \delta \quad \forall \quad x' \in \mathring{\mathbf{B}}(x, \rho). \tag{2.3.1}$$

A function $f: \mathbf{R}^n \to \mathbf{R}^m$ is said to be *continuous* if it is continuous at all $x \in \mathbf{R}^n$. ■

Exercise 2.3.1 : Show that $f: \mathbf{R}^n \to \mathbf{R}^m$ is continuous at $\hat{x} \Leftrightarrow$ for any sequence $\{x_i\}_{i \in \mathbf{N}}$ in $\mathbf{R}^n$ such that $x_i \to \hat{x}$ as $i \to \infty$, $f(x_i) \to f(\hat{x})$ as $i \to \infty$. ■

Definition 2.3.2 : A function $f: \mathbf{R}^n \to \mathbf{R}^m$ is said to be *uniformly continuous* on a subset $\mathbf{X} \subset \mathbf{R}^n$ if for any $\delta > 0$ there exists a $\rho > 0$ such that for any x', $x'' \in \mathbf{X}$ satisfying $|x' - x''| < \rho$,

$$|f(x') - f(x'')| < \delta. \tag{2.3.2}$$

■

Proposition 2.3.1 : Suppose that $f: \mathbf{R}^n \to \mathbf{R}^m$ is continuous and that $\mathbf{X} \subset \mathbf{R}^n$ is compact. Then $f(\cdot)$ is uniformly continuous on $\mathbf{X}$.

Proof : For the sake of contradiction, suppose that $f(\cdot)$ is *not* uniformly continuous on $\mathbf{X}$. Then, for some $\delta > 0$, there exist sequences $\{x'_i\}_{i \in \mathbf{N}}$, $\{x''_i\}_{i \in \mathbf{N}}$ in $\mathbf{X}$ such that

$$\| x'_i - x''_i \| < \frac{1}{i}, \forall\, i \in \mathbf{N}, \tag{2.3.3a}$$

but

$$\| f(x'_i) - f(x''_i) \| > \delta, \forall\, i \in \mathbf{N}. \tag{2.3.3b}$$

Since $\mathbf{X}$ is compact, there must exist a subsequence $\{x'_i\}_{i \in K}$ such that $x'_i \overset{K}{\to} x^* \in \mathbf{X}$ as $i \to \infty$. Furthermore, because of (2.3.3a), $x''_i \overset{K}{\to} x^*$ as $i \to \infty$ also holds. Hence, since $f(\cdot)$ is continuous, we must have $f(x'_i) \overset{K}{\to} f(x^*)$ and $f(x''_i) \overset{K}{\to} f(x^*)$ as $i \to \infty$. Therefore, there exists a $i_0 \in K$ such that for all $i \in K, i \geq i_0$.

$$\| f(x'_i) - f(x''_i) \| \leq \| f(x'_i) - f(x^*) \| + \| f(x^*) - f(x''_i) \| < \delta/2, \tag{2.3.4}$$

contradicting (2.3.3b). This completes our proof. ■

Proposition 2.3.2 : Suppose that $\mathbf{X} \subset \mathbf{R}^n$ is compact and that $f: \mathbf{R}^n \to \mathbf{R}^m$ is continuous. Then the set

$$f(\mathbf{X}) \triangleq \{y \in \mathbf{R}^m \mid y = f(x), x \in \mathbf{X}\} \tag{2.3.5}$$

is compact.

Proof : (a) First we show that $f(\mathbf{X})$ is closed. Thus, let $\{f(x_i)\}_{i \in \mathbf{N}}$, with $x_i \in \mathbf{X}$, be any sequence in $f(\mathbf{X})$ such that $f(x_i) \to y$ as $i \to \infty$. Since $\{x_i\}_{i \in \mathbf{N}}$ is in a compact set $\mathbf{X}$, there exists a subsequence $\{x_i\}_{i \in K}$ such that $x_i \overset{K}{\to} x^* \in \mathbf{X}$ as $i \to \infty$. Since $f(\cdot)$ is continuous, $f(x_i) \overset{K}{\to} f(x^*)$ as $i \to \infty$. But y is the limit point of $\{f(x_i)\}_{i \in \mathbf{N}}$ and hence it is the limit point of any subsequence of $\{f(x_i)\}_{i \in \mathbf{N}}$. We conclude that $y = f(x^*)$ and hence that $y \in f(\mathbf{X})$, i.e., $f(\mathbf{X})$ is closed.

(b) Next, we prove that $f(\mathbf{X})$ is bounded. Suppose $f(\mathbf{X})$ is not bounded. Then there exists a sequence $\{x_i\}_{i \in \mathbf{N}}$ such that $\| f(x_i) \| \geq i$ for all $i \in \mathbf{N}$. Now, since $\{x_i\}_{i \in \mathbf{N}}$ is in a compact set, there exists a subsequence $\{x_i\}_{i \in K}$ such that $x_i \overset{K}{\to} x^*$, as $i \to \infty$, with $x^* \in \mathbf{X}$, and $f(x_i) \overset{K}{\to} f(x^*)$, as $i \to \infty$, by continuity of $f(\cdot)$. Hence there exists an i_0 such that for any $j > i > i_0$, $j, i \in K$,

$$\|f(x_j) - f(x_i)\| \leq \|f(x_j) - f(x^*)\| + \|f(x_i) - f(x^*)\| < 1/2. \tag{2.3.6}$$

Let $i \geq i_0$ be given. By hypothesis there exists a $j \in K, j \geq i$ such that $\|f(x_j)\| \geq j \geq \|f(x_i)\| + 1$. Hence

$$\|f(x_j) - f(x_i)\| \geq \| \|f(x_j)\| - \|f(x_i)\| \| \geq 1, \tag{2.3.7}$$

which contradicts (2.3.6). Thus $f(\mathbf{X})$ must be bounded. This completes our proof. ■

Proposition 2.3.3 : Suppose that $f: \mathbf{R}^n \to \mathbf{R}$ is continuous and that $\mathbf{X} \subset \mathbf{R}^n$ is compact. Then there exists an $\hat{x} \in \mathbf{X}$ such that

$$f(\hat{x}) = \inf_{x \in \mathbf{X}} f(x), \tag{2.3.8}$$

i. e., $\min_{x \in \mathbf{X}} f(x)$ is well defined.

Proof : Since X is compact, $f(X)$ is bounded. Hence $\inf_{x \in X} f(x) = \alpha$ is finite. Let $\{x_i\}_{i \in \mathbf{N}}$ be a sequence in X such that $f(x_i) \downarrow \alpha$ as $i \to \infty$. Since X is compact, there exists a converging subsequence $\{x_i\}_{i \in K}$ such that $x_i \overset{K}{\to} \hat{x} \in X$. By continuity, $f(x_i) \overset{K}{\to} f(\hat{x})$ as $i \to \infty$. It now follows from Proposition 2.2.1 that $f(x_i) \to f(\hat{x})$ as $i \to \infty$. Since $\{f(x_i)\}_{i \in \mathbf{N}}$ has a unique limit point, we conclude that $f(\hat{x}) = \alpha$. ■

Exercise 2.3.2 : Prove Proposition 2.3.3 by making use of the fact that $f(X)$ is compact. ■

2.4. DERIVATIVES

We shall now present a few results involving derivatives that we will need in our study of optimization.

Definition 2.4.1 : Let $f:\mathbf{R}^n \to \mathbf{R}^m$. We say that $Df:\mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}^m$ is a *differential* for $f(\cdot)$ at $\hat{x} \in \mathbf{R}^n$ if

a) $Df(\hat{x};\cdot)$ is linear.

$$b) \quad \lim_{\|h\| \to 0} \frac{\|f(\hat{x}+h) - f(\hat{x}) - Df(\hat{x};h)\|}{\|h\|} = 0 . \tag{2.4.1}$$

When $f:\mathbf{R}^n \to \mathbf{R}^m$ has a differential at all $x \in \mathbf{R}^n$, we say that $f(\cdot)$ is *differentiable*. ■

Since $Df(\hat{x};\cdot)$ is a linear map from $\mathbf{R}^n$ into $\mathbf{R}^m$, there exists an $m \times n$ matrix $\partial f(\hat{x})/\partial x$ such that $Df(\hat{x};h) = \partial f(\hat{x})/\partial x h$ for all $h \in \mathbf{R}^n$; $\partial f(\hat{x})/\partial x$ is called a *Jacobian* matrix.

When $f:\mathbf{R}^n \to \mathbf{R}$ is differentiable, we use the notation $\nabla f(x) = \partial f(x)^T/\partial x$, and call $\nabla f(\cdot)$ the *gradient* of $f(\cdot)$.

Proposition 2.4.1 : Suppose that the function $f:\mathbf{R}^n \to \mathbf{R}^m$ has a differential $Df(\hat{x};h)$ at $\hat{x}$. Then the ij–th component of the Jacobian $\partial f(\hat{x})/\partial x$ is the partial derivative $\partial f^i(\hat{x})/\partial x^j$.

Proof : Set $h = te_j$, where e_j is the j–th unit vector in $\mathbf{R}^n$. Then $\partial f(\hat{x})/\partial x e_j = \left[\partial f(\hat{x})/\partial x\right]_{\cdot j}$, the j–th column of $\partial f(\hat{x})/\partial x$, and hence, from (2.4.1), for $j = 1,2,\cdots,n$,

$$\lim_{t \to 0} \frac{\left| f^i(\hat{x} + te_j) - f^i(\hat{x}) - t\left[\dfrac{\partial f(\hat{x})}{\partial x}\right]_{ij} \right|}{t} = 0, \tag{2.4.2}$$

i.e., $\left[\partial f(\hat{x})/\partial x\right]_{ij} = \partial f^i(\hat{x})/\partial x^j$. ■

Definition 2.4.2 : We say that $f:\mathbf{R}^n \to \mathbf{R}^m$ is *locally Lipschitz continuous* at $\hat{x}$ if there exist $L \in [0,\infty)$, $\hat{\rho}>0$ such that

$$\|f(x) - f(x')\| \le L\|x - x'\|, \ \forall\ x,x' \in B(\hat{x},\hat{\rho}) . \tag{2.4.3a}$$

■

Exercise 2.4.1 : Suppose that $f:\mathbf{R}^n \to \mathbf{R}^m$ has a continuous differential $Df(\cdot,\cdot)$ in a neighborhood of

$\hat{x}$. Show that f is locally Lipschitz continuous at $\hat{x}$. ■

It should be noted that the existence of partial derivatives does not ensure the existence of a differential (see e.g. Apostol p. 103). Thus consider the function

$$f(x,y) = x + y \text{ if } x = 0 \text{ or } y = 0, \tag{2.4.3b}$$

$$f(x,y) = 1 \text{ otherwise .} \tag{2.4.3c}$$

In this case

$$\frac{\partial f(0,0)}{\partial x} = \lim_{t \to 0} \frac{f(t,0) - f(0,0)}{t} = 1, \tag{2.4.4a}$$

$$\frac{\partial f(0,0)}{\partial y} = \lim_{t \to 0} \frac{f(0,t) - f(0,0)}{t} = 1, \tag{2.4.4b}$$

but the function is not even continuous at (0,0). In view of this, the following result is of interest (see Apostol p. 118).

Proposition 2.4.2 : Consider a function $f:\mathbb{R}^n \to \mathbb{R}^m$ such that the partial derivatives $\partial f^i(x)/dx^j$ exist in a neighborhood of $\hat{x}$, for $i = 1,2,\cdots,n$, $j = 1,2,\cdots,m$. If these partial derivatives are *continuous* at $\hat{x}$, then the differential $Df(\hat{x};h)$ exists. ■

The following *chain rule* holds.

Proposition 2.4.3 : Suppose that $f:\mathbb{R}^n \to \mathbb{R}^m$ is defined by $f(x) = h(g(x))$ with both $h:\mathbb{R}^l \to \mathbb{R}^m$ and $g:\mathbb{R}^n \to \mathbb{R}^l$ differentiable. Then

$$\frac{\partial f(\hat{x})}{\partial x} = \frac{\partial h(g(\hat{x}))}{\partial x} \frac{\partial g(\hat{x})}{\partial x} . \tag{2.4.5}$$
■

We make frequent use of Taylor's formula with remainder up to order 2. It comes in two forms: the first is in terms of an intermediate point, while the second one is in integral form (see Apostol[1] p. 124 and Dieudonne[2] p. 186. Also, refer to Apostol p. 124 for exposition on higher order differentials). We denote by $D^k f(\cdot;\cdot)$ the differential of order k of $f(\cdot)$.

Proposition 2.4.4 : Consider a function $f:\mathbb{R}^n \to \mathbb{R}$. Suppose that $f(\cdot)$ has continuous partial derivatives of order p at each point x of $\mathbb{R}^n$. Then for any $x,y \in \mathbb{R}^n$

$$f(y) - f(x) = \sum_{k=1}^{p-1} \frac{1}{k!} D^k f(x;y-x) + \frac{1}{p!} D^p f(z;y-x), \tag{2.4.6a}$$

for some $z = x + t(y - x)$, $t \in [0,1]$. ■

When $p = 1$, we recognize (2.4.6a) as being simply the mean value theorem. For $p = 2$, $D^2 f(x;y-x) = \langle y-x, \partial^2 f(x)/\partial x^2 (y-x)\rangle$ where $\partial^2 f(x)/\partial x^2$ is a matrix of second partial derivatives, i.e., $\left[\partial^2 f(x)/\partial x^2\right]_{ij} = \partial^2 f(x)/\partial x^i \partial x^j$.

[1] T. M. Apostol, *Mathematical Analysis*, Addision-Wesley, 1960

For functions $f:\mathbb{R}^n \to \mathbb{R}^m$, with $m > 1$, formula (2.4.6a) is not valid since there is no z of the form stated that works for all the components of $f(\cdot)$. Instead we use the following result (see Dieudonne p. 186).

Proposition 2.4.5 : Consider a function $f:\mathbb{R}^n \to \mathbb{R}^m$. Suppose that $f(\cdot)$ has continuous partial derivatives of order p at each point x of $\mathbb{R}^n$. Then for any $x,y \in \mathbb{R}^n$,

$$f(y) - f(x) = \sum_{k=1}^{p-1} \frac{1}{k!} D^k f(x;y-x) + \frac{1}{(p-1)!}\int_0^1 (1-s)^{p-1} D^p f(x+s(y-x);y-x)ds\,. \tag{2.4.6b}$$

Proof : We shall prove (2.4.6b) only for $p \le 2$. For $p = 1$, consider the function $g(s) = f(x + s(y - x))$. Then $g(1) = f(y)$, $g(0) = f(x)$ and

$$g(1) - g(0) = \int_0^1 g'(s)ds$$

$$= \int_0^1 Df(x + s(y - x);y - x)ds, \tag{2.4.7a}$$

which completes the proof for $p = 1$.

Next, let $p = 2$. Then we have

$$g''(s)(1 - s) = \frac{d}{ds}[g'(s)(1 - s) + g(s)]\,. \tag{2.4.7b}$$

Integrating (2.4.7b) from 0 to 1 we get

$$g(1) - g(0) - g'(0) = \int_0^1 (1 - s)g''(s)ds, \tag{2.4.7c}$$

which, on rearranging, we recognize as being

$$f(y) - f(x) = \langle \nabla f(x), y - x\rangle + \int_0^1 (1 - s)\, D^2 f(x + s(y - x);(y - x))ds, \tag{2.4.7d}$$

after substitution for $g(s)$. ■

Finally, we define directional derivatives which may exist even when a function fails to have a differential.

Definition 2.4.3 : Let $f:\mathbb{R}^n \to \mathbb{R}^m$. We define the *directional derivative* of $f(\cdot)$ at a point $\hat{x} \in \mathbb{R}^n$ in the direction $h \in \mathbb{R}^n$ $(h \neq 0)$ by

$$df(\hat{x};h) \triangleq \lim_{\substack{t \downarrow 0 \\ t>0}} \frac{f(\hat{x} + th) - f(\hat{x})}{t}, \tag{2.4.8}$$

if this limit exists. Note that $t > 0$ is required. ■

---

[2] J. Dieudonne, *Foundations of Modern Analysis*, Academic Press, 1960.

Exercise 2.4.2 : Suppose that $f:\mathbb{R}^n \to \mathbb{R}^m$ has a differential at $\hat{x}$. Show that for any h, the directional derivative $df(\hat{x};h)$ exists and is given by

$$df(\hat{x};h) = Df(\hat{x};h) = \frac{\partial f(x)}{\partial x} h \, .$$

■

As we shall see later, directional derivatives play a very important part in the theory of optimization.

2.5. CONVEX SETS AND CONVEX FUNCTIONS

Convexity is an enormous subject (e.g. see Rockafellar[3]). We collect here only a few essential results that we will need in our study of optimization (For further details see Rockafellar). We begin with convex sets.

Definition 2.5.1 : A set $S \subset \mathbb{R}^n$ is said to be *convex* if for any $x', x'' \in S$ and $\lambda \in [0,1]$, $[\lambda x' + (1-\lambda)x''] \in S$. ■

Exercise 2.5.1 : Suppose $S \subset \mathbb{R}^n$ is convex. Let $\{x_i\}_{i=1}^k$ be points in S and let $\{\mu^i\}_{i=1}^k$ be scalars such that $\mu^i \geq 0$ for $i = 1,2,\cdots,k$ and $\sum_{i=1}^k \mu^i = 1$. Show that

$$\left[\sum_{i=1}^k \mu^i x_i\right] \in S \, . \tag{2.5.1}$$

■

Definition 2.5.2 : Let S be a subset of $\mathbb{R}^n$. We say that coS is the *convex hull* of S if it is the smallest convex set containing S. ■

Theorem 2.5.1 (Caratheodory) : Let S be a subset of $\mathbb{R}^n$. If $\bar{x} \in$ co#S, then there exists at most $(n+1)$ distinct points, $\{x_i\}_{i=1}^{n+1}$, in S such that $\bar{x} = \sum_{i=1}^{n+1} \mu^i x_i$, $\mu^i > 0$, $\sum_{i=1}^{n+1} \mu^i = 1$.

Proof : Consider the set

$$C_S \triangleq \{x \mid x = \sum_{i=1}^{k_x} \mu^i x_i \, , \, x_i \in S \, , \, \mu^i \geq 0 \, , \, \sum_{i=1}^{k_x} \mu^i = 1, \, k_x \in \mathbb{N}\} \, . \tag{2.5.2}$$

First, it is clear that $S \subset C_S$. Next, since for any x', $x'' \in C_S$, $\lambda x' + (1-\lambda x'') \in C_S$, for $\lambda \in [0,1]$, it follows that C_S is convex. Hence we must have that $coS \subset C_S$. However, because C_S consists of all the convex combinations of poins in S, we must also have that $C_S \subset coS$. Hence $C_S = coS$.

Now suppose that

$$\bar{x} = \sum_{i=1}^{\bar{k}} \Pi^i x_i \, , \tag{2.5.3}$$

with $\Pi^i \geq 0$, $i = 1,2,\cdots,\bar{k}$, $\sum_{i=1}^{\bar{k}} \Pi^i = 1$. Then the following system of equations is satisfied

$$\sum_{i=1}^{\bar{k}} \pi^i \begin{bmatrix} x_i \\ 1 \end{bmatrix} = \begin{bmatrix} \bar{x} \\ 1 \end{bmatrix}, \tag{2.5.4}$$

with $\pi^i \geq 0$. Suppose that $\bar{k} > n+1$. Then there exist coefficients α^j, $j = 1,2,\cdots,\bar{k}$, not all zero, such that

$$\sum_{i=1}^{\bar{k}} \alpha^i \begin{bmatrix} x_i \\ 1 \end{bmatrix} = 0\,. \tag{2.5.5}$$

Adding (2.5.5) multiplied by θ to (2.5.4) we get

$$\sum_{i=1}^{\bar{k}} (\pi^i + \theta\alpha^i) \begin{bmatrix} x_i \\ 1 \end{bmatrix} = \begin{bmatrix} \bar{x} \\ 1 \end{bmatrix}. \tag{2.5.6}$$

Suppose (w.l.o.g.) that at least one $\alpha^i < 0$. Then there exists a $\bar{\theta} > 0$ such that $\pi^j + \bar{\theta}\alpha^j = 0$ for some j while $\pi^i + \bar{\theta}\alpha^i \geq 0$ for all other i. Thus we have succeeded in expressing $\bar{x}$ as a convex combination of $\bar{k} - 1$ vectors in S. Clearly, these reductions can go on as long as $\bar{x}$ is expressed in terms of more than $(n+1)$ vectors in S. This completes our proof. ■

Definition 2.5.5 : Let S_1, S_2 be any two sets in $\mathbb{R}^n$. We say that the hyperplane

$$H = \{x \in \mathbb{R}^n \mid \langle x, v\rangle = \alpha\} \tag{2.5.7}$$

separates S_1 and S_2 if

$$\langle x, v\rangle \geq \alpha \;\; \forall \;\; x \in S_1 \tag{2.5.8a}$$

$$\langle y, v\rangle \leq \alpha \;\; \forall \;\; y \in S_2 \tag{2.5.8b}$$

The separation is said to be *strong* if there exists an $\varepsilon > 0$ such that

$$\langle x, v\rangle \geq \alpha + \varepsilon \;\; \forall \; x \in S_1 \tag{2.5.8c}$$

$$\langle y, v\rangle \leq \alpha - \varepsilon \;\; \forall \; y \in S_2 \tag{2.5.8d}$$

Theorem 2.5.2 (Separation of Convex Sets) : Let S_1, S_2 be two convex sets in $\mathbb{R}^n$ such that $S_1 \cap S_2 = \emptyset$ Then there exists a hyperplane which separates S_1 and S_2. Furthermore, if S_1 and S_2 are closed and either S_1 or S_2 is compact, then the separation can be made strict. ■

Theorem 2.5.3 : Suppose that $S \subset \mathbb{R}^n$ is closed and convex and $0 \notin S$. Let

$$\hat{x} = \text{argmin}\{|x|^2 \mid x \in S\}\,. \tag{2.5.9}$$

Then

$$H = \{x \mid \langle \hat{x}, x\rangle = |\hat{x}|^2\} \tag{2.5.10}$$

separates S from 0, i.e., $\langle \hat{x}, x\rangle \geq |\hat{x}|^2$ for all $x \in S$.

---

[3] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

Proof : Let $x \in S$ be arbitrary. Then, since S is convex, $[\hat{x} + \lambda(x - \hat{x})] \in S$ for all $\lambda \in [0,1]$. By definition of $\hat{x}$, we must have

$$0 < |\hat{x}|^2 \leq |\hat{x} + \lambda(x - \hat{x})|^2$$

$$= |\hat{x}|^2 + 2\lambda\langle\hat{x}, x - \hat{x}\rangle + \lambda^2|x - \hat{x}|^2 . \tag{2.5.11a}$$

Hence, for all $\lambda \in (0,1]$,

$$0 \leq 2\langle\hat{x}, x - \hat{x}\rangle + \lambda|x - \hat{x}|^2 . \tag{2.5.11b}$$

Letting $\lambda \to 0$ we get the desired result. ■

Theorem 2.5.3 can be used to prove the following special case of Theorem 2.5.2:

Corollary 2.5.1 : Let S_1, S_2 be two compact convex sets in $\mathbb{R}^n$ such that $S_1 \cap S_2 = \emptyset$. Then there exists a hyperplane which separates S_1 and S_2.

Proof : Let $C = S_1 - S_2 \triangleq \{ x \in \mathbb{R}^n \mid x = x_1 - x_2, x_1 \in S_1, x_2 \in S_2 \}$. Then C is convex and compact and $0 \notin C$. Let $\hat{x} = (\hat{x}_1 - \hat{x}_2) = \text{argmin}\{ |x|^2 \mid x \in C \}$, where $\hat{x}_1 \in S_1$ and $\hat{x}_2 \in S_2$. Then, by Theorem 2.5.3,

$$\langle x - \hat{x}, \hat{x}\rangle \geq 0, \forall\, x \in C . \tag{2.5.12a}$$

Let $x = x_1 - \hat{x}_2$, with $x_1 \in S_1$. Then (2.5.12a) leads to

$$\langle x_1 - \hat{x}_2, \hat{x}\rangle \geq |\hat{x}|^2, \forall\, x_1 \in C_1, \tag{2.5.12b}$$

and, for $x = \hat{x}_1 - x_2$, with $x_2 \in S_2$,

$$\langle \hat{x}_1 - x_2, \hat{x}\rangle \geq |\hat{x}|^2, \tag{2.5.12c}$$

which implies that

$$\langle (\hat{x}_1 - \hat{x}_2) + \hat{x}_2 - x_2, \hat{x}\rangle \geq |\hat{x}|^2, \tag{2.5.12d}$$

i.e., that

$$\langle x_2 - \hat{x}_2, \hat{x}\rangle \leq 0, \tag{2.5.12e}$$

which completes our proof. ■

Definition 2.5.6 : Suppose $S \subset \mathbb{R}^n$ is convex. We say that $H = \{x \mid \langle x - \bar{x}, v\rangle = 0\}$ is a *support hyperplane* to S through $\bar{x}$ with *inward (outward) normal* v if $\bar{x} \in \bar{S}$ (the closure of S) and

$$\langle x - \bar{x}, v\rangle \geq 0 \ (\leq 0) \ \forall \ x \in S . \tag{2.5.12f}$$

■

Theorem 2.5.4 : A closed convex set is equal to the intersection of the half spaces which contain it.

Proof : Let C be a closed convex set and A the intersection of half spaces containing C. Then

clearly $C \subset A$. Now suppose $\bar{x} \notin C$. Then there exists a support hyperplane H which separates strictly $\bar{x}$ and C, i.e., $\bar{x}$ does not belong to one subspace containing C, i.e., $\bar{x} \notin A$. Hence $C^c \subset A^c$ which leads to the conclusion that $A \subset C$. ■

Next we turn to convex functions. For an example see Fig. 2.5.1.

Definition 2.5.7 : A function $f:\mathbb{R}^n \to \mathbb{R}$ is said to be *convex* if for any $x',x'' \in \mathbb{R}^n$ and $\lambda \in [0,1]$,

$$f(\lambda x' + (1-\lambda)x'') \leq \lambda f(x') + (1-\lambda)f(x'') \tag{2.5.13}$$

A function $f:\mathbb{R}^n \to \mathbb{R}$ is said to be *concave* if $-f(\cdot)$ is convex. ■

Exercise 2.5.3 : The *epigraph* of a function $f:\mathbb{R}^n \to \mathbb{R}$ is defined by

$$Epi(f) \triangleq \{(x,y) \in \mathbb{R}^n \times \mathbb{R} \mid y \geq f(x)\} . \tag{2.5.13a}$$

Show that $f:\mathbb{R}^n \to \mathbb{R}$ is convex if and only if its epigraph is convex. ■

Theorem 2.5.5 : Suppose $f:\mathbb{R}^n \to \mathbb{R}$ is convex. Then $f(\cdot)$ is continuous. (For a proof, see Berge p. 193). ■

The following property can be deduced from Fig. 2.5.1.

Theorem 2.5.6 : Suppose $f:\mathbb{R}^n \to \mathbb{R}$ is differentiable. Then $f(\cdot)$ is convex if and only if

$$f(y) - f(x) \geq \langle \nabla f(x), y-x \rangle \quad \forall \quad x,y \in \mathbb{R}^n . \tag{2.5.14}$$

Proof : $\Rightarrow$ Suppose $f(\cdot)$ is convex. Then for any $x,y \in \mathbb{R}^n$, $\lambda \in [0,1]$,

$$f(x + \lambda(y-x)) \leq (1-\lambda)f(x) + \lambda f(y) . \tag{2.5.15}$$

Rearranging (2.5.15) we get

$$\frac{f(x+\lambda(y-x)) - f(x)}{\lambda} \leq f(y) - f(x) \quad \forall \quad \lambda \in [0,1] . \tag{2.5.16}$$

Taking the limit as $\lambda \to 0$ we get (2.5.14).

$\Leftarrow$ Suppose (2.5.14) holds. Then for any $\lambda \in [0,1]$, $x,y \in \mathbb{R}^n$

$$f(y) - f(x + \lambda(y-x)) \geq \langle \nabla f(x + \lambda(y-x)), y-x \rangle (1-\lambda), \tag{2.5.17a}$$

$$f(x) - f(x + \lambda(y-x)) \geq \langle \nabla f(x + \lambda(y-x)), y-x \rangle (-\lambda) . \tag{2.5.17b}$$

Multiplying (2.5.17a) by λ, (2.5.17b) by $(1-\lambda)$ and adding, we get (2.5.15), i.e., $f(\cdot)$ is convex. ■

Theorem 2.5.7 : Suppose that $f:\mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. Then $f(\cdot)$ is convex if and only if the Hessian (second derivative) matrix $\partial^2 f(x)/\partial x^2$ is positive semi-definite for all $x \in \mathbb{R}^n$, i.e., $\langle y, \partial^2 f(x)/\partial x^2 y \rangle \geq 0$ for all $x, y \in \mathbb{R}^n$.

Proof : $\Rightarrow$ Suppose $f(\cdot)$ is convex. Then for any $x,y \in \mathbb{R}^n$, because of Theorem 2.5.6 and Proposition 2.4.5

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle$$

$$= \int_0^1 (1 - s) \langle y - x, \frac{\partial^2 f(x + s(y - x))}{\partial x^2}(y - x) \rangle ds . \tag{2.5.18}$$

Hence, dividing by $\|y - x\|^2$ and letting $y \to x$, we obtain that $\partial^2 f(x)/\partial x^2$ is positive semi-definite.

$\Leftarrow$ Suppose that $\partial^2 f(x)/\partial x^2$ is positive semi-definite for all $x \in \mathbf{R}$. Then it follows directly from the equality in (2.5.18) and Theorem 2.5.6 that $f(\cdot)$ is convex. ■

Exercise 2.5.2 : Suppose that $f : \mathbf{R}^n \to \mathbf{R}$ is twice continuously differentiable and that for some $\infty > M \geq m > 0$, $M\|y\|^2 \geq \langle y, \partial^2 f/\partial x^2(x) y \rangle \geq m\|y\|^2$ for all $x, y \in \mathbf{R}^n$. Show that the level sets of $f(\cdot)$ are convex and compact and that $f(\cdot)$ attains its infinimum. ■

Exercise 2.5.3 : Suppose $f^i : \mathbf{R}^n \to \mathbf{R}$, $i = 1,2, \cdots, m$ are convex and that $\mathbf{m} \triangleq \{ 1,2, \cdots, m \}$. Show that

$$\psi^1(x) \triangleq \max_{i \in \mathbf{m}} f^i(x), \tag{2.5.19a}$$

$$\psi^2(x) \triangleq \sum_{i=1}^{m} f^i(x), \tag{2.5.19b}$$

are both convex. ■

3. UNCONSTRAINED OPTIMIZATION

In this Lecture, we shall be concerned with the geometry and characterization of solutions of optimization problems of the form

$$\min_{x \in \mathbf{R}^n} f(x) \tag{3.1.1}$$

with $f:\mathbf{R}^n \to \mathbf{R}$ at least once continuously differentiable.

3.1. GEOMETRY OF THE PROBLEM

Definition 3.1.1 : Given a function $f:\mathbf{R}^n \to \mathbf{R}$ and an $\alpha \in \mathbf{R}$, we shall say that the set $\mathbf{L}_\alpha \subset \mathbf{R}^n$, defined by

$$\mathbf{L}_\alpha \triangleq \{x \mid f(x) \leq \alpha\}, \tag{3.1.2}$$

is a *level set* (parametrized by α). ■

The level sets have the following properties:

(a) If $\alpha_1 > \alpha_2$, then $\mathbf{L}_{\alpha_2} \subset \mathbf{L}_{\alpha_1}$, i.e., the level sets are nested;

(b) If $\hat{\alpha} = \min_{x \in \mathbf{R}^n} f(x)$, then the *solution set* to (3.1.1) is the set $\mathbf{L}_{\hat{\alpha}}$;

(c) An algorithm for solving (3.1.1) is called a *descent method* if it constructs sequences $\{x_i\}_{i=0}^{\infty}$ such that $f(x_{i+1}) < f(x_i)$ for all $i \in \mathbf{N}$, i.e., if it constructs points which descend into ever lower level sets.

The boundary $\partial\mathbf{L}_\alpha$, of a level set $\mathbf{L}_\alpha$ (see Fig. 3.1.1) can be visualized as a constant altitude line on a topological map. Points on the boundary of $\mathbf{L}_\alpha$ satisfy the equation

$$f(x) = \alpha\,. \tag{3.1.3a}$$

Definition 3.1.2 : The *set of tangents* to the boundary $\partial\mathbf{L}_\alpha$, at a point $\hat{x} \in \partial\mathbf{L}_\alpha$ is defined by

$$\mathbf{T}(\hat{x}, \partial\mathbf{L}_\alpha) \triangleq \{y \in \mathbf{R}^n \mid \beta y = \lim_{x_i \to \hat{x}} (x_i - \hat{x})/|x_i - \hat{x}|,\ x_i \in \partial\mathbf{L}_\alpha,\ \beta > 0\}\,. \tag{3.1.3b}$$

The *set of normals* to the boundary $\partial\mathbf{L}_\alpha$, at a point $\hat{x} \in \partial\mathbf{L}_\alpha$ is defined by

$$\mathbf{N}(\hat{x}, \partial\mathbf{L}_\alpha) \triangleq \{v \in \mathbf{R}^n \mid \langle v, y\rangle = 0,\ \forall\, y \in \mathbf{T}(\hat{x}, \partial\mathbf{L}_\alpha)\}\,. \tag{3.1.3c}$$

We shall say that v is an *outward* normal to $\partial\mathbf{L}_\alpha$ at $\hat{x}$ if $f(\hat{x} + \lambda v) > f(\hat{x})$ for all sufficiently small $\lambda > 0$■

Proposition 3.1.1 : Suppose that $x \in \mathbf{R}^n$ is such that $f(x) = \alpha$. If the vector $\nabla f(x) \neq 0$, then $\nabla f(x)$ is an outward normal to the boundary of the level set $\mathbf{L}_\alpha$ at x.

Proof :

(i) Consider any sequence of points $\{ \Delta x_i \}_{i=0}^{\infty}$ such that $x + \Delta x_i \in \partial L_\alpha$ for all $i \in \mathbf{N}$ (i.e., $f(x + \Delta x_i) = \alpha$) and $\Delta x_i \to 0$ as $i \to \infty$. Hence, by the mean value theorem, we get that for some $s_i \in (0,1)$

$$f(x + \Delta x_i) = f(x) + \langle \nabla f(x + s_i \Delta x_i), \Delta x_i \rangle = \alpha . \tag{3.1.4}$$

But $f(x) = \alpha$ and hence

$$\frac{1}{\|\Delta x_i\|} \langle \nabla f(x + s_i \Delta x_i), \Delta x_i \rangle = \langle \nabla f(x + s_i \Delta x_i), \frac{\Delta x_i}{\|\Delta x_i\|} \rangle = 0 . \tag{3.1.5}$$

Since the unit ball is compact, without loss of generality, we can assume that $\Delta x_i / \|\Delta x_i\| \to y \in \mathbf{R}^n$ as $i \to \infty$. Hence, from (3.1.5), we get that $\langle \nabla f(x), y \rangle = 0$. Clearly, $y \in T(x, \partial L_\alpha)$, i.e., it is tangent to the boundary of the level set L_α. Since, by definition, all unit vectors which are tangent to the boundary of L_α at x are limit points of sequences $\{ \Delta x_i \}_{i=0}^{\infty}$, as above, it follows that $\nabla f(x) \in N(x, \partial L_\alpha)$, i.e., that it is normal to the boundary of L_α at x.

(ii) To show that $\nabla f(x)$ is an outward normal, let $\lambda > 0$ and consider

$$f(x + \lambda \nabla f(x)) = f(x) + \lambda \langle \nabla f(x + s\lambda \nabla f(x)), \nabla f(x) \rangle , \tag{3.1.6}$$

where $s \in (0,1)$, by the mean value theorem. Letting $\lambda \to 0$, we conclude, because of the continuity of $\nabla f(\cdot)$, that there exists a $\bar{\lambda} > 0$ such that for all $\lambda \in [\, 0, \bar{\lambda} \,]$

$$f(x + \lambda \nabla f(x)) \geq f(x) + \lambda \|\nabla f(x)\|^2 / 2, \tag{3.1.7}$$

which completes our proof. ■

Corollary 3.1.1 : If $x \in \mathbf{R}^n$ is such that $\nabla f(x) \neq 0$, then any vector $h \in \mathbf{R}^n$ such that $\langle \nabla f(x), h \rangle < 0$ is a *descent direction* for $f(\cdot)$ at x, i.e., there exists a $\hat{\lambda} > 0$ such that $f(x + \hat{\lambda} h) - f(x) < 0$.

Proof : Recall that the directional derivative at x, in the direction h is given by

$$df(x;h) = \langle \nabla f(x), h \rangle < 0 . \tag{3.1.8a}$$

Now, by definition of the directional derivative,

$$\lim_{\lambda \downarrow 0} [\frac{f(x + \lambda h) - f(x)}{\lambda} - df(\hat{x};h) \,] = 0 . \tag{3.1.8b}$$

Hence there exists a $\hat{\lambda} > 0$ such that

$$\left[\frac{f(x + \hat{\lambda} h) - f(x)}{\hat{\lambda}} - df(x;h) \right] \leq -\langle \nabla f(x), h \rangle / 2 , \tag{3.1.8c}$$

and therefore

$$f(x + \hat{\lambda} h) - f(x) \leq \hat{\lambda} \langle \nabla f(x), h \rangle / 2 < 0, \tag{3.1.8d}$$

which shows that h is a descent direction at $\hat{x}$. ■

3.2. FIRST AND SECOND ORDER OPTIMALITY CONDITIONS

We return to the problem (3.1.1).

Definition 3.2.1 : We shall say that $\hat{x}$ is a *global* minimizer for problem (3.1.1) if

$$f(\hat{x}) \leq f(x), \quad \forall \ x \in \mathbb{R}^n . \tag{3.2.1a}$$

We shall say that $\hat{x}$ is a *local* minimizer for problem (3.1.1) if there exists a $\hat{\rho} > 0$ such that

$$f(\hat{x}) \leq f(x) \quad \forall \ x \in B(\hat{x},\hat{\rho}) . \tag{3.2.1b}$$

Theorem 3.2.1 : Suppose that $f(\cdot)$ in (3.1.1) is continuously differentiable and that $\hat{x}$ is a local minimizer for (3.1.1), with associated radius $\hat{\rho} > 0$. Then $\nabla f(\hat{x}) = 0$.

Proof : To obtain a contradiction, suppose that $\nabla f(\hat{x}) \neq 0$. Then, letting $h = -\nabla f(\hat{x})$, we obtain

$$df(\hat{x};h) = \langle \nabla f(\hat{x}), -\nabla f(\hat{x}) \rangle < 0 . \tag{3.2.2a}$$

But, by definition of the directional derivative,

$$\lim_{\lambda \downarrow 0} \left[\frac{f(\hat{x} + \lambda h) - f(\hat{x})}{\lambda} - df(\hat{x};h) \right] = 0 . \tag{3.2.2b}$$

Hence there exists a $\hat{\lambda} \in (0,\hat{\rho}]$ such that

$$\left| \frac{f(\hat{x} + \hat{\lambda} h) - f(\hat{x})}{\hat{\lambda}} - df(\hat{x};h) \right| \leq -df(\hat{x};h)/2 , \tag{3.2.3a}$$

and therefore

$$f(\hat{x} + \hat{\lambda} h) - f(\hat{x}) \leq \hat{\lambda} df(\hat{x};h)/2 < 0 , \tag{3.2.3b}$$

which contradicts the optimality of $\hat{x}$. ■

Theorem 3.2.2 : Suppose that $f(\cdot)$ is twice continuously differentiable, that $\hat{x}$ is a global minimizer for problem (3.1.1), with associated radius $\hat{\rho} > 0$, and that $H(x) \triangleq \partial^2 f(x)/\partial x^2$. Then

$$\langle h, H(\hat{x}) h \rangle \geq 0 \quad \forall \ h \in \mathbb{R}^n . \tag{3.2.4}$$

Proof : Since by Theorem 3.2.1, $\nabla f(\hat{x}) = 0$, it follows from the optimality of $\hat{x}$ that for any $h \in \mathbb{R}^n$, $\lambda > 0$, such that $\lambda |h| \leq \hat{\rho}$,

$$f(\hat{x} + \lambda h) - f(\hat{x}) = \lambda^2 \int_0^1 (1 - s) \langle h, H(\hat{x} + s \lambda h) h \rangle ds$$

$$\geq 0 . \tag{3.2.5}$$

Letting $\lambda \to 0$, we obtain the desired result.

■

Theorem 3.2.3 : Suppose that $f(\cdot)$ is twice continuously differentiable and that $\hat{x} \in \mathbf{R}^n$ is such that $\nabla f(\hat{x}) = 0$, and $H(\hat{x}) > 0$, where $H(x) = \partial^2 f(x)/\partial x^2$, as before. Then $\hat{x}$ is a local minimizer for problem (3.1.1).

Proof : Suppose the theorem is false. Then there exists a sequence $\{x_i\}_{i=0}^{\infty}$ such that $x_i \to \hat{x}$ as $i \to \infty$ and $f(x_i) < f(\hat{x})$ for all $i \in \mathbf{N}$. Now, by (2.4.7d),

$$f(x_i) - f(\hat{x}) = \langle \nabla f(\hat{x}), x_i - \hat{x} \rangle + \int_0^1 (1-s)\langle (x_i - \hat{x}), H(\hat{x} + s(x_i - \hat{x}))(x_i - \hat{x}) \rangle ds \tag{3.2.6}$$

By assumption $H(\cdot)$ is continuous and $H(\hat{x}) > 0$. Hence, (a) there exists an $m > 0$ such that

$$\langle h, H(\hat{x})h \rangle \geq m|h|^2, \quad \forall \ h \in \mathbf{R}^n, \tag{3.2.7}$$

and (b) the function $\langle h, H(x)h \rangle$ of (h,x), is uniformly continuous on the compact set $B(0,1) \times B(\hat{x},1)$. Hence there exists an i_0 such that for all $i \geq i_0$ and $s \in (0,1)$,

$$|\langle (x_i - \hat{x})H(\hat{x} + s(x_i - \hat{x}))(x_i - \hat{x}) \rangle - \langle (x_i - \hat{x})H(\hat{x})(x_i - \hat{x}) \rangle| \leq \frac{m}{2}|x_i - \hat{x}|^2, \tag{3.2.8a}$$

which, because of (3.2.7), leads to the conclusion that

$$\langle (x_i - \hat{x}), H(\hat{x} + s(x_i - \hat{x}))(x_i - \hat{x}) \rangle \geq \frac{m}{2}|x_i - \hat{x}|^2 . \tag{3.2.8b}$$

Therefore, from (3.2.6)

$$f(x_i) - f(\hat{x}) \geq \frac{m}{4}|x_i - \hat{x}|^2 > 0, \tag{3.2.9}$$

which contradicts our assumption that $f(x_i) - f(\hat{x}) < 0$. Thus, the theorem must be true. ■

For convex functions we get some additional results.

Theorem 3.2.4 : Suppose that $f: \mathbf{R}^n \to \mathbf{R}$ is continuously differentiable and convex. If $\hat{x} \in \mathbf{R}^n$ is such that $\nabla f(\hat{x}) = 0$, then $\hat{x}$ is a global minimizer of $f(\cdot)$.

Proof : By Proposition 2.5.5, since $f(\cdot)$ is convex, for all $x \in \mathbf{R}^n$, we must have

$$f(x) - f(\hat{x}) \geq \langle \nabla f(\hat{x}), x - \hat{x} \rangle = 0 . \tag{3.2.10}$$

Since $\nabla f(\hat{x}) = 0$, we conclude that $\hat{x}$ is a global minimizer. ■

Theorem 3.2.5 : Suppose that $f(\cdot)$ is strictly convex, then it can have at most one global minimizer.

Proof : Suppose that x^*, x^{**} are two global minimizers of $f(\cdot)$. Then $f(x^*) = f(x^{**})$ must hold, and hence, since $f(\cdot)$ is strictly convex, for any $\lambda \in [0,1]$,

$$f(\lambda x^{**} + (1-\lambda)x^*)) < \lambda f(x^{**}) + (1-\lambda) f(x^*) = f(x^*), \tag{3.2.11}$$

which contradicts the fact that x^* is a global minimizer. ■

Corollary 3.2.1 : Suppose that $f:\mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and that there exists an $m > 0$ such that for all $x \in \mathbb{R}^n$, $h \in \mathbb{R}^n$, $\langle h, H(x)h\rangle \geq m|h|^2$. Then $f(\cdot)$ has a unique global minimizer.

Proof : We know from Proposition 2.5.7 that $f(\cdot)$ must be strictly convex. Hence, if it has a global minimizer, that minimizer is unique. Thus, we only need to prove that a global minimizer exists. We do this by showing that for any $x_0 \in \mathbb{R}^n$, the level set $\mathbf{L} \triangleq \{x \mid f(x) \leq f(x_0)\}$ is compact.

(a) Let $\{x_i\}_{i=0}^{\infty} \subset \mathbf{L}$ be any converging sequence with limit point $\hat{x}$, i.e., $x_i \to \hat{x}$ as $i \to \infty$. Then $f(x_i) \leq f(x_0)$ for all $i \in \mathbb{N}$. It follows by continuity of $f(\cdot)$ that $f(\hat{x}) \leq f(x_0)$ i. e., that $\mathbf{L}$ is closed.

(b) For any $x \in \mathbf{L}$, we have

$$0 \geq f(x) - f(x_0) = \langle \nabla f(x_0), x - x_0\rangle + \int_0^1 (1-s)\langle (x-x_0), H(x_0 + s(x-x_0))(x-x_0)\rangle ds$$

$$\geq \langle \nabla f(x_0), x - x_0\rangle + \frac{m}{2}|x - x_0|^2$$

$$\geq -|\nabla f(x_0)||x - x_0| + \frac{m}{2}|x - x_0|^2 . \tag{3.2.12}$$

Hence we must have that $|x - x_0| \leq \dfrac{2|\nabla f(x_0)|}{m}$, for all $x \in \mathbf{L}$.

(c) The existence of a global minimizer now follows from the fact that

$$\inf_{x \in \mathbb{R}^n} f(x) = \inf_{x \in \mathbf{L}} f(x) = \min_{x \in \mathbf{L}} f(x), \tag{3.2.13}$$

because $\mathbf{L}$ is compact (see Proposition 2.3.2). ■

3.3. GRADIENT METHODS

We shall now see how optimality conditions lead to computational methods. Consider again the problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{3.3.1}$$

with $f:\mathbb{R}^n \to \mathbb{R}$ continuously differentiable. We have seen in the proof of Theorem 3.2.1 that whenever $\nabla f(x) \neq 0$, the direction $h = -\nabla f(x)$, results in the directional derivative $df(x;h) = \langle \nabla f(x), h\rangle < 0$, i.e., $-\nabla f(x)$ is a descent direction for the cost function. The easiest way to transform this observation into an algorithm for solving (3.3.1) is as follows.

Steepest Descent Algorithm 3.3.1 :

Data : $x_0 \in \mathbb{R}^n$.

Step 0 : Set $i = 0$.

Step 1 : Compute the *search direction*

Armijo Gradient Algorithm 3.3.2 :

Parameters : $\alpha, \beta \in (0,1)$.

Data: $x_0 \in \mathbb{R}^n$.

Step 0 : set $i = 0$.

Step 1 : Compute *the search direction*

$$h_i = h(x_i) \triangleq -\nabla f(x_i) . \tag{3.3.4a}$$

Stop if $\nabla f(x_i) = 0$.

Step 2 : Compute the *step size*

$$\lambda_i = \beta^{k_i} \triangleq \arg\max_{k \in \mathbb{N}} \left\{ \beta^k \mid f(x_i + \beta^k h_i) - f(x_i) \leq -\beta^k \alpha \|\nabla f(x_i)\|^2 \right\} \tag{3.3.4b}$$

Step 3 : *Update*

$$x_{i+1} = x_i + \lambda_i h_i, \tag{3.3.4c}$$

replace i by $i + 1$ and go to step 1. ■

To develop a geometric understanding of the Armijo stepsize rule, we define the function $\phi:\mathbb{R} \to \mathbb{R}$ by

$$\phi(\lambda) = f(x_i - \lambda \nabla f(x_i)) - f(x_i) . \tag{3.3.5}$$

Then $\phi(0) = 0$ and $\phi'(\lambda)\,|_{\lambda=0} = -\|\nabla f(x_i)\|^2$, by the chain rule. Hence the graphical interpretation of the Armijo step size rule is as shown in Fig. 3.3.1.

The following theorem holds under the assumption that $f(\cdot)$ is only once continuously differentiable. However, to obtain a simpler proof, *we will assume that $f(\cdot)$ is twice continuously differentiable* and that its second derivative is bounded in a region of interest, i.e., we shall assume that there exists an $M < \infty$, such that

$$\|H(x)\| \triangleq \max\{ \|H(x)y\| \mid \|y\| = 1 \} \leq M \;\forall\; x \in \mathbb{R}^n , \tag{3.3.6}$$

where $H(x) \triangleq \partial^2 f(x)/\partial x^2$.

Theorem 3.3.2 : Suppose that (3.3.6) holds. Then,

(a) The Armijo step size rule is well defined.

(b) If $\{x_i\}_{i=0}^{\infty}$ is an infinite sequence constructed by Algorithm 3.3.2, then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies $\nabla f(\hat{x}) = 0$.

(c) If the set $\{x \in \mathbb{R}^n \mid \nabla f(x) = 0\}$ contains only a *finite number* of points, then any bounded sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 3.3.2 must converge to a point $\hat{x}$ such that $\nabla f(\hat{x}) = 0$.

(d) If $x^* \neq x^{**}$ are two accumulation points of a sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 3.3.2, then $f(x^*) = f(x^{**})$.

Proof :

(a) We must show that when $\nabla f(x_i) \neq 0$, $\lambda_i > 0$, i.e., that the solution k_i of (3.3.4b) is finite. First, from (3.3.4b) we see that the stepsize λ_i must satisfy the inequality

$$f(x_i - \lambda \nabla f(x_i)) - f(x_i) + \lambda\alpha|\nabla f(x_i)|^2 \leq 0, \tag{3.3.7}$$

for $\lambda = \lambda_i = \beta^{k_i}$. Expanding the left hand side of (3.3.7) to second order, we get

$$\begin{aligned} &f(x_i - \lambda \nabla f(x_i)) - f(x_i) + \lambda\alpha|\nabla f(x_i)|^2 \\ &\quad = -\lambda(1-\alpha)|\nabla f(x_i)|^2 + \lambda^2 \textstyle\int_0^1 (1-s)\, \langle \nabla f(x_i), H(x_i - s\lambda \nabla f(x_i))\nabla f(x_i)\rangle ds \\ &\quad \leq -\lambda(1-\alpha)|\nabla f(x_i)|^2 + \frac{\lambda^2 M}{2}|\nabla f(x_i)|^2 \\ &\quad = \lambda \frac{M}{2}|\nabla f(x_i)|^2 \{ \frac{-2(1-\alpha)}{M} + \lambda \}. \end{aligned} \tag{3.3.8}$$

Clearly, there exists a $\bar{k} \in \mathbf{N}$ such that

$$-\frac{2(1-\alpha)}{M} + \beta^{\bar{k}} \leq 0\,. \tag{3.3.9}$$

Consequently, $\lambda_i = \beta^{k_i} \geq \beta^{\bar{k}}$, and hence we see that the step size is well defined.

(b) To obtain a contradiction, suppose that $x_i \overset{K}{\to} \hat{x}$ and that $\nabla f(\hat{x}) \neq 0$. First we recall that by (a) above, λ_i satisfies $\lambda_i \geq \beta^{\bar{k}}$ for all $i \in \mathbf{N}$. Next, since $x_i \overset{K}{\to} \hat{x}$, and $\nabla f(\cdot)$ is continuous, there exists an $i_0 \in K$, such that for all $i \in K$, $i \geq i_0$, $|\nabla f(x_i)|^2 \geq |\nabla f(\hat{x})|^2/2$. Consequently, for all $i \in K$, $i \geq i_0$,

$$\begin{aligned} f(x_{i+1}) - f(x_i) &\leq -\lambda_i \alpha |\nabla f(x_i)|^2 \\ &\leq -\beta^{\bar{k}} \alpha\, |\nabla f(\hat{x})|^2/2 \triangleq -\hat{\delta} < 0\,. \end{aligned} \tag{3.3.10}$$

Because $f(\cdot)$ is continuous and $x_i \overset{K}{\to} \hat{x}$, we must have that $f(x_i) \overset{K}{\to} f(\hat{x})$ and hence, because $\{f(x_i)\}_{i=0}^{\infty}$ is monotonically decreasing (see Proposition 2.2.1), that $f(x_i) \to f(\hat{x})$, which contradicts (3.3.10). Therefore we must have that $\nabla f(\hat{x}) = 0$.

(c) Suppose, without loss of generality, that the solution set

$$\Delta \triangleq \{x \in \mathbf{R}^n \mid \nabla f(x) = 0\} \tag{3.3.11}$$

contains only two points, x^*, x^{**}. By assumption, the sequence $\{x_i\}_{i=0}^{\infty}$, constructed by Algorithm 3.3.2 is bounded and hence it must have accumulation points. By part (b) of this theorem, all the accumulation points of $\{x_i\}_{i=0}^{\infty}$ must be in the set Δ. Hence we must have that $\min\{|x_i - x^*|, |x_i - x^{**}|\} \to 0$ as $i \to \infty$ (because otherwise it would be possible to show that there exists a third distinct accumulation point x^{***}). Let $\rho > 0$ be such that $\rho < |x^* - x^{**}|/4$. Then there exists an i_0 such that for all $i \geq i_0$, $\min\{|x_i - x^*|, |x_i - x^{**}|\} \leq \rho$, i.e., $x_i \in B(x^*, \rho)$ or $x_i \in B(x^{**}, \rho)$, for all $i \geq i_0$, (see Fig. 3.3.2).

Next, since $\nabla f(\cdot)$ is continuous, $\nabla f(x_i) \to 0$, and $\lambda_i \leq 1$ for all $i \in \mathbb{N}$, by construction. Therefore $|x_{i+1} - x_i| = \lambda_i |\nabla f(x_i)| \to 0$ as $i \to \infty$. Hence there exists an $i_1 \geq i_0$ such that for all $i \geq i_1$, $|x_i - x_{i+1}| \leq \rho$. Consequently, for all $i \geq i_1$, if $x_i \in B(x^*,\rho)$ $(B(x^{**},\rho))$, then $x_{i+1} \in B(x^*,\rho)$ $(B(x^{**},\rho))$ and thus the entire sequence converges to x^* (x^{**}) .

(d) Since the sequence $\{f(x_i)\}_{i=0}^{\infty}$ is monotone decreasing, it follows that if $x_i \xrightarrow{K} x^*$, then $f(x_i) \to f(x^*)$ as $i \to \infty$ and hence if x^{**} is also an accumulation point of $\{x_i\}_{i=0}^{\infty}$, then we must have that $f(x^{**}) = f(x^*)$. ■

Comment 3.3.1 : The computation of $\lambda_i = \beta^{k_i}$ need not be performed by trying $k = 0,1,2,....$, until (3.3.6) is satisfied. A much more efficient method is to start with $\lambda = \beta^{k_{i-1}}$ and, if (3.3.6) is satisfied, try $k_{i-1} - 1$, $k_{i-1} - 2$, . . . until it fails, and then back up one step. If (3.3.6) is not satisfied with $\lambda = \beta^{k_{i-1}}$, then try $k_{i-1} + 1$, $k_{i-1} + 2$, . . . until (3.3.6) is satisfied. Also, one may set $\lambda_i = \beta^{k_i} B$, with $B > 0$ selected on the basis of experience. ■

Comment 3.3.2 : The fact that an algorithm constructs descent directions along which it then minimizes is not sufficient to guarantee that it has useful convergence properties. For example, consider the recursion

$$x_{i+1} = x_i + \lambda_i h_i, \quad i = 1,2,.... \tag{3.3.12a}$$

where

$$\lambda_i \in \lambda(x_i) \triangleq \operatorname*{argmin}_{\lambda \geq 0} f(x_i + \lambda h_i) , \tag{3.3.12b}$$

and h_i is such that

$$\langle \nabla f(x_i), h_i \rangle = -\frac{2}{3(2i+1)} |\nabla f(x_i)| |h_i| . \tag{3.3.12c}$$

Since $\cos\theta = \prod_{i=1}^{\infty} \left[1 - \frac{4\theta^2}{(2i+1)^2\pi^2} \right]$, so that $\cos\pi/3 = \frac{1}{2} = \prod_{i=1}^{\infty} \left[1 - \frac{4}{9(2i+1)^2} \right]$, it is easy to see that when applied to the minimization of $|x|^2$, from an x_0 such that $|x_0| = 1$, the recursion (3.3.12a) constructs a sequence which converges to a point $\hat{x}$ such that $|\hat{x}| = \sqrt{1/2}$. ■

Exercise 3.3.1 : Consider the function $f(x) \triangleq e^{-|x|^2}$, with $x \in \mathbb{R}^n$. Show that for this function the Armijo Gradient Method constructs a sequence $\{x_i\}_{i=0}^{\infty}$ such that $|x_i| \to \infty$ and $f(x_i) \to 0$ as $i \to \infty$. ■

Exercise 3.3.2 : Consider the function $f(x) = x^2 e^{-x} - x$, with $x \in \mathbb{R}$. Determine the behaviour of the Armijo Gradient Method on this function when $x_0 = 0.5$ and when $x_0 = 2$. ■

Exercise 3.3.3 : Show that whenever the Armijo method constructs a bounded sequence $\{x_i\}_{i=0}^{\infty}$, we must have $\nabla f(x_i) \to 0$, as $i \to \infty$. ■

Exercise 3.3.4 : The geometry of the behavior of the Armijo method is best seen in terms of level sets and trajectories. Show that the speed of the Armijo method is affected by the "narrowness" of the level sets: the closer the level sets are to the spheres, the better the behavior of the Armijo method. ■

Exercise 3.3.5 : Suppose that $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by the Armijo Gradient Algorithm 3.3.2 in solving problem (3.3.1), with $f:\mathbf{R}^n \to \mathbf{R}$ twice continuously differentiable, and that $\{x_i\}_{i=0}^{\infty}$ has an accumulation point $\hat{x}$ such that $\partial^2 f(\hat{x})/\partial x^2$ is positive definite. Show that the sequence $\{x_i\}_{i=0}^{\infty}$ converges to $\hat{x}$. ■

The proofs of convergence both for the method of steepest descent and for the Armijo method followed the same pattern which we will also use in the proof of convergence of a number of other algorithms. This pattern is best studied in an abstract setting, as follows. First we observe that the λ_i computed by the steepest descent algorithm in (3.3.2) need not be unique and hence the successor point of x_i is not unique. We therefore see that the algorithm is defined by an *iteration map* which is *set valued*, i.e., which is a *multifunction*, so that a relationship of the form $x_{i+1} \in a(x_i)$ holds, rather than $x_{i+1} = a(x_i)$. Multifunctions play an important role in optimization. An example we have seen earlier is that of the level set of a continuous function $f:\mathbf{R}^n \to \mathbf{R}$, i.e., $L(x) \triangleq \{x' \in \mathbf{R}^n \mid f(x') \leq f(x)\}$. In the literature we find concepts of upper and lower continuity, continuity and differentiablity of multifunctions.

Abstract Problem 3.3.1 : Let $\mathbf{D} \subset \mathbf{R}^n$ be the set of *desirable points* in $\mathbf{R}^n$. Find a point $\hat{x}$ in D. ■

Let $c:\mathbf{R}^n \to \mathbf{R}$ be a *surrogate cost* function and let $a:2^n \to \mathbf{R}^{\mathbf{R}^n}$ be a set valued iteration map. We propose the following algorithm for solving the Abstract Problem 3.3.1:

Model Algorithm 3.3.1 :

Data: $x_0 \in \mathbf{R}^n$.

Step 0 : set $i = 0$.

Step 1 : Compute a *candidate successor*

$$y \in a(x_i) . \tag{3.3.13a}$$

Step 2 : If

$$c(y) < c(x_i), \tag{3.3.13b}$$

set

$$x_{i+1} = y , \tag{3.3.13c}$$

replace i by $i + 1$ and go to step 1.

Else stop. ■

Theorem 3.3.3 : Suppose that $c(\cdot)$ is continuous and that for every $x \in \mathbf{D}$ there exist an $\rho(x) > 0$ and a $\delta(x) > 0$ such that

$$c(x'') - c(x') \leq -\delta(x) < 0, \ \forall\, x' \in B(x,\rho(x)), \forall\, x'' \in a(x') . \tag{3.3.14}$$

Then either the sequence $\{x_i\}$, constructed by the Model Algorithm is finite and its last element is in D

or it is infinite and every accumulation point of $\{x_i\}$ is in **D**.

Proof : First, we conclude from (3.3.14) that for all $x \in \mathbf{D}$, we must have $c(a(x)) \geq c(x)$. Hence if the sequence $\{x_i\}$ is finite, its last element must be in **D**. Next, suppose that $\hat{x}$ is an accumulation point of $\{x_i\}_{i=0}^{\infty}$, i.e., $x_i \xrightarrow{K} \hat{x}$ as $i \to \infty$. Then, because the sequence $\{c(x_i)\}_{i=0}^{\infty}$ is monotone decreasing and $c(\cdot)$ is continuous, we must have that $c(x_i) \to c(\hat{x})$ as $i \to \infty$. Now suppose that $\hat{x} \in \mathbf{D}$. Then there exist $\hat{\delta} > 0$, $\hat{\rho} > 0$ such that

$$c(x') - c(x) \leq -\hat{\delta} < 0, \quad \forall\, x \in B(\hat{x},\hat{\rho}), \forall\, x' \in a(x). \tag{3.3.15a}$$

Therefore there must exist an i_o such that for all $i \geq i_o$, $i \in K$, $x_i \in B(\hat{x},\hat{\rho})$ and hence

$$c(x_{i+1}) - c(x_i) \leq -\hat{\delta} < 0, \tag{3.3.15b}$$

which contradicts the convergence of $c(x_i)$ to $c(\hat{x})$. This completes our proof. ■

4. RATE OF CONVERGENCE AND EFFICIENCY

The most reliable way of evaluating the relative merits of two algorithms is to apply both of them to a set of problems of interest and to compare the cpu times needed to solve these problems. Such an exhaustive comparison is not always possible. Hence it is useful to have some mathematical measures of algorithm performance, which can be used to make qualitative distinctions between algorithms. We shall now discuss two of these performance measures.

4.1. RATE OF CONVERGENCE OF SEQUENCES

Definition 4.1.1 : We say that a sequence $\{x_i\}_{i=0}^{\infty}$, in $\mathbb{R}^n$, converges to a point $\hat{x}$ at least with *root rate* $r \geq 1$ if there exist $M \in (0,\infty)$, $\delta \in (0,1)$ and $i_0 \in \mathbb{N}$ such that for all $i \geq i_0$,

$$|x_i - \hat{x}| \leq M\delta^i, \quad \text{if } r = 1, \tag{4.1.1a}$$

$$|x_i - \hat{x}| \leq M\delta^{r^i}, \quad \text{if } r > 1 . \tag{4.1.1b}$$

When $r = 1$, we say that the convergence is *linear*. When $r > 1$, we say that the convergence is *superlinear*. ■

When plotted on a semilog scale, a linearly converging sequence produces a graph as shown in Fig. 4.1.1a, while a superlinearly converging sequence produces a graph as shown in Fig. 4.1.1b.

Remark 4.1.1 : It is possible for a sequence $\{x_i\}_{i=0}^{\infty}$ to converge to a point $\hat{x}$ *slower* then linearly, e.g., when $|x_i - \hat{x}| = k/i$. ■

The following result is basic to the study of rate of convergence of algorithms.

Theorem 4.1.1 : Let $\{x_i\}_{i=0}^{\infty}$ be a sequence in $\mathbb{R}^n$.

(a) If there exists a $\delta \in (0,1)$ and an $i_0 \in \mathbb{N}$ such that

$$|x_{i+1} - x_i| \leq \delta |x_i - x_{i-1}|, \quad \forall \ i \geq i_0+1, \tag{4.1.2a}$$

then there exists an $\hat{x} \in \mathbb{R}^n$ such that $x_i \to \hat{x}$ as $i \to \infty$, at least linearly.

(b) If there exists an $M \in (0,\infty)$, an $r > 1$, and an $i_0 \in \mathbb{N}$ such that

$$M^{\left[\frac{1}{r-1}\right]} |x_{i_0+1} - x_{i_0}| \leq 1, \tag{4.1.2b}$$

and

$$|x_{i+1} - x_i| \leq M|x_i - x_{i-1}|^r, \quad \forall \ i \geq i_0+1, \tag{4.1.2c}$$

then there exists an $\hat{x} \in \mathbb{R}^n$ such that $x_i \to \hat{x}$ as $i \to \infty$ superlinearly, with root rate at least r.

Proof :

(a) For $i = 0, 1, 2...$, let $e_i \triangleq |x_{i+1} - x_i|$. Then (4.1.2a) becomes

$$e_{i+1} \leq \delta e_i \quad \forall \quad i \geq i_0 . \tag{4.1.3}$$

Hence, by induction,

$$\begin{aligned} &e_{i_0+1} \leq \delta e_{i_0}, \\ &e_{i_0+2} \leq \delta e_{i_0+1} \leq \delta^2 e_{i_0}, \\ &\cdot \\ &\cdot \\ &\cdot \\ &e_{i_0+j} \leq \delta^j e_{i_0}, \end{aligned} \tag{4.1.4}$$

and, since $\delta \in (0,1)$, it follows that $e_i \to 0$ as $i \to \infty$. Hence, for any $j > k \geq i_0$, we get that

$$\begin{aligned} |x_j - x_k| &= |(x_j - x_{j-1}) + (x_{j-1} - x_{j-2}) + ... + (x_{k+1} - x_k)| \\ &\leq \sum_{i=k}^{j-1} e_i \leq \sum_{i=k}^{\infty} e_i \\ &= \sum_{j=k-i_0}^{\infty} e_{i_0}\delta^j = \delta^{k-i_0}\left(\frac{e_{i_0}}{1-\delta}\right), \end{aligned} \tag{4.1.5}$$

i.e., $|x_j - x_k| \to 0$ for $j > k$ as $k \to \infty$, uniformly in j. This shows that $\{x_i\}_{i=0}^{\infty}$ is *Cauchy* and hence that it must converge to a point $\hat{x}$.

Next, letting $j \to \infty$, we can replace x_j by $\hat{x}$ in (4.1.5) to obtain that

$$|\hat{x} - x_k| \leq \left[\frac{e_{i_0}}{1-\delta}\,\delta^{-i_0}\right]\delta^k, \tag{4.1.6}$$

which proves that $x_k \to \hat{x}$ as $k \to \infty$ linearly.

(b) Next, again with $e_i = |x_{i+1} - x_i|$, we get from (4.1.2c) that

$$e_{i+1} \leq M e_i^r, \quad \forall \quad i \geq i_0 . \tag{4.1.7a}$$

Hence, multiplying both sides of (4.1.7a) by $M^{\left[\frac{1}{r-1}\right]}$ we get

$$M^{\left[\frac{1}{r-1}\right]} e_{i+1} \leq M^{\left[\frac{1}{r-1}\right]} M e_i^r = \left(M^{\left[\frac{1}{r-1}\right]} e_i\right)^r, \quad \forall \quad i \geq i_0 . \tag{4.1.7b}$$

For $i = 0,1,2...$, let $\mu_i \triangleq M^{\left[\frac{1}{r-1}\right]} e_i$. Then, from (4.1.7b) we get that

$$\mu_{i+1} \leq \mu_i^r, \quad \forall \quad i \geq i_0 . \tag{4.1.8a}$$

Letting $w_i = \ln \mu_i$, we get from (4.1.8a) that

$$w_{i+1} \leq r w_i, \quad \forall \ i \geq i_0 . \tag{4.1.8b}$$

Hence

$$w_i \leq r^{i-i_0} w_{i_0}, \quad \forall \ i \geq i_0, \tag{4.1.9a}$$

which leads to the conclusion that

$$\mu_i \leq \mu_{i_0}^{r^{i-i_0}}, \quad \forall \ i \geq i_0 . \tag{4.1.9b}$$

Now (4.1.9b) can be rewritten as

$$\mu_i \leq \left[\mu_{i_0}^{(1/r)^{i_0}}\right]^{r^i}, \quad \forall \ i \geq i_0 . \tag{4.1.9c}$$

Substituting for the μ_i in (4.1.9c), we get that

$$e_i \ \leq \ (1/M)^{\left[\frac{1}{r-1}\right]} \left[\left[M^{\frac{1}{r-1}} e_0\right]^{(1/r)^{i_0}}\right]^{r^i}$$

$$\triangleq \ c\,\delta^{r^i}, \quad \forall \ i \geq i_0, \tag{4.1.10}$$

with $\delta = \mu_{i_0}^{(1/r)^{i_0}}$. Next we note that because of (4.1.2b), $\delta \in (0,1)$. Therefore, since $r > 1$, there exists an $i_r \geq i_0$ such that $\delta^{(r^j - r^k)} \leq \delta^{(j-k)}$ for all $i \geq i_r$, and hence, by arguments analogous to the ones used in obtaining (4.1.5), we get that for some $c' < \infty$ and all $j > k \geq i_0$,

$$|x_j - x_k| \leq \sum_{i=k}^{\infty} e_i \leq c \sum_{i=k}^{\infty} \delta^{r^i} \leq c\delta^{r^k} \left[\sum_{i=k}^{k+i_r-1} \delta^{(r^i - r^k)} + \sum_{i=i_r}^{\infty} \delta^{(i-i_r)}\right] \leq c'\delta^{r^k}, \tag{4.1.11a}$$

which proves that $\{x_k\}$ is Cauchy, so that it must have a limit point $\hat{x}$. Letting $j \to \infty$, it now follows from (4.1.11a) that

$$|\hat{x} - x_k| \leq c'\delta^{r^k}, \quad \forall \ k \geq i_0, \tag{4.1.11b}$$

which shows that $x_k \to \hat{x}$, as $i \to \infty$, with rate at least r. ■

4.2. EFFICIENCY

Next we turn to the task of estimating the work needed to solve an optimization problem to prescribed precision.

Thus suppose that in solving a particular problem, an optimization algorithm produces a sequence $\{x_i\}_{i=0}^{\infty}$ which converges to a solution $\hat{x}$ linearly. To reduce the error from an initial value of $|x_0 - \hat{x}|$ to $\alpha|x_0 - \hat{x}|$, for a given $\alpha \in (0,1)$, requires a number of iterations. This number can be estimated if we know constants $\delta \in (0,1)$ and $M \in (0,\infty)$ such that $|x_i - \hat{x}| \leq M\delta^i$, for all $i \in \mathbf{N}$ (i.e., we are setting $i_0 = 0$). Clearly, in this case we must have that $|x_0 - \hat{x}| \leq M$ and a bound on the number of iterations

needed to reduce the initial error by the factor α is given by the smallest solution, $i^* \in \mathbb{N}$, of the inequality

$$\delta^i \leq \alpha . \tag{4.2.2a}$$

Taking logarithms of both sides, (4.2.2a) yields (since both δ , $\alpha \in (0,1)$) that i^* is the smallest integer such that

$$i^* \geq \frac{\ln\alpha}{\ln\delta} . \tag{4.2.2b}$$

Assuming that it takes w units of work (say cpu seconds) to construct x_i, and that $\ln\alpha/\ln\delta$ is much greater than 1, so that $i^* \equiv \ln\alpha/\ln\delta$, an estimate for the total work performed in reducing the error by the factor α is given by

$$W \equiv \frac{w}{\ln\delta}\ln\alpha . \tag{4.2.3}$$

The factor

$$\eta \triangleq -\frac{\ln\delta}{w} > 0 \tag{4.2.4}$$

is called the *efficiency* of the process that constructed the sequence $\{x_i\}_{i=0}^{\infty}$.

When $x_i \to \hat{x}$ superlinearly (with root rate $r > 1$), the *efficiency* of the process which constructed the sequence $\{x_i\}_{i=0}^{\infty}$ is sometimes defined by

$$\eta \triangleq \frac{\ln r}{w} . \tag{4.2.5}$$

The expression (4.2.5) expresses the work required to reduce the error $|x_i - \hat{x}|$ by a specificed factor for processes with the same M, δ. It does not take into account the fact that both M and δ can depend on $|x_0 - \hat{x}|$, as is the case in (4.1.10). Hence it must be used with caution.

Exercise 4.2.1 : (a) Obtain a formula which estimates the work needed by an algorithm satisfying a relation of the form (4.1.1b), with $r > 1$, to reduce an initial error $|x_0 - \hat{x}|$ by a factor $\alpha \in (0,1)$.

(b) Justify the definition (4.2.5). ■

4.3. RATE OF CONVERGENCE OF ARMIJO GRADIENT METHOD

To conclude this section, we shall establish the rate of convergence of the Armijo Gradient Algorithm 3.3.2.

Theorem 4.3.1 : Suppose that $f:\mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and that there exist $0 < m \leq M < \infty$ such that

$$m|v|^2 \leq \langle v, \frac{\partial^2 f(x)}{\partial x^2} v\rangle \leq M|v|^2 \tag{4.3.1}$$

holds for all $x, v \in \mathbb{R}^n$. If $\{x^i\}_{i=0}^{\infty}$ is a sequence constructed by the Armijo Gradient Algorithm 3.3.2, in

solving $\min_{x \in \mathbb{R}^n} f(x)$, then

(a) $x_i \to \hat{x}$ as $i \to \infty$, with $\hat{x}$ the unique minimizer of $f(\cdot)$, and

(b) $x_i \to \hat{x}$ as $i \to \infty$ linearly, with convergence constant $\delta \leq 1 - 4m\beta\alpha(1-\alpha)/M$.

Proof :

(a) From Exercise 2.4.3, the level sets of $f(\cdot)$ are compact. Hence a solution $\hat{x}$ to $\min_{x \in \mathbb{R}^n} f(x)$ must exist. Because $f(\cdot)$ is strictly convex, $\hat{x}$ is a unique solution. Since every accumulation point x^* of $\{x_i\}_{i=0}^{\infty}$, constructed by the Armijo Gradient Algorithm 3.3.2, must satisfy $\nabla f(x^*) = 0$, it follows that $x^* = \hat{x}$ is the only accumulation point of this sequence.

(b) To establish the rate of convergence we need three results:

(i) For any x_i, by the second order expansion formula (2.4.7d),

$$f(x_i) - f(\hat{x}) = \int_0^1 (1-s)\langle (x_i - \hat{x}), H(\hat{x} + s(x_i - \hat{x}))(x_i - \hat{x})\rangle ds \,. \tag{4.3.2a}$$

Because of (4.3.1), (4.3.2a) leads to

$$m|x_i - \hat{x}|^2 \leq 2[f(x_i) - f(\hat{x})] \leq M|x_i - \hat{x}|^2 \,. \tag{4.3.2b}$$

(ii) Making use of the first order expansion formula (2.4.7a) and the fact that $\nabla f(\hat{x}) = 0$, we obtain

$$\nabla f(x_i) = \int_0^1 H(\hat{x} + s(x_i - \hat{x}))(x_i - \hat{x}) ds \,. \tag{4.3.3a}$$

It now follows from (4.3.1) and the Schwartz inequality that

$$m|x_i - \hat{x}|^2 \leq \langle \nabla f(x_i), x_i - \hat{x}\rangle \leq |\nabla f(x_i)||x_i - \hat{x}| \,. \tag{4.3.3b}$$

(iii) Expanding to second order the formula used in the Armijo step size calculation (3.3.4b), we obtain

$$\begin{aligned} & f(x_i - \lambda \nabla f(x_i)) - f(x_i) + \lambda\alpha|\nabla f(x_i)|^2 \\ & \quad = -\lambda \Big[(1-\alpha)|\nabla f(x_i)|^2 \\ & \qquad - \lambda \int_0^1 (1-s)\langle \nabla f(x_i), H(x_i - s\lambda \nabla f(x_i))\, \nabla f(x_i)\rangle ds\Big] \\ & \quad \leq -\lambda|\nabla f(x_i)|^2[(1-\alpha) - \lambda M/2] \,. \end{aligned} \tag{4.3.4a}$$

The right hand side of (4.3.4a) is negative for all $\lambda \in [0, 2(1-\alpha)/M]$. Hence we must have that $\beta^{k_i} \geq \frac{2\beta}{M}(1-\alpha)$. To see this, observe that $k_i \leq \hat{k}$, with $\hat{k}$ such that $\beta^{\hat{k}} \leq \frac{2(1-\alpha)}{M}$ and $\beta^{\hat{k}-1} \geq 2(1-\alpha)/M$. Consequently we get that

$$h_i = h(x_i) \triangleq -\nabla f(x_i) . \tag{3.3.2a}$$

Stop if $\nabla f(x_i) = 0$.

Step 2 : Compute the *step size*

$$\lambda_i \in \lambda(x_i) \triangleq \operatorname{argmin}_{\lambda \geq 0} f(x_i + \lambda h_i) . \tag{3.3.2b}$$

Step 3 : *Update*

$$x_{i+1} = x_i + \lambda_i h_i . \tag{3.3.2c}$$

Replace i by $i + 1$ and go to Step 1. ■

All of our convergence theorems will be stated in terms of subsequences constructed by an algorithm. Hence we shall be using the following notation.

Notation 3.3.1 : Given a sequence $\{x_i\}_{i=0}^{\infty}$ and an infinite subset $K \subset \mathbb{N}$ we shall denote by $x_i \overset{K}{\to} \hat{x}$ (as $i \to \infty$) the fact that the subsequence $\{x_i\}_{i \in K}$ converges to $\hat{x}$. ■

Without making additional assumptions on the function $f(\cdot)$, it is not possible to be sure that a sequence $\{x_i\}_{i=0}^{\infty}$, constructed by the steepest descent Algorithm 3.3.1, is bounded or that it converges. Hence, we must content ourselves with the following, quite typical, milder convergence result.

Theorem 3.3.1 : If $\{x_i\}_{i=0}^{\infty}$ is an infinite sequence constructed by Algorithm 3.3.1, then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies $\nabla f(\hat{x}) = 0$.

Proof : Suppose that $x_i \overset{K}{\to} \hat{x}$ as $i \to \infty$ and that $\nabla f(\hat{x}) \neq 0$. Then

$$df(\hat{x}; h(\hat{x})) = -\|\nabla f(\hat{x})\|^2 < 0 . \tag{3.3.3a}$$

Hence any $\hat{\lambda} \in \lambda(\hat{x})$ satisfies $\hat{\lambda} > 0$ and there exists a $\hat{\delta} > 0$ such that

$$f(\hat{x} + \hat{\lambda} h(\hat{x})) - f(\hat{x}) = -\hat{\delta} < 0 . \tag{3.3.3b}$$

Since $h(\cdot) = -\nabla f(\cdot)$ is continuous by assumption, the function $f(x + \hat{\lambda} h(x)) - f(x)$ is continuous in x and hence there exists an i_0 such that for all $i \in K$, $i \geq i_0$,

$$f(x_{i+1}) - f(x_i) \leq f(x_i + \hat{\lambda} h(x_i)) - f(x_i) \leq -\frac{\hat{\delta}}{2} . \tag{3.3.3c}$$

Now, by construction, $\{f(x_i)\}_{i=0}^{\infty}$ is monotone decreasing and $f(x_i) \overset{K}{\to} f(\hat{x})$ as $i \to \infty$ by continuity of $f(\cdot)$. Therefore, by Proposition 2.2.1, we must have that $f(x_i) \to f(\hat{x})$ as $i \to \infty$. But this contradicts (3.3.3c). Hence $\nabla f(\hat{x}) = 0$ must hold. ■

The main objection to the method of steepest descent is that it contains a non-implementable step size rule. To get around this difficulty, several alternatives have been proposed. We find the following one particularly efficient.

$$f(x_{i+1}) - f(x_i) \leq -\frac{2\beta\alpha(1-\alpha)}{M}|\nabla f(x_i)|^2, \tag{4.3.4b}$$

for all $i \in \mathbf{N}$. Now,

$$f(\hat{x}) - f(x_i) = \langle \nabla f(x_i), \hat{x} - x_i \rangle + \int_0^1 (1-s)\langle (\hat{x} - x_i, H(x_i + s(\hat{x} - x_i))(\hat{x} - x_i)\rangle ds$$

$$\geq \langle \nabla f(x_i), \hat{x} - x_i \rangle + \frac{m}{2}|\hat{x} - x_i|^2$$

$$\geq \min_{h \in \mathbf{R}^n} \langle \nabla f(x_i), h \rangle + \frac{m}{2}|h|^2$$

$$= -\frac{1}{2m}|\nabla f(x_i)|^2 . \tag{4.3.5}$$

Substituting for $|\nabla f(x_i)|^2$ in (4.3.4b) from (4.3.5), we get that

$$f(x_{i+1}) - f(x_i) \leq \beta\alpha(1-\alpha)\frac{m}{M}[f(\hat{x}) - f(x_i)] . \tag{4.3.6a}$$

Subtracting $f(\hat{x})$ from both sides of (4.3.6a) and rearranging terms, we get that for all $i \in \mathbf{N}$,

$$f(x_{i+1}) - f(\hat{x}) \leq [1 - \frac{4m\beta\alpha(1-\alpha)}{M}][f(x_i) - f(\hat{x})] . \tag{4.3.6b}$$

Since $[1 - 4m\beta\alpha(1-\alpha)/M] \in (0,1)$, we find from Theorem 4.1 that for all $i \geq 0$,

$$0 < f(x_i) - f(\hat{x}) \leq \delta^i[f(x_0) - f(\hat{x})], \tag{4.3.7a}$$

with $\delta = 1 - 4m\beta\alpha(1-\alpha)/M$. Hence, making use of (4.3.2b) we obtain that

$$|x_i - \hat{x}| \leq \left[\frac{2}{m}[f(x_0) - f(\hat{x})]\right]^{1/2} (\delta^{1/2})^i, \quad \forall \ i \geq 0, \tag{4.3.7b}$$

which completes our proof. ■

Remark 4.3.1 : It is possible to obtain a much less conservative result than (4.3.7a, b) for the Armijo Gradient Algorithm 3.3.2, with $1 - \frac{m}{M}$ in (4.3.6) replaced by the tighter bound $(\frac{M-m}{M+m})^2$ (c.f. Luenberger, Introduction to Linear and Nonlinear Programming, pp. 148-154). ■

Exercise 4.3.1 :

(a) Prove that setting $\alpha = 1/2$, in the Armijo step size rule, is a good idea.

(b) What is the trade off in making β small or large in the Armijo step size rule? ■

Exercise 4.3.2 : Consider the Steepest Descent Algorithm 3.3.1 and suppose that it is applied to the solution of the problem $\min_{x \in \mathbf{R}^n} f(x)$, with $f(\cdot)$ satisfying the assumptions of Theorem 4.3.1.

(a) Prove that if it constructs a sequence $\{x_i\}_{i=0}^{\infty}$, then $x_i \to \hat{x}$, as $i \to \infty$ the unique minimization of $f(\cdot)$, linearly.

(b) Compare the speed of convergence of Algorithm 3.3.1 with that of Algorithm 3.3.2 and estimate their relative efficiences. ■

Exercise 4.3.3 : Suppose that $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by the Armijo Gradient Algorithm 3.3.2 in solving problem (3.3.1), with $f:\mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable, and that $\{x_i\}_{i=0}^{\infty}$ has an accumulation point $\hat{x}$ such that $\partial^2 f(\hat{x})/\partial x^2$ is positive definite. Show that the sequence $\{x_i\}_{i=0}^{\infty}$ converges linearly to $\hat{x}$. ■

5. NEWTON'S METHOD

Newton's method is one of the very oldest and best methods for solving many root finding and optimization problems. However, in its simplest form it converges only if the initial guess is sufficiently close to a solution, as will soon become apparent. We will examine both the simplest *(local)* version as well as *stabilized* versions which have global convergence properties.

5.1 THE LOCAL NEWTON METHOD

We return to the problem

$$\min_{x \in \mathbb{R}^n} f(x) . \tag{5.1.1}$$

Assumption 5.1.1 :

(a) The function $f:\mathbb{R}^n \to \mathbb{R}$ is twice continuously locally Lipschitz differentiable, i.e., given any bounded set $S \subset \mathbb{R}^n$ there exists an $L < \infty$ such that

$$|H(x') - H(x)| \leq L|x - x'| , \tag{5.1.2a}$$

for all $x, x' \in S$, with $H(x) \triangleq \partial^2 f(x)/\partial x^2$, as before.

The norm in (5.1.2a) will be assumed to be the induced L_2 norm which is defined by

$$|H| \triangleq \max \{ |Hx| \mid |x| = 1 \} . \tag{5.1.2b}$$

(The use of the L_2 norm is convenient, but not essential. Other induced matrix norms can also be used.)

(b) We will assume that (5.1.1) has a local minimizer $\hat{x}$ satisfying the second order sufficiency condition (3.2.7), and hence that there exist constants $0 < m \leq M < \infty$ such that

$$m|y|^2 \leq \langle y, H(\hat{x})y \rangle \leq M|y|^2, \quad \forall \ y \in \mathbb{R}^n . \tag{5.1.3}$$

■

Exercise 5.1.1 : Suppose that H is an $n \times n$ real symmetric matrix and that there exist $0 < m \leq M < \infty$ such that for all $y \in \mathbb{R}^n$,

$$m|y|^2 \leq \langle y, Hy \rangle \leq M|y|^2 . \tag{5.1.4}$$

Show that

$$|H| \leq M, \tag{5.1.5a}$$

$$|H^{-1}| \leq \frac{1}{m} . \tag{5.1.5b}$$

■

The basic idea behind the local Newton method is as follows. Given a current estimate x_i, of the

local minimizer $\hat{x}$, we expand $f(\cdot)$, approximately, to second order terms about x_i, to obtain

$$f(x) \simeq f(x_i) + \langle(\nabla f(x_i),(x - x_i)\rangle + \frac{1}{2}\langle(x - x_i),H(x_i)\,(x - x_i)\rangle. \tag{5.1.6}$$

If we minimize the right hand side of (5.1.6) w.r.t. x, we find that we can compute the minimizer x_{i+1} of the right hand side, by making use of the first order optimality condition, Theorem 3.1., viz.,

$$\nabla f(x_i) + H(x_i)\,(x_{i+1} - x_i) = 0\,. \tag{5.1.7}$$

Since $H(x_i)$ must be nonsingular for x_i close enough to $\hat{x}$, (5.1.7) defines the iterative process

$$x_{i+1} = x_i - H(x_i)^{-1}\,\nabla f(x_i),\quad i = 0, 1, 2,.... \tag{5.1.8}$$

We can restate (5.1.8) in the form of an algorithm, as follows:

Local Newton Algorithm 5.1.1 :

Data : $x_0 \in \mathbf{R}^n$.

Step 0 : Set $i = 0$.

Step 1 : Compute the *search direction*

$$h_i = -H(x_i)^{-1}\nabla f(x_i)\,. \tag{5.1.9a}$$

Step 2 : *Update*

$$x_{i+1} = x_i + h_i, \tag{5.1.9b}$$

replace i by $i + 1$ and go to step 1. ■

Theorem 5.1.1 : Suppose that Assumption 5.1.1 is satisfied. Then there exists a $\hat{\rho} > 0$ such that if $x_0 \in \mathbf{B}(\hat{x},\hat{\rho})$, then the sequence $\{\,x_i\,\}_{i=0}^{\infty}$, constructed by the Local Newton Algorithm 5.1.1, converges to $\hat{x}$ quadratically (with root rate 2).

Proof : Since $H(\cdot)$ is continuous, and Assumption 5.1 holds, there exist a $\rho > 0$ and an $L < \infty$ such that

$$\frac{m}{2}|y|^2 \le \langle y,H(x)y\rangle \le 2M|y|^2,\quad \forall\ x \in \mathbf{B}(\hat{x},\rho)\ \text{and}\ \forall\ y \in \mathbf{R}^n, \tag{5.1.10a}$$

$$|H(x') - H(x)| \le L|x - x'|,\quad \forall\ x', x \in \mathbf{B}(\hat{x},\rho)\,. \tag{5.1.10b}$$

Suppose that $x_i \in \mathbf{B}(\hat{x},\rho)$, then, from (5.1.7), we obtain that

$$H(x_i)(x_{i+1} - x_i) = -\nabla f(x_i) + \nabla f(\hat{x}) = -\int_0^1 H(x_i - s(x_i - \hat{x}))ds(x_i - \hat{x})\,. \tag{5.1.11a}$$

Hence

$$|x_{i+1} - \hat{x}| \le |H(x_i)^{-1}| \int_0^1 |H(x_i) - H(x_i - s(x_i - \hat{x}))|ds|x_i - \hat{x}|$$

$$\leq \frac{L}{m}|x_i - \hat{x}|^2 . \tag{5.1.11b}$$

Therefore, if $(L/m)|x_i - \hat{x}| < 1$, we must have that $|x_{i+1} - \hat{x}| < |x_i - \hat{x}|$ and hence that $x_{i+1} \in B(\hat{x},\rho)$. Therefore, if for any $\alpha \in (0,1)$, we define

$$\hat{\rho} \triangleq \min\{ \rho, m\alpha/L \}, \tag{5.1.12}$$

we obtain, by induction, that if $x_0 \in B(\hat{x},\hat{\rho})$, then the entire sequence $\{ x_i \}_{i=0}^{\infty}$, constructed by the Local Newton Algorithm, is well defined and contained in $B(\hat{x},\hat{\rho})$ and satisfies the relation (5.1.11b).

For $i=0,1,2,\ldots$, let

$$\mu_i = \frac{L}{m}|x_i - \hat{x}| . \tag{5.1.13}$$

Then, from (5.1.11b), we get that

$$\mu_{i+1} \leq \mu_i^2, \quad \text{for } i = 0, 1, 2, \ldots , \tag{5.1.14a}$$

so that

$$\mu_i \leq \mu_0^{2^i}, \quad \text{for } i = 0, 1, 2, \ldots , \tag{5.1.14b}$$

which proves that the sequence $\{ x_i \}_{i=0}^{\infty}$ converges to $\hat{x}$ quadratically. ■

Remark 5.1.1 : An examination of (5.1.8) shows that Newton's method is in reality a method for finding a solution to $\nabla f(x) = 0$, i.e., it is a method for solving a system of n equations in n variables. ■

Exercise 5.1.2 : Let $g:\mathbb{R}^n \to \mathbb{R}^n$ be locally Lipschitz continuously differentiable. Suppose that $\hat{x} \in \mathbb{R}^n$ is such that $g(\hat{x}) = 0$ and $g_x(\hat{x}) \triangleq \partial g(\hat{x})/\partial x$ is nonsingular. Show that if x_0 is sufficiently close to $\hat{x}$, then the sequence $\{ x_i \}_{i=0}^{\infty}$, constructed according to the Newtonian recursion

$$x_{i+1} = x_i - g_x(x_i)^{-1} g(x_i), \quad i = 0,1,2,..., \tag{5.1.15}$$

converges to $\hat{x}$ quadratically. ■

Exercise 5.1.3 : Let $g:\mathbb{R}^n \to \mathbb{R}^n$ be Lipschitz continuously differentiable, with *global* Lipschitz constant L. Suppose that $g_x(x)$ is nonsingular for all $x \in \mathbb{R}^n$ and that there exists an $M < \infty$ such that $|g_x(x)^{-1}| \leq M$ for all $x \in \mathbb{R}^n$. Show that if x_0 is such that

$$\frac{LM}{2}|g_x(x_0)^{-1} g(x_0)| < 1, \tag{5.1.16a}$$

then the sequence $\{ x_i \}_{i=0}^{\infty}$ constructed according to the Newtonian recursion (5.1.15) converges quadratically to a point $\hat{x} \in \mathbb{R}^n$ such that $g(\hat{x}) = 0$.

Hint: First show that

$$g_x(x_i)(x_{i+1} - x_i) = -g(x_i) + g_x(x_{i-1})(x_i - x_{i-1}) + g(x_{i-1}) . \tag{5.1.16b}$$

Then use a first order expansion of $g(x_i)$ about x_{i-1} to obtain that

$$|x_{i+1} - x_i| \leq \frac{LM}{2}|x_i - x_{i-1}|^2 . \tag{5.1.16c}$$

Finally, make use of Theorem 5.1.1 (b). ■

Remark 5.1.2 : The convergence and divergence properties of the local Newton method are demonstrated in Fig. 5.1.1 for the case where $x \in \mathbb{R}$ (with $g(x) \triangleq f'(x)$) Thus the lack of *global convergence* is a fact that is demonstratable empirically. ■

Exercise 5.1.4 : Use Newton's method to prove the Implicit Function Theorem.

Hint : Given $(\bar{x}, \bar{y})$ such that $g(\bar{x},\bar{y}) = 0$, show that Newton's method (5.1.15) constructs a solution $x(\bar{y} + \delta y)$ of the equation $f(x,\bar{y} + \delta y) = 0$ for all δy such that $\left|\left[\frac{\partial g(\bar{x},\bar{y} + \delta y)}{\partial x}\right]^{-1} g(\bar{x},\bar{y} + \delta y)\right| < \hat{\rho}$ where $\hat{\rho}$ is an appropriate constant. ■

5.2 GLOBAL NEWTON METHOD FOR CONVEX FUNCTIONS

We will now show that the local Newton method can be "globalized" for the case where the function $f:\mathbb{R}^n \to \mathbb{R}$ in (5.1.1) is strictly convex, by adding an Armijo type step size rule to the local method. Furthermore, we will see that the global method converges with quadratic rate.

Assumption 5.2.1 :

(a) The function $f:\mathbb{R}^n \to \mathbb{R}$ is twice continuously locally Lipschitz differentiable, i.e., given any bounded set $S \subset \mathbb{R}^n$ there exists an $L_S < \infty$ such that

$$|H(x') - H(x)| \leq L_S|x - x'| , \tag{5.2.2a}$$

for all $x, x' \in S$.

(b) There exists an $m > 0$, such that

$$m|y|^2 \leq \langle y, H(x)y\rangle , \quad \forall \ x, y \in \mathbb{R}^n . \tag{5.2.2b}$$

■

Proposition 5.2.2 : Suppose that Assumption 5.2.1 is satisfied. Then (a) the level sets of $f(\cdot)$ are compact and, (b) problem (5.1.1) has a unique minimizer.

Proof : (a) Suppose that $x_0 \in \mathbb{R}^n$ is arbitrary and that $x \in L_0 \triangleq \{ x \in \mathbb{R}^n \mid f(x) \leq f(x_0) \}$. Then we must have that

$$0 \geq f(x) - f(x_0) = \langle \nabla f(x_0), x - x_0\rangle + \int_0^1 (1 - s)\langle (x - x_0), H(x_0 + s(x - x_0))(x - x_0)\rangle ds$$

$$\geq \left[-|\nabla f(x_0)| + \frac{m}{2}|x - x_0|\right]|x - x_0| , \tag{5.2.3a}$$

which leads to the conclusion that $|x - x_0| \leq 2|\nabla f(x_0)|/m$ must hold, i.e. that L_0 is bounded. The fact that L_0 is closed follows from the continuity of $f(\cdot)$.

(b) Since the level sets of $f(\cdot)$ are compact, (5.1.1) must have at least one global minimizer. Suppose that x^*, x^{**} are two minimizers. Then, because of (5.2.2b), the fact that $\nabla f(x^*) = 0$ and (2.4.7d), we must have that

$$f(x^{**}) - f(x^*) = \int_0^1 (1-s)\langle (x^{**} - x^*), H(x^* + s(x^* - x^{**}))(x^* - x^{**})\rangle ds$$

$$\geq \frac{m}{2}|x^{**} - x^*|^2 . \tag{5.2.3b}$$

Since $f(x^*) = f(x^{**})$, we conclude from (5.2.3b) that $x^* = x^{**}$, i.e., that there is only one global minimizer, which completes our proof. ■

Armijo-Newton Algorithm 5.2.1 :

Parameters : $\alpha \in (0,1/2)$, $\beta \in (0,1)$.

Data : $x_0 \in \mathbb{R}^n$.

Step 0 : Set $i = 0$.

Step 1 : Compute the Newton *search direction*

$$h_i \triangleq -H(x_i)^{-1} \nabla f(x_i) . \tag{5.2.4a}$$

Stop if $h_i = 0$.

Step 2 : Compute the Armijo *step size*

$$\lambda_i = \max_{k \in \mathbf{N}} \{ \beta^k \mid f(x_i + \beta^k h_i) - f(x_i) \leq \alpha\beta^k \langle h_i, \nabla f(x_i)\rangle \} . \tag{5.2.4b}$$

Step 3 : *Update*

$$x_{i+1} = x_i + \lambda_i h_i, \tag{5.2.4c}$$

replace i by $i + 1$ and go to Step 1. ■

Theorem 5.2.1 : Suppose that Assumption 5.2.1 is satisfied. If $\{ x_i \}_{i=0}^{\infty}$ is a sequence constructed by the Armijo-Newton Algorithm, then $x_i \to \hat{x}$ as $i \to \infty$, *quadratically*, where $\hat{x}$ is the unique minimizer of $f(\cdot)$.

Proof : The proof is in two parts. First we prove that $x_i \to \hat{x}$ as $i \to \infty$, with $\nabla f(\hat{x}) = 0$, then we show that this convergence is quadratic.

(a) By Proposition 5.2.1, the level sets of $f(\cdot)$ are compact. Let $\mathbf{L}_0 \triangleq \{ x \in \mathbb{R}^n \mid f(x) \leq f(x_0) \}$. Let $h : \mathbb{R}^n \to \mathbb{R}$ be defined by $h(x) \triangleq -H(x)^{-1}\nabla f(x)$ and let

$$\gamma \triangleq \max\{ f(x + \lambda h(x)) \mid x \in \mathbf{L}_0,\ \lambda \in [0,1] \}. \tag{5.2.5}$$

Then (i) since $H(\cdot)$ is continuous by assumption, there exists an $M \in [m,\infty)$, such that

$$\langle y, H(x)y\rangle \leq M|y|^2 \ \forall\ y \in \mathbb{R}^n ,\ x \in \{ x \mid f(x) \leq \gamma \} ; \tag{5.2.6}$$

and, (ii) $\{ x_i \}_{i=0}^{\infty}$ must have accumulation points, all of which are in the level set $\mathbf{L}_0$.

Suppose that $\hat{x}$ is an accumulation point of $\{x_i\}_{i=0}^{\infty}$ such that $\nabla f(\hat{x}) \neq 0$. Then for any $x \in \mathbf{L}_0$ such that $\nabla f(x) \neq 0$, expanding $f(\cdot)$ to second order and making use of (5.2.2b) and (5.2.6), we get that

$$f(x + \lambda h(x)) - f(x) - \alpha\lambda \langle \nabla f(x), h(x)\rangle = \lambda\left[(1-\alpha)\langle \nabla f(x), h(x)\rangle + \lambda\int_0^1 (1-s)\langle h(x), H(x + s\lambda h(x))h(x)\rangle\, ds\right]$$

$$\leq \lambda\left[-(1-\alpha)\langle \nabla f(x), H(x)^{-1}\nabla f(x)\rangle + \frac{\lambda M}{2}\|H(x)^{-1}\nabla f(x)\|^2\right]$$

$$\leq \lambda\|\nabla f(x)\|^2\left[-\frac{(1-\alpha)}{M} + \frac{\lambda M}{2m^2}\right]. \tag{5.2.7}$$

Hence (just as for the Armijo method) there exists a $\hat{\lambda} = \beta^{\hat{i}} > 2\beta(1-\alpha)m^2/M^2 > 0$ such that for all $x \in \mathbb{R}^n$ satisfying $\nabla f(x) \neq 0$,

$$f(x + \hat{\lambda}h(x)) - f(x) - \alpha\hat{\lambda}\langle \nabla f(x), h(x)\rangle < 0\,. \tag{5.2.8}$$

It follows from (5.2.8) that for all $i \in \mathbb{N}$, $\lambda_i \geq \hat{\lambda}$. Now, by assumption, $\nabla f(\hat{x}) \neq 0$ and hence $h(\hat{x}) \neq 0$. It now follows from the continuity of $h(\cdot)$ and $\nabla f(\cdot)$ that if $x_i \overset{K}{\to} \hat{x}$ as $i \to \infty$, then there exists an i_0 such that for all $i \geq i_0$, $i \in K$, $\langle \nabla f(x_i), h_i\rangle \leq \frac{1}{2}\langle \nabla f(\hat{x}), h(\hat{x})\rangle < 0$, so that

$$f(x_{i+1}) - f(x_i) \leq \alpha\lambda_i\langle \nabla f(x_i), h_i\rangle$$

$$\leq \frac{\alpha\hat{\lambda}}{2}\langle \nabla f(\hat{x}), \hat{h}\rangle < 0, \quad \forall\, i \in K, i \geq i_0\,. \tag{5.2.9}$$

Since $\{f(x_i)\}_{i=0}^{\infty}$ is monotone decreasing by construction, (5.2.9) shows that $f(x_i) \to -\infty$. However, $x_i \overset{K}{\to} \hat{x}$ implies that $f(x_i) \to f(\hat{x})$, and thus we have a contradiction. Consequently, we must have that $\nabla f(\hat{x}) = 0$. Since by Proposition (5.2.1), there is only one point $\hat{x}$ such that $\nabla f(\hat{x}) = 0$, we must have that $x_i \to \hat{x}$ as $i \to \infty$.

(b) To show that $x_i \to \hat{x}$ quadratically, it is necessary to repeat the calculation in (5.2.7) with greater care and to make use of the fact that $\alpha \in (0,0.5)$.

First we observe that because $H(\cdot)$ is locally Lipschitz continuous, there exists an $L \in (0,\infty)$ such that

$$\|H(x') - H(x)\| \leq L\|x - x'\|, \tag{5.2.10}$$

for all $x, x' \in \{x \mid f(x) \leq \gamma\}$. Next, setting $\lambda = 1$ in the expression for the Armijo step size rule (see (5.2.4b)), we obtain, upon expanding to second order terms, and adding and subtracting the term $\langle \nabla f(x_i), H(x_i)^{-1}\nabla f(x_i)\rangle$, that

$$f(x_i + h_i) - f(x_i) - \alpha\langle \nabla f(x_i), h_i\rangle$$

$$= (1-\alpha)\langle \nabla f(x_i), h_i\rangle + \int_0^1 (1-s)\langle h_i, H(x_i+sh_i)h_i\rangle ds$$

$$= -(1-\frac{1}{2}-\alpha)\langle \nabla f(x_i), H(x_i)^{-1}\nabla f(x_i)\rangle$$

$$+ \int_0^1 (1-s)\langle H(x_i)^{-1}\nabla f(x_i), \Big[H(x_i+s\lambda h_i) - H(x_i)\Big]H(x_i)^{-1}\nabla f(x_i)\rangle ds$$

$$\leq |\nabla f(x_i)|^2\left[-(1-2\alpha)\frac{1}{2M} + \frac{L}{6m^3}|\nabla f(x_i)|\right]. \tag{5.2.11}$$

Since $\alpha \in (0,0.5)$ and $\nabla f(x_i) \to 0$ as $i \to \infty$, it follows that there must exist an i_0 such that for all $i \geq i_0$,

$$f(x_i + h_i) - f(x_i) - \alpha\langle \nabla f(x_i), h_i\rangle \leq 0, \tag{5.2.12}$$

which shows that $\lambda_i = 1$ for all $i \geq i_0$. Since this shows that the global method degenerates to the local method as the solution $\hat{x}$ is approached, the rate of convergence result follows. ■

Exercise 5.2.1 : (a) Prove that if an algorithm for minimizing a function $f:\mathbb{R}^n \to \mathbb{R}$ follows a recursion of the form

$$x_{i+1} \in A(x_i) \tag{5.2.13}$$

with $A: \mathbb{R}^n \to 2^{\mathbb{R}^n}$ such that for each x^* satisfying $\nabla f(x^*) \neq 0$, there exist $\rho^* > 0$, $\delta^* > 0$ such that

$$f(x') - f(x) \leq -\delta^*, \quad \forall\, x' \in A(x), \; x \in B(x^*,\rho^*), \tag{5.2.14}$$

then $x_i \overset{K}{\to} \hat{x}$ as $i \to \infty$ implies that $\nabla f(\hat{x}) = 0$.

(b) Show that this principle was used in proving the convergence both of the Armijo Gradient and the Armijo-Newton Algorithms. ■

Exercise 5.2.2 : Use the above result to combine the Armijo gradient and Newton Algorithms into an algorithm that works for convex as well as nonconvex functions. ■

5.3 AN AID FOR GLOBAL STABILIZATION OF LOCAL ALGORITHMS

The observation outlined in Exercise 5.2.1 is certainly very helpful in the construction of new algorithms. However, its use can be made considerably easier if we specialize it to particular cases. We shall now present such a specialization to be used in unconstrained optimization.

Theorem 5.3.1 : (Polak-Sargent-Sebastian).

Let $f:\mathbb{R}^n \to \mathbb{R}$ be twice continuously† differentiable. Consider an algorithm for the minimization of $f(\cdot)$ which uses a search direction map $h:\mathbb{R}^n \to \mathbb{R}^n$ and the Armijo step size rule: viz., with $\alpha, \beta \in (0,1)$, given x_i, the algorithm constructs x_{i+1} according to the recursion

† It is also possible to prove this theorem under the weaker assumption that $f(\cdot)$ is only once continuously differentiable.

$$x_{i+1} = x_i + \lambda_i h(x_i), \tag{5.3.1a}$$

where

$$\lambda_i = \max_{k \in \mathbf{N}} \{ \beta^k \mid f(x_i + \beta^k h(x_i)) - f(x_i) \leq -\beta^k \alpha \langle \nabla f(x_i), h(x_i) \rangle \} . \tag{5.3.1b}$$

Suppose that there exist two continuous functions $N_1 : \mathbf{R}^n \to \mathbf{R}$, $N_2 : \mathbf{R}^n \to \mathbf{R}$ such that for all x satisfying $\nabla f(x) \neq 0$, (i) $N_1(x) > 0$, $N_2(x) >$ and, (ii) for all $x \in \mathbf{R}^n$

$$\langle \nabla f(x), h(x) \rangle \leq -N_1(x), \tag{5.3.2a}$$

$$|h(x)| \leq N_2(x) . \tag{5.3.2b}$$

If $\{ x_i \}_{i=0}^{\infty}$ is a sequence constructed by this algorithm, then any accumulation point $\hat{x}$ of this sequence satisfies $\nabla f(\hat{x}) = 0$.

Proof: Suppose that $x_i \xrightarrow{K} \hat{x}$ as $i \to \infty$ and $\nabla f(\hat{x}) \neq 0$. Then $|h(\hat{x})| \leq N_2(\hat{x})$ and $\langle \nabla f(\hat{x}), h(\hat{x}) \rangle \leq -N_1(\hat{x}) < 0$. Clearly, this implies that $|h(\hat{x})| > 0$. Consider the computation of the step size λ_i, for $x_i \in B(\hat{x}, \hat{\rho})$, where $\hat{\rho} > 0$ is such that $N_1(x) \geq \frac{1}{2} N_1(\hat{x}) > 0$ and $N_2(x) \leq \frac{3}{2} N_2(\hat{x})$ hold. Note that because $f(\cdot)$ is twice continuously differentiable and $N_2(x) \leq \frac{3}{2} N_2(\hat{x})$ holds for all $x \in B(\hat{x}, \hat{\rho})$, the bound

$$M \triangleq \sup \{ \|H(x')\| \mid x' = x + \mu h(x), \mu \in [0,1], x \in B(\hat{x}, \hat{\rho}) \} < \infty , \tag{5.3.3a}$$

where $H(x) \triangleq \partial^2 f(x) / \partial x^2$. Since $x_i \xrightarrow{K} \hat{x}$, as $i \to \infty$, there exists an i_0 such that for all $i \geq i_0$, $i \in K$, $x_i \in B(\hat{x}, \hat{\rho})$ and therefore

$$\begin{aligned} & f(x_i + \lambda h(x_i)) - f(x_i) - \lambda \alpha \langle \nabla f(x_i), h(x_i) \rangle \\ & \quad = \lambda (1 - \alpha) \langle \nabla f(x_i), h(x_i) \rangle + \lambda^2 \int_0^1 (1 - s) \langle H(x_i + s\lambda h(x_i)) h(x_i), h(x_i) \rangle ds \\ & \quad \leq \lambda \left[-\frac{1}{2}(1 - \alpha) N_1(\hat{x}) + \frac{9}{4} \lambda M N_2(\hat{x})^2 \right] . \end{aligned} \tag{5.3.3b}$$

It follows form (5.3.3b) that there exists a $\hat{k} \in \mathbf{N}$, $\hat{k} > 0$ such that for all $x_i \in B(\hat{x}, \hat{\rho})$, the step length λ_i must satisfy $\lambda_i \geq \beta^{\hat{k}} \geq \min\{ 1, 2\beta(1 - \alpha) N_1(\hat{x}) / 9 M N_2(\hat{x})^2 \}$. Consequently, for all $i \in K$, $i \geq i_0$ (so that $x_i \in B(\hat{x}, \hat{\rho})$),

$$\begin{aligned} & f(x_{i+1}) - f(x_i) \leq \lambda_i \alpha \langle \nabla f(x_i), h(x_i) \rangle \\ & \quad \leq -\beta^{\hat{k}} \alpha N_1(\hat{x}) < 0 . \end{aligned} \tag{5.3.3c}$$

Since $\{ f(x_i) \}_{i=0}^{\infty}$ is monotone decreasing, (5.3.3c) implies that $f(x_i) \to -\infty$ as $i \to \infty$. However, $f(\cdot)$ is continuous and $x_i \xrightarrow{K} \hat{x}$ as $i \to \infty$, which implies that $f(x_i) \to f(\hat{x})$ as $i \to \infty$. Thus we have a contradiction, which completes our proof.

■

Remark 5.3.1 : Note that the above proof did not require $h(x)$ to be continuous! ■

An alternative result, to be used with an exact minimization stepsize is given in the following

Theorem 5.3.2 : Let $f:\mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Consider an algorithm for the minimization of $f(\cdot)$ which uses a search direction map $h:\mathbb{R}^n \to \mathbb{R}^n$ and an exact minimization step size rule: viz., the algorithm constructs x_{i+1} according to the recursion

$$x_{i+1} = x_i + \lambda_i h(x_i), \tag{5.3.4a}$$

where

$$\lambda_i = \arg\min_{\lambda > 0} f(x_i + \lambda h(x_i)) - f(x_i) . \tag{5.3.4b}$$

Suppose that (i) $h(x) \neq 0$ for all $x \in \mathbb{R}^n$ such that $\nabla f(x) \neq 0$, and (ii) there exists a $\rho \in (0,1]$ such that for all $i \in \mathbb{N}$

$$\langle \nabla f(x_i), h(x_i) \rangle \leq -\rho |\nabla f(x_i)| |h(x_i)| . \tag{5.3.4c}$$

If $\{ x_i \}_{i=0}^{\infty}$ is a sequence constructed by this algorithm, then any accumulation point $\hat{x}$ of this sequence satisfies $\nabla f(\hat{x}) = 0$. ■

Exercise 5.3.1 : Prove Theorem 5.3.2. ■

Exercise 5.3.2 : Show that the following stabilized version of the local Newton method satisfies the hypotheses of Theorem 5.3.1 for functions $f:\mathbb{R}^n \to \mathbb{R}$ that are twice locally Lischitz differentiable and that it will converge quadratically under suitable assumptions.

Stabilized Armijo Newton Algorithm 5.3.1 :

Parameters : $\alpha \in (0,1/2)$, $\beta \in (0,1)$.

Data: $x_0 \in \mathbb{R}^n$.

Step 0: Set $i = 0$.

Step 1: Compute the Hessian matrix $H(x_i)$ and its largest and smallest eigenvalues, $\lambda_{max}(x_i)$, $\lambda_{min}(x_i)$.

If $\lambda_{min}(x_i) \geq 10^{-8}$ and $\lambda_{max}(x_i)/\lambda_{min}(x_i) \leq 10^7$, set the *search direction* to be

$$h_i \triangleq -H(x_i)^{-1} \nabla f(x_i) . \tag{5.3.5a}$$

Else, set the *search direction* to be

$$h_i \triangleq -\nabla f(x_i) . \tag{5.3.5b}$$

Step 2 : Compute the Armijo *step size*

$$\lambda_i = \max_{k \in \mathbb{N}} \{ \beta^k \mid f(x_i + \beta^k h_i) - f(x_i) \leq \alpha \beta^k \langle h_i, \nabla f(x_i) \rangle \} . \tag{5.3.5c}$$

Step 3 : *Update*:

$$x_{i+1} = x_i + \lambda_i h_i, \tag{5.3.5d}$$

replace i by $i + 1$ and go to Step 1. ■

6. METHODS OF CONJUGATE DIRECTIONS

6.0. INTRODUCTION

As we shall shortly see, when applied to convex problems of the form $\min_{x \in \mathbb{R}^n} f(x)$, with $f:\mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable, methods of conjugate directions accomplish in n iterations what Newton's method accomplishes in one. However, for large problems, because they do not require the computation of the hessian matrix, their efficiency is considerably better than that of Newton's method. Methods of conjugate gradients emanate from an iterative formula, for constructing, simultaneously, both an orthogonal and a conjugate basis, by using a positive definite matrix. A conjugate gradient method was first proposed by Hestenes and Stiefel as a method for solving equations of the form $Hh = g$, when the matrix H is positive definite. The formula makes use of the fact that the two problems $\mathbf{P}_1$: find an $h \in \mathbb{R}^n$ such that $Hh = g$ and $\mathbf{P}_2$: $\min_{h \in \mathbb{R}^n} |Hh - g|^2$ have the same solution.

6.1. METHODS OF CONJUGATE DIRECTIONS : QUADRATIC FUNCTIONS

We begin with definitions and a few preliminary results.

Definition 6.1.1 : A basis $\{ u_i \}_{i=1}^n$, for $\mathbb{R}^n$ is said to be *orthogonal* if $\langle u_i, u_j \rangle = 0$ for all $i \neq j$. ■

Definition 6.1.2 : Given a symmetric, positive definite matrix H, a basis $\{ u_i \}_{i=1}^n$ for $\mathbb{R}^n$ is said to be *H–conjugate* (or simply *conjugate*) if $\langle u_i, Hu_j \rangle = 0$ for all $i \neq j$. ■

Exercise 6.1.1 : Suppose that $\{ u_i \}_{i=1}^n$ is a basis for $\mathbb{R}^n$. Show that the following process constructs an orthogonal basis $\{ v_i \}_{i=1}^n$ for $\mathbb{R}^n$:

$$v_1 = u_1, \tag{6.1.4a}$$

$$v_i = u_i + \sum_{j=1}^{i-1} \lambda_{ij} v_j, \tag{6.1.4b}$$

with λ_{ij} determined from the equation

$$0 = \langle v_k, v_j \rangle = \langle v_j, u_k \rangle + \lambda_{ij} \langle v_j, v_j \rangle. \tag{6.1.4c}$$

■

Exercise 6.1.2 : Let H be a symmetric, positive definite $n \times n$ matrix and suppose that $g_0 \in \mathbb{R}^n$, not an eigenvector of H, is given. Show that if the following process does not stop because $g_i = 0$, then it constructs *simultaneously* both an orthogonal basis $\{ g_i \}_{i=1}^n$ and an H–conjugate basis $\{ h_i \}_{i=1}^n$ basis for $\mathbb{R}^n$:

$$\left.\begin{aligned} &g_0 \text{ given,}\\ &g_{i+1} = g_i + \lambda_i H h_i,\\ &\lambda_i = -\|g_i\|^2/\langle g_i, H h_i\rangle \end{aligned}\right\},\ i = 0,1,\ldots,n-1; \tag{6.1.2a}$$

$$\left.\begin{aligned} &h_0 = g_0,\\ &h_{i+1} = g_{i+1} + \gamma_i h_i\\ &\gamma_i = -\langle H h_i, g_{i+1}\rangle / \langle H h_i, h_i\rangle \end{aligned}\right\},\ i = 0,1,\ldots,n-1\ . \tag{6.1.2b}$$

■

The next two results complete our presentation of the basic facts which make conjugate gradient methods work.

Theorem 6.1.1 : Let $f:\mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Suppose that $\hat{x}$ minimizes $f(\cdot)$ on the subspace spanned by the vectors $y_1, y_2, \ldots, y_k$. then $\langle \nabla f(\hat{x}), y_i\rangle = 0$ for $i = 1, 2, \ldots, k$.

Proof : The subspace in question is

$$\{\ x \mid x = Y\alpha\ \}, \tag{6.1.3}$$

where Y is an $n \times k$ matrix with columns y_i, $i = 1, 2, \ldots, k$. Hence $\hat{x} = Y\hat{\alpha}$, with $\hat{\alpha} \in \mathbb{R}^k$, solves the problem

$$\min_{\alpha \in \mathbb{R}^k} f(Y\alpha)\ , \tag{6.1.4}$$

i.e., $\hat{\alpha}$ solves the problem $\min_{\alpha \in \mathbb{R}^k} g(\alpha)$, with $g(\alpha) \triangleq f(Y\alpha)$. Hence we must have that

$$\begin{aligned} 0 = \nabla g(\hat{\alpha}) &= Y^T \nabla f(Y\hat{\alpha})\\ &= Y^T \nabla f(\hat{x}). \end{aligned} \tag{6.1.5}$$

■

The desired result now follows immediately.

The following result is known as the Expanding Subspace Theorem.

Theorem 6.1.2 : Let H be a symmetric, positive definite, $n \times n$ matrix, and let $d \in \mathbb{R}^n$ be arbitrary. Let

$$f(x) \triangleq \tfrac{1}{2}\langle x, Hx\rangle + \langle d, x\rangle \tag{6.1.6}$$

and let $\{\ h_i\ \}_{i=0}^{n-1}$ be an H–conjugate basis for $\mathbb{R}^n$. If $\{\ x_i\ \}_{i=0}^{n}$ is a sequence constructed according to the recursion

$$\left.\begin{aligned} &x_0 \text{,given,}\\ &x_{i+1} = x_i - \lambda_i h_i,\\ &\lambda_i = \operatorname*{argmin}_{\lambda \geq 0} f(x_i - \lambda\, h_i)\ , \end{aligned}\right\},\ i = 0,1,\ldots,n, \tag{6.1.7}$$

then x_i minimizes $f(\cdot)$ on the subspace spanned by $h_0, h_1, \ldots, h_{i-1}$.

Proof : For $i = 0, 1, ...,n$, let $g_i \triangleq \nabla f(x_i) = Hx_i + d$. Then, because of of the rule for λ_i in (6.1.7), for $i = 0,1,2,....,n-1$,

$$\frac{\partial f(x_i - \lambda h_i)}{\partial \lambda}\Big|_{\lambda = \lambda_i} = -\langle \nabla f(x_{i+1}), h_i \rangle$$

$$= -\langle g_{i+1}, h_i \rangle = 0 . \tag{6.1.8}$$

Also,

$$x_i = x_{i-1} - \lambda_{i-1} h_{i-1}$$

$$= x_{i-2} - \lambda_{i-2} h_{i-2} - \lambda_{i-1} h_{i-1}$$

$$= x_0 - \lambda_0 h_0 - \lambda_1 h_1 - \ldots\ldots - \lambda_{i-1} h_{i-1} . \tag{6.1.9a}$$

Similarly, making use of (6.1.9a),

$$g_i = Hx_i + d$$

$$= Hx_0 + d - \lambda_0 Hh_0 - \lambda_1 Hh_1 - \ldots\ldots - \lambda_{i-1} Hh_{i-1}$$

$$= g_0 - \lambda_0 Hh_0 - \lambda_1 Hh_1 - \ldots\ldots - \lambda_{i-1} Hh_{i-1} . \tag{6.1.9b}$$

Since the h_j are H–conjugate, it follows that for any $0 \le k < i$,

$$\langle g_i, h_k \rangle = \langle g_0, h_k \rangle - \lambda_k \langle h_k, Hh_k \rangle . \tag{6.1.10}$$

Because of (6.1.8), we get from (6.1.10) that

$$0 = \langle g_{k+1}, h_k \rangle = \langle g_0, h_k \rangle - \lambda_k \langle h_k, Hh_k \rangle \tag{6.1.11a}$$

and hence, from (6.1.10), that

$$\langle g_i, h_k \rangle = 0 , \quad \text{for } k = 0, 1, ..., i - 1 . \tag{6.1.11b}$$

Thus x_i satisfies the necessary conditions of optimality stated in Theorem 6.1.1. Since, $f(\cdot)$ is strictly convex, this condition is not only necessary, but also sufficient, i.e., x_i minimizes $f(\cdot)$ on the subspace spanned by $h_0, h_1, ... h_{i-1}$. ■

Corollary : The vector x_n, constructed as above, minimizes the quadratic function

$$f(x) = \tfrac{1}{2}\langle x, Hx \rangle + \langle d, x \rangle \tag{6.1.12}$$

on $\mathbb{R}^n$. ■

It is possible to construct several algorithms which produce H–conjugate directions while minimizing a quadratic function $f(x) = \frac{1}{2}\langle x, Hx \rangle + \langle d, x \rangle$. The different conjugate gradient algorithms use different formulae for computing the parameter γ_i for (6.1.13), below. The search directions produced by these algorithms are identical when minimizing a quadratic cost and hence they construct identical

sequences of points x_i. However, when $f(\cdot)$ is not quadratic, their behavior can be quite different. We state these algorithms in the form of a *master algorithm* which does not use a specific formula for computing the parameter γ_i, and which is applicable *only* to quadratic functions of the form $f(x) = \frac{1}{2}\langle x, Hx\rangle + \langle d, x\rangle$ with $H > 0$ and symmetric. Extensions to more general functions will be described in the next section.

Master Conjugate Gradient Algorithm 6.1.1 :

(Solves $\min_{x \in \mathbb{R}^n} \frac{1}{2}\langle x, Hx\rangle + \langle d, x\rangle$, $H > 0$, $n \times n$ symmetric.)

Data : $x_0 \in \mathbb{R}^n$.

Step 0 : Set $i = 0$, $h_0 = g_0 \triangleq Hx_0 + d$.

Step 1 : Compute the *step length*

$$\lambda_i = \arg \min_{\lambda \geq 0} f(x_i - \lambda h_i) . \tag{6.1.13a}$$

Step 2 : *Update:* set

$$\begin{cases} x_{i+1} = x_i - \lambda_i h_i, \\ g_{i+1} \triangleq Hx_{i+1} + d, \\ h_{i+1} = g_{i+1} + \gamma_i h_i, \end{cases} \tag{6.1.13b}$$

with

$$\gamma_i = \frac{\langle Hh_i, g_{i+1}\rangle}{\langle h_i, Hh_i\rangle} , \tag{6.1.13c}$$

so that

$$\langle h_{i+1}, Hh_i\rangle = 0 . \tag{6.1.13d}$$

Step 3 : Replace i by $i + 1$ and go to Step 1. ■

Theorem 6.1.3 : Suppose that H is an $n \times n$ symmetric matrix and $d \in \mathbb{R}^n$. Then The Master Conjugate Gradient Algorithm solves the problem

$$\min_{x \in \mathbb{R}_n} \{ \tfrac{1}{2}\langle x, Hx\rangle + \langle d, x\rangle \} \tag{6.1.14}$$

in at most n iterations.

Proof : In view of Theorem 6.1.2, we only need to show that $\langle h_i, Hh_j\rangle = 0$ for all $i \neq j$, with h_i, g_i constructed by the algorithm. In the process we will also show that $\langle g_i, g_j\rangle = 0$, for all $i \neq j$.

Our proof proceeds by induction. Since $h_0 = g_0$,

$$\langle g_1, h_0\rangle = \langle g_1, g_0\rangle = 0, \tag{6.1.15a}$$

by construction of λ_0. Next, by construction of γ_0,

$$\langle h_1, Hh_0\rangle = 0\,. \tag{6.1.15b}$$

Hence, suppose that

$$\left.\begin{aligned}\langle g_i, g_j\rangle &= 0\\ \langle h_i, Hh_j\rangle &= 0\end{aligned}\right\},\ \forall\ 0 \le i, j \le k < n,\ i \neq j\,. \tag{6.1.16}$$

First, by construction of λ_i, $\langle g_{i+1}, h_i\rangle = 0$ for all i. Next,

$$g_{i+1} = Hx_{i+1} + d = H(x_i - \lambda_i h_i) + d = g_i - \lambda_i Hh_i,\ \text{ for } i = 0,1,\ldots,n\,. \tag{6.1.17}$$

Consequently,

$$\langle g_{i+1}, h_i\rangle = 0 = \langle h_i, g_i\rangle - \lambda_i\langle h_i, Hh_i\rangle,\ \text{ for } i = 0,1,\ldots,n, \tag{6.1.18}$$

so that

$$\lambda_i = \frac{\langle h_i, g_i\rangle}{\langle h_i, Hh_i\rangle},\ \text{ for } i = 0,1,\ldots,n\,. \tag{6.1.19}$$

Thus, from (6.1.17) and (6.1.19), we get that for all $i = 0,1,\ldots,n-1$,

$$\langle g_i, g_{i+1}\rangle = \langle g_i, g_i\rangle - \frac{\langle h_i, g_i\rangle\langle g_i, Hh_i\rangle}{\langle h_i, Hh_i\rangle}\,. \tag{6.1.20a}$$

Now

$$\langle h_i, g_i\rangle = \langle g_i + \gamma_{i-1}h_{i-1}, g_i\rangle = \langle g_i, g_i\rangle \tag{6.1.20b}$$

$$\langle h_i, Hh_i\rangle = \langle g_i + \gamma_{i-1}h_{i-1}, Hh_i\rangle = \langle g_i, Hh_i\rangle. \tag{6.1.20c}$$

Substituting into (6.1.20a), we get

$$\langle g_i, g_{i+1}\rangle = 0,\ \text{ for } i = 0,1,\ldots,n, \tag{6.1.21}$$

so that $\langle g_{k+1}, g_k\rangle = 0$.

Next for all $i \neq 0$, $i < k$, since $\langle g_i, g_k\rangle = 0$ and $\langle h_i, Hh_k\rangle = 0$,

$$\begin{aligned}\langle g_{k+1}, g_i\rangle &= \langle g_k - \lambda_k Hh_k, g_i\rangle\\ &= -\lambda_k\langle Hh_k, g_i\rangle\\ &= -\lambda_k\langle Hh_k, h_i - \gamma_{i-1}h_{i-1}\rangle = 0\,.\end{aligned} \tag{6.1.22}$$

Finally, since $g_0 = h_0$,

$$\begin{aligned}\langle g_{k+1}, g_0\rangle &= -\lambda\langle Hh_k, g_0\rangle\\ &= -\lambda_k\langle Hh_k, h_0\rangle = 0,\end{aligned} \tag{6.1.23}$$

which completes our proof that the vectors in the set $\{\, g_i \,\}_{i=0}^{n-1}$ are orthogonal.

Next, for $i = k$, $\langle h_{k+1}, Hh_k\rangle = 0$, by construction of γ_k. For $0 \le i < k$,

$$\langle h_{k+1}, Hh_i\rangle = \langle g_{k+1} + \gamma_k h_k, Hh_i\rangle$$

$$= \langle g_{k+1}, Hh_i\rangle$$

$$= \langle g_{k+1}, \frac{1}{\lambda_i}(g_{i+1} - g_i)\rangle$$

$$= 0, \tag{6.1.24}$$

■

which shows that the $\{ h_i \}_{i=0}^{n-1}$ are H–conjugate. This completes our proof.

Thus we see that a conjugate gradient algorithm takes at most n iterations in minimizing a quadratic function on $\mathbb{R}^n$. Newton's method requires only one iteration on this problem. Hence, we may hope that on general problems, properly constructed versions of conjugate gradient algorithms may turn out to be n–step quadratically convergent. This is in fact true.

6.2. METHODS OF CONJUGATE DIRECTIONS : GENERAL FUNCTIONS

As we have seen in the preceding section, even for the quadratic problem $\min_{x \in \mathbb{R}^n} \frac{1}{2}\langle x, Hx\rangle + \langle d, x\rangle$ with H symmetrical and positive definite, the Master Conjugate Gradient method requires n iterations to produce the same result as Newton's method does in one iteration. Furthermore, it consumes a lot of computing time in finding the step size. Hence the use of a conjugate gradient method can be justified only if the Hessian matrix calculation can be avoided. If we were to extend *formally* the Master Conjugate Gradient Algorithm 6.1.1 to the solution of

$$\min_{x \in \mathbb{R}^n} f(x), \tag{6.2.1}$$

with $f: \mathbb{R}^n \to \mathbb{R}$ continuously differentiable, then, referring to (6.1.13c), we would have to set $\gamma_i = -\langle H(x_i)h_i, \nabla f(x_{i+1})\rangle / \langle H(x_i)h_i, h_i\rangle$, which involves the Hessian matrix. We shall now develop alternative formulas for γ_i, which are equivalent to (6.1.23c) for the quadratic case, but which enable us to apply conjugate gradient algorithms to the problem (6.2.1) without computing Hessians.

First, *for the quadratic case*, where $f(x) = \frac{1}{2}\langle x, Hx\rangle + \langle d, x\rangle$, with $H > 0$, symmetrical, we obtain from (6.1.13c) that

$$\gamma_i = -\frac{\langle Hh_i, g_{i+1}\rangle}{\langle Hh_i, h_i\rangle}. \tag{6.2.2}$$

Since by construction

$$\lambda_i Hh_i = g_{i+1} - g_i, \tag{6.2.3}$$

we obtain from (6.2.2) and (6.2.3) that

$$\gamma_i = -\frac{\langle g_{i+1} - g_i, g_{i+1}\rangle}{\langle g_{i+1} - g_i, h_i\rangle}$$

$$= \frac{\langle g_{i+1} - g_i, g_{i+1}\rangle}{\langle g_i, h_i\rangle}, \tag{6.2.4}$$

because $\langle g_{i+1}, h_i\rangle = 0$, by construction of λ_i. Now, by (6.1.13b), $h_i = g_i + \gamma_{i-1} h_{i-1}$, and $\langle g_i, h_{i-1}\rangle = 0$, by construction of λ_{i-1}. Hence, from (6.2.4), we obtain that

$$\textbf{PR} \qquad \gamma_i = \frac{\langle g_{i+1} - g_i, g_{i+1}\rangle}{|g_i|^2}. \tag{6.2.5}$$

Formula (6.2.5) is known as the *Polak-Ribiere* formula. When used in the Master Conjugate gradient algorithm, with $g_i \triangleq \nabla f(x_i)$, it defines the *Polak-Ribiere* method of conjugate gradients for solving (6.2.1).

Now, again for the quadratic case, $\langle g_i, g_{i+1}\rangle = 0$ for all i, and hence (6.2.5) becomes

$$\textbf{FR} \qquad \gamma_i = \frac{|g_{i+1}|^2}{|g_i|^2}, \tag{6.2.6}$$

which is the *Fletcher-Reeves* formula. It defines the Fletcher-Reeves method of conjugate gradients for solving (6.2.1).

We can summarize the two conjugate gradient methods for solving problem (6.2.1) in the following shorthand manner:

Polak-Ribiere Conjugate Gradient Algorithm 6.2.1

Data : $x_0 \in \mathbb{R}^n$, $h_0 = \nabla f(x_0)$.

$$\lambda_i = \arg\min_{\lambda \geq 0} f(x_i - \lambda h_i), \tag{6.2.7a}$$

$$x_{i+1} = x_i - \lambda_i h_i, \tag{6.2.7b}$$

$$\gamma_i^{PR} = \frac{\langle \nabla f(x_{i+1}) - \nabla f(x_i), \nabla f(x_{i+1})\rangle}{|\nabla f(x_i)|^2}, \tag{6.2.7c}$$

$$h_{i+1} = \nabla f(x_{i+1}) + \gamma_i^{PR} h_i. \tag{6.2.7d}$$

■

Fletcher-Reeves Conjugate Gradient Algorithm 6.2.2.

Data : $x_0 \in \mathbb{R}^n$, $h_0 = \nabla f(x_0)$.

$$\lambda_i = \arg\min_{\lambda \geq 0} f(x_i - \lambda h_i), \tag{6.2.8a}$$

$$x_{i+1} = x_i - \lambda_i h_i, \tag{6.2.8b}$$

$$\gamma_i^{FR} = |\nabla f(x_{i+1})|^2/|\nabla f(x_i)|^2 . \tag{6.2.8c}$$

$$h_{i+1} = \nabla f(x_{i+1}) + \gamma_i^{FR} h_i, \tag{6.2.8d}$$

■

Remark 6.2.1 : When applied to quadratic functions the Polak-Ribiere and Fletcher-Reeves methods produce identical sequences $\{ x_i \}_{i=0}^{n}$. However, when solving the general case of problem (6.2.1), the two methods produce different sequences. There is empirical evidence indicating that the Polak-Ribiere method is superior to the Fletcher-Reeves method. The reason for this appears to be the fact that the Polak-Ribiere method satisfies the assumptions of the Polak-Sargent-Sebastian theorem, while the Fletcher-Reeves method does not. At present we find more complex conjugate gradient methods, such as the one due to Nazareth, which build on the Polak-Ribiere method, and which are less sensitive to step length errors. ■

Theorem 6.2.1 : Suppose that $f:\mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and that there exist $0 < m \leq M < \infty$, such that for all $x, y \in \mathbb{R}^n$,

$$m|y|^2 \leq \langle y, H(x)y \rangle \leq M|y|^2 . \tag{6.2.9}$$

If the Polak-Ribiere Algorithm is applied to $\min_{x \in \mathbb{R}^n} f(x)$, producing a sequence $\{ x_i \}_{i=0}^{\infty}$, then

(a) There exists a $\rho \in (0,1)$ such that

$$\langle \nabla f(x_i), h_i \rangle \geq \rho |\nabla f(x_i)||h_i|; \tag{6.2.10}$$

(b) The sequence $\{ x_i \}_{i=0}^{\infty}$ converges to $\hat{x}$, the unique minimizer of $f(\cdot)$. ■

Proof : (a) Letting $g(x) = \nabla f(x)$, $g_i = \nabla f(x_i)$, we obtain that

$$g_{i+1} = g(x_{i+1}) = g(x_i - \lambda_i h_i)$$

$$= g_i - \lambda_i \int_0^1 H(x_i - s\lambda_i h_i)\, ds h_i . \tag{6.2.11}$$

Since $\langle g_{i+1}, h_i \rangle = 0$, by construction of λ_i, we get that

$$\lambda_i = \frac{\langle h_i, g_i \rangle}{\langle h_i, H_i h_i \rangle} = \frac{\langle g_i, g_i \rangle}{\langle h_i, H_i h_i \rangle}, \tag{6.2.12}$$

where

$$H_i = \int_0^1 H(x_i - s\lambda_i h_i) ds . \tag{6.2.13}$$

Hence

$$\gamma_i^{PR} = \frac{\langle g_{i+1} - g_i, g_{i+1} \rangle}{|g_i|^2}$$

$$= - \lambda_i \frac{\langle H_i\, h_i, g_{i+1} \rangle}{|g_i|^2}$$

$$= - \frac{\langle H_i h_i, g_{i+1} \rangle}{\langle h_i, H_i h_i \rangle} . \tag{6.2.14}$$

Therefore

$$|\gamma_i^{PR}| \leq \leq M|g_{i+1}|/m|h_i| . \tag{6.2.15}$$

Hence,

$$|h_{i+1}| \leq |g_{i+1}| + |\gamma_i^{PR}||h_i|$$

$$\leq |g_{i+1}|(1 + \frac{M}{m}) . \tag{6.2.16}$$

Finally,

$$\langle g_{i+1}, h_{i+1} \rangle = \langle g_{i+1}, g_{i+1} + \gamma_i^{PR} h_i \rangle$$

$$= |g_{i+1}|^2 . \tag{6.2.17}$$

Consequently, making use of (6.2.16), we get

$$\frac{\langle g_{i+1}, h_{i+1} \rangle}{|g_{i+1}||h_{i+1}|} = \frac{|g_{i+1}|}{|h_{i+1}|} \geq \frac{1}{1 + \frac{M}{m}} \triangleq \rho, \tag{6.2.18}$$

which completes the proof of part (a).

(b) The fact that the sequence $\{ x_i \}_{i=0}^{\infty}$ converges to the unique minimizer of $f(\cdot)$ follows from the modified Polak-Sargent-Sebastian Theorem (6.3.2). ■

The properties of the Fletcher-Reeves method can be summarized follows:

Theorem 6.2.1 : Suppose that $f: \mathbf{R}^n \to \mathbf{R}$ is twice continuously differentiable and that there exist $0 < m \leq M < \infty$, such that (6.2.9) holds. If the Fletcher-Reeves Algorithm is applied to $\min_{x \in \mathbf{R}^n} f(x)$, producing a sequence $\{ x_i \}_{i=0}^{\infty}$, then

(a) There exists no sequence $\{ t_i \}_{i=0}^{\infty}$, such that

$$(i) \quad t_i > 0, \text{ for all } i \in \mathbf{N}, \tag{6.2.19a}$$

$$(ii) \quad t_i \to 0 \text{ as } i \to \infty, \tag{6.2.19b}$$

$$(iii) \quad \sum_{i=0}^{k} t_i^2 \to \infty \text{ as } k \to \infty, \tag{6.2.19c}$$

and

$$(iv) \quad \langle \nabla f(x_i), h_i \rangle \geq t_i |\nabla f(x_i)||h_i|, \text{ for all } i \in \mathbf{N} . \tag{6.2.19d}$$

(b) The sequence $\{ x_i \}_{i=0}^{\infty}$ converges to $\hat{x}$, the unique minimizer of $f(\cdot)$. ■

Remark 6.2.2 : It is clear from the above theorem that there is a distinct possibility that in the Fletcher-Reeves method, the angle between the gradient g_i at x_i and the search direction h_i may

approach 90° as $i \to \infty$, while in the Polak-Ribiere method this angle is well bounded away from 90°. As we have already mentioned, this fact makes the Polak-Ribiere method somewhat less sensitive to numerical errors. ■

6.3. PARTIAL CONJUGATE GRADIENT METHODS

We note that even in the quadratic case, the finite convergence of the conjugate gradient methods depends on setting $h_0 = g_0$. Now, near a minimizer $\hat{x}$ which satisfies the second order sufficiency condition, stated in Theorem 3.2.3, a function $f(\cdot)$ does have a reasonably good quadratic approximation and it may be conjectured that, near the minimizer $\hat{x}$, if we reinitialize a conjugate gradient method by setting $h_i = g_i$, from time to time, we might get better performance then by constructing h_i by one of the standard formulae all the time. There are two interesting results dealing with this case (For a simplified exposition see Luenberger[1]

Theorem 6.3.1 : Suppose that (i) $f:\mathbb{R}^n \to \mathbb{R}$ is twice continuously Lipschitz differentiable and that there exist $0 < m \leq M < \infty$ such that

$$m|y|^2 \leq \langle y, H(x)y \rangle \leq M|y|^2 \quad \forall \quad x, y \in \mathbb{R}^n ,$$

and (ii) that the Polak-Ribiere (Fletcher-Rieves) conjugate gradient method is modified so that for a given $k \in \mathbb{N}$, $k \geq 1$, whenever $(i+1)/k$ is an integer, h_{i+1} is constructed according to

$$h_{i+1} = g_{i+1} \tag{6.3.1}$$

(rather than not according to (6.2.7c), (6.2.7d)), and

$$h_{i+1} = g_{i+1} + \gamma_i^{PR} h_i \tag{6.3.2}$$

otherwise.

(a) If $k = n$, then

$$\frac{|x_{i(n+1)} - x_{in}|}{|x_{in} - x_{i(n-1)}|^2} \to 0 \text{ as } i \to \infty, \tag{6.3.3}$$

where $\hat{x}$ is the global minimizer of $f(\cdot)$ i.e., $x_i \to \hat{x}$ n-step *quotient* quadratically.

(b) If $k < n$, then

$$f(x_{i+1}) - f(\hat{x}) \leq \left[\frac{b-a}{b+a}\right]^2 [f(x_i) - f(\hat{x})], \tag{6.3.4}$$

$$|x_i - \hat{x}| \leq \left[\frac{2}{m}\right]^{1/2} \left[f(x_0) - f(\hat{x})\right]^{1/2} \left[\frac{b-a}{b+a}\right]^i \tag{6.3.5}$$

where $m = a < b < M$ are such that $(n-k)$ eigenvalues of $H(\hat{x})$ are contained in the interval $[a,b]$ and the remaining k eigenvalues are larger than b. (Thus the effect of the k worst eigenvalues has been

[1] D. G. Luenberger, *Introduction to Linear and Nnlinear Programming* , Addison-Wesley, 1973.

removed.) ■

7. ONE DIMENSIONAL OPTIMIZATION

We saw that conjugate gradient methods require the solution of one dimensional minimization problems of the form

$$\min_{\lambda \geq 0} f(x + \lambda h), \tag{7.1a}$$

where $f:\mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. We can write (7.1a) as

$$\min_{\lambda \geq 0} \phi(\lambda) . \tag{7.1b}$$

Note that $\phi'(\lambda)$ is given by

$$\phi'(\lambda) = \langle \nabla f(x + \lambda h), h \rangle \tag{7.1c}$$

and that it can be quite expensive to compute if the formula (7.1c) is used. When high precision is not important, it is much cheaper to evaluate $\phi'(\lambda)$ by finite differences, i.e., by making use of a formula such as

$$\phi'(\lambda) \approx \frac{\phi(\lambda + \varepsilon) - \phi(\lambda)}{\varepsilon}, \tag{7.1d}$$

or

$$\phi'(\lambda) \approx \frac{\phi(\lambda + \varepsilon) - \phi(\lambda - \varepsilon)}{2\varepsilon} . \tag{7.1e}$$

We shall discuss two commonly used methods for solving (7.1a).

7.1. THE GOLDEN SECTION SEARCH

The golden section search method is to be used when the function $\phi:\mathbb{R} \to \mathbb{R}$ is differentiable and *unimodal*, i.e., there is a unique point $\hat{\lambda}$ such that $\phi(\hat{\lambda}) = 0$, which is also the unique global minimizer of $\phi(\cdot)$. The golden section search is based on two observations.

(a) First, see Fig. 7.1.1, suppose that we have an interval $[a,b]$ such that $\hat{\lambda}$, the global minimizer of $\phi(\cdot)$, satisfies $\hat{\lambda} \in [a,b]$ and that we have two additional points a', b' such that $a < a' < b' < b$ and either $\phi(a') \leq \min\{ \phi(a),\phi(b) \}$, or $\phi(b') \leq \min\{ \phi(a),\phi(b) \}$.

First suppose that $\phi(a') \leq \min\{ \phi(a),\phi(b) \}$.

Case 1: Suppose that $\phi(a') \leq \phi(b')$. Then it follows from the mean value theorem that there is a $\lambda_1 \in [a',b']$ such that $\phi'(\lambda_1) \geq 0$. Since $\phi(a') \leq \phi(a)$, by assumption, it follows again from the mean value theorem that there is a $\lambda_2 \in [a,a']$ such that $\phi'(\lambda_2) \leq 0$. Hence, since $\phi'(\cdot)$ is continuous, the interval $[\lambda_2,\lambda_1] \subset [a,b']$ must contain a zero of $\phi'(\cdot)$, and hence $\hat{\lambda} \in [a,b']$, must hold. Thus, we have succeded in reducing the initial *bracket* $[a,b]$, containing the global minimizer $\hat{\lambda}$, to the smaller bracket,

$[a,b']$ which also contains $\hat{\lambda}$.

Case 2: Suppose that $\phi(a') > \phi(b')$. Then it follows from the mean value theorem that there is a $\lambda_1 \in [a',b']$ such that $\phi'(\lambda_1) < 0$. Since $\phi(a') \leq \phi(b)$, by assumption, it follows that $\phi(b') \leq \phi(b)$. It follows again from the mean value theorem that there is a $\lambda_2 \in [b',b]$ such that $\phi'(\lambda_2) \leq 0$. Hence, since $\phi'(\cdot)$ is continuous, the interval $[lambda_1,\lambda_2] \subset [a',b]$ must contain a zero of $\phi'(\cdot)$, and hence $\hat{\lambda} \in [a',b]$, must hold. Thus, we have again succeded in reducing the initial *bracket* $[a,b]$, containing the global minimizer $\hat{\lambda}$, to the smaller bracket, $[a',b]$ which also contains $\hat{\lambda}$.

The situation when $\phi(a') \leq \min\{\ \phi(a), \phi(b)\ \}$ holds, leads to similar conclusions.

(b) Second, the process of reducing the bracket $[a,b]$, containing the global minimizer $\hat{\lambda}$, can be made more efficient by making use of the following observation. Suppose that we wish to construct a sequence of nested intervals $[a_i,b_i] \subset [a,b]$, $i = 0,1,2,3...$, such that either $\phi(a_{i+1}) \leq \min\{\ \phi(a_i),\phi(b_i)\ \}$, or $\phi(b_{i+1}) \leq \min\{\ \phi(a_i),\phi(b_i)\ \}$, so that each of these intervals also contains the global minimizer $\hat{\lambda}$.

In keeping with the preceeding discussion, we assume that at each stage we construct two points $a'_i < b'_i \in (a_i,b_i)$ and hence that either $[a_{i+1},b_{i+1}] = [a_i,b'_i]$ or $[a_{i+1},b_{i+1}] = [a'_i,b_i]$ holds.

If the points a'_i,b'_i are placed symmetrically, we can ensure that either $b'_i = a'_{i+1}$ or $a'_i = b'_{i+1}$ holds by requiring that for some $F \in (0,1)$,

$$l_{i+1} = Fl_i, \tag{7.1.1}$$

$$l_{i+1} - (1 - F)l_i = (1 - F)l_{i+1} \tag{7.1.2}$$

is satisfied at each stage. Eliminating l_i and l_{i+1} from (7.1.1), (7.1.2) we get

$$F^2 + F - 1 = 0\ . \tag{7.1.3}$$

Hence

$$F = \tfrac{1}{2}(5^{\frac{1}{2}} - 1) = 0.61804\ . \tag{7.1.4}$$

Because of the symmetry of placement of the points a'_i, b'_i and the choice of F, either a'_i or b'_i can be reused in the construction of $[a_{i+2},b_{i+2}]$, whereas any other scheme would involve the placement of *two* fresh points a'_{i+1}, b'_{i+1}. If we associate with each placement of an a'_i or b'_i an evaluation of the function $\phi(\cdot)$, we can obtain a comparison of the efficiency of the golden section search with any other "two point" scheme. It should be clear that no matter how we place two additional points in the interval $[a_i,b_i]$, of length l_i, the next interval will have length $l_{i+1} > \tfrac{1}{2}l_i$. Hence, a limit on the efficiency of a two point scheme, using two function evaluations, per iteration is $-\tfrac{1}{2} \ln \tfrac{1}{2} = 0.3466$, whereas the efficiency of the golden section search, which uses only one function evaluation per iteration, is $-\ln 0.61804 = 0.4812$, which is better. We now state the golden section search formally.

Golden Section Algorithm 7.1.1 :

Step 0 : Compute a bracket $[a_0,b_0]$, containing $\hat{\lambda}$, the miniminizer of $\phi(\lambda)$, and set $i = 0$.

Step 1 : Set $l_i = b_i - a_i$, and compute

$$a'_i = a_i + l_i(1 - F), \tag{7.1.5a}$$

$$b'_i = b_i - l_i(1 - F) . \tag{7.1.5b}$$

Step 2 : If $\phi(b'_i) \leq \min\{\phi(a'_i),\phi(b_i)\}$, set $a_{i+1} = a'_i$, $b_{i+1} = b_i$. Else set $a_{i+1} = a_i$, $b_{i+1} = b'_i$.

Step 3 : Set $i = i + 1$ and go to Step 1.

■

Note that because of (7.1.1), the bracket lengths shrink linearly, with constant 0.61804, i.e.,

$$l_i = F^i\, l_0 = (0.61804)^i l_0 . \tag{7.1.6}$$

Hence the precision of identification of $\hat{\lambda}$ increases very rapidly with i, e. g .,

$$l_{15} = 0.00073\; l_0, \tag{7.1.7}$$

so that if $l_0 = 1$, then, $\hat{\lambda} = (a_{15} + b_{15}) / 2 \pm 0.000365$.

A technique for obtaining an initial bracket $[a_0,b_0]$ containing the minimizer $\hat{\lambda}$, is to start with a λ_0 such that $\phi'(\lambda_0) < 0$ and to evaluate $\phi(\lambda_i)$ for $\lambda_i = \lambda_0 + i\Delta$, with $i = 0,1,2...$, until a "triangle" of values is obtained, such that $\phi(\lambda_{i-2}) > \phi(\lambda_{i-1})$ and $\phi(\lambda_{i-1}) < \phi(\lambda_i)$, so that we have $\lambda_{i-2} \leq \hat{\lambda} \leq \lambda_i$.

7.2. SUCCESIVE QUADRATIC INTERPOLATION

Again we consider the problem

$$\min_{\lambda \in \mathbf{R}} \phi(\lambda), \tag{7.2.1}$$

where $\phi:\mathbf{R} \to \mathbf{R}$.

Assumption 7.2.1 : We shall assume that (i) $\phi(\cdot)$ is continuously differentiable and unimodal, with unique local minimizer $\hat{\lambda}$, and (ii) $\phi'(\lambda) \neq 0$ for all $\lambda \neq \hat{\lambda}$. ■

Given *three distinct points* $z^1 < z^2 < z^3$ in $\mathbf{R}$, we can construct a unique quadratic polynomial $q(\lambda;z)$, in λ, parametrized by the vector $z = (z^1,z^2,z^3)$, such that $q(z^i;z) = \phi(z^i)$, $i = 1, 2, 3$.

The Lagrange *interpolation formula* defining this quadratic polynomial is as follows:

$$q(\lambda;z) = \phi(z^1)\frac{(\lambda - z^2)(\lambda - z^3)}{(z^1 - z^2)(z^1 - z^3)} + \phi(z^2)\frac{(\lambda - z^1)(\lambda - z^3)}{(z^2 - z^1)(z^2 - z^3)}$$

$$+ \phi(z^3)\frac{(\lambda - z^1)(\lambda - z^2)}{(z^3 - z^1)(z^3 - z^2)} . \tag{7.2.2a}$$

The polynomical $q(\lambda;z)$ is called an *interpolating polynomial.*

Given *two distinct points* $z^1 < z^3$, we can construct a quadratic interpolating polynomial $q(\lambda;z)$ for $\phi(\cdot)$, such that $q(z^i;z) = \phi(z^i)$, $i = 1, 3$ and $q'(z^i;z) = \phi'(z^i)$ for $i = 1$ *or* 3. For the case where $z^2 = z^1$, the Lagrange interpolation formula for this polynomial is

$$q(\lambda;z) = \phi(z^1)\frac{(\lambda - z^3)^2}{(z^1 - z^3)^2} + \phi(z^3)\frac{(\lambda - z^1)^2}{(z^1 - z^3)^2}$$

$$+ \left[\frac{\phi'(z^1)}{z^1 - z^3} - \frac{2\phi(z^1)}{(z^1 - z^3)^2}\right](\lambda - z^1)(\lambda - z^3)\,. \qquad (7.2.2b)$$

For the case where $z^2 = z^3$, the Lagrange interpolation formula for this polynomial is

$$q(\lambda;z) = \phi(z^1)\frac{(\lambda - z^3)^2}{(z^1 - z^3)^2} + \phi(z^3)\frac{(\lambda - z^1)^2}{(z^1 - z^3)^2}$$

$$+ \left[\frac{\phi'(z^3)}{z^3 - z^1} - \frac{2\phi(z^3)}{(z^3 - z^1)^2}\right](\lambda - z^1)(\lambda - z^3)\,. \qquad (7.2.2c)$$

Our application of the above interpolation formulae will be confined to the set of vectors $z \in \mathbb{R}^3$ which define an interval $[z^1,z^3]$ that contains the minimizer $\hat{\lambda}$, viz. to the set $T \subset \mathbb{R}^3$ defined by

$$T \triangleq \{z \in \mathbb{R}^3 \mid z^1 < z^2 < z^3 \text{ and } \phi(z^2) \le \min\{\phi(z^1),\phi(z^3)\}\} \cup$$

$$\{z \in \mathbb{R}^3 \mid z^1 = z^2 < z^3,\ \phi'(z^1) \le 0,\ \phi(z^3) \ge \phi(z^1)\} \cup$$

$$\{z \in \mathbb{R}^3 \mid z^1 < z^2 = z^3,\ \phi'(z^3) \ge 0,\ \phi(z^1) \ge \phi(z^3)\} \cup$$

$$\{\hat{z} \in \mathbb{R}^3 \mid z^1 = z^2 = z^3 = \hat{\lambda}\} \qquad (7.2.3)$$

Proposition 7.2.1 : (a) For every $z \in T$, $z^1 \le \hat{\lambda} \le z^3$ holds. (b) The set T defined by (7.2.3) is closed.

Proof : (a) This part follows directly from the mean value theorem.

(b) Suppose that $\{z_i\}_{i=0}^{\infty} \subset T$ is such that $z_i \to \hat{z}$ as $i \to \infty$. Since $z_i^1 \le z_i^2 \le z_i^3$ holds for all i, we must have that $\hat{z}^1 \le \hat{z}^2 \le \hat{z}^3$. We now consider the various possibilities.

(i) Suppose that $\hat{z}^1 < \hat{z}^2 < \hat{z}^3$. Then there must exist an i_0 such that $z_i^1 < z_i^2 < z_i^3$ for all $i \ge i_0$, and hence $\phi(z_i^2) \le \min\{\phi(z_i^1),\phi(z_i^3)\}$ for all $i \ge i_0$. It now follows from the continuity of $\phi(\cdot)$ that $\phi(\hat{z}^2) \le \min\{\phi(\hat{z}^1),\phi(\hat{z}^3)\}$ and hence that $\hat{z} \in T$.

(ii) Next suppose that $\hat{z}^1 < \hat{z}^2 = \hat{z}^3$. Then we need to consider two subcases:

(a) There is an infinite subsequence $\{z_i\}_{i \in K}$ such that $z_i^1 < z_i^2 = z_i^3$. In this case, we have that $\phi'(z_i^3) \ge 0$ and $\phi(z_i^1) \ge \phi(z_i^3)$ for all $i \in K$, and hence it follows from the continuity of $\phi(\cdot)$, $\phi'(\cdot)$ that $\phi'(\hat{z}^3) \ge 0$ and $\phi(\hat{z}^1) \ge \phi(\hat{z}^3)$, so that $\hat{z} \in T$.

(b) There exists an i_0 such that $z_i^1 < z_i^2 < z_i^3$ holds for all $i \geq i_0$. Hence, for all $i \geq i_0$, we must have that $\phi(z_i^2) \leq \phi(z_i^1)$ and $[\phi(z_i^2) - \phi(z_i^3)]/[z_i^3 - z_i^2] \leq 0$. Hence, in the limit, since $z_i^2 \to \hat{z}^3$ and $z_i^3 \to \hat{z}^3$ as $i \to \infty$, we get that $\phi(\hat{z}^3) \leq \phi(\hat{z}^1)$ and $\phi'(\hat{z}^3) \geq 0$, i.e., $\hat{z} \in \mathbf{T}$.

(iii) The case where $\hat{z}^1 = \hat{z}^2 < \hat{z}^3$ follows by symmetry from (ii).

(iv) Finally consider the case where $\hat{z}^1 = \hat{z}^2 = \hat{z}^3$. Since we must have for all $i \in \mathbf{N}$ that $z_i^1 \leq \hat{\lambda} \leq z_i^3$, it is clear that $\hat{z}^j = \hat{\lambda}$ for $j = 1, 2, 3$ and hence that $\hat{z} \in \mathbf{T}$.

We therefore conclude that $\mathbf{T}$ is closed. ■

Theorem 7.2.1 : For any $z \in \mathbf{T}$, consider the polynominal $q(\lambda;z)$, defined by (7.2.2a) when $z^1 < z^2 < z^3$, by (7.2.2b) when $z^1 = z^2 < z^3$ or $z^1 < z^2 = z^3$ and by $q(\lambda;z) \triangleq \phi(\hat{\lambda})$ when $z^1 = z^2 = z^3$. Then

(a) The coefficients of $q(\lambda;z)$ are continuous in z on $\mathbf{T}$,

(b) The miniminizer $\lambda^*(z)$, of $q(\lambda;z)$, is continuous in z on $\mathbf{T}$.

(c) The minimizer $\lambda^*(z)$, of $q(\lambda;z)$, satisfies $\lambda^*(z) \in [z^1, z^3]$.

Proof : (a) Let the coefficients $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ of $q(\lambda;z)$ be defined (as functions of z) by

$$q(\lambda;z) = a(z)\lambda^2 + b(z)\lambda + c(z) . \tag{7.2.4a}$$

First suppose that $z \in \mathbf{R}^3$ is such that $z^1 < z^2 < z^3$. Then the coeficients of $q(\lambda;z)$ are determined by the equation

$$\begin{bmatrix} 1 & z^1 & (z^1)^2 \\ 1 & z^2 & (z^2)^2 \\ 1 & z^3 & (z^3)^2 \end{bmatrix} \begin{bmatrix} c \\ b \\ a \end{bmatrix} = \begin{bmatrix} \phi(z^1) \\ \phi(z^2) \\ \phi(z^3) \end{bmatrix} . \tag{7.2.4b}$$

The matrix in (7.2.4b) is a Van der Monde matrix and hence nonsingular, because $z^1 < z^2 < z^3$.

Now suppose that $z \in \mathbf{R}^3$ is such that $z^1 < z^2 = z^3$. Then the coeficients of $q(\lambda;z)$ are determined by the equation

$$\begin{bmatrix} 1 & z^1 & (z^1)^2 \\ 1 & z^3 & (z^3)^2 \\ 0 & 1 & 2z^3 \end{bmatrix} \begin{bmatrix} c \\ b \\ a \end{bmatrix} = \begin{bmatrix} \phi(z^1) \\ \phi(z^3) \\ \phi'(z^3) \end{bmatrix} . \tag{7.2.4c}$$

The matrix in (7.2.4b) is nonsingular because $z^1 < z^3$. An expression similar to (7.2.4b) holds for the case where $z \in \mathbf{R}^3$ is such that $z^1 = z^2 < z^3$.

Now suppose that $\{z_i\}_{i=0}^{\infty}$ is such that $z_i \to \hat{z}$ as $i \to \infty$, with $\hat{z}^1 < \hat{z}^2 < \hat{z}^3$. Then there has to be an i_0 such that $z^1 < z^2 < z^3$ for all $i \geq i_0$ and hence the coefficients $a(x_i)$, $b(z_i)$, $c(z_i)$ are determined by (9.2.4b). Since the matrix in (7.2.4b) is nonsingular and continuous, and since the right hand side in (7.2.4b) is continuous, it follows that $a(z_i) \to a(\hat{z})$, $b(z_i) \to b(\hat{z})$ and $c(z_i) \to c(\hat{z})$ as $i \to \infty$.

Next, suppose that $\{z_i\}_{i=0}^{\infty}$ is such that $z_i \to \hat{z}$ as $i \to \infty$, with $\hat{z}^1 < \hat{z}^2 = \hat{z}^3$. Clearly, the last equation in (7.2.4b) can be replaced by the result of subtracting the last equation from the second one, i.e., by

$$(z^2 - z^3)b + (z^2 - z^3)(z^2 + z^3)a = \phi(z^2) - \phi(z^3), \tag{7.2.5a}$$

which, when $z^2 \neq z^3$ can be rewritten as

$$b + (z^2 + z^3)a = \frac{\phi(z^2) - \phi(z^3)}{z^2 - z^3} . \tag{7.2.5b}$$

Hence, when $z^2 \neq z^3$, (7.2.4b) can be replaced by

$$\begin{bmatrix} 1 & z^1 & (z^1)^2 \\ 1 & z^2 & (z^2)^2 \\ 0 & 1 & z^2 + z^3 \end{bmatrix} \begin{bmatrix} c \\ b \\ a \end{bmatrix} = \begin{bmatrix} \phi(z^1) \\ \phi(z^3) \\ [\phi(z^2) - \phi(z^3)]/[z^2 - z^3] \end{bmatrix} . \tag{7.2.6}$$

It now follows from the continuity and nonsingularity of the matrix and right hand side in (7.2.6) and from the continuity and nonsingularity of the matrix and right hand side in (7.2.4b) that $a(z_i) \to a(\hat{z})$, $b(z_i) \to b(\hat{z})$ and $c(z_i) \to c(\hat{z})$ as $i \to \infty$. Since all other cases follow in a similar manner, we conclude that the coefficients $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ of $q(\lambda;z)$ are continuous.

(b) The minimizer $\lambda^*(z)$, of $q(\lambda;z)$, is given by

$$\hat{\lambda}(z) = - b(z) / 2a(z), \tag{7.2.7}$$

and hence is continuous as long as $a(z) \neq 0$, because $a(\cdot)$, $b(\cdot)$ are continuous.

(c) Since $q(z^2;z) \leq \min\{ q(z^1;z), q(z^3;z) \}$, and $q(\cdot;z)$ is convex, it follows that $\lambda^*(z) \in [z^1, z^3]$. ■

In order to define an algorithm we need two more quantities. First we define the candidate triplets that might replace a $z \in \mathbf{T}$ and define a smaller interval containing $\hat{\lambda}$ than z, by

$$u_1(z) = (z^1, \lambda(z), z^2)^T, \tag{7.2.8a}$$

$$u_2(z) = (z^2, \lambda(z), z^3)^T, \tag{7.2.8b}$$

$$u_3(z) = (\lambda(z), z^2, z^3)^T, \tag{7.2.8c}$$

$$u_4(z) = (z^1, z^2, \lambda(z))^T . \tag{7.2.8d}$$

We now define the set of *admissible* replacement triplets by

$$\mathbf{A}(z) \triangleq \mathbf{T} \cap \{u_1(z), u_2(z), u_3(z), u_4(z)\} . \tag{7.2.8e}$$

Finally, we define the *surrogate cost* function $c:\mathbb{R}^3 \to \mathbb{R}$ by

$$c(z) \triangleq \phi(z^1) + \phi(z^2) + \phi(z^3) . \tag{7.2.8f}$$

Algorithm 7.2 : (One dimensional minimization via SQI)

Data : $z_0 \in \mathbf{T}$.

Step 0 : Set $i = 0$.

Step 1 : Compute the quadratic interpolating polynominal $q(\lambda;z_i)$.

Step 2 : Compute $\lambda(z_i) = \arg \min_{\lambda \in \mathbb{R}} q(\lambda;z_i)$. If $\lambda(z_i) = z_i^1$ or $\lambda(z_i) = z_i^3$, stop.

Note : $\lambda(z_i) \in [z_i^1, z_i^3]$.

Step 3 : Construct the vectors in $\mathbb{R}^3$:

$$u_1(z_i) = (z_i^1, \lambda(z_i), z_i^2)^T, \tag{7.2.9a}$$

$$u_2(z_i) = (z_i^2, \lambda(z_i), z_i^3)^T, \tag{7.2.9b}$$

$$u_3(z_i) = (\lambda(z_i), z_i^2, z_i^3)^T, \tag{7.2.9c}$$

$$u_4(z_i) = (z_i^1, z_i^2, \lambda(z_i))^T, \tag{7.2.9d}$$

and set

$$\mathbf{A}(z_i) = \mathbf{T} \cap \{u_1(z_i), u_2(z_i), u_3(z_i), u_4(z_i)\} . \tag{7.2.9e}$$

Step 4 : Compute

$$z_{i+1} \in \operatorname{argmin} \{c(z) \mid z \in \mathbf{A}(z_i)\} . \tag{7.2.10}$$

Step 5 : Set $i = i + i$ and go to Step 1. ■

The following concept is helpful in showing that the above algorithm will find the minimizer $\hat{\lambda}$.

Definition 7.2.1 : Let $\mathbf{A}:\mathbb{R}^n \to 2^{\mathbb{R}^n}$ be a set valued map. We say that $\mathbf{A}$ is *closed* if for every pair of sequences $\{ x_i \}$, $\{ y_i \}$ such that $x_i \to x^*$, $y_i \in \mathbf{A}(x_i)$ and $y_i \to y^*$ as $i \to \infty$, $y^* \in \mathbf{A}(x^*)$ holds. ■

Proposition 7.2.2 : (a) For every $z \in \mathbf{T}$, the set $\mathbf{A}(z)$, defined by (7.2.8e), is nonempty. (b) The set valued map $\mathbf{A}(\cdot)$ is closed.

Proof : (a) Suppose $\lambda^*(z) \in [z^1, z^2]$. Then $A(z)$ is empty if and only if $\phi(\lambda^*(z)) > \min\{ \phi(z^1), \phi(z^2) \} = \phi(z^2)$, and also $\phi(z^2) > \min\{ \phi(\lambda^*(z)), \phi(z^3) \} = \phi(\lambda^*(z))$ (because $\phi(z_2) \le \phi(z^3)$), which is clearly impossible. The case where $\lambda^*(z) \in [z^2, z^3]$ can be disposed of similarly.

(b) Suppose that $\{z_i\}_{i=0}^{\infty} \subset \mathbf{T}$ is such that $z_i \to z^*$ and $z_{i+1} \to z^{**}$ as $i \to \infty$, with $z_{i+1} \in \mathbf{A}(z_i)$ for all $i \in \mathbb{N}$. Then there must exists a $k \in \{1,2,3,4\}$ and an infinite subset $K \subset \mathbb{N}$ such that $z_{i+1} = u_k(z_i)$ for all $i \in K$. Since $u_k(\cdot)$ is continuous, $u_k(z_i) \to u_k(z^*)$ as $i \to \infty$, and since $\mathbf{T}$ is closed, $u_k(z^*) \in \mathbf{T}$. Hence $z^{**} = u_k(z^*) \in \mathbf{A}(z^*)$, which shows that $\mathbf{A}(\cdot)$ is closed. ■

Theorem 7.2.2 : Let the solution set be defined by

$$\Delta \triangleq \{ z \in \mathbf{T} \mid \phi'(z^1) = 0, \text{ or } \phi'(z^2) = 0, \text{ or } \phi'(z^3) = 0 \} . \tag{7.2.11}$$

(a) For every $z \in \mathbf{T}$ which is not in Δ, and any $y \in \mathbf{A}(z)$, $c(y) < c(z)$ holds.

(b) Let $\{z_i\}_{i=0}^{\infty}$ be a sequence constructed by Algorithm 7.2.1 for $\phi:\mathbf{R} \to \mathbf{R}$ continuously differentiable and unimodal. Then every accumulation point $\hat{z}$ of $\{ z_i \}_{i=0}^{\infty}$ is in the solution set Δ.

Proof : (a) Suppose that $z \notin \mathbf{T} \cap \Delta$. Then $\lambda^*(z) \neq z^1$ and $\lambda^*(z) \neq z^1$.

(i) If $\lambda^*(z) \in (z^1, z^2]$, then we must have that $\phi(z^2) < \phi(z^3)$. Furthermore, in this case, only $u_1(z)$ and $u_3(z)$ can be in $\mathbf{T}$, and by Proposition 7.2.2(a), at least one of them must be in $\mathbf{T}$. If $u_1(z) \in \mathbf{T}$, then, by definition, $\phi(\lambda^*(z)) \leq \min\{ \phi(z^1), \phi(z^2) \}$ and hence

$$c(u_1(z)) = \phi(z^1) + \phi(\lambda^*(z)) + \phi(z^2) \leq \phi(z^1) + \phi(z^2) + \phi(z^2)$$

$$< \phi(z^1) + \phi(z^2) + \phi(z^3) = c(z) . \tag{7.2.12}$$

Next, suppose that $u_3(z) \in T$. We will show that $\phi(\lambda^*(z)) < \phi(z^1)$. For the sake of contradiction, suppose that $\phi(\lambda^*(z)) \geq \phi(z^1)$. Then, since $\phi(z^2) \leq \phi(z^1)$, and $\lambda^*(z) \in (z^1, z^2]$, it follows that $\phi(\cdot)$ has a local maximum in $[z^1, z^2]$, which contradicts the unimodality of $\phi(\cdot)$. Hence $c(u_3(z)) < c(z)$ in this case.

(ii) If $\lambda^*(z) \in (z^2, z^3)$, then we must have that $\phi(z^2) < \phi(z^1)$. Furthermore, in this case, only $u_2(z)$ and $u_4(z)$ can be in $\mathbf{T}$. The rest of the proof follows from (i), by symmetry.

(b) By construction, $z_i^1, z_i^2, z_i^3 \in [z_0^1, z_0^3]$, for all $i \in \mathbf{N}$, so that $\{z_i\}_{i=0}^{\infty}$ is bounded and hence must have accumulation points. Since $\{c(z_i)\}_{i=0}^{\infty}$ is monotone decreasing by construction, and $c(\cdot)$ is continuous, it follows that if $\hat{z}$ is an accumulation point of $\{z_i\}_{i=0}^{\infty}$, we must have that $c(z_i) \to c(\hat{z})$, as $i \to \infty$. Suppose that $K \subset \mathbf{N}$ is such that $z_i \overset{K}{\to} \hat{z}$ as $i \to \infty$ and that $\hat{z} \notin \Delta$. Then for any $z' \in \mathbf{A}(\hat{z})$ we must have that $c(z') < c(\hat{z})$. Clearly, there exist an infinite subset $K' \subset K$ and a $z^* \in \mathbf{T}$, such that $z_{i+1} \overset{K'}{\to} z^*$. Then since $z_i \overset{K'}{\to} \hat{z}$ also holds, and since $\mathbf{A}(\cdot)$ is closed, $z^* \in \mathbf{A}(\hat{z})$ and therefore $c(z^*) - c(\hat{z}) = -\delta < 0$. Hence, by continuity of $c(\cdot)$, there must exist an i_0 such that $c(z_{i+1}) - c(z_i) \leq -\delta / 2$ for all $i \in K'$, $i \geq i_0$, which contradicts the fact that $c(z_i) \to c(\hat{z})$ as $i \to \infty$. Hence the theorem must be true. ■

8. QUASI-NEWTON METHODS

Quasi-Newton methods were invented as a second-derivative-free approximation to Newton's method. As such, they are related to secant methods, from which they differ by the formulae used to construct approximations to hessian matrices. Just like conjugate gradient methods, quasi-Newton methods must be explained in two steps. First as methods for minimizing quadratic functions and then as methods for general unconstrained optimization.

8.1. THE VARIABLE METRIC CONCEPT

Consider the problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{8.1.1a}$$

where

$$f(x) \triangleq \tfrac{1}{2}\langle x, Hx\rangle + \langle d, x\rangle, \tag{8.1.1b}$$

with H an $n \times n$ positive definite, symmetric matrix. Hence

$$\nabla f(x) = Hx + d\,. \tag{8.1.1c}$$

Given a positive definite, symmetric $n \times n$ matrix Q, the steepest descent direction *with respect to the Q norm* is defined by

$$h(x) \triangleq \arg\min_{h \in \mathbb{R}^n} \{\tfrac{1}{2}|h|_Q^2 + df(x;h)\}, \tag{8.1.2}$$

where $|h|_Q^2 \triangleq \langle h, Qh\rangle$ and $df(x;h)$ denotes the directional derivative of $f(\cdot)$. Hence the steepest descent direction *with respect to the Euclidean norm* is given by

$$\begin{aligned} h(x) &= \arg\min_{h \in \mathbb{R}^n} \{\tfrac{1}{2}|h|^2 + df(x;h)\} \\ &= \arg\min_{h \in \mathbb{R}^n} \{\tfrac{1}{2}|h|^2 + \langle Hx + d, h\rangle\}\,. \end{aligned} \tag{8.1.3a}$$

Applying the first order optimality condition, Theorem 3.1, to problem (8.1.3a), we get that

$$h(x) + Hx + d = 0, \tag{8.1.3b}$$

which implies that

$$h(x) = -[Qx + d] = -\nabla f(x), \tag{8.1.3c}$$

i.e., that $h(x)$ is the steepest descent direction that we saw in Section 4.

The steepest descent direction *with respect to the H norm* is given by

$$h(x) = \arg \min_{h \in \mathbb{R}^n} \{\tfrac{1}{2}|h|_H^2 + df(x;h)\}$$

$$= \arg \min_{h \in \mathbb{R}^n} \{\tfrac{1}{2}\langle h, Hh\rangle + \langle Hx + d, h\rangle\} . \tag{8.1.4a}$$

Applying the first order optimality condition, Theorem 3.1, to problem (8.1.4a), we get that

$$Hh(x) + Hx + d = 0 , \tag{8.1.4b}$$

from which we conclude that

$$h(x) = -[x + H^{-1}d] = -\left[\frac{\partial^2 f(x)}{\partial x^2}\right]^{-1} \nabla f(x), \tag{8.1.4c}$$

i.e., that $h(x)$ is the Newton direction.

If we update the matrix Q, defining the norm, at each iteration, we get *a variable metric* method. The idea behind the variable metric methods, for minimizing quadratic functions, as in (8.1.1b), is to keep updating the matrix Q in such a way that the matrix H^{-1} is eventually constructed. For the case of quadratic functions on $\mathbb{R}^n$, this process requires n iterations. The following observation is a key to this construction.

Let $x_0, x_1, x_2, \ldots, x_n$ be a set of distinct vectors and let $B \triangleq H^{-1}$. Then, setting

$$\begin{aligned} g_i &\triangleq \nabla f(x_i), \\ \Delta x_i &\triangleq x_{i+1} - x_i, \\ \Delta g_i &\triangleq g_{i+1} - g_i, \end{aligned} \tag{8.1.5}$$

we find that

$$B\Delta g_i = \Delta x_i, \quad i = 0,1,2,\ldots,n-1 . \tag{8.1.6}$$

Hence, if we define the $n \times n$ matrices ΔG, ΔX, by $\Delta G = [\Delta g_0, \ldots, \Delta g_{n-1}]$, $\Delta X = [\Delta x_0, \ldots, \Delta x_{n-1}]$, we obtain that

$$B\Delta G = \Delta X . \tag{8.1.7a}$$

Assuming that the matrices ΔX, ΔG are of rank n, we find that

$$H^{-1} = B = \Delta X \Delta G^{-1} . \tag{8.1.7b}$$

In solving problem (8.1.1a), with $f(\cdot)$ as in (8.1.1b), by means of a quasi-Newton method, one starts with an initial guess at a solution, x_0, and a symmetric, positive definite matrix B_0, which is an initial guess at H^{-1}, and one uses an update formula of the form

$$x_{i+1} = x_i - \lambda_i B_i g_i, \quad i = 0,1,2,\ldots, \tag{8.1.8a}$$

with

$$\lambda_i = \arg \min_{\lambda} f(x_i - \lambda B_i g_i) . \tag{8.1.8b}$$

The matrix B_{i+1} is chosen so that the following *the quasi-Newton property* (c.f. (8.1.6)) is satisfied:

$$B_{i+1}\Delta g_k = \Delta x_k, \quad \text{for } k = 0,1,\ldots,i\ . \tag{8.1.9}$$

Note that if the difference vectors $\{\Delta x_k\}_{k=0}^{n-1}$ turn out to be linearly independent, then the fact that

$$B_n \Delta G = \Delta X \tag{8.1.10}$$

holds, implies that

$$B_n = (\Delta X)\Delta G^{-1}\ . \tag{8.1.11}$$

Comparing with (8.1.7b), we conclude that $B_n = H^{-1}$. Since

$$x_{n+1} = x_n - \lambda_n B_n g_n, \tag{8.1.12}$$

and since $B_n = H^{-1}$, if $\lambda_n = 1$, then x_n is the minimizer of the quadratic function $f(\cdot)$. Hence, we must, in fact, have that $\lambda_n = 1 = \arg\min_{\lambda} f(x_n - \lambda B_n g_n)$.

We see from this that, just like a conjugate gradient method, a quasi-Newton method may take up to n iterations to solve an unconstrained optimization problem with a quadratic cost function, as compared to the one iteration required by the Newton method.

There are several methods for generating the matrices B_i, required by a quasi-Newton method, and we shall discuss these in the following two sections.

8.2. A RANK-ONE METHOD FOR GENERATING MATRICES B_i

Both rank-one and rank-two methods for generating matrices for quasi-Newton methods stem from the following technique for inverting perturbed matrices.

Proposition 8.2.1: Suppose that H is an $n \times n$ nonsingular matrix, and let $B \triangleq H^{-1}$, and let

$$H^* = H + ab^T, \tag{8.2.1a}$$

for some $a, b \in \mathbb{R}^n$. If $B^* \triangleq H^{*-1}$ exists, then it is given by

$$B^* = B - \frac{1}{1 + b^T Ba}(Ba)(b^T B)\ . \tag{8.2.1b}$$

Proof: Suppose that B^* exists. Then we can write $B^* = B + \Delta B$. Hence we must have that

$$I = B^*H^* = BH + Bab^T + \Delta B(H + ab^T)\ . \tag{8.2.1c}$$

Since $BH = I$, we conclude that

$$0 = (Ba)b^T + \Delta BH^*\ , \tag{8.2.1d}$$

which shows that $\Delta B = -(Ba)b^T B^*$, i.e., that ΔB is of the form

$$\Delta B = (Ba)c^T, \tag{8.2.1e}$$

for some $c \in \mathbb{R}^n$. Substituting for ΔB into (8.2.1d), we now obtain that

$$0 = (Ba)b^T + (Ba)c^T(H + ab^T), \tag{8.2.1f}$$

which yields that

$$c^T = -(1 + c^T a)b^T B \ . \tag{8.2.1g}$$

Hence

$$c^T a = -(1 + c^T a)b^T B a \ . \tag{8.2.1h}$$

Solving (8.2.1h) for $c^T a$ and substituting into (8.2.1g), leads to (8.2.1b), which completes our proof. ■

Exercise 8.2.1: Let $H = [h_1, h_2, \ldots h_n]$ be an $n \times n$ nonsingular matrix, with inverse B, and suppose that $H^* = [h_1, h_2, \ldots, h_{j-1}, h^*_j, h_{j+1}, \ldots, h_n]$, is an $n \times n$ nonsingular matrix which differs from H only in that it has a different j–th column. Use Proposition 8.2.1 to show that it inverse B^* is given by

$$B^* = B - \frac{1}{1 + e_j^T B h^*_j}(B(h^*_j - h_j)(e_j^T B) \ , \tag{8.2.2}$$

where e_j is the j–th column of the $n \times n$ identity matrix.

Hence show that it may be possible to invert an $n \times n$ nonsingular matrix H by means of n rank one corrections. ■

Exercise 8.2.2: Let H be an $n \times n$, symmetric, nonsingular matrix, with inverse B, and suppose that $H^* = H + aa^T + bb^T$ is also nonsingular, with inverse B^*. Show that B^* must be of the form

$$B^* = B + \alpha(Ba)(Ba)^T + \beta(Bb + \alpha(Ba)(Ba)^T b)(Bb + \alpha(Ba)(Ba)^T b)^T \ . \tag{8.2.3}$$

i.e., that the inverse B^* is given by a rank two correction of B. ■

Returning to problem (8.1.1a), (8.1.1b), and the iterative process defined by (8.1.8a), (8.1.8b), since the $n \times n$ matrix H in (8.1.1b) is symmetric, we can attempt to compute $B = H^{-1}$ in n steps, by setting $B_0 = I$, and using the update

$$B_{i+1} = B_i + \alpha_i z_i z_i^T, \ \text{for } i = 0,1,2,\ldots \tag{8.2.4a}$$

(i.e., by using a symmetric rank one update formula) with B_{i+1} required to satisfy (8.1.6) (and, hopefully, as a result also the quasi-Newton requirement (8.1.9)), i.e.:

$$B_{i+1}\Delta g_i = \Delta x_i, \ \text{for } i = 0,1,2,\ldots \ , \tag{8.2.4b}$$

where x_i, Δx_i, Δg_i are constructed according to (8.1.8a,b) and (8.1.5).

We shall now show that (8.2.2) and (8.2.3) define B_{i+1} uniquely. Let B_i be an $n \times n$, symmetric positive definite matrix. Then, because of (8.2.4b) we have that

$$\Delta x_i = B_{i+1}\Delta g_i = B_i \Delta g_i + \alpha_i z_i \langle z_i, \Delta g_i \rangle \ . \tag{8.2.5a}$$

Hence

$$\alpha_i z_i = \frac{1}{\langle z_i, \Delta g_i \rangle}[\Delta x_i - B_i \Delta g_i] \ . \tag{8.2.5b}$$

Next, from (8.2.5a)

$$\langle \Delta g_i, \Delta x_i \rangle = \langle \Delta g_i, B_i \Delta g_i \rangle + \alpha_i \langle \Delta g_i, z_i \rangle^2, \tag{8.2.5c}$$

so that

$$\alpha_i \langle \Delta g_i, z_i \rangle^2 = \langle \Delta g_i, \Delta x_i \rangle - \langle \Delta g_i, B_i \Delta g_i \rangle$$

$$= \langle \Delta g_i, \Delta x_i - B_i \Delta g_i \rangle \ . \tag{8.2.5d}$$

Substituting from (8.2.5b), (8.2.5d) into (8.2.4), we get

$$B_{i+1} = B_i + \frac{\alpha_i}{\alpha_i^2 \langle z_i, \Delta g_i \rangle^2} [\Delta x_i - B_i \Delta g_i] [\Delta x_i - B_i \Delta g_i]^T$$

$$= B_i + \frac{1}{\langle \Delta g_i, \Delta x_i - B_i \Delta g_i \rangle} [\Delta x_i - B_i \Delta g_i] [\Delta x_i - B_i \Delta g_i]^T \ . \tag{8.2.6}$$

Theorem 8.2.1 : The matrices B_i, constructed according to (8.2.6), (8.2.4), (8.1.5), and (8.1.8a, b), satisfy the quasi-Newton property (8.1.9) for $i = 0,1,\ldots,n-1$.

Proof : Clearly, for $i = 0$, we get $B_1 \Delta g_0 = \Delta x_0$, by construction of B_1. Next, proceeding by induction, we assume that (8.1.9) holds for $i = 0,1,2,\ldots,k-1$, $k \leq n-1$. Then for $i \leq k-1$,

$$B_{k+1} \Delta g_i = B_k \Delta g_i + y_k \langle \Delta x_k - B_k \Delta g_k, \Delta g_i \rangle, \tag{8.2.7}$$

with y_k defined by

$$y_k = \frac{1}{\langle \Delta g_k, \Delta x_k - B_k \Delta g_k \rangle} [\Delta x_k - B_k \Delta g_k] \ . \tag{8.2.8}$$

Since $B_k \Delta g_i = \Delta x_i$, by hypothesis, and B_k is symmetric, we obtain from (8.2.7) that

$$\langle \Delta x_k - B_k \Delta g_k, \Delta g_i \rangle = \langle \Delta x_k, \Delta g_i \rangle - \langle \Delta g_k, B_k \Delta g_i \rangle$$

$$= \langle \Delta x_k, H \Delta x_i \rangle - \langle H \Delta x_k, \Delta x_i \rangle = 0 \ . \tag{8.2.9}$$

Consequently,

$$B_{k+1} \Delta g_i = \Delta x_i, \text{ for } i = 0,1,\ldots,k-1 \ . \tag{8.2.10}$$

Since $B_{k+1} \Delta g_k = \Delta x_k$ by construction of B_{k+1}, the theorem is proved. ■

Thus, the construction of B_{i+1} defined in (8.2.6) has one desirable property. Unfortunately, it is possible for $\langle \Delta g_i, \Delta x_i - B_i \Delta g_i \rangle$ to be zero, at which point the construction breaks down. This fact has led to the development of rank-two update formulae which are more complex, but also more robust.

8.3. RANK-TWO METHODS FOR GENERATING B_i

Next we turn to rank-two update methods which overcome the shortcomings of the rank-one methods. The rank-two methods are derived from the rank-one methods as follows. If we expand (8.2.6), we get an expression of the form

$$B_{i+1} = B_i + \beta_i \Delta x_i \, \Delta x_i^T + \gamma_i (B_i \Delta g_i) \, (B_i \Delta g_i)^T + \delta_i [\, \Delta x_i \, (B_i \Delta g_i)^T + (B_i \Delta g_i) \, \Delta x_i^T \,], \tag{8.3.1}$$

where β_i, γ_i, δ_i are coefficients determined from (8.2.6). If we suppress the nonsymmetrical terms $\Delta x_i \, (B_i \Delta g_i)^T$ in (8.3.1), we get the following candidate, symmetric rank-two update formula which was invented by Davidon and popularized by Fletcher and Powell:

$$B_{i+1} = B_i + \beta_i \Delta x_i \, \Delta x_i^T + \gamma_i (B_i \Delta g_i) \, (B_i \Delta g_i)^T . \tag{8.3.2}$$

Since we need $B_{i+1} \Delta g_i = \Delta x_i$ to hold, we require that

$$\Delta x_i = B_i \Delta g_i + \beta_i \Delta x_i \, \langle \Delta x_i, \Delta g_i \rangle + \gamma_i (B_i \Delta g_i) \, \langle B_i \Delta g_i, \Delta g_i \rangle . \tag{8.3.3}$$

If we set $\beta_i = 1 \, / \, \langle \Delta x_i, \Delta g_i \rangle$ and $\gamma_i = -1 / \langle B_i \Delta g_i, \Delta g_i \rangle$, we find that $B_{i+1} \Delta g_i = \Delta x_i$ holds. With these values of β_i, γ_i, (8.3.2) becomes the Davidon-Fletcher-Powell update formula:

$$B_{i+1} = B_i + \frac{1}{\langle \Delta x_i, \Delta g_i \rangle} \Delta x_i \, \Delta x_i^T - \frac{1}{\langle B_i \Delta g_i, \Delta g_i \rangle} (B_i \Delta g_i) \, (B_i \Delta g_i)^T . \tag{8.3.4}$$

Formula (8.3.4) is by no means the only valid rank-two update formula[1] To exhibit the nonuniqueness of (8.3.4), we observe that we can always write

$$B_{i+1} = B_i + \Delta B_i, \; i = 0, 1, \ldots , \tag{8.3.5a}$$

and require that the relationship $B_{i+1} \Delta g_i = \Delta x_i$ be satisfied. Hence we require that

$$B_i \Delta g_i + \Delta B_i \Delta g_i = \Delta x_i \tag{8.3.5b}$$

hold. Rearranging (8.3.5b) we obtain that

$$\Delta B_i \Delta g_i = \Delta x_i - B_i \Delta g_i \tag{8.3.5c}$$

must hold. For any $t \geq 0$, we get from (8.3.5c) that

$$\Delta B_i \Delta g_i = t \Delta x_i + [(1 - t) \Delta x_i - B_i \Delta g_i] . \tag{8.3.5d}$$

Clearly, (8.3.5d) will hold if we set

$$\Delta B_i = \frac{t}{\langle \Delta x_i, \Delta g_i \rangle} \Delta x_i \Delta x_i^T + \frac{1}{\langle p_i, \Delta g_i \rangle} p_i \, p_i^T , \tag{8.3.6a}$$

with

$$p_i \triangleq (1 - t) \Delta x_i - B_i \Delta g_i . \tag{8.3.6b}$$

We note that setting $t = 0$ reduces (8.3.5a), (8.3.6a, b) to (8.2.6), while setting $t = 1$ produces (8.3.4).

[1] For a nice exposition of variable metric methods see D. F.Shanno, "Conditioning of Quasi-Newton Methods for Function Minimization", *Mathematics of Computation*, Vol. 24, No. 111 pp. 647-656, 1970; and J. E. Dennis Jr. and J. J. More', "Quasi-

At present it is felt that a quasi-Newton method, much superior to the ones indicated above, is obtained by using the rank-two Broyden-Fletcher-Goldfarb-Shanno (BFGS) update formula which updates estimates of H rather than of its inverse B, as follows:

$$H_{i+1} = H_i + \frac{1}{\langle \Delta g_i, \Delta x_i \rangle} \Delta g_i \, \Delta g_i^T - \frac{1}{\langle H_i \Delta x_i, \Delta x_i \rangle} (H_i \Delta x_i)(H_i \Delta x_i)^T , \tag{8.3.7}$$

with H_0 a symmetric, $n \times$ positive definite matrix. Note that the BFGS formula can be obtained from the DFP formula by interchanging B with H, and Δx_i with Δg_i. The matrices B_i, for use in a variable metric algorithm, such as the one below, can be obtained from the BFGS matrices H_i by means of formula (8.3). As we shall see at the end of this section, the variable metric algorithm, based on the BFGS formula, differs from the DFP variable metric algorithm, below, only in the manner in which B_i is computed. We will present proofs for the DFP method and leave the corresponding proofs for the BFGS method as an easy exercise for the reader.

Before exploring the properties of the Davidon-Fletcher-Powell rank-two update formula, we find it convenient to state the variable metric algorithm which is based on this formula. The algorithm, as stated below, can be used to solve the problem $\min_{x \in \mathbb{R}^n} f(x)$, where $f: \mathbb{R}^n \to \mathbb{R}$, is twice continuously differentiable. We continue to use the notation $g_i = \nabla f(x_i)$, etc.

DFP Variable Metric Algorithm 8.3.1

Data : $x_0 \in \mathbb{R}^n$, B_0, a symmetric $n \times$ positive definite matrix.

Step 0 : Set $i = 0$.

Step 1 : If $g_i = 0$ stop. Else, compute

$$\lambda_i = \arg \min_{\lambda \geq 0} f(x_i - \lambda B_i g_i) . \tag{8.3.8a}$$

Step 2 : Compute

$$x_{i+1} = x_i - \lambda_i B_i g_i , \tag{8.3.8b}$$

$$\Delta x_i = x_{i+1} - x_i , \quad \Delta g_i = g_{i+1} - g_i , \tag{8.3.8c}$$

$$B_{i+1} = B_i + \frac{1}{\langle \Delta g_i, \Delta x_i \rangle} \Delta x_i \, \Delta x_i^T - \frac{1}{\langle \Delta g_i, B_i \Delta g_i \rangle} (B_i \Delta g_i)(B_i \Delta g_i)^T . \tag{8.3.8d}$$

Step 3 : Set $i = i + 1$ and go to Step 1. ■

Theorem 8.3.1 : Suppose that $f(x) = \frac{1}{2}\langle x, Hx \rangle + \langle d, x \rangle$ with H a symmetric, positive definite $n \times n$ matrix and that for any $i \in \mathbb{N}$ that B_i is a symmetric, positive definite $n \times$ matrix. Then B_{i+1}, defined by (8.3.8d), is also positive definite.

Proof : First we note that $\langle \Delta g_i, \Delta x_i \rangle = \langle H \Delta x_i, \Delta x_i \rangle > 0$, whenever $\Delta x_i \neq 0$, and, similarly, that $\langle \Delta g_i, B_i \Delta g_i \rangle > 0$, whenever $\Delta g_i \neq 0$, since H and B_i are both positive definite by assumption. Hence, B_{i+1} is well defined by (8.3.8d). Next, for any $y \in \mathbb{R}^n$, $y \neq 0$,

$$\langle y,B_{i+1}y\rangle = \langle y,B_iy\rangle + \langle y,\Delta x_i\rangle^2 / \langle \Delta g_i,\Delta x_i\rangle - \langle y,B_i\Delta g_i\rangle^2 / \langle \Delta g_i,B_i\Delta g_i\rangle. \tag{8.3.9}$$

Let $a = B_i^{1/2}y$, $b = B_i^{1/2}\Delta g_i$, then (8.3.9) becomes

$$\langle y,B_{i+1}y\rangle = \frac{|a|^2|b|^2 - \langle a,b\rangle^2}{|b|^2} + \frac{\langle y,\Delta x_i\rangle^2}{\langle \Delta x_i,H\Delta x_i\rangle}. \tag{8.3.10}$$

By the Schwartz inequality

$$|a|^2|b|^2 - \langle a,b\rangle^2 \geq 0, \tag{8.3.11}$$

and hence $\langle y,B_{i+1}y\rangle \geq 0$. Now suppose that $\langle y,B_{i+1}y\rangle = 0$ for some $y \neq 0$. Then *both* terms in the RHS of (8.3.10) must be zero. But $|a|^2|b|^2 - \langle a,b\rangle^2 = 0$ implies that $a = \alpha b$ for some $\alpha \in \mathbb{R}$; i.e., that $y = \alpha\Delta g_i$. But then

$$\begin{aligned}\langle y,\Delta x_i\rangle^2 &= \alpha^2 \langle \Delta g_i,\Delta x_i\rangle^2 \\ &= \alpha^2 \langle H\Delta x_i,\Delta x_i\rangle^2 > 0,\end{aligned} \tag{8.3.12}$$

which contradicts our assumption that $\langle y,B_{i+1}y\rangle = 0$. Hence the theorem must be true. ■

Thus, unlike the rank-one formula, the DFP rank-two formula does not break down as the computation proceeds.

Theorem 8.3.2 : Suppose that $f(x) = \frac{1}{2}\langle x,Hx\rangle + \langle d,x\rangle$ with H a symmetric, positive definite $n \times n$ matrix. Then for x_i, B_i, constructed by Algorithm 8.3.1,

(a) the quasi-Newton property holds, i.e.,

$$B_{i+1}\Delta g_k = \Delta x_k, \quad \forall\ i \geq k \geq 0; \tag{8.3.13}$$

(b) the Δx_i are H conjugate, i.e.,

$$\langle \Delta x_i,H\Delta x_j\rangle = 0, \quad \forall\ i \neq j; \tag{8.3.14}$$

(c) x_{n+1} is the minimizer of $f(\cdot)$ over $\mathbb{R}^n$.

Proof : We shall prove (8.3.13) and (8.3.14) together, by induction. (Recall that $\Delta g_i = H\Delta x_i$ because of the form of $f(\cdot)$ and because $B_{i+1}\Delta g_i = \Delta x_i$ by construction.)

First,

$$\begin{aligned}\langle \Delta x_0,H\Delta x_1\rangle &= \langle H\Delta x_0,-\lambda_1B_1g_1\rangle \\ &= -\lambda_1 \langle H\Delta x_0,B_1g_1\rangle \\ &= -\lambda_1 \langle B_1\Delta g_0,g_1\rangle \\ &= -\lambda_1 \langle \Delta x_0,g_1\rangle = 0,\end{aligned} \tag{8.3.15}$$

because of the manner of computing λ_0 and B_1.

Next,

$$B_1\Delta g_0 = \Delta x_0, \tag{8.3.16a}$$

and

$$B_2\Delta g_1 = \Delta x_1, \tag{8.3.16b}$$

by construction. Proceeding further (with $\Delta g_0 = H\Delta x_0$),

$$B_2\Delta g_0 = B_1\Delta g_0 + \frac{1}{\langle \Delta g_1, \Delta x_1\rangle}\Delta x_1 \langle \Delta x_1, H\Delta x_0\rangle$$

$$+ \frac{1}{\langle \Delta g_1, B_1\Delta g_1\rangle}(B_1\Delta g_1)\langle B_1\Delta g_1, \Delta g_0\rangle . \tag{8.3.16c}$$

Now, by (8.3.15), $\langle \Delta x_0, H\Delta x_1\rangle = 0$, and

$$\langle B_1\Delta g_1, \Delta g_0\rangle = \langle \Delta g_1, B_1\Delta g_0\rangle$$

$$= \langle H\Delta x_1, \Delta x_0\rangle = 0 . \tag{8.3.16d}$$

Hence

$$B_2\Delta g_0 = \Delta x_0 . \tag{8.3.16e}$$

Thus we have initialized the induction process in (8.3.13), (8.3.14), with $i = 1$. Consequently, suppose that (8.3.13) holds for all $0 \le i \le l < n$, and (8.3.14) holds for all $0 \le i,j \le l < n$. For any $i \in \{0,1,\ldots,l\}$, (by adding and subtracting terms), we get that

$$g_{i+1} = g_{i+1} + H(\Delta x_{i+1} + \Delta x_{i+} + \ldots + \Delta x_l) . \tag{8.3.17}$$

Since $\langle \Delta x_i, g_{i+1}\rangle = 0$ by construction of λ_i, we get that for any $i \le l$,

$$\langle \Delta x_i, g_{i+1}\rangle = \langle \Delta x_i, g_{i+1}\rangle + \langle \Delta x_i, \sum_{j=i+1}^{l} H\Delta x_j\rangle = 0 . \tag{8.3.18}$$

(We conclude from (8.3.18) that x_{l+1} minimizes $f(x)$ on the $(l+1)$ dimensional subspace spanned by $\Delta x_0, \Delta x_1,\ldots,\Delta x_l$.) Hence, for any $i \in \{0,1,\ldots,l\}$,

$$\langle \Delta x_i, H\Delta x_{l+1}\rangle = -\lambda_{l+1}\langle \Delta x_i, HB_{l+1}g_{l+1}\rangle$$

$$= -\lambda_{l+1}\langle B_{l+1}H\Delta x_i, g_{l+1}\rangle$$

$$= -\lambda_{l+1}\langle B_{l+1}\Delta g_i, g_{l+1}\rangle$$

$$= -\lambda_{l+1}\langle \Delta x_i, g_{l+1}\rangle$$

$$= 0, \tag{8.3.19}$$

i.e., the vectors $\{\Delta x_i\}_{i=0}^{l+1}$ are H-conjugate for $0 \le i \le l+1$.

Next, for $0 \le i \le l$,

$$B_{i+2}\Delta g_i = B_{i+2}H\Delta x_i$$

$$= B_{i+1}H\Delta x_i + \frac{1}{\langle \Delta x_{i+1}, \Delta g_{i+1}\rangle}\Delta x_{i+1}\langle \Delta x_{i+1}, H\Delta x_i\rangle$$

$$+ \frac{1}{\langle \Delta g_{i+1}, B_{i+1}\Delta g_{i+1}\rangle}(B_{i+1})\Delta g_{i+1}\langle B_{i+1}\Delta g_{i+1}, H\Delta x_i\rangle$$

$$= \Delta x_i + \frac{1}{\langle \Delta g_{i+1}, B_{i+1}\Delta g_{i+1}\rangle}(B_{i+1}\Delta g_{i+1})\langle H\Delta x_{i+1}, B_{i+1}\Delta g_i\rangle$$

$$= \Delta x_i, \tag{8.3.20}$$

i.e., (8.3.13) holds for $i = l + 1$. This completes our proof of (8.3.13) and (8.3.14). Part (c) of the theorem follows from (8.3.18). Hence our proof is completed. ■

This concludes our exposition of the behavior of the DFP variable metric method on quadratic functions.

The BFGS variable metric method has the following form:

BFGS Variable Metric Algorithm 8.3.2

Data : $x_0 \in \mathbb{R}^n$, H_0, a symmetric $n \times$ positive definite matrix.

Step 0 : Set $i = 0$.

Step 1 : If $g_i = 0$ stop. Else, compute

$$\lambda_i = \arg\min_{\lambda \ge 0} f(x_i - \lambda H_i^{-1} g_i)\,. \tag{8.3.21a}$$

Step 2 : Compute

$$x_{i+1} = x_i - \lambda_i H_i^{-1} g_i, \tag{8.3.21b}$$

$$\Delta x_i = x_{i+1} - x_i, \quad \Delta g_i = g_{i+1} - g_i, \tag{8.3.21c}$$

$$H_{i+1} = H_i + \frac{1}{\langle \Delta g_i, \Delta x_i\rangle}\Delta g_i \Delta g_i^T - \frac{1}{\langle H_i\Delta x_i, \Delta x_i\rangle}(H_i\Delta x_i)(H_i\Delta x_i)^T\,, \tag{8.3.21d}$$

Step 3 : Set $i = i + 1$ and go to Step 1. ■

Exercise 8.3.1: Suppose that $f(x) = \frac{1}{2}\langle x, Hx\rangle + \langle d, x\rangle$ with H a symmetric, positive definite $n \times$ matrix and that for any $i \in \mathbb{N}$ that H_i is a symmetric, positive definite $n \times$ matrix. Show that H_{i+1}, defined by (8.3.7), is also positive definite. ■

Exercise 8.3.2 : Prove the following result:

Newton Methods, Motivation and Theory", *SIAM Review*, Vol. 19, pp.46-89, Jan. 1977.

Theorem 8.3.3 : Suppose that $f(x) = \frac{1}{2}\langle x,Hx\rangle + \langle d,x\rangle$ with H a symmetric, positive definite $n \times$ matrix. Then for x_i, H_i, constructed by Algorithm 8.3.2,

(a) the quasi-Newton property holds, i.e.,

$$H_{i+1}^{-1}\,\Delta g_k = \Delta x_k,\ \forall\ i \geq k \geq 0; \tag{8.3.22a}$$

(b) the Δx_i are H conjugate, i.e.,

$$\langle \Delta x_i, H\Delta x_j\rangle = 0,\ \forall\ i \neq j \in \{0,1,\ldots n\}; \tag{8.3.22b}$$

(c) x_{n+1} is the minimizer of $f(\cdot)$ over $\mathbb{R}^n$. ■

Exercise 8.3.3 : Show that the parametrized formula (8.3.6a) yields results similar to Theorem 8.3.3. for a range of $0 < t < 1$. *Hint: look up Shanno's paper.* ■

It was shown by Geraldine Meyer[2] that when the cost function is quadratic and the DFP method is initialized with $B_0 = I$, then the DFP method, the Polak-Ribiere method and the Fletcher-Reeves method all produce identical trajectories. A similar result for the BFGS method was shown by Larry Nazareth[3]

For *the nonquadratic cost function case*, it was shown by M.J.D. Powell[4] that the variable metric method using the DFP update formula converges globally, under the assumption that $f(\cdot)$ is twice continuously differentiable and convex. Current experience indicates that the BFGS variable metric method is far less sensitive to the precision of step size calculation than the DFP method and that it is much faster than the DFP method. In fact, it is not uncommon to use the BFGS method with the Wolfe step size formula, which is similar to the Armijo step size formula.

In practice, variable metric methods are often found to be considerably more efficient than Newton's method. They share with Newton's method the disadvantage of requiring the storage of the matrices B_i, which may be large.

---

[2] G. E. Meyer, "Properties of the Conjugate Gradient and Davidon Methods", Analytical Mechanics Associates, Inc., Wesbury, N.Y., 1967, mimeo.

[3] J. L. Nazareth, "A Relationship between the BFGS and Conjugate Gradient Algorithms and its Implications for New Algorithms", *SIAM J. Numer. Analysis.*, Vol. 16, No. 5, pp. 794-800, October 1979.

[4] See E. Polak, *Computational Methods in Optimization*, Academic Press, New York, 1971.

9. MINIMIZATION OF MAX FUNCTIONS

9.0. INTRODUCTION

Eventually, we want to be able to solve problems of the form

$$\min\{ f^0(x) \mid f^j(x) \leq 0, j \in \underline{m}; g^k(x) = 0, k \in \underline{l} \}, \tag{9.0.1a}$$

with the f^j, g^k:$\mathbf{R}^n \to \mathbf{R}$ continuously differentiable. In engineering design, the following special form of (9.0.1b) is frequently encountered:

$$\min\{ f^0(x) \mid f^j(x) \leq 0, j \in \underline{m} \}. \tag{9.0.1b}$$

Since the inequalities $f^j(x) \leq 0$ represent design requirements, the first thing we may try to do is to find a *feasible* design, i.e., a vector $\hat{x}$ such that $f^j(\hat{x}) \leq 0$ for all $j \in \underline{m}$. Note that a feasible vector must satisfy

$$\max_{j \in \underline{m}} f^j(x) \leq 0 . \tag{9.0.2}$$

It follows from (9.0.2) that a *feasible vector* can be obtained by solving the unconstrained optimization problem

$$\min_{x \in \mathbf{R}^n} \psi(x) \tag{9.0.3a}$$

with

$$\psi(x) \triangleq \max_{j \in \underline{m}} f^j(x) . \tag{9.0.3b}$$

We shall therefore devote this lecture to the solution of (9.0.3a). We shall later see that a lot of what we learn in the process carries over to the solution of problem (9.0.1b).

Exercise 9.0.1 : Consider the problem (9.0.3a) and suppose that there is an $\hat{x}$ such that $\psi(\hat{x}) < 0$. Show that if (9.0.3) is solved by a descent algorithm which produces a sequence $\{x_i\}_{i=0}^{\infty}$, then there is a finite $i_0 \in \mathbf{N}$ such that $\psi(x_{i_0}) \leq 0$, i.e., that the computation of a feasible design is a finite process. ■

Before proceeding further, let us examine the geometry of minimax problems. First, Fig. 9.0.1 shows that the graph of $\psi(\cdot)$ has corners and hence it is not differentiable everywhere. However, its directional derivatives appear to exist and should be given by the formula $d\psi(x,h) = \max_{j \in I(x)} df^j(x;h)$, where $I(x) \triangleq \{ j \in \underline{m} \mid f^j(x) = \psi(x)$. Next, we note that the level sets, L_α^ψ, of $\psi(\cdot)$, which, for any $\alpha \in \mathbf{R}$ are defined by

$$L_\alpha^\psi \triangleq \{ x \in \mathbf{R}^n \mid \psi(x) \leq \alpha \} = \{ x \in \mathbf{R}^n \mid f^j(x) \leq \alpha, j \in \underline{m} \}, \tag{9.0.4a}$$

are the intersection of the level sets, L_α^j, of the functions $f^j(\cdot)$, $j = 1,2,\ldots,m$, i.e., that

$$\mathbf{L}^{\psi}_{\alpha} = \cap_{j \in \mathbf{m}} \mathbf{L}^{f^j}_{\alpha} . \tag{9.0.4b}$$

Hence, referring to Fig. 9.0.2, we see that the boundary of $\mathbf{L}^{\psi}_{\alpha}$ has corners.

Next, we will show that the geometry of the level sets of $\psi(\cdot)$ suggests a natural extension of the method of steepest descent, discussed in Lecture 4. We begin with a geometric interpretation of the Steepest Descent Algorithm 3.3.1. Given a point x_i, this algorithm approximates the differentiable function $f(\cdot)$ by the quadratic function $q_{x_i}(\cdot)$ defined by

$$q_{x_i}(x) \triangleq f(x_i) + \langle \nabla f(x_i), (x - x_i) \rangle + \tfrac{1}{2}\|x - x_i\|^2 . \tag{9.0.5a)}$$

Note that $q_{x_i}(x_i) = f(x_i)$ and $\nabla q_{x_i}(x_i) = \nabla f(x_i)$, so that $q_{x_i}(\cdot)$ is a first order *quadratic* approximation to $f(\cdot)$ at x_i. Its smallest level set containing x_i,

$$\mathbf{L}^{q_{x_i}}_{f(x_i)} \triangleq \{ x \in \mathbf{R}^n \mid q_{x_i}(x) \leq f(x_i) \} \tag{9.0.5b}$$

is a ball which is tangent at x_i to

$$\mathbf{L}^{f}_{f(x_i)} \triangleq \{ x \in \mathbf{R}^n \mid f(x) \leq f(x_i) \}, \tag{9.0.5c}$$

which is the smallest level set of $f(\cdot)$ containing the point x_i, see Fig.9.0.3.

We can think of any minimizer $\hat{x}$ of $f(\cdot)$ as defining a "center" of $\mathbf{L}^{f}_{f(x_i)}$. The point $(x_i - \nabla f(x_i))$, which minimizes $q_{x_i}(x)$, is the center of the ball $\mathbf{L}^{q_{x_i}}_{f(x_i)}$. Since $(x_i - \nabla f(x_i))$ is a poor approximation to $\hat{x}$, the Steepest Descent Algorithm performs a line search along the line passing through x_i and $(x_i - \nabla f(x_i))$, i.e., along the direction $h_i = -gradf(x_i)$, to obtain a somewhat better approximation to $\hat{x}$, x_{i+1}.

We now return to the function $\psi(\cdot)$ in (9.0.3b). Proceeding geometrically to obtain an extension of the Steepest Descent Algorithm for solving (9.0.3a), given $x_i \in \mathbf{R}^n$, we approximate each function $f^j(\cdot)$, $j \in \mathbf{m}$, by the first order quadratic approximation

$$q^j_{x_i}(x) \triangleq f^j(x_i) + \langle \nabla f^j(x_i), (x - x_i) \rangle + \tfrac{1}{2}\|x - x_i\|^2 , \tag{9.0.6a}$$

and we approximate $\psi(\cdot)$ by the first order *convex* approximation to it:

$$\bar{\psi}_{x_i}(x) \triangleq \max_{j \in \mathbf{m}} q^j_{x_i}(x) . \tag{9.0.6b}$$

Note that $\bar{\psi}_{x_i}(x_i) = \psi(x_i)$[1]. Next, we approximate the level set $\mathbf{L}^{\psi}_{\psi(x_i)}$ by the corresponding level set of $\bar{\psi}(\cdot;x_i)$:

$$\bar{\mathbf{L}}^{\psi}_{\psi(x_i)} \triangleq \{ x \in \mathbf{R}^n \mid q^j_{x_i}(x) \leq \psi(x_i), j \in \mathbf{m} \}$$

[1] In the next section we shall show that the directional derivatives of $\psi()$ and $\bar{\psi}_x(\cdot)$ exist. It should then become obvious that for any $h \in \mathbf{R}^n$, $d\bar{\psi}_{x_i}(x_i;h) = d\psi(x_i;h)$.

$$= \bigcap_{j \in \underline{m}} L^{\phi^j_{x_i}}_{\psi(x_i)} . \tag{9.0.6c}$$

The last relationship shows that $L^{\bar{\psi}_{x_i}}_{\psi(x_i)}$, is the intersection of the balls $L^{\phi^j_{x_i}}_{\psi(x_i)}$.

To obtain an extension of the Steepest Descent Algorithm, we will think of any $\hat{x}$ which minimizes $\psi(\cdot)$ as a "center" of $L^{\psi}_{\psi(\hat{x})}$, and we will approximate it by the "center" $(x_i + h(x_i))$ of $L^{\bar{\psi}_{x_i}}_{\psi(x_i)}$ which, by analogy with the above geometric interpretation of the Steepest Descent Algorithm, is defined as the solution of the search direction finding problem

$$\min_{x \in \mathbb{R}^n} \max_{j \in \underline{m}} q^j_{x_i}(x) . \tag{9.0.7}$$

Since, again, the point $(x_i + h(x_i))$ is a poor approximation to $\hat{x}$, we will add a line search for step size calculations.

We now have to establish a a theoretical framework which will enable us to transform the above observations into a well justified algorithm.

9.1 CONTINUITY AND DIRECTIONAL DIFFERENTIABILITY OF MAX FUNCTIONS

We begin by establishing the continuity and directional differentiability of functions of the form

$$\psi(x) \triangleq \max_{j \in \underline{m}} f^j(x) \tag{9.1.1}$$

Where $f^j : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable and

$$\underline{m} \triangleq \{1,2,....,m\} . \tag{9.1.2}$$

First, we recall the following facts. Let $\{\alpha_i\}_{i \in \mathbb{N}}$ be a bounded sequence of real numbers and let S be the set of all accumulation points of $\{\alpha_i\}_{i \in \mathbb{N}}$. Then S is compact and

$$\overline{\lim}\ \alpha_i = \max \{\alpha \mid \alpha \in S\}, \tag{9.1.3a}$$

$$\underline{\lim}\ \alpha_i = \min \{\alpha \mid \alpha \in S\} . \tag{9.1.3b}$$

Furthermore, the sequence $\{\alpha_i\}_{i \in \mathbb{N}}$ converges to an $\hat{\alpha}$ if and only if $\hat{\alpha} = \overline{\lim}\ \alpha_i = \underline{\lim}\ \alpha_i$ holds.

We are now ready to establish the continuity of the function $\psi(\cdot)$.

Theorem 9.1.1 : Suppose that for $j \in \underline{m}$, the functions $f^j : \mathbb{R}^n \to \mathbb{R}$ are continuous. Then the function $\psi(x) \triangleq \max_{j \in \underline{m}} f^j(x)$ is continuous.

Proof : Let $\hat{x} \in \mathbb{R}^n$ be arbitrary and let $\{x_i\}_{i \in \mathbb{N}}$ be any sequence in $\mathbb{R}^n$ which converges to $\hat{x}$. To show that $\psi(\cdot)$ is continuous, we must show that $\psi(x_i) \to \psi(\hat{x})$ as $i \to \infty$.

First, $\psi(\hat{x}) = \hat{f}^{\hat{k}}(\hat{x})$, for some $\hat{k} \in \underline{m}$. Since $\hat{k} \in \underline{m}$, we must have

$$\psi(x_i) = \max_{j \in \underline{m}} f^j(x_i) \geq \hat{f}^{\hat{k}}(x_i), \quad \forall\, i \in \mathbf{N}\,. \tag{9.1.4a}$$

Therefore, since $\hat{f}^{\hat{k}}(\cdot)$ is continuous,

$$\underline{\lim}\ \psi(x_i) \geq \underline{\lim}\ \hat{f}^{\hat{k}}(x_i) = \lim_{i \to \infty} \hat{f}^{\hat{k}}(x_i) = \hat{f}^{\hat{k}}(\hat{x}) = \psi(\hat{x})\,. \tag{9.1.4b}$$

Next we need to show that $\overline{\lim}\psi(x_i) \leq \psi(\hat{x})$. To obtain a contradiction, suppose that

$$\overline{\lim}\ \psi(x_i) > \hat{f}^{\hat{k}}(\hat{x}) = \psi(\hat{x})\,. \tag{9.1.5a}$$

($\overline{\lim}\ \psi(x_i)$ must be finite since all the sequences $\{f^j(x_i)\}_{i \in \mathbf{N}}$ are bounded.) Now, for each $i \in \mathbf{N}$, $\psi(x_i) = f^{j_i}(x_i)$, for some $j_i \in \underline{m}$. Since $\overline{\lim}\ \psi(x_i) = \lim_{i \in K} \psi(x_i)$ for some infinite subset $K \subset \mathbf{N}$, and $\underline{m}$ is a *finite* set, there exists an infinite subset $K' \subset K$ and an index $\hat{j} \in \underline{m}$ such that $j_i = \hat{j}$ for all $i \in K'$. Hence we must have that

$$\overline{\lim}\ \psi(x_i) = \lim_{i \in K} \psi(x_i)$$

$$= \lim_{i \in K'} \hat{f}^{\hat{j}}(x_i) = \hat{f}^{\hat{j}}(\hat{x}) > \hat{f}^{\hat{k}}(xhat) = \psi(\hat{x})\,. \tag{9.1.5b}$$

But this contradicts the definition of $\hat{k}$ and hence (9.1.5a) cannot hold. Thus we must have

$$\psi(\hat{x}) \leq \underline{\lim}\psi(x_i) \leq \overline{\lim}\ \psi(x_i) \leq \psi(\hat{x}), \tag{9.1.5c}$$

and we conclude that $\psi(x_i) \to \psi(xhat)$ as $i \to \infty$, which completes our proof. ■

Exercise 9.1.1 : Suppose that for $j \in \underline{m}$, the functions $f^j:\mathbf{R}^n \to \mathbf{R}$ are locally Lipschitz continuous. Show that the function $\psi(x) \triangleq \max_{j \in \underline{m}} f^j(x)$ is locally Lipshitz continuous. ■

Next, we establish an important property of the maximizing set

$$I(x) \triangleq \{\, j \in \underline{m} \mid \psi(x) = f^j(x) \,\}\,. \tag{9.1.6}$$

Proposition 9.1.1: Suppose that for $j \in \underline{m}$, the functions $f^j:\mathbf{R}^n \to \mathbf{R}$ are continuous. Let $\hat{x} \in \mathbf{R}^n$ be arbitrary. Then there exists a $\hat{\rho} > 0$ such that $I(x) \subset I(\hat{x})$ for all $x \in \mathbf{B}(\hat{x},\hat{\rho})$.

Proof: Suppose that $j \notin I(\hat{x})$. Then $\psi(\hat{x}) - f^j(\hat{x}) = \delta^j > 0$ and, since $\psi(\cdot)$ and $f^j(\cdot)$ are both continuous, there exists a $\rho^j > 0$ such that

$$\psi(x) - f^j(x) \geq \delta^j/2 > 0, \quad \forall \quad x \in \mathbf{B}(\hat{x},\rho^j)\,. \tag{9.1.7}$$

Let $\hat{\rho} \triangleq \min\{\rho^j \mid j \notin I(\hat{x})\}$. Then (9.1.7) implies that $I^c(\hat{x}) \subset I^c(x)$ for all $x \in \mathbf{B}(\hat{x},\hat{\rho})$, where I^c denotes the complement of I in $\underline{m}$. Consequently, $I(x) \subset I(\hat{x})$ for all $x \in \mathbf{B}(\hat{x},\hat{\rho})$. ■

We shall now establish the directional differentiability of functions of the form (9.1.1), for the case where the functions $f^j:\mathbf{R}^n \to \mathbf{R}$ are continuously differentiable.

Theorem 9.1.2: Consider the function $\psi(x) = \max_{j \in \underline{m}} f^j(x)$, with $f^j: \mathbb{R}^n \to \mathbb{R}$ continuously differentiable. Then the directional derivative $d\psi(x;h)$ exists for all $x, h \in \mathbb{R}^n$ and is given by

$$d\psi(x;h) = \max_{j \in I(x)} \langle \nabla f^j(x), h \rangle . \tag{9.1.8}$$

Proof: First, since by Exercise 9.1.1, $\psi(\cdot)$ is locally Lipshitz continuous, we must have that for any $x, h \in \mathbb{R}^n$,

$$-\|h\| \leq \underline{\lim}_{t \downarrow 0} \frac{\psi(x+th) - \psi(x)}{t} \leq \overline{\lim}_{t \downarrow 0} \frac{\psi(x+th) - \psi(x)}{t} \leq \|h\| . \tag{9.1.9}$$

Since $f^j(x) \leq \psi(x)$ for all $j \in \underline{m}$, and because of Proposition 9.1.1, we must have, for sufficiently small t, that

$$\begin{aligned} \frac{\psi(x+th) - \psi(x)}{t} &= \max_{j \in \underline{m}} \frac{f^j(x+th) - \psi(x)}{t} \\ &= \max_{j \in I(x+th)} \frac{f^j(x+th) - \psi(x)}{t} \\ &\leq \max_{j \in I(x)} \frac{f^j(x+th) - f^j(x)}{t} . \end{aligned} \tag{9.1.10a}$$

Now the functions $g^j(t) \triangleq [f^j(x+th) - f^j(x)]/t$ are continuous, provided we define $g^j(0) = df^j(x;h)$, and hence the max in (9.1.10a) is also continuous. Consequently,

$$\overline{\lim}_{t \downarrow 0} \frac{\psi(x+th) - \psi(x)}{t} \leq \max_{j \in I(x)} df^j(x;h) . \tag{9.1.10b}$$

Next,

$$\begin{aligned} \frac{\psi(x+th) - \psi(x)}{t} &= \max_{j \in \underline{m}} \frac{f^j(x+th) - \psi(x)}{t} \\ &\geq \max_{j \in I(x)} \frac{f^j(x+th) - f^j(x)}{t} \end{aligned} \tag{9.1.11a}$$

because $\psi(x) = f^j(x)$ for all $j \in I(x)$ and $I(x) \subset \underline{m}$. Hence we must have (by same arguments as before) that

$$\underline{\lim}_{t \downarrow 0} \frac{\psi(x+th) - \psi(x)}{t} \geq \max_{j \in I(x)} df^j(x;h) . \tag{9.1.11b}$$

We conclude that

$$d\psi(x;h) = \lim_{t \downarrow 0} \frac{\psi(x+th) - \psi(x)}{t} = \max_{j \in I(x)} df^j(x;h) = \max_{j \in I(x)} \langle \nabla f^j(x), k \rangle , \tag{9.1.11c}$$

which completes our proof. ■

9.2. AN OPTIMALITY FUNCTION

We shall now develop two equivalent first order optimality conditions for the problem

$$\min_{x \in \mathbf{R}^n} \psi(x) \tag{9.2.1a}$$

with

$$\psi(x) = \max_{j \in \underline{m}} f^j(x), \tag{9.2.1b}$$

and the functions $f^j:\mathbf{R}^n \to \mathbf{R}$ continuously differentiable. We shall see that the simpler one of these two conditions does not lead to continuous descent directions for the function $\psi(x)$, while the more complicated one does.

Theorem 9.2.1 (Danskin): Suppose that the functions $f^j:\mathbf{R}^n \to \mathbf{R}$ in (9.2.1b) are continuously differentiable and that $\hat{x}$ is a local minimizer of $\psi(\cdot)$. Then

$$d\psi(\hat{x};h) \geq 0, \quad \forall\, h \in \mathbf{R}^n. \tag{9.2.2}$$

Proof: Suppose that there exists an $\hat{h} \in \mathbf{R}^n$ such that

$$d\psi(\hat{x};\hat{h}) < 0. \tag{9.2.3a}$$

Then, by definition of the directional derivative, there must exist a $\hat{\lambda}$ such that for all $\lambda \in (0,\hat{\lambda})$,

$$\psi(\hat{x} + \lambda\hat{h}) - \psi(\hat{x}) \leq \tfrac{1}{2}\lambda d\psi(\hat{x};\hat{h}) < 0, \tag{9.2.3b}$$

which contradicts the optimality of $\hat{x}$. ■

Corollary 9.2.1 : Suppose that the functions $f^j:\mathbf{R}^n \to \mathbf{R}$ in (9.2.1b) are convex and continuously differentiable. Then $\hat{x}$ is a global minimizer of $\psi(\cdot)$ if and only if (9.2.2) holds. ■

Exercise 9.2.1: Prove Corollary 9.2.1. ■

Proposition 9.2.1: Suppose that the functions $f^j:\mathbf{R}^n \to \mathbf{R}$ in (9.2.1b) are continuously differentiable. Then (9.2.2) holds at $\hat{x}$ if and only if

$$0 \in \partial\psi(\hat{x}), \tag{9.2.4a}$$

where

$$\partial\psi(\hat{x}) \triangleq \operatorname*{co}_{j \in I(\hat{x})} \{ \nabla f^j(\hat{x}) \}, \tag{9.2.4b}$$

with

$$I(\hat{x}) \triangleq \{ j \in \underline{m} \mid f^j(\hat{x}) = \psi(\hat{x}) \}. \tag{9.2.4c}$$

Proof: We introduce the notation

$$Nr[\partial\psi(x)] \triangleq \arg\min\{ |\xi| \mid \xi \in \partial\psi(x) \}. \tag{9.2.5}$$

($\Rightarrow$) Suppose that (9.2.2) holds, but (9.2.4a) is not true. Let $\hat{h} = -Nr[\partial\psi(\hat{x})]$. Then we must have that $\hat{h} \neq 0$ and hence, by Theorem 2.5.3, that

$$d\psi(\hat{x},\hat{h}) = \max_{\xi \in \partial\psi(\hat{x})} \langle \xi,\hat{h}\rangle \leq -|\hat{h}|^2 < 0, \tag{9.2.6a}$$

which contradicts (9.2.2).

($\Leftarrow$) Next suppose that (9.2.4a) holds, but that there is an $\hat{h} \in \mathbf{R}^n$ such that

$$d\psi(\hat{x},\hat{h}) = \max_{\xi \in \partial\psi(\hat{x})} \langle \xi,\hat{h}\rangle < 0\,. \tag{9.2.6b}$$

Then we see that the origin is strictly separated from $\partial\psi(\hat{x})$ and we have a contradiction. ■

Definition: For any $x \in \mathbf{R}^n$, the set valued map $\partial\psi(x) \triangleq \mathrm{co}_{j \in I(x)}\{\nabla f^j(x)\}$, where $I(x) \triangleq \{j \in \underline{m} \mid f^j(x) = \psi(x)\}$, is called the *generalized gradient* of $\psi(\cdot)$ at x. ■

Corollary 9.2.2: Suppose that the functions $f^j:\mathbf{R}^n \to \mathbf{R}$ in (9.2.1b) are continuously differentiable and that $\hat{x}$ is a local minimizer of $\psi(\cdot)$. Then there exists a multiplier vector $\mu \in \Sigma$, where *the unit simplex*

$$\Sigma \triangleq \{\mu \in \mathbf{R}^m \mid \mu^j \geq 0 \ \forall\, j \in \underline{m},\ \sum_{j=1}^{m} \mu^j = 1\}, \tag{9.2.7a}$$

such that

$$\sum_{j=1}^{m} \mu^j \nabla f^j(\hat{x}) = 0, \tag{9.2.7b}$$

and

$$\sum_{j=1}^{m} \mu^j [\psi(\hat{x}) - f^j(\hat{x})] = 0\,. \tag{9.2.7c}$$

Exercise 9.2.2: Prove corollary 9.2.2. ■

Remark 9.2.1: We shall refer to the multipliers μ^j in (9.2.7a), (9.2.7b) as *Danskin* multipliers. ■

We can simplify the remainder of our presentation considerably by making use of minimax theorems. The following results are among the best known[2].

Theorem 9.2.2 (von Neumann): Let $\phi:\mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ be such that $\phi(x,y)$ is convex in x and concave in y and let $\mathbf{X} \subset \mathbf{R}^n$, $\mathbf{Y} \subset \mathbf{R}^m$ be compact convex sets. Then

$$\min_{x \in \mathbf{X}} \max_{y \in \mathbf{Y}} \phi(x,y) = \max_{y \in \mathbf{Y}} \min_{x \in \mathbf{X}} \phi(x,y)\,. \tag{9.2.8a}$$

Furthermore, $\hat{x} \in \mathbf{X}$, $\hat{y} \in \mathbf{Y}$ satisfy

$$\phi(\hat{x},\hat{y}) = \min_{x \in \mathbf{X}} \max_{y \in \mathbf{Y}} \phi(x,y) \tag{9.2.8b}$$

if and only if

[2]See p. 204, C. Berge, *Topological Spaces*, The MacMillan Co., New York, 1963.

$$\phi(\hat{x},\hat{y}) = \max_{y \in Y} \min_{x \in X} \phi(x,y) . \tag{9.2.8c}$$

■

It is easy to obtain the following extension of the von Neumann Theorem for the case where either X or Y is unbounded, but $\phi(\cdot,\cdot)$ satisfies a growth condition, as follows.

Corollary 9.2.3: Let $\phi: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be such that $\phi(x,y)$ is convex in x and concave in y and let Y be a compact, convex set in $\mathbb{R}^m$. Suppose that $\phi(x,y) \to \infty$ as $|x| \to \infty$, for all $y \in Y$. Then

$$\min_{x \in \mathbb{R}^n} \max_{y \in Y} \phi(x,y) = \max_{y \in Y} \min_{x \in \mathbb{R}^n} \phi(x,y) . \tag{9.2.9a}$$

Furthermore, $\hat{x} \in X$, $\hat{y} \in Y$ satisfy

$$\phi(\hat{x},\hat{y}) = \min_{x \in X} \max_{y \in Y} \phi(x,y) \tag{9.2.9b}$$

if and only if

$$\phi(\hat{x},\hat{y}) = \max_{y \in Y} \min_{x \in X} \phi(x,y) . \tag{9.2.9c}$$

■

A result similar to Corollary 11.2.3, for the case where X compact and $Y = \mathbb{R}^m$ is obtained by assuming that $\phi(x,y) \to -\infty$ as $\|y\| \to \infty$, for all $x \in X$.

Exercise 9.2.3 : Consider problem (9.2.1a) with the assumptions stated.

(a) Suppose that $x \in \mathbb{R}^n$ is such that $0 \notin \partial\psi(x)$. Show that $h(x) = -Nr[\partial\psi(x)]$ is a descent direction for $\psi(\cdot)$ at x.

(b) Show that for any $I \subset \underline{m}$, x, $h \in \mathbb{R}^n$,

$$\max_{j \in I} \langle \nabla f^j(x), h \rangle = \max_{\mu \in \Sigma_I} \sum_{j \in I} \mu^j \langle \nabla f^j(x), h \rangle , \tag{9.2.10a}$$

where

$$\Sigma_I \triangleq \{ \mu \in \mathbb{R}^m \mid \mu^j \geq 0, \forall j \in \underline{m}, \mu^j = 0 \ \forall j \notin I, \sum_{j=1}^{m} \mu^j = 1 \} . \tag{9.2.10b}$$

(c) Use Corollary (9.2.3) and (9.2.10a) to show that

$$\min_{h \in \mathbb{R}^n} \{ \tfrac{1}{2}|h|^2 + d\psi(x;h) \} = -\tfrac{1}{2}\|Nr[\partial\psi(x)]\|^2, \tag{9.2.10c}$$

and that

$$h(x) = \arg \min_{h \in \mathbb{R}^n} \{ \tfrac{1}{2}|h|^2 + d\psi(x;h) \} = -Nr[\partial\psi(x)], \tag{9.2.10d}$$

and hence that $h(x)$ is unique. ■

Although $h(x)$, defined by (9.2.10d), is a descent direction, it is not continuous, because the active function index set $I(x)$, used in the definition of $\partial\psi(x)$, can change abruptly. Hence, when used in a steepest descent type algorithm for solving (9.2.1a), it can cause the algorithm to converge to points which go not even satisfy our first order optimality conditions. Examples of such undesirable behavior have been published[3].

We shall therefore develop an alternative optimality condition which does yield continuous descent directions. However, first we must establish an extension of Theorem 9.1.1.

Lemma 9.2.1: Suppose that $\phi:\mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is continuous and $Y \subset \mathbb{R}^m$ is compact. Then the function $\zeta:\mathbb{R}^n \to \mathbb{R}$, defined by

$$\zeta(x) \triangleq \max_{y \in Y} \phi(x,y) \tag{9.2.11}$$

is continuous.

Proof: Let $\hat{x} \in \mathbb{R}^n$ be arbitrary and let $\{x_i\}_{i \in \mathbb{N}}$ be any sequence converging to $\hat{x}$. Suppose that the $y_i \in Y$ are such that $\zeta(x_i) = \phi(x_i,y_i)$ for all $i \in \mathbb{N}$. Then, because $\phi(\cdot,\cdot)$ is continuous and Y is compact, there exists a $K \subset \mathbb{N}$ and a $y^* \in Y$ such that $y_i \overset{K}{\to} y^*$ as $i \to \infty$ and

$$\begin{aligned}\overline{\lim_{i \to \infty}} \zeta(x_i) &= \lim_{\substack{K \\ i \to \infty}} \zeta(x_i) \\ &= \lim_{\substack{K \\ i \to \infty}} \phi(x_i,y_i) = \phi(\hat{x},y^*) \leq \zeta(\hat{x}) .\end{aligned} \tag{9.2.12a}$$

Next, let $\hat{y} \in Y$ be such that $\zeta(\hat{x}) = \phi(\hat{x},\hat{y})$. Then we must have that

$$\zeta(x_i) \geq \phi(x_i,\hat{y}) \quad \forall\, i \in \mathbb{N}, \tag{9.2.12b}$$

and hence, because because $\phi(\cdot,\cdot)$ is continuous that

$$\underline{\lim}\zeta(x_i) \geq \zeta(\hat{x}) . \tag{9.2.12c}$$

Combining (9.2.12a) and (9.2.12b), we conclude that $\zeta(x_i) \to \zeta(\hat{x})$ as $i \to \infty$, i.e., that $\zeta(\cdot)$ is continuous at $\hat{x}$. Since $\hat{x}$ was arbitrary, our proof is complete. ■

We now return to problem (9.2.1a), (9.2.1b), with the assumptions stated. Normalizing the search direction finding problem (9.0.7) so that its value is always nonpositive, we define for this problem the *optimality function* $\theta:\mathbb{R}^n \to \mathbb{R}$ and the associated *search direction function* $h:\mathbb{R}^n \to \mathbb{R}^n$ by

$$\theta(x) \triangleq \min_{h \in \mathbb{R}^n} \max_{j \in \mathbb{m}} \{ f^j(x) - \psi(x) + \langle \nabla f^j(x),h\rangle + \tfrac{1}{2}|h|^2 \}, \tag{9.2.13a}$$

[3] Philip Wolfe, "On the Convergence of Gradient Methods under Constraint", *IBM J. Res. Develp.*, July, 1972, pp 407-411.

$$h(x) \triangleq \arg \min_{h \in \mathbf{R}^n} \max_{j \in \underline{m}} \{ f^j(x) - \psi(x) + \langle \nabla f^j(x), h \rangle + \tfrac{1}{2}|h|^2 \}. \tag{9.2.13b}$$

Theorem 9.2.3: Consider the functions $\theta(\cdot)$ and $h(\cdot)$ defined by (9.2.13a) and (9.2.13b). Then,

(a) For all $x \in \mathbf{R}^n$,

$$\theta(x) \leq 0; \tag{9.2.14a}$$

(b) For all $x \in \mathbf{R}^n$,

$$d\psi(x;h(x)) \leq \theta(x); \tag{9.2.14b}$$

(c) Alternative expressions for $\theta(x)$ and $h(x)$ are given by[4]

$$\theta(x) = -\min_{\mu \in \Sigma} \{ \sum_{j=1}^{m} \mu^j [f^j(x) - \psi(x)] + \tfrac{1}{2}|\sum_{j=1}^{m} \mu^j \nabla f^j(x)|^2 \}, \tag{9.2.14c}$$

where Σ was defined in (11.2.7a); and

$$h(x) = -\sum_{j=1}^{m} \mu_x^j \nabla f^j(x), \tag{9.2.14d}$$

where the μ_x is any solution of (9.2.14c).

Equivalently, $\theta(x)$ and $h(x)$ can be expressed in the form

$$\theta(x) = -\min_{\xi \in \bar{G}\psi(x)} \{ \xi^0 + \tfrac{1}{2}|\xi|^2 \}, \tag{9.2.14e}$$

$$\bar{h}(x) = (h^0(x), h(x)) = -\arg \min_{\xi \in \bar{G}\psi(x)} \{ \xi^0 + \tfrac{1}{2}|\xi|^2 \}, \tag{9.2.14f}$$

where $\bar{G}\psi(x) \subset \mathbf{R}^{m+1}$ has elements denoted by $\bar{\xi} = (\xi^0, \xi)$, with $\xi^0 \in \mathbf{R}$, $\xi \in \mathbf{R}^n$ and is defined by

$$\bar{G}\psi(x) \triangleq \operatorname{co}_{j \in \underline{m}} \left\{ \begin{bmatrix} \psi(x) - f^j(x) \\ \nabla f^j(x) \end{bmatrix} \right\}. \tag{9.2.14g}$$

(d) For any $x \in \mathbf{R}^n$, $0 \in \partial\psi(x) \Leftrightarrow \theta(x) = 0$.

(e) Both $\theta(\cdot)$ and $h(\cdot)$ are continuous.

Proof:

(a) Since setting $h = 0$ in the max part of (9.2.13a) makes this part zero, the result follows.

(b) From (9.2.13b) we have that

$$\theta(x) = \max_{j \in \underline{m}} \{ f^j(x) - \psi(x) + \langle \nabla f^j(x), h(x) \rangle + \tfrac{1}{2}|h(x)|^2 \}. \tag{9.2.15a}$$

Hence, since for every $j \in I(x)$, $f^j(x) - \psi(x) = 0$, we have that

$$d\psi(x;h(x)) = \max_{j \in I(x)} \langle \nabla f^j(x), h(x) \rangle \leq \theta(x) - \tfrac{1}{2}|h(x)|^2 \leq \theta(x). \tag{9.2.15b}$$

(c) Next, making use of (9.2.10a), we obtain that

[4] The form (9.2.14c) is also suggested by (9.2.7b), (9.2.7c).

$$\theta(x) = \min_{h \in \mathbb{R}^n} \max_{\mu \in \Sigma} \{ \sum_{j=1}^{m} \mu^j [f^j(x) - \psi(x)] + \sum_{j=1}^{m} \mu^j \langle \nabla f^j(x), h \rangle + \tfrac{1}{2}|h|^2 \}. \tag{9.2.15c}$$

Applying Corollary 9.2.3 to (9.2.15c), we get that

$$\theta(x) = \max_{\mu \in \Sigma} \min_{h \in \mathbb{R}^n} \{ \sum_{j=1}^{m} \mu^j [f^j(x) - \psi(x)] + \sum_{j=1}^{m} \mu^j \langle \nabla f^j(x), h \rangle + \tfrac{1}{2}|h|^2 \}. \tag{9.2.15d}$$

Solving the unconstrained min for h in terms of μ, we obtain that

$$h = -\sum_{j=1}^{m} \mu^j \nabla f^j(x) . \tag{9.2.15e}$$

Substituting back into (9.2.15d), we now obtain (9.2.14c).

The expression (9.2.14e) follows from (9.2.14c) by inspection, while (9.2.14d) follows from (9.2.15e) and Corollary 9.2.2.

(d) Since $\overline{\xi} = (\xi^0, \xi) \in \overline{G}\psi(x)$ implies that $\xi^0 \geq 0$, it follows that $0 \in \partial\psi(x) \Leftrightarrow 0 \in \overline{G}\psi(x)$ and also, from (9.2.14e), that $\theta(x) = 0 \Leftrightarrow 0 \in \overline{G}\psi(x)$. Hence (d) is proved.

(e) The continuity of $\theta(\cdot)$ follows from Lemma 9.2.1 and the form (9.2.14c). To establish the continuity of $h(\cdot)$ we make use of the form (9.2.13a), which we rewrite as

$$\theta(x) = \min_{h \in \mathbb{R}^n} \phi(x,h), \tag{9.2.15f}$$

with

$$\phi(x,h) \triangleq \max_{j \in m} \{ f^j(x) - \psi(x) + \langle \nabla f^j(x), h \rangle + \tfrac{1}{2}|h|^2 \} . \tag{9.2.15g}$$

Next, it follows from Theorem 9.1.1 that $\phi(\cdot,\cdot)$ is continuous. Now suppose that $\hat{x} \in \mathbb{R}^n$ is arbitrary and that $\{x_i\}_{i \in \mathbb{N}}$ is any sequence converging to $\hat{x}$. Then $\theta(x_i) = \phi(x_i, h(x_i))$ holds for all $i \in \mathbb{N}$. Furthermore, because the $\nabla f^j(\cdot)$ are continuous, and because of (9.2.14d), it follows that the sequence $\{h(x_i)\}_{i \in \mathbb{N}}$ is bounded, and hence that it must have at least one accumulation point $\hat{h}$. Thus, suppose that $K \subset \mathbb{N}$ is such that $h(x_i) \to \hat{h}$ as $i \to \infty$. Then, because $\theta(\cdot)$ and $\phi(\cdot,\cdot)$ are continuous, it follows that $\theta(x_i) = \phi(x_i, h(x_i)) \overset{K}{\to} \phi(\hat{x},\hat{h}) = \theta(\hat{x}) = \phi(\hat{x}, h(\hat{x}))$. Since there is only one vector $h(\hat{x})$ such that $\theta(\hat{x}) = \phi(\hat{x}, h(\hat{x}))$, it follows that $\hat{h} = h(\hat{x})$, i.e., that the sequence $\{h(x_i)\}_{i \in \mathbb{N}}$ has only one accumulation point, $h(\hat{x})$ and hence that it converges to it. Consequently, since x was arbitrary, we coonclude that $h(\cdot)$ is continuous, which completes our proof. ■

9.3. UNCONSTRAINED MINIMAX ALGORITHMS

Next we shall describe two algorithms for solving the unconstrained minimax problem

$$\min_{x \in \mathbb{R}^n} \psi(x) \tag{9.3.1a}$$

with

$$\psi(x) = \max_{j \in \mathbb{m}} f^j(x), \tag{9.3.1b}$$

and the functions $f^j: \mathbb{R}^n \to \mathbb{R}$ continuously differentiable. In the form stated, these algorithms were first proposed by Pshenichnyi[5], as his method of linearizations. They can also be traced to the Pironneau-Polak[6] method of feasible directions, which, in turn, is evolved from the Huard[7] method of centers. The form of these algorithms is based on those of the steepest descent algorithm and the Armijo gradient algorithm.

Algorithm 9.3.1 : *(Exact Line Search).*

Data : $x_0 \in \mathbb{R}^n$

Step 0 : Set $i = 0$.

Step 1 : Compute the *search direction*

$$h_i = h(x_i) \triangleq \arg \min_{h \in \mathbb{R}^n} \max_{j \in \mathbb{m}} \{ f^j(x_i) - \psi(x_i) + \langle \nabla f^j(x_i), h \rangle + \tfrac{1}{2}|h|^2 \}. \tag{9.3.2a}$$

Step 2 : Compute the *step size*

$$\lambda_i \in \arg \min_{\lambda \geq 0} \psi(x_i + \lambda h_i) . \tag{9.3.2b}$$

Step 3 : *Update:* set

$$x_{i+1} = x_i + \lambda_i h_i, \tag{9.3.2c}$$

replace i by $i + 1$ and go to Step 1. ■

In view of the continuity of the search direction $h(\cdot)$, established in the preceeding section, the following result is hardly surprising since it can be established by mimicking the proof of convergence of the steepest descent algorithm for differentiable unconstrained optimization.

Theorem 9.3.1: Consider problem (9.3.1a) with the assumptions stated. If in solving problem (9.3.1a), Algorithm 9.3.1 constructs a sequence $\{x_i\}_{i=0}^{\infty}$, then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies the first order optimality condition $\theta(\hat{x}) = 0$. ■

Exercise 9.3.1: Prove Theorem 9.3.1. ■

It is also possible to propose a minimax algorithm which uses an Armijo type step size rule, as follows.

Algorithm 9.3.2: *(Armijo Line Search).*

Parameters : $\alpha, \beta \in (0,1)$.

---

[5]B. N. Pshenichnyi and Yu. M. Danilin *Numerical Methods in Extremal Problems*, Moscow, Mir Publishers, 1978.

[6]O. Pironneau and E. Polak, "On the Rate of Convergence of Certain Methods of Centers", *Mathematical Programming*, Vol. 2, No. 2, pp. 230-258, 1972.

[7]P. Huard, "Programmation Mathematique Convex", *Rev. Franc. Inform. Recher. Operationelle*, Vol. 7, pp. 43-59, 1968.

Data : $x_0 \in \mathbb{R}^n$

Step 0 : Set $i = 0$.

Step 1 : Compute the *search direction*

$$h_i = h(x_i) \triangleq \arg \min_{h \in \mathbb{R}^n} \max_{j \in \underline{m}} \{ f^j(x_i) - \psi(x_i) + \langle \nabla f^j(x_i), h \rangle + \tfrac{1}{2}|h|^2 \}. \tag{9.3.3a}$$

Step 2 : Compute the *step size*

$$\lambda_i = \arg \max_{k \in \mathbf{N}} \{ \beta^k \mid \psi(x_i + \beta^k h_i) - \psi(x_i) - \beta^k \alpha \theta(x_i) \leq 0 \}. \tag{9.3.3b}$$

Step 3 : *Update:* set

$$x_{i+1} = x_i + \lambda_i h_i, \tag{9.3.3c}$$

replace i by $i + 1$ and go to Step 1. ■

The convergence properties of Algorithm 9.3.2 can be established by making use of the fact that $h(\cdot)$ is continuous and mimicking the proof of convergence for the Armijo gradient method. The result of such an exercise is the following:

Theorem 9.3.2 : Consider problem (9.3.1a) with the assumptions stated. Suppose that Algorithm 9.3.2 constructs a sequence $\{x_i\}_{i=0}^{\infty}$. Then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ satisfies the first order optimality condition $\theta(\hat{x}) = 0$. ■

Exercise 9.3.2: Prove Theorem 9.3.2. ■

To conclude this lecture, we must discuss methods for computing the search direction $h(x)$ and evaluating the optimality function $\theta(x)$. First, observe that (9.2.14c) is a standard quadratic program and hence our first instinct would be to try to solve it by commercially available code and then obtain the search direction from (9.2.14d). Unfortunately, the matrix Q, defined by

$$Q \triangleq \sum_{j=1}^{m} \nabla f^j(x) \nabla f^j(x)^T, \tag{9.3.4}$$

which apppears implicitly in (9.2.14c), is often only positive semidefinite. As a result, standard quadratic programming codes fail to solve (9.2.14c) from time to time. Because of this, it is preferable to use a "child" (such as the Wolfe[8] or van Hohenbalken[9] algorithms) of the Gilbert algorithm[10] which we will now describe. This algorithm solves problem (9.2.14e). The geometry of this algorithm is illustrated in Fig. 9.2.1. The data for this algorithm consists of the vectors $\xi_j = (\xi_j^0, \xi_j) \in \mathbb{R}^{n+1}$, $j \in \underline{m}$, with $\xi_j^0 = \psi(x_i) - f^j(x_i)$ and $\xi_j = \nabla f^j(x_i)$. In the form stated, the algorithm solves the problem

---

[8]Wolfe, Ph., "Finding the Nearest Point in a Polytope", *Math. Programming*, Vol. 11, pp 128-149, 1976.

[9]B. von-Hohenbalken, "A Finite Algorithm to Maximize Certain Pseudoconcave Functions on Polytopes", *Mathematical Programming*, Vol. 9, pp. 189-206, 1975.

[10] E. G. Gilbert, "An iterative Procedure for Computing the Minimum of a Quadratic Form on a Convex Set", *SIAM J. Control*, Vol. 4, No. 1, 1966, pp 61-79.

$$\min\{\, q(\bar{\xi}) \mid \bar{\xi} \in \operatorname*{co}_{j \in \underline{m}} \{\, \bar{\xi}_j \,\} \,\}, \tag{9.3.5a}$$

where $q:\mathbf{R}^{n+1} \to \mathbf{R}$ is defined by

$$q(\bar{\xi}) \triangleq \xi^0 + \tfrac{1}{2}\|\xi\|^2 . \tag{9.3.5b}$$

Search Direction Algorithm 9.3.3 (E. Gilbert) :

Data : $\{\bar{\xi}_j\}_{j=1}^{m} \subset \mathbf{R}^{n+1}$.

Step 0 : Set $i = 0$, set $\bar{s}_0 = \bar{\xi}_1$.

Step 1 : Compute a $j_i \in \underline{m}$ such that

$$\langle \nabla q(\bar{s}_i), \bar{\xi}_{j_i} \rangle = \min_{j \in \underline{m}} \langle \nabla q(\bar{s}_i), \bar{\xi}_j \rangle . \tag{9.3.6a}$$

Step 2 : Set

$$\bar{\xi}(\lambda) \triangleq \lambda \bar{s}_i + (1 - \lambda)\bar{\xi}_{j_i}, \tag{9.3.6b}$$

and compute

$$\lambda_i = \arg \min_{\lambda \in [0,1]} \{\, \xi^0(\lambda) + \tfrac{1}{2}\|\xi(\lambda)\|^2 \,\}. \tag{9.3.6c}$$

Step 3 : Set $\bar{s}_{i+1} = \bar{\xi}(\lambda_i)$, replace i by $i+1$ and go to Step 1. ■

Remark 9.3.1: Note that the computation of λ_i in (9.3.6c) is very simple because either $\lambda_i \in (0,1)$, in which case its value is obtained from

$$\frac{d}{dt}\left[\, \xi^0(\lambda) + \tfrac{1}{2}\|\xi(\lambda)\|^2 \,\right] = 0, \tag{9.3.7}$$

or $\lambda_i \in \{\, 0,1 \,\}$. Thus λ_i can be computed in at most three evaluations of a simple function. ■

Remark 9.3.2 : In practice, the construction in Algorithm 9.3.3 must be stopped at some point. Since (see Fig. 9.3.2) we always have an over-estimate of $-\theta \triangleq \min\{\, \xi^0 + \|\xi\|^2 \mid \bar{\xi} \in \operatorname*{co}_{j \in \underline{m}} \{\, \bar{\xi}_j \,\} \,\}$ in $-\bar{\theta}_i \triangleq s_i^0 + \tfrac{1}{2}\|s_i\|^2$ and an easily computable under-estimate in $-\underline{\theta}_i \triangleq \min\{\, \xi^0 + \tfrac{1}{2}\|\xi\|^2 \mid \langle \nabla q(\bar{s}_i), \bar{\xi} - \bar{\xi}_{j_i} \rangle = 0 \,\}$, we propose to stop computation in Algorithm 9.3.3 when $[\bar{\theta}_i - \underline{\theta}_i] / \underline{\theta}_i \le \delta$, where $\delta > 0$ is a preassigned tolerance. Since $\underline{\theta}$ must approach zero as a solution of problem (9.3.1a) is approached, we see that this test automatically increases the precision of our evaluation of $\theta(x_i)$ as x_i approaches a solution $\hat{x}$.

A proof of convergence for Algorithm 9.3.1 using approximate evaluations of the search direction can be produced, but it is beyond the scope of this course. ■

Theorem 9.3.2: The Sequence $\{\bar{s}_i\}_{i=0}^{\infty}$, constructed by Algorithm 9.3.3, converges to the unique solution $\bar{s}^*$ of the problem (9.3.5). ■

Exercise 9.3.3: Prove Theorem 9.3.2.

[Hint: write Algorithm 9.3.3 in the form $\bar{s}_{i+1} \in A(\bar{s}_i)$ and show that $A(\cdot)$ is a closed map.] ■

Corollary 9.3.1: When the vectors $\{\bar{\xi}_j\}_{j=1}^m$ are such that $\bar{\xi}_j = (\xi_j^0, \xi_j) \in \mathbb{R}^{n+1}$, with $\xi_j^0 = \psi(x_i) - f^j(x_i)$ and $\xi_j = \nabla f^j(x_i)$, then the solution point $\bar{s}^*$ of (9.3.5) satisfies $\theta(x_i) = -(s^{*0} + \frac{1}{2}|s^*|^2)$, and $h(x_i) = -s^*$. ■

9.4. RATE OF CONVERGENCE OF MINIMAX ALGORITHM 9.4.1

we will now show that the rate of convergence of the Minimax Algorithm 9.3.1 is similar to that of the of the Steepest Descent Algorithm 3.3.1. We will need an assumption which generalizes (4.3.1) that was used in establishing the rate of convergence of the Steepest Descent Algorithm 3.3.1, i.e.,

Assumption 9.4.1 : In problem (9.0.3a), each $f^j : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, and there exist $0 < c \leq 1 \leq C < \infty$, such that

$$c|y|^2 \leq \langle y, \frac{\partial^2 f^j(x)}{\partial x^2} y \rangle \leq C|y|^2 , \quad \forall\, j \in \underline{m},\ x,y \in \mathbb{R}^n . \tag{9.4.1}$$

■

We note that under Assumption 9.4.1, the function $\psi(x) \triangleq \max_{j \in \underline{m}} f^j(x)$ is *strictly convex* and hence it has a unique minimizer $\hat{x}$. For any $x \in \mathbb{R}^n$, Assumption 9.4.1 enables us to get a useful estimate of the quantity $\psi(\hat{x}) - \psi(x)$, as we shall now see.

Lemma 9.4.1: For any $x,x' \in \mathbb{R}^n$ and any $\mu \in \Sigma \triangleq \{ \mu \in \mathbb{R}^m \mid \sum_{j=1}^m \mu^j = 1,\ \mu^j \geq 0\ \forall\, j \in \underline{m} \}$,

$$\psi(x') - \psi(x) \geq \sum_{j \in \underline{m}} \mu^j \{ f^j(x) - \psi(x) + \langle \nabla f^j(x), x' - x \rangle + \frac{1}{2} c|x' - x|^2 \} . \tag{9.4.2}$$

Proof : First, note that

$$\psi(x') - \psi(x) = \max_{j \in \underline{m}} \{ f^j(x') - \psi(x) \}$$

$$= \max_{\mu \in \Sigma} \{ \sum_{j \in \underline{m}} \mu^j f^j(x') - \psi(x) \}. \tag{9.4.3}$$

Next, making use of the second order Taylor expansion (2.4.6b) and (9.4.1), we obtain that for any $j \in \underline{m}$,

$$f^j(x') = f^j(x) + \langle \nabla f^j(x), (x' - x) \rangle + \int_0^1 (1 - s) \langle (x' - x), \frac{\partial^2 f^j(x + s(x' - x)}{\partial x^2} (x' - x) \rangle ds$$

$$\geq f^j(x) + \langle \nabla f^j(x), (x' - x) \rangle + \frac{1}{2} c|x' - x|^2 . \tag{9.4.4}$$

Hence, from (9.4.3) and (9.4.4) we obtain that

$$\psi(x') - \psi(x) \geq \max_{\mu \in \Sigma} \{ \sum_{j \in \underline{m}} \mu^j \{ f^j(x) - \psi(x) + \langle \nabla f^j(x), x' - x \rangle + \frac{1}{2} c|x' - x|^2 \} .$$

$$\geq \sum_{j \in \mathbf{m}} \mu^j \{ f^j(x) - \psi(x) + \langle \nabla f^j(x), x' - x \rangle + \frac{1}{2} m |x' - x|^2 \} , \tag{9.4.5}$$

for any $\mu \in \Sigma$, which completes our proof. ■

Theorem 9.4.1 (Linear Convergence) : Suppose that Assumption 9.4.1 is satisfied. If the Minimax Algorithm 9.3.1 constructs a sequence $\{x_i\}_{i=0}^{\infty}$, then,

(a) $x_i \to \hat{x}$ as $i \to \infty$, and

(b)

$$[\psi(x_{i+1}) - \psi(\hat{x})] \leq \delta[\psi(x_i) - \psi(\hat{x})] , \ \forall \ i \in \mathbf{N}_+ , \tag{9.4.6a}$$

where

$$\delta \triangleq 1 - \frac{c}{C} < 1 . \tag{9.4.6b}$$

(c) There exists a constant $K < \infty$ such that

$$|x_i - \hat{x}| \leq K(\delta^{1/2})^i \ \forall \ i \in \mathbf{N}_+ . \tag{9.4.6c}$$

Proof:

(a) Because $\psi(\cdot)$ is strictly convex under Assumption 9.4.1, the level set $L \triangleq \{ x \in \mathbf{R}^n \mid \psi(x) \leq \psi(x_0) \}$ is compact. Hence the sequnce $\{x_i\}_{i=0}^{\infty}$ must have accumulation points $\hat{x}$, all of which satisfy $0 \in \partial\psi(\hat{x})$. Since $\psi(\cdot)$ is strictly convex, it follows that $\hat{x} = \text{argmin}_{x \in \mathbf{R}^n} \psi(x)$ and is unique. Therefore $x_i \to \hat{x}$ as $i \to \infty$.

(b) (i) First we obtain a bound on the decrease in $\psi(x)$ at iteration i. For all $\lambda \in [0,1]$,

$$\psi(x_i + \lambda h_i) - \psi(x_i) = \max_{j \in \mathbf{m}} f^j(x_i + \lambda h_i) - \psi(x_i)$$

$$\leq \max_{j \in \mathbf{m}} f^j(x_i) + \langle \nabla f^j(x_i), \lambda h_i \rangle - \psi(x_i) + \tfrac{1}{2} C \lambda^2 |h_i|^2$$

$$\leq \lambda \left[\max_{j \in \mathbf{m}} f^j(x_i) + \langle \nabla f^j(x_i), h_i \rangle - \psi(x_i) + \tfrac{1}{2} C \lambda |h_i|^2 \right] , \tag{9.4.7}$$

because $\lambda \in [0,1]$ and $f^j(x_i) \leq \psi(x_i)$. Therefore, if $\lambda \leq 1/C$,

$$\psi(x_i + \lambda h_i) - \psi(x_i) \leq \lambda \left[\max_{j \in \mathbf{m}} f^j(x_i) + \langle \nabla f^j(x_i), h_i \rangle - \psi(x_i) + \tfrac{1}{2} |h_i|^2 \right]$$

$$= \lambda \theta(x_i) < 0 . \tag{9.4.8}$$

Thus

$$\psi(x_{i+1}) - \psi(x_i) \leq \frac{1}{C}\theta(x_i) . \tag{9.4.9}$$

(ii) Next we relate $\theta(x_i)$, defined in (2.8a), to $\psi(\hat{x}) - \psi(x_i)$. For any $\mu_i \in \mu(x_i)$,

$$\theta(x_i) = \min_{h \in R^n} \sum_{j \in m} \mu_i^j [f^j(x_i) + \langle \nabla f^j(x_i), h\rangle - \psi(x_i) + \tfrac{1}{2}|h|^2] . \tag{9.4.10}$$

Replacing h by $c(\hat{x} - x_i)$ in (9.4.10), we obtain that

$$\theta(x_i) \leq \sum_{j \in m} \mu_i^j [f^j(x_i) + \langle \nabla f^j(x_i), c(\hat{x} - x_i)\rangle - \psi(x_i)] + \tfrac{1}{2}|c(\hat{x} - x_i)|^2$$

$$\leq c\{ \sum_{j \in m} \mu_i^j [f^j(x_i) + \langle \nabla f^j(x_i), (\hat{x} - x_i)\rangle - \psi(x_i)] + \frac{1}{2}c|(\hat{x} - x_i)|^2\} . \tag{9.4.11}$$

Making use of Lemma 9.4.1, we obtain that

$$\theta(x_i) \leq c[\psi(\hat{x}) - \psi(x_i)] . \tag{9.4.12}$$

Combining (9.4.12) with (9.4.9) yields

$$\psi(x_{i+1}) - \psi(x_i) \leq \frac{c}{C} [\psi(\hat{x}) - \psi(x_i)] . \tag{9.4.13}$$

Relation (9.4.6a) now follows directly.

(c) First, it follows from (9.4.6a) that

$$\psi(x_i) - \psi(\hat{x}) \leq [\psi(x_0) - \psi(\hat{x})]\delta^i . \tag{9.4.14}$$

Setting $x'=x_i$, $x=\hat{x}$, and $\mu=\hat{\mu} \in \Sigma$, an optimal multiplier at $\hat{x}$, we obtain from (9.4.2) that

$$|x_i - \hat{x}| \leq \{\frac{2}{c}[\psi(x_i) - \psi(\hat{x})]\}^{1/2} . \tag{9.4.15}$$

The relation (9.4.6c) now follows directly from (9.4.14). ■

We are now ready to consider inequality constrained optimization problems.

10. CONSTRAINED OPTIMIZATION : INEQUALITY CONSTRAINTS

We now turn to constrained optimization problems with inequality constraints only. We shall develop first order optimality conditions both in classical multiplier form and in optimality function form; then we shall present an algorithm. Second order conditions will be dealt with separately later.

10.1. FIRST ORDER OPTIMALITY CONDITIONS

Consider the inequality constrained optimization problem

$$P_I \qquad \min\{ f^0(x) \mid f^j(x) \leq 0, j \in \underline{m} \}, \tag{10.1.1}$$

where the functions $f^j:\mathbb{R}^n \to \mathbb{R}$ are continuously differentiable.

Definition 10.1.1 : We say that $\hat{x}$ is a *local minimizer* of P_I, if there exists a $\hat{\rho} > 0$ such that $f^0(\hat{x}) \leq f^0(x)$ for all $x \in \mathbf{B}(\hat{x},\rho) \cap \{ x \mid f^j(x) \leq 0, j \in \underline{m} \}$. ■

Theorem 10.1.1 (F. John) : Suppose that $\hat{x}$ is a local miniminizer for P_I (with associated radius $\hat{\rho} \geq 0$). Then there exist multipliers $\mu^0 \geq 0$, $\mu^1 \geq 0$,.....,$\mu^m \geq 0$, not all zero (alternatively, such that $\sum_{j=0}^{m} \mu^j = 1$), such that

$$\sum_{j=0}^{m} \mu^j \nabla f^j(\hat{x}) = 0, \tag{10.1.2a}$$

and

$$\mu^j f^j(\hat{x}) = 0 \ \forall \ j \in \underline{m} . \tag{10.1.2b}$$

Proof : Consider the function

$$F_{\hat{x}}(x) \triangleq \max\{ f^0(x) - f^0(\hat{x}); f^j(x), j \in \underline{m} \}. \tag{10.1.3}$$

Then $F_{\hat{x}}(\hat{x}) = 0$ and, for all $x \in \mathbf{B}(\hat{x},\hat{\rho})$, $F_{\hat{x}}(x) \geq 0$, either because $f^0(x) - f^0(\hat{x}) \geq 0$, or because $f^j(x) \geq 0$ for some $j \in \underline{m}$. Hence $\hat{x}$ is also a local miniminizer of $F_{\hat{x}}(\cdot)$. Consequently, by Proposition 10.2.1, we must have that

$$0 \in \partial F_{\hat{x}}(\hat{x}) . \tag{10.1.4}$$

Since

$$\partial F_{\hat{x}}(\hat{x}) = \operatorname*{co}_{j \in J(\hat{x})} \{ \nabla f^j(\hat{x}) \}, \tag{10.1.5a}$$

where

$$J(\hat{x}) \triangleq \{0\} \cup \{ j \in \underline{m} \mid f^j(\hat{x}) = 0 \}, \tag{10.1.5b}$$

it follows from Corollary 10.2.2, that there exists a mutiplier vector $\Pi \in \bar{\Sigma}$, where

$$\bar{\Sigma} \triangleq \{ (\mu^0,\mu) \in \mathbb{R}^{m+1} \mid \mu^j \geq 0 \text{ for } j = 0,1,\cdots,m, \ \sum_{j=0}^{m} \mu^j = 1 \}, \tag{10.1.6}$$

such that

$$\sum_{j=0}^{m} \mu^j \nabla \tilde{f}^j(\hat{x}) = 0, \tag{10.1.7a}$$

$$\sum_{j=0}^{m} \mu^j [F_{\hat{x}}(\hat{x}) - \tilde{f}^j(\hat{x})] = 0, \tag{10.1.7b}$$

where $\tilde{f}^0(x) \triangleq f^0(x) - f^0(\hat{x})$ and $\tilde{f}^j(x) \triangleq f^j(x)$ for all $j \in \underline{m}$. Since $F_{\hat{x}}(\hat{x}) = 0$ and since $\tilde{f}^j(\hat{x}) \leq 0$ for $j = 0,1,\cdots,m$, we see that (10.1.7a), (10.1.7b) are equivalent to (10.1.2a), (10.1.2b). ■

Remark 10.1.1 : We shall refer to the multipliers μ^j in (10.1.2a), (10.1.2b) as *F. John multipliers.* ■

Originally, both the F. John theorem and the following very important special case were obtained in a very different manner from the one that we have employed. At this point, we reintroduce the notation which we used in Lecture 10, viz., for all $x \in \mathbb{R}^n$, we define the max function $\psi:\mathbb{R}^n \to \mathbb{R}$ by

$$\psi(x) \triangleq \max_{j \in \underline{m}} f^j(x), \tag{10.1.8a}$$

and, as before, we denote its generalized gradient by

$$\partial\psi(x) \triangleq \operatorname*{co}_{j \in I(x)} \{ \nabla f^j(x) \}, \tag{10.1.8b}$$

where

$$I(x) \triangleq \{ j \in \underline{m} \mid f^j(x) = \psi(x) \}. \tag{10.1.8c}$$

Corollary 10.1.1 (Kuhn-Tucker) : Suppose that $\hat{x}$ is a local miniminizer for P_I (with associated radius $\hat{\rho} \geq 0$) and that $0 \notin \partial\psi(\hat{x})$. Then there exist multipliers $\mu^1 \geq 0,\ldots,\mu^m \geq 0$, such that

$$\nabla f^0(\hat{x}) + \sum_{j=1}^{m} \mu^j \nabla f^j(\hat{x}) = 0, \tag{10.1.9a}$$

and

$$\mu^j f^j(\hat{x}) = 0 \ \forall \ j \in \underline{m}. \tag{10.1.9b}$$

■

Exercise 10.1.2 : Prove Corollary 10.1.1. ■

Theorem 10.1.2 : Suppose that the functions $f^j(\cdot)$, $j = 0,1,\ldots,m$, in (10.1.1), are convex and continuously differentiable and that $\hat{x} \in \mathbb{R}^n$ satisfies $\psi(\hat{x}) \leq 0$, as well as (10.1.9a), (10.1.9b). Then $\hat{x}$ is a global minimizer for P_I.

Proof : Let the μ^j, $j \in \underline{m}$, be as in (10.1.9a), (10.1.9b), and consider the *Lagrangian* $L:\mathbb{R}^n \to \mathbb{R}$, defined by

$$L(x) \triangleq f^0(x) + \sum_{j=1}^{m} \mu^j f^j(x) . \tag{10.1.10a}$$

Then $L(\cdot)$ is convex and, by (10.1.9a), $\nabla L(\hat{x}) = 0$. Hence $\hat{x}$ is a global minimizer of $L(\cdot)$. Since (10.1.9b) holds, we have that

$$f^0(\hat{x}) = L(\hat{x}) \leq L(x) \quad \forall\, x \in \mathbb{R}^n . \tag{10.1.10b}$$

Since for all $x \in \mathbb{R}^n$ such that $\psi(x) \leq 0$, $L(x) \leq f^0(x)$, it now follows that

$$f^0(\hat{x}) \leq f^0(x) \quad \forall\, x \in \{\, x \in \mathbb{R}^n \mid \psi(x) \leq 0 \,\}, \tag{10.1.10c}$$

which completes our proof. ■

Remark 10.1.2 : We shall refer to the multipliers μ^j in (10.1.9b), (10.1.9c) as *Kuhn-Tucker multipliers.* ■

Exercise 10.1.2 : Supppose that the functions $f^j(\cdot)$ $j = 0,1, \cdots, m$ are convex and continuously differentiable and that $\hat{x} \in \mathbb{R}^n$ is such that $\psi(\hat{x}) \leq 0$ and $0 \in \partial F_{\hat{x}}(\hat{x})$, where $F_{\hat{x}}(\cdot)$ is as in (10.1.3). Show by example that $\hat{x}$ need not be a minimizer for (10.1.1). ■

10.2. AN OPTIMALITY FUNCTION

Constrained minimax algorithms can be obtained as implementations of the following phase I - phase II[1] *conceptual* method of centers which has a simple geometric interpretation, see Fig. 10.2.1a, 10.2.1b. The method below is a straightforward generalization of the Huard method of centers[2].

Conceptual Method of Centers 10.2.1 :

Step 0 : Select $x_0 \in \mathbb{R}^n$ and set $i = 0$.

Step 1 : Compute

$$x_{i+1} = \arg\min_{x \in \mathbb{R}^n} \max\{\psi^0(x) - \psi^0(x_i) - \psi_+(x_i), \psi^j(x); j \in \underline{m}\} . \tag{10.2.1}$$

Step 2 : Set $i = i + 1$ and go to Step 1. ■

fBTheorem 10.2.1 : Suppose that

(a) For every $x' \in \mathbb{R}^n$, the level sets $\{\, x \in \mathbb{R}^n \mid F_{x'}(x) \leq F_{x'}(x') \,\}$ are compact, where $F_{x'}(\dot{c})$ was defined in (9.1.3);

---

[1] The reason for calling this method *phase I - phase II* is that it combines the operation of finding a feasible point (phase I) with that of minimizing the cost while maintaining feasibility (phase II).

[2] P. Huard, "Programmation Mathematic Convex", *Rev. Fr. Inform. Rech. Operation.*, Vol. 7, pp. 43-59, 1968.

(b) For every $x' \in \mathbb{R}^n$ which is not a local minimizer of (10.1.1b),

$$\Delta(x') \triangleq \min_{x \in \mathbb{R}^n} F_{x'}(x) - F_{x'}(x') < 0 . \tag{10.2.2a}$$

If $\{ x_i \}_{i=0}^{\infty}$ is an infinite sequence constructed by the Conceptual Method of Centers 10.2.1, then every accumulation point $\hat{x}$ of $\{x_i\}_{i=0}^{\infty}$ is a local minimizer for (10.1.1).

Proof : First we note that if $x \in \mathbb{R}^n$ is such that $\psi(x) > 0$, then $F_x(x) = \psi(x) = \psi_+(x)$. If $x \in \mathbb{R}^n$ is such that $\psi(x) \leq 0$, then $F_x(x) = 0$. Since for any $x,x' \in \mathbb{R}^n$, $\psi(x) \leq F_{x'}(x)$, we see that if $\psi(x_i) \leq 0$, then $\psi(x_{i+1}) \leq F_{x_i}(x_{i+1}) \leq F_{x_i}(x_i)$. This leads to the conclusion that if the Conceptual Method of Centers 10.2.1 constructs a sequence $\{x_i\}_{i=0}^{\infty}$, such that for some i_0, $\psi(x_{i_0}) \leq 0$, then $\psi(x_i) \leq 0$ for all $i \geq i_0$.

Next we note that because of assumption (a), and Lemma 10.2.1, $\Delta(\cdot)$ is continuous. Now suppose that $\{x_i\}_{i=0}^{\infty}$ is an infinite sequence constructed by the Conceptual Method of Centers 10.2.1, such that $x_i \xrightarrow{K} \hat{x}$, with $0 \notin \partial F_{\hat{x}}(\hat{x})$. Then we must have that $\Delta(\hat{x}) = -\delta < 0$, by assumption (b), and hence, by continuity of $\Delta(\cdot)$, there must exist an i_1 such that

$$F_{x_i}(x_{i+1}) - F_{x_i}(x_i) = \Delta(x_i) \leq -\tfrac{1}{2}\delta \tag{10.2.2b}$$

for all $i \in K$ such that $i \geq i_1$. Now suppose that $\psi(x_i) > 0$ for all i. Then $\{\psi(x_i)\}_{i=0}^{\infty}$ is a monotone decreasing sequence with accumulation point $\psi(\hat{x})$. Hence we must have that $\psi(x_i) \to \psi(\hat{x})$ as $i \to \infty$. However, it follows from (10.2.2b) that

$$\psi(x_{i+1}) - \psi(x_i) = F_{x_i}(x_{i+1}) - F_{x_i}(x_i) \leq -\tfrac{1}{2}\delta \tag{10.2.2c}$$

for all $i \in K$ such that $i \geq i_1$, which leads to a contradiction.

Next suppose that there is an i_0 such that $\psi(x_{i_0}) \leq 0$ for all $i > i_0$. Then

$$\psi^0(x_{i+1}) - \psi^0(x_i) \leq F_{x_i}(x_{i+1}) - F_{x_i}(x_i) \leq -\tfrac{1}{2}\delta \tag{10.2.2d}$$

for all $i \in K$ such that $i \geq i_2 = \max\{i_0,i_1\}$, i.e., the sequence $\{\psi^0(x_i)\}_{i=i_2}^{\infty}$ is monotone decreasing to $-\infty$. However, $\{\psi^0(x_i)\}_{i=i_2}^{\infty}$ must converge to the accumulation point $\psi^0(\hat{x})$, and hence again we obtain a contradiction. This completes our proof. ■

The simplest implementation of the Conceptual Method of Centers consists of replacing the update formula (10.2.1) by one iteration of the Minimax Algorithm 9.4.1. This leads us to the following optimality function and associated search direction function for the problem P_I defined in (10.1.1). Since there is little likelihood of confusion, we shall reuse the symbols θ and h which were first used in Lecture 9.2. We need to introduce one more function:

$$\psi_+(x) \triangleq \max\{ 0,\psi(x) \}, \tag{10.2.3a}$$

where $\psi(\cdot)$ was defined in (10.1.8a), and an arbitrary constant $\gamma > 0$. We now define for problem P_I the *optimality function* $\theta:\mathbb{R}^n \to \mathbb{R}$ and the associated *search direction function* $h:\mathbb{R}^n \to \mathbb{R}^n$ by

$$\theta(x) \triangleq \min_{h \in \mathbb{R}^n} \max \{ -\gamma\psi_+(x) + \langle \nabla f^0(x), h\rangle + \tfrac{1}{2}|h|^2;$$

$$f^j(x) - \psi_+(x) + \langle \nabla f^j(x), h\rangle + \tfrac{1}{2}|h|^2, \ j \in \underline{m} \}, \tag{10.2.3b}$$

$$h(x) \triangleq \arg \min_{h \in \mathbb{R}^n} \max \{ -\gamma\psi_+(x) + \langle \nabla f^0(x), h\rangle + \tfrac{1}{2}|h|^2;$$

$$f^j(x) - \psi_+(x) + \langle \nabla f^j(x), h\rangle + \tfrac{1}{2}|h|^2, \ j \in \underline{m} \}. \tag{10.2.3c}$$

We shall see in the next section that the constant γ can be used to control the trade-off involved between finding *some* feasible solution as rapidly as possible and finding a *low-cost* feasible solution.

To simplify the process of deducing the properties of the optimality function $\theta(\cdot)$ and associated search direction function $h(\cdot)$ for P_I from those defined in (9.2.13a), (9.2.13b), we define

$$\tilde{f}^0(x) \equiv 0, \tag{10.2.4a}$$

$$\tilde{f}^j(x) \equiv f^j(x), \quad \forall\, j \in \underline{m}, \tag{10.2.4b}$$

$$\bar{\underline{m}} \triangleq \{ 0, 1, \ldots, m \}, \tag{10.2.4c}$$

and, finally, we define $\gamma^0 = \gamma$, $\gamma^j = 1$ for all $j \in \underline{m}$.

Making use of these definitions, (10.2.2), (10.2.3) can be rewritten as

$$\theta(x) \triangleq \min_{h \in \mathbb{R}^n} \max_{j \in \bar{\underline{m}}} \{ \gamma^j[\tilde{f}^j(x) - \psi_+(x)] + \langle \nabla f^j(x), h\rangle + \tfrac{1}{2}|h|^2 \}, \tag{10.2.5a}$$

$$h(x) \triangleq \arg \min_{h \in \mathbb{R}^n} \max_{j \in \bar{\underline{m}}} \{ \gamma^j[\tilde{f}^j(x) - \psi_+(x)] + \langle \nabla f^j(x), h\rangle + \tfrac{1}{2}|h|^2 \}. \tag{10.2.5b}$$

Theorem 10.2.1 : Consider the functions $\theta(\cdot)$ and $h(\cdot)$ defined by (10.2.3a) and (10.2.3b). Then,

(a) For all $x \in \mathbb{R}^n$,

$$\theta(x) \le 0; \tag{10.2.6a}$$

(b) For all $x \in \mathbb{R}^n$,

$$d\psi(x; h(x)) \le \theta(x) - [\psi(x) - \psi_+(x)]; \tag{10.2.6b}$$

(c) For all $x \in \mathbb{R}^n$,

$$df^0(x; h(x)) \le \theta(x) + \gamma^0\psi_+(x); \tag{10.2.6c}$$

(d) Alternative expressions for $\theta(x)$ and $h(x)$ are given by

$$\theta(x) = -\min_{\mu \in \bar{\Sigma}} \{ \sum_{j=0}^{m} \mu^j\gamma^j[\psi_+(x) - \tilde{f}^j(x)] + \tfrac{1}{2}|\sum_{j=0}^{m} \mu^j \nabla f^j(x)|^2 \}, \tag{10.2.6d}$$

$$h(x) = -\sum_{j=0}^{m} \mu_x^j \nabla f^j(x) \,. \tag{10.2.6e}$$

where the μ_x is any solution of (10.2.6d).

Equivalently, let $\overline{G}^{0,\Psi}(x) \subset \mathbb{R}^{n+1}$ be a set with elements denoted by $\overline{\xi} = (\xi^0, \xi)$, with $\xi^0 \in \mathbb{R}$, $\xi \in \mathbb{R}^n$, and defined by

$$\overline{G}^{0,\Psi}(x) \triangleq \underset{j \in \overline{m}}{\text{co}} \left\{ \begin{bmatrix} \gamma^j[\psi_+(x) - \overline{f}^j(x)] \\ \nabla f^j(x) \end{bmatrix} \right\}. \tag{10.2.6f}$$

Then,

$$\theta(x) = - \min_{\xi \in \overline{G}^{0,\Psi}(x)} \{ \xi^0 + \tfrac{1}{2}|\xi|^2 \}, \tag{10.2.6g}$$

$$h(x) = -\xi(x), \tag{10.2.6h}$$

where

$$\overline{\xi}(x) = (\xi^0(x), \xi(x)) = \arg \min_{\xi \in \overline{G}^{0,\Psi}(x)} \{ \xi^0 + \tfrac{1}{2}|\xi|^2 \}. \tag{10.2.6i}$$

(e) For any $x \in \mathbb{R}^n$ such that $\psi(x) \geq 0$, $0 \in \partial\psi(x) \Rightarrow \theta(x) = 0$.

(f) For any $x \in \mathbb{R}^n$ such that $\psi(x) > 0$, $\theta(x) = 0 \Rightarrow 0 \in \partial\psi(x)$.

(g) Both $\theta(\cdot)$ and $h(\cdot)$ are continuous. ■

Exercise 10.2.1 : Follow the proof of Theorem 10.2.3, to construct a proof for Theorem 10.2.1. ■

10.3. PHASE I - PHASE II METHODS OF FEASIBLE DIRECTIONS

To conclude this lecture, we shall describe two phase I - phase II algorithms for solving the problem P_I (10.2.1), which we rewrite in the more compact form

$$P_I \quad \min\{ f^0(x) \mid \psi(x) \leq 0 \}, \tag{10.3.1a}$$

with

$$\psi(x) = \max_{j \in \overline{m}} f^j(x) \,. \tag{10.3.1b}$$

To simplify the proofs of convergence to follow, we will assume in this section that the functions $f^j: \mathbb{R}^n \to \mathbb{R}$ are *twice continuously differentiable*[3].

As we have already pointed out in the preceeding section, an algorithm for solving (10.1.1) which combines the operation of finding a feasible point (phase I) with that of minimizing the cost while maintaining feasibility (phase II), is usually referred to as a *phase I - phase II* algorithm. Phase II

[3] Referring to E. Polak, R. Trahan and D. Q. Mayne, "Combined Phase I - Phase II Methods of Feasible Directions", *Mathematical Programming*, Vol. 17, No. 1, pp. 32-61, 1979, we see that convergence can be proved also under the assumption

versions of the algorithms below, were first proposed by Pironneau and Polak[4] as an implementation of the Huard method of centers[5]. In the form given below, these algorithms were first described Polak, Trahan, and Mayne[6], and were obtained as reasonably straightforward generalizations of Algorithms 9.3.1 and 9.3.2, via the Conceptual Method of Centers 10.1.1.

We continue using the notation introduced in (10.2.4a) - (10.2.5c).

Algorithm 10.3.1 (Pironneau-Polak) : *(Exact Line Search).*

Parameters : $\gamma^0 > 0, \gamma^j = 1, j \in \underline{m}$.

Data : $x_0 \in \mathbb{R}^n$.

Step 0 : Set $i = 0$.

Step 1 : Compute the *search direction*

$$h_i = h(x_i) \triangleq \arg \min_{h \in \mathbb{R}^n} \max_{j \in \bar{m}} \{ \gamma^j [\bar{f}^j(x_i) - \psi(x_i)_+] + \langle \nabla f^j(x_i), h \rangle + \tfrac{1}{2}|h|^2 \} . \tag{10.3.2a}$$

Step 2 : Compute the *step size*:

if $\psi(x_i) > 0$,

$$\lambda_i \in \arg \min_{\lambda \geq 0} \psi(x_i + \lambda h_i), \tag{10.3.2b}$$

if $\psi(x_i) < 0$,

$$\lambda_i \in \arg \min_{\lambda \geq 0} \{ f^0(x_i + \lambda h_i) \mid \psi(x_i + \lambda h_i) \leq 0 \}, \tag{10.3.2c}$$

Step 3 : *Update:* set

$$x_{i+1} = x_i + \lambda_i h_i, \tag{10.3.2d}$$

replace i by $i + 1$ and go to Step 1. ■

The interesting features of Algorithm 10.3.1 are: (i) it does not have to be initialized with a feasible point (i.e., it is not necessary to have $\psi(x_0) \leq 0$), and (ii) once a feasible point x_{i_0} has been constructed, the following points x_i, with $i > i_0$, are also all feasible.

Theorem 10.3.1 : Consider problem (10.3.1a) with the assumptions stated and, in addition, assume that $0 \notin \partial\psi(x)$ for all $x \in \mathbb{R}^n$ such that $\psi(x) > 0$. Then every accumulation point $\hat{x}$, of a sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 10.3.1, satisfies $\psi(\hat{x}) \leq 0$ and the first order optimality condition $\theta(\hat{x}) = 0$, with $\theta(\cdot)$ defined by (10.2.3a).

---

that the functions $f^j(\cdot)$ are only once continuously differentiable.

[4] O. Pironneau and E. Polak, "On the Rate of Convergence of Certain Methods of Centers", *Mathematical Programming*, Vol. 2, No. 2, pp. 230-258, 1972.

[5] P. Huard, "Programmation Mathematic Convex", *Rev. Fr. Inform. Rech. Operation.*, Vol. 7, pp. 43-59, 1968.

[6] E. Polak, R. Trahan and D. Q. Mayne, "Combined Phase I - Phase II Methods of Feasible Directions", *Mathematical Programming*, Vol. 17, No. 1, pp. 32-61, 1979.

Proof : Suppose that Algorithm 10.3.1 has constructed a sequence $\{x_i\}_{i=0}^{\infty}$ which has an accumulation point $\hat{x}$ such that $\theta(\hat{x}) < 0$.

Case 1: Suppose that $\psi(x_i) > 0$ for all $i \in \mathbf{N}$. Then $\{\psi(x_i)\}_{i=0}^{\infty}$ is a monotone decreasing sequence, and hence, since $\psi(\cdot)$ is continuous and $x_i \xrightarrow{K} \hat{x}$ as $i \to \infty$, for some $K \subset \mathbf{N}$, $\psi(x_i) \to \psi(\hat{x})$ as $i \to \infty$. Clearly, since $\psi(\cdot)$ is continuous, we must have that $\psi(\hat{x}) \geq 0$.

Since $\psi(\hat{x}) \geq 0$ and $\theta(\hat{x}) < 0$, by assumption, we have, by (10.2.6b), that

$$d\psi(\hat{x};h(\hat{x})) \leq \theta(\hat{x}) \triangleq -2\delta < 0 . \tag{10.3.4a}$$

Hence there exists a $\hat{\lambda} > 0$ such that

$$\psi(\hat{x} + \hat{\lambda}h(\hat{x})) - \psi(\hat{x}) \leq -\hat{\lambda}\delta . \tag{10.3.4b}$$

Hence, because both $\psi(\cdot)$ and $h(\cdot)$ are continuous, there exists an $i_0 \in \mathbf{N}$ such that for all $i \in K$, $i \geq i_0$,

$$\psi(x_{i+1}) - \psi(x_i) \leq \psi(x_i + \hat{\lambda}h(x_i)) - \psi(x_i) \leq -\hat{\lambda}\delta/2, \tag{10.3.4c}$$

which leads to the conclusion that $\psi(x_i) \to -\infty$, as $i \to \infty$, and we have a contradiction. Consequently, $\theta(\hat{x}) = 0$ must be true. Since, by assumption, $0 \notin \partial\psi(x)$ whenever $\psi(x) > 0$, it follows from the fact that $\theta(\hat{x}) = 0$ that $\psi(\hat{x}) = 0$ also.

Case 2: Suppose there is an $i_0 \in \mathbf{N}$ such that $\psi(x_i) \leq 0$ for all $i \geq i_0$. Then, because $\psi(\cdot)$ is continuous, we must have that $\psi(\hat{x}) \leq 0$. Since $\theta(\cdot)$ and $h(\cdot)$ are continuous, there exists a $\hat{\rho} > 0$ such that $\theta(x) \leq \theta(\hat{x})/2 < 0$ and $|h(x)| \leq 2|h(\hat{x})|$ for all $x \in \mathbf{B}(\hat{x},\hat{\rho})$. Hence, since all the functions $f^j(\cdot)$ are twice continuously differentiable, there exists an $1 \leq M < \infty$, such that (with $H^j(\cdot) \triangleq \partial^2 f^j(\cdot)/\partial x^2$) for all $x \in \mathbf{B}(\hat{x},\hat{\rho})$ and $\lambda \in [0,1]$, $|H^j(x + \lambda h(x))| \leq M$. Therefore, for all $x \in \mathbf{B}(\hat{x},\hat{\rho})$ and $\lambda \in [0,1/M]$,

$$\begin{aligned}
\psi(x + \lambda h(x)) - \psi(x)_+ &= \max_{j \in \bar{m}} f^j(x + \lambda h(x)) - \psi(x)_+ \\
&= \max_{j \in \bar{m}} \{ f^j(x) - \psi(x)_+ + \lambda \langle \nabla f^j(x), h(x) \rangle + \lambda^2 \int_0^1 (1 - s) \langle H^j(x + s\lambda h(x))h(x), h(x) \rangle ds \} \\
&\leq \max_{j \in \bar{m}} \{ \gamma^j [f^j(x) - \psi(x)_+] + \lambda \langle \nabla f^j(x), h(x) \rangle + \lambda^2 \int_0^1 (1 - s) \langle H^j(x + s\lambda h(x))h(x), h(x) \rangle ds \} \\
&\leq \max_{j \in \bar{m}} \{ \gamma^j [f^j(x) - \psi(x)_+] + \lambda \langle \nabla f^j(x), h(x) \rangle + \frac{\lambda^2 M}{2} |h(x)|^2 \} \\
&\leq \lambda\theta(x) \leq \lambda\theta(\hat{x})/2 < 0 .
\end{aligned} \tag{10.3.4d}$$

Next, it follows from (10.2.6c) that

$$df^0(\hat{x};h(\hat{x})) \leq \theta(\hat{x}) \triangleq -2\delta < 0 . \tag{10.3.4e}$$

Hence there exists a $\hat{\lambda} \in (0,1/M]$ such that

$$f^0(\hat{x} + \hat{\lambda}h(\hat{x})) - f(\hat{x}) \leq -\hat{\lambda}\delta . \tag{10.3.4f}$$

Since $x_i \overset{K}{\to} \hat{x}$ as $i \to \infty$, for some $K \subset \mathbf{N}$, there exists an $i_1 \in \mathbf{N}$ such that for all $i \in K$, $i \geq i_1$, $x_i \in \mathbf{B}(\hat{x},\hat{\rho})$, and hence it follows from (10.3.4d) that for all $i \in K$, $i \geq i_1$,

$$\psi(x_i + \lambda h(x_i)) \leq 0 . \tag{10.3.4g}$$

Next, since both $f^0(\cdot)$ and $h(\cdot)$ are continuous, it follows from (10.3.4f) that there exists an $i_2 \geq i_1$ such that for all $i \in K$, $i \geq i_2$,

$$f^0(x_{i+1} + \hat{\lambda}h(x_i)) - f^0(x_i) \leq -\hat{\lambda}\delta/2 . \tag{10.3.4h}$$

Clearly, (10.3.4g) and (10.3.4h) imply that for all $i \in K$, $i \geq i_2$,

$$f^0(x_{i+1}) - f^0(x_i) \leq -\hat{\lambda}\delta/2 . \tag{10.3.4i}$$

Since the sequence $\{f^0(x_i)\}_{i=0}^{\infty}$ is monotone decreasing, it follows from (10.3.4i) that $f(x_i) \to -\infty$ as $i \to \infty$. However, because the sequence $\{f^0(x_i)\}_{i=0}^{\infty}$ is monotone decreasing and $x_i \overset{K}{\to} \hat{x}$ and $f^0(\cdot)$ is continuous, we must have that $f^0(x_i) \to f^0(\hat{x})$ as $i \to \infty$, and hence we have a contradiction, which completes our proof. ■

It is also possible to propose a phase I - phase II algorithm which uses an Armijo type step size rule, as follows.

Algorithm 10.3.2 (Pironneau-Polak) : *(Armijo Line Search).*

Parameters : $\alpha, \beta \in (0,1)$.

Data : $x_0 \in \mathbf{R}^n$

Step 0 : Set $i = 0$.

Step 1 : Compute the *search direction*

$$h_i = h(x_i) \triangleq \arg \min_{h \in \mathbf{R}^n} \max_{j \in \bar{m}} \{ \gamma^j[f^j(x_i) - \psi(x_i)] + \langle \nabla f^j(x_i), h \rangle + \tfrac{1}{2}|h|^2 \} . \tag{10.3.5a}$$

Step 2 : Compute the *step size*

if $\psi(x_i) > 0$,

$$\lambda_i = \underset{k \in \mathbf{N}}{\operatorname{argmax}} \{ \beta^k \mid \psi(x_i + \beta^k h_i) - \psi(x_i) - \beta^k \alpha\theta(x_i) \leq 0 \} . \tag{10.3.5b}$$

if $\psi(x_i) \leq 0$,

$$\lambda_i = \underset{k \in \mathbf{N}}{\operatorname{argmax}} \{ \beta^k \mid f^0(x_i + \beta^k h_i) - f^0(x_i) - \beta^k \alpha\theta(x_i) \leq 0,\ \psi(x_i + \beta^k h_i) \leq 0 \}, \tag{10.3.5c}$$

Step 3 : *Update:* set

$$x_{i+1} = x_i + \lambda_i h_i, \tag{10.3.5d}$$

replace i by $i + 1$ and go to Step 1. ■

Theorem 10.3.2 : Consider problem (10.3.1a) with the assumptions stated and, in addition, assume that $0 \notin \partial\psi(x)$ for all $x \in \mathbb{R}^n$ such that $\psi(x) \geq 0$. Then every accumulation point $\hat{x}$ of a sequence $\{x_i\}_{i=0}^{\infty}$, constructed by Algorithm 10.3.1, satisfies $\psi(\hat{x}) \leq 0$ and the first order optimality condition $\theta(\hat{x}) = 0$, with $\theta(\cdot)$ defined by (10.2.3a). ■

Exercise 10.3.1 : Prove Theorem 10.3.2. ■

Exercise 10.3.2 : For any $x_i \in \mathbb{R}^n$, let $F_{x_i}:\mathbb{R}^n \to \mathbb{R}$ be defined by

$$F_{x_i}(x) = \max\{ f^0(x) - f^0(x_i) - \psi(x_i)_+ \, , f^j(x) - \psi(x_i)_+, \, j \in \underline{m} \}. \tag{10.3.6}$$

Prove that Theorem 10.3.2 also applies to the following Phase I - Phase II method which uses a single *surrogate* cost function for step length calculations:

Algorithm 10.3.3 : (*Single-Cost Armijo Line Search*).

Parameters : $\alpha, \beta \in (0,1)$.

Data : $x_0 \in \mathbb{R}^n$

Step 0 : Set $i = 0$.

Step 1 : Compute the *search direction*

$$h_i = h(x_i) \triangleq \arg \min_{h \in \mathbb{R}^n} \max_{j \in \bar{m}} \{ \gamma^j[\tilde{f}^j(x_i) - \psi(x_i)] + \langle \nabla f^j(x_i), h \rangle + \tfrac{1}{2}|h|^2 \} . \tag{10.3.7a}$$

Step 2 : Compute the *step size*

$$\lambda_i = \operatorname{argmax}_{k \in \mathbb{N}} \{ \beta^k \mid F_{x_i}(x_i + \beta^k h_i) - F_{x_i}(x_i) - \beta^k \alpha \theta(x_i) \leq 0 \} . \tag{10.3.7b}$$

Step 3 : *Update:* set

$$x_{i+1} = x_i + \lambda_i h_i, \tag{10.3.7c}$$

replace i by $i + 1$ and go to Step 1. ■

11. FIRST AND SECOND ORDER OPTIMALITY CONDITIONS : MIXED CONSTRAINTS

We now return to the study of first and second order optimality conditions for the full nonlinear programming problem

$$\mathbf{P_{IE}}: \quad \min\{ f^0(x) \mid f^j(x) \le 0, j \in \underline{m}; g^k(x) = 0, k \in \underline{l} \}. \tag{11.0.1}$$

11.1. FIRST ORDER OPTIMALITY CONDITIONS : MIXED CONSTRAINTS.

To reduce the amount of mathematical baggage that we need to manipulate at one time, let us first consider the special case of problem (11.0.1),

$$\mathbf{P_E}: \quad \min \{ f^0(x) \mid g^k(x) = 0, k \in \underline{l} \}, \tag{11.1.1}$$

where $f^0: \mathbb{R}^n \to \mathbb{R}$, $g^k: \mathbb{R}^n \to \mathbb{R}$, $k \in \underline{l}$ are all continuously differentiable functions.

Definition 11.1.1 : We shall say that $\hat{x}$ is a local miniminizer of $\mathbf{P_E}$ if $g^k(\hat{x}) = 0$ for all $k \in \underline{l}$ and there exists a $\hat{\rho} > 0$ such that $f^0(x) \ge f^0(\hat{x})$ for all $x \in \{ x \in \mathbb{R}^n \mid |x-\hat{x}| \le \hat{\rho}, g^k(x) = 0, k \in \underline{l} \}$. ■

Theorem 11.1.1 (FONC, $\mathbf{P_E}$: Suppose that $\hat{x}$ is a local miniminizer for $\mathbf{P_E}$. Then there exist multipliers $\psi^0, \psi^1, \ldots, \psi^l$, not all zero, such that

$$\psi^0 \nabla f^0(\hat{x}) + \sum_{k=1}^{l} \psi^k \nabla g^k(\hat{x}) = 0 . \tag{11.1.2}$$

Proof : Let $\hat{\rho} > 0$ be the radius associated with $\hat{x}$. First, consider the problem

$$P'_E: \quad \min\{ f^0(x) + |x - \hat{x}|^2 \mid g^k(x) = 0, k \in \underline{l} \}. \tag{11.1.3a}$$

Clearly, $\hat{x}$ is also a local miniminizer for P'_E and, in addition, for all $x \neq \hat{x}$ such that $x \in B(\hat{x},\hat{\rho}) \cap \{ x \mid g^k(x) = 0, k \in \underline{l} \}$,

$$f^0(x) + |x - \hat{x}|^2 > f^0(\hat{x}), \tag{11.1.3b}$$

i.e., $\hat{x}$ is the *only* local miniminizer of $\mathbf{P_E}'$ in the ball $B(\hat{x},\hat{\rho})$.

Now consider the family of inequality constrained problems

$$P^\varepsilon_f: \quad \min \{ f^0(x) + |x - \hat{x}|^2 \mid -\varepsilon \le g^k(x) \le \varepsilon, k \in \underline{l}, |x - \hat{x}|^2 \le \hat{\rho} \}, \tag{11.1.4}$$

with $\varepsilon \ge 0$, and let x_ε denote the solution of P^ε_f. Then we have

(i) $-\varepsilon \le g^k(\hat{x}) \le \varepsilon$, for all $k \in \underline{l}$, for all $\varepsilon \ge 0$, i.e., $\hat{x}$ is feasible for all the problems P^ε_f.

(ii) Suppose $\varepsilon_i \to 0$, as $i \to \infty$. Since the sequence $\{ x_{\varepsilon_i} \}_{i \in \mathbf{N}} \subset B(\hat{x},\hat{\rho})$ is bounded, it must have at least one accumulation point, say x^*. Clearly, $g^k(x^*) = 0$ for all $k \in \underline{l}$, and, since $f^0(x_{\varepsilon_i}) \le f^0(\hat{x})$ must

hold for all $i \in \mathbb{N}$, we must have that

$$f^0(x^*) \leq f^0(\hat{x}) . \tag{11.1.5a}$$

(iii) Next we show that $x^* = \hat{x}$. For suppose that $x^* \neq \hat{x}$. Then x^* cannot be a local miniminimizer of P'_ε, because $\hat{x}$ is the *only local miniminizer* of P'_ε in $B(\hat{x},\hat{\rho})$. Hence we must have that

$$f^0(\hat{x}) < f^0(x^*), \tag{11.1.5b}$$

which contradicts (11.1.5a)

(iv) Since x_{ε_i} is a local miniminizer of $P_I^{\varepsilon_i}$, it follows from Corollary 9.2.2. that there exist multiplier vectors μ_{ε_i}, with components $\mu^0_{\varepsilon_i}$, $\mu^k_{\varepsilon_{i+}}$, $\mu^k_{\varepsilon_{i-}} \geq 0$, $k \in \underline{l}$, such that[1]

$$\mu^0_{\varepsilon_i}[\nabla f^0(x_{\varepsilon_i}) + 2(x_{\varepsilon_i} - \hat{x})] + \sum_{k \in \underline{l}} \mu^k_{\varepsilon_{i+}} \nabla g^k(x_{\varepsilon_i}) - \sum_{k \in \underline{l}} \mu^k_{\varepsilon_{i-}} \nabla g^k(x_{\varepsilon_i}) = 0, \tag{11.1.6a}$$

$$\mu^k_{\varepsilon_{i+}}[g^k(x_{\varepsilon_i}) - \varepsilon_i] = 0, \tag{11.1.6b}$$

$$\mu^k_{\varepsilon_{i-}}[- g^k(x_{\varepsilon_i}) - \varepsilon_i] = 0, \tag{11.1.6c}$$

$$\mu^0_{\varepsilon_i} + \sum_{k \in \underline{l}} \mu^k_{\varepsilon_{i+}} + \sum_{k \in \underline{l}} \mu^k_{\varepsilon_{i-}} = 1 . \tag{11.1.6d}$$

Since all the components of μ_{ε_i}'s are in [0,1], they must have accumulation points. Hence there must be an infinite $K \subset \mathbb{N}$ such that $\mu_{\varepsilon_i} \overset{K}{\to} \mu$ as $i \to \infty$, with $\mu = (\mu^0,\mu^1_+,,....,\mu^l_+,\mu^1_-, \cdots ,\mu^l_-)$ satisfying

$$\mu^0 \nabla f^0(\hat{x}) + \sum_{k \in \underline{l}} (\mu^k_+ - \mu^k_-) \nabla g^k(\hat{x}) = 0, \tag{11.1.8a}$$

and

$$\mu^0 + \sum_{k=1}^{l} \mu^k_+ + \sum_{k=1}^{l} \mu^k_- = 1 . \tag{11.1.8b}$$

It remains to show that not all the coefficients in (11.1.8a) are zero. Since for all $i \in K$, either $\mu^i_{\varepsilon_{i+}}$ or $\mu^i_{\varepsilon_{i-}}$ are zero, or both, it follows that for all $k \in \underline{l}$, $\mu^k_+ \mu^k_- = 0$ and hence that one of these two coefficients must be zero. Hence it is not possible for all coeficients in (11.1.8a) to be zero, which completes our proof. ■

Exercise 11.1.1 : Suppose that the gradients $\{ \nabla g^k(\hat{x}) \}_{k \in \underline{l}}$ are linearly independent. Show that ψ^0 can be chosen to be 1 in (11.1.2). ■

Exercise 11.1.2 : Suppose that $f^0(\cdot)$ is convex, that the functions $g^k(\cdot)$ are affine (i.e., they are of the form $g^k(x) = A_k x + b_k$), and that $\hat{x}$ is such that (i) $g^k(\hat{x}) = 0$ for all $k \in \underline{l}$, that (11.1.1) is satisfied with

[1] Note that the multiplier associated with the constraint $|x - \hat{x}|^2 \leq \hat{\rho}$ can be assumed to be zero because this inequality is slack for all i sufficiently large.

multipliers which are not all zero and that the gradients $\{ \nabla g^k(\hat{x}) \}_{k \in \underline{l}}$ are linearly independent. Show that $\hat{x}$ is a global minimizer for the problem P_E. ■

We are now in a position to state a first order optimality condition for the problem P_{IE} in (11.0.1).

Definition 11.1.2 : We shall say that $\hat{x}$ is a local miniminizer for P_{IE} if $f^j(\hat{x}) \leq 0$ for all $j \in \underline{m}$, $g^k(\hat{x}) = 0$ for all $k \in \underline{l}$ and there exists a $\hat{\rho} > 0$ such that $f^0(x) \geq f^0(\hat{x})$ for all $x \in \{ x \in \mathbb{R}^n \mid |x-\hat{x}| \leq \rho, f^j(x) \leq 0, j \in \underline{m}, g^k(x) = 0, k \in \underline{l} \}$. ■

Theorem 11.1.2 (FONC, P_{IE}): 0.*rConsiderproblem*(11.0.1)and*supposethatthefunctionsf* #^sup j : reals sup n~->~reals,j ~=~0 , 1 , . . . , m,andg sup k : reals sup n~->~reals,k member l under,*areallcontinuouslydifferentiable.Ifx* hat*isalocalminimizer*for(11.0.1),*thenthereexistmultipliers*mu sup j ge 0,j ~=~0 , 1 , . . . , m,andpsi sup k,k member l under,*notallzero,suchthat.EQI*(11.1.9*a*)$\mu^0 \nabla f^0(\hat{x}) + \sum_{j=1}^{m} \mu^j \nabla f^j(\hat{x}) + \sum_{k=1}^{l} \psi^k \nabla g^k(\hat{x}) = 0$,.EQ I (11.1.9b) mu sup j f sup j (x hat)~=~0 ,~~j~=~1 , 2 , . . . , m~.

(11.1.8b)

Exercise 11.1.3 : Prove Theorem 11.1.2. ■

Exercise 11.1.4 : Consider the problem

$$\min\{ f(x) \mid Ax - b = 0 \}, \tag{11.1.10a}$$

where $f:\mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, $x = (x_1, x_2)$, with $x_1 \in \mathbb{R}^l$, and $A = [A_1, A_2]$ an $l \times n$ matrix such that the $l \times l$ matrix A_1 is nonsingular.

(a) Show that the problem (11.1.10a) is equivalent to the problem

$$\min_{x \in \mathbb{R}^{n-l}} \tilde{f}(x_2), \tag{11.1.10b}$$

where $\tilde{f}(x_2) \triangleq f(-A_1^{-1}A_2x_2 + b, x_2)$.

(b) Show that the optimality condition for (11.1.10b)

$$\nabla \tilde{f}(\hat{x}_2) = 0 \tag{11.1.10c}$$

is equivalent to the optimality condition for (11.1.10a)

$$\begin{aligned} A\hat{x} - b &= 0 \\ \nabla f(\hat{x}) + A^T\psi &= 0, \end{aligned} \tag{11.1.10d}$$

for some $\psi \in \mathbb{R}^l$.

(c) show that the optimality condition for (11.1.10b) (11.1.10c) together with

$$\langle y, \frac{\partial^2 f(\hat{x}_2)}{\partial x_2^2} y \rangle \geq 0 \quad \forall\ y \in \mathbf{R}^{n-l}, \tag{11.1.10e}$$

is equivalent to the optimality condition for (11.1.10a), consisting of (11.1.10d) together with

$$\langle y, \frac{\partial^2 f(\hat{x})}{\partial x^2} y \rangle \geq 0 \quad \forall\ y \in M_E(\hat{x}), \tag{11.1.10f}$$

where $M_E(\hat{x}) = \{\, y \in \mathbf{R}^n \mid Ay = 0 \,\}$. ■

11.2. SECOND ORDER OPTIMALITY CONDITIONS : MIXED CONSTRAINTS

We shall now establish second order optimality conditions for the full nonlinear programming problem (11.0.1), i.e., for

$$P_{IE}: \qquad \min\{\, f^0(x) \mid f^j(x) \leq 0,\ j \in \underline{m};\ g^k(x) = 0,\ k \in \underline{l} \,\}\ . \tag{11.2.1}$$

We proceed in two stages: first we assume that there are only equality constraints in (11.2.1), as in (11.1.1), and then we consider the full problem. However, we must first digress to recall the Implicit Function Theorem and some of its consequences.

Implicit Function Theorem 11.2.1 : Suppose that $g: \mathbf{R}^l \times \mathbf{R}^{n-l} \to \mathbf{R}^l$ is k times continously differentiable. If $\hat{x}_1 \in \mathbf{R}^l$, $\hat{x}_2 \in \mathbf{R}^{n-l}$ are such that $g(\hat{x}_1, \hat{x}_2) = 0$ and the matrix $\partial g(\hat{x}_1, \hat{x}_2) / \partial x_1$ is non-singular, then there exists a $\hat{\rho} > 0$ and a k times continuously differentiable function $\phi: \mathbf{B}(\hat{x}_2, \hat{\rho}) \to \mathbf{B}(\hat{x}_1, \hat{\rho})$ such that $\phi(\hat{x}_2) = \hat{x}_1$,

$$\frac{\partial \phi(\hat{x}_2)}{\partial x_2} = -\left[\frac{\partial g(\hat{x}_1, \hat{x}_2)}{\partial x_1}\right]^{-1} \frac{\partial g(\hat{x}_1, \hat{x}_2)}{\partial x_2}, \tag{11.2.2a}$$

and

$$g(\phi(y), y) = 0 \quad \forall\ y \in \mathbf{B}(\hat{x}_2, \hat{\rho})\ . \tag{11.2.2b}$$

■

Corollary 11.2.1 : Suppose that $g: \mathbf{R}^n \to \mathbf{R}^l$ is twice continuously differentiable and that $\hat{x} \in \mathbf{R}^n$ is such that $g(\hat{x}) = 0$, and $\partial g(\hat{x}) / \partial x$ has row rank l. Then, given any $h \neq 0$ in $\mathbf{R}^n$ such that $\partial g(\hat{x}) / \partial x h = 0$, there exists a $t_h > 0$ and a twice continuously differentiable function $s: [0, t_h] \to \mathbf{R}^n$ such that (i) $s(0) = \hat{x}$, (ii) $s'(0) = h$, and (iii) $g(s(t)) = 0$ for all $t \in [0, t_h]$.

Proof : Let $h \in \mathbf{R}^n$, $h \neq 0$, be given. Without loss of generality we can assume that we can partition vectors $x \in \mathbf{R}^n$ into two parts, so that $x = (x_1^T, x_2^T)^T$, with $x_1 \in \mathbf{R}^l$, $x_2 \in \mathbf{R}^{n-l}$, with the partition such that $\partial g(\hat{x}) / \partial x_1$ is invertible. To simplify notation, we shall write $x = (x_1, x_2)$. Let $\hat{\rho} > 0$ be a radius and let $\phi: \mathbf{B}(\hat{x}_2, \hat{\rho}) \to \mathbf{B}(\hat{x}_1, \hat{\rho})$ the corresponding twice differentiable function, as postulated in the Implicit Function Theorem. Next, let $t_h \in (0, \hat{\rho}/|h|)$. Then, using the partition $h = (h_1, h_2)$, the function $s: [0, t_h] \to \mathbf{R}^n$

given by

$$s(t) \triangleq (\phi(\hat{x}_2 + th_2)\hat{x}_2 + th_2) \tag{11.2.2c}$$

is well defined and twice continuously differentiable. Clearly, $s(0) = \hat{x}$, $s'(0) = (\frac{\partial\phi(\hat{x}_2)}{\partial x_2}h_2, h_2)$ and $g(s(t)) = 0$ for all $t \in [0, t_h]$.

Now, since $\frac{\partial g(\hat{x})}{\partial x}h = 0$ by assumption, we must have, in partitioned form, that

$$\frac{\partial g(\hat{x})}{\partial x_1}h_1 + \frac{\partial g(\hat{x})}{\partial x_2}h_2 = 0, \tag{11.2.2d}$$

i.e., that

$$h_1 = -\left[\frac{\partial g(\hat{x})}{\partial x_1}\right]^{-1}\frac{\partial g(\hat{x})}{\partial x_2}h_2 = \frac{\partial\phi(\hat{x}_2)}{\partial x_2}h_2 . \tag{11.2.2e}$$

Hence we see that $s'(0) = h$, which completes our proof. ■

Lemma 11.2.1: Suppose that $g:\mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and that for some $\hat{x} \in \mathbb{R}^n$, the matrix $\partial g(\hat{x})/\partial x$. has maximum row rank. Let

$$M_E(\hat{x}) = \{\, y \in \mathbb{R}^n \mid \frac{\partial g(\hat{x})}{\partial x}y = 0 \,\} . \tag{11.2.3a}$$

Then

$$M_E(\hat{x}) = \{\, y \in \mathbb{R}^n \mid \alpha y = \lim_{i \to \infty} (x_i - \hat{x})/|(x_i - \hat{x})|,\ x_i \to \hat{x},\ \text{as } i \to \infty,\ g(x_i) = 0, \forall\ i \in \mathbb{N},\ \alpha \in \mathbb{R} \,\} \tag{11.2.3b}$$

Proof : Suppose that $\{x_i\}_{i=0}^{\infty}$ is such that $x_i \to \hat{x}$ as $i \to \infty$ and $g(x_i) = 0$ for all $i \in \mathbb{N}$. Then, for all $i \in \mathbb{N}$,

$$0 = g(x_i) = g(\hat{x}) + \int_0^1 \frac{\partial g(\hat{x} + s(x_i - \hat{x}))}{\partial x}\, ds\, (x_i - \hat{x}) . \tag{11.2.3c}$$

Let $y_i = (x_i - \hat{x})/|(x_i - \hat{x})|$. Then it follows from (11.2.3c) that any accumulation point y of the sequence $\{y_i\}_{i=0}^{\infty}$ must be in $M_E(\hat{x})$. Next, suppose that $y \in M_E(\hat{x})$ is arbitrary, but nonzero. Without loss of generality, we may assume that $|y| = 1$. Then, by Corollary 11.2.1, there exists a $t_y > 0$ and a function $s:[0, t_y] \to \mathbb{R}^n$ such that (i) $g(0) = \hat{x}$, (ii) $g(s(t)) \equiv 0$, and (iii) $s'(0) = y$. Let $\{t_i\}_{i=0}^{\infty} \subset [0, t_y]$ be such that $t_i \to 0$ as $i \to \infty$. Then if $x_i \triangleq s(t_i)$, $x_i \to \hat{x}$, as $i \to \infty$. Let $y_i \triangleq (x_i - \hat{x})/|(x_i - \hat{x})|$. Then, by the Mean Value Theorem, $y_i = s'(\lambda_i t_i)/|s'(\lambda_i t_i)|$, with $\lambda_i \in [0,1]$. It follows that $y_i \to y$, $i \to \infty$. Hence we conclude that (11.2.3b) holds. ■

Theorem 11.2.2 (SONC, P_E) : Consider the problem

$$\mathbf{P_E}: \quad \min\{ f^0(x) \mid g^k(x) = 0, k \in \underline{l} \}, \tag{11.2.4}$$

with $f^0:\mathbf{R}^n \to \mathbf{R}$ and all the $g^k:\mathbf{R}^n \to \mathbf{R}$ twice continuously differentiable.

Suppose that $\hat{x}$ is a local miniminizer for (11.2.3a) and that the matrix $\partial g(\hat{x})/\partial x$ has maximum row rank. Then there exists a multiplier vector $\psi \in \mathbf{R}^l$ such that the Lagrangian $L:\mathbf{R}^n \to \mathbf{R}$ defined by

$$L(x) \triangleq f^0(x) + \langle \psi, g(x) \rangle \tag{11.2.5a}$$

satisfies

$$\nabla L(\hat{x}) = \nabla f^0(\hat{x}) + \left[\frac{\partial g(\hat{x})}{\partial x}\right]^T \psi = 0, \tag{11.2.5b}$$

and

$$\langle y, \frac{\partial^2 L(\hat{x})}{\partial x^2} y \rangle \geq 0 \ \forall \ y \in \mathbf{M}_E(\hat{x}), \tag{11.2.6}$$

where

$$\mathbf{M}_E(\hat{x}) = \{ y \in \mathbf{R}^n \mid \frac{\partial g(\hat{x})}{\partial x} y = 0 \}. \tag{11.2.7}$$

Proof : We only need to establish (11.2.6) since (11.2.5b) was established in Theorem 11.1.1. Suppose that $y \in \mathbf{M}_E(\hat{x})$, is arbitrary, but nonzero. Let $\{x_i\}_{i=0}^{\infty}$ be such that (i) $x_i \to \hat{x}$ as $i \to \infty$, (ii) $g(x_i) = 0$, for all $i \in \mathbf{N}$, and (iii) $y = \lim_{i \to \infty} (x_i - \hat{x})/\|(x_i - \hat{x})\|$. Then, since $\hat{x}$ is a local minimizer, there exists an $i_0 \in \mathbf{N}$ such that $f^0(x_i) \geq f^0(\hat{x})$, for all $i \geq i_0$. Equivalently, since $g(x_i) = 0$ for all i,

$$f^0(\hat{x}) \leq L(x_i), \ \forall \ i \geq i_0 . \tag{11.2.8}$$

Expanding the right hand side of (11.2.9a) about $\hat{x}$ to second order, we obtain that

$$f^0(\hat{x}) \leq L(\hat{x}) + \langle \nabla L(\hat{x}), x_i - \hat{x} \rangle + \int_0^1 (1 - s) \langle (x_i - \hat{x}), \frac{\partial^2 L(\hat{x} + s(x_i - \hat{x}))}{\partial x^2} (x_i - \hat{x}) \rangle ds . \tag{11.2.9}$$

Now, $L(\hat{x}) = f^0(\hat{x})$, and $\nabla L(\hat{x}) = 0$ by (11.2.5b). Hence, taking limits, (11.2.9) leads to (11.2.6), which completes our proof. ■

We are now ready to consider problem (11.2.1) in full, and develop both necessary and sufficient second order conditions.

Theorem 11.2.3 (SONC, $\mathbf{P_{EI}}$) : Consider the problem

$$\mathbf{P_{IE}}: \quad \min\{ f^0(x) \mid f^j(x) \leq 0, j \in \underline{m}, g(x) = 0 \}, \tag{11.2.10}$$

where the functions $f^j:\mathbf{R}^n \to \mathbf{R}, j = 0,1, ...,m$ and $g:\mathbf{R}^n \to \mathbf{R}^l$ are twice continuously differentable. Suppose that $\hat{x}$ is a local optimal solution of (11.2.10) and that the gradients $\nabla g^i(\hat{x}), j \in \underline{l}$ together with the gradients $\nabla f^j(\hat{x}), j \in I(\hat{x}) \triangleq \{ j \in \underline{m} \mid f^j(\hat{x}) = 0 \}$, are linearly independent. Then there exist

multiplier vectors $\mu \in \mathbb{R}^m$, $\mu \geq 0$, $\psi \in \mathbb{R}^l$ such that

$$\nabla f^0(\hat{x}) + \sum_{j \in \underline{m}} \mu^j \nabla f^j(\hat{x}) + \frac{\partial g(\hat{x})^T}{\partial x} \psi = 0, \tag{11.2.11a}$$

$$\mu^j f^j(\hat{x}) = 0, \ \forall \ j \in \underline{m} \tag{11.2.11b}$$

and the lagrangian $L:\mathbb{R}^n \to \mathbb{R}$ defined by

$$L(x) \triangleq f^0(x) + \sum_{j \in \underline{m}} \mu^j f^j(x) + \sum_{k \in \underline{l}} \psi^k g^k(x) \tag{11.2.11c}$$

satisfies

$$\langle y, \frac{\partial^2 L(\hat{x})}{\partial x^2} y \rangle \geq 0 \quad \forall \ y \in \mathbf{M}_{IE}, \tag{11.2.11d}$$

where

$$\mathbf{M}_{IE} \triangleq \{ \ y \in \mathbb{R}^n \mid \frac{\partial g(x)}{\partial x} y = 0, \langle \nabla f^j(\hat{x}), y \rangle = 0, \forall \ j \in I(\hat{x}) \ \} . \tag{11.2.11e}$$

Proof : Clearly, if $\hat{x}$ solves (11.2.10), it must also solve

$$\min\{ \ f^0(x) \mid f^j(x) = 0, j \in I(\hat{x}), \ g(x) = 0 \ \} . \tag{11.2.11d}$$

The desired result now follows from Theorem 11.2.2. ■

Exercise 11.2.1 : Use the fact that the problem $\min_{x \in \mathbb{R}^n} \max_{j \in \underline{m}} f^j(x)$ is equivalent to the problem $\min\{ \ x^0 \mid f^j(x) - x^0 \leq 0 \ \}$ to prove the following theorem:

Thereom 11.2.4 (SONC for minimax): Consider the problem

$$\min_{x \in \mathbb{R}^n} \max_{j \in \underline{m}} f^j(x), \tag{11.2.12a}$$

where the functions $f^j:\mathbb{R}^n \to \mathbb{R}$ are twice continuously differentiable. Suppose that $\hat{x}$ is a local minimizer of $\psi(x) \triangleq \max_{j \in \underline{m}} f^j(x)$, and that the vectors $(1, \nabla f^j(\hat{x}), \ j \in I(\hat{x}) = \{ \ j \in \underline{m} \mid f^j(\hat{x}) = \psi(\hat{x}) \ \}$, are linearly independent. Then there exists a multiplier vector $\hat{\mu} \in \Sigma$ such that

$$\sum_{j=1}^{m} \hat{\mu}^j \nabla f^j(\hat{x}) = 0, \tag{11.2.12b}$$

$$\hat{\mu}^j (f^j(\hat{x}) - \psi(\hat{x})) = 0, \forall \ j \in \underline{m}, \tag{11.2.12c}$$

and, with the Lagrangian defined by $L(x) \triangleq \sum_{j=1}^{m} \hat{\mu}^j f^j(x)$,

$$\langle h, \frac{\partial^2 L(\hat{x})}{\partial x^2} h \rangle \geq 0, \ \forall \ h \in \mathbf{M}, \tag{11.2.12d}$$

where

$$\mathbf{M}_N \triangleq \{ h \in \mathbf{R}^n \mid \nabla f^j(\hat{x}), h\rangle = 0, j \in I(\hat{x}) \}. \tag{11.2.12f}$$

■

Next we develop second order sufficiency conditions. In this case there is no advantage in dealing with the equality constrained case first and hence we go directly to problem $\mathbf{P}_{\mathbf{IE}}$.

Theorem 11.2.5 (SOSC, $\mathbf{P}_{\mathbf{IE}}$) : Consider Problem (11.2.10) and suppose that the functions $f^j:\mathbf{R}^n \to \mathbf{R}$, $j = 0,1, ...,m$ and $g:\mathbf{R}^n \to \mathbf{R}^l$ are twice continuously differentable. Suppose that $\hat{x}$ is such that (i) $g(\hat{x}) = 0, f^j(\hat{x}) \leq 0$ for all $j \in \underline{m}$ and (ii) the gradients $\nabla g^j(\hat{x}), j \in \underline{l}$ together with the gradients $\nabla f^j(\hat{x}), j \in I(\hat{x}) \triangleq \{ \in \underline{m} \mid f^j(\hat{x}) = 0 \}$, are linearly independent. Suppose there exist multiplier vectors $\mu \in \mathbf{R}^m, \mu \geq 0, \psi \in \mathbf{R}^l$ such that

$$\nabla f^0(\hat{x}) + \sum_{j \in \underline{m}} \mu^j \nabla f^j(\hat{x}) + \frac{\partial g(\hat{x})^T}{\partial x}\psi = 0, \tag{11.2.13a}$$

$$\mu^j f^j(\hat{x}) = 0, \ \forall \ j \in \underline{m}, \tag{11.2.13b}$$

and, for some $m > 0$

$$\langle y, \frac{\partial^2 L(\hat{x})}{\partial x^2} y\rangle \geq m|y|^2, \ \ \forall \ y \in \bar{\mathbf{M}}_{IE}, \tag{11.2.13c}$$

where $L(x)$ is defined as in (11.2.11c) and

$$\bar{\mathbf{M}}_{IE} \triangleq \{ y \in \mathbf{R}^n \mid \frac{\partial g(\hat{x})}{\partial x} y = 0, \langle \nabla f^j(\hat{x}), y\rangle = 0 \ \forall \ j \in I(\hat{x}) \text{ such that } \mu^j > 0 \}. \tag{11.2.13d}$$

Then $\hat{x}$ is a local miniminizer for (11.2.10).

Proof : To obtain a contradiction, suppose that $\hat{x}$ is not a local minimizer for (11.2.10). Then there exists a sequence $\{ x_i \}_{i=0}^{\infty}$ such that $x_i \to \hat{x}$ as $i \to \infty$ with $g(x_i) = 0, f^j(x_i) \leq 0$ for all $j \in \underline{m}$, and $f^0(x_i) < f^0(\hat{x})$ for all $i \in \mathbf{N}$. We can write $x_i = \hat{x} + \delta_i h_i$, with $|h_i| = 1, \delta_i > 0$,so that $\delta_i \to 0$ as $i \to \infty$. Without loss of generality, we assume that $h_i \to \hat{h}$ as $i \to \infty$. Then we must have :

$$f^j(x_i) - f^j(\hat{x}) = \delta_i \int_0^1 \langle \nabla f^j(\hat{x} + s\delta_i h_i), h_i\rangle ds \leq 0, \ \ \forall \ i \in \mathbf{N}, \forall \ j \in \{ 0 \} \cup I(\hat{x}), \tag{11.2.14a}$$

and

$$g(x_i) = \delta_i \int_0^1 \frac{\partial g(\hat{x} + s\delta_i h_i)}{\partial x} h_i \, ds = 0, \ \ \forall \ i \in \mathbf{N}. \tag{11.2.14b}$$

Dividing (11.2.14a) - (11.2.14b) by δ_i and letting $i \to \infty$ we obtain that

$$\langle \nabla f^j(\hat{x}), \hat{h}\rangle \leq 0 \ \forall \ j \in \{ 0 \} \cup I(\hat{x}) \tag{11.2.15a}$$

and

$$\frac{\partial g(\hat{x})}{\partial x}\hat{h} = 0\,. \tag{11.2.15b}$$

Now, either $\hat{h} \in \tilde{M}_{IE}$ or not. If $\hat{h} \notin \tilde{M}_{IE}$, then there exists a $\hat{j} \in I(\hat{x})$ such that $\hat{\mu}^{\hat{j}} > 0$ and $\langle \nabla f^{\hat{j}}(\hat{x}), \hat{h}\rangle < 0$. Consequently, (11.2.13a) and (11.2.13b) yield that

$$0 = \langle \nabla f^0(\hat{x}), \hat{h}\rangle + \sum_{\substack{j \in I(\hat{x}) \\ \mu^j > 0}} \mu^j \langle \nabla f^j(\hat{x}), \hat{h}\rangle < 0, \tag{11.2.16}$$

which is clearly impossible. Hence we must assume that $\hat{h} \in \tilde{M}_{IE}$.

Next, we note that because $g(x_i) = 0$ and $\mu^j f^j(x_i) \le 0$ for all $j \in \mathbf{m}$, $L(x_i) \le f^0(x_i) < f^0(\hat{x}) = L(\hat{x})$. Hence expanding $L(x_i)$ about $\hat{x}$ to second order, we obtain that

$$L(x_i) - L(\hat{x}) = \delta_i \langle \nabla L(\hat{x}), h_i \rangle + \delta_i^2 \int_0^1 (1-s)\langle h_i, \frac{\partial^2 L(\hat{x} + s\delta h_i)}{\partial x^2} h_i\rangle ds < 0 \;\forall\; i \in \mathbf{N}\,. \tag{11.2.17a}$$

Since by (11.2.13a) $\nabla L(\hat{x}) = 0$, it follows from (11.2.17a) that

$$\int_0^1 (1-s)\langle h_i, \frac{\partial^2 L(\hat{x} + s\delta h_i)}{\partial x^2} h_i\rangle ds < 0, \;\forall\; i \in \mathbf{N}\,. \tag{11.2.17b}$$

Letting $i \to \infty$, we conclude from (11.2.17b) that

$$\langle \hat{h}, \frac{\partial^2 L(\hat{x})}{\partial x^2}\hat{h}\rangle \le 0\,. \tag{11.2.18}$$

Since this contradicts (11.2.13c), we see that our proof is complete. ■

Exercise 11.2.2 : Mimick the above proof to establish the second order sufficiency conditions stated below for minimax problems. Note that it does seem to be possible to deduce these conditions directly from Theorem 11.2.5, above.

Thereom 11.2.4 (SOSC for minimax): Consider the problem

$$\min_{x \in \mathbf{R}^n} \max_{j \in \mathbf{m}} f^j(x), \tag{11.2.19a}$$

where the functions $f^j : \mathbf{R}^n \to \mathbf{R}$ are twice continuously differentiable. Suppose that $\hat{x}$ is such that there exists a multiplier vector $\hat{\mu} \in \Sigma$ satisfying

$$\sum_{j=1}^{m} \hat{\mu}^j \nabla f^j(\hat{x}) = 0, \tag{11.2.19b}$$

$$\hat{\mu}^j (f^j(\hat{x}) - \psi(\hat{x})) = 0, \;\forall\; j \in \mathbf{m}, \tag{11.2.19c}$$

and, with the Lagrangian defined by $L(x) \triangleq \sum_{j=1}^{m} \hat{\mu}^j f^j(x)$,

$$\langle h, \frac{\partial^2 L(\hat{x})}{\partial x^2} h \rangle \geq c|h|^2, \ \forall \ h \in M_S, \tag{11.2.19d}$$

where $c > 0$ and

$$M_S \triangleq \{ h \in \mathbb{R}^n \mid \langle \nabla f^j(\hat{x}), h \rangle = 0, j \in \{ j \in I(\hat{x}) \mid \hat{\mu}^j > 0 \} \}. \tag{11.2.19f}$$

Then $\hat{x}$ is a local minimizer of $\psi(x) \triangleq \max_{j \in \underline{m}} f^j(x)$. ■

12. EXACT PENALTY FUNCTIONS, SENSITIVITY AND DUALITY

We will now present three results which can be explained using the the same geometric setting: the "$f - g$" diagram.

12.1 EXACT PENALTY FUNCTIONS

We begin with exact penalty functions which have many uses in optimization, and, in particular, in the solution of equality constrained optimization problems. For further details consult Han-Mangasarian [1]. Consider the problem

$$\min\{ f(x) \mid g(x) = 0 \}, \tag{12.1.1a}$$

where $f:\mathbb{R}^n \to \mathbb{R}$ and $g:\mathbb{R}^n \to \mathbb{R}^l$, with $l < n$ are twice locally Lipschitz continuously differentiable.

Definition 12.1.1 : For any $c > 0$, the function

$$f_c(x) \triangleq f(x) + c \max_{j \in l} |g^j(x)| \tag{12.1.1b}$$

will be called an *exact penalty function* for the problem (12.1.1a). ■

We begin with an heuristic exploration of the properties of exact penalty functions. To this end, we suppose that $l = 1$ and that $\hat{x}$ is optimal for (12.1.1a) Then, letting

$$F(x) \triangleq \begin{bmatrix} f(x) \\ g(x) \end{bmatrix}, \tag{12.1.2}$$

we can draw $F(\mathbb{R}^n)$, the image of $\mathbb{R}^n$ under $F(\cdot)$ in $\mathbb{R}^2$, as shown in Fig. 12.1.1.

Suppose that the slope of the boundary of $F(\mathbb{R}^n)$, at $(f(\hat{x}),0)$, is finite. Then the line tangent to the boundary at this point has the equation

$$f(x) + \psi g(x) = f(\hat{x}) , \tag{12.1.3a}$$

where $-\psi$ is the slope of the line. Since *locally* all of $F(\mathbb{R}^n)$ lies to one side of this line we get, to first order terms that for some $\rho > 0$,

$$f(\hat{x}) + \langle \nabla f(\hat{x}) + \psi \nabla g(\hat{x}), \delta x \rangle \geq f(\hat{x}), \quad \forall\ \delta x \in \mathbf{B}(\hat{x},\rho), \tag{12.1.3b}$$

This leads to the conclusion that

$$\nabla f(\hat{x}) + \psi \nabla g(\hat{x}) = 0 \tag{12.1.3c}$$

must hold, i.e., that ψ is the Lagrange multiplier at $\hat{x}$. Next, if $c > |\psi|$, then it follows from Fig. 12.1.1 that

---

[1] S-P. Han and O. L. Mangasarian, "Exact penalty Functions in Nonlinear Programming", *Mathematical Programming*, Vol.

$$\min_{x \in \mathbb{R}^n} \{ f(x) + c|g(x)| \} = f(\hat{x}), \tag{12.1.3d}$$

i.e., that there is an unconstrained, but nondifferentiable optimization problem with the same solution point $\hat{x}$ and optimal value $f(\hat{x})$ as (12.1.1a). In formalizing these observations, we shall need the following result.

Lemma 12.1.1 : Let Q be an $n \times n$ matrix and let G be an $l \times n$. Then the function $\psi:\mathbb{R}_+ \to \mathbb{R}$ defined by

$$\psi(\varepsilon) \triangleq \min\{ \langle h,Qh\rangle | \; |h| = 1, \; |Gh|_\infty \leq \varepsilon \} \tag{12.1.4}$$

is continuous at 0.

Proof : Suppose that $\varepsilon_i \to 0$ as $i \to \infty$. Then there exist h_i such that $|h_i| = 1$, $|Gh_i|_\infty \leq \varepsilon_i$, and $\psi(\varepsilon_i) = \langle h_i,Qh_i\rangle$ Without loss of generality, we may assume that $h_i \to \hat{h}$ as $i \to \infty$. Then, by continuity, $|\hat{h}| = 1$ and $|G\hat{h}|_\infty = 0$. Hence $\psi(\hat{\varepsilon}) \leq \langle \hat{h},Q\hat{h}\rangle$ Suppose that

$$\psi(\hat{\varepsilon}) < \langle \hat{h},Q\hat{h}\rangle. \tag{12.1.5a}$$

Then there exists an h^* such that $|h^*| = 1$, $|Gh^*|_\infty = 0$ and $\psi(\hat{\varepsilon}) = \langle h^*,Qh^*\rangle$ Now, because $|Gh^*|_\infty \leq \varepsilon i$ for all i, we must have that $\psi(\varepsilon_i) \leq \langle h^*,Qh^*\rangle$ for all i. But, from (12.1.5a) we get, by continuity, that

$$\langle h^*,Qh^*\rangle < \lim_{i \to \infty} \langle h_i,Qh_i\rangle \tag{12.1.5b}$$

which contradicts the optimality of the h_i for i sufficiently large. Hence our proof is complete. ■

Theorem 12.1.1 : (a) Suppose that for some $\hat{c} \geq 0$, a point $\hat{x} \in \mathbb{R}^n$ satisfies $g(\hat{x}) = 0$ and $\hat{x} \in \arg\min_{x \in \mathbb{R}^n} f_{\hat{c}}(x)$. Then $\hat{x}$ is optimal for (12.1.1a).

(b) Suppose that (i) $\hat{x} \in \mathbb{R}^n$ is such that $g(\hat{x}) = 0$, and (ii) there exists a $\psi \in \mathbb{R}^l$ satisfying

$$\nabla f(\hat{x}) + \frac{\partial g(\hat{x})^T}{\partial x}\psi = 0 \; . \tag{12.1.6a}$$

Then there exists a $\hat{c} \geq 0$ such that for all $c \geq \hat{c}$, $df_c(\hat{x};h) \geq 0$ for all $h \in \mathbb{R}^n$.

(c) Suppose that $\hat{x}$ is such that (i) $g(\hat{x}) = 0$, (ii) for some $\psi \in \mathbb{R}^l$ (12.1.6a) is satisfied, and (iii) there exists an $m > 0$, such that for $L(x) \triangleq f(x) + \langle \psi,g(x)\rangle$,

$$\langle y, \frac{\partial^2 L(\hat{x})}{\partial x^2} y\rangle \geq m|y|^2, \tag{12.1.6b}$$

for all y such that $\frac{\partial g(\hat{x})}{\partial x} y = 0$. Then there exists a $\hat{c} \geq 0$ such that for all $c \geq \hat{c}$, $\hat{x}$ is also a local

17, No.3, pp. 251-270, 1979.

miniminizer of $f_c(\cdot)$.

Proof : (a) Since $f_c(x) = f(x)$ for all x such that $g(x) = 0$, this part is obvious.

(b) First note that $f_c(\cdot)$ can be written alternatively in the form

$$f_c(x) = \max_{j \in \underline{2l}} \{ f(x) + cg^j(x) \}, \tag{12.1.7a}$$

where we define $g^{j+l}(x) = -g^j(x)$ for all $j \in \underline{l}$. Hence, since $g(\hat{x}) = 0$, for any $c \geq 0$, we have

$$df_c(\hat{x};h) = \langle \nabla f(x),h\rangle + c\max_{j \in \underline{l}} |\langle \nabla g^j(\hat{x}),h\rangle| . \tag{12.1.7b}$$

Because of (12.1.6a), we have for any $h \in \mathbb{R}^n$

$$\langle \nabla f(\hat{x}),h\rangle = -\sum_{j \in \underline{l}} \psi^j \langle \nabla g^j(\hat{x}),h\rangle. \tag{12.1.7c}$$

Hence, $df_c(\hat{x};h) = 0$ for all h such that $\frac{\partial g(\hat{x})}{\partial x} h = 0$. Furthermore, substituting from (12.1.7c) into (12.1.7b), we get

$$\begin{aligned} df_c(\hat{x};h) &= -\sum_{j \in \underline{l}} \psi^j \langle \nabla g^j(\hat{x}),h\rangle + c\max_{j \in \underline{l}} |\langle \nabla g^j(\hat{x}),h\rangle| \\ &\geq -\sum_{j \in \underline{l}} |\psi^j| \, |\langle \nabla g^j(\hat{x}),h\rangle| + c\max_{j \in \underline{l}} |\langle \nabla g^j(\hat{x}),h\rangle| \\ &\geq (c - \sum_{j \in \underline{l}} |\psi^j|) \max_{j \in \underline{l}} |\langle \nabla g^j(\hat{x}),h\rangle| . \end{aligned} \tag{12.1.7d}$$

Let $\hat{c} \geq \sum_{j \in \underline{l}} |\psi^j|$. Then we obtain that for all $c \geq \hat{c}$, $df_c(\hat{x};h) \geq 0$ for all $h \in \mathbb{R}^n$.

(c) Let $\hat{\rho} > 0$ and let $K \in (0,\infty)$ be a common Lipschitz constant for $\frac{\partial^2 f(\cdot)}{\partial x^2}$, $\frac{\partial^2 g^j(\cdot)}{\partial x^2}$, $j \in \underline{l}$, in $B(\hat{x},\hat{\rho})$. Then given any $h \in \mathbb{R}^n$ such that $|h| = 1$ and $t \in (0,\hat{\rho})$,

$$\begin{aligned} f(\hat{x} + th) - f(\hat{x}) &= t\langle \nabla f(\hat{x}),h\rangle + \frac{t^2}{2}\langle h, \frac{\partial^2 f(\hat{x})}{\partial x^2} h\rangle \\ &\quad + t^2 \int_0^1 (1-s)\langle h, \left[\frac{\partial^2 f(\hat{x} + sth)}{\partial x^2} - \frac{\partial^2 f(\hat{x})}{\partial x^2}\right] h\rangle ds \\ &\geq t\langle \nabla f(\hat{x}),h\rangle + \frac{t^2}{2}\langle h, \frac{\partial^2 f(\hat{x})}{\partial x^2} h\rangle - \frac{K}{2} t^3 . \end{aligned} \tag{12.1.8a}$$

Similarly, taking into account the fact that $g(\hat{x}) = 0$,

$$|g^j(\hat{x} + th)| \geq |t\langle \nabla g^j(\hat{x}),h\rangle + \frac{t^2}{2}\langle h, \frac{\partial^2 g^j(\hat{x})}{\partial x^2} h\rangle| - \frac{K}{2} t^3, \quad \forall\, j \in \underline{l} . \tag{12.1.8b}$$

Now let $\hat{c} > \sum_{j \in \underline{l}} |\psi^j|$, and suppose that $x = \hat{x} + th$, with $t \in (0,\hat{\rho})$ and $h \in \mathbb{R}^n$ such that $|h| = 1$ arbitrary. Then

$$f_c(\hat{x} + th) - f_c(\hat{x}) \geq t \langle \nabla f(\hat{x}), h \rangle + \frac{t^2}{2} \langle h, \frac{\partial^2 f(\hat{x})}{\partial x^2} h \rangle$$

$$+ c \max_{j \in \underline{l}} | t \langle \nabla g^j(\hat{x}), h \rangle + \frac{t^2}{2} \langle h, \frac{\partial^2 g^j(\hat{x})}{\partial x^2} h \rangle | - K' t^3, \tag{12.1.8c}$$

where $K' = K(1 + c)$. Adding and subtracting

$$t \sum_{j \in \underline{l}} \psi^j \langle \nabla g^j(\hat{x}), h \rangle + \frac{t^2}{2} \sum_{j \in \underline{l}} \psi^j \langle h, \frac{\partial^2 g^j(\hat{x})}{\partial x^2} h \rangle \tag{12.1.8d}$$

to the right hand side of (12.1.8c), we obtain that

$$f_c(\hat{x} + th) - f_c(\hat{x}) \geq t \langle \nabla f(\hat{x}) + \sum_{j \in \underline{l}} \psi^j \nabla g^j(\hat{x}), h \rangle + \frac{t^2}{2} \langle h, \frac{\partial^2 L(\hat{x})}{\partial x^2} h \rangle$$

$$+ [c - \sum_{j \in \underline{l}} |\psi^j|] \max_{j \in \underline{l}} | t \langle \nabla g^j(\hat{x}), h \rangle + \frac{t^2}{2} \langle h, \frac{\partial^2 g_j(\hat{x})}{\partial x^2} h \rangle | - K' t^3 . \tag{12.1.8e}$$

To complete our demonstration that $\hat{x}$ is a local minimizer for $f_c(x)$, with $c \geq \hat{c}$, we must show that there exists a $\hat{t} \in (0,\hat{\rho}]$ such that $f_c(x + th) \geq f_c(\hat{x})$ for all $t \in [0,\hat{t}]$ and all $h \in \mathbb{R}^n$ such that $|h| = 1$.

Now, by assumption, there exists an $m > 0$ such that $\langle h, \frac{\partial^2 L(\hat{x})}{\partial x^2} h \rangle \geq m$, for all $h \in \mathbb{R}^n$ such that $|h| = 1$ and $\partial g(\hat{x}) \partial x h = 0$. Hence, by Lemma 12.1.1, there exists an $\varepsilon > 0$ such that $\langle h, \frac{\partial^2 L(\hat{x})}{\partial x^2} h \rangle \geq m/2$ for all $h \in \mathbb{R}^n$ such that $|h| = 1$ and $|\frac{\partial g(\hat{x})}{\partial x} h|_\infty \leq \varepsilon$. Thus, suppose that $h \in \mathbb{R}^n$ is such that $|h| = 1$ and $|\frac{\partial g(\hat{x})}{\partial x} h|_\infty \leq \varepsilon$. Then , since $\langle \nabla f(\hat{x}) + \sum_{j \in \underline{l}} \psi^j \nabla g^j(\hat{x}), h \rangle = 0$, if we set $t' = m/4K$, it follows from (12.1.8e) that $f_c(x + th) \geq f_c(\hat{x})$ for all $t \in [0,t']$ and $c \geq \hat{c}$, as before.

Next, suppose that $h \in \mathbb{R}^n$ is such that $|h| = 1$ and $|\frac{\partial g(\hat{x})}{\partial x} h|_\infty > \varepsilon$. Then there exists a $t'' \in (0,t')$ such that $t \max_{j \in \underline{l}} \max_{|h|=1} | \langle h, \partial^2 g^j(\hat{x})/\partial x^2 h \rangle | = t \max_{j \in \underline{l}} |\partial^2 g^j(\hat{x})/\partial x^2| \leq \varepsilon$ for all $t \in [0,t'']$. Hence for all $t \in [0,t'']$,

$$\max_{j \in \underline{l}} | \langle \nabla g^j(\hat{x}), h \rangle + \frac{t}{2} \langle h, \frac{\partial^2 g^j(\hat{x})}{\partial x^2} h \rangle | \geq \frac{1}{2} \max_{j \in \underline{l}} | \langle \nabla g^j(\hat{x}), h \rangle | \tag{12.1.8f}$$

and hence, since $\langle \nabla f(\hat{x}) + \sum_{j \in \underline{l}} \psi^j \nabla g^j(\hat{x}), h \rangle = 0$,

$$f_c(\hat{x}+th)-f_c(\hat{x}) \geq \frac{t^2}{2}\langle h,\frac{\partial^2 L(\hat{x})}{\partial x^2}h\rangle+\frac{t}{2}\left[c-\sum_{j\in \underline{l}}|\psi^j|\right]\max_{j\in \underline{l}}|\langle \nabla g^j(\hat{x}),h\rangle|-Kt^3$$

$$\geq \frac{t^2}{2}\langle h,\frac{\partial^2 L(\hat{x})}{\partial x^2}h\rangle+\frac{t}{2}[c-\sum_{j\in \underline{l}}|\psi^j|]\varepsilon-Kt^3\,. \tag{12.1.8g}$$

Since $\max_{|h|=1}\langle h,\frac{\partial^2 L(\hat{x})}{\partial x^2}h\rangle$ is finite, there exists a $\hat{t}\in(0,t')$ such that $f_c(\hat{x}+th)-f_c(\hat{x})\geq 0$ for all $t\in[0,\hat{t}]$, which completes our proof of that $\hat{x}$ is also a local miniminizer of $f_c(\cdot)$, for all $c\geq\hat{c}$.

This completes the proof of the theorem. ■

Exercise 12.1.1 : Consider the exact penalty function $f_c(x)$ defined in (12.1.1b). Show that if x^* is such that $g(x^*)\neq 0$ and $\partial g(x^*)/\partial x$ has maximum row rank, then there exists a $c^*>0$ such that $0\notin\partial f_c(x^*)$ for all $c\geq c^*$. ■

Exercise 12.1.2 : Consider the problem

$$\min\{\, f^0(x)\mid f^j(x)\leq 0,\ j=1,2,\ldots,m,\ g(x)=0\,\}, \tag{12.1.9}$$

where $f^j:\mathbb{R}^n\to\mathbb{R}$, $j=0,1,\ldots,m$, and $g:\mathbb{R}^n\to\mathbb{R}^l$ are twice continuously differentiable, and $l<n$. Prove the following theorem:

Theorem 12.1.2 : For any $c\geq 0$, let

$$f_c(x)\triangleq f^0(x)+c\max_{\substack{j\in \underline{l}\\ j\in \underline{m}}}\{\,|g^j(x)|\,,f^j(x)_+\,\}, \tag{12.1.10a}$$

where $f^j(x)_+\triangleq\max\{\,f^j(x)\,,0\,\}$.

(a) Suppose that for some $\hat{c}\geq 0$, a point $\hat{x}\in\mathbb{R}^n$ satisfies $g(\hat{x})=0$, $f^j(\hat{x})\leq 0$, for all $j\in\underline{m}$, and $\hat{x}\in\arg\min_{x\in\mathbb{R}^n}f_c(x)$. Then $\hat{x}$ is optimal for (12.1.9).

(b) Suppose that $\hat{x}\in\mathbb{R}^n$ is such that (i) $g(\hat{x})=0$, $f^j(\hat{x})\leq 0$, for all $j\in\underline{m}$, and (ii) there exist multiplier vectors $\psi\in\mathbb{R}^l$ and $\mu\in\mathbb{R}^m$, with $\mu\geq 0$, satisfying

$$\nabla f^0(\hat{x})+\frac{\partial g(\hat{x})^T}{\partial x}\psi+\frac{\partial f(\hat{x})^T}{\partial x}\mu=0\,, \tag{12.1.10b}$$

where $f=(f^1\,f^2.\cdots f^m)^T$, and

$$\mu^j f^j(\hat{x})=0,\ \text{for } j=1,2,\ldots,m\,. \tag{12.1.10c}$$

Then there exists a $\hat{c}\geq 0$ such that for all $c\geq\hat{c}$, $df_c(\hat{x};h)\geq 0$ for all $h\in\mathbb{R}^n$.

(c) Suppose that $\hat{x}$ is such that (i) $g(\hat{x})=0$ and $f^j(\hat{x})\leq 0$ for all $j\in\underline{m}$, (ii) for some multipliers $\psi\in\mathbb{R}^l$, $\mu\in\mathbb{R}^m$, $\mu\geq 0$, (12.1.10b,c) are satisfied, (iii) there exists a $b>0$ such that

$$\min_{h \in M_E} \max_{j \in I(\hat{x})} \langle \nabla f^j(\hat{x}), h \rangle \geq b|h|, \tag{12.1.10d}$$

where $M_E \triangleq \{ h \in \mathbb{R}^n \mid [\partial g(\hat{x})/\partial x]h = 0 \}$, and (iv) there exists an $m > 0$, such that for $L(x) \triangleq f(x) + \langle \psi, g(x) \rangle + \langle \mu, f(x) \rangle$

$$\langle y, \frac{\partial^2 L(\hat{x})}{\partial x^2} y \rangle \geq m|y|^2, \tag{12.1.10d}$$

for all y such that $\frac{\partial g(\hat{x})}{\partial x} y = 0$ and $\langle \nabla f^j(\hat{x}), y \rangle = 0$, for all $j \in \underline{m}$ such that $f^j(\hat{x}) = 0$ and $\mu^j > 0$. Then there exists a $\hat{c} \geq 0$ such that for all $c \geq \hat{c}$, $\hat{x}$ is also a local miniminizer of $f_c(\cdot)$. ■

Exact penalty functions can be combined with the Minimax Algorithm 9.3.2, to produce an algorithm for solving equality constrained optimization problems, of the form (12.1.1a), as follows:

Algorithm 12.1.1: (*Armijo Line Search*).

Parameters : $\alpha, \beta \in (0,1)$, $c_{-1} > 0$, $\delta > 0$.

Data : $x_0 \in \mathbb{R}^n$

Step 0 : Set $i = 0$.

Step 1 : Compute the multiplier vector

$$\psi_i \triangleq -\left[\frac{\partial g(x_i)^T}{\partial x} \frac{\partial g(x_i)}{\partial x}\right]^{-1} \frac{\partial g(x_i)}{\partial x} \nabla f(x_i), \tag{12.1.11a}$$

which solves the multiplier problem

$$\min_{\psi \in \mathbb{R}^l} |\nabla f(x_i) + \frac{\partial g(x_i)^T}{\partial x} \psi|^2 . \tag{12.1.11b}$$

Step 2 : If $c_{i-1} \geq \sum_{j=1}^{l} |\psi_i^j|$, set $c_i = c_{i-1}$, else set $c_i = \sum_{j=1}^{l} |\psi_i^j| + \delta$.

Step 3 : Compute the *search direction*

$$\begin{aligned} h_i = h(x_i) \triangleq \arg \min_{h \in \mathbb{R}^n} \max_{j \in \underline{m}} \max \{ & f(x_i) + c_i g^j(x_i) - f_{c_i}(x_i) \\ & + \langle \nabla f(x_i), h \rangle + c_i \langle \nabla g^j(x_i), h \rangle + \tfrac{1}{2}|h|^2, \\ & f(x_i) - c_i g^j(x_i) - f_{c_i}(x_i) \\ & + \langle \nabla f(x_i), h \rangle - c_i \langle \nabla g^j(x_i), h \rangle + \tfrac{1}{2}|h|^2 \}, \end{aligned} \tag{12.1.11c}$$

(where $f_{c_i}(x_i) = f(x_i) + \max_{j \in \underline{l}} |g^j(x_i)|$), and the value of the optimality function $\theta_{c_i}(x_i)$:

$$\theta_i = \theta_{c_i}(x_i) \triangleq \min_{h \in \mathbb{R}^n} \max_{j \in \underline{m}} \max \{ f(x_i) + c_i g^j(x_i) - f_{c_i}(x_i)$$

$$+ \langle \nabla f(x_i), h \rangle + c_i \langle \nabla g^j(x_i), h \rangle + \tfrac{1}{2}|h|^2,$$

$$f(x_i) - c_i g^j(x_i) - f_{c_i}(x_i)$$

$$+ \langle \nabla f(x_i), h \rangle - c_i \langle \nabla g^j(x_i), h \rangle + \tfrac{1}{2}|h|^2 \}. \qquad (12.1.11d)$$

Step 4 : Compute the *step size*

$$\lambda_i = \arg \max_{k \in \mathbf{N}} \{ \beta^k \mid f_{c_i}(x_i + \beta^k h_i) - f_{c_i}(x_i) - \beta^k \alpha \theta_i \leq 0 \}. \qquad (12.1.11e)$$

Step 5 : *Update:* set

$$x_{i+1} = x_i + \lambda_i h_i, \qquad 12.1.11f)$$

replace i by $i + 1$ and go to Step 1. ■

The properties of Algorithm 12.1.1 can be summarized as follows[2]:

Theorem 12.1.3 : Consider problem (12.1.1a), and in addition to the assumptions stated, assume that $\partial g(x)/\partial x$ has full row rank for all $x \in \mathbf{R}^n$ (so that $\psi(x)$ is well defined by (12.1.11a)).

(a) If Algorithm 12.1.1 constructs and infinite sequence $\{ x_i \}_{i=0}^{\infty}$, increasing c_i a finite number of times only (at $i_1, i_2, \ldots, i_N$), then any accumulation point $\hat{x}$ of $\{ x_i \}_{i=0}^{\infty}$ satisfies the first order optimality condition $g(\hat{x}) = 0$, and $\nabla f(\hat{x}) + [\partial g(\hat{x})/\partial x]^T \psi(\hat{x}) = 0$.

(b) If Algorithm 12.1.1 constructs and infinite sequence $\{ x_i \}_{i=0}^{\infty}$, increasing c_i an infinite number of times, so that $c_i \to \infty$ as $i \to \infty$, at $i_1, i_2, i_3, \ldots$, then the subsequence $\{ x_{i_k} \}_{k=1}^{\infty}$, at which c_i was increased, has no accumulation points. ■

Exercise 12.1.3 : Modify the proofs in Mayne-Polak to construct a proof for Theorem 12.1.3. ■

12.2. SENSITIVITY: EQUALITY CONSTRAINED PROBLEMS

We now turn to the second result that we were able to deduce heuristically from the f–g diagram in Lecture 12.1, viz., that the optimality condition multipliers are related to the sensitivity of the value function with respect to constraint perturbations.

First we shall consider equality constrained optimization problems with parameters and we shall show that under certain assumptions, the solutions of these problems are differentiable functions of the parameters. Furthermore, we shall see that the multipliers associated with the solutions of parametrized problems are, in fact the derivatives of the value function.

Thus consider the parametrized optimization problem

---

[2] These properties can be established by adapting the proofs in: D. Q. Mayne and E. Polak, "Feasible Directions Algorithms for Optimization Problems with Equalitiy and Inequality Constraints", *Mathematical Programming*, Vol. 11, No.1 , pp 67-81, 1976.

$$v(b) \triangleq \min\{ f(x) \mid g(x) = b\} \tag{12.2.1}$$

where $f:\mathbf{R}^n \to \mathbf{R}$, $g:\mathbf{R}^n \to \mathbf{R}^l$ are twice continuously differentiable and $\partial g(x)/\partial x$ has maximum row rank for all $x \in \mathbf{R}^n$ and $b \in \mathbf{R}^l$ is a parameter vector. The function $v:\mathbf{R}^l \to \mathbf{R}$ will be referred to as the *value function*.

The solutions of (12.2.1) obviously depend on b. We shall now show that under certain conditions the solution $x(b)$ is unique and is a differentiable function of the parameter $b \in \mathbf{R}^l$. We shall see that this fact depends crucially on the applicability of the implicit function theorem.

Theorem 12.2.1 : Suppose that $x(0)$ solves (12.2.1) for $b = 0$ and that second order sufficiency conditions are satisfied by $x(0) \in \mathbf{R}^n$ and the corresponding multiplier $\psi(0) \in \mathbf{R}^l$, viz. for

$$L(x,\psi) \triangleq f(x) + \langle \psi, f(x)\rangle \ , \tag{12.2.2a}$$

$$\nabla_x L(x(0),\psi(0)) = 0, \tag{12.2.2b}$$

$$\nabla_\psi L(x(0),\psi(0)) = g(x(0)) = 0, \tag{12.2.2c}$$

and there exists an $m > 0$ such that

$$\langle y, \frac{\partial^2 L(x(0)),\psi(0)}{\partial x^2} y\rangle \geq m|y|^2, \ \forall\ y \in \{y' \mid \frac{\partial g(x(0))}{\partial x} y' = 0\} \ . \tag{12.2.2d}$$

Then $(x(b),\psi(b))$, the solution and corresponding multiplier of (12.2.1), are diffferentiable functions of b, in a neighborhood of $b = 0$.

Proof : By assumption, $x(0)$ is a local solution of problem (12.2.1) at $b = 0$ and $\psi(0)$ is the corresponding multiplier (it must be unique because of the rank assumption on $\frac{\partial g(x)}{\partial x}$). Next, the pair $(x(b),\psi(b))$ must satisfy the necessary conditions

$$\nabla_x L(x,\psi) \ = 0, \tag{12.2.3a}$$

$$g(x) - b \ = 0 \ . \tag{12.2.3b}$$

This is a set of equations which has a solution $(x(0),\psi(0),0)$ at $b = 0$. By the Implicit Function Theorem, $x(b)$, $\psi(b)$ are differentiable functions of b on a neighborhood of $b = 0$, if the matrix H, defined below, is nonsingular:

$$H \triangleq \begin{bmatrix} \frac{\partial^2 L(x(0),\psi(0))}{\partial x^2} & \frac{\partial g(x(0))}{\partial x} \\ \frac{\partial g(x(0))^T}{\partial x} & 0 \end{bmatrix} . \tag{12.2.4}$$

Suppose that $z \triangleq (\overline{x},\overline{\psi})$ is such that $Hz = 0$. Then we have that

$$\frac{\partial g(x(0))}{\partial x}\overline{x} = 0, \tag{12.2.5a}$$

and

$$\frac{\partial^2 L(x(0),\psi(0))}{\partial x^2}\bar{x} + \frac{\partial g(x(0))^T}{\partial x}\bar{\psi} = 0\,. \tag{12.2.5b}$$

Taking the scalar product of (12.2.5b) with $\bar{x}$, we get, because of (12.2.5a) that

$$\langle \bar{x}, \frac{\partial^2 L(x(0)),\psi(0))}{\partial x^2}\bar{x}\rangle = 0\,. \tag{12.2.5c}$$

In view of (12.2.2d), we conclude that $\bar{x} = 0$. Since $\frac{\partial g(x(0))^T}{\partial x}$ is of maximum rank, (12.2.5b) leads to the conclusion that $\bar{\psi} = 0$. Since $Hz = 0$ is possible only for $z = 0$, H must be nonsingular. To complete our proof we must show that $x(b)$ not only satisfies necessary conditions of optimality, for b in a neighborhood of $b = 0$, but also sufficient conditions. For this we need the following result.

For any $x \in \mathbb{R}^n$, let

$$\mathrm{N}(x) \triangleq \{y \in \mathbb{R}^n \mid \frac{\partial g(x)}{\partial x}y = 0, |y| = 1\}\,. \tag{12.2.6}$$

We will show that given $\hat{x} \triangleq x(0)$ and any $\delta > 0$, there exists a $\hat{\rho} > 0$ such that for all $x \in \mathrm{B}(\hat{x},\hat{\rho})$, if $y \in \mathrm{N}(x)$, then there exists a $\hat{y} \in \mathrm{N}(\hat{x})$ such that $|y - \hat{y}| < \delta$. For suppose that this is false. Then we must have a sequence $x_i \to \hat{x}$ as $i \to \infty$ and a corresponding sequence $y_i \in \mathrm{N}(x_i)$, $i = 0, 1, 2,....$, such that $|y_i - y| \geq \delta$ for all $y \in \mathrm{N}(\hat{x})$.

Now, since $|y_i| = 1$, there must be a subsequence $\{y_i\}_{i \in \mathcal{K}}$ such that $y_i \xrightarrow{\mathcal{K}} y^*$. By continuity of $\frac{\partial g(x)}{\partial x}$ we get that $y^* \in \mathrm{N}(\hat{x})$, which leads to a contradiction.

Hence, since $\mathrm{N}(x) \to \mathrm{N}(\hat{x})$ as $x \to \hat{x}$, we obtain from (12.2.2d), by the continuity of $\frac{\partial^2 L(x(b),\psi(b))}{\partial x^2}$ that there exists a neighborhood of $b = 0$ such that

$$\langle y, \frac{\partial^2 L(x(b),\psi(b))}{\partial x^2}y\rangle \geq m/2 \tag{12.2.7}$$

for all $y \in \mathrm{N}(x(b))$, i.e., $x(b)$ solves (12.2). ■

Since under the conditions of Theorem 12.2 $x(b),\psi(b)$ are differentiable functions in a neighborhood of $b = 0$, we find that $v(b) = f(x(b))$ is also differentable and

$$\frac{\partial v(b)}{\partial b}\Big|_{b=0} = \frac{\partial f(x(b))}{\partial x}\frac{\partial x(b)}{\partial b}\Big|_{b=0}\,. \tag{12.2.8}$$

Now, from (12.2.3b),

$$\frac{\partial g(x(b))}{\partial x}\frac{\partial x(b)}{\partial b} - I = 0 \tag{12.2.9a}$$

and from (12.2.3a)

$$\frac{\partial f(x(b))}{\partial x} = -\psi(b)^T \frac{\partial g(x(b))}{\partial x} . \tag{12.2.9b}$$

Substituting from (12.2.9a), (12.2.9b) into (12.2.8), we obtain

$$\frac{\partial v(b)}{\partial b}\Big|_{b=0} = -\psi(0)^T . \tag{12.20}$$

We have thus proved the following result.

Theorem 12.2.2 : Under the conditions of Theorem 12.2.1, the value function $v(b)$ is differentiable at $b = 0$ and $\frac{\partial v(b)}{\partial b}\Big|_{b=0} = -\psi(0)^T$. ■

This result could have been anticipated from the diagrams which are drawn for the illustration of exact penalty functions as shown in Fig. 12.1.1.

Exercise 12.2.1 : Consider the problem

$$v(b) \triangleq \min\{f(x) \mid g(x,b) = 0\} \tag{12.1.21}$$

where $f:\mathbb{R}^n \to \mathbb{R}$, $g:\mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^l$ are all twice continuously differentiable. Develop conditions for $v(b)$ to be differentiable at $b = \hat{b}$ and obtain a formula for $\frac{\partial v(b)}{\partial b}$. ■

When inequalities are present, the situation becomes somewhat more complicated.

12.3. SENSITIVITY: EQUALITY AND INEQUALITY CONSTRAINED PROBLEMS

Next we shall determine the sensitivity of the value functions of a problem with both equality and inequality constraints:

$$v(b) = \min\{f^0(x) \mid f(x) \le b_1, g(x) = b_2\}, \tag{12.3.1}$$

where $f^0:\mathbb{R}^n \to \mathbb{R}$, $f:\mathbb{R}^n \to \mathbb{R}^m$, $g:\mathbb{R}^n \to \mathbb{R}^l$ are all twice continuously differentiable; $b_1 \in \mathbb{R}^l$, $b_2 \in \mathbb{R}^m$ and $b \triangleq (b_1,b_2)$.

For any $b = (b_1,b_2) \in \mathbb{R}^m \times \mathbb{R}^l$, let $\mathbf{F}(b) \triangleq \{x \in \mathbb{R}^n \mid f(x) \le b_1, g(x) = b_2\}$ and let $\mathbf{B} \triangleq \{(b_1,b_2) \in \mathbb{R}^m \times \mathbb{R}^l \mid \mathbf{F}(b) \neq \emptyset\}$. For any $b \in \mathbf{B}$, we shall denote by $x(b)$ a solution of (12.3.1) and we define

$$I(b) \triangleq \{j \in \underline{m} \mid f^j(x(b)) = b_1^j\}, \tag{12.3.2a}$$

$$M(x(b)) \triangleq \{y \mid \frac{\partial g(x(b))}{\partial x} y = 0, \langle \nabla f^j(x(b)), y\rangle = 0, j \in I(b)\} . \tag{12.3.2b}$$

For the equalities only case, (12.2.1), we saw that the solution $x(b)$ of (12.2.1) was differentiable at $b = 0$ if $\frac{\partial g(x(0))}{\partial x}$ had maximum row rank and second order sufficiency conditions were satisfied at $x(0)$. In the case of (12.3.1) we need slightly stronger conditions.

Theorem 12.3.1 : Suppose that $x(0)$ is a local solution of (12.3.1) and that

(i) The gradients $\{\nabla g^j(x(0))\}_{j \in \underline{m}}$ together with the gradients $\{\nabla f^j(x(0)\}_{j \in I(0)}$ are linearly independent.

(ii) For $L(x,\mu,\psi) \triangleq f^0(x) + \langle \mu, f(x)\rangle + \langle \psi, g(x)\rangle$, there exist $\mu(0) \in \mathbf{R}^m$, $\psi(0) \in \mathbf{R}^l$ and an $m > 0$ such that

$$\nabla_x L(x(0),\mu(0),\psi(0)) = 0, \tag{12.3.3a}$$

$$\mu^j(0) f^j(x(0)) = 0, \forall\, j \in \underline{m}, \tag{12.3.3b}$$

$$\mu^j(0) > 0, \forall\, j \in I(0), \tag{12.3.3c}$$

$$\langle y, \frac{\partial^2 l(x(0),\mu(0),\psi(0))}{\partial x^2} y\rangle \geq m|y|^2, \forall\, y \in M(x(0)) . \tag{12.3.3d}$$

Then there exists a differentiable function $x(b)$, together with corresponding differentiable multiplier functions $\mu(b)$, $\psi(b)$, all defined on a neighborhood of $b = 0$, which solves (12.3.1).

Proof : To simplify notation (which can always be obtained simply by renumbering the functions), suppose that $I(0) = \underline{k}$, with $k \leq m$. Let $\bar{f}: \mathbf{R}^n \to \mathbf{R}^k$ be defined by $\bar{f}^j(x) = f^j(x)$ for all $j \in \underline{k}$, let $\Pi \in \mathbf{R}^k$ and let $\bar{L}(x,\Pi,\psi) \triangleq f^0(x) + \langle \Pi, \bar{f}(x)\rangle + \langle \psi, g(x)\rangle$. Now consider the system of equations:

$$\nabla_x \bar{L}(x,\bar{\Pi},\psi) = 0, \tag{12.3.4a}$$

$$\Pi^j(\bar{f}^j(x) - b_1^j) = 0, \forall\, j \in \underline{k}, \tag{12.3.4b}$$

$$g(x) = b_2 . \tag{12.3.4c}$$

Let $M(\Pi) \triangleq \underset{j \in \underline{k}}{\text{diag}}(\Pi^j)$ and let $F(x) = \underset{j \in \underline{k}}{\text{diag}}(f^j(x))$. Then, by the implicit function theorem, the system (12.3.4a - c) has a differentiable solution $x(b)$, $\Pi(b)$, $\psi(b)$, in a neighborhood of $b = 0$, if the matrix

$$H \triangleq \begin{bmatrix} \dfrac{\partial^2 \bar{L}(x(0),\bar{\Pi}(0),\psi(0))}{\partial x^2} & \dfrac{\partial \bar{f}(x(0))^T}{\partial x} & \dfrac{\partial g(x(0))^T}{\partial x} \\ M(\bar{u}(0))\dfrac{\partial \bar{f}(x(0))}{\partial x} & F(x(0)) & 0 \\ \dfrac{\partial g(x(0))}{\partial x} & 0 & 0 \end{bmatrix}, \tag{12.3.4d}$$

is nonsingular. Since $f^j(x(0)) = 0$ for all $j \in I(0)$, we see that $F(x(0)) = 0$. Next, because of (12.3.2c), $M(\mu(0))$ is nonsingular. Let $z \triangleq (u,v,w) \in \mathbf{R}^n \times \mathbf{R}^k \times \mathbf{R}^k$ be such that $Hz = 0$. Then we have

$$\frac{\partial g(x(0))}{\partial x} u = 0, \tag{12.3.5a}$$

$$M(\Pi(0))\frac{\partial \bar{f}(x(0))}{\partial x} u = 0, \tag{12.3.5b}$$

$$\frac{\partial^2 L(x(0),\mu(0),\psi(0)}{\partial x^2}u + \frac{\partial \overline{f}(x(0)))^T}{\partial x}v + \frac{\partial g(x(0))^T}{\partial x}w = 0\,. \tag{12.3.5c}$$

Taking the scalar product of (12.3.5c) with u, we get because of (12.3.5a, b) that

$$\langle u, \frac{\partial^2 L(x(0),\mu(0),\psi(0))}{\partial x^2}u\rangle = 0\,. \tag{12.3.6}$$

Because of (12.3.3d), (12.3.5a,b) and (12.3.6) imply that $u = 0$. Since, by assumption, the matrix $\left[\frac{\partial \overline{f}(x(0))^T}{\partial x} \mid \frac{\partial g(x(0))^T}{\partial x}\right]$ has linearly independent columns, it follows from (12.3.5c) that $(v,w) = 0$. Hence, we see that H is nonsingular, and therefore the differentiable functions $x(b),\overline{\mu}(b),\psi(b)$ exist in a neighborhood of $b = 0$ and satisfy (12.3.3a - c). Now, since $\mu^j(0) > 0$ for all $j \in \underline{k}$, by continuity of $\overline{\mu}(\cdot)$, there exists a $\rho_1 > 0$ such that $\overline{\mu}^j(b) > 0$ for all $|b| \le \rho_1$, and $j \in \underline{k}$. Hence, from (12.3.4b), $\overline{f}^j(x(b)) = b^j$, for all $j \in \underline{k}$ and $|b| \le \rho_1$. Next, since $f^j(x(0)) < 0$ for all $j \notin \underline{k}$, there exists a $\rho_2 \in (0,\rho_1)$ such that $f^j(x(b)) < b_1^j$ for all $j \notin \underline{k}$ and $|b| \le \rho_2$. Therefore, for all $b = (b_1,b_2)$ such that $|b| \le \rho_2$, there exists a continuous function $x(b)$ solving (12.3.4a - c) such that

$$\left.\begin{aligned} f^j(x(b)) &\le b_1 \quad \forall\, j \in \underline{m} \\ g(x(b)) &= b_2 \end{aligned}\right\} \tag{12.3.7}$$

i.e., $x(b) \in \mathbf{F}(b)$ for all $|b| \le \rho_2$. Next, because by (12.3.4a, 4b)

$$\nabla_x \overline{L}(x(b),\overline{\mu}(b),\psi(b)) = 0, \tag{12.3.8a}$$

and hence $x(b)$, $\overline{\mu}(b)$, $\psi(b)$ satisfy first order conditions. Referring to (12.3.2d), it follows by continuity that there exists a $\hat{\rho} \in (0,\rho_2)$ such that

$$\langle y, \frac{\partial^2 \overline{L}(x(b),\overline{\mu}(b),\psi(b))}{\partial x^2}y\rangle \ge \frac{m}{2}|y|^2, \forall\, y \in M(x(b))\,. \tag{12.3.8b}$$

But this implies that for $|b| \le \hat{\rho}$, $x(b)$ together with $\mu(b) = (\overline{\mu}(b),0)$ and $\psi(b)$, satisfy second order sufficiency conditions and hence $x(b)$ is a local solution of (12.3.1). ■

Since $x(b)$ is differentiable at $b = 0$, $v(b)$ is differentable at $b = 0$ and

$$\frac{\partial v(b)}{\partial b}\Big|_{b=0} = \frac{\partial f^0(x(b))}{\partial x}\frac{\partial x(b))}{\partial b}\Big|_{b=0}\,. \tag{12.3.9}$$

Now, because of (12.3.8a),

$$\frac{\partial f^0(x(0))}{\partial x} = -\left[\overline{\mu}^T(0)\frac{\partial f(x(0))}{\partial x} + \psi(0)^T\frac{\partial g(x(0))}{\partial x}\right]. \tag{12.3.10}$$

From (12.3.3c) we have that

$$\frac{\partial g(x(0))}{\partial x}\frac{\partial x(0))}{\partial b} = [0 \mid I], \tag{12.3.11a}$$

and from (12.3.3b) we have that

$$\langle \Pi(b) f(x(b)) - b_1 \rangle = 0, \tag{12.3.11b}$$

so that

$$\Pi(0)^T \left[\frac{\partial f(x(0))}{\partial x} \frac{\partial x(0)}{\partial b} - [I \mid 0] \right] + \bar{f}(x(0))^T \frac{\partial \Pi(0)}{\partial b} = 0 . \tag{12.3.11c}$$

Substiting into (12.3.9) we finally get

$$\frac{\partial v(b)}{\partial b} \Big|_{b=0} = - (\mu(0)^T \mid \psi(0))^T) . \tag{12.3.12}$$

We can summarize our findings in the form of a theorem:

Theorem 12.3.2 : Under the conditions of Theorem 12.3.1, the value function $v(b)$ is differentiable at $b = 0$ and $\frac{\partial v(b)}{\partial b} \Big|_{b=0} = - (\mu(0)^T \mid \psi(0))^T)$. ■

12.4. DUALITY

The final result that we are going to obtaing from $f - g$ type diagrams deals with duals of optimization problems. Thus, consider the parametrized problem:

$$\mathbf{P} : \quad \min\{ f^0(x) \mid f^j(x) \le b^j, j = 1, \cdots, m, x \in \mathbf{X} \}, \tag{12.4.1}$$

where the set $\mathbf{X} \subset \mathbf{R}^n$ is convex, and the functions $f^0 : \mathbf{R}^n \to \mathbf{R}$, and $f^j : \mathbf{R}^n$, $j = 0,1, \cdots, m$, are convex (and hence continuous) on $\mathbf{X} \to \mathbf{R}^m$. We shall use the notation $f = (f^1, \ldots f^m)^T$.

Assumption 12.4.1 : (i) There exists an $x_0 \in \mathbf{X}$ such that $f^j(x_0) < 0$ for all $j \in \underline{m}$; (ii) for every $b \in \mathbf{R}^m$ such that the set $\{ x \mid f(x) \le b, x \in \mathbf{X} \}$ is nonempty, the minimum in (12.4.1) is achieved. ■

For any $b \in \mathbf{R}^m$, we define the feasible set

$$\mathbf{F}_b \triangleq \{ x \mid f(x) \le b, x \in \mathbf{X} \}, \tag{12.4.2a}$$

and we define the set

$$\mathbf{B} \triangleq \{ b \in \mathbf{R}^m \mid \mathbf{F}_b \neq \emptyset \}. \tag{12.4.2b}$$

Note that by Assumtion 12.4.1 (i), $0 \in \mathbf{B}$.

Finally, we define the value function $v_p : \mathbf{R}^m \to \mathbf{R}$ of the *primal problem* by

$$v_p(b) = \inf_{x \in \mathbf{X}} \sup_{\mu \ge 0} (f^0(x) + \langle \mu, f(x) - b \rangle) , \tag{12.4.3a}$$

and we define the value function $v_d : \mathbf{R}^m \to \mathbf{R}$ of the *dual problem* by

$$v_d(b) = \sup_{\mu \ge 0} \inf_{x \in \mathbf{X}} (f^0(x) + \langle \mu, f(x) - b \rangle) . \tag{12.4.3b}$$

We now proceed to show that $\mathbf{B}$ is convex, that $v_p(\cdot)$, restricted to $\mathbf{B}$, is convex, and that for $b \in \mathbf{B}$, $v_p(b) = \min\{ f^0(x) \mid f^j(x) \le b^j, j = 0,1, \cdots, m, x \in \mathbf{X} \}$. Then we will show that $v_d(0) = v_p(0)$,

and that we can solve (12.4.1) by solving the dual problem (12.4.3b).

Lemma 12.4.1 : The set **B** is convex.

Proof : Let $b_1, b_2 \in \mathbf{B}$ and let $\lambda \in [0,1]$. By definition of **B**, there exist $x_1, x_2 \in \mathbf{X}$ such that $f(x_1) \leq b_1, f(x_2) \leq b_2$. Since the components of $f(\cdot)$ are convex by assumption, it follows that

$$f^j(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f^j(x_1) + (1-\lambda)f(x_2) \leq \lambda b_1^j + (1-\lambda)b_2^j, \ \forall\, j \in \underline{m}\,. \tag{12.4.4}$$

Since **X** is convex by assumption, $\lambda x_1 + (1-\lambda)x_2 \in \mathbf{X}$, and hence we see that $\lambda b_1 + (1-\lambda)b_2 \in \mathbf{B}$. ■

Lemma 12.4.2 : Suppose that $b \in \mathbf{B}$, then

$$v_p(b) = \min\{\, f^0(x) \mid f^j(x) \leq b^j,\ j = 1, \cdots, m,\ x \in \mathbf{X}\,\}. \tag{12.4.5}$$

Proof : Suppose that $x \in \mathbf{X}$ is such that $f^j(x) > b^j$ for some $j \in \underline{m}$. Then, for this x,

$$\sup_{\mu \geq 0} (f^0(x) + \langle \mu, f(x) - b \rangle) = \infty\,. \tag{12.4.6a}$$

Hence we conclude that me must have

$$\begin{aligned} v_p(b) &= \inf_{\substack{x \in \mathbf{X} \\ f(x) \leq b}} \sup_{\mu \geq 0} (f^0(x) + \langle \mu, f(x) - b \rangle) \\ &= \inf\{\, f^0 \mid f(x) \leq b,\ x \in \mathbf{X}\,\}. \end{aligned} \tag{12.4.6b}$$

Since by Assumption 12.4.1 (ii), the infimum in (12.4.6b) is achieved, our proof is complete. ■

Lemma 12.4.3 : The value function $v_p(\cdot)$ is convex on **B**.

Proof : Let $b_1, b_2 \in \mathbf{B}$ be arbitrary, and let $\lambda \in [0,1]$. Then, making use of Lemma 12.4.2, we obtain that

$$\lambda v_p(b_1) + (1-\lambda)v_p(b_2) = \min\{\, \lambda f^0(x_1) \mid f(x_1) \leq b_1, x_1 \in \mathbf{X}\,\} + \min\{\, (1-\lambda) f^0(x_2) \mid f(x_2) \leq b_2, x_2 \in \mathbf{X}\,\}$$

$$= \min\{\, \lambda f^0(x_1) + (1-\lambda) f^0(x_2) \mid f(x_1) \leq b_1,\ f(x_2) \leq b_2,\ x_1, x_2 \in \mathbf{X}\,\}$$

$$\geq \min\{\, \lambda f^0(x_1) + (1-\lambda) f^0(x_2) \mid \lambda f(x_1) + (1-\lambda) f(x_2) \leq \lambda b_1 + (1-\lambda) b_2,\ x_1, x_2 \in \mathbf{X}\,\}$$

$$\geq \min\{\, f^0(x) \mid f(x) \leq \lambda b_1 + (1-\lambda) b_2,\ x \in \mathbf{X}\,\}$$

$$= v_p(\lambda b_1 + (1-\lambda) b_2)\,, \tag{12.4.7}$$

which shows that $v_p(\cdot)$ is convex. ■

The remainder of our analysis will take place in $\mathbf{R}^{m+1}$. We will denote vectors in $\mathbf{R}^{m+1}$ by $\bar{z} = (z^0, z)$, where $z^0 \in \mathbf{R}$ and $z \in \mathbf{R}^m$. As before, we define the function $F: \mathbf{R}^n \to \mathbf{R}^{m+1}$ by

$$F(x) \triangleq \begin{bmatrix} f^0(x) \\ f(x) \end{bmatrix}. \tag{12.4.8}$$

For $m = 1$ the set $F(X)$ and its relation to B are shown in Fig. 12.4.1.

Definition 12.4.1 : The *graph* of $v_p(\cdot)$ is the set

$$\Gamma \triangleq \{ \bar{z} \in \mathbb{R}^{m+1} \mid z \in B, z^0 = v_p(z) \}. \tag{12.4.9a}$$

The *epigraph* of $v_p(\cdot)$ is the set (see Fig. 12.4.2))

$$\Gamma_0 \triangleq \{ \bar{z} \in \mathbb{R}^{m+1} \mid z \in B, z^0 \geq v_p(z) \}. \tag{12.4.9b}$$

Note that Γ_0 is convex because $v_p(\cdot)$ is convex.

Next we observe that the following relation holds:

$$\begin{aligned} v_d(b) &= \sup_{\mu \geq 0} \inf_{x \in X} (f^0(x) + \langle \mu, f(x) - b \rangle) \\ &= \sup_{\mu \geq 0} \inf_{\bar{z} \in F(X)} (z^0 + \langle \mu, z - b \rangle). \end{aligned} \tag{12.4.10}$$

Next, let

$$\begin{aligned} d(\mu) &\triangleq \inf_{x \in X} (f^0(x) + \langle \mu, f(x) \rangle) \\ &= \inf_{\bar{z} \in F(X)} (z^0 + \langle \mu, z \rangle), \end{aligned} \tag{12.4.11}$$

then we see that $v_d(0) = \sup_{\mu \geq 0} d(\mu)$. The quantity $d(\mu)$ has an important geometrical interpretation. First, it is the value of minimizing the *linear* cost function $z^0 + \langle \mu, z \rangle$ on the convex set $F(X)$. Hence, if the minimum is achieved at the point $\bar{z}^*$, then the hyperplane $\{ \bar{z} \in \mathbb{R}^{m+1} \mid z^0 + \langle \mu, z \rangle = d(\mu) \}$ is tangent to $F(X)$ at $\bar{z}^*$. Furthermore, this hyperplane intercepts the line $\{ \bar{z} \in \mathbb{R}^{m+1} \mid z = 0 \}$ at the point $(d(\mu), 0)$, as shown in Fig. 12.4.2. It follows immediately that $d(\mu) \leq v_p(0)$. In fact, we will prove the following result.

Lemma 12.4.3 : For every $b \in B$, $v_d(b) \leq v_p(b)$.

Proof : Let $b \in B$ be arbitrary. Then for all $x' \in X$ and $\mu \geq 0$,

$$\inf_{x \in X} (f^0(x) + \langle \mu, f(x) - b \rangle) \leq f^0(x') + \langle \mu, f(x') - b \rangle, \tag{12.4.12a}$$

which leads to the conclusion that

$$\begin{aligned} v_d(b) &= \sup_{\mu \geq 0} \inf_{x \in X} (f^0(x) + \langle \mu, f(x) - b \rangle) \\ &\leq \sup_{\mu \geq 0} (f^0(x') + \langle \mu, f(x') - b \rangle), \ \forall x' \in X. \end{aligned} \tag{133.4.12b}$$

Hence

$$v_d(b) \leq \inf_{x' \in X} \sup_{\mu \geq 0} (f^0(x') + \langle \mu, f(x') - b \rangle)$$

$$= v_p(b) . \qquad (12.4.12c)$$

■

Definition 12.4.2 : We define the *negative octant* Q_- by

$$Q_- \triangleq \{ \overline{z} \in \mathbb{R}^{m+1} \mid z^j < 0, j = 0,1,2,...,m \}. \qquad (12.4.13)$$

Lemma 12.4.4 : suppose that $\overline{z}^* = (z^{0*}, b) \in \mathbb{R}^{m+1}$ is such that $\overline{z}^* \in \Gamma$. Then

$$\left[\Gamma_0 - \{ \overline{z}^* \} \right] \cap Q_- = \phi. \qquad (12.4.14a)$$

Proof : Suppose not. Then there exists a $\overline{z}^{**} \in \Gamma_0$ such that

$$z^{**0} - z^{*0} < 0, \qquad (12.4.14b)$$

$$z^{**j} - b^j < 0, \ j = 1,2,...,m . \qquad (12.4.14c)$$

By definition of Γ and Γ_0, $z^{*0} = v_p(b)$ and $z^{**0} \geq v_p(z^{**})$. Also, since $z^{**} < b$, we must have that $v_p(z^{**}) \geq v_p(b)$. Hence

$$z^{*0} = v_p(b) > z^{**0} \geq v_p(z^{**}) \geq v_p(b) , \qquad (12.4.14c)$$

which is a contradiction. ■

Theorem 12.4.1 : Suppose that Assumption 12.4.1 is satisfied. Then $v_p(0) = v_d(0)$.

Proof : Consider the point $\overline{z}^* = (v_p(0),0) \in \Gamma$. Then, by Lemma 12.4.4, $(\Gamma_0 - \{ \overline{z}^* \}) \cap Q_- = \phi$ and both Q_- and $\Gamma_0 - \{ \overline{z}^* \}$ are convex. Therefore $\Gamma_0 - \{ \overline{z}^* \}$ and Q_- can be separated, and hence so can be their closures, i.e., there exists a nonzero vector $\overline{\pi} \in \mathbb{R}^{m+1}$ and an $\alpha \in \mathbb{R}$ such that

$$\langle \overline{z}, \overline{\pi} \rangle \geq \alpha, \quad \forall \ \overline{z} \in \Gamma_0 - \{ \overline{z}^* \}, \qquad (12.4.15a)$$

$$\langle \overline{z}, \overline{\pi} \rangle \leq \alpha, \quad \forall \ \overline{z} \in \overline{Q}_- , \qquad (12.4.15b)$$

where $\overline{Q}_-$ denotes the closure of Q_-. Since $0 \in (\Gamma_0 - \{ \overline{z}^* \}) \cap \overline{Q}_-$, it follows that $\alpha = 0$. Since the $\mathbb{R}^{m+1}$ unit vectors $-\overline{e}_j \in \overline{Q}_-$, for $j = 0,1,2,...,m$, it follows from (12.4.15b) that

$$\langle -\overline{e}_j, \overline{\pi} \rangle = -\pi^j \leq 0, \ \text{for } j = 0,1,2,...,m , \qquad (12.4.16)$$

which shows that $\pi^j \geq 0$ for $j = 0,1,2,...,m$. We shall now show that $\pi^0 > 0$. Suppose not. Then $\pi^0 = 0$ and hence (12.4.15a) yields that

$$\sum_{j=1}^{m} \pi^j (z^j - 0) = \sum_{j=1}^{m} \pi^j z^j \geq 0 \ \forall \ \overline{z} \in \Gamma_0 . \qquad (12.4.17)$$

But by Assumption 12.4.1(i), there exists a $b' \in B$ such that $b' < 0$. Since $(v_p(b'), b') \in \Gamma_0$, it follows from (12.4.17) that

$$\sum_{j=1}^{m} \pi^j b^{ij} \geq 0 . \tag{12.4.18}$$

Since the $\pi^j \geq 0$ for $j = 0,1,2,....,m$, it follows that $\pi^j = 0$ for $j = 0,1,2,....,m$, and hence that $\bar{\pi} = 0$, which is a contradiction. Hence we must have $\pi^0 > 0$. We now define

$$\hat{\mu}^j = \pi^j/\pi^0, \quad j = 1,2,....,m . \tag{12.4.19}$$

Then, from (12.4.15a), we obtain that

$$\frac{1}{\pi^0}\langle \bar{\pi}, \bar{z} - \bar{z}^* \rangle = z^0 - z^{*0} + \langle \hat{\mu}, z - z^* \rangle \geq 0, \ \forall \ \bar{z} \in \Gamma_0 . \tag{12.4.20a}$$

Since $z^{*0} = v_p(0)$ and $z^* = 0$, (12.4.20a) yields

$$z^0 + \langle \hat{\mu}, z \rangle \geq v_p(0), \ \forall \ \bar{z} \in \Gamma_0 . \tag{12.4.20b}$$

Since $F(X) \subset \Gamma_0$, it follows that

$$\begin{aligned} v_d(0) &= \sup_{\mu \geq 0} \inf_{\bar{z} \in F(X)} (z^0 + \langle \mu, z - b \rangle) \\ &\geq \sup_{\mu \geq 0} \inf_{\bar{z} \in \Gamma_0} (z^0 + \langle \mu, z - b \rangle) \\ &\geq \inf_{\bar{z} \in \Gamma_0} (z^0 + \langle \hat{\mu}, z - b \rangle) \geq v_p(0) . \end{aligned} \tag{12.4.21}$$

In view of Lemma 12.4.3, we must therefore have that $v_p(0) = v_d(0)$. ■

Remark 12.4.1 : Note that the above result was established without any differentiability assumptions■

One of the ways in which duality is used is to estimate "cost-to-go" in an optimization and to use this estimate to stop a computation. Thus, suppose that we have a "primal-dual algorithm" which constructs a sequence of multipliers $\{ \mu_i \}$ which are dual feasible and a sequence of points $\{ x_i \}$ which are primal feasible for problem (12.4.1). Then $d(\mu_i) \leq v_p(0) \leq f^0(x_i)$ for all i. Hence when duality applies, $f^0(x_i) - d(\mu_i)$ is a good measure of the cost-to-go, and when it is below a threshold, computation can be stopped.

Next, suppose that we apply an algorithm to the dual and construct a sequence $\{ \mu_i \}$ which is dual feasible. If we use the μ_i to construct a sequence $\{ x_i \}$ satisfying

$$d(\mu) = (f^0(x_i) + \langle \mu, f(x_i) \rangle), \tag{12.4.22}$$

then, quite likely we will have that $f^0(x_i) < v_p(0)$. But then the x_i cannot be primal feasible. Thus, by solving the dual, we probably approach primal feasibility in the limit. This can be a real disadvantage in real-time operations.

Finally, consider the quadratic programming problem

$$\min\{\ \frac{1}{2}\langle x,Qx\rangle + \langle c,x\rangle \mid \langle a_j,x\rangle \le 0,\quad j = 1,2,....,m\ \}. \tag{12.4.23a}$$

where Q is a positive definite $n \times n$ matrix. Let A be an $m \times n$ matrix whose ith row is a_j^T. Then the dual of (12.4.23a) is

$$\sup_{\mu \ge 0} \inf_{x \in \mathbb{R}^n} \{\ \frac{1}{2}\langle x,Qx\rangle + \langle c,x\rangle + \langle \mu,Ax\rangle\}. \tag{12.4.23b}$$

Since the function in (12.4.23b) is convex and differentiable in x, given μ we can compute the corresponding mimimizer x_μ from the optimality condition

$$Qx_\mu + c + A^T\mu = 0\ . \tag{12.4.24a}$$

Hence

$$x_\mu = -Q^{-1}(c + A^T\mu)\ . \tag{12.4.24b}$$

We now compute that

$$x_\mu^T Q x_\mu = (c + A^T\mu)^T Q^{-1}(c + A^T\mu)$$

$$= c^T Q^{-1} c^T + 2c^T Q^{-1} A^T\mu + \mu^T A Q^{-1} A^T\mu\ ; \tag{12.4.25a}$$

$$\mu^T A x_\mu = -\mu^T A Q^{-1}(c + A^T\mu)$$

$$= \mu^T A Q^{-1} c - \mu^T A Q^{-1} A^T\mu; \tag{12.4.25b}$$

$$c^T x_\mu = -c^T Q^{-1} c - c^T Q^{-1} A^T\mu\ . \tag{12.4.25c}$$

Substituting into (12.4.23b), we obtain that

$$v_d(0) = \sup_{\mu \ge 0}\{\ -\frac{1}{2}\langle \mu,AQ^{-1}A^T\mu\rangle - \langle c,Q^{-1}A^T\mu\rangle - \frac{1}{2}\langle c,Q^{-1}c\rangle\ \}\ . \tag{12.4.26}$$

Note that (12.4.26) is a quadratic program with simple constraints ($\mu \ge 0$), but we must compute Q^{-1}.

Exercise 12.4.1 : Consider the search direction finding problem

$$\min_{h \in \mathbb{R}^n} \max_{j \in \underline{m}} \{\ f^j(x) - \psi(x) + \langle \nabla f^j(x),h\rangle + \frac{1}{2}|h|^2\ \}\ . \tag{12.4.27a}$$

Use the Duality Theorem 12.4.1 to show that its solution h^* is given by $h^* = -\sum_{j=1}^{m} \mu^{*j}\nabla f^j(x)$, where the μ^{*j} are any solution to the dual problem:

$$\max_{\mu \in \Sigma}\{\ \sum_{j=1}^{m} \mu^j(f^j(x) - \psi(x)) - \frac{1}{2}|\sum_{j=1}^{m} \mu^j\nabla f^j(x)|\ \}, \tag{12.4.27b}$$

where $\Sigma \triangleq \{\ \mu \in \mathbb{R}^m \mid \mu^j \ge 0, j \in \underline{m}, \sum_{j=1}^{m} \mu^j = 1\ \}$. ■

13. UNCONSTRAINED OPTIMAL CONTROL

We shall now show that the optimality conditions which we have derived for finite dimensional optimization problems have obvious extensions to optimal control problems.

13.1. FIRST ORDER EXPANSIONS OF SOLUTIONS OF DIFFERENTIAL EQUATIONS

Before we can derive optimality conditions for optimal control problems with nonlinear dynamics, we need to develop formulae for first order expansions of solutions of ordinary differential equations. Hence consider the differential equation

$$\frac{d}{dt}x(t) = h(x(t),u(t)),\ t \in [0,T], \tag{13.1.1}$$

where $h:\mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$ is continuously differentiable and $u(\cdot)$ is a piecewise continuous function from $[0,T]$ into U, a subset of $\mathbf{R}^m$. We shall denote by $\mathbf{U}[0,T]$ the set of piecewise continuous functions from $[0,T]$ into U, and we shall denote by $x(t,x_0,u)$ the solution of (13.1.1) corresponding to an initial state x_0 and a control $u \in \mathbf{U}[0,T]$.

It takes several applications of the Bellman-Gronwall inequality to show first that $x(t,x_0,u)$ is Lipschitz continuous and then that it is continuously differentiable in (x_0,u), in the $L_\infty[0,T]$ topology. We omit the laborious proof of these facts, and content ourselves with deriving the formula for the differential of $x(t,x_0,u)$ with respect to $(\delta x_0,\delta u)$.

First we recall that, given a Banach space $\mathbf{X}$, the differential of a function $f:\mathbf{X} \to \mathbf{R}^n$, at $x \in \mathbf{X}$, is defined as the *linear functional* $\partial f(x;\cdot)$ with the property that

$$\lim_{h\to 0} \frac{\|f(x+h) - f(x) - \partial f(x;h)\|}{\|h\|} = 0\ . \tag{13.1.2}$$

When the differential of $f:\mathbf{X} \to \mathbf{R}^n$ exists, the directional derivative is equal to the differential, i.e., $df(x;h) = \partial f(x;h)$, and a formula for both can be obtained from the fact that for any $x, h \in \mathbf{X}$, $\partial f(x;h) = \partial f(x+sh)/\partial s$, evaluated at $s = 0$. Hence we proceed to find a formula for the differential of $x(t,x_0,u)$, as follows.

Proposition 13.1.1: Let $\mathbf{PC}[0,T]$ denote the space of piecewise continuous functions from $[0,T]$ into $\mathbf{R}^m$. Suppose that $u,\delta u \in \mathbf{PC}[0,T]$, and $x_0,\delta x \in \mathbf{R}^n$. Let the differential of $x(t,x_0,u)$ be denoted by $\delta x(t,x_0,u;\delta x_0,\delta u)$. Then

$$\delta x(t,x_0,u;\delta x_0,\delta u) \triangleq \left.\frac{\partial x(t,x_0+\delta x_0,u+s\delta u)}{\partial s}\right|_{s=0}, \tag{13.1.3a}$$

and $\delta x(t,x_0,u;\delta u)$ is the solution of the following linear differential equation

$$\frac{d}{dt}\delta x(t) = \frac{\partial h(x(t,x_0,u),u(t))}{\partial x}\,\delta x(t) + \frac{\partial h(x(t,x_0,u),u(t))}{\partial u}\,\delta u(t),\ t \in [0,T], \tag{13.1.3b}$$

$$\delta x(0) = \delta x_0 . \tag{13.1.3c}$$

Proof : By the definition of $x(t,x_0,u + s\delta u)$, we have that for every $t \in [0,T]$

$$x(t,x_0 + s\delta x_0,u + s\delta u) = x_0 + s\delta x_0 + \int_0^t h(x(\tau,x_0 + s\delta x_0,u + s\delta u),u(\tau) + s\delta u(\tau))d\tau\ ,\ t \in [0,T] \tag{13.1.4}$$

Taking the partial derivative of both sides of equation (13.1.4) with respect to s and evaluating at $s = 0$, we obtain

$$\delta x(t,x_0,u;\delta x_0,\delta u) = \delta x_0 + \int_0^t \left\{ \frac{\partial h(x(\tau,x_0 + \delta x_0,u + s\delta u),u(\tau) + s\delta u(\tau))}{\partial x} \frac{\partial x(t,x_0 + \delta x_0,u + s\delta u)}{\partial s} \right.$$

$$\left. + \frac{\partial h(x(\tau,x_0 + \delta x_0,u + s\delta u),u(\tau) + s\delta u(\tau))}{\partial u}\delta u(\tau) \right\} \Big|_{s=0} d\tau$$

$$= \delta x_0 + \int_0^t \left\{ \frac{\partial h(x(\tau,x_0,u),u(\tau))}{\partial x}\delta x(t,x_0,u;\delta u) + \frac{\partial h(x(\tau,x_0,u),u(\tau))}{\partial u}\delta u(\tau) \right\} d\tau\ . \tag{13.1.5}$$

The desired result now follows by inspection. ■

Corollary 13.1.1 : Let $\Phi(t,\tau)$ denote the state transition matrix of the linear system (13.1.3a), i.e,

$$\frac{\partial \Phi(t,\tau)}{\partial t} = \frac{\partial h(x(t,x_0,u),u(t))}{\partial x}\Phi(t,\tau),\ \text{ for } t,\tau \in [0,T]\ , \tag{13.1.6a}$$

$$\Phi(t,t) = I\ ,\ \text{for } t \in [0,T]\ . \tag{13.1.6b}$$

Then for any $t \in [0,T]$,

$$\delta x(t,x_0,u;\delta x_0,\delta u) = \Phi(t,0)\delta x_0 + \int_0^t \Phi(t,\tau)\frac{\partial h(x(\tau,x_0,u),u(\tau))}{\partial u}\delta u(\tau)d\tau\ , \tag{13.1.6c}$$

i.e., $\Phi(t,0)$ is the jacobian of $x(t,x_0,u)$ with respect to x_0 and $\partial h(x(\tau,x_0,u),u(\tau))/\partial u$ is the "jacobian" of $x(t,x_0,u)$ with respect to u. ■

Note that the jacobian of $x(t,x_0,u)$ with respect to x_0 is a matrix, since all linear operators on $\mathbb{R}^n$ have matrix realizations, while the "jacobian" of $x(t,x_0,u)$ with respect to u is a *kernel* for a linear operation defined through integration.

13.2. FIRST ORDER OPTIMALITY CONDITION

Now we consider following optimal control problem

$$\min\{f^0(x(T)) \mid \dot{x}(t) = h(x(t),u(t)),\ \text{for } t \in [0,T]\ ,\ x(0) = x_0\ ,\ u \in \mathrm{U}[0,T]\ \}\ , \tag{13.2.1}$$

where $f^0:\mathbb{R}^n \to \mathbb{R}$ and $h:\mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ are continuously differentiable, the initial state x_0 is given, and $\mathrm{U}[0,T]$ the set of piecewise continuous functions from $[0,T]$ into U, a subset of $\mathbb{R}^m$.

Note that $U[0,T] \subset L_\infty$ is not compact in the L_∞ topology, and hence it is not clear whether (13.2.1) has a solution. We now proceed to derive first order optimality conditions for problem (13.2.1) on the assumption that it does have a solution.

Theorem 13.2.1 (First Order Optimality Conditions) : Suppose that $\hat{u} \in U[0,T]$ is an optimal control for (13.2.1) and $\hat{x}(\cdot)$ is the corresponding optimal trajectory. Let $\hat{p}(\cdot)$ be the solution of the adjoint equation

$$\dot{p}(t) = -\left[\frac{\partial h(\hat{x}(t),\hat{u}(t))}{\partial x}\right]^T p(t), \text{ for } t \in [0,T] , \tag{13.2.2a}$$

$$p(T) = \nabla f^0(\hat{x}(T)) . \tag{13.2.2b}$$

Then for any δu such that $\hat{u} + s\delta u \in U[0,T]$ for $s \in [0,s_{\delta u}]$, with $s_{\delta u} > 0$,

$$\int_0^T \langle \hat{p}(\tau), \frac{\partial h(\hat{x}(\tau),\hat{u}(\tau))}{\partial u}\delta u(\tau)\rangle \, d\tau \geq 0 . \tag{13.2.3}$$

Proof : Let δu be as above. Since $\hat{u}(\cdot)$ is an optimal control for the problem (13.2.1),

$$f^0(x(T,x_0,\hat{u} + s\delta u)) - f^0(x(T,x_0,\hat{u})) \geq 0, \forall \ s \in [0,s_{\delta u}] . \tag{13.2.4}$$

Dividing both sides of inequality (13.2.4) by s and letting s tend to 0, we obtain

$$\nabla f^0(\hat{x}(T))^T \, \delta x(T,x_0,\hat{u};\delta u) \geq 0, \tag{13.2.5}$$

where $\delta x(\cdot,x_0,\hat{u};\delta u)$ is defined by (13.1.2). Making use of Proposition 13.1.1, we conclude that

$$\delta x(T,x_0,\hat{u};\delta u) = \int_0^T \Phi(T,\tau)\frac{\partial h(\hat{x}(\tau),\hat{u}(\tau))}{\partial u}\delta u(\tau)d\tau, \tag{13.2.6}$$

where $\Phi(\cdot,\cdot)$ is the state transition matrix for the linear system (13.1.3a) with $u(\cdot) = \hat{u}(\cdot)$, i.e.,

$$\frac{\partial \Phi(t,\tau)}{\partial t} = \frac{\partial h(\hat{x}(t),\hat{u}(t))}{\partial x}\Phi(t,\tau), \text{ for } t,\tau \in [0,T] , \tag{13.2.7a}$$

$$\Phi(t,t) = I , \text{ for } t \in [0,T] . \tag{13.2.7b}$$

Thus, (13.2.3) follows from (13.2.4) and (13.2.6) and the fact that $\hat{p}(t) = \Phi(T,t)^T \nabla f^0(\hat{x}(T))$. ■

Corollary 13.2.1 : Suppose that $U = \mathbb{R}^m$, $\hat{u} \in U[0,T]$ is an optimal control for (13.2.1) and that $\hat{x}(\cdot)$ is the corresponding optimal trajectory. Let $\hat{p}(\cdot)$ be the solution of the adjoint equation (13.2.2a) and (13.2.2b). Then for every $t \in [0,T]$,

$$\left\{\frac{\partial h(\hat{x}(t),\hat{u}(t))}{\partial u}\right\}^T \hat{p}(t) = 0 . \tag{13.2.8}$$

Proof : Let

$$\delta\hat{u}(t) = -\left\{\frac{\partial h(\hat{x}(t),\hat{u}(t)}{\partial u}\right\}^T \hat{p}(t)\ ,\ t \in [0,T]\ . \tag{13.2.9}$$

Since $U = \mathbf{R}^m$, $\hat{u} + s\delta\hat{u} \in \mathbf{U}[0,T]$ for all $s \in \mathbf{R}$. Hence, it follows from Theorem 13.2.1 that

$$0 \leq \int_0^T \langle \hat{p}(\tau), \frac{\partial h(\hat{x}(\tau),\hat{u}(\tau))}{\partial u}\delta\hat{u}(\tau)\rangle\ d\tau$$

$$= -\int_0^T |\delta\hat{u}(\tau)|^2\ d\tau\ \leq\ 0\ . \tag{13.2.10}$$

Thus

$$\int_0^T |\delta\hat{u}(\tau)|^2\ d\tau = 0\ . \tag{13.2.11}$$

Since $\delta\hat{u}(\cdot)$ is piecewise continuous, we conclude that $\delta\hat{u}(t) = 0$ for all $t \in [0,T]$. ■

13.3. GRADIENT METHODS

We shall now consider the special case of problem (13.2.1)

$$\min\{\, f^0(x(T)) \mid \dot{x}(t) = h(x(t),u(t)),\ \text{for } t \in [0,T]\ ,\ x(0) = x_0\ ,\ u \in \mathbf{U}[0,T]\ \}\ , \tag{13.3.1a}$$

where $f^0{:}\mathbf{R}^n \to \mathbf{R}$ and $h{:}\mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$ are continuously differentiable, the initial state x_0 is given, and $\mathbf{U}[0,T]$ the set of piecewise continuous functions from $[0,T]$ into $\mathbf{R}^m$. As before, we shall denote by $x(t,x_0,u)$ the solution of (13.1.1) corresponding to an initial state x_0 and a control $u \in \mathbf{U}[0,T]$.

First we note that problem (13.3.1a) can be rewritten as an *unconstrained* problem in the space $\mathbf{U}[0,T]$, as follows:

$$\min_{u \in \mathbf{U}[0,T]} f(u), \tag{13.3.1b}$$

where $f{:}\mathbf{U}[0,T] \to \mathbf{R}$ is defined by

$$f(u) \triangleq f^0(x(T,x_0,u))\ . \tag{13.3.1c}$$

Since both $f^0(\cdot)$ and $x(T,x_0,\cdot)$ are differentiable, it follows that the function $f(\cdot)$ is differentiable, and that, by the chain rule, its differential is given (via Corollary 13.1.1) by

$$\partial f(u;\delta u) = \langle \nabla f^0(x(T,x_0,u)), \delta x(T,x_0,u;0,\delta u)\rangle$$

$$= \int_0^T \langle \nabla f^0(x(T,x_0,u)), \Phi(t,\tau)\frac{\partial h(x(\tau,x_0,u),u(\tau))}{\partial u}\delta u(\tau)\rangle\ d\tau$$

$$= \int_0^T \langle \frac{\partial h(x(\tau,x_0,u),u(\tau))^T}{\partial u}\Phi(t,\tau)^T\nabla f^0(x(T,x_0,u)), \delta u(\tau)\rangle\ d\tau$$

$$= \int_0^T \langle \nabla f(u)(\tau), \delta\tau \rangle \, d\tau, \tag{13.3.2}$$

where

$$\nabla f(u)(\tau) \triangleq \frac{\partial h(x(\tau,x_0,u),u(\tau))^T}{\partial u} \Phi(t,\tau)^T \nabla f^0(x(T,x_0,u))$$

$$= \frac{\partial h(x(\tau,x_0,u),u(\tau))^T}{\partial u} p(\tau), \tag{13.3.3a}$$

with $p(\cdot)$ the solution of the adjoint equation

$$\dot{p}(t) = -\left[\frac{\partial h(x(t,x_0,u),u(t))}{\partial x}\right]^T p(t), \quad \text{for } t \in [0,T], \tag{13.3.3b}$$

$$p(T) = \nabla f^0(x(T,u,x_0)). \tag{13.3.3c}$$

If we define the scalar product in $U[0,T]$ by

$$\langle u,v \rangle_2 \triangleq \int_0^T \langle u(t),v(t) \rangle dt, \tag{13.3.4}$$

then we see that the function $\nabla f(u)$ is the function space analog of the gradient of a function from $\mathbb{R}^n$ into $\mathbb{R}$. Furthermore, it should be clear that it is continuous with respect to the $L_\infty[0,T]$ norm. Hence we get the following obvious extension of the Armijo Gradient Method 3.3.2 to the problem (13.3.1b).

Armijo Gradient Algorithm 13.3.1 :

Parameters : $\alpha, \beta \in (0,1)$.

Data: $u_0 \in U[0,T]$.

Step 0 : set $i = 0$.

Step 1 : Compute *the search direction*

$$h_i = h(u_i) \triangleq -\nabla f(u_i). \tag{13.3.5a}$$

Stop if $\nabla f(u_i) = 0$.

Step 2 : Compute the *step size*

$$\lambda_i = \beta^{k_i} \triangleq \arg\max_{k \in \mathbf{N}} \left\{ \beta^k \mid f(u_i + \beta^k h_i) - f(u_i) \leq -\beta^k \alpha \langle \nabla f(u_i), \nabla f(u_i) \rangle_2 \right\} \tag{13.3.5b}$$

Step 3 : *Update*

$$x_{i+1} = x_i + \lambda_i h_i, \tag{13.3.5c}$$

replace i by $i + 1$ and go to step 1.

■

Note again that the evaluation of $\nabla f(u_i)$ in the algorithm above requires several operations: (i) the differential equation (13.1.1) has to be solved for $u = u_i$, from the given initial state x_0; (ii) the adjoint equation (13.3.3b), (13.3.3c) has to be solved, with $u = u_i$, for $p(t,u_i)$; and finally, $\nabla f(u_i)$ can be computed using (13.3.3a), again with $u = u_i$ and $p(t) = p(t,u_i)$. Similarly, each test in the Armijo step size rule requires the solution of the differential equation (13.1.1), from the given initial state x_0, using the control $u(t) = u_i(t) - \beta^k \nabla f(u)(t)$.

In view of the continuity of the gradient function $\nabla f(\cdot)$, the following theorem should be obvious.

Theorem 13.3.1 :

(a) The Armijo step size rule is well defined.

(b) If $\{ u_i \}_{i=0}^{\infty}$ is an infinite sequence constructed by Algorithm 13.3.1, then every accumulation point $\hat{u}$ (in the $L_\infty[0,T]$ sense) of $\{u_i\}_{i=0}^{\infty}$ satisfies $\nabla f(\hat{u}) = 0$. ■

Unlike the finite dimensional case, a bounded sequence of controls in U[0,T] need not have accumulation points in the $L_\infty[0,T]$ sense. Hence the above theorem seems to be weak. Fortunately, there is a topology, call the the *relaxed controls topology* in which such a sequence always has an accumulation point, and it can be shown that the above theorem remains valid in the above topology as well.

Exercise 13.3.1 : Construct a formal extension of the Polak-Ribiere conjugate gradient method for solving the optimal control problem (13.3.1b). ■

13.4. LINEAR-QUADRATIC REGULATOR PROBLEM

In this section, we shall obtain analytical results for the following linear-quadratic regulator problem with time-invariant dynamics:

$$\min\{\int_0^T (\tfrac{1}{2}\langle x(t),Qx(t)\rangle + \tfrac{1}{2}\langle u(t),Ru(t)\rangle)dt \mid \dot{x}(t) = Ax(t) + Bu(t) \text{ , for } t \in [0,T] \text{ ,}$$

$$x(0) = x_0 \text{ , } u \in U[0,T] \} \text{ ,} \qquad (13.4.1)$$

where $U = \mathbb{R}^m$, x_0 is given, R is a symmetric, positive definite $n \times n$ matrix and Q is a symmetric, semi-positive$m \times m$ definite matrix.

Remark 13.4.1 : It can be shown that problem (13.4.1) always has a solution. ■

Assumption 13.4.1 : We shall assume that the pair (A,B) in (13.4.1) is completely controllable, and that the pair $(A,Q^{1/2})$ is completely observable. ■

First, by augmenting the state variables of the linear dynamics in (13.4.1), we transform problem (13.4.1) into a problem in form of (13.2.1):

$$\min\{ f^0(\bar{x}(T)) \mid \dot{\bar{x}}(t) = (x^0(t),x(t)) \text{ ,}$$

$$\dot{x}^0(t) = \tfrac{1}{2}\langle x(t), Qx(t)\rangle + \tfrac{1}{2}\langle u(t), Ru(t)\rangle, \text{ for } t \in [0,T],$$

$$\dot{x}(t) = Ax(t) + Bu(t), \text{ for } t \in [0,T],$$

$$x^0(0) = 0,\ x(0) = x_0,\ u \in U[0,T] \}, \tag{13.4.2}$$

where $f^0(\bar{x}) = \langle e_0, \bar{x}\rangle = x_0$ and $e_0 = (1,0,\cdots,0)^T \in \mathbb{R}^{n+1}$.

Suppose that $\hat{u} \in U[0,T]$ is an optimal control for problem (13.4.1) and that $\hat{x}(\cdot)$ is the corresponding optimal trajectory. Then, it follows from Corollary 13.2.1 that for all $t \in [0,T]$,

$$\begin{bmatrix} \hat{u}(t)^T R \\ B \end{bmatrix}^T \bar{p}(t) = 0, \tag{13.4.3}$$

where $\bar{p}(\cdot)$ satisfies following adjoint differential equation

$$\frac{d\bar{p}(t)}{dt} = -\begin{bmatrix} 0 & \hat{x}^T(t)Q \\ 0 & A \end{bmatrix}^T p(t), \text{ for } t \in [0,T], \tag{13.4.4a}$$

$$\bar{p}(T) = e_0. \tag{13.4.4b}$$

Let $\bar{p}(t) = (p^0(t), p^T(t))^T$, where $p^0(t) \in \mathbb{R}$ and $p(t) \in \mathbb{R}^n$. Then equations (13.4.4a) and (13.4.4b) become

$$\dot{p}^0(t) = 0,\ t \in [0,T],\ \ p^0(T) = 1, \tag{13.4.5a}$$

$$\dot{p}(t) = -A^T p(t) - Q\hat{x}(t)p^0(t),\ t \in [0,T],\ \ p(T) = 0. \tag{13.4.5b}$$

Hence $p^0(t) = 1$ for all $t \in [0,T]$. Thus, making use of (13.4.3), we obtain

$$\hat{u}(t) = -R^{-1}B^T p(t), \text{ for } t \in [0,T]. \tag{13.4.6}$$

Making use of (13.4.5b), (13.4.6) and the fact that $\hat{x}(\cdot)$ is an optimal trajectory, we conclude that $\hat{x}(\cdot)$ and $p(\cdot)$ satisfy following linear differential equation:

$$\frac{d}{dt}\begin{bmatrix} \hat{x}(t) \\ p(t) \end{bmatrix} = \begin{bmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{bmatrix}\begin{bmatrix} \hat{x}(t) \\ p(t) \end{bmatrix}, \text{ for } t \in [0,T]. \tag{13.4.7a}$$

$$\hat{x}(0) = x_0,\ \ p(T) = 0. \tag{13.4.7b}$$

Proposition 13.4.1 : If $\hat{u}(\cdot)$ is the optimal control for (13.4.1) and $\hat{x}(\cdot)$ is the corresponding optimal trajectory, then the following relationship holds between the optimal cost and the adjoint vector $p(\cdot)$ at $t = 0$:

$$\int_0^T (\tfrac{1}{2}\langle \hat{x}(t), Q\hat{x}(t)\rangle + \tfrac{1}{2}\langle \hat{u}(t), R\hat{u}(t)\rangle)dt = \tfrac{1}{2}\langle x(0), p(0)\rangle. \tag{13.4.8a}$$

Proof : Making use of (13.4.7a), we observe that

$$\frac{d}{dt}\langle p(t),\hat{x}(t)\rangle = \langle \dot{p}(t),\hat{x}(t)\rangle + \langle p(t),\frac{d\hat{x}(t)}{dt}\rangle$$

$$= -\langle Q\hat{x}(t),\hat{x}(t)\rangle - \langle A^T p(t),\hat{x}(t)\rangle + \langle p(t),A\hat{x}(t)\rangle - \langle p(t),BR^{-1}B^T p(t)\rangle$$

$$= -\langle \hat{x}(t),Q\hat{x}(t)\rangle - \langle p(t),BR^{-1}B^T p(t)\rangle. \tag{13.4.8b}$$

Since $\hat{u}(t) = -R^{-1}B^T p(t)$ and $p(T) = 0$ by (13.4.7b), the desired result follows. ■

We now proceed to express $p(t)$ in terms of $x(t)$. Let $\Phi(t,\tau)$ be the $2n \times 2n$ state transition matrix for the system (13.4.7a) and let the $2n \times 2n$ matrix H be defined by:

$$H \triangleq \begin{bmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{bmatrix} \tag{13.4.9a}$$

Then we must have that

$$\frac{d}{dt}\Phi(t,\tau) = H\Phi(t,\tau), \text{ for } t \in [\tau,T] \; ; \; \Phi(\tau,\tau) = I \; ; \tag{13.4.9b}$$

$$\frac{d}{d\tau}\Phi(t,\tau)^T = -H^T\Phi(t,\tau)^T, \text{ for } \tau \in [t,T] \; ; \; \Phi(t,t) = I \; ; \tag{13.4.9c}$$

$$\begin{bmatrix} \hat{x}(T) \\ p(T) \end{bmatrix} = \Phi(T,t)\begin{bmatrix} \hat{x}(t) \\ p(t) \end{bmatrix}, \text{ for } t \in [0,T] . \tag{13.4.9d}$$

We partition the $2n \times 2n$ matrix $\Phi(T,t)$ into four $n \times n$ submatrices as follows:

$$\Phi(T,t) = \begin{bmatrix} \Phi_{11}(T,t) & \Phi_{12}(T,t) \\ \Phi_{21}(T,t) & \Phi_{22}(T,t) \end{bmatrix}. \tag{13.4.10a}$$

Then, because of (13.4.7b), the bottom part of the equation (13.4.9d) can be rewritten as follows:

$$0 = p(T) = \Phi_{21}(T,t)\hat{x}(t) + \Phi_{22}(T,t)p(t) \text{ , for } t \in [0,T] . \tag{13.4.10b}$$

Assuming that $\Phi_{22}(T,t)$ is nonsingular, we find that

$$p(t) = P_T(t)x(t) \text{ , } \tag{13.4.11a}$$

where

$$P_T(t) = -\Phi_{22}(T,t)^{-1}\Phi_{21}(T,t) . \tag{13.4.11b}$$

The matrix $P_T(t)$ depends upon the terminal time T, and on the matrices A, B, Q and R, but not on the initial state x_0. In view of Proposition 13.4.1, we get the following result:

Corollary 13.4.1 : If $\hat{u}(\cdot)$ is the optimal control for (13.4.1) with corresponding optimal trajectory $\hat{x}(\cdot)$ and $P_T(t)$ is defined by (13.4.11b), then for all nonzero $x(0) \in \mathbb{R}^n$

$$0 < \int_0^T (\tfrac{1}{2}\langle \hat{x}(t), Q\hat{x}(t)\rangle + \tfrac{1}{2}\langle \hat{u}(t), R\hat{u}(t)\rangle)dt = \tfrac{1}{2}\langle x(0), p(0)\rangle = \tfrac{1}{2}\langle x(0), P_T(0)x(0)\rangle \ . \tag{13.4.11c}$$

■

Before proceeding further, we will show that the matrix $\Phi_{22}(T,t)$ is nonsingular.

Lemma 13.4.1 : The matrix $\Phi_{22}(T,t)$ is nonsingular for all $t \in [0,T]$.

Proof : Suppose that $\Phi_{22}(T,t)$ is singular at $t_1 \in [0,T]$, and hence so is its transpose. Therefore there exists a nonzero vector $\xi \in \mathbb{R}^n$ such that $\Phi_{22}(T,t_1)^T\xi = 0$. For $t \in [t_1,T]$, let $w_1(\cdot) = \Phi_{21}(T,\cdot)^T\xi$ and let $w_2(\cdot) = \Phi_{22}(T,\cdot)^T\xi$. Then, making use of (13.4.9c), we obtain that

$$\frac{d}{dt}\langle w_1(t), w_2(t)\rangle = \langle \dot{w}_1(t), w_2(t)\rangle + \langle w_1(t), \dot{w}_2(t)\rangle$$

$$= -\langle A^T w_1(t) - Qw_2(t), w_2(t)\rangle - \langle w_1(t), -BR^{-1}B^T w_1(t) - Aw_2(t)\rangle$$

$$= \langle Qw_2(t), w_2(t)\rangle + \langle w_1(t), BR^{-1}B^T w_1(t)\rangle. \tag{13.4.12a}$$

Hence, because $w_1(T) = w_2(t_1) = 0$, we obtain that

$$0 = \int_{t_1}^T \frac{d}{dt}\langle w_1(t), w_2(t)\rangle dt = \int_{t_1}^T \langle Qw_2(t), w_2(t)\rangle + \langle w_1(t), BR^{-1}B^T w_1(t)\rangle dt \ , \tag{13.4.12b}$$

which leads to the conclusion that $\langle Qw_2(t), w_2(t)\rangle = 0$ and $\langle w_1(t), BR^{-1}B^T w_1(t)\rangle = 0$ for all $t \in [t_1,T]$. Since R is positive definite, by assumption, we conclude that $B^T w_1(t) = 0$ for all $t \in [t_1,T]$. It now follows from (13.4.9c) that

$$\dot{w}_2(t) = Aw_2(t), \quad w_2(t_1) = 0 \ , \tag{13.4.12c}$$

and hence that $w_2(t) = 0$ for all $t \in [t_1,T]$. In particular, we must have that $w_2(T) = \Phi_{22}^T(T,T)\xi = 0$. Since $\Phi_{22}^T(T,T) = I$, it follows that $\xi = 0$ and hence we have a contradiction, which completes our proof.

Next, we shall derive a matrix differential equation which $P_T(t)$ must satisfy. By differentiating equation (13.4.11a) and making use of equations (13.4.7a) and (13.4.11a), we obtain

$$\dot{p}(t) = \dot{P}_T(t)\hat{x}(t) + P_T(t)\frac{d\hat{x}(t)}{dt}$$

$$= \dot{P}_T(t)\hat{x}(t) + P_T(t)(A\hat{x}(t) - BR^{-1}B^T p(t))$$

$$= \dot{P}_T(t)\hat{x}(t) + P_T(t)(A\hat{x}(t) - BR^{-1}B^T P_T(t)\hat{x}(t)) \ . \tag{13.4.13}$$

But, substituting (13.4.11a) into (13.4.7a), we find that

$$\dot{p}(t) = -Q\hat{x}(t) - A^T P_T(t)\hat{x}(t) . \tag{13.4.14}$$

From equations (13.4.13) and (13.4.14), we conclude that

$$[\dot{P}_T(t) + P_T(t)A + A^T P_T(t) - P_T(t)BR^{-1}B^T P_T(t) + Q]\hat{x}(t) = 0 . \tag{13.4.15}$$

Since $P_T(t)$ does not depend upon initial state, and since $\hat{x}(t)$ is a solution of the homogeneous equation

$$\frac{d\hat{x}(t)}{dt} = (A - BR^{-1}B^T P_T(t))\hat{x}(t) , \tag{13.4.16}$$

we must have that

$$\hat{x}(t) = \Psi(t,0)x_0 , \tag{13.4.17}$$

where $\Psi(t,\tau)$ is the $n \times n$ state transition matrix for the linear system (13.4.16). Substituting (13.4.17) into (13.4.15), we obtain

$$[\dot{P}_T(t) + P_T(t)A + A^T P_T(t) - P_T(t)BR^{-1}B^T P_T(t) + Q]\Psi(t,0)x_0 = 0 , \quad \forall\, t \in [0,T] . \tag{13.4.18}$$

Making use of the fact that $\Psi(t,0)$ is nonsingular and the fact that equation (13.4.18) holds for any initial state x_0, we conclude that $P_T(\cdot)$ satisfies the Riccati equation

$$\dot{P}_T(t) = -P_T(t)A - A^T P_T(t) + P_T(t)BR^{-1}B^T P_T(t) - Q , \tag{13.4.19}$$

with boundary condition (from (13.4.11b)) $P_T(T) = 0$. Taking the transpose of both sides of equation (13.4.19), we obtain that both $P_T(t)$ and $P_T(t)^T$ satisfy the same equation and the same boundary condition. It follows from the uniqueness of the solution of differential equation that $P_T(t) = P_T(t)^T$. Therefore $P_T(t)$ is symmetric.

Thus, we have established the following result.

Theorem 13.4.1: The linear-quadratic regulator problem (13.4.1) has an unique optimal control $\hat{u}(\cdot)$ defined by

$$\hat{u}(t) = -R^{-1}B^T P_T(t)\hat{x}(t) , \tag{13.4.20a}$$

where $P_T(t)$ is the symmetric, positive definite solution of the Riccati differential equation (13.4.19) with boundary condition $P_T(T) = 0$, and $\hat{x}(\cdot)$ is the corresponding optimal trajectory satisfying following linear differential equation

$$\frac{d\hat{x}(t)}{dt} = (A - BR^{-1}B^T P_T(t))\hat{x}(t) , \text{ for } t \in [0,T] , \; x(0) = x_0 . \tag{13.4.20b}$$

■

Lemma 13.4.2 : For any $x_0 \in \mathbb{R}^n$ and $T > 0$, let $J(x_0,T)$ denote the value of the problem (13.4.1), i.e.,

$$J(x_0,T) = \min\{\int_0^T (\tfrac{1}{2}\langle x(t),Qx(t)\rangle + \tfrac{1}{2}\langle u(t),Ru(t)\rangle)dt \mid \dot{x}(t) = Ax(t) + Bu(t) \text{ , for } t \in [0,T] \text{ ,}$$

$$x(0) = x_0 \text{ , } u \in \mathbf{U}[0,T] \} \text{ .} \tag{13.4.21}$$

(a) For any $x_0 \in \mathbf{R}^n$, if $T' > T''$, then $J(x_0,T') \geq J(x_0,T'')$;

(b) There exists an $\alpha \in (0,\infty)$ such that $J(x_0,T) \leq \alpha|x_0|^2$ for all $T \in (0,\infty)$.

Proof : (a) Let $\hat{u}'(\cdot)$, $\hat{u}''(\cdot)$ be the optimal controls and $\hat{x}'(\cdot)$, $\hat{x}''(\cdot)$ be the corresponding optimal trajectories for (13.4.1), with $T = T'$, T'' respectively. Then

$$J(x_0,T') = \int_0^{T'} (\tfrac{1}{2}\langle \hat{x}'(t),Q\hat{x}'(t)\rangle + \tfrac{1}{2}\langle \hat{u}'(t),R\hat{u}'(t)\rangle)dt$$

$$\leq \int_0^{T'} (\tfrac{1}{2}\langle \hat{x}''(t),Q\hat{x}''(t)\rangle + \tfrac{1}{2}\langle \hat{u}''(t),R\hat{u}''(t)\rangle)dt$$

$$\leq \int_0^{T''} (\tfrac{1}{2}\langle \hat{x}''(t),Q\hat{x}''(t)\rangle + \tfrac{1}{2}\langle \hat{u}''(t),R\hat{u}''(t)\rangle)dt = J(x_0,T'') \text{ .} \tag{13.4.22}$$

(b) Since, by assumption, (A,B) is completely controllable, there exists an $n \times n$ matrix K such that the real parts of the eigenvalues of $A + BK$ are all less than $-\gamma < 0$, so that $|e^{t(A+BK)}| \leq \beta e^{-\gamma t}$, for some $\beta < \infty$. The corresponding trajectory is given by $x(t) = e^{t(A+BK)}x_0$ and the corresponding control is given by $u(t) = Kx(t) = Ke^{t(A+BK)}x_0$. Since the constant feedback control law, defined by $u(t) = Kx(t)$ defines a *feasible* control for (13.4.1), we must have that

$$J(x_0,T) \leq \int_0^{T} (\tfrac{1}{2}\langle e^{t(A+BK)}x_0,Qe^{t(A+BK)}x_0\rangle + \tfrac{1}{2}\langle Ke^{t(A+BK)}x_0,RKe^{t(A+BK)}x_0\rangle)dt$$

$$\leq |x_0|^2 \int_0^{\infty} \tfrac{1}{2}(|Q| + |K^TRK|)\, |e^{t(A+BK)}|^2 dt$$

$$\leq |x_0|^2 \int_0^{\infty} \tfrac{1}{2}(|Q| + |K^TRK|)\, \beta^2 e^{-2\gamma t} dt \text{ .} \tag{13.4.23}$$

Since the last integral is finite, the desired result follows. ■

Theorem 13.4.2 : (a) When $T \to \infty$, the feedback matrices $P_T(0)$, defined by (13.4.11b) converge to a symmetric, positive definite matrix $\hat{P}$ which is a solution of the algebraic Riccati equation

$$-\hat{P}A - A^T\hat{P} + \hat{P}BR^{-1}B^T\hat{P} - Q = 0 \tag{13.4.24a}$$

(b) The matrix $\hat{K} = -R^{-1}B^T\hat{P}$ defines a stabilizing feedback-control law for the system

$$\dot{x} = Ax + Bu \tag{13.4.24b}$$

(c) The state feedback control $\hat{u}(t) = \hat{K}\hat{x}(t)$ is the optimal control for the infinite horizon linear-quadratic regulator problem

$$\min\{ \int_0^{+\infty} (\tfrac{1}{2}\langle x(t), Qx(t)\rangle + \tfrac{1}{2}\langle u(t), Ru(t)\rangle) dt \mid \dot{x}(t) = Ax(t) + Bu(t) \text{ , for } t \in [0,+\infty) \text{ ,}$$

$$x(0) = x_0 \text{ , } u \in \mathbf{U}[0,+\infty) \} \text{ .} \tag{13.4.24c}$$

Proof : (a) By Lemma 13.4.2 and (13.4.11c), for any $x_0 \in \mathbb{R}^n$, $\{ \langle x_0, K_T(0)x_0\rangle \}_{T>0}$ is a monotone increasing sequence which is bounded from above. Hence the sequence $\{ \langle x_0, P_T(0)x_0\rangle \}_{T>0}$ converges. It also follows from Lemma 13.4.2 and (13.4.11c) that $\langle x_0, P_T(0)x_0\rangle \leq \alpha |x_0|^2$ for all $T > 0$ and all $x_0 \in \mathbb{R}^n$. Hence we must have that $|P_T(0)| \leq \alpha$ for all $T > 0$. Since the matrices $P_T(0) \in \mathbb{R}^{n\times n}$, it follows that the sequence $\{ P_T(0) \}_{T>0}$ must have accumulation points, which must all be symmetric, positive definite matrices. Suppose that P', P'' are two accumulation points of this sequence. Then, for any $x_0 \in \mathbb{R}^n$, because $\{ \langle x_0, P_T(0)x_0\rangle \}_{T>0}$ converges as $T \to \infty$, we must have that $\langle x_0, (P' - P'')x_0\rangle = 0$. Since the matrices P', P'' are symmetric, we conclude that $P' = P'' \triangleq \hat{P}$, and hence that $P_T(0) \to \hat{P}$ as $T \to \infty$.

Since the Riccati equation (13.4.19) is time invariant, we see that for every $T > 0$, $P_T = P(-T,0)$, where $P(t,0)$ is the solution of (13.4.19) from the initial condition $P(0,0) = 0$. Since $P_T \to \hat{P}$, a constant, as $T \to \infty$, we must have that $\dot{P}(-T,0) \to 0$ as $t \to \infty$. It now follows from (13.4.19) that $\hat{P}$ must satisfy (13.4.24a).

(b) Suppose that $\hat{K}$ does not stabilize the system. Then there exists a nonzero $x_0 \in \mathbb{C}^n$ and $\lambda \in \mathbb{C}$ with $\mathrm{Re}(\lambda) \geq 0$ such that

$$(A - BR^{-1}B^T\hat{P})x_0 = \lambda x_0 \text{ .} \tag{13.4.25}$$

Let $x(t)$ be the solution of the linear differential equation

$$\dot{x}(t) = (A - BR^{-1}B^T\hat{P})x(t), \quad t \in [0,+\infty), \quad x(0) = x_0 \text{ .} \tag{13.4.26}$$

Because of (13.4.25), $x(t) = e^{\lambda t}x_0$. Let $x^*(t)$ denote the complex conjugate transpose of $x(t)$. Then, making use of (13.4.26) and the fact that $\hat{P}$ satisfies (13.4.24a), we obtain

$$\frac{d}{dt}(x^*(t)\hat{P}x(t)) = \dot{x}^*(t)\hat{P}x(t) + x^*\hat{P}\dot{x}(t)$$

$$= x^*(t)(A^T - \hat{P}BR^{-1}B^T)\hat{P}x(t) + x^*(t)\hat{P}(A - BR^{-1}B^T\hat{P})x(t)$$

$$= - \|Q^{1/2}x(t)\|^2 - \|R^{-1/2}B^T\hat{P}x(t)\|^2 . \tag{13.4.27}$$

Substituting $x(t) = e^{\lambda t}x_0$ into (13.4.27), we find that

$$2\mathrm{Re}(\lambda)e^{2\mathrm{Re}(\lambda)t}x_0^*\hat{P}x_0 = - e^{2\mathrm{Re}(\lambda)t}\|Q^{1/2}x_0\|^2 - e^{2\mathrm{Re}(\lambda)t}\|R^{-1/2}B^T\hat{P}x_0\|^2 . \tag{13.4.28}$$

Since $\mathrm{Re}(\lambda) \geq 0$, the left hand side of (13.4.28) is non-negative. Hence $Q^{1/2}x_0 = 0$ and $R^{-1/2}B^T\hat{P}x_0 = 0$. Therefore, because of (13.2.25), $Ax_0 = \lambda x_0$, and hence, since $Q^{1/2}x_0 = 0$, we have obtained a contradiction of the fact that $(A, Q^{1/2})$ is completely observable.

(c) Let $u(\cdot)$ be any feasible control and let $x(\cdot)$ be the corresponding trajectory for (13.4.24c). Then, it follows from Corollary 13.4.1 that for any $T \geq 0$

$$\int_0^{+\infty} (\tfrac{1}{2}\langle x(t), Qx(t)\rangle + \tfrac{1}{2}\langle u(t), Ru(t)\rangle)dt \geq \int_0^{T} (\tfrac{1}{2}\langle x(t), Qx(t)\rangle + \tfrac{1}{2}\langle u(t), Ru(t)\rangle)dt$$

$$\geq \tfrac{1}{2}\langle x_0, P_T(0)x_0\rangle . \tag{13.4.29}$$

Since $P_T(0) \to \hat{P}$ as $T \to \infty$, we find that

$$\int_0^{+\infty} (\tfrac{1}{2}\langle x(t), Qx(t)\rangle + \tfrac{1}{2}\langle u(t), Ru(t)\rangle)dt \geq \tfrac{1}{2}\langle x_0, \hat{P}x_0\rangle . \tag{13.4.30}$$

On the other hand, making use of (13.4.24a) and the fact that $\hat{u}(t) = R^{-1}B^T\hat{P}\hat{x}(t)$, we obtain that for any $T \geq 0$

$$\int_0^{T} (\tfrac{1}{2}\langle \hat{x}(t), Q\hat{x}(t)\rangle + \tfrac{1}{2}\langle \hat{u}(t), R\hat{u}(t)\rangle)dt = \int_0^{T} \tfrac{1}{2}\langle \hat{x}(t), (Q + \hat{P}BR^{-1}B^T\hat{P})\hat{x}(t)\rangle dt$$

$$= - \int_0^{T} (\tfrac{1}{2}\langle \hat{x}(t), \hat{P}(A - BR^{-1}B^T\hat{P})\hat{x}(t)\rangle + \tfrac{1}{2}\langle (A - BR^{-1}B^T\hat{P})\hat{x}(t), \hat{P}x(t)\rangle dt . \tag{13.4.31}$$

Hence, because $d\hat{x}(t)/dt = (A - BR^{-1}B^T\hat{P})\hat{x}(t)$, we find that

$$\int_0^{T} (\tfrac{1}{2}\langle \hat{x}(t), Q\hat{x}(t)\rangle + \tfrac{1}{2}\langle \hat{u}(t), R\hat{u}(t)\rangle)dt = - \int_0^{T} (\tfrac{1}{2}\langle \frac{d}{dt}\hat{x}(t), \hat{P}\hat{x}(t)\rangle + \tfrac{1}{2}\langle \hat{x}(t), \hat{P}\frac{d}{dt}\hat{x}(t)\rangle) dt$$

$$= - \int_0^{T} \tfrac{1}{2}\frac{d}{dt}\langle \hat{x}(t), \hat{P}\hat{x}(t)\rangle dt$$

$$= \tfrac{1}{2}\langle x_0, \hat{P}x_0\rangle - \tfrac{1}{2}\langle \hat{x}(T), \hat{P}\hat{x}(T)\rangle . \tag{13.4.32}$$

Since $\hat{K}$ stabilizes the system, $\hat{x}(T) \to 0$ as $T \to \infty$. Hence, letting $T \to \infty$, we obtain that the cost corresponding to $\hat{u}(\cdot)$ and $\hat{x}(\cdot)$ equals $\tfrac{1}{2}\langle x_0, \hat{P}x_0\rangle$, which was shown to be a lower bound for the optimal cost of (13.4.24c). Therefore, $\hat{u}(\cdot)$ is the optimal control for (13.4.24c).