Chapter 0

Mathematical Preliminaries

0.1 Norms

Throughout this course we will be working with the vector space $\mathbb{R}^n$. For this reason we begin with a brief review of its metric space properties

DEFINITION 0.1.1 (VECTOR NORM) *A function $\nu : \mathbb{R}^n \to \mathbb{R}$ is a vector norm on $\mathbb{R}^n$ if*

i. $\nu(x) \geq 0 \ \forall \ x \in \mathbb{R}^n$ *with equality iff* $x = 0$.

ii. $\nu(\alpha x) = |\alpha|\nu(x) \ \forall \ x \in \mathbb{R}^n \ \alpha \in \mathbb{R}$

iii. $\nu(x + y) \leq \nu(x) + \nu(y) \ \forall \ x, y \in \mathbb{R}^n$

We usually denote $\nu(x)$ by $\|x\|$. Norms are convex functions.

EXAMPLE: l_p norms

$$\begin{array}{rcl} \|x\|_p & := & (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}, \quad 1 \leq p < \infty \\ \|x\|_\infty & = & \max_{i=1,\ldots,n} |x_i| \end{array}$$

– $P = 1, 2, \infty$ are most important cases

$\|x\|_1 = 1$ $\|x\|_2 = 1$ $\|x\|_\infty = 1$

– The unit ball of a norm is a convex set.

0.1.1 Equivalence of Norms

$$\alpha(p,q)\|x\|_q \le \|x\|_p \le \beta(p,q)\|x\|_q$$

$\alpha(p,q)$

p \ q	1	2	3
1	1	1	1
2	$n^{-\frac{1}{2}}$	1	1
3	n^{-1}	$n^{-\frac{1}{2}}$	1

$\beta(p,q)$

p \ q	1	2	3
1	1	$n^{\frac{1}{2}}$	n
2	1	1	$n^{\frac{1}{2}}$
3	1	1	1

0.2 Open, Closed, and Compact Sets

– A subset $D \subset \mathbb{R}^n$ is said to be <u>open</u> if for every $x \in D$ there exists $\epsilon > 0$ such that $x + \epsilon\mathbb{B} \subset D$ where

$$x + \epsilon\mathbb{B} = \{x + \epsilon u : u \in \mathbb{B}\}$$

and $\mathbb{B}$ is the unit ball of some given norm on $\mathbb{R}^n$.

– A point $\overline{x}$ is said to be a cluster point (or accumulation point) of the set $D \subset \mathbb{R}^n$ if

$$(\overline{x} + \epsilon\mathbb{B}) \cap D \neq \emptyset$$

for every $\epsilon > 0$.

– A subset $D \subset \mathbb{R}^n$ is said to be closed if it contains all of its cluster points.

– A subset $D \subset \mathbb{R}^n$ is said to be bounded if there exists $m > 0$ such that

$$\|x\| \le m \text{ for all } x \in D.$$

– A subset $D \subset \mathbb{R}^n$ is said to be compact, if it is closed and bounded.

FACT: [Bolzano–Weierstrass Compactness Theorem] A set $D \subset \mathbb{R}^n$ is compact if and only if every infinite subset of D has a cluster point in D.

0.3 Continuity and the Existence of Extrema

– The mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is said to be continuous at the point $\overline{x}$ if

$$\lim_{\|x-\overline{x}\|\to 0} \|F(x) - F(\overline{x})\| = 0,$$

or equivalently, for every $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\|F(x) - F(\overline{x})\| < \epsilon$$

whenever $\|x - \overline{x}\| < \delta$. The function F is said to be continuous on a set $D \subset \mathbb{R}^n$ if F is continuous at every point of D.

Weierstrass Extreme Value Theorem *Every continuous function on a compact set attains its extreme values on that set.*

0.4 Dual Norms

Let $\|\cdot\|$ be a given norm on $\mathbb{R}^n$ with associated closed unit ball $\mathbb{B}$. For each $x \in \mathbb{R}^n$ define

$$\|x\|_0 := \max\{x^T y : \|y\| \le 1\}.$$

Since the transformation $y \mapsto x^T y$ is continuous (in fact, linear) and $\mathbb{B}$ is compact, Weierstrass's Theorem says that the maximum in the definition of $\|x\|_0$ is attained. Thus, in particular, the function $x \to \|x\|_0$ is well defined and finite-valued. Indeed, the mapping defines a norm on $\mathbb{R}^n$. This norm is said to be the norm dual to the norm $\|\cdot\|$. Thus, every norm has a norm dual to it.

We now show that the mapping $x \mapsto \|x\|_0$ is a norm.

(a) It is easily seen that $\|x\|_0 = 0$ if and only if $x = 0$. If $x \neq 0$, then

$$\|x\|_0 = \max\{x^T y : \|y\| \le 1\} \ge x^T \left(\frac{x}{\|x\|}\right) = \frac{\|x\|_2}{\|x\|} > 0.$$

(b) From (a), $\|0 \cdot x\|_0 = 0 = 0 \cdot \|x\|_0$. Next suppose $\alpha \in \mathbb{R}$ with $\alpha \neq 0$. Then

$$\begin{aligned}
\|\alpha x\|_0 &= \max\{x^T(\alpha y) : \|y\| \le 1\}, (z = \alpha y) \\
&= \max\left\{x^T z : 1 \le \left\|\tfrac{z}{\alpha}\right\| = \tfrac{1}{|\alpha|}\|z\| = \left\|\tfrac{z}{|\alpha|}\right\|\right\}, \left(w = \tfrac{z}{|\alpha|}\right) \\
&= \max\{x^T(|\alpha| z) : 1 \ge \|w\|\} \\
&= |\alpha| \, \|x\|_0.
\end{aligned}$$

In order to establish the triangle inequality, we make use of the following elementary, but very useful, fact.

FACT: If $f : \mathbb{R}^n \to \mathbb{R}$ and $C \subset D \subset \mathbb{R}^n$, then

$$\sup_{x \in C} f(x) \leq \sup_{x \in D} f(x).$$

That is, the supremum over a larger set must be larger. Similarly, the infimum over a larger set must be smaller.

(c)
$$\begin{aligned} \|x+z\|_0 &= \max\{x^T y + z^T y : \|y\| \leq 1\} \\ &= \max\left\{ x^T y_1 + z^T y_2 : \begin{array}{l} \|y_1\| \leq 1 \\ \|y_2\| \leq 1 \end{array}, y_1 = y_2 \right\} \end{aligned}$$

(max over a larger set)
$$\begin{aligned} &= \leq \max\{x^T y_1 + z^T y_2 : \|y_1\| \leq 1, \|y_2\| \leq 1\} \\ &= \|x\|_0 + \|z\|_0 \end{aligned}$$

FACTS:

(i) $x^T y \leq \|x\| \, \|y\|_0$ (apply definition)

(ii) $\|x\|_\infty = \|x\|$

(iii) $(\|x\|_p)_0 = \|x\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$, $1 \leq p \leq \infty$

(iv) Hölder's Inequality: $|x^T y| \leq \|x\|_p \|y\|_q$

$$\frac{1}{p} + \frac{1}{q} = 1$$

(v) Cauchy-Schwartz Inequality:

$$|x^T y| \leq \|x\|_2 \|y\|_2$$

0.5 Operators

0.5.1 Operator Norms

$A \in \mathbb{R}^{m \times n}$

$$\|A\|_{(a,b)} = \max\{\|Ax\|_{(a)} : \|x\|_{(b)} \leq 1\}$$

EXAMPLE:
$$\begin{aligned} \|A\|_2 &= \max\{\|Ax\|_2 : \|x\|_2 \le 1\} \\ \|A\|_\infty &= \max\{\|Ax\|_\infty : \|x\|_\infty \le 1\} \\ &= \max_{1\le i\le m} \textstyle\sum_{j=1}^n |a_{ij}|, \text{max row sum} \\ \|A\|_1 &= \max\{\|Ax\|_1 : \|x\|_1 \le 1\} \\ &= \max_{1\le j\le n} \textstyle\sum_{i=1}^m |a_{ij}|, \text{max column sum} \end{aligned}$$

FACT: $\|Ax\|_{(a)} \le \|A\|_{(a,b)}\|x\|_{(b)}$.

(a) $\|A\| \ge 0$ with equality $\Leftrightarrow \|Ax\| = 0 \ \forall\ x$ or $A \equiv 0$.

(b)
$$\begin{aligned} \|\alpha A\| &= \max\{\|\alpha Ax\| : \|x\| \le 1\} \\ &= \max\{|\alpha|\,\|Ax\| : \|\alpha\| \le 1\} = |\alpha|\,\|A\| \end{aligned}$$

(c)
$$\begin{aligned} \|A+B\| &= \max\{\|Ax+Bx\| : \|x\| \le 1\} \le \max\{\|Ax\| + \|Bx\|\, A \le 1\} \\ &= \max\{\|Ax_1\| + \|Bx_2\| : x_1 = x_2, \|x_1\| \le 1, \|x_2\| \le 1\} \\ &\le \max\{\|Ax_1\| + \|Bx_2\| : \|x_1\| \le 1, \|x_2\| \le 1\} \\ &= \|A\| + \|B\| \end{aligned}$$

0.5.2 Spectral Radius

$A \in \mathbb{R}^{n\times n}$

$$\rho(A) := \max\{|\lambda| : \lambda \in \Sigma(A)\}$$

$\Sigma(A) = \{\lambda \in \mathbb{C} : Ax = \lambda x \text{ for some } x \ne 0)$.

$\rho(A) \sim$ spectral radius of A

$\Sigma(A) \sim$ spectrum of A

FACT:

(i) $\|A\|_2 = (\rho(A^TA))^{\frac{1}{2}}$

(ii) $\rho(A) < 1 \Leftrightarrow \lim_{k\to\infty} A^k = 0$

(iii) $\rho(A) < 1 \Rightarrow (I-A)^{-1} = \sum_{i=0}^{\infty} A^i$ (Neumann Lemma)

0.5.3 Condition number

$A \in \mathbb{R}^{n\times n}$

$$\kappa(A) = \begin{cases} \|A\|\,\|A^{-1}\| & \text{if } A^{-1} \text{ exists} \\ \infty & \text{otherwise} \end{cases}$$

FACT: [Error estimates in the solution of linear equations] If $Ax_1 = b$ and $Ax_2 = b + e$, then

$$\frac{\|x_1 - x_2\|}{\|x_1\|} \leq \kappa(A)\frac{\|e\|}{\|b\|}$$

PROOF: $\|b\| = \|Ax_1\| \leq \|A\|\,\|x_1\| \Rightarrow \frac{1}{\|x_1\|} \leq \frac{\|A\|}{\|b\|}$, so

$$\frac{\|x_1 - x_2\|}{\|x_1\|} \leq \frac{\|A\|}{\|b\|}\|A^{-1}(A(x_1 - x_2)\| \leq \|A\|\,\|A^{-1}\|\frac{1}{\|b\|}\|Ax_1 - Ax_2\|$$

■

0.5.4 The Frobenius Norm

There is one further norm for matrices that is very useful. It is called the Frobenius norm. Observe that we can identify $\mathbb{R}^{m\times n}$ with $\mathbb{R}^{(mn)}$ by simply stacking the columns of a matrix one on top of the other to create a very long vector in $\mathbb{R}^{(mn)}$. The mapping from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^{(mn)}$ defined in this way is denoted by $\text{vec}\,(\cdot)$. The Frobenius norm of a matrix $A \in \mathbb{R}^{m\times n}$ is then the 2-norm of $\text{vec}\,(A)$. It can be verified that

$$\|A\|_F = \text{tr}\,(A^T A).$$

0.6 Review of Differentiation

1) Let $F : \mathbb{R}^n \to \mathbb{R}^m$ and let $x, d \in \mathbb{R}^n$. If the limit

$$\lim_{t\downarrow 0} \frac{F(x + td) - F(x)}{t} =: F'(x; d)$$

exists, it is called the directional derivative of F at x in the direction h. If this limit exists for all $d \in \mathbb{R}^n$ and is linear in the d argument,

$$F'(x; \alpha d_1 + \beta d_2) = \alpha F'(x; d_1) + \beta F'(x; d_2),$$

then F is said to be Gâteaux differentiable at x.

2) Let $F : \mathbb{R}^n \to \mathbb{R}^m$ and let $x \in \mathbb{R}^n$. If there exists $J \in \mathbb{R}^{m\times n}$ such that

$$\lim_{\|y-x\|\to 0} \frac{\|F(y) - (F(x) + J(y-x))\|}{\|y-x\|} = 0,$$

then F is said to be Fréchet differentiable at x and J is said to be its "Fréchet derivative". We denote J by $J = F'(x)$.

FACTS:

(i) If $F'(x)$ exists, it is unique.

(ii) If $F'(x)$ exists, then $F'(x;d)$ exists for all d and

$$F'(x;d) = F'(x)d.$$

(iii) If $F'(x)$ exists, then F is continuous at x.

(iv) (Matrix Representation)

Suppose $F'(x)$ exists for all x near $\overline{x}$ and that the mapping $x \mapsto F'(x)$ is continuous at $\overline{x}$,

$$\lim_{\|x-\overline{x}\|\to 0} \|F'(x) - F'(\overline{x})\| = 0,$$

then $\partial F_i/\partial x_j$ exist for each $i = 1,\dots,m$, $j = 1,\dots,n$ and with respect to the standard basis the linear operator $F'(\overline{x})$ has the representation

$$\nabla F(\overline{x}) = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} & \dots & \frac{\partial F_1}{\partial x_n} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} & \dots & \frac{\partial F_2}{\partial x_n} \\ \vdots & & & \\ \frac{\partial F_n}{\partial x_1} & \dots & \dots & \frac{\partial F_m}{\partial x_n} \end{bmatrix}^T = \left[\frac{\partial F_i}{\partial x_j}\right]^T$$

where each partial derivative is evaluated at $\overline{x} = (\overline{x}_1,\dots,\overline{x}_n)^T$. This matrix is called the Jacobian matrix for F at $\overline{x}$.

NOTATION: For $f : \mathbb{R}^n \to \mathbb{R}$, $f'(x) = \left[\frac{\partial f_1}{\partial x_1},\dots,\frac{\partial f_*}{\partial x_n}\right]$ we write $\nabla f(x) = f'(x)^T$.

(v) If $F : \mathbb{R}^n \to \mathbb{R}^m$ has continuous partials $\partial F_i/\partial x_i$ on an open set $D \subset \mathbb{R}^n$, then F is differentiable on D. Moreover, in the standard basis the matrix representation for $F'(x)$ is the Jacobian of F at x.

(vi) (Chain Rule) Let $F : A \subset \mathbb{R}^m \to \mathbb{R}^k$ be differentiable on the open set A and let $G : B \subset \mathbb{R}^k \to \mathbb{R}^n$ be differentiable on the open set B. If $F(A) \subset B$, then the composite function $G \circ F$ is differentiable on A and

$$(G \circ F)'(x_0) = G'(F(x_0)) \circ F'(x_0).$$

REMARKS: Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be differentiable. If $L(\mathbb{R}^n, \mathbb{R}^m)$ denotes the set of linear maps from $\mathbb{R}^n$ to $\mathbb{R}^m$, then

$$F' : \mathbb{R}^n \to L(\mathbb{R}^n, \mathbb{R}^m).$$

(In a standard basis we usually identify $L(\mathbb{R}^n, \mathbb{R}^m)$ with $\mathbb{R}^{m\times n}$.) Therefore hierarchy for higher derivatives:

$$\begin{array}{lcll} F & : & \mathbb{R}^n \to \mathbb{R}^m & \\ F' & : & \mathbb{R}^n \to L(\mathbb{R}^n, \mathbb{R}^m) & \approx \quad \mathbb{R}^{m\times n} \\ F'' & : & \mathbb{R}^n \to L(\mathbb{R}^n, L(\mathbb{R}^n, \mathbb{R}^m)) & \approx \quad \mathbb{R}^{m\times n\times n} \\ F''' & : & \mathbb{R}^n \to L(\mathbb{R}^n, L(\mathbb{R}^n, L(\mathbb{R}^n, \mathbb{R}^m))) & \approx \quad \mathbb{R}^{m\times n\times n\times n} \\ & \vdots & & \end{array}$$

(v) The Mean Value Theorem:

(a) If $f : \mathbb{R} \to \mathbb{R}$ is differentiable, then for every $x, y \in \mathbb{R}$ there exists z between x and y such that

$$f(y) = f(x) + f'(z)(y - x).$$

(b) If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then for every $x, y \in \mathbb{R}$ there is a $z \in [x, y]$ such that

$$f(y) = f(x) + \nabla f(z)^T (y - x).$$

(c) If $F : \mathbb{R}^n \to \mathbb{R}^m$ continuously differentiable, then for every $x, y \in \mathbb{R}$

$$\|F(y) - F(x)\| \leq \left[\sup_{z\in[x,y]} \|F'(z)\|\right] \|x - y\|.$$

PROOF OF (b): Set $\varphi(t) = f(x + t(y - x))$. Then, by the chain rule, $\varphi'(t) = \nabla f(x + t(y - x))^T (y - x)$ so that φ is differentiable. Moreover, $\varphi : \mathbb{R} \to \mathbb{R}$. Thus, by (a), there exists $\bar{t} \in (0, 1)$ such that

$$\varphi(1) = \varphi(0) + \varphi'(\bar{t})(1 - 0),$$

or equivalently,

$$f(y) = f(x) + \nabla f(z)^T (y - x)$$

where $z = x + \bar{t}(y - x)$. ■

0.6.1 The Implicit Function Theorem

Let $F : \mathbb{R}^{n+m} \to \mathbb{R}^n$ be continuously differentiable on an open set $E \subset \mathbb{R}^{n+m}$. Further suppose that there is a point $(\bar{x}, \bar{y}) \in \mathbb{R}^{n+m}$ at which $F(\bar{x}, \bar{y}) = 0$. If $\nabla_x F(\bar{x}, \bar{y})$ is invertable, then there exist open sets $U \subset \mathbb{R}^{n+m}$ and $W \subset \mathbb{R}^m$, with $(\bar{x}, \bar{y}) \in U$ and $\bar{y} \in W$, having the following property:
To every $y \in W$ corresponds a unique $x \in \mathbb{R}^n$ such that

$$(x, y) \in U \qquad \text{and} \qquad F(x, y) = 0\,.$$

Moreover, if x is defined to be $G(y)$, then G is a continuously differentiable mapping of W into $\mathbb{R}^n$ satisfying

$$G(\bar{y}) = \bar{x}, \quad F(G(y), x) = 0 \ \forall \ y \in W, \quad \text{and} \quad G'(\bar{y}) = -(\nabla_x F(\bar{x}, \bar{y}))^{-1} \nabla_y F(\bar{x}, \bar{y}) \ .$$

0.6.2 Some facts about the Second Derivative

Let $f : \mathbb{R}^n \to \mathbb{R}$ so that $f' : \mathbb{R}^n \to L(\mathbb{R}^n, \mathbb{R})(\approx \mathbb{R}^{n\times 1} = \mathbb{R}^2)$ and

$$f'' : \mathbb{R}^n \to L(\mathbb{R}^n, L(\mathbb{R}^n, \mathbb{R}))(\approx \mathbb{R}^{n\times n\times 1} = \mathbb{R}^{n\times n}).$$

(i) If f'' exists and is continuous at x_0, then in the standard basis

$$f''(x_0) \approx \nabla^2 f(x_0) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]_{x=x_0}$$

Moreover, $\frac{\partial f}{\partial x_i \partial x_j} = \frac{\partial f}{\partial x_j \partial x_i}$ for all $i, j = 1, \ldots, n$. The matrix $\nabla^2 f(x_2)$ is called the Hessian of f at x_0. It is a symmetric matrix.

(ii) Second-Order Taylor Theorem:

If $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable on an open set containing $[x, y]$, then there is a $z \in [x, y]$ such that

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x).$$

We also obtain

$$\|f(7) - (f(x) + f'(x)(y - x))\| \leq \frac{1}{2}\|x - y\|^2 \sup_{z\in[x,y]} \|f''(z)\|.$$

0.6.3 Integration

Let $f : \mathbb{R}^n \to \mathbb{R}^1$ be differentiable and set $\varphi(t) = f(x + t(y - x))$ so that $\varphi : \mathbb{R} \to \mathbb{R}$. Then

$$\begin{aligned} f(y) - f(x) &= \varphi(1) - \varphi(0) = \textstyle\int_0^1 \varphi'(t)dt \\ &= \textstyle\int_0^1 \nabla f(x + t(y - x))^T (y - x)dt \end{aligned}$$

Similarly, if $F : \mathbb{R}^n \to \mathbb{R}^m$, then

$$\begin{aligned} F(y) - F(x) &= \begin{bmatrix} \int_0^1 \nabla F_1(x + t(y - x))^T (y - x)dt \\ \vdots \\ \int_0^1 \nabla F_m(x + t(y - x))^T (y - x)dt \end{bmatrix} \\ &= \textstyle\int_0^1 F'(x + t(y - x))(y - x)dt \end{aligned}$$

0.6.4 More Facts about Continuity

Let $F : \mathbb{R}^n \to \mathbb{R}^m$.

– We say that F is continuous on a set $D \subset \mathbb{R}^n$ if for every $x \in D$ and $\epsilon > 0$ there exists a $\delta(x, \epsilon) > 0$ such that

$$\|F(y) - F(x)\| \leq \epsilon \text{ whenever } \|y - x\| \leq \delta(x, \epsilon).$$

– We say that F is uniformly continuous on $D \subset \mathbb{R}^n$ if for every $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that

$$\|F(y) - F(x)\| \leq \epsilon \text{ whenever } \|y - x\| \leq \delta(\epsilon).$$

FACT: If F is continuous on a compact set $D \subset \mathbb{R}^n$, then F is uniformly continuous on D.

– We say that F is Lipschitz continuous on a set $D \subset \mathbb{R}^n$ if there exists a constant $K \geq 0$ such that

$$\|F(x) - F(y)\| \leq K\|x - y\|$$

for all $x, y \in D$.

FACT: Lipschitz continuity implies uniform continuity.

PROOF: $\delta = \epsilon/K$. ■

EXAMPLES:

1. $f(x) = x^{-1}$ is continuous on $(0, 1)$, but it is not uniformly continuous on $(0, 1)$.
2. $f(x) = \sqrt{x}$ is uniformly continuous on $[0, 1]$, but it is not Lipschitz continuous on $[0, 1]$.

FACT: If F' exists and is continuous on a compact convex set $D \subset \mathbb{R}^m$, then F is Lipschitz continuous on D.

PROOF: Mean value Theorem:

$$\|F(x) - F(y)\| \leq \left(\sup_{z \in [x,y]} \|F'(z)\| \right) \|x - y\|.$$

Lipschitz continuity is almost but not quite a differentiability hypothesis. The Lipschitz constant provides bounds on rate of change.

■

Quadratic Bound Lemma

LEMMA 0.6.1 *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be such that F' is Lipschitz continuous on the convex set $D \subset \mathbb{R}^n$. Then*

$$\|F(y) - (F(x) + F'(x)(y-x))\| \leq \frac{K}{2}\|y-x\|^2$$

for all $x, y \in D$ where K is a Lipschitz constant for F' on D.

PROOF:
$$\begin{aligned} F(y) - F(x) - F'(x)(y-x) &= \int_0^1 F'(x+t(y-x))(y-x)dt - F'(x)(y-x) \\ &= \int_0^1 [F'(x+t(y-x)) - F'(x)](y-x)dt \end{aligned}$$

$$\begin{aligned} \|F(y) - (F(x) + F'(x)(y-x))\| &= \| \int_0^1 [F'(x+t(y-x)) - F'(x)](y-x)dt\| \\ &\leq \int_0^1 \|(F'(x+t(y-x) - F'(x))(y-x)\|dt \\ &\leq \int_0^1 \|F'(x+t(y-x)) - F'(x)\| \, \|y-x\|dt \\ &\leq \int_0^1 Kt\|y-x\|^2 dt \\ &= \tfrac{K}{2}\|y-x\|^2. \end{aligned}$$

■

Extended Quadratic Bound Lemma

LEMMA 0.6.2 *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable in an open convex set $D \subset \mathbb{R}^n$. If we assume that F' is Lipschitz continuous in D with Lipschitz constant $K > 0$, then for all $x, y, z \in D$ we have*

$$\begin{aligned} \|F(y) - F(x) &- F'(z)(y-x)\| \\ &\leq K\frac{\|x-z\| + \|y-z\|}{2}\|x-y\| \end{aligned}$$

PROOF: Just as in the proof of the quadratic bound lemma

$$F(y) - F(x) - F'(z)(y-x) = \int_0^1 (F'(x+t(y-x)) - F'(z))(y-x)dt.$$

Therefore,

$$\begin{aligned}
\|F(y) - F(x) - F'(z)(y-x)\| &\leq \|y-x\| \int_0^t \|x + t(y-x) - z\| dt \\
&= \|y-x\| \int_0^t K\|t(y-z) + (1-t)(x-z)\| dt \\
&\leq \|y-x\| K \int_0^t t\|y-z\| + (1-t)\|x-z\| dt \\
&= K \frac{\|y-z\| + \|x-z\|}{2} \|y-x\|.
\end{aligned}$$

■

0.6.5 Some Facts about Symmetric Matrices

Let $H \in \mathbb{R}^{n \times n}$ be symmetric, i.e. $H^T = H$

1. There exists an orthonormal basis of eigen-vectors for H, i.e. if $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the n eigenvalues of H (not necessarily distinct), then there exist vectors $q_1, \ldots, q_n$ such that $\lambda_i q_i = H q_i \; i = 1, \ldots, n$ with $q_i^T q_j = \delta_{ij}$. Equivalently, there exists a unitary transformation $Q = \{q_1, \ldots, q_n\}$ such that

   $$H = Q \Lambda Q^T$$

   where $\Lambda = \text{diag}[\lambda_1, \ldots, \lambda_n]$.

2. $H \in \mathbb{R}^{n \times n}$ is positive semi-definite, i.e.

   $$x^T H x \geq 0 \text{ for all } x \in \mathbb{R}^n,$$

   if and only if $\forall \; \lambda \in \Sigma \left(\frac{1}{2}(H + H^T) \right) \qquad \lambda \geq 0.$

Chapter 1

Optimality Conditions: Unconstrained Optimization

1.1 Differentiable Problems

Consider the problem of minimizing the function $f : \mathbb{R}^n \to \mathbb{R}$ where f is twice continuously differentiable on $\mathbb{R}^n$:

$$\mathcal{P} \quad \begin{array}{l} \text{minimize} \\ x \in \mathbb{R}^n \end{array} \quad f(x)$$

We wish to obtain constructible first– and second–order necessary and sufficient conditions for optimality. Recall the following elementary results.

THEOREM 1.1.1 *[First– Order Necessary Conditions for Optimality]*
Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at a point $\overline{x} \in \mathbb{R}^n$. If $\overline{x}$ is a local solution to the problem $\mathcal{P}$, then $\nabla f(\overline{x}) = 0$.

PROOF: From the definition of the derivative we have that

$$f(x) = f(\overline{x}) + \nabla f(\overline{x})^T (x - \overline{x}) + o(\|x - \overline{x}\|)$$

where $\lim_{x \to \overline{x}} \dfrac{o(\|x - \overline{x}\|)}{\|x - \overline{x}\|} = 0$. Let $x := \overline{x} - t\nabla f(\overline{x})$. Then

$$0 \leq \frac{f(\overline{x} - t\nabla f(\overline{x})) - f(\overline{x})}{t} = -\|\nabla f(\overline{x})\|^2 + \frac{o(t\|\nabla f(\overline{x})\|)}{t}.$$

Taking the limit as $t \downarrow 0$ we obtain

$$0 \leq -\|\nabla f(\overline{x})\|^2 \leq 0.$$

Hence $\nabla f(\overline{x}) = 0$. ■

THEOREM 1.1.2 *[Second–Order Optimality Conditions]*
Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable at the point $\overline{x} \in \mathbb{R}^n$.

1. *(necessity) If $\overline{x}$ is a local solution to the problem $\mathcal{P}$, then $\nabla f(\overline{x}) = 0$ and $\nabla^2 f(\overline{x})$ is positive semi-definite.*

2. *(sufficiency) If $\nabla f(\overline{x}) = 0$ and $\nabla^2 f(\overline{x})$ is positive definite, then there is an $\alpha > 0$ such that $f(x) \geq f(\overline{x}) + \alpha \|x - \overline{x}\|^2$ for all x near $\overline{x}$.*

PROOF:

1. We make use of the second–order Taylor series expansion

$$f(x) = f(\overline{x}) + \nabla f(\overline{x})^T (x - \overline{x}) + \frac{1}{2}(x - \overline{x})^T \nabla^2 f(\overline{x})(x - \overline{x}) + o(\|x - \overline{x}\|^2) \; . \tag{1.1.1}$$

   Given $d \in \mathbb{R}^n$ and $t > 0$ set $x := \overline{x} + td$, plugging this into (1.1.1) we find that

$$0 \leq \frac{f(x + td) - f(x)}{t^2} = \frac{1}{2} d^T \nabla^2 f(\overline{x}) d + \frac{o(t^2)}{t^2}$$

   since $\nabla f(\overline{x}) = 0$ by Theorem 1.1.1. Taking the limit as $t \to 0$ we get that

$$0 \leq d^T \nabla^2 f(\overline{x}) d.$$

   Now since d was chosen arbitrarily we have that $\nabla^2 f(\overline{x})$ is positive semi-definite.

2. From (1.1.1) we have that

$$\frac{f(x) - f(\overline{x})}{\|x - \overline{x}\|^2} = \frac{1}{2} \frac{(x - \overline{x})^T}{\|x - \overline{x}\|} \nabla^2 f(\overline{x}) \frac{(x - \overline{x})}{\|x - \overline{x}\|} + \frac{o(\|x - \overline{x}\|^2)}{\|x - \overline{x}\|^2}. \tag{1.1.2}$$

   If $\lambda > 0$ is the smallest eigenvalue of $\nabla^2 f(\overline{x})$, choose $\epsilon > 0$ so that

$$\left| \frac{o(\|x - \overline{x}\|^2)}{\|x - \overline{x}\|^2} \right| \leq \frac{\lambda}{4} \tag{1.1.3}$$

   whenever $\|x - \overline{x}\| < \epsilon$. Then for all $\|x - \overline{x}\| < \epsilon$ we have from (1.1.2) and (1.1.3) that

$$\begin{array}{rcl} \frac{f(x) - f(\overline{x})}{\|x - \overline{x}\|^2} & \geq & \frac{1}{2}\lambda + \frac{0(\|x - \overline{x}\|^2)}{\|x - \overline{x}\|^2} \\ & \geq & \frac{1}{4}\lambda. \end{array}$$

   Consequently, if we set $\alpha = \frac{1}{4}\lambda$, then

$$f(x) \geq f(\overline{x}) + \alpha \|x - \overline{x}\|^2$$

   whenever $\|x - \overline{x}\| < \epsilon$.

■

1.2 Convex Problems

Observe that Theorem 1.1.1 establishes first–order necessary conditions while Theorem 1.1.2 establishes both second–order necessary and sufficient conditions. What about first–order sufficiency conditions? For this we introduce the following definitions.

DEFINITION 1.2.1 *[Convex Sets and Functions]*

1. *A subset $C \subset \mathbb{R}^n$ is said to be convex is for every pair of points x and y taken from C, the entire line segment connecting x and y is also contained in C, i.e.,*

$$[x, y] \subset C \qquad \text{where} \qquad [x, y] = \{(1 - \lambda)x + \lambda y \,:\, 0 \leq \lambda \leq 1\}.$$

2. *A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ is said to be convex if the set*

$$epi(f) = \{(\mu, x) \,:\, f(x) \leq \mu\}$$

is a convex subset of $\mathbb{R}^{1+n}$. In this context, we also define the set

$$dom(f) = \{x \in \mathbb{R}^n \,:\, f(x) < +\infty\}$$

to be the essential domain of f.

LEMMA 1.2.1 *The function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for every two points $x_1, x_2 \in dom(f)$ and $\lambda \in [0, 1]$ we have*

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2).$$

That is, the secant line connecting $(x_1, f(x_1))$ and $(x_2, f(x_2))$ lies above the graph of f.

EXAMPLE: The following functions are examples of convex functions: $c^T x$, $\|x\|$, e^x, x^2

The significance of convexity in optimization theory is illustrated in the following result.

THEOREM 1.2.1 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ be convex. If $\overline{x} \in \operatorname{dom}(f)$ is a local solution to the problem $\mathcal{P}$, then $\overline{x}$ is a global solution to the problem $\mathcal{P}$.*

PROOF: If $f(\bar{x}) = -\infty$ we are done, so let us assume that $-\infty < f(\bar{x})$. Suppose there is a $\widehat{x} \in \mathbb{R}^n$ with $f(\widehat{x}) < f(\overline{x})$. Let $\epsilon > 0$ be such that $f(\overline{x}) \leq f(x)$ whenever $\|x - \overline{x}\| \leq \epsilon$. Set $\lambda := \epsilon(2\|\overline{x} - \widehat{x}\|)^{-1}$ and $x_\lambda := \overline{x} + \lambda(\widehat{x} - \overline{x})$. Then $\|x_\lambda - \overline{x}\| \leq \epsilon/2$ and $f(x_\lambda) \leq (1-\lambda)f(\overline{x}) + \lambda f(\widehat{x}) < f(\overline{x})$. This contradicts the choice of ϵ, hence no such $\widehat{x}$ exists. ■

If f is a differentiable convex function, then a better result can be established. In order to obtain this result we need the following lemma.

LEMMA 1.2.2 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be convex.*

1. *Given $x \in \operatorname{dom}(f)$ and $d \in \mathbb{R}^n$ the difference quotient*

$$\frac{f(x+td) - f(x)}{t} \tag{1.2.4}$$

is a non-decreasing function of t on $(0, +\infty)$.

2. *For every $x \in \operatorname{dom}(f)$ and $d \in \mathbb{R}^n$ the directional derivative $f'(x;d)$ always exists and is given by*

$$f'(x;d) := \inf_{t>0} \frac{f(x+td) - f(x)}{t}. \tag{1.2.5}$$

3. *For every $x \in \operatorname{dom}(f)$, the function $f'(x;\cdot)$ is sublinear, i.e. $f'(x;\cdot)$ is positively homogeneous,*

$$f'(x;\alpha d) = \alpha f'(x;d) \quad \forall\ d \in \mathbb{R}^n,\ 0 \leq \alpha,$$

and subadditive,

$$f'(x;u+v) \leq f'(x;u) + f'(x;v).$$

PROOF: We assume (1.2.4) is true and show (1.2.5). If $x + td \notin \operatorname{dom}(f)$ for all $t > 0$, then the result obviously true. Therefore, we may as well assume that there is a $\bar{t} > 0$ such that $x + td \in \operatorname{dom}(f)$ for all $t \in (0, \bar{t}]$. Recall that

$$f'(x;d) := \lim_{t \downarrow 0} \frac{f(x+td) - f(x)}{t}. \tag{1.2.6}$$

Now if the difference quotient (1.2.4) is non-decreasing in t on $(0, +\infty)$, then the limit in (1.2.6) is necessarily given by the infimum in (1.2.5). This infimum always exists and so $f'(x;d)$ always exists and is given by (1.2.5).

We now prove (1.2.4). Let $x \in \operatorname{dom}(f)$ and $d \in \mathbb{R}^n$. If $x + td \notin \operatorname{dom}(f)$ for all $t > 0$, then the result is obviously true. Thus, we may assume that

$$0 < \bar{t} = \sup\{t : x + td \in \operatorname{dom}(f)\}.$$

Let $0 < t_1 < t_2 < \bar{t}$ (we allow the possibility that $t_2 = \bar{t}$ if $\bar{t} < +\infty$). Then

$$\begin{aligned} f(x + t_1 d) &= f\left(x + \left(\tfrac{t_1}{t_2}\right) t_2 d\right) \\ &= f\left[\left(1 - \left(\tfrac{t_1}{t_2}\right)\right) x + \left(\tfrac{t_1}{t_2}\right)(x + t_2 d)\right] \\ &\leq \left(1 - \tfrac{t_1}{t_2}\right) f(x) + \left(\tfrac{t_1}{t_2}\right) f(x + t_2 d). \end{aligned}$$

Hence

$$\frac{f(x + t_1 d) - f(x)}{t_1} \leq \frac{f(x + t_2 d) - f(x)}{t_2}.$$

We now show Part 3 of this result. To see that $f'(x; \cdot)$ is positively homogeneous let $d \in \mathbb{R}^n$ and $\alpha > 0$ and note that

$$f'(x; \alpha d) = \alpha \lim_{t \downarrow 0} \frac{f(x + (t\alpha)d) - f(x)}{(t\alpha)} = \alpha f'(x; d).$$

To see that $f'(x; \cdot)$ is subadditive let $u, v \in \mathbb{R}^n$, then

$$\begin{aligned} f'(x; u + v) &= \lim_{t \downarrow 0} \frac{f(x + t(u + v)) - f(x)}{t} \\ &= \lim_{t \downarrow 0} \frac{f(x + \frac{t}{2}(u + v)) - f(x)}{t/2} \\ &= \lim_{t \downarrow 0} 2 \frac{f(\frac{1}{2}(x + tu) + \frac{1}{2}(x + tv)) - f(x)}{t} \\ &\leq \lim_{t \downarrow 0} 2 \frac{\frac{1}{2} f(x + tu) + \frac{1}{2} f(x + tv) - f(x)}{t} \\ &= \lim_{t \downarrow 0} \frac{f(x + tu) - f(x)}{t} + \frac{f(x + tv) - f(x)}{t} \\ &= f'(x; u) + f(x; v)\ . \end{aligned}$$

■

From Lemma 1.2.2 we immediately obtain the following result.

THEOREM 1.2.2 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be convex and suppose that $\overline{x} \in \mathbb{R}^n$ is a point at which f is differentiable. Then $\overline{x}$ is a global solution to the problem $\mathcal{P}$ if and only if $\nabla f(\overline{x}) = 0$.*

PROOF: If $\overline{x}$ is a global solution to the problem $\mathcal{P}$, then, in particular, $\overline{x}$ is a local solution to the problem $\mathcal{P}$ and so $\nabla f(\overline{x}) = 0$ by Theorem 1.1.1. Conversely, if $\nabla f(\overline{x}) = 0$, then, by setting $t := 1$, $x := \overline{x}$, and $d := y - \overline{x}$ in (1.2.5), we get that

$$0 \leq f(y) - f(\overline{x}),$$

or $f(\overline{x}) \leq f(y)$. Since y was chosen arbitrarily, the result follows. ■

As Theorems 1.2.1 and 1.2.2 demonstrate, convex functions are very nice functions indeed. This is especially so with regard to optimization theory. Thus, it is important that we be able to recognize when a function is convex. For this reason we give the following result.

THEOREM 1.2.3 *Let $f : \mathbb{R}^n \to \mathbb{R}$.*

1. *If f is differentiable on $\mathbb{R}^n$, then the following statements are equivalent:*
   - *(a) f is convex,*
   - *(b) $f(y) \geq f(x) + \nabla f(x)^T(y - x)$ for all $x, y \in \mathbb{R}^n$*
   - *(c) $(\nabla f(x) - \nabla f(y))^T(x - y) \geq 0$ for all $x, y \in \mathbb{R}^n$.*
2. *If f is twice differentiable then f is convex if and only if f is positive semi-definite for all $x \in \mathbb{R}^n$.*

PROOF: (a) $\Rightarrow$ (b) If f is convex, then 1.2.3 holds. By setting $t := 1$ and $d := y - x$ we obtain (b).

(b) $\Rightarrow$ (c) Let $x, y \in \mathbb{R}^n$. From (b) we have

$$f(y) \geq f(x) + \nabla f(x)^T(y - x)$$

and

$$f(x) \geq f(y) + \nabla f(y)^T(x - y).$$

By adding these two inequalities we obtain (c).

(c) $\Rightarrow$ (b) Let $x, y \in \mathbb{R}^n$. By the mean value theorem there exists $0 < \lambda < 1$ such that

$$f(y) - f(x) = \nabla f(x_\lambda)^T(y - x)$$

where $x_\lambda := \lambda y + (1 - \lambda)x$. By hypothesis,

$$\begin{aligned} 0 &\leq [\nabla f(x_\lambda) - \nabla f(x)]^T(x_\lambda - x) \\ &= \lambda[\nabla f(x_\lambda) - \nabla f(x)]^T(y - x) \\ &= \lambda[f(y) - f(x) - \nabla f(x)^T(y - x)]. \end{aligned}$$

Hence $f(y) \geq f(x) + \nabla f(x)^T(y - x)$.

(b) $\Rightarrow$ (a) Let $x, y \in \mathbb{R}^n$ and set

$$\alpha := \max_{\lambda \in [0,1]} \varphi(\lambda) := [f(\lambda y + (1 - \lambda)x) - (\lambda f(y) + (1 - \lambda)f(x))].$$

We need to show that $\alpha \leq 0$. Since $[0, 1]$ is compact and φ is continuous, there is a $\lambda \in [0, 1]$ such that $\varphi(\lambda) = \alpha$. If λ equals zero or one, we are done. Hence we may as well assume that $0 < \lambda < 1$ in which case

$$0 = \varphi'(\lambda) = \nabla f(x_\lambda)^T(y - x) + f(x) - f(y)$$

where $x_\lambda = x + \lambda(y - x)$, or equivalently

$$\lambda f(y) = \lambda f(x) - \nabla f(x_\lambda)^T (x - x_\lambda).$$

But then

$$\begin{aligned} \alpha &= f(x_\lambda) - (f(x) + \lambda(f(y) - f(x))) \\ &= f(x_\lambda) + \nabla f(x_\lambda)^T (x - x_\lambda) - f(x) \\ &\leq 0 \end{aligned}$$

by (b).

2) Suppose f is convex and let $x, d \in \mathbb{R}^n$, then by (b) of Part (1),

$$f(x + td) \geq f(x) + t\nabla f(x)^T d$$

for all $t \in \mathbb{R}$. Replacing the left hand side of this inequality with its second–order Taylor expansion yields the inequality

$$f(x) + t\nabla f(x)^T d + \frac{t^2}{2} d^T \nabla^2 f(x) d + o(t^2) \geq f(x) + t\nabla f(x)^T d$$

or equivalently

$$\frac{1}{2} d^t \nabla^2 f(x) d + \frac{o(t^2)}{t^2} \geq 0.$$

Letting $t \to 0$ yields the inequality

$$d^T \nabla^2 f(x) d \geq 0.$$

Since d was arbitrary, $\nabla^2 f(x)$ is positive semi-definite.

Conversely, if $x, y \in \mathbb{R}^n$, then by the mean value theorem there is a $\lambda \in (0, 1)$ such that

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x_\lambda)(y - x)$$

where $x_\lambda = \lambda y + (1 - \lambda)x$. Hence

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

since $\nabla^2 f(x_\lambda)$ is positive semi-definite. Therefore, f is convex by (b) of Part (1).

■

We have established that $f'(x; d)$ exists for all $x \in \text{ dom}\,(f)$ and $d \in \mathbb{R}^n$, but we have not yet discussed to continuity properties of f. We give a partial result in this direction in the next lemma.

LEMMA 1.2.3 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be convex. Then f if bounded in a neighborhood of a point $\bar{x}$ if and only if f is Lipschitz in a neighborhood of $\bar{x}$.*

PROOF: If f is Lipschitz in a neighborhood of $\bar{x}$, then f is clearly bounded above in a neighborhood of $\bar{x}$. Therefore, we assume local boundedness and establish Lipschitz continuity.

Let $\epsilon > 0$ and $M > 0$ be such that $|f(x)| \leq M$ for all $x \in \bar{x} + 2\epsilon\mathbb{B}$. Set $g(x) = f(x + \bar{x}) - f(\bar{x})$. It is sufficient to show that g is Lipschitz on $\epsilon\mathbb{B}$. First note that for all $x \in 2\epsilon\mathbb{B}$

$$0 = g(0) = g(\frac{1}{2}x + \frac{1}{2}(-x)) \leq \frac{1}{2}g(x) + \frac{1}{2}g(-x),$$

and so $-g(x) \leq g(-x)$ for all $x \in 2\epsilon\mathbb{B}$. Next, let $x, y \in \epsilon\mathbb{B}$ with $x \neq y$ and set $\alpha = \|x - y\|$. Then $w = y + \epsilon\alpha^{-1}(y - x) \in 2\epsilon\mathbb{B}$, and so

$$g(y) = g\left(\frac{1}{1 + \epsilon^{-1}\alpha}x + \frac{\epsilon^{-1}\alpha}{1 + \epsilon^{-1}\alpha}w\right) \leq \frac{1}{1 + \epsilon^{-1}\alpha}g(x) + \frac{\epsilon^{-1}\alpha}{1 + \epsilon^{-1}\alpha}g(w).$$

Consequently,

$$g(y) - g(x) \leq \frac{\epsilon^{-1}\alpha}{1 + \epsilon^{-1}\alpha}(g(w) - g(x)) \leq 2M\epsilon^{-1}\alpha = 2M\epsilon^{-1}\|x - y\|.$$

Since this inequality is symmetric in x and y, we obtain the result. ■

1.3 Convex Composite Problems

Convex composite optimization is concerned with the minimization of functions of the form $f(x) := h(F(x))$ where $h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is a proper convex function and $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable. Most problems from nonlinear programming can be cast in this framework.

EXAMPLES:

(1) Let $F : \mathbb{R}^n \to \mathbb{R}^m$ where $m > n$, and consider the equation $F(x) = 0$. Since $m > n$ it is highly unlikely that a solution to this equation exists. However, one might try to obtain a *best* approximate solution by solving the problem $\min\{\|F(x)\| : x \in \mathbb{R}^n\}$. This is a convex composite optimization problem since the norm is a convex function.

(2) Again let $F : \mathbb{R}^n \to \mathbb{R}^m$ where $m > n$, and consider the inclusion $F(x) \in C$, where $C \subset \mathbb{R}^n$ is a non-empty closed convex set. One can pose this inclusion as the optimization problem $\min\{\text{dist}(F(x)|C) : x \in \mathbb{R}^n\}$. This is a convex composite optimization problem since the distance function

$$\text{dist}(y \mid C) := \inf_{z \in C} \|y - z\|$$

is a convex function.

(3) Let $F : \mathbb{R}^n \to \mathbb{R}^m$, $C \subset \mathbb{R}^n$ a non-empty closed convex set, and $f_0 : \mathbb{R}^n \to \mathbb{R}$, and consider the constrained optimization problem $\min\{f_0(x) : F(x) \in C\}$. One can approximate this problem by the unconstrained optimization problem

$$\min\{f_0(x) + \alpha \text{dist}(f(x)|C) : x \in \mathbb{R}^n\}.$$

This is a convex composite optimization problem where $h(\eta, y) = \eta + \alpha \text{dist}(y|C)$ is a convex function. The function $f_0(x) + \alpha \text{dist}(f(x)|C)$ is called an *exact penalty function* for the problem $\min\{f_0(x) : F(x) \in C\}$. We will review the theory of such functions in a later section.

Most of the first-order theory for convex composite functions is easily derived from the observation that

$$f(y) = h(F(y)) = h(F(x) + F'(x)(y - x)) + o(\|y - x\|). \tag{1.3.7}$$

This local representation for f is a direct consequence of h being locally Lipschitz:

$$\begin{array}{rcl} |h(F(y)) & - & h(F(x) + F'(x)(y - x))| \\ & \leq & K\|y - x\| \int_0^1 \|F'(x + t(y - x)) - F'(x)\| dt \end{array}$$

for some $K \geq 0$. Equation (1) can be written equivalently as

$$h(F(x + d)) = h(F(x)) + \Delta f(x; d) + o(\|d\|) \tag{1.3.8}$$

where

$$\Delta f(x; d) := h(F(x) + F'(x)d) - h(F(x)).$$

From 1.3.8, one immediately obtains the following result.

LEMMA 1.3.1 *Let $h : \mathbb{R}^n \to \mathbb{R}$ be convex and let $F : \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable. Then the function $f = h \circ F$ is everywhere directional differentiable and one has*

$$\begin{array}{rcl} f'(x; d) & = & h'(F(x); F'(x)d) \\ & = & \inf_{\lambda > 0} \frac{\Delta f(x; \lambda d)}{\lambda}. \end{array} \tag{1.3.9}$$

This result yields the following optimality condition for convex composite optimization problems.

THEOREM 1.3.1 *Let $h : \mathbb{R}^m \to \mathbb{R}$ be convex and $F : \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable. If $\bar{x}$ is a local solution to the problem $\min\{h(F(x))\}$, then $d = 0$ is a global solution to the problem*

$$\min_{d \in \mathbb{R}^n} h(F(\bar{x}) + F'(\bar{x})d). \tag{1.3.10}$$

There are various ways to test condition 1.3.8. A few of these are given below.

LEMMA 1.3.2 *Let h and F be as in Theorem 1.3.1. The following conditions are equivalent*

(a) $d = 0$ *is a global solution to 1.3.10.*

(b) $0 \leq h'(F(x); F'(x)d)$ *for all* $d \in \mathbb{R}^n$.

(c) $0 \leq \Delta f(x; d)$ *for all* $d \in \mathbb{R}^n$.

PROOF:The equivalence of (a) and (b) follows immediately from convexity. Indeed, this equivalence is the heart of the proof of Theorem 1.3.1. The equivalence of (b) and (c) is an immediate consequence of 1.3.2. ■

In the sequel, we say that $x \in \mathbb{R}^n$ satisfies the first-order condition for optimality for the convex composite optimization problem if it satisfies any of the three conditions (a)–(c) of Lemma 1.3.2.

1.3.1 A Note on Directional Derivatives

Recall that if $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then the function $f'(x; d)$ is linear in d:

$$f'(x; \alpha d_1 + \beta d_2) = \alpha f'(x; d_1) + \beta f'(x; d_2) \ .$$

If f is only assumed to be convex and not necessarily differentiable, then $f'(x; \cdot)$ is sublinear and hence convex. Finally, if $f = h \circ F$ is convex composite with $h : \mathbb{R}^m \to \mathbb{R}$ convex and $F : \mathbb{R}^n \to \mathbb{R}^m$ continuously differentiable, then, by Lemma (1.3.1), $f'(x; \cdot)$ is also sublinear and hence convex. Moreover, the approximate directional derivative $\Delta f(x; d)$ satisfies

$$\lambda_1^{-1} \Delta f(x; \lambda_1 d) \leq \lambda_2^{-1} \Delta f(x; \lambda_2 d) \qquad \text{for } 0 < \lambda_1 \leq \lambda_2,$$

by the non–decreasing nature of the difference quotients. Thus, in particular,

$$\Delta f(x; \lambda d) \leq \lambda \Delta f(x; d) \qquad \text{for all } \lambda \in [0, 1].$$

Chapter 2

Basic Convergence Theory

2.1 Global Theory

2.1.1 Line–Search Methods

In this section we consider the problem of minimizing a function $f : \mathbb{R}^n \to \mathbb{R}$. In particular, we are interested in iterative schemes of the form

$$x_{k+1} := x_k + \lambda_k d_k,$$

where it is intended that $f(x_{k+1}) < f(x_k)$. Such methods are called descent methods. The scalar λ_k is called the *step length* and the vector d_k is called the *search direction*. It is easily seen from the definition of the directional derivative that

$$\{d : f'(x; d) < 0\} \subset \{d : \exists\, \bar{\lambda} > 0, \text{ s.t. } f(x + \lambda d) < f(x)\ \forall\, \lambda \in (0, \bar{\lambda})\}.$$

Thus one way to implement a descent method is to choose the search direction from the set $\{d : f'(x_0; d) < 0\}$. For example, one could take d_k as the solution to the problem

$$\min\{f'(x_k; d) : \|d\| = 1\}. \tag{2.1.1}$$

The search direction d_k obtained in this way is called the direction of steepest descent, or the Cauchy direction. If f is differentiable at x_k and $\nabla f(x_k) \neq 0$, then the solution to (2.1.1) is

$$d_k := -\nabla f(x_k)\|\nabla f(x_k)\|^{-1}. \tag{2.1.2}$$

The Cauchy direction is only one of many choices that we will consider. The common feature in all of these methods is that

$$f'(x_k; d_k) < 0 \tag{2.1.3}$$

unless $\nabla f(x_k) = 0$. In this regard, we have the following general convergence result.

THEOREM 2.1.1 (DIFFERENTIABLE OBJECTIVE) *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $x_0 \in \mathbb{R}^n$ be such that f is differentiable on $\mathbb{R}^n$ with f' uniformly continuous on $\overline{co}\{x : f(x) \leq f(x_0)\}$. Consider the following algorithm.*

Choose $\gamma \in (0,1)$, $c \in (0,1)$. Having x_k determine x_{k+1} as follows:

1. *Let D_k be a subset of $\{d : f'(x_k : d) < 0\}$. If $D_k = \emptyset$ stop; otherwise choose $d_k \in D_k$.*

2. *Set*

$$\begin{array}{rl} \lambda_k := & \max \gamma^s \\ & \text{subject to } s \in \mathbb{N} := \{0, 1, 2, \ldots\} \\ & f(x_k + \gamma^s d_k) - f(x_k) \leq c\gamma^s f'(x_k; d_k). \end{array} \tag{2.1.4}$$

3. *Set $x_{k+1} := x_k + \lambda_k d_k$.*

If $\{x_k\}$ is the sequence generated by the algorithm, then one of the following must occur:

(i) There is a k_0 such that $D_{k_0} = \emptyset$;

(ii) $f(x_k) \downarrow -\infty$;

(iii) The sequence $\{\|d_k\|\}$ diverges to $+\infty$;

(iv) For every subsequence $J \subset \mathbb{N}$ for which $\{d_k\}_J$ is bounded, we have that $\lim_J f'(x_k; d_k) = 0$.

Remarks

1. The set D_k is introduced at each interation to represent general termination criteria. We do not specify these criteria at this time as they will depend on the specific type of problem under consideration.

2. If (ii) occurs, then the end result of the iteration is considered to be successful.

3. Depending on the structure of D_k, it is possible to prevent (iii) from occurring. For example, one could take

$$D_k := \begin{cases} -\nabla f(x_k)/\|\nabla f(x_0)\|, & \text{if } \nabla f(x_k) \neq 0; \\ \emptyset & \text{otherwise.} \end{cases}$$

   in which case $\{d_k\}$ is bounded so that $f'(x_k; d_k) \to 0$.

4. Since $c \in (0,1)$, the process of determining λ_k in (2.1.4) is finite. In order to see this simply divide the inequality in (2.1.4) by γ^s to obtain

$$\frac{f(x_k) + \gamma^s d_n) - f(x_k)}{\gamma^s} \leq cf'(x_k; d_k). \tag{2.1.5}$$

   Since the left hand side of this inequality converges to $f'(x_k; d_k)$ as $s \to \infty$ and $f'(x_i; d_k) < 0$, inequality (2.1.5) is valid for all s sufficiently large.

5. One should think of (iv) as a limiting stationarity condition. For example, if D_k is as given in remark (3) above then

$$f'(x_k; d_k) = -\|\nabla f(x_k)\|.$$

Hence (iv) implies that $\|\nabla f(x_k)\| \to 0$.

PROOF: We will assume that none of the conclusions (i)–(iv) occur and establish a contradiction. Since (iii) and (iv) do not occur, there is a subsequence $J \subset \mathbb{N}$ and a vector $\overline{d} \in \mathbb{R}^n$ with $d_k \xrightarrow{J} \overline{d}$ and $\sup_J f'(x_k; d_k) < \beta < 0$. Moreover, as (ii) does not occur, $f(x_k) \downarrow f^* \in \mathbb{R}$, and so $(f(x_{k+1}) - f(x_k)) \to 0$. Step 2 of the algorithm now implies that

$$\lambda_k f'(x_k; d_k) \to 0.$$

Therefore $\lambda \xrightarrow{J} 0$, and so with no loss of generality, $\lambda_k < 1$ for all $k \in J$. Hence

$$c\lambda_k\gamma^{-1} f'(x_k; d_k) < f(x_k + \lambda_k\gamma^{-1} d_k) - f(x_k), \tag{2.1.6}$$

for all $k \in J$. Now, since f' is uniformly continuous on $\overline{\text{co}}\{x : f(x) \leq f(x_0)\}$ we have that

$$f(x_k + \lambda_k\gamma^{-1} d_k) - f(x_k) \leq \gamma^{-1}\lambda_k[f'(x_k; d_k) + \omega(\gamma^{-1}\lambda_k\|d_k\|)], \tag{2.1.7}$$

where ω is the modulus of continuity for f'. Inequalities (2.1.6) and (2.1.7) yield the inequality

$$0 < (1 - c)\beta + \omega(\gamma^{-1}\lambda_k\|d_k\|).$$

Taking the limit over $k \in J$, we have that $\gamma^{-1}\lambda_k\|d_k\| \to 0$ and so

$$\omega(\gamma^{-1}\lambda_k\|d_k\|) \to 0.$$

This yields the contradiction

$$0 < (1 - c)\beta < 0.$$

■

This convergence result will be refered to repreatedly throughout the course. It allows us to dispense with discussions of the global behavior for various algorithms very quickly. For example, we have the following result.

COROLLARY 2.1.1.1 *Let f and $\{x_k\}$ be as in Theorem 2.1.1 and suppose that*

1. *f is bounded below, and*
2. $D_k := \begin{cases} -\nabla f(x_k)/\|\nabla f(x_k)\| & \text{if } \nabla f(x_k) \neq \emptyset \\ 0 & \text{else.} \end{cases}$

Then $\|\nabla f(x_k)\| \to 0$.

PROOF: Since $\{d_k\}$ is bounded and f is bounded below neither (ii) or (iii) of Theorem 2.1.1 can occur. If (i) occurs, then $\nabla f(x_{k_0}) = 0$; otherwise (iv) occurs in which case

$$-\|\nabla f(x_k)\| = f'(x_k; d_k) \to 0.$$

■

Observe that Theorem 2.1.1 says nothing about the convergence of the sequence $\{x_k\}$. Indeed, the sequence $\{x_k\}$ may diverge, e.g. $f(x) = e^x$. But if $\{x_k\}$ has a cluster point $\overline{x}$, then we know that $f'(x_k; d_k) \xrightarrow{J} f'(\overline{x}; \overline{d}) = 0$. Hence, depending on the limit $\overline{d}$, stationarity criteria for $\overline{x}$ can be obtained via the theorem.

We now establish a similar global convergence result for the problem of convex composite optimization.

THEOREM 2.1.2 (CONVEX COMPOSITE OBJECTIVE) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(x) = h(F(x))$ where $h : \mathbb{R}^m \to \mathbb{R}$ is convex and $F : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable. Let $x_0 \in \mathbb{R}^n$ and assume that*

(a) h is Lipschitz continuous on the set $\{y : h(y) \leq h(F(x_0))\}$, and

(b) F' is uniformly continuous on the set $\overline{co}\{x : h(F(x)) \leq h(F(x_0))\}$.

Consider the following algorithm:

Choose $\gamma \in (0,1)$ and $c \in (0,1)$. Having x_k determine x_{k+1} as follows:

1) Let D_k be a subset of $\{d : \Delta f(x_k; d) < 0\}$ where $\Delta f(x; d) := h(F(x)+F'(x)d)-h(F(x))$. If $D_k = \emptyset$ stop; otherwise choose $d_k \in D_k$.

2) Set $\lambda_k := \max \gamma^s$ *subject to $s \in \{0, 1, 2, \ldots\}$ and* $h(F(x + \gamma^s d)) \leq h(F(x)) + c\gamma^s \Delta f(x_k d_k)$.

3) Set $x_{k+1} := x_k + \lambda_k d_k$

If $\{x_k\}$ is the sequence generated by the algorithm initiated at x_0, then one of the following must occur:

(i) There is a k_0 such that $D_{k_0} = \emptyset$;

(ii) $f(x_n) \downarrow -\infty$

(iii) The sequence $\{\|d_k\|\}$ diverges to $+\infty$;

(iv) For every subsequence $J \subset \mathbb{N}$ for which $\{d_k\}_J$ is bounded, we have

$$\lim_J \Delta f(x_k; d_k) = 0.$$

PROOF: Suppose to the contrary that none of (i) – (iv) occur. Then there is a subsequence $J \subset \mathbb{N}$ such that $\{d_j\}_J$ is bounded and there is a $\beta > 0$ with

$$\sup_J \Delta f(x_j; d_j) \leq -\beta < 0.$$

Now $\{f(x_j)\}$ is a decreasing sequence that is bounded below, hence $f(x_j) \to f^*$ for some $f^* \in \mathbb{R}$. Consequently, $(f(x_{j+1}) - f(x_j)) \to 0$. The choice of λ_k implies that $\lambda_j \Delta f(x_j; d_j) \to 0$. Therefore, $\lambda_j \xrightarrow{J} 0$ so with no loss in generality we assume that $\lambda_j < 1$ for all $j \in J$. Again, the choice of λ_j implies that

$$c\lambda_j \gamma^{-1} \Delta f(x_j; d_j) \leq f(x_j + \lambda_j \gamma^{-1} d_j) - f(x_j)$$

for all $j \in J$. But,

$$\begin{array}{rcl} f(x_j + \lambda_j\gamma^{-1}d_j) - f(x_j) & & \\ \leq \lambda_j\gamma^{-1}\Delta f(x_j; d_j) + K\|F(x_j + \lambda_j\gamma^{-1}d_j) - (F(x_j) + \lambda_j\gamma^{-1}F'(x_j)d_j)\| & & \\ \leq \lambda_j\gamma^{-1}\Delta f(x_j; d_j) + K\lambda_j\gamma^{-1}\|d_j\| \int_0^1 \|F'(x_j + \tau\gamma^{-1}\lambda_j d_j) - F'(x_j)\| d\tau & & \\ \leq \lambda_j\gamma^{-1}\{\Delta f(x_j; d_j) + K\|d_j\|\omega(\gamma^{-1}\lambda_j\|d_j\|)] & & \end{array}$$

for all $j \in J$, where K is a Lipschitz constant for h and ω is the modulus of continuity for F'. Therefore,

$$\begin{array}{rcl} 0 & < & (1-c)\Delta f(x_j; d_j) + K\omega(\lambda_j\gamma^{-1}\|d_j\|)\|d_j\| \\ & \leq & (c-1)\beta + K\omega(\lambda_j\gamma^{-1}\|d_j\|)\|d_j\| \end{array}$$

for all $j \in J$. Letting $j \in J$ go to ∞, we obtain the contradiction

$$0 \leq (c-1)\beta < 0.$$

■

It should be noted that the line search procedure in Step (2) of the algorithm is finitely terminating since $f'(x; d) \leq \Delta f(x; d)$. As an illustration of how the above result can be used we consider an instance of the choice of set D_k that corresponds to steepest descent in the differentiable case if h is the identity map on $\mathbb{R}$.

COROLLARY 2.1.2.1 *Let f and $\{x_k\}$ be as in the statement of Theorem 2.1.2 and suppose that*

(a) f is bounded below, and

(b) $D_k := \arg\min\{h(F(x_k) + F'(x_k)d_k) : \|d_k\| \leq 1\}$.

Then every cluster, $\overline{x}$, point of the sequence $\{x_j\}$ satisfies

$$f'(\overline{x}; d) \geq 0 \qquad \forall\ d \in \mathbb{R}^n,$$

i.e., $\overline{x}$ satisfies first–order optimality conditions for the convex composite optimization problem.

PROOF: First note that Theorem (2.1.2) indicates that

$$\Delta f(x_j; d_j) \to 0.$$

If $J \subset \mathbb{N}$ is such that $x_j \xrightarrow{J} \overline{x}$ we can always refine J if necessary to get that $d_j \xrightarrow{J} \overline{d}$ for some $\overline{d}$ with $\|\overline{d}\| \leq 1$. But then

$$\Delta f(\overline{x}; \overline{d}) = 0$$

or

$$h(F(\overline{x}) + F'(\overline{x})\overline{d}) = h(F(\overline{x})).$$

Further note that for all $d \in \mathbb{B}$

$$h(F(x_j) + F'(x_j)d_j) \leq h(F(x_j) + F'(x_j)d).$$

Hence, in the limit over J

$$h(F(\overline{x}) + F'(\overline{x})\overline{d}) \leq h(F(\overline{x}) + F'(\overline{x})d).$$

Consequently, $\overline{d} \in \arg\min\{h(F(\overline{x}) + F'(\overline{x})d) : \|d\| \leq 1\}$. But $h(F(\overline{x})) = h(F(\overline{x}) + F'(\overline{x})\overline{d})$ so that $0 \in \arg\min\{h(F(\overline{x}) + F'(\overline{x})d) : \|d\| \leq 1\}$ as well. Therefore $d = 0$ is a local solution to the problem $\min\{h(F(\overline{x}) + F'(\overline{x})d)\}$. As the function $h(F(\overline{x}) + F'(\overline{x})d)$ is convex in d, we have that $d = 0$ is actually a global minimum so that

$$f'(\overline{x}; d) \geq 0 \qquad \forall\ d \in \mathbb{R}^n$$

by Lemma 3. ■

2.1.2 Trust–Region Methods

We again consider the problem of minimizing a function $f : \mathbb{R}^n \to \mathbb{R}$, but this time we require that a step-size of 1 must be taken at each iteration. In order to guarantee that the method is a descent method we take greater care in the selection of the search direction or step. In this context we label the search direction s_k to emphasize that it is the step to the new point and not just a direction to search along. The direction finding subproblem takes the form

$$\mathcal{P}(x, \delta) : \min_{\|s\| \leq \delta} \phi(x; s)$$

where $\phi(x; s) = f(x) + \nabla f(x)^T s + \frac{1}{2} s^T H s$ is a quadratic approximation to the function f at x. For obvious reasons, a good choice for H is $\nabla^2 f(x)$. The parameter δ is called the trust-region radius. At a given point x we require the step s to be an approximate solution to $\mathcal{P}(x, \delta)$. More specifically we make the following assumption.

Basic Assumption on the Trust-Region Step

For all $\epsilon > 0$ there exist constants κ_1, $\kappa_2 > 0$ such that

$$\nabla f(x_k)^T s_k + \frac{1}{2} s_k^T H_k s_k \leq -\kappa_1 \min\{\kappa_2, \delta_k\}$$

whenever $\epsilon < \|\nabla f(x_k)\|_o$.

This assumption guarentees that the step s_k must be at least as effective as a step optained by taking $H_k = 0$. In order to illustrate this comment and to show that this assumption can be satisfied for some choice of s_k, we give the following lemma.

LEMMA 2.1.1 *Let $H_k \in \mathbb{R}^{n\times n}$. If $\widehat{s}_k$ sovles $\min\{\nabla f(x_k)^T s_k : \|s_k\| \leq \delta_k\}$, then there exists $\widehat{t} \in (0, 1]$ such that*

$$t\nabla f(x_k)^T \widehat{s}_k + \frac{\widehat{t}^2}{2} \widehat{s}_k^T H_k \widehat{s}_k \leq -\frac{1}{2}\|\nabla f(x_k)\|_o \min\left\{\frac{\|\nabla f(x_k)\|_o}{\sigma^2 \|H_k\|_2}, \delta_k\right\}$$

where $\sigma > 0$ is such that $\|s\| \leq \sigma\|s\|_2$.

PROOF: $\widehat{s}_k$ solves $\min\{\nabla f(x_k)^T s : \|s\| \leq \delta_k\}$ if and only if $\nabla f(x_k)^T \widehat{s}_k = -\delta_k \|\nabla f(x_k)\|_o$. Next note that

$$\begin{array}{rcl} t\nabla f(x_k)^T \widehat{s}_k + \frac{t^2}{2} \widehat{s}_k^T H_k \widehat{s}_k & \leq & t\nabla f(x_k)^T \widehat{s}_k + \frac{t^2}{2}(\sigma^2 \delta_k^2)\|H_k\|_2 \\ & =: & \alpha t + \frac{\beta}{2} t^2 \end{array}$$

with $\alpha < 0$ and $\beta > 0$. One directly verifies that

$$\min\left\{\frac{-\alpha}{\beta}, 1\right\} = \arg\min_{[0,1]}\{\alpha t + \frac{\beta}{2} t^2\}.$$

Case 1: $\left(\frac{-\alpha}{\beta} \leq 1\right)$. Then set $\widehat{t} = \frac{-\alpha}{\beta}$ to get

$$\widehat{t}\nabla f(x_k)^T \widehat{s}_k + \frac{\widehat{t}^2}{2} \widehat{s}_k^T H_k \widehat{s}_k \leq \frac{1}{2} \nabla f(x_k)^T \widehat{s}_k \frac{|\nabla f(x_k)^T \widehat{s}_k|}{\sigma^2 \delta_k^2 \|H_k\|_2}.$$

Case 2: $\left(1 < \frac{-\alpha}{\beta}\right)$. Then set $\widehat{t} = 1$ to get

$$\begin{array}{rcl} \widehat{t}\nabla f(x_2)^T \widehat{s}_k + \frac{\widehat{t}^2}{2} \widehat{s}_k H_k \widehat{s}_k & \leq & \nabla f(x_k)^T \widehat{s}_k + \frac{1}{2}\sigma^2 \delta_k^2 \cdot \|H_k\|_2 \\ & \leq & \frac{1}{2} \nabla f(x_k)^T \widehat{s}_k. \end{array}$$

In either case we obtain

$$\widehat{t}\nabla f(x_k)^T \widehat{s}_k + \frac{\widehat{t}^2}{2} \widehat{s}_k^T H_k \widehat{s}_k \leq \frac{1}{2} \nabla f(x_k)^T \widehat{s}_k \min\left\{1, \frac{|\nabla f(x_k)^T s_k|}{\sigma^2 \delta_k^2 \|H_k\|_2}\right\}.$$

Now, by employing the relation

$$\nabla f(x_k)^T \widehat{s}_k = -\delta_k \|\nabla f(x_k)\|_o$$

we obtain

$$\begin{aligned} \widehat{t}\nabla f(x_k)^T \widehat{s}_k + \frac{\widehat{t}^2}{2}\widehat{s}_k^T H_k \widehat{s}_k &\leq -\tfrac{1}{2}\delta_k \|\nabla f(x_k)\|_o \min\left\{1, \tfrac{\delta_k \|\nabla f(x_k)\|_o}{\sigma^2 \delta_k^2 \|H_k\|_2}\right\} \\ &= \tfrac{-1}{2}\|\nabla f(x_k)\|_o \min\left\{\delta_k, \tfrac{\|\nabla f(x_k)\|_o}{\sigma^2 \|H_k\|_2}\right\} \end{aligned}$$

■

Before proceeding to the main result, we need the following technical lemma.

LEMMA 2.1.2 *Let $H \in \mathbb{R}^{n\times n}$ $0 < \overline{\beta}_1 \leq \overline{\beta}_2 < 1$ and α, κ_1, $\kappa_2 > 0$. Choose $\overline{\delta} > 0$ so that*

$$\kappa_1(1-\overline{\beta}_2)\min\{\kappa_2, \delta\} \geq \delta\omega_s(\delta) + \frac{1}{2}\sigma^2\delta^2\|H\|_2$$

for all $\delta \in [0, \overline{\delta}]$, where $\sigma > 0$ satisfies $\|z\| \leq \sigma\|z\|_2$ and

$$\omega_x(\delta) = \max\{\|\nabla f(x) - \nabla f(y)\|_o : y \in x + \delta\mathbb{B}\}.$$

Thus for every $\delta \in [0, \overline{\delta}]$ and $s \in \delta\mathbb{B}$ for which

$$\nabla f(x)^T s + \frac{1}{2}s^T H s \leq -\kappa_1 \min\{\kappa_2, \delta\}$$

one has

$$f(x+s) - f(x) \leq \overline{\beta}_1[\nabla f(x)^T s + \frac{1}{2}s^T H s].$$

PROOF:

$$\begin{aligned} f(x+s) - f(x) &\leq \nabla f(x)^T s + |f(x+s) - (f(x) + \nabla f(x)^T s)| \\ &\leq \nabla f(x)^T s + \|s\|\omega_x(\|s\|) \\ &\leq \nabla f(x)^T s + \tfrac{1}{2}s^T H s + \delta\omega_x(\delta) + \tfrac{1}{2}\sigma^2\delta^2\|H\|_2 \\ &\leq \nabla f(x)^T s + \tfrac{1}{2}s^T H s + \kappa_1(1-\overline{\beta}_1)\min\{\kappa_2, \delta\} \\ &\leq \nabla f(x)^T s + \tfrac{1}{2}s^T H s - (1-\overline{\beta}_1)[\nabla f(x)^T s + \tfrac{1}{2}s^T H s] \\ &= \overline{\beta}_1[\nabla f(x)^T s + \tfrac{1}{2}s^T H s] \end{aligned}$$

■

THEOREM 2.1.3 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and let $x_0 \in \mathbb{R}^n$ be such that ∇f is uniformly continuous on the set $\bar{co}\{x : f(x) \leq f(x_0)\}$.*
Consider the following algorithm:

Initialization: *Choose* $H_0 \in \mathbb{R}^{n\times n}$, $\delta_0 > 0$,

$$0 < \gamma_1 \le \gamma_2 < 1 \le \gamma_3, 0 < \beta_1 \le \beta_2 < \beta_3 \le 1.$$

Having x_k *obtain* x_{k+1} *as follows.*

Step 1: *Choose* $s_k \in D_k \subset \{s : \|s\| \le \delta_k,\ \nabla f(x_k)^T s + \frac{1}{2} s^T H_k s < 0\}$. *If* $D_k = \emptyset$ *then stop.*

Step 2: *Set* $r_k = \frac{f(x_k+s_k)-f(x_k)}{\nabla f(x_k)^T s_k + \frac{1}{2} s_k^T H_k s_k}$

If $r_k > \beta_3$ *choose* $\delta_{k+1} \in [\delta_k, \gamma_3 \delta_k]$.

If $\beta_2 \le r_k \le \beta_3$, *set* $\delta_{k+1} = \delta_k$.

If $r_k < \beta_2$, *choose* $\delta_{k+1} \in [\gamma_1 \delta_k, \gamma_2 \delta_k]$.

Step 3: *If* $r_k < \beta_1$, *set* $x_{k+1} = x_k$, $H_{k+1} = H_k$; *otherwise, set* $x_{k+1} = x_k + s_k$ *and choose* $H_{k+1} \in \mathbb{R}^{n\times n}$.

If the sequence $\{H_k\}$ *is bounded and the basic trust-region assumption is satisfied, then at least one of the following must occur:*

(1) $D_k = \emptyset$ *for some* k.

(2) $f(x_k) \searrow -\infty$

(3) $\|\nabla f(x_k)\|_o \to 0$

PROOF: We assume that none of (1)–(3) occur and derive a contradiction.

The sequence $\{x_k\}$ is infinite and

$$2\zeta < \|\nabla f(x_k)\|_o \qquad k \in J \tag{A}$$

for some $\zeta > 0$ and subsequence $J \subset \mathbb{N}$.

Thus, by the basic assumption there are constants κ_1 and $\kappa_2 > 0$ such that

$$\nabla f(x_k)^T s_k + \frac{1}{2} s_k^T H_k s_k \le -\kappa_1 \min\{\kappa_2, \delta_k\} \tag{B}$$

whenever $\|\nabla f(x_k)\|_o > \zeta$. In particular, (B) holds $\forall\ k \in J$. The technical lemma and the uniform continuity of ∇f yield the existence of a $\widehat{\delta}$ such that

$$r_k \ge \beta_1 \text{ and } x_{k+1} = x_k + s_k \tag{C}$$

whenever $\delta_k \le \widehat{\delta}$. We now show that $f(x_k) \downarrow -\infty$ to establish the contradiction:

Suppose there is a subsequence $\widehat{J} \subset J$ such that

$$\inf\{\delta_k : k \in \widehat{J}\} > \xi > 0. \tag{D}$$

Then for each $k \in \widehat{J}$ let $\sigma(k)$ be the first integer greater than or equal to k for which

$$x_{\sigma(k)+1} = x_{\sigma(k)} + s_{\sigma(k)}$$

and consider the subsequence

$$\widehat{J}_\sigma := \{\sigma(k) : k \in \widehat{J}\}.$$

Observe that for each $k \in \widehat{J}_\sigma$ we have

$$\delta_k \geq \min\{\gamma_1\widehat{\delta}, \gamma_1\xi\}. \quad (\text{ by (C)})$$

Consequently, we have from (B) that for each $k \in \widehat{J}_\sigma$

$$f(x_{k+1}) \leq f(x_k) - \kappa_1\beta_1 \min\{\kappa_2, \gamma_1\widehat{\delta}, \gamma_1\xi\}.$$

But then $f(x_k) \downarrow -\infty$ which is a contradiction. Therefore, we can assume that $\delta_k \leq \widehat{\delta}$ $\forall\ k \in J$ and $\lim_J \delta_k = 0$.

We obtain from the uniform continuity of ∇f the existence of an $\epsilon > 0$ such that

$$|\,\|\nabla f(x_i)\|_o - \|\nabla f(x_j)\|_o| < \zeta \tag{E}$$

whenever $\|x_i - x_j\| < \epsilon$ $i, j \in \mathbb{N}$. Given $k \in J$, let $v(k)$ be the first integer greater than k for which one of

$$\|x_{v(k)} - x_k\| \leq \epsilon \tag{F}$$

and

$$\delta_{v(k)} \leq \widehat{\delta} \tag{G}$$

is violated. Let us first show that $v(k)$ is well-defined and finite. Indeed, if $\|x_j - x_k\| \leq \epsilon$ $\forall\ j \geq k$ and $\delta_j \leq \widehat{\delta}$ $\forall\ j \geq k$, then, by (A) and (E)

$$\|\nabla f(x_j)\|_o > \zeta > 0 \quad \forall\ j \geq k.$$

Therefore, (B) and (C) hold for all $j \geq k$. Now take $\overline{\beta}_1 = \overline{\beta}_2 = \beta_2$ in the technical lemma to obtain the existence of a $0 < \widetilde{\delta} < \widehat{\delta}$ such that

$$r_k \geq \beta_2 \quad \text{whenever } \delta_k < \widetilde{\delta}.$$

Hence

$$f(x_{j+1}) \leq f(x_j) - \kappa_1 \min\{\kappa_2, \gamma_1\widetilde{\delta}\}$$

for all $j \geq k$, so $f(x_k) \downarrow -\infty$. This contradiction implies that $v(k)$ is well defined and finite for all $k \in J$.

Let $k \in J$ and consider $v(k)$. If (F) is violated, then by (A)–(E)–(B)

$$f(x_{l+1}) \leq f(x_l) - \kappa_1\beta_1 \min\{x_2, \delta_l\}$$

and

$$\delta_l \leq \delta_{l+1}$$

for $l = k, \ldots, v(k) - 1$. Hence

(H) $$f(x_{v(k)}) \leq f(x_k) - \kappa_1 \beta_1 \min\{\kappa_2, \epsilon\}$$

since

$$\sum_{l=k}^{v(k)} \delta_k \geq \|x_{v(k)} - x_k\| \geq \epsilon.$$

If (G) is violated, then

$$f(x_{v(k)}) \leq f(x_k) - \kappa_1 \beta_1 \min\{\kappa_2, \gamma_3^{-1} \widehat{\delta}\}.$$

In either case,

$$f(x_{v(k)}) \leq f(x_k) - \kappa_1 \beta_1 \min\{\kappa_2, \epsilon, \gamma_3^{-1} \widehat{\delta}\}$$

so that $f(x_k) \searrow -\infty$. This contradiction establishes the result. ■

A similar result holds for convex composite objective functions $f(x) = h(F(x))$. However, in this context we take

$$\phi(x; s) = h(F(x) + F'(x)s) + \frac{1}{3} s^T H s$$

and take

$$D_k \subset \{s : \|s\| \leq \delta_k, \ \Delta f(x_k : s) + \frac{1}{2} s^T H_k s < 0\}$$

where

$$\Delta f(x_k; s) = h(F(x) + F'(x_k)s) - h(F(x)).$$

One shows that either $D_k = \emptyset$ for some k, $f(x_k) \downarrow -\infty$, or $\Delta f(x_k, \delta_k) \to 0$ where

$$\Delta f(x_k, \delta_k) = \inf\{\Delta f(x_k : s) : \|s\| \leq \delta\}.$$

2.2 Local Theory

In this section we make assumptions that guarantee that the algorithm of the previous section converges to some point $\overline{x}$ and then study the *rate* or speed of convergence. The key assumption for the investigations of this section is that of strong convexity.

2.2.1 Strong Convexity

DEFINITION 2.2.1 *The function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be strongly convex if there is a $\delta > 0$ such that*

$$f(x + \lambda(y - x)) \leq f(x) + \lambda[f(y) - f(x)] - \frac{\delta}{2}\lambda(1 - \lambda)\|y - x\|^2 \tag{2.2.8}$$

for every $x, y \in \mathbb{R}^n$ and $\lambda \subset [0, 1]$. The parameter δ is called the modulus of strong convexity.

THEOREM 2.2.1 *(Characterizations of strongly convex functions) Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Then the following statements are equivalent.*

1. *f is strongly convex with modulus δ.*
2. *$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\delta}{2}\|y - x\|^2$ for all $x, \in \mathbb{R}^n$.*
3. *$(\nabla f(x) - \nabla f(y))^T(x - y) \geq \delta\|y - x\|^2$ for all $x, y \in \mathbb{R}^n$.*

If it is further assumed that f is twice continuously differentiable on $\mathbb{R}^n$, then then the above conditions are also equivalent to the following statement:

$$\inf\{u^T\nabla^2 f(x)u \,:\, \|u\| = 1,\ x \in \mathbb{R}^n\} \geq \delta \ .$$

That is, the spectrum of the Hessian of f is uniformly bounded below bu δ on $\mathbb{R}^n$.

PROOF: [(1) $\Longrightarrow$ (2)] From inequality (2.2.8) we have that

$$\frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \leq f(y) - f(x) - \frac{\delta}{2}(1 - \lambda)\|y - x\|^2$$

for $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$. Taking the limit as $\lambda \downarrow 0$ we obtain (2).
[(2) $\Longrightarrow$ (3)] Simply add the two inequalities

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\delta}{2}\|y - x\|^2$$

and

$$f(x) \geq f(y) + \nabla f(y)^T(y - x) + \frac{\delta}{2}\|y - x\|^2$$

to obtain the result.
[(3) $\Longrightarrow$ (2)] Let $x, y \in \mathbb{R}^n$ and define

$$x_\lambda := x + \lambda(y - x) \text{ for } \lambda \in \mathbb{R}.$$

By the mean value Theorem we have for some $\lambda \in (0, 1)$ that

$$\begin{aligned}
f(y) - f(x) &= \nabla f(x_\lambda)^T(y - x) \\
&= \nabla f(x)^T(y - x) + [\nabla f(x_\lambda) - \nabla f(x)]^T(y - x) \\
&= \nabla f(x)^T(y - x) + \lambda^{-1}[\nabla f(x_\lambda) - \nabla f(x)]^T(x_\lambda - x) \\
&\geq \nabla f(x)^T(y - x) + \lambda^{-1}\delta\|x_\lambda - x\|^2 \\
&\geq \nabla f(x)^T(y - x) + \frac{\delta}{2}|y - x\|^2
\end{aligned}$$

which proves the implication (3) $\Longrightarrow$ (2).
[(2) $\Longrightarrow$ (1)] Multiply the inequality

$$f(x) \geq f(x_\lambda) + \nabla f(x_\lambda)^T(x - x_\lambda) + \frac{\delta}{2}\|x - x_\lambda\|^2$$

by $(1 - \lambda)$ to obtain the inequality

$$(1-\lambda)f(x) \geq (1-\lambda)f(x_\lambda) - \lambda(1-\lambda)\nabla f(x_\lambda)^T(y-x) + \frac{\delta}{2}\lambda^2(1-\lambda\|y-x\|^2. \tag{2.2.9}$$

Then multiply the inequality

$$f(y) \geq f(x_\lambda) + \nabla f(x_\lambda)^T(y - x_\lambda) + \frac{\delta}{2}\|y - x_\lambda\|^2$$

by λ to obtain the inequality

$$\lambda f(y) \geq \lambda f(x_\lambda) + \lambda(1-\lambda)\nabla f(x_\lambda)^T(y-x) + \frac{\delta}{2}\lambda(1-\lambda)^2\|y-x\|^2. \tag{2.2.10}$$

Adding (2.2.9) and (2.2.10) yields the result.

The final statement of the theorem is an immediate consequence of 3. and the second–order Taylor series expansion of f. ■

Strongly convex functions possess two properties that are significant to our study.

THEOREM 2.2.2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be strongly convex.*

1. *The sets $\{x : f(x) \leq \alpha\}$ are compact convex sets for each $\alpha \in \mathbb{R}$.*
2. *The problem* $\min f$ *has a unique global optimal solution on $\mathbb{R}^n$.*

EXERCISE: Prove Theorem 2.2.2.

2.2.2 Linear Convergence

We employ two basic assumptions for the convergence analysis of this section:

Basic Assumptions:

(2.2.11)
1. f' is Lipschitz continuous with modulus $K > 0$ on an open convex set S containing the set $\{x : f(x) \leq f(x_0)\}$, and
2. f is strongly convex on S with modulus $\delta > 0$.

If f is twice differentiable on $\mathbb{R}^n$, then the constants K and δ have an interpretation in terms of the spectral structure of the hessian matrices $\nabla^2 f(x)$.

THEOREM 2.2.3 *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the basic assumptions (2.2.11). If f is twice differentiable on S, then*

1. $\delta\|z\|^2 \leq z^T\nabla^2 f(x)z \leq K\|z\|^2$ *for all* $x \in \mathbb{R}^n$ *and* $x \in S$,
2. *if* λ *is an eigenvalue of* $\nabla^2 f(x)$ *for some* $x \in S$, *then* $\delta \leq \lambda \leq K$, *and*
3. $\sup_{x\in S} \kappa(\nabla^2 f(x)) \leq \delta^{-1}K$ *where* $\kappa(\nabla^2 f(x))$ *is the condition number of the matrix* $\nabla^2 f(x)$.

PROOF: We only show (1). By (2) and (3) of Theorem 2.2.1 and the definition of Lipschitz continuity we know that

$$\delta\lambda^2\|z\|^2 \leq \lambda[\nabla f(x + \lambda z) - \nabla f(x)]^T z \leq K\lambda^2\|z\|^2$$

for $x \in S$, $z \in \mathbb{R}^n$, and $\lambda > 0$ sufficiently small. Dividing this expression by λ^2 and taking the limit as $\lambda \downarrow 0$ yields the result. ■

The condition number refered to in the above result is defined in the supplement on matrices. In this supplement the significance of the condition number to numerical computation is discussed. The condition number of a matrix is directly correlated to the stability of linear systems associated with said matrix. Under the assumption that the function f is twice continuously differntiable, the final statement in the above theorem implies that the Basic Assumptions in (**??**) are equivalent to the statement that the norm and condition number of the Hessian of f must both be uniformly bounded on $\mathbb{R}^n$. The condition number of the hessian matrix will also be seen to be closely correlated to the convergence rates of the various optimization algorithms that we will consider. Our first result along these lines now follows.

THEOREM 2.2.4 *(Linear Convergence Theorem) Let $f : \mathbb{R}^n \to \mathbb{R}$, $\{x_k\} \subset \mathbb{R}^n$, and $S := \{x : f(x) \leq f(x_0)\}$. Suppose that the basic assumptions (2.2.11) hold. Moreover, assume that the sequence $\{x_n\}$ is such that*

$$f(x_k) - f(x_{k+1}) \geq \alpha\|\nabla f(x_k)\|^2, \tag{2.2.12}$$

for some $\alpha > 0$ and each $k = 1, 2, \ldots$. Then the sequence $\{x_k\}$ converges to a unique solution $\overline{x}$ of the problem $\min\{f(x) : x \in \mathbb{R}^n\}$ at the linear root rate

$$\|x_i - \overline{x}\| \leq \left[\frac{2}{\delta}(f(x_0) - f(\overline{x})\right]^{1/2} \left[1 - \frac{2\alpha\delta^2}{K}\right]^{i/2}. \tag{2.2.13}$$

PROOF: By Theorem 2.2.2, S is compact and the problem $\min f$ has a unique global solution, $\overline{x}$, on S. Hence there is at least one cluster point $\hat{x}$. Since the sequence $\{f(x_k)\}$ is bounded

below by $f(\hat{x})$, we have $(f(x_k) - f(x_{k+1})) \to 0$ and so $\nabla f(\hat{x}) = 0$ by (2.2.12). Therefore $\overline{x} = \hat{x}$ and $x_k \to \overline{x}$. We now establish the rate. By Theorem 2.2.1,

$$\begin{aligned}\|\nabla f(x_k)\|\|x_k - \overline{x}\| &= \|\nabla f(x_k) - \nabla f(\overline{x})\|\|x_k - \overline{x}\| \\ &\geq (\nabla f(x_k) - \nabla f(\overline{x}))(x_k - \overline{x}) \\ &\geq \delta\|x_k - \bar{x}\|^2.\end{aligned}$$

Hence, by (2.2.12),

$$f(x_k) - f(x_{k+1}) \geq \alpha\delta^2\|x_k - \overline{x}\|^2.$$

The quadratic bound Lemma (B.2) thus yields

$$f(x_k) - f(x_{k+1}) \geq 2\alpha K^{-1}\delta^2(f(x_k) - f(\overline{x}))$$

which is equivalent to

$$(1 - 2\alpha K^{-1}\delta^2)(f(x_k) - f(\overline{x})) \geq (f(x_{k+1}) - f(\overline{x})).$$

By induction we obtain

$$(f(x_i) - f(\overline{x})) \leq (1 - 2\alpha K^{-1}\delta^2)^i(f(x_0) - f(\overline{x})).$$

But, by Theorem 2.2.1,

$$\frac{\delta}{2}\|x_i - \overline{x}\|^2 + \nabla f(\bar{x})^T(x_i - \overline{x}) \leq f(x_i) - f(\overline{x})$$

or

$$\|x_i - \overline{x}\| \leq (\frac{2}{\delta}(f(x_i) - f(\overline{x}))^{1/2}).$$

Using this last ineqality we obtain

$$\|x_i - \overline{x}\| \leq ((1 - 2\alpha K^{-1}\delta^2)^{1/2})^i(\frac{2}{\delta}(f(x_0) - f(\overline{x})))^{1/2}.$$

■

We now apply this result to the algorithm described in Theorem 2.1.1.

COROLLARY 2.2.4.1 *Let the assumptions of Theorem 2.2.4 hold except for hypothesis (2.2.12). Suppose that the sequence $\{x_k\}$ is generated by the algorithm of Theorem 2.1.1 where $d_i \in D_i$ implies that*

1. $-\nabla f(x_i)^T d_i \geq \nu\|\nabla f(x_i)\|\|d_i\|$ *for some* $\nu \in [0, 1]$, *and*
2. $\|d_i\| \geq \rho\|\nabla f(x_i)\|$ *for some* $\rho > 0$.

Then $x_i \to \overline{x}$ where $\overline{x}$ is the unique global solution of $\min f$ *and*

$$\|x_i - \overline{x}\| \leq \left[\frac{2}{\delta}(f(x_0) - f(\overline{x})\right]^{1/2} \left[1 - \frac{2\alpha\delta^2}{K}\right]^{i/2}, \tag{2.2.14}$$

where

$$\alpha := \min\{\frac{2\gamma c(1-c)\nu^2}{K}, c\gamma\rho\}. \tag{2.2.15}$$

PROOF: By the Armijo inequality

$$f(x_{i+1}) - f(x_i) \leq \frac{c\lambda_i \nabla f(x_i)^T d_i}{\|\nabla f(x_i)\|^2} \|\nabla f(x_i)\|^2.$$

Thus the result will follow from Theorem 2.2.4 if we can show that

$$\alpha \leq \frac{-c\lambda_i \nabla f(x_i)^T d_i}{\|\nabla f(x_i)\|^2}.$$

To this end, observe that by the quadratic bound lemma (Appendix B)

$$f(x_i + \lambda d_i) - f(x_i) - \lambda \nabla f(x_i)^T d_i \leq \frac{K}{2}\lambda^2 \|d_i\|^2$$

and so

$$\begin{aligned} f(x_i) - f(x_i + \lambda_i d_i) &\geq -\lambda \nabla f(x_i)^T d_i - \frac{K}{2}\lambda^2 \|d_i\|^2 \\ &= \lambda[(1-c)(-\nabla f(x_i)^T d_i) - \frac{K}{2}\lambda \|d_i\|^2 - c\nabla f(x_i)^T d_i] \end{aligned}$$

Hence the Arimijo inequality in Step 2 of the algorithm is satisfied if

$$(1-c)(-\nabla f(x_i)^T d_i) - \frac{K}{2}\lambda \|d_i\|^2 \geq 0$$

or equivalently if

$$\lambda \leq \frac{2}{K}(1-c)\frac{(-\nabla f(x_i)^T d_i)}{\|d_i\|^2}.$$

Therefore, by the maximality of λ_i, we know that

$$\lambda_i \geq \min\{1, \frac{2\gamma(1-c)}{K}\frac{(-\nabla f(x_i)^T d_i)}{\|d_i\|^2}\}.$$

Consequently, by (1) and (2),

$$\begin{aligned} \frac{-c\lambda_i \nabla f(x_i)^T d_i}{\|\nabla f(x_i)\|^2} &\geq \min\{\tfrac{-c\nabla f(x_i)^T d_i}{\|\nabla f(x_i)\|^2}, \tfrac{2\gamma c(1-c)}{K}\tfrac{(\nabla f(x_i)^T d_i)^2}{(\|\nabla f(x_i)\|\|d_i\|)^2}\} \\ &\geq \min\{c\nu\rho, \tfrac{2\gamma c(1-c)\nu^2}{K}\} \\ &= \alpha. \end{aligned}$$

■

We now relate the convergence rate (2.2.14) to the condition of the hessian matrices $\nabla^2 f(x)$. In order to simplify the discussion we assume that

$$\nu \sim \rho \sim 1, K > 1,$$

and

$$\frac{K}{\delta} \sim \kappa_{\max} := \sup\{\kappa(\nabla^2 f(x)) : x \in S\}.$$

In later sections it will be shown that these assumption are quite reasonable and provide an accurate description of most of the algorithms that we consider. With these assumption the parameter α in (2.2.15) is

$$\alpha = \frac{2\gamma c(1-c)}{K}.$$

Thus (2.2.14) becomes

$$\|x_i - \overline{x}\| \leq (\frac{2}{\delta}(f(x_0) - f(\overline{x})))^{1/2}(1 - 4\gamma c(1-c)\kappa_{\max}^{-2})^{i/2}.$$

Now in order to achieve the fastest rate of convergence we would like

$$(1 - 4\gamma c(1-c)\kappa_{\max}^{-2})$$

to be as close to zero as possible, and so $\gamma c(1-c)$ should be as large as possible with $0 < \gamma < 1$, $0 < c < 1$. The supremum of $\gamma c(1-c)$ over the allowable values is $1/4$. Thus the most optimistic rate is

$$\|x_i - \overline{x}\| \leq (\frac{2}{\delta}(f(x_0) - f(\overline{x})))(1 - [\kappa_{\max}]^{-2})^{i/2}.$$

Thus we see that the smaller δ is and the larger $\kappa_{\max}$ is the slower the convergence. Moreover, the convergence is clearly most sensitive to $\kappa_{\max}$.

Let us now consider an explicit example and examine the actual convergence behavior.

EXAMPLE: Let $f(x) = x^2 + e^x$. Then $f'(x) = 2x + e^x$ and $f''(x) = 2 + e^x$. Hence f is strongly convex on $\mathbb{R}$. If we take $x_0 = 1$, $c = .01$, $\gamma = \frac{1}{2}$, and $D_i = \{-\nabla f(x_i)/\|\nabla f(x_i)\|\}$ then one can show that the parameters appearing in Corollary 2.1.1. can be taken to be $K = 4$, $\delta = 2$, $\nu = 1$, $\rho = 1$, and consequently $\alpha = .0025$. Hence $(1 - \frac{2\alpha\delta^2}{K})^{1/2} \cong .997$ and so in the limit we get

$$\|x_i - \overline{x}\| \leq \sqrt{2}(.997)^i.$$

Therefore, in order to obtain $\|x_i - \overline{x}\| \leq .01$, this inequality implies that we should compute $i = 1649$ iterations. This convergence behavior on such a nice function is terrifyingly slow. Let us now look at the actual performance.

K	X	$f(x)$	$f'(x)$	s
0	1	.37182818	4.7182818	0
1	0	1	1	0
2	$-.5$	.8565307	-0.3934693	1
3	$-.25$	.8413008	0.2788008	2
4	$-.375$	.8279143	$-.0627107$	3
5	$-.34075$	.8273473	.0297367	5
6	$-.356375$	.8272131	$-.01254$	6
7	$-.3485625$	.8271976	.0085768	7
8	$-.3524688$	.8271848	$-.001987$	8
9	$-.3514922$	.8271841	.0006528	10
10	$-.3517364$	.827184	$-.0000072$	12

The behavior is clearly not as bad as that which is predicted by Corollary 2.2.4.1. It is rather slow and requires 55 function evaluations. In the next section we consider a much faster proceedure.

2.3 Newton's Method

2.3.1 Newton's Method for Equation Solving

In this section we study the following problem:

$\mathcal{E}$ **:** Given $g : \mathbb{R}^n \to \mathbb{R}^n$, find $x \in \mathbb{R}^n$ for which $g(x) = 0$.

In the context of optimization, this problem is significant for many reasons. In particular, it is directly related to the first–order necessary conditions in unconstrained optimization, i.e. $\nabla f(x) = 0$. In our discussion of $\mathcal{E}$ we always assume that g is C^1.

Suppose one is given an approximate solution $x_0 \in \mathbb{R}^n$ to $\mathcal{E}$ and wishes to improve upon it. If $\overline{x}$ is an actual solution to $\mathcal{E}$, then

$$0 = g(\overline{x}) = g(x_0) + g'(x_0)(\overline{x} - x_0) + o\|\overline{x} - x_0\|.$$

Thus, if x_0 is "close" to $\overline{x}$, it is reasonable to suppose that the solution to the linearized system

$$0 = g(x_0) + g'(x_0)(x - x_0) \tag{2.3.16}$$

is even closer. This proceedure is known as Newton's method for finding the roots of the equation $g(x) = 0$. It has one obvious pitfall. Equation (2.3.16) may not be consistent. That is, there may not exist an x solving (2.3.16). In general, the set of solutions to (2.3.16) is either

1. the empty set,

2. an infinite set, or

3. a single point.

For the sake of the present argument, we assume that (3) holds, i.e. $g'(x_0)^{-1}$ exists. Under this assumption (2.3.16) defines the iteration scheme,

$$x_{k+1} := x_k - [g'(x_i)]^{-1} g(x_k), \tag{2.3.17}$$

called the Newton iteration. The associated direction

$$d := -[g'(x_k)]^{-1} g(x_k). \tag{2.3.18}$$

is called the Newton direction. We analyze the convergence behavior of this scheme under the additional assumption that only an approximation to $g'(x_k)$ is available. We denote this approximation by J_k. The resulting iteration scheme is

$$x_{k+1} := x_k - J_k^{-1} g(x_k). \tag{2.3.19}$$

Methods of this type are called Newton-Like methods.

THEOREM 2.3.1 *Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be differentiable, $x_0 \in \mathbb{R}^n$, and $J_0 \in \mathbb{R}^{n\times n}$. Suppose that there exists $\bar{x}$, $x_0 \in \mathbb{R}^n$, and $\epsilon > 0$ with $\|x_0 - \bar{x}\| < \epsilon$ such that*

1. *$g(\overline{x}) = 0$,*
2. *$g'(x)^{-1}$ exists for $x \in B(\overline{x};\epsilon) := \{x \in \mathbb{R}^n : \|x - \overline{x}\| < \epsilon\}$ with*

$$\sup\{\|g'(x)^{-1}\| : x \in B(\overline{x};\epsilon)] \le M_1$$

3. *g' is Lipschitz continuous on $c\ell B(\overline{x};\epsilon)$ with Lipschitz constant L, and*
4. *$\theta_0 := \frac{LM_1}{2}\|x_0 - \overline{x}\| + M_0 K < 1$ where $K \ge \|(g'(x_0)^{-1} - J_0^{-1})y^0\|$, $y^0 := g(x^0)/\|g(x^0)\|$, and $M_0 = \max\{\|g'(x)\| : x \in B(\overline{x};\epsilon)\}$.*

Further suppose that iteration (2.3.19) is initiated at x_0 where the J_k's are chosen to satisfy one of the following conditions;

(i) $\|(g'(x_k)^{-1} - J_k^{-1})y_k\| \le K$,

(ii) $\|(g'(x_k)^{-1} - J_k^{-1})y_k\| \le \theta_1^k K$ for some $\theta_1 \in (0,1)$,

(iii) $\|(g'(x_k)^{-1} - J_k^{-1})y_k\| \le \min\{M_2\|x_k - x_{k-1}\|, K\}$, for some $M_2 > 0$, or

(iv) $\|(g'(x_k)^{-1} - J_k^{-1})y_k\| \le \min\{M_2\|g(x_k)\|, K\}$, for some $M_3 > 0$,

where for each $k = 1, 2, \ldots$, $y_k := g(x_k)/\|g(x_k)\|$.

These hypotheses on the accuracy of the approximations J_k yield the following conclusions about the rate of convergence of the iterates x_k.

(a) If (i) holds, then $x_k \to \overline{x}$ linearly.

(b) If (ii) holds, then $x_k \to \overline{x}$ superlinearly.

(c) If (iii) holds, then $x_k \to \overline{x}$ two step quadratically.

(d) If (iv) holds, then $x_i \to \overline{x}$ quadratically.

PROOF: We begin by establishing the basic inequalities

$$\|x_{k+1} - \overline{x}\| \leq \frac{LM_1}{2}\|x_k - \overline{x}\|^2 + \|(g'(x_k)^{-1} - J_k^{-1})g(x_k)\|, \tag{2.3.20}$$

and

$$\|x_{k+1} - \overline{x}\| \leq \theta_0\|x_k - \overline{x}\| \tag{2.3.21}$$

and the inclusion

$$x_{k+1} \in B(\bar{x}; \epsilon) \tag{2.3.22}$$

by induction on k. For $k = 0$ we have

$$\begin{aligned} x_1 - \overline{x} &= x_0 - \overline{x} - g'(x_0)^{-1}g(x_0) + [g'(x_0)^{-1} - J_0^{-1}]g(x_0) \\ &= g'(x_0)^{-1}[g(\overline{x}) - (g(x_0) + g'(x_0)(\overline{x} - x_0))] \\ &\quad + [g'(x_0)^{-1} - J_0^{-1}]g(x_0), \end{aligned}$$

since $g'(x_0)^{-1}$ exists by the hypotheses. Consequently, the hypothese (1)–(4) plus the quadratic bound lemma imply that

$$\begin{aligned} \|x_{k+1} - \overline{x}\| &\leq \|g'(x_0)^{-1}\|\|g(\overline{x}) - (g(x_0) + g'(x_0)(\overline{x} - x_0))\| \\ &\quad + \|(g'(x_0)^{-1} - J_0^{-1})g(x_0)\| \\ &\leq \frac{M_1 L}{2}\|x_0 - \overline{x}\|^2 + K\|g(x_0) - g(\overline{x})\| \\ &\leq \frac{M_1 L}{2}\|x_0 - \overline{x}\|^2 + M_0 K\|x_0 - \overline{x}\| \\ &\leq \theta_0\|x_0 - \overline{x}\| < \epsilon, \end{aligned}$$

whereby (2.3.20) – (2.3.21) are established for $k = 0$.

Next suppose that (2.3.20) – (2.3.21) hold for $k = 0, 1, \ldots, s-1$. We show that (2.3.20) – (2.3.21) hold at $k = s$. Since $x_s \in B(\overline{x}, \epsilon)$, hypotheses (2)–(4) hold at x_s, one can proceed exactly as in the case $k = 0$ to obtain (2.3.20). Now if any one of (i)–(iv) holds, then (i) holds. Thus, by (2.3.20), we find that

$$\begin{aligned} \|x_{s+1} - \overline{x}\| &\leq \tfrac{M_1 L}{2}\|x_s - \overline{x}\|^2 + \|(g'(x_s)^{-1} - J_s^{-1})g(x_s)\| \\ &\leq [\tfrac{M_1 L}{2}\theta_0^s\|x_0 - \overline{x}\| + M_0 K]\|x_s - \overline{x}\| \\ &\leq [\tfrac{M_1 L}{2}\|x_0 - \overline{x}\| + M_0 K]\|x_s - \overline{x}\| \\ &= \theta_0\|x_s - \overline{x}\|. \end{aligned}$$

Hence $\|x_{s+1}-\overline{x}\| \leq \theta_0\|x_s-\overline{x}\| \leq \theta_0\epsilon < \epsilon$ and so $x_{s+1} \in B(\overline{x},\epsilon)$. We now proceed to establish (a)–(d).

(a) This clearly holds since the induction above established that

$$\|x_{k+1}-\overline{x}\| \leq \theta_0\|x_k-\overline{x}\|.$$

(b) From (2.3.20), we have

$$\begin{aligned}\|x_{k+1}-\overline{x}\| &\leq \frac{LM_1}{2}\|x_k-\overline{x}\|^2 + \|(g'(x_k)^{-1}-J_k^{-1})g(x_k)\| \\ &\leq \frac{LM_1}{2}\|x_k-\overline{x}\|^2 + \theta_1^k K\|g(x_k)\| \\ &\leq [\frac{LM_1}{2}\theta_0^k\|x_0-\overline{x}\| + \theta_1^k M_0 K]\|x_k-\bar{x}\|\end{aligned}$$

Hence $x_k \to \overline{x}$ superlinearly.

(c) From (2.3.20) and the fact that $x_k \to \bar{x}$, we eventually have

$$\begin{aligned}\|x_{k+1}-\overline{x}\| &\leq \frac{LM_1}{2}\|x_k-\overline{x}\|^2 + \|(g'(x_k)^{-1}-J_k^{-1})g(x_k)\| \\ &\leq \frac{LM_1}{2}\|x_k-\overline{x}\|^2 + M_2\|x_k-x_{k-1}\|\|g(x_k)\| \\ &\leq [\frac{LM_1}{2}\|x_k-\overline{x}\| + M_0M_2[\|x_{k-1}-\bar{x}\| + \|x_k-\bar{x}\|]]\|x_k-\bar{x}\| \\ &\leq [\frac{LM_1}{2}\theta_0\|x_{k-1}-\overline{x}\| + M_0M_2(1+\theta_0)\|x_{k-1}-\overline{x}\|] \\ &\qquad \times\theta_0\|x_{k-1}-\overline{x}\| \\ &= [\frac{LM_1}{2}\theta_0 + M_0M_2(1+\theta_0)]\theta_0\|x_{k-1}-\overline{x}\|^2.\end{aligned}$$

Hence $x_k \to \overline{x}$ two step quadratically.

(d) Again by (2.3.20) and the fact that $x_k \to \bar{x}$, we eventually have

$$\begin{aligned}\|x_{k+1}-\overline{x}\| &\leq \frac{LM_1}{2}\|x_k-\overline{x}\|^2 + \|(g'(x_k)^{-1}-J_k^{-1})g(x_k)\| \\ &\leq \frac{LM_1}{2}\|x_k-\overline{x}\|^2 + M_2\|g(x_k)\|^2 \\ &\leq [\frac{LM_1}{2} + M_2M_0^2]\|x_k-\overline{x}\|^2 \ .\end{aligned}$$

■

Note that the conditions required for the approximations to the Jacobian matrices $g'(x_k)$ given in (i)–(ii) do not imply that $J_k \to g'(\bar{x})$. The stronger conditions

$(i)'$ $\|g'(x_k)^{-1} - J_k^{-1}\| \leq \|g'(x_0)^{-1} - J_0^{-1}\|$,

$(ii)'$ $\|g'(x_{k+1})^{-1} - J_{k+1}^{-1}\| \leq \theta_1 \|g'(x_k)^{-1} - J_k^{-1}\|$ for some $\theta_1 \in (0,1)$,

$(iii)'$ $\|g'(x_k)^{-1} - J_k^{-1}\| \leq \min\{M_2\|x_{k+1} - x_k\|, \|g'(x_0)^{-1} - J_0^{-1}\|\}$ for some $M_2 > 0$, or

$(iv)'$ $g'(x_k) = J_k$,

which imply the conditions (i) through (iv) of Theorem 2.3.1 respectively, all imply the convergence of the Jacobian approximates to $g'(\bar{x})$. Clearly the conditions $(i)'$–$(iv)'$ are not as desirable since they require a great deal more expense and care in the construction of the Jacobian approximates.

2.3.2 Newton's Method for Minimization

In this section we interpret the results of previous section in the context of minimization. That is, we apply them to the equation $\nabla f(x) = 0$ in the context of minimization. The translation of Theorem 2.3.1 for minimization follows.

THEOREM 2.3.2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable, $x_0 \in \mathbb{R}^n$, and $H_0 \in \mathbb{R}^{n\times n}$. Suppose that*

1. *there exists $\overline{x} \in \mathbb{R}^n$ and $\epsilon > \|x_0 - \bar{x}\|$ such that $f(\overline{x}) \leq f(x)$ whenever $\|x - \bar{x}\| \leq \epsilon$,*
2. *there is a $\delta > 0$ such that $\delta\|z\|_2^2 \leq z^T\nabla^2 f(x)z$ for all $x \in B(\overline{x}, \epsilon)$,*
3. *$\nabla^2 f$ is Lipschitz continuous on $clB(\overline{x}; \epsilon)$ with Lipschitz constant L, and*
4. *$\theta_0 := \frac{L}{2\delta}\|x_0 - \overline{x}\| + M_0 K < 1$ where $M_0 > 0$ satisfies $z^T\nabla^2 f(x)z \leq M_0\|z\|_2^2$ for all $x \in B(\overline{x}, \epsilon)$ and $K \geq \|(\nabla^2 f(x_0)^{-1} - H_0^{-1})y^0\|$ with $y^0 = \nabla f(x^0)/norm\nabla f(x^0)$.*

Further, suppose that the iteration

$$x_{k+1} := x_k - H_k^{-1}\nabla f(x_k) \tag{2.3.23}$$

is initiated at x_0 where the H_k's are chosen to satisfy one of the following conditions:

(i) $\|(\nabla^2 f(x_k)^{-1} - H_k^{-1})y_k\| \leq K$,

(ii) $\|(\nabla^2 f(x_k)^{-1} - H_k^{-1})y_k\| \leq \theta_1^k K$ *for some* $\theta_1 \in (0,1)$,

(iii) $\|(\nabla^2 f(x_k)^{-1} - H_k^{-1})y_k\| \leq \min\{M_2\|x_k - x_{k-1}\|, K\}$, *for some* $M_2 > 0$, *or*

(iv) $\|(\nabla^2 f(x_k)^{-1} - H_k^{-1})y_k\| \leq \min\{M_2\|\nabla f(x_k)\|, K\}$, *for some* $M_3 > 0$,

where for each $k = 1, 2, \ldots$ $y_k := \nabla f(x_k)/\|\nabla f(x_k)\|$.

These hypotheses on the accuracy of the approximations H_k yield the following conclusions about the rate of convergence of the iterates x_k.

(a) If (i) holds, then $x_k \to \overline{x}$ linearly.

(b) If (ii) holds, then $x_k \to \overline{x}$ superlinearly.

(c) If (iii) holds, then $x_\epsilon \to \overline{x}$ two step quadratically.

(d) If (iv) holds, then $x_k \to \overline{k}$ quadradically.

In order to more fully understand the convergence behavior described in the above result a careful study of the role of the controling parameters L, M_0, and M_1 needs to be made. Although we do not attempt this study, we do make a few observations. First observe that since L is a Lipschitz constant for $\nabla^2 f$ it represents a bound on the third–order behavior of f. Thus the assumptions for convergence make implicit demands on the third derivative. Next, the constant δ in the context of minimization represents a local uniform lower bound on the eigenvalues of $\nabla^2 f$. That is, f behaves locally as if it were a strongly convex function with modulus δ. Finally, M_0 can be interpreted as a local Lipschitz constant for ∇f and only plays a role when $\nabla^2 f$ is approximated inexactly by H_k's. Let us now consider the convergence behavior of this preceedure when applied to the example of Section 2.2.2.

EXAMPLE: Let $f(x) = x^2 + e^x$. Then $f'(x) = 2x + e^x$, $f''(x) = 2 + e^x$, $f'''(x) = e^x$. Given $x_0 = 1$ we may take $L = 2$, $M_0 = 4$, and $M_1 = \frac{1}{2}$. Hence the pure Newton strategy should converge to $\overline{x} \approx -0.3517337$ with

$$\|x_k - \overline{x}\| \leq 2(.676)^{2^k}.$$

The actual interates are given in the following table.

x	$f'(x)$
1	4.7182818
0	1
$-1/3$	.0498646
$-.3516893$	.00012
$-.3517337$	.00000000064

2.4 Linking Global and Local Methods

Recall that in the global theory we chose our search direction d_k from the set

$$\{d : f'(x_k; d) < 0\}.$$

If $\nabla^2 f(x_k)$ is positive definite, then the solution to the problem

$$\min_{d \in \mathbb{R}^n} f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2 f(x_k) d$$

is the Newton step $d_k^N = -\nabla^2 f(x_k)^{-1}\nabla f(x_k)$. In this case, we have

$$f'(x_k; d_k^N) = -\nabla f(x_k)^T \nabla^2 f(x_k)^T \nabla f(x_k) < 0$$

as long as $\nabla f(x_k) \neq 0$. That is,

$$d_k^N \in \{d; f'(x_k; d) < 0\}$$

and so is potentially a candidate direction for both the line-search and trust-region based methods. However, in line search methods it may be that a unit step length is not chosen and in trust-region methods the trust-region radius is too small, in which case the method may not attain the excellent local convergence rate of Newton's method. In this section we show that the full Newton step is locally acceptable under certain conditions. This in turn implies a local rate of convergence result for these global methods. We begin our analysis with the following local attraction result. The result says that, under reasonable assumptions on the construction of the iterates, any cluster point satisfying second–order sufficient conditions for optimality is actually the unique limit point of the sequence.

LEMMA 2.4.1 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Suppose that $\{x_k\} \subset \mathbb{R}^n$ is a sequence generated to satisfy*

(i) $\nabla f(x_k)^T(x_{k+1} - x_k) + \frac{1}{2}(x_{k+1} - x_k)^T \nabla^2 f(x_k)(x_{k+1} - x_k) \leq 0$,

(ii) $f(x_{k+1}) \leq f(x_k)$, *and*

(iii) $\nabla f(x_k) \to 0$.

If $\overline{x}$ is a cluster point of $\{x_k\}$ at which $\nabla^2 f(\overline{x})$ is positive definite, then it must be the case that $x_k \to \overline{x}$.

PROOF: Since $\nabla^2 f(\overline{x})$ is positive definite, there is an $\overline{\epsilon} > 0$ and $\delta > 0$ such that

$$s^T \nabla^2 f(x) s \geq \delta \|s\|^2 \; \forall \; x \in \overline{x} + \overline{\epsilon}\mathbb{B}.$$

Also note that $\nabla f(\overline{x}) = 0$ by (ii). For all $k > 0$ define

$$J_\epsilon := \{k : x_k \in \overline{x} + \epsilon\mathbb{B}\} \text{ and let } \overline{k} = \inf\{k : k \in J_{\overline{\epsilon}}\}.$$

Clearly, J_ϵ is infinite for all $\epsilon > 0$ since $\overline{x}$ is a cluster point of $\{x_k\}$. Define $s_k := x_{k+1} - x_k$. Then, by hypothesis (i),

$$\frac{\delta}{2}\|s_k\|^2 \leq \frac{1}{2} s_k^T \nabla^2 f(x_k) s_k \leq -\nabla f(x_k)^T s_k \leq \|\nabla f(x_k)\| \, \|s_k\|$$

for all $k \in J_{\overline{\epsilon}}$. Consequently

$$\frac{\delta}{2}\|s_k\| \leq \|\nabla f(x_k)\|$$

for all $k \in J_{\bar{\epsilon}}$.

Since $\nabla f(x_k) \to 0$, there is a $\widehat{k} \geq \overline{k}$, $\widehat{k} \in J_{\bar{\epsilon}/2}$ such that, for all $k \geq \widehat{k}$, $\|\nabla f(x_k)\| \leq \frac{\bar{\epsilon}\delta}{4}$. We now claim that for any $k \geq \widehat{k}$, $k \in J_{\bar{\epsilon}/2}$ we have

$$\{x \in \overline{x} + \bar{\epsilon}\mathbb{B} : f(x) \leq f(x_k)\} \subset \overline{x} + \frac{\bar{\epsilon}}{2}\mathbb{B}.$$

Indeed, if $x \in \{x \in \overline{x} + \bar{\epsilon}\mathbb{B} : f(x) \leq f(x_k)\}$, then

$$\begin{array}{rcl} f(x_k) \geq f(x) & \geq & f(\overline{x}) + \frac{\delta}{2}\|x - \overline{x}\|^2 \\ & \geq & f(x_k) + \nabla f(x_k)^T(\overline{x} - x_k) + \frac{\delta}{2}\|x - \overline{x}\| \end{array}$$

so that

$$\|\nabla f(x_k)\| \, \|\overline{x} - x_k\| \geq \frac{\delta}{2}\|x - \overline{x}\|^2 \, .$$

Since $x_k \in \bar{x} + \epsilon\mathbb{B}$ and $x + \epsilon\mathbb{B}$, we obtain

$$\frac{\bar{\epsilon}\delta}{4} \geq \frac{\delta}{2}\|x - \overline{x}\|$$

or

$$\frac{\epsilon}{2} \geq \|x - \overline{x}\|.$$

But if $k \geq \overline{k}$, $k \in J_{\bar{\epsilon}/2}$, then $f(x_{k+1}) \leq f(x_k)$ and

$$\frac{\delta}{2}\|s_k\| \leq \|\nabla f(x_k)\| \leq \frac{\bar{\epsilon}\delta}{4} \, ,$$

or equivalently $\|s_k\| \leq \frac{\bar{\epsilon}}{2}$, so that $\|x_{k+1} - \overline{x}\| \leq \bar{\epsilon}$. Hence, $\|x_{k+1} - \overline{x}\| \leq \frac{\bar{\epsilon}}{2}$. An easy induction yields $x_k \in \overline{x} + \frac{\bar{\epsilon}}{2}\mathbb{B}$ for all $k \geq \overline{k}$. Letting $\bar{\epsilon} \to 0$ establishes the result. ■

2.4.1 Line Search Methods

There are many approaches to safe–guarding the Newton search direction in the context of a line search strategy. We consider one such approach known as the *dog-leg strategy*. In this context the word dog-leg refers to a golfing term for a fairway of a given shape. The originator of this term is the well–known golfer M.J.D. Powell.

Dog-Leg Search Direction: Let $\delta > 0$ be given.

Step 1 For each pair $(x, H) \in \mathbb{R}^n \times \mathbb{R}^{n \times n}_s$ define

$$d^{SD} = \begin{cases} 0 & \text{if } \nabla f(x) = 0 \\ \bar{t}u & \text{else} \end{cases}$$

where $u = -\nabla f(x)/\|\nabla f(x)\|_2$ and $\bar{t}$ solves

$$\min\{t\nabla f(x)^T u + \frac{t^2}{2}u^T\nabla^2 f(x)u : t \in [0, \delta]\}.$$

Step 2 Solve $\nabla^2 f(x)d = -\nabla f(x)$ for d^N. If in the solution process one finds that $\nabla^2 f(x)$ is not positive definite, then terminate the computation and set $d^{DL} := d^{SD}$.

Step 3 If the computation of d^N in Step 2 was successful, then set $d^{DL} := d^N$ if $\|d^N\| \leq \delta$; otherwise set $d^{DL} := \lambda d^{SD} + (1-\lambda)d^N$ where λ is chosen so that $\|d^{DL}\|_2 = \delta$.

The resulting direction d^{DL} is called a dog-leg search direction.

The dog-leg search direction can be used in the line-search algorithm of Theorem 3.1.1 to yield the following result.

THEOREM 2.4.1 *Let the hypotheses of Theorem 3.1.1 be satisfied and let it be further assumed that $\nabla^2 f(x)$ is bounded on the set $\overline{co}\{x : f(x) \leq f(x_0)\}$ and that the search direction d_n be taken to be the dog-leg direction. Then one of the following must occur:*

(i) $\nabla f(x_k) = 0$ *for some* k,

(ii) $f(x_k) \searrow \infty$,

(iii) $\|\nabla f(x_k)\| \to 0$.

PROOF: Using the notation given in the definition of the dog-leg search direction, set

$$d_k = \lambda_k d_k^{SD} + (1-\lambda_k)d_k^N,$$

$$d_k^{SD} = t_k u_k, \text{ and}$$

$$u_k = -\nabla f(x_k)/\|\nabla f(x_k)\|$$

where

$$t_k = \begin{cases} 1, & \text{if } u_k^T \nabla f(x_k) u_k \leq 0 \\ \min\left\{\frac{\|\nabla f(x_k)\|}{|u_k^T \nabla^2 f(x_k) u_k|}, 1\right\}, & \text{otherwise,} \end{cases}$$

$\lambda_k \in [0,1]$ with $\lambda_k = 1$ if d_k^N is not computed.

We will assume that (i) and (ii) do not occur and show that (iii) must occur. Since $\{d_k\}$ is bounded, we have from Theorem 3.1.1 that $f'(x_k; d_k) \to 0$, where

$$f'(x_k; d_k) = -\lambda_k t_k \|\nabla f(x_k)\| - (1-\lambda_k)\zeta_k$$

with

$$\zeta_k = \begin{cases} 0, & \text{if } \lambda_k = 1 \\ |\nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k)|, & \text{otherwise.} \end{cases}$$

Since $\{\nabla^2 f(x_k)\}$ is bounded, for any subsequence $J \subset \mathbb{N}$ one has that $t_k \xrightarrow{J} 0$ only if $\|\nabla f(x_k)\| \xrightarrow{J} 0$. Therefore, for every subsequence $J \subset \mathbb{N}$ for which $\inf_J \lambda_k > 0$ one has $\|\nabla f(x_k)\| \xrightarrow{J} 0$.

Let $J \subset \mathbb{N}$ be any subsequence for which $\lambda_k \xrightarrow{J} 0$ and $\lambda_k \neq 1$ for all $k \in J$. Then $\zeta_k = |\nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k)|$ and, by construction, $\nabla^2 f(x_k)$ is positive definite for all $k \in J$. Moreover, $\zeta_k \to 0$ since $f'(x_k; d_k) \to 0$. Since $\nabla^2 f(x_k)$ is bounded there exists $M > 0$ such that $v^T \nabla^2 f(x_k) v \leq M\|v\|^2$, or equivalently, $v^T \nabla^2 f(x_k)^{-1} v^T \geq M^{-1}\|v\|^2$, for all $k \in J$. But then,

$$0 < M^{-1}\|\nabla f(x_k)\|^2 \leq \nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k) \to 0.$$

Therefore, $\|\nabla f(x_k)\| \xrightarrow{J} 0$.

Now, since $\|\nabla f(x_k)\| \xrightarrow{J} 0$ for every subsequence for which $\lambda_k \to 0$ and for every subsequence on which λ_k is bounded away from zero, we have that $\|\nabla f(x_k)\| \to 0$. ■

We can now apply Lemma 2.4.1 to obtain a local rate of convergence for the dog-leg method.

THEOREM 2.4.2 *Let the hypotheses of Theorem 2.4.1 be satisfied and let it further be assumed that the backtracking parameter c satisfies $0 < c < \frac{1}{2}$. If $\overline{x}$ is a cluster point of the sequence $\{x_k\}$ at which $\nabla^2 f(\overline{x})$ is positive definite, then $x_k \to \overline{x}$ at a quadratic rate.*

PROOF: By construction, the sequence $\{f(x_k)\}$ is bounded below by $f(\overline{x})$. Therefore, Theorem 2.4.1 and Lemma 2.4.1 combine to imply that $x_k \to \overline{x}$. Since $\nabla^2 f(\overline{x})$ is positive definite and $\nabla f(x_k) \to 0$, we have that d_k^N exists with $\|d_k^N\| < \delta$ for all k large. We now show that the step length $\lambda_k = 1$ for all k large. This will establish the result since then the method is locally equivalent to Newton's method and so $x_k \to \overline{x}$ quadratically.

Since $\nabla^2 f$ is continuous, it is uniformly continuous in a neighborhood of $\overline{x}$. Let $\omega(t)$ be the modulus of continuity for $\nabla^2 f$ near $\overline{x}$ and let $\mu > 0$ and $M > 0$ be such that

$$\mu\|v\|^2 \leq v^T \nabla^2 f(x) v \leq M\|v\|$$

for all x near $\overline{x}$. Let $\bar{t} > 0$ be such that

$$\omega(t) \leq \left(\frac{1}{2} - c\right)\frac{\mu^2}{M}$$

for all $t \in [0, \bar{t}]$ and let $\bar{k}$ be such that $\|d_k^N\| \leq \bar{t}$ for all $k \geq \bar{k}$. Then, for $k \geq \bar{k}$ and some $z_k \in [x_k, x_k + d_k^N]$,

$$\begin{aligned}
f(x_k + d_k^N) - f(x_k) &= \nabla f(x_k)^T d_k^N + \tfrac{1}{2} d_k^{N^T} \nabla^2 f(z_k) d_k^N \\
&\leq \nabla f(x_k)^T d_k^N + \tfrac{1}{2} d_k^{N^T} \nabla^2 f(x_k) d_k^N + \omega(\|d_k^N\|)\|d_k^N\|^2 \\
&\leq c\nabla f(x_k)^T d_k^N + (1-c)\nabla f(x_k)^T d_k^N - \tfrac{1}{2}\nabla f(x_k)^T d_k^N \\
&\qquad + \left(\tfrac{1}{2} - c\right) \tfrac{\mu^2}{M} \|d_k^N\|^2 \\
&= c\nabla f(x_k)^T d_k^N + \left(c - \tfrac{1}{2}\right) \nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k) \\
&\qquad + \left(\tfrac{1}{2} - c\right) \tfrac{\mu^2}{M} \|\nabla^2 f(x_k)^{-1} \nabla f(x_k)\|^2 \\
&\leq c\nabla f(x_k)^T d_k^N + \left(c - \tfrac{1}{2}\right) M^{-1} \|\nabla f(x_k)\|^2 \\
&\qquad + \left(\tfrac{1}{2} - c\right) M^{-1} \|\nabla f(x_k)\|^2 \\
&= c\nabla f(x_k)^T d_k^N.
\end{aligned}$$

Therefore, the step length $\lambda_k = 1$ for all k large. ■

2.4.2 Trust–Region Methods

A similar result is easily obtained for trust–region methods by making more explicit the requirement that the step s_k be at least as effective as a step based on linear information alone. This is done by requiring that the step satisfy the inequality established in Lemma 2.1.1. This inequality is a refinement of the Basic Assumption on the Trust–Region Step.

THEOREM 2.4.3 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and let $x_0 \in \mathbb{R}^n$ be such that ∇f is uniformly continuous on the set $\overline{co}\{x : f(x) \leq f(x_0)\}$. Suppose $\{x_k\}$ is a sequence generated by the trust-region algorithm of Theorem 2.1.3 with $H_k = \nabla^2 f(x_k)$ and s_k satisfying*

$$\nabla f(x_k)^T s_k + \frac{1}{2} s_k^T \nabla^2 f(x_k) s_k \leq -\frac{1}{2} \|\nabla f(x_k)\|_o \min\left\{ \frac{\|\nabla f(x_k)\|_o}{\sigma^2 \|\nabla^2 f(x_k)\|^2}\, ,\, \delta_k \right\} \tag{2.4.1}$$

for all $k = 1, 2, \ldots$, where $\sigma > 0$ is chosen to satisfy $\|s_k\|_2 \leq \sigma \|s_k\|$ (note that such an s_k is guaranteed to exist by Lemma 3.1.1). If $\bar{x}$ is a cluster point of $\{x_k\}$ at which $\nabla^2 f(\bar{x})$ is positive definite, then $x_k \to \bar{x}$ and $r_k \to 1$.

PROOF: Theorem 2.1.2 and Lemma 2.4.1 combine to imply that $x_k \to \overline{x}$. Let $\epsilon > 0$ and $\delta > 0$ be chosen so that

$$s^T \nabla^2 f(x) s \geq \delta \|s\|^2 \quad \forall\, x \in \overline{x} + \epsilon \mathbb{B}.$$

Let $\overline{k}$ be such that $x_k \in \overline{x} + \epsilon\mathbb{B}$ for all $k \geq \overline{k}$. Then for $k \geq \overline{k}$

$$\frac{\delta}{2}\|s_k\|_2^2 \leq \frac{1}{2} s_k^T \nabla^2 f(x_k) s_k \leq -\nabla f(x_k)^T s_2 \leq \|\nabla f(x_k)\|_2 \|s_k\|_2$$

so that

$$\frac{\delta}{2}\|s_k\|_2 \leq \|\nabla f(x_k)\|_2. \tag{2.4.2}$$

Let $\sigma_2 > 0$ be such that $\sigma_2\|s\| \leq \|s\|_2$. Note that $\sigma_2\mathbb{B}_2 \subset \mathbb{B}$ since if $u \in \sigma_2\mathbb{B}_2$, then $\|u\| \leq \sigma_2^{-1}\|u\|_2 \leq \sigma_2^{-1}\sigma_2 = 1$. Therefore,

$$\|v\|_o = \sup\{\langle v, u\rangle : u \in \mathbb{B}\} \geq \sup\{\langle v, u\rangle : u \in \sigma_2\mathbb{B}_2\} = \sigma_2\|v\|_2.$$

Thus, (2.4.2) implies that

$$\frac{\delta\sigma_2}{2}\|s\| \leq \sigma_2^{-1}\|\nabla f(x_k)\|_o,$$

or

$$\frac{\delta}{2}\sigma_2^2\|s\| \leq \|\nabla f(x_k)\|_o.$$

Combining this with assumption (2.4.1) yields the existence of a constant $\kappa > 0$ such that

$$-[\nabla f(x_k)^T s_k + \frac{1}{2} s_k^T \nabla^2 f(x_k) s_k] \geq \kappa \|s_k\|^2$$

for all $k \geq \overline{k}$. Therefore,

$$\begin{aligned} |r_k - 1| &= \left| \frac{f(x_2+s_k) - [f(x_k) + \nabla f(x_k)^T s_2 + \frac{1}{2} s_k^T \nabla^2 f(x_k) s_k]}{-[\nabla f(x_k)^T s_k + \frac{1}{2} s_k^T \nabla^2 f(x_k) s_k]} \right| \\ &\leq \frac{|f(x_k+s_k) - (f(x_k) + \nabla f(x_k)^T s_k + \frac{1}{2} s_2^T \nabla^2 f(x_k) s_k)|}{\kappa \|s_k\|^2} \\ &\to 0 \qquad \text{as } k \to \infty. \end{aligned}$$

■

COROLLARY 2.4.3.1 *Let the hypotheses of the Theorem hold. If it is further assume that s_k is chosen as the solution to the subproblem $\mathcal{P}(x_k, \delta_k)$, then $x_\epsilon \to \overline{x}$ quadratically.*

PROOF: Since $r_k \to 1$ and $s_k \to 0$, we have $\|s_k\| < \delta_k$ for all k large. In this case, the method is locally equivalent to Newton's Method so the result follows from Theorem 2.3.2. ■

Chapter 3

Conjugate Direction Methods

3.1 General Discussion

In this section we are again concerned with the problem of unconstrained optimization:

$$\begin{array}{ll} \mathcal{P}: & \text{minimize } f(x) \\ & \text{subject to } x \in \mathbb{R}^n \end{array}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is C^2. However, the emphasis will be on local quadratic approximations to f. In particular, we study the problem $\mathcal{P}$ when f has the form

$$f(x) := \frac{1}{2}x^T Q x - b^T x, \tag{3.1.1}$$

where Q is a symmetric positive definite matrix. In this regard the notion of Q-conjugacy plays a key role.

DEFINITION 3.1.1 *Let $Q \in \mathbb{R}^{n\times n}$ be symmetric and positive definite. We say that the vectors $x, y \in \mathbb{R}^n \backslash \{0\}$ are Q-conjugate (or Q-orthogonal) if $x^T Q y = 0$.*

PROPOSITION 3.1.2 *If $Q \in \mathbb{R}^{n\times n}$ is positive definite and the set of nonzero vectors d_0, $d_1, \ldots, d_k$ are (pairwise) Q-conjugate, then these vectors are linearly independent.*

PROOF: If $0 = \sum_{i=0}^{k} \alpha_i d_i$, then for $i_0 \in \{0, 1, \ldots, k\}$

$$0 = d_{i_0}^T Q[\sum_{i=0}^{k} \alpha_i d_i] = \alpha_{i_0} d_{i_0}^T Q d_i,$$

Hence $\alpha_i = 0$ for each $i = 0, \ldots, k$. ∎

Observe that the unique solution to $\mathcal{P}$ when f is given by (3.1.1) is

$$x^* = Q^{-1}b.$$

If $\{d_0, d_1, \ldots, d_{n-1}\}$ is a Q-conjugate basis for $\mathbb{R}^n$, there are scalars $\alpha_0, \ldots, \alpha_{n-1}$ such that

$$x^* = \alpha_0 d_0 + \ldots + \alpha_{n-1} d_{n-1}. \tag{3.1.2}$$

Multiplying this expression through by Qd_i for each $i = 0, \ldots, n-1$ we find that

$$\alpha_i = \frac{d_i^T Q x^*}{d_i^T Q d_i} = \frac{d_i^T b}{d_i^T Q d_i}$$

for each $i = 0, \ldots, n-1$. Therefore

$$x^* = \sum_{i=0}^{n-1} \frac{d_i^T b}{d_i^T Q d_i} = \left[\sum_{i=0}^{n-1} \frac{d_i d_i^T}{d_i^T Q d_i}\right] b$$

so that

$$Q^{-1} = \sum_{i=0}^{n-1} \frac{d_i d_i^T}{d_i^T Q d_i}.$$

It is important to note that the coefficients α_i in the representation (3.1.2) can be computed without knowledge of x^*. This observation is the basis of the following result.

THEOREM 3.1.1 *Let $\{d_i\}_{i=0}^{n-1}$ be a set of nonzero Q-conjugate vectors. For any $x_0 \in \mathbb{R}^n$ the sequence $\{x_k\}$ generated according to*

$$x_{k+1} := x_k + \alpha_k d_k, \quad k \geq 0$$

with

$$\alpha_k := \arg\min\{f(x_k + \alpha d_k) : \alpha \in \mathbb{R}\}$$

converges to the unique solution, x^* of $\mathcal{P}$ with f given by (3.1.1) after n steps, that is $x_n = x^*$.

PROOF: Let us first compute the value of the α_k's. Set

$$\begin{aligned} \varphi_k(\alpha) &= f(x_i + \alpha d_k) \\ &= \tfrac{\alpha^2}{2} d_k^T Q d_k + \alpha g_k^T d_k + f(x_k), \end{aligned}$$

where $g_k = \nabla f(x_k) = Qx_k - b$. Then $\varphi_k'(\alpha) = \alpha d_k^T Q d_k + g_k^T d_k$, hence

$$\alpha_k = -\frac{g_k^T d_k}{d_k^T Q d_k}.$$

Now suppose $x^* - x_0$ has representation

$$x^* - x_0 = \hat{\alpha}_0 d_0 + \hat{\alpha}_1 d_1 + \ldots + \hat{\alpha}_{n-1} d_{n-1}. \tag{3.1.3}$$

Since $x_n = x_0 + \alpha_0 d_0 + \ldots + \alpha_{n-1} d_{n-1}$, the result is established if we can show that $\hat{\alpha}_k = \alpha_k$ for each $k = 0, 1, \ldots, n-1$. Multiplying (1.3) through by Qd_k yields

$$\hat{\alpha}_k = \frac{d_k^T Q(x^* - x_0)}{d_k^T Q d_k}. \tag{3.1.4}$$

But $Qx^* = b$ and

$$\begin{aligned} d_k^T Q x_0 &= d_k^T Q(x_0 + \alpha_0 d_0 + \ldots + \alpha_{k-1} d_{k-1}) \\ &= d_k^T Q d_k. \end{aligned}$$

Therefore

$$\begin{aligned} \hat{\alpha}_k &= -\frac{d_k^T Q(x_0 - x^*)}{d_k^T Q d_k} \\ &= -\frac{d_k^T (Q x_k - b)}{d_k^T Q d_k} \\ &= -\frac{d_k^T g_k}{d_k^T Q d_k} = \alpha_k. \end{aligned}$$

■

The following result provides further geometric insight into how the algorithm is proceeding.

THEOREM 3.1.2 (Expanding Subspace Theorem) *Let $\{d_i\}_{i=0}^{n-1}$ be a sequence of nonzero Q-conjugate vectors in $\mathbb{R}^n$. Then for any $x_0 \in \mathbb{R}^n$ the sequence $\{x_k\}$ generated according to*

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k d_k \\ \alpha_k &= -\frac{g_k^T d_k}{d_k^T Q d_k} \end{aligned}$$

has the property that $f(x) = \frac{1}{2} x^T Q x - b^T x$ attains its minimum value on the affine set $x_0 +$ Span $\{d_0, \ldots, d_{k-1}\}$ at the point x_k.

PROOF: We establish the result by directly computing the solution to

$$\begin{aligned} &\min f(x) \\ &\text{subject to } x - x_0 \in \text{ Span } \{d_0, d_1, \ldots, d_{k-1}\}. \end{aligned} \tag{3.1.5}$$

By setting $D_k = [d_0, d_1, \ldots, d_{k-1}]$ and $z = x - x_0$ we can rewrite (3.1.5) as

$$\begin{aligned} &\min_{(z,y)} f(z + x_0) \\ &\text{subject to } z = D_k y, \end{aligned}$$

which can be written as

$$\min_y f(D_k y + x_0). \tag{3.1.6}$$

Writing

$$\begin{aligned} \varphi(y) &= f(D_k y + x_0) \\ &= \tfrac{1}{2} y^T D_k^T Q D_k Y + g_0^T D_k y + f(x_0), \end{aligned}$$

where $g_0 = \nabla f(x_0)$, we see that the solution to (3.1.6) is obtained by setting

$$0 = \nabla \varphi(y) = D_k^T Q D_k y + D_k^T g_0.$$

Now

$$\begin{aligned} D_k^T Q D_k &= [d_i^T Q d_j]_{i,j=0}^{k-1} \\ &= \text{diag}[d_i^T Q d_i]_{i=0}^{k-1}, \end{aligned}$$

and

$$D_k^T g_0 = [d_0^T g_0, d_1^T g_0, \ldots, d_{k-1}^T g,]^T.$$

Hence

$$y_i = \frac{-d_i^T g_0}{d_i^T Q d_i} \quad \text{for } i = 0, \ldots, k-1.$$

Therefore, the solution (3.1.5) is

$$x_k^* = x_0 + \sum_{i=0}^{k-1} -\frac{d_i^T g_0}{d_i^T Q d_i} d_i.$$

Consequently, the result will be established if we can show that

$$-\frac{d_i^T g_0}{d_i^T Q d_i} = -\frac{d_i^T g_i}{d_i^T Q d_i} \tag{3.1.7}$$

for each $i = 0, 1, \ldots, k-1$. But this follows immediately from (3.1.4) since

$$\begin{aligned} \alpha_i = -\frac{d_i^T g_i}{d_i^T Q d_i} &= \frac{d_i^T (Q x_i - b)}{d_i^T Q d_i} \\ &= -\frac{d_i^T (Q(x_0 + \alpha_0 d_0 + \alpha_1 d_1 + \ldots + \alpha_{i-1} d_{i-1} - b)}{d_i^T Q d_i} \\ &= -\frac{d_i^T g_0}{d_i^T Q d_i} \end{aligned}$$

where the global minimum value of f is attained at x^* and so satisfies $Qx^* = b$. ■

COROLLARY 3.1.2.1 *In the method of Conjugate directions the gradients g_k, $k = 0, 1, \ldots, n$ satisfy*

$$g_k^T d_i = 0 \quad \textit{for } i < k.$$

PROOF: This follows from a general property of minimization on affine sets. Consider the problem

$$\begin{aligned} &\min \varphi(x) \\ &\text{subject to } x \in x_0 + S, \end{aligned}$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}$ is C^1 and S is the subspace $S :=$ span $\{v_1, \ldots, v_k\}$. If V is the matrix whose columns are given by $v_1, \ldots, v_k$, then this problem is equivalent to the problem

$$\begin{aligned} &\min \varphi(x_0 + Vz) \\ &\text{subject to } z \in \mathbb{R}^k \ . \end{aligned}$$

Setting $\hat{\phi}(z) = \varphi(x_0 + Vz)$, we get that if $\bar{z}$ solves the latter problem, then $V^T \nabla \varphi(x_0 + V\bar{z}) = \nabla \bar{\phi}(\bar{z}) = 0$. Setting $\bar{x} = x_0 + V\bar{z}$, we conclude that $\bar{x}$ solves the original problem if and only if $\bar{z}$ solves the latter problem in which case $V^T \nabla \varphi(\bar{x}) = 0$, or equivalently, $v_i^T \nabla \varphi(\bar{x}) = 0$ for $i = 1, 2, \ldots \ , k$. ■

3.2 The Conjugate Gradient Algorithm

The conjugate direction algorithm of the previous section appears to be seriously flawed in that one must have on hand a set of conjugate directions $\{d_0, \ldots, d_{n-1}\}$ in order to apply it. However, one builds a set of Q–conjugate directions as the algorithm proceeds. The example of such a procedure studied in this section is called the conjugate gradient algorithm.

The C-G Algorithm:

Initialization: $x_0 \in \mathbb{R}^n$, $d_0 = -g_0 = -\nabla f(x_0) = b - Qx_0$.

For $k = 0, 1, 2, \ldots$

$$\begin{aligned} \alpha_k &:= -g_k^T d_k / d_k^T Q d_k \\ x_{k+1} &:= x_k + \alpha_k d_k \\ g_{k+1} &:= Q x_{k+1} - b \\ \beta_k &:= g_{k+1}^T Q d_k / d_k^T Q d_k \\ d_{k+1} &:= -g_{k+1} + \beta_k d_k \\ k &:= k + 1. \end{aligned}$$

THEOREM 3.2.1 **Conjugate Gradient Theorem**
The C-G algorithm is a conjugate direction method. If it does not terminate at x_k, then

1. *Span* $[g_0, g_1, \ldots, g_k] =$ *span* $[g_0, Qg_0, \ldots, Q^k g_0]$
2. *Span* $[d_0, d_1, \ldots, d_k] =$ *span* $[g_0, Qg_0, \ldots, Q^k g_0]$
3. $d_k^T Q d_i = 0$ *for* $i \leq k - 1$
4. $\alpha_k = g_k^T g_k / d_k^T Q d_k$
5. $\beta_k = g_{k+1}^T g_{k+1} / g_k^T g_k$.

PROOF: We first prove (1)-(3) by induction. The results are clearly true for $k = 0$. Now suppose they are true for k, we show they are true for $k + 1$. First observe that

$$g_{k+1} = g_k + \alpha_k Q d_k$$

so that $g_{k+1} \in$ Span$[g_0, \ldots, Q^{k+1} g_0]$ by the induction hypothesis on (1) and (2). Also $g_{k+1} \notin$ Span $[d_0, \ldots, d_k]$ otherwise $g_{k+1} = 0$ (by Theorem 3.1.2.1 since the method is a conjugate direction method up to step k by the induction hypothesis. Hence $g_{k+1} \notin$ Span $[g_0, \ldots, Q^k g_0]$ and so Span $[g_0, g_1, \ldots, g_{k+1}] =$ Span $[g_0, \ldots, Q^{k+1} g_0]$, which proves (1).

To prove (2) write

$$d_{k+1} = -g_{k+1} + \beta_k d_k$$

so that (2) follows from (1) and the induction hypothesis on (2).

To see (3) observe that

$$d_{k+1}^T Q d_i = -g_{k+1} Q d_i + \beta_k d_k^T Q d_i.$$

For $i = k$ the right hand side is zero by the definition of β_k. For $i < k$ both terms vanish. The term $g_{k+1}^T Q d_i = 0$ by Theorem 3.1.2 since $Qd_i \in \text{Span}[d_0, \ldots, d_k]$ by (1) and (2). The term $d_i^T Q d_i$ vanishes by the induction hypothesis on (3).

To prove (4) write

$$-g_k^T d_k = g_k^T g_k - \beta_{k-1} g_k^T d_{k-1}$$

where $g_k^T d_{k-1} = 0$ by Theorem 3.1.2.

To prove (5) note that $g_{k+1}^T g_k = 0$ by Theorem 3.1.2 because $g_k \in \text{Span}[d_0, \ldots, d_k]$. Hence

$$g_{k+1}^T Q d_k = \frac{1}{\alpha_k} g_{k+1}^T [g_{k+1} - g_k] = \frac{1}{\alpha_k} g_{k+1}^T g_{k+1}.$$

Therefore,

$$\beta_k = \frac{1}{\alpha_k} \frac{g_{k+1}^T g_{k+1}}{d_k^T Q d_k} = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}.$$

■

Remarks:

1. The C–G method decribed above is a descent method since the values $f(x_0),\ f(x_1),\ \ldots, f(x_n)$ form a decreasing sequence. Moreover, note that

   $$\nabla f(x_k)^T d_k = -g_k^T g_k \quad \text{and} \quad \alpha_k > 0\ .$$

   Thus, the C–G method behaves very much like the methods discussed at the beginning of Chapter 2.

2. It should be observed that due to the occurrence of round-off error the C-G algorithm is best implemented as an iterative method. That is, at the end of n steps, f may not attain its global minimum at x_n and the intervening directions d_k may not be Q-conjugate. Consequently, at the end of the n^{th} step one should check the value $\|\nabla f(x_n)\|$. If it is sufficiently small, then accept x_n as the point at which f attains its global minimum value; otherwise, reset $x_0 := x_n$ and run the algorithm again. Due to the observations in remark above, this approach is guarenteed to continue to reduce the function value if possible since the overall method is a descent method. In this sense the C–G algorithm is self correcting.

3.3 Extensions to Non-Quadratic Problems

If $f : \mathbb{R}^n \to \mathbb{R}$ is not quadratic, then the Hessian matrix $\nabla^2 f(x_k)$ changes with k. Hence the C-G method needs modification in this case. An obvious approach is to replace Q by $\nabla^2 f(x_k)$ everywhere it occurs in the C-G algorithm. However, this approach is fundamentally flawed in its explicit use of $\nabla^2 f$. By using parts (4) and (5) of the conjugate gradient Theorem 3.2.1 and by trying to mimic the descent features of the C–G method, one can obtain a workable approximation of the C–G algorithm in the non–quadratic case.

The Non-Quadratic C-G Algorithm

Initialization: $x_0 \in \mathbb{R}^n$, $g_0 = \nabla f(x_0)$, $d_0 = -g_0$, $0 < c < \beta < 1$.
Having x_k otain x_{k+1} as follows:
Check restart criteria. If a restart condition is satisfied, then reset $x_0 = x_n$, $g_0 = \nabla f(x_0)$, $d_0 = -g_0$; otherwise, set

$$\begin{aligned}
\alpha_k &\in \left\{ \lambda \;\middle|\; \begin{array}{c} \lambda > 0, \nabla f(x_k + \lambda d_k)^T d \geq \beta \nabla f(x_k)^T d_k, \text{ and} \\ f(x_k + \lambda d_k) - f(x_k) \leq c\lambda \nabla f(x_k)^T d_k \end{array} \right\} \\
x_{k+1} &:= x_k + \alpha_k d_k \\
g_{k+1} &:= \nabla f(x_{k+1}) \\
\beta_k &:= \begin{cases} \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k} & \text{Fletcher-Reeves} \\ \max\left\{0, \frac{g_{k+1}^T (g_{k+1} - g_k)}{g_k^T g_k}\right\} & \text{Polak-Ribiere} \end{cases} \\
d_{k+1} &:= -g_{k+1} + \beta_k d_k \\
k &:= k + 1.
\end{aligned}$$

Remarks

1. The Polak-Ribiere update for β_k has a demonstrated experimental superiority. One way to see why this might be true is to observe that

$$g_{k+1}^T (g_{k+1} - g_k) \approx \alpha_k g_{k+1}^T \nabla^2 f(x_k) d_k$$

   thereby yielding a better second–order approximation. Indeed, the formula for β_k in in the quadratic case is precisely

$$\frac{\alpha_k g_{k+1}^T \nabla^2 f(x_k) d_k}{g_k^T g_k} .$$

2. Observe that the Hessian is never explicitly refered to in the above algorithm.

3. At any given iteration the procedure requires the storage of only 2 vectors if Fletcher-Reeves is used and 3 vectors if Polak-Ribiere is used. This is of great significance if n is very large, say $n = 50,000$. Thus we see that one of the advantages of the C-G method is that it can be practically applied to very large scale problems.

4. Aside from the cost of gradient and function evaluations the greatest cost lies in the line search employed for the computation of α_k.

We now consider appropriate restart criteria. Clearly, we should restart when $k = n$ since this is what we do in the quadratic case. But there are other issues to take into consideration. First, since $\nabla^2 f(x_k)$ changes with each iteration, there is no reason to think that we are preserving any sort of conjugacy relation from one iteration to the next. In order to get some kind of control on this behavior, we define a *measure* of conjugacy and if this

measure is violated, then we restart. Second, we need to make sure that the search directions d_k are descent directions. Moreover, (a) the angle between these directions and the negative gradient should be bounded away from zero in order to force the gradient to zero, and (b) the directions should have a magnitude that is comparable to that of the gradient in order to prevent ill–conditioning. The precise restart conditions are given below.

Restart Conditions

1. $k = n$
2. $|g_{k+1}^T g_k| \geq 0.2 g_k^T g_k$
3. $-2g_k^T g_k \geq g_k^T d_k \geq -0.2 g_k^T g_k$

Conditions (2) and (3) above are known as the Powell restart conditions.

Chapter 4

Matrix Secant Methods

4.1 Equation Solving

In this section we again study the problem of finding $\overline{x} \in \mathbb{R}^n$ such that $g(\overline{x}) = 0$ where $g : \mathbb{R}^n \to \mathbb{R}^n$ is C^1. Specifically, we will consider Newton-Like methods of a special type. Recall that in a Newton-Like method the iteration scheme takes the form

$$x_{k+1} := x_k - M_k^{-1} g(x_k), \tag{4.1.1}$$

where M_k is meant to approximate $g'(x_k)$. In the one dimensional case, a choice of particular note is the secant approximation

$$M_k = \frac{g(x_{k-1}) - g(x_k)}{x_{k-1} - x_k}. \tag{4.1.2}$$

With this approximation one has

$$g'(x_k)^{-1} - M_k^{-1} = \frac{g(x_{k-1}) - [g(x_k) + g'(x_k)(x_{k-1} - x_k)]}{g'(x_k)[g(x_{k-1}) - g(x_k)]}.$$

Also, near a point x^* at which g' is non–singular there exists an $\alpha > 0$ such that

$$\alpha \|x - y\| \leq \|g(x) - g(y)\|.$$

Consequently, by the Quadratic Bound Lemma,

$$\|g'(x_k)^{-1} - M_k^{-1}\| \leq \frac{\frac{L}{2}\|x_{k-1} - x_k\|^2}{\alpha \|g'(x_k)\| \|x_{k-1} - x_k\|} \leq K \|x_{k-1} - x_k\|$$

for some constant $K > 0$ whenever x_k and x_{k-1} are sufficiently close to x^*. Therefore, by Theorem 2.3.1, there secant method is locally two step quadratically convergent to a non–singular solution of the equation $g(x) = 0$. An additional advantage of this approach is that no extra function evaluations are required to obtain the approximation M_k.

Unfortunately, the secant approximation (4.1.2 is meaningless in the $n > 1$ dimensional case since division by vectors is undefined. However, this can be rectified by simply writing

$$M_k(x_{k-1} - x_k) = g(x_{k-1}) - g(x_k). \tag{4.1.3}$$

Equation (4.1.3) is called the Quasi-Newton equation (QNE) at x_k and it determines M_k along an n dimensional manifold in $\mathbb{R}^{n\times n}$. Thus equation (4.1.3) is not enough to uniquely determine M_k since (4.1.3) is n linear equations in n^2 unknowns. Consequently, we may place further conditions on the update M_k if we wish to do so. In order to see what further properties one would like the update to possess, let us consider an overall iteration scheme based on (4.1.1). At every iteration we have (x_k, M_k) and compute x_{k+1} by (4.1.1). Then M_{k+1} is constructed to satisfy (4.1.3). If M_k is close to $g'(x_k)$ and x_{k+1} is close to x_k, then M_{k+1} should be chosen not only to satisfy (4.1.3) but also to be as "close" to M_k as possible. In what sense should we mean "close" here? In order to facilitate the computations it is reasonable to mean "algebraically" close in the sense that M_{k+1} is only a rank 1 modification of M_k, i.e. there are vectors $u, v \in \mathbb{R}^n$ such that

$$M_{k+1} = M_k + uv^T. \tag{4.1.4}$$

Multiplying (1.3) by

$$s_k := x_{k+1} - x_k$$

and using (4.1.3) we find that

$$y_k = M_{k+1}s_k = M_k s_k + uv^T s_k$$

where $y_k := g(x_{k+1}) - g(x_k)$. Hence, if $v^T s_k \neq 0$, we obtain

$$u = \frac{y_k - M_k s_k}{v^T s_k}$$

and

$$M_{k+1} = M_k + \frac{(y_k - M_k s_k)v^T}{v^T s_k}. \tag{4.1.5}$$

Equation (4.1.5) determines a whole class of rank one updates that satisfy the QNE where one is allowed to choose $v \in \mathbb{R}^n$ as long as $v^T s_k \neq 0$. If $s_k \neq 0$, then an obvious choice for v is s_k yielding the update

$$M_{k+1} = M_k = \frac{(y_k - M_k s_k)s_k^T}{s_k^T s_k}. \tag{4.1.6}$$

This is known as Broyden's update. Given the algebraically "close" updates in (4.1.5), it is reasonable to ask whether there related updates that are analytically close.

THEOREM 4.1.1 *Let $A \in \mathbb{R}^{n\times n}$, s, $y \in \mathbb{R}^n$, $s \neq 0$. Then for any matrix norms $\|\cdot\|$ and $|||\cdot|||$ such that*

$$\|AB\| \leq \|A\| \, |||B|||$$

and

$$|||\frac{vv^T}{v^Tv}||| \leq 1,$$

the solution to

$$\min\{\|B - A\| : Bs = y\} \tag{4.1.7}$$

is

$$A_+ = A + \frac{(y - As)s^T}{s^Ts}. \tag{4.1.8}$$

In particular, (4.1.8) solves (4.1.7) when $\|\cdot\|$ is the ℓ_2 matrix norm, and (4.1.8) solves (4.1.7) uniquely when $\|\cdot\|$ is the Frobenius norm.

PROOF: Let $B \in \{B \in \mathbb{R}^{n\times n} : Bs = y\}$, then

$$\begin{aligned} \|A_+ - A\| &= \|\frac{(y - As)s^T}{s^Ts}\| = \|(B - A)\frac{ss^T}{s^Ts}\| \\ &\leq \|B - A\|\,|||\frac{ss^T}{s^Ts}||| \leq \|B - A\|. \end{aligned}$$

Note that if $|||\cdot||| = \|\cdot\|_2$, then

$$\begin{aligned} \|\frac{vv^T}{v^Tv}\|_2 &= \sup\{\|\frac{vv^T}{v^Tv}x\|_2 : \|x\|_2 = 1\} \\ &= \sup\{\sqrt{\frac{(v^Tx)^2}{\|v\|^2}} : \|x\|_2 = 1\} \\ &= 1, \end{aligned}$$

so that the conclusion of the result is not vacuous. For uniqueness observe that the Frobenius norm is strictly convex and $\|A \cdot B\|_F \leq \|A\|_F\|B\|_2$. ■

Therefore, the Broyden update (4.1.6) is both algebraically and analytically close to M_k. These properties indicate that it should perform well in practice and indeed it does.

Algorithm: Broyden's Method

Initialization: $x_0 \in \mathbb{R}^n$, $M_0 \in \mathbb{R}^{n\times n}$

Having (x_k, M_k) compute (x_{k+1}, M_{x+1}) as follows:

Solve $M_ks_k = -g(x_k)$ for s_k and set

$$\begin{aligned} x_{k+1} : &= x_k + s_k \\ y_k : &= g(x_k) - g(x_{k+1}) \\ M_{k+1} : &= M_k + \frac{(y_k - M_ks_k)s_k^T}{s_k^Ts_k}. \end{aligned}$$

Due to its derivation we call methods based up (4.1.5) matrix secant methods. In the literature they are also called Quasi-Newton methods.

Observe that the derivation of (4.1.5) only relied upon the relations

$$M_{k+1}s_k = y_k$$

and

$$M_{k+1} = M_k + uv^T.$$

Thus by switching the roles of s_k and y_k it is possible to obtain an inverse updating scheme. That is if instead of (4.1.1) we write

$$x_{k+1} := x_k - W_k g(x_k)$$

where $W_k \approx [g'(x_k)]^{-1}$, then a matrix secant method for updating W_k would be

$$W_{k+1} := W_k + \frac{(s_k - W_k y_k)y_k^T}{y_k^T y_k}, \tag{4.1.9}$$

since we want the QNE

$$x_{k+1} - x_k = s_k - W_{k+1}y_k = W_{k+1}(g(x_{k+1}) - g(x_k))$$

to hold. It would be interesting to determine if $W_k = M_k^{-1}$, For this we require the following well-know lemma.

LEMMA 4.1.1 (Sherman-Morrison-Woodbury) *Suppose $A \in \mathbb{R}^{n\times n}$, $U \in \mathbb{R}^{n\times k}$, $V \in \mathbb{R}^{n\times k}$ are such that both A^{-1} and $(I + V^T A^{-1}U)^{-1}$ exist, then*

$$(A + UV^T)^{-1} = A^{-1} - A^{-1}U(I + V^T A^{-1}U)^{-1}V^T A^{-1}$$

EXERCISE: Prove Lemma 4.1.1.

The above lemma verifies that if M^{-1} exists and $s^T M^{-1} y \neq 0$, then

$$[M + \frac{(y - Ms)s^T}{s^T s}]^{-1} = M^{-1} + \frac{(s - M^{-1}y)s^T M^{-1}}{s^T M^{-1} y}. \tag{4.1.10}$$

Consequently, it is not true that the W_k's obtained from (4.1.9) and the M_k's from (4.1.6) satisfy

$$W_k = M_k^{-1}.$$

However, (4.1.10) does indicate a variation on both (4.1.6) and (4.1.9). Specifically, in (4.1.5) one could choose $v = M_k y_k$, in which case one does obtain the inverse of (4.1.9). Conversely, one could replace (4.1.9) with

$$W_{k+1} := W_k + \frac{(s_k - W_k y_k)s_k^T W_k}{s_k^T W_k y_k} \tag{4.1.11}$$

yielding the inverse of (4.1.6).

On the surface computing the inverse updates appears to be more attractive since then we need not solve the equation

$$B_k s_k = y_k$$

for s_k at every iteration. However, this approach can suffer from fatal numerical instabilities if $g'(x^*)$ is singular or nearly singular.

Although we do not pause to establish the convergence rates here, we do give the following result due to Dennis and Moré (1974).

THEOREM 4.1.2 *Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable in an open convex set $D \subset \mathbb{R}^n$. Assume that there exists $x^* \in \mathbb{R}^n$ and $r, \beta > 0$ such that $x^* + r\mathbb{B} \subset D$, $g(x^*) = 0$, $g'(x^*)^{-1}$ exists with $\|g'(x^*)^{-1}\| \leq \beta$, and g' is Lipschitz continuous on $x^* + r\mathbb{B}$ with Lipschitz constant $\gamma > 0$. Then there exist positive constants ϵ and δ such that if $\|x_0 - x^*\|_2 \leq \epsilon$ and $\|B_0 - g'(x_0)\| \leq \delta$, then the sequence $\{x_k\}$ generated by the iteration*

$$\left[\begin{array}{rcl} x_{k+1} & := & x_k + s_k \text{ where } s_k \text{ solves } 0 = g(x_k) + B_k s \\ B_{k+1} & := & B_k + \frac{(y_k - B_k s_k) s_k^T}{s_k^T s_k} \text{ where } y_k = g(x_{k+1}) - g(x_k) \end{array} \right.$$

is well-defined with $x_k \to x^*$ superlinearly.

4.2 Minimization

In this section the underlying problem is one of minimization:

$$\mathcal{P} : \underset{x \in \mathbb{R}^n}{\text{minimize}} \ f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is C^2. The basic idea is to modify and/or extend the matrix secant methods of the previous section to the setting of minimization where one wishes to solve the equation $\nabla f(x) = 0$. In this context the QNE becomes

$$M_{k+1} s_k = y_k$$

where $s_k := x_{k+1} - x_k$ and

$$y_k := \nabla f(x_{k+1}) - \nabla f(x_k).$$

A straightforward application of Broyden's method would yield the update

$$M_{k+1} = M_k + \frac{(y_k - M_k s_k) s_k^T}{s_k^T s_k}.$$

However, this is unsatisfactory for two reasons:

1. Since M_k is intended to approximate $\nabla^2 f(x_k)$ it is desirable that M_k be symmetric.

2. Since we are concerned with minimization, then at least locally one can assume the second order sufficiency condition holds. Consequently, we would like the M_k's to be positive definite.

To address problem 1 above, one could return to equation (4.1.5) for an alternate update that preserves symmetry. Such an update is uniquely obtained by setting

$$v = (y_k - M_k s_k).$$

This is called the symmetric rank 1 update or SR1. Although this update can on occasion exhibit problems with numerical stability, it has recently received a great deal of renewed interest. The stability problems occur whenever

$$v^T s_k = \nabla f(x_k)^T M_k^{-1}(\nabla f(x_k) - \nabla f(x_{k+1}))$$

tends to zero faster than $\|\nabla f(x_k)\|^2$.

The following alternate strategy has been proposed by Powell. One begins by symmetrizing the Broyden update

$$\overline{M}_1 = M + \frac{(y - Ms)s^T}{s^T s}.$$

This is done by replacing $\overline{M}_1$ with its symmetric part

$$\overline{M}_2 = \frac{1}{2}(\overline{M}_1 + \overline{M}_1^T).$$

But then the QNE fails. To remedy this set

$$\overline{M}_3 = \overline{M}_2 + \frac{(y - \overline{M}_2 s)s^T}{s^T s}.$$

But again symmetry fails so set

$$\overline{M}_4 = \frac{1}{2}(\overline{M}_3 + \overline{M}_3^T).$$

Proceeding in this way we get a sequence $\{\overline{M}_k\}$ with

$$\begin{cases} \overline{M}_{2k+1} = \overline{M}_{2k} + \frac{(y - \overline{M}_{2k} s)s^T}{s^T s} \\ \overline{M}_{2(K+1)} = \frac{1}{2}(\overline{M}_{2k+1} + \overline{M}_{2k+1}^T) \end{cases}$$

for $k = 0, 1, \ldots$. Since the set $S_1 = \{M \in \mathbb{R}^{n\times n} : Ms = y\}$ is an affine subset of $\mathbb{R}^{n\times n}$ and the set $S_2 = \{M \in \mathbb{R}^{n\times n} : M \text{ is symmetric}\}$ is a subspace of $\mathbb{R}^{n\times n}$, and the equations (4.2) represent a sequence of alternating orthogonal projections in $\mathbb{R}^{n\times n}$ onto S_1, and S_2 respectively, the sequence $\{\overline{M}_k\}$ must converge to a fixed point solving the proximation problem

$$\begin{array}{ll} \min & \|\overline{M} - M\|_F \\ \text{subject to} & \overline{M} \in S_1, (M - \overline{M}) \in S_2. \end{array}$$

It is possible to show that the solution is

$$\overline{M} = M + \frac{(y - Ms)s^T + s(y - Ms)^T}{s^T s} - \frac{(y - Ms)^T s s s^T}{(s^T s)^2}. \tag{4.2.12}$$

The update (4.2.12) is called the Powell-symmetric-Broyden (PSB) update.

Update (4.2.12) was derived to preserve both symmetry and stability, however there is no guarantee that if M is positive definite, then $\overline{M}$ is also. We now address the question of when this is possible. That is, suppose $M \in \mathbb{R}^{n\times n}$ symmetric and positive definite, we wish to find $\overline{M}$ satisfying the QNE such that $\overline{M}$ is also symmetric and positive definite. Let $M = LL^T$ be the Cholesky factorization of M. If $\overline{M}$ is to be symmetric and positive definite then there is a matrix $J \in \mathbb{R}^{n\times n}$ such that $\overline{M} = JJ^T$. The QNE implies that if

$$J^T s = v \tag{4.2.13}$$

then

$$Jv = y. \tag{4.2.14}$$

Let us try to apply the Broyden update technique to (4.2.14), J, and L. That is, suppose that

$$J = L + \frac{(y - Lv)v^T}{v^T v}. \tag{4.2.15}$$

Then by (4.2.13)

$$v = J^T s = L^T s + \frac{v(y - Lv)^T s}{v^T v}. \tag{4.2.16}$$

This expression implies that v must have the form

$$v = \alpha L^T s$$

for some $\alpha \in \mathbb{R}$. Substituting this back into (4.2.16) we get

$$\alpha L^T s = L^T s + \frac{\alpha L^T s(y - \alpha LL^T s)^T s}{\alpha^2 s^T LL^T s}.$$

Hence

$$\alpha^2 = \left[\frac{s^T y}{s^T Ms}\right].$$

Consequently, such a matrix J satisfying (4.2.16) exists only if $s^T y > 0$ in which case

$$J = L + \frac{(y - \alpha Ms)s^T L}{\alpha s^T Ms},$$

with

$$\alpha = \left[\frac{s^T y}{s^T Ms}\right]^{1/2},$$

yielding

$$\overline{M} = M + \frac{yy^T}{y^T s} - \frac{Mss^T M}{s^T Ms}. \tag{4.2.17}$$

Moreover, the Cholesky factorization for $\overline{M}$ can be obtained directly from the matrices J. Specifically, if the QR factorization of J^T is $J^T = QR$, we can set $\overline{L} = R$ yielding

$$\overline{M} = JJ^T = R^T Q^T QR = \overline{L}\,\overline{L}^T.$$

The question of course remains as to how one can assure the positivity of the product $s^T y$. Recall that in the iterative context

$$s = s_k = -\lambda_k M_k^{-1} \nabla f(x_k)$$

and

$$y = y_k = \nabla f(x_{k+1}) - \nabla f(x_k).$$

Hence

$$\begin{aligned} y^T s = y_k^T s_k &= \nabla f(x_{k+1})^T s_k - \nabla f(x_k)^T s_k \\ &= \lambda_k \nabla f(x_k + \lambda_k d_k)^T d_k - \lambda_k \nabla f(x_k)^T k_k \ , \end{aligned}$$

where $d_k := -M_k^{-1} \nabla f(x_k)$. Now since M_k is positive definite the direction d_k is a descent direction for f at x_k and so i $\lambda_k > 0$. Therefore, to assure that $y^T s > 0$ we need only show that we can choose $\lambda_k > 0$ so that

$$\nabla f(x_k + \lambda_k d_k)^T d_k \geq \beta \nabla f(x_k)^T d_k \tag{4.2.18}$$

for some $\beta \in (0, 1)$.

Note that for any descent direction d of f at x_k we can choose

$$\overline{\lambda} := \arg\min\{f(x_k + \lambda d) : \lambda > 0\}$$

in which case $\overline{\lambda} = +\infty$ or

$$\nabla f(x_k + \overline{\lambda} d)^T d = 0.$$

Therefore, λ_k can always be chosen to make $y_k^T s_k > 0$, and so the updating strategy (4.2.17) can be used to guarantee both symmetry and positive definiteness if a *suitable* line search is employed. We return to the question of what a suitable line search is later in this section.

The update (4.2.17) is called the BFGS (Broyden-Fletcher-Goldfarb-Shanno) update and is currently considered the best available matrix secant type update for minimization. Observe in (4.2.17) that if both $\overline{M}$ and M are positive definite, then they are both invertible. The Sherman-Morrison-Woodbury formula shows that the inverse is given by

$$\begin{aligned} \overline{M}^{-1} = M^{-1} &+ \frac{(s - M^{-1}y)s^T + s(s - M^{-1}y)^T}{y^T s} \\ &- \frac{(s - M^{-1}y)^T y s s^T}{(y^T s)^2}. \end{aligned}$$

Thus the corresponding inverse updating scheme for the BFGS update is

$$\begin{aligned}\overline{W} = W &+ \frac{(s - Wy)s^T + s(s - Wy)^T}{y^T s} \\ &- \frac{(s - Wy)^T y s s^T}{(y^T s)^2}.\end{aligned}$$

One can now use this representation of the inverse to write down an alternate update of M, it is

$$\begin{aligned}\hat{M} = M &+ \frac{(y - Ms)y^T + y(y - Ms)^T}{y^T s} \\ &- \frac{(y - Ms)^T y y^T}{(y^T s)^2}.\end{aligned}$$

This is known as the DFP formula named after Davidon-Fletcher-Powell.

One can show that the DFP, BFGS, and SR1 updates are all members of a one parameter family of updates known as the *Broyden family*. In order to see this set

$$a := s^T M s,\ b := y^T s,\ c := y^T W y,$$

and, assuming $a,\ b$, and c are nonzero, define two vectors

$$m := \frac{y}{b} - \frac{Ms}{a}$$

and

$$w := \frac{s}{b} - \frac{Wy}{c}$$

satisfying $m^T s = 0 = w^T y$. Then define two parameterized families of matrices by

$$\bar{M}(\mu) := M - \frac{M s s^T M}{a} + \frac{y y^T}{b} + \mu a m m^T, \tag{4.2.19}$$

and

$$\bar{W}(\nu) := W - \frac{W y y^T W}{c} + \frac{s s^T}{b} + \nu c w w^T. \tag{4.2.20}$$

The following table illustrates the relationship between the updates and various values of μ and ν:

Update	μ	ν
DFP	1	0
BFGS	0	1
SR1	$b/(b-a)$	$b/(b-c)$

We now turn to the study of an appropriate line search procedure. In particular, this procedure should enforce inequality (4.2.18). The line search that we consider is a combination of the Armijo-Goldstein procedure and (4.2.18).

LEMMA 4.2.1 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1 and let $x, d \in \mathbb{R}^n$ be given so that $\nabla f(x)^T d < 0$ and the set $\{f(x + \lambda d) : \lambda > 0\}$ is bounded below. Then for every choice of the scalars $0 < c < \beta < 1$ the set*

$$\left\{ \lambda > 0 \,\middle|\, \begin{array}{c} f(x + \lambda d) - f(x) \leq c\lambda \nabla f(x)^T d, \text{ and} \\ \nabla f(x + \lambda d)^T d \geq \beta \nabla f(x)^T d \end{array} \right\} \tag{4.2.21}$$

is non-empty.

EXERCISE: Prove Lemma 4.2.1.

Thus it is possible to choose a steplength λ such that both the Armijo inequality

$$f(x + \lambda d) - f(x) \leq c\lambda \nabla f(x)^T d$$

and inequality (4.2.18) are satisfied. Concerning this steplength Powell [1976] has established the following convergence result.

THEOREM 4.2.1 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^2. Assume that*

1. *$\nabla^2 f$ is Lipschitz continuous, and*
2. *$\nabla^2 f$ is strongly convex.*

Let $x_0 \in \mathbb{R}^n$, $H_0 \in \mathbb{R}^{n \times n}$ symmetric and positive definite, and $\{x_k\}$ be a sequence defined by

$$s_k := -M_k^{-1} \nabla f(x_k), \; x_{k+1} = x_k + \lambda_k s_k$$

where λ_k is chosen from the set defined in (4.2.21) with $\lambda_k = 1$ being used whenever it is a permissible value, and M_{k+1} is defined as the BFGS update

$$M_{k+1} := M_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{M_k s_K s_k^T M_k}{s_k^T H_k s_k}.$$

Then the sequences $\{x_k\}$ and $\{M_k\}$ are well defined and $\{x_k\}$ converges q-superlinearly to $\overline{x}$ the unique point at which f attains its global minimum value.

We omit the proof of the above result as it is rather involved. It may be found in

M.J.D. Powell, "Some global convergence properties of a variable metric algorithm without exact line searches," in Nonlinear Programming. R. Cottle and C. Lemke, eds. AMS, Providence, R.I. (1976) 53–72.

We mention though that it is still an open problem as to whether a similar result holds for the DFP update.

The variable metric methods discussed in this section are by far the methods of choice for most unconstrained optimization problems when good derivative approximations are available. Observe that one could employ the PSB, BFGS, or DFP formulas to do inverse

updating, but in general this is not done due to possible numerical instabilities that can arise when the hessians are nearly singular. If one wishes to investigate these methods further an excellent survey article is

J.E. Dennis, Jr. and J. J. Moré, "Quasi-Newton methods, motivation and theory," SIAM Review 19 (1977) 46–89.

also see

J.E. Dennis, Jr., and R. B. Schnabel, "Numerical Methods for Unconstrained Optimization and Nonlinear Equations," Prentice-Hall Inc. (1983).

Chapter 5

Optimality Conditions: Constrained Optimization

5.1 First–Order Conditions

In this section we consider first–order optimality conditions for the constrained problem

$$\begin{array}{rll} \mathcal{P}: & \text{minimize} & f_0(x) \\ & \text{subject to} & x \in \Omega, \end{array}$$

where $f_0 : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and $\Omega \subset \mathbb{R}^n$ is closed and non-empty. A very general result can be obtained by simply observing that $\bar{x}$ is a local solution to $\mathcal{P}$ if $f_0'(\bar{x} : d) \geq 0$ for all directions d *pointing into* Ω. To make this statement more precise we define the tangent cone to Ω at a point $x \in \Omega$ to be the set of limiting directions obtained from sequences in Ω that converge to x. Specifically, the tangent cone is given by

$$T_\Omega(x) := \{d : \exists \tau_i \searrow 0, \text{ and } \{x_i\} \subset \Omega, \text{ with } x_i \to x, \text{ such that } \tau_i^{-1}(x_i - x) \to d\}.$$

THEOREM 5.1.1 *If $\bar{x}$ is a local solution to $\mathcal{P}$, the*

$$f_0'(\bar{x}; d) \geq 0 \quad \textit{for all } d \in T_\Omega(\bar{x}).$$

PROOF: The result follows immediately from the fact that

$$f_0'(x; d) = \lim_{\tau \searrow 0} \frac{f_0(\bar{x} + \tau d) - f_0(\bar{x})}{\tau} = \lim_{\substack{s \to d \\ \tau \searrow 0}} \frac{f_0(\bar{x} + \tau s) - f_0(\bar{x})}{\tau}$$

due to the fact that f_0 is continuously differentiable (just apply the Mean–Value Theorem). ■

Although this theorem is a first–order optimality condition, it is not particularly useful without a more concrete description for the set $T_\Omega(\bar{x})$. Recall that in the case of unconstrained optimization we derived the first–order condition $\nabla f_0(\bar{x}) = 0$. This condition is

testable and provides a basis for stopping criteria for our algorithms. In its current form, the condition given by Theorem 5.1.1 has niether of these properties. The derivation of a more useful first–order condition requires a more concrete description of the set Ω.

We begin by assuming that Ω has the form

$$\Omega := \{x : f_i(x) \leq 0, i = 1, \ldots, s, f_i(x) = 0, i = s+1, \ldots, m\}, \tag{5.1.1}$$

where each $f_i : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable on $\mathbb{R}^n$. Observe that if $x \in \Omega$ and $d \in T_\Omega(x)$ then there are sequences $\{x_k\} \subset \Omega$ and $\tau_k \searrow 0$ with $x_k \to x$ such that $\tau_k^{-1}(x_k - x) \to x$. Setting $d_k = \tau_k^{-1}(x_k - x)$ for all k we have that

$$f_i'(x; d) = \lim \frac{f_i(x + \tau_k d_k) - f(x)}{\tau_k}$$

equals 0 for $i \in \{s+1, \ldots, m\}$ and is less than or equal to 0 for $i \in A(x)$ where

$$A(x) := \{i \, : \, i \in \{1, \ldots, s\}, \; f_i(x) = 0\} \; .$$

Consequently,

$$T_\Omega(x) \subset \{d \, : \, \nabla f_i(x)^T d \leq 0, \, i \in A(x), \; \nabla f_i(x)^T d = 0, i = s+1, \ldots, m\} \; .$$

Clearly the set on the right hand side of the inclusion above is a much more tractable representation with respect to the optimality condition given in Theorem 5.1.1. Moreover, it is a somewhat unusual case where these two sets differ. For this reason we make the following definition.

DEFINITION 5.1.1 *We say that the set Ω is regular at $x \in \Omega$ if*

$$T_\Omega(x) = \{d \in \mathbb{R}^n : f_i'(x; d) \leq 0, i \in A(x), f_i'(x; d) = 0 \; i = s+1, \ldots, m\}.$$

Unfortunately, not every set is regular.

EXERCISE: Show that the set

$$\Omega := \{x \in \mathbb{R}^2 | - x_1^3 \leq x_2 \leq x_1^3\}$$

is not regular at the origin. Graph the set Ω.

Before discussing conditions under which one can guarantee the regularity of Ω at a point $x \in \Omega$, we will use regularity to develop a set of tractable first–order necessary conditions for optimality in $\mathcal{P}$. In order to do this we need to recall the strong duality theorem of linear programming.

THEOREM 5.1.2 *(The Duality Theorem of Linear Programming) Let $A \in \mathbb{R}^{s \times n}$, $a \in \mathbb{R}^s$, $B \in \mathbb{R}^{(m-s) \times n}$, $b \in \mathbb{R}^{m-s}$, and $c \in \mathbb{R}^n$, and consider the pair of linear programs*

$$\begin{array}{ll} \max & c^T x \\ \textit{subject to} & Ax \leq a, Bx = b \end{array} \tag{5.1.2}$$

and

$$\begin{array}{ll} \min & a^T u + b^T v \\ \textit{subject to} & A^T u + B^T v = c, u \geq 0. \end{array} \tag{5.1.3}$$

The linear program (5.1.3) is called the dual of (5.1.2). They are related as follows:

1. *The L.P. (5.1.2) has finite value if and only if (5.1.3) has finite value in which case these values coincide and there exist an $\overline{x} \in \mathbb{R}^n$ feasible for (5.1.2) and dual variables $(\overline{u}, \overline{v})$ feasible for (5.1.3) such that $\overline{x}$ solves (5.1.2), $(\overline{u}, \overline{v})$ solves (5.1.3), and*

$$c^T \overline{x} = \overline{u}^T A\overline{x} + \overline{v}^T B\overline{x} = a^T \overline{u} + b^T \overline{v}.$$

2. *If (5.1.2) [(5.1.3)] is infeasible then either (5.1.3) [(5.1.2)] is infeasible or the optimal value in (5.1.3) [(5.1.2)] is $-\infty[+\infty]$.* ■

Although we will not take the time to establish this fundamental result we do observe that if x is feasible for (5.1.2) and (u, v) is feasible for (5.1.3) then

$$\begin{aligned} c^T x \;\; &= [A^T u + B^T v]^T x = u^T(Ax) + v^T(Bx) \\ &\leq a^T u + b^T v. \end{aligned}$$

Hence (5.1.2) $\leq$ (5.1.3) and every dual feasible pair (u, v) provide an upper bound to (5.1.2).

We will now apply Theorem 5.1.1 in conjunction with Theorem 5.1.2 to obtain the main result of this section. Let $\overline{x} \in \Omega$ be a local solution to $\mathcal{P}$ at which Ω is regular and consider the the linear program

$$\begin{array}{lll} \max & (-\nabla f_0(\overline{x}))^T d & \\ \text{subject to} & \nabla f_i(x_0)^T d \leq 0 & i \in I(x_0) \\ & \nabla f_i(x_0)^T d = 0 & i = s+1, \ldots, m. \end{array} \tag{5.1.4}$$

According to Theorem 5.1.2, the dual of (5.1.4) is the linear program

$$\begin{array}{ll} \min\ 0 & \\ \text{subject to} & \sum_{i \in I(x_0)} u_i \nabla f_i(x_0) + \sum_{i=s+1}^m u_i \nabla f_i(x_0) = -\nabla f_0(x_0) \\ & 0 \leq u_i,\ i \in I(x_0). \end{array} \tag{5.1.5}$$

From our assumptions on x_0, Theorem 5.1.1 tells us that the maximum in (5.1.4) is less than or equal to zero. But $d = 0$ is feasible for (5.1.4), hence the maximum value in (5.1.4) is zero. Therefore, by Theorem 5.1.2, the linear program (5.1.5) is feasible, that is, there exist scalars $u_i,\ i \in I(x_0) \cup \{s+1, \ldots, m\}$ with $u_i \geq 0$ for $i \in I(x_0)$ such that

$$0 = \nabla f_0(x_0) + \sum_{i \in I(x_0)} u_i \nabla f_i(x_0) + \sum_{i=s+1}^m u_i \nabla f_i(x_0). \tag{5.1.6}$$

This observation yields the following result.

THEOREM 5.1.3 *Let $x_0 \subset \Omega$ be a local solution to $\mathcal{P}$ at which Ω is regular. Then there exist $u \in \mathbb{R}^m$ such that*

1. $0 = \nabla_x L(x_0, u)$,
2. $0 = u_i f_i(x_0)$ *for* $i = 1, \ldots, s$, *and*
3. $0 \leq u_i$, $i = 1, \ldots, s$,

where the mapping $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is defined by

$$L(x, u) := f_0(x) + \sum_{i=1}^{m} u_i f_i(x)$$

and is called the Lagrangian for the problem $\mathcal{P}$.

PROOF: For $i \in I(x_0) \cup \{s+1, \ldots, m\}$ let u_i be as given in (5.1.6) and for $i \in \{1, \ldots, s\} \setminus I(x_0)$ set $u_i = 0$. Then this choice of $u \in \mathbb{R}^m$ satisfies (1)–(3) above. ■

DEFINITION 5.1.2 *Let $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$. We say that (x, u) is a Kuhn-Tucker pair for $\mathcal{P}$ if*

1. $f_i(x_0) \leq 0$ $i = 1, \ldots, s$, $f_i(x_0) = 0$ $i = s+1, \ldots, m$ *(Primal feasibility),*
2. $u_i \geq 0$ *for* $i = 1, \ldots, s$ *(Dual feasibility),*
3. $0 = u_i f_i(x)$ *for* $i = 1, \ldots, s$ *(complementarity), and*
4. $0 = \nabla_x L(x, u)$ *(stationarity of the Lagrangian).*

Given $x \in \mathbb{R}^n$, if there is a $u \in \mathbb{R}^m$ such that (x, u) is a Kuhn-Tucker pair for $\mathcal{P}$, then we say that x is a stationary point for $\mathcal{P}$. ■

5.2 Regularity and Constraint Qualifications

We now briefly discuss conditions that yield the regularity of Ω at a point $x \in \Omega$. These conditions should be testable in the sense that there is a finitely terminating algorithm that can determine whether they are satisfied or not satisfied. The condition that we will concentrate on is the so called *Mangasarian-Fromovitz constraint qualification* (MFCQ).

DEFINITION 5.2.1 *We say that a point $x \in \Omega$ satisfies the Mangasarian-Fromovitz constraint qualification (or MFCQ) at x if*

1. *there is a $d \in \mathbb{R}^n$ such that*

$$\nabla f_i(x)^T d < 0 \text{ for } i \in I(x),$$
$$\nabla f_i(x)^T d = 0 \text{ for } i = s+1, \cdots, m,$$

and

2. the gradients $\{\nabla f_i(x) | i = s+1, \cdots, m\}$ *are linearly independent.*

We have the following key result which we shall not prove.

THEOREM 5.2.1 *(***MFCQ** $\rightarrow$ **regularity***) Let* $f_i : \mathbb{R}^n \rightarrow \mathbb{R},\ i = 1, 2, \cdots, m$ *be* C^1 *near* $\bar{x} \in \Omega$*. If the MFCQ holds at* $\bar{x}$*, then* Ω *is regular at* $\bar{x}$*.*

The MFCQ is algorithmically verifiable. This is seen by considering the LP

$$\begin{array}{lll} \min & 0 & \\ \text{subject to} & \nabla f_i(\bar{x})^T d \leq -1 & i \in I(\bar{x}) \\ & \nabla f_i(\bar{x})^T d = 0 & i = s+1, \cdots, m. \end{array} \tag{5.2.7}$$

Cleary, if the MFCQ is satisfied at $\bar{x}$ if and only if the above LP is feasible and the gradients $\{\nabla f_i(\bar{x}) \mid i = s+1, \cdots, m\}$ are linearly independent. This observation also leads to a *dual* characterization of the MFCQ by considering the dual of the LP (5.2.7).

PROPOSITION 5.2.2 *The MFCQ is satisfied at a point* $\bar{x} \in \Omega$ *if and only if the only solution to the system*

$$\begin{aligned} \sum_{i=1}^{m} u_i \nabla f_i(\bar{x}) &= 0, \\ u_i f_i(\bar{x}) &= 0 \quad i = 1, 2, \cdots, s, \ \textit{and} \\ u_i &\geq 0 \qquad i = 1, 2, \cdots, s, \end{aligned}$$

is $u_i = 0,\ i = 1, 2, \cdots, m$*.*

PROOF: The dual the LP (5.2.7) is the LP

$$\begin{array}{lll} \min & & \sum_{i \in I(\bar{x})} u_i \\ \text{subject to} & \sum_{i \in I(\bar{x})} u_i \nabla f_i(\bar{x}) + \sum_{i=s+1}^{m} u_i \nabla f_i(\bar{x}) = 0 \\ & & 0 \leq u_i,\ i \in I(\bar{x}). \end{array} \tag{5.2.8}$$

This LP is always feasible, simply take all u_i's equal to zero. Hence, by Theorem 5.1.2, the LP (5.2.7) is feasible if and only if the LP (5.2.8) is finite valued in which case the optimal value in both is zero. That is, the MFCQ holds at $\bar{x}$ if and only if the optimal value in (5.2.8) is zero and the gradients $\{\nabla f_i(\bar{x}) \mid i = s+1, \cdots, m\}$ are linearly independent. The latter statement is equvalent to the statement that the only solution to the system

$$\begin{aligned} \sum_{i=1}^{m} u_i \nabla f_i(\bar{x}) &= 0, \\ u_i f_i(\bar{x}) &= 0 \quad i = 1, 2, \cdots, s, \ \text{and} \\ u_i &\geq 0 \qquad i = 1, 2, \cdots, s, \end{aligned}$$

is $u_i = 0,\ i = 1, 2, \cdots, m$. ■

Techniques similar to these show that the MFCQ is a local property. That is, if it is satisfied at a point then it must be satisfied on a neighborhood of that point. The MFCQ is a powerful tool in the analysis of constraint systems as it implies many useful properties. One such property is established in the following result.

THEOREM 5.2.2 *Let $\bar{x} \in \Omega$ be a local solution to $\mathcal{P}$ at which the set of Kuhn-Tucker multipliers*

$$K\text{–}T(\bar{x}) := \left\{ u \in \mathbb{R}^m \;\middle|\; \begin{array}{c} \nabla_x L(\bar{x}, u) = 0 \\ u_i f_i(\bar{x}) = 0, \;\; i = 1, 2, \cdots, s, \\ 0 \leq u_i, \;\; i = 1, 2, \cdots, s \end{array} \right\} \tag{5.2.9}$$

is non-empty. Then K–$T(\bar{x})$ is a compact set if and only if the MFCQ is satisfied at $\bar{x}$.

PROOF: ($\Rightarrow$) If MFCQ is not satisfied at $\bar{x}$, then from Theorem 5.1.2, Proposition 5.2.2, and the LP (5.2.8) the existence of a non-zero vector $\bar{u} \in \mathbb{R}^m$ satisfying

$$\sum_{i=1}^{m} u_i \nabla f_i(\bar{x}) = 0 \text{ and } 0 \leq u_i \text{ with } 0 = u_i f_i(\bar{x}) \text{ for } i = 1, 2, \cdots, s.$$

Then for each $u \in$ K–T($\bar{x}$) we have that $u + t\bar{u} \in$ K–T($\bar{x}$) for all $t > 0$. Consequently, K–T($\bar{x}$) cannot be compact.
($\Leftarrow$) If K–T($\bar{x}$) is not compact, there is a sequence $\{u^j\} \subset$ K–T($\bar{x}$) with $\|u^j\| \uparrow +\infty$. With no loss is generality, we may assume that

$$\frac{u^j}{\|u^j\|} \to u.$$

But then

$$\begin{array}{rrl} u_i \geq 0, & i = 1, 2, \cdots, s, & \\ u_i f_i(\bar{x}) = & \lim_{i\to\infty} \frac{u^j}{\|u^j\|} f_i(\bar{x}) = & 0, \;\; i = 1, 2, \cdots, s, \text{ and} \\ \sum_{i=1}^m u_i f_i(\bar{x}) = & \lim_{i\to\infty} \frac{\nabla_x L(\bar{x}, u^j)}{\|u^j\|} = & 0. \end{array}$$

Hence, by Proposition 5.2.2, the MFCQ cannot be satisfied at $\bar{x}$. ■

Before closing this section we introduce one more constraint qualification. This is the so called *LI* condition and is associated with the uniqueness of the multipliers..

DEFINITION 5.2.3 *The LI condition is said to be satisfied at the point $x \in \Omega$ if the constraint gradients*

$$\{\nabla f_i(x) \mid i \in I(x) \cup \{s+1, \cdots, m\}\}$$

are linearly independent.

Clearly, the LI condition implies the MFCQ. However, it is a much stronger condition in the presence of inequality constraints. In particular, the LI condition implies the uniqueness of the multipliers at a local solution to $\mathcal{P}$.

5.3 Second–Order Conditions

Second–order conditions are introduced by way of the Lagrangian. As is illustrated in the following result, the multipliers provide a natural way to incorporate the curvature of the constraints.

THEOREM 5.3.1 *Let Ω have representation (5.1.1) and suppose that each of the functions f_i, $i = 0, 1, 2, \ldots m$ are C^2. Let $\overline{x} \in \Omega$ be such that Ω is regular at $\overline{x}$. If $(\overline{x}, \overline{u}) \in \mathbb{R}^n \times \mathbb{R}^m$ is a Kuhn-Tucker pair for $\mathcal{P}$ such that*

$$d^T \nabla_x^2 L(\overline{x}, \overline{u}) d > 0$$

for all $d \in T_\Omega(\overline{x})$, $d \neq 0$, with $\nabla f_0(\overline{x})^T d = 0$, then there is an $\epsilon > 0$ and $\nu > 0$ such that

$$f_0(\overline{x}) \leq f_0(x) - \nu \|x - \overline{x}\|^2$$

for every $x \in \Omega$ with $\|x - \overline{x}\| \leq \epsilon$, in particular $\overline{x}$ is a strict local solution to $\mathcal{P}$.

PROOF: Suppose to the contrary that no such $\epsilon > 0$ and $\nu > 0$ exist, then there exist sequences $\{x_k\} \subset \Omega$, $\{\nu_k\} \subset \mathbb{R}_+$ such that $x_k \to \overline{x}$, $\nu_k \downarrow 0$, and

$$f_0(x_k) \leq f_0(\overline{x}) + \nu_k \|x_k - \overline{x}\|^2$$

for all $k = 1, 2, \ldots$. Now for every $x \in \Omega$ we know that $\overline{u}^T f(x) \leq 0$ and $0 = \overline{u}^T f(\overline{x})$ where the ith component of $f : \mathbb{R}^n \to \mathbb{R}^m$ is f_i. Hence

$$\begin{aligned} L(x_k, \overline{u}) \leq f_0(x_k) \quad &\leq f_0(\overline{x}) + \nu_k \|x_k - \overline{x}\|^2 \\ &= L(\overline{x}, \overline{u}) + \nu_k \|x_k - \overline{x}\|^2. \end{aligned}$$

Therefore,

$$f_0(\overline{x}) + \nabla f_0(\overline{x})^T (x_k - \overline{x}) + o(\|x_k - \overline{x}\|) \leq f_0(\overline{x}) + \nu_k \|x_k - \overline{x}\|^2 \tag{5.3.10}$$

and

$$\begin{aligned} L(\overline{x}, \overline{u}) \quad &+ \nabla_x L(\overline{x}, \overline{u})^T (x_k - \overline{x}) \\ &+ \tfrac{1}{2} (x_k - \overline{x})^T \nabla_x^2 L(\overline{x}, \overline{u})(x_k - \overline{x}) + o(\|x_k - \overline{x}\|^2) \\ &\leq L(\overline{x}, \overline{u}) + \nu_k \|x_k - \overline{x}\|^2 \ . \end{aligned} \tag{5.3.11}$$

Now, with no loss of generality, we can assume that

$$d_k := \frac{x_k - \overline{x}}{\|x_k - \overline{x}\|} \to \overline{d} \in T_\Omega(\overline{x}).$$

By dividing (5.3.10) through by $\|x_k - \overline{x}\|$ and taking the limit we find that $\nabla f_0(x)^T \overline{d} \leq 0$. Since Ω is regular at $\overline{x}$, we know that

$$T_\Omega(\overline{x}) = \{d : \nabla f_i(\overline{x})^T d \leq 0,\ i \in I(\overline{x}),\ \nabla f_i(\overline{x})^T d = 0,\ i = s+1, \ldots, m\}.$$

Therefore $\nabla f_i(x)^T\overline{d} \leq 0,\ i \in I(\overline{x}) \cup \{0\}$ and $\nabla f_i(x)^T\overline{d} = 0$ for $i = s+1, \ldots, m$. On the other land, $(\overline{x}, \overline{u})$ is a Kuhn-Tucker point so

$$\nabla f_0(\overline{x})^T\overline{d} = -\sum_{i \in I(\overline{x})} \overline{u}_i \nabla f_i(\overline{x})^T\overline{d} \geq 0.$$

Hence $\nabla f_0(\overline{x})^T\overline{d} = 0$, so that

$$\overline{d}^T \nabla_x^2 L(\overline{x}, \overline{u})\overline{d} > 0.$$

But if we divide (5.3.11) by $\|x_k - \overline{x}\|^2$ and take the limit, we arrive at the contradiction

$$\frac{1}{2}\overline{d}^T \nabla_x^2 L(\overline{x}, \overline{u})\overline{d} \leq 0,$$

whereby the result is established. ■

The assumptions required to establish Theorem 5.3.1 are somewhat strong but they do lead to a very practical and, in many cases, satisfactory second-order sufficiency result. In order to improve on this result one requires a much more sophisticated mathematical machinery. We do not take the time to develop this machinery. Instead we simply state a very general result. The statement of this result employs the entire set of Kuhn-Tucker multipliers K–T($\bar{x}$).

THEOREM 5.3.2 *Let $\overline{x} \in \Omega$ be a point at which Ω is regular.*

1. *If $\overline{x}$ is a local solution to $\mathcal{P}$, then K–T($\overline{x}$) $\neq \emptyset$, and for every $d \in T_\Omega(\bar{x})$ there is a $u \in$ K–T($\overline{x}$) such that*
$$d^T \nabla_x^2 L(\overline{x}, u)d \geq 0.$$

2. *If K–T($\overline{x}$) $\neq \emptyset$, and for every $d \in T_\Omega(\bar{x})$, $d \neq 0$, for which $\nabla f_0(\overline{x})^T d = 0$ there is a $u \in$ K–T($\overline{x}$) such that*
$$d^T \nabla_x^2 L(\overline{x}, u)d > 0,$$
then there is an $\epsilon > 0$ and $\nu > 0$ such that
$$f_0(\overline{x}) \leq f_0(x) - \nu\|x - \overline{x}\|^2$$
for every $x \in \Omega$ with $\|x - \overline{x}\| \leq \epsilon$, in particular $\overline{x}$ is a strict local solution to $\mathcal{P}$.

5.4 Optimality Conditions in the Presence of Convexity

Just as in the unconstrained case, necessary conditions for optimality become sufficient conditions in the presence of convexity. A key observation in this regard is the equivalence

$$T_\Omega(x) = \overline{\bigcup_{\lambda \geq 0} (\Omega - x)}\ . \tag{5.4.12}$$

This equivalence can be refined to

$$T_\Omega(x) = \bigcup_{\lambda \geq 0} (\Omega - x) \ . \tag{5.4.13}$$

if it is further assumed that Ω is polyhedral.

THEOREM 5.4.1 *If in the problem $\mathcal{P}$ both the function f_0 and the set Ω are assumed to be convex, then x_0 is the global solution to $\mathcal{P}$ if and only if*

$$f_0'(x_0; y - x_0) \geq 0$$

for all $y \in \Omega$.

PROOF: By Theorem 5.1.1, we know that if x_0 is the global solution to $\mathcal{P}$, then $f_0'(x_0; d) \geq 0$ for all $d \in T_\Omega(x)$. Thus, in particular, this inequality holds for $d = y - x_0$ for all $y \in \Omega$ due to the equivalence 5.4.12.

In order to see the reverse implication recall that $f_0'(x; y - x) \leq f_0(y) - f_0(x)$ for all $x, y \in \mathbb{R}^n$. Therefore,

$$0 \leq f_0'(x_0; y - x) \leq f_0(y) - f_0(x_0)$$

for all $y \in \mathbb{R}^n$. ■

The assumption of convexity also yields a somewhat different second–order result. In particular, we now obtain a second–order necessary condition.

THEOREM 5.4.2 *Let $f_0 : \mathbb{R}^n \to \mathbb{R}$ be C^2 and $\overline{x}$ be an element of the convex set Ω.*

1. *(necessity) If $\overline{x} \in \mathbb{R}^n$ is a local solution to $\mathcal{P}$ with Ω a polyhedral convex set, then $\nabla f_0(\overline{x})^T (y - \overline{x}) \geq 0$ for all $d \in T_\Omega(\bar{x})$ and*

$$d^T \nabla^2 f_0(\overline{x}) d \geq 0$$

for all $d \in T_\Omega(\bar{x})$ with $\nabla f(\overline{x})^T d = 0$.

2. *(sufficiency) If $\overline{x} \in \mathbb{R}^n$ is such that $\nabla f_0(\overline{x})^T (y - \overline{x}) \geq 0$ for all $d \in T_\Omega(\bar{x})$ and*

$$d^T \nabla^2 f_0(\overline{x}) d > 0$$

for all $d \in T_\Omega(\bar{x}) \backslash \{0\}$ with $\nabla f_0(\overline{x})^T d = 0$, then there exist $\epsilon, \nu > 0$ such that

$$f_0(\overline{x}) \leq f_0(x) - \nu \|x - \overline{x}\|^2$$

for all $x \in \Omega$ with $\|x - \bar{x}\| \leq \epsilon$.

PROOF: (1) Let $\epsilon > 0$ be such that $f_0(\overline{x}) \leq f_0(x)$ for all $x \in \Omega$ with $\|x - \overline{x}\| \leq \epsilon$ and let $d \in T_\Omega(\bar{x}) = \bigcup_{\lambda \geq 0}(\Omega - \bar{x})$. Then there is a $y \in \Omega$, $y \neq \overline{x}$, and a $\lambda_0 > 0$ such that $d = \lambda_0(y - \overline{x})$. Set $\overline{\lambda} = \min\{\lambda_0, \epsilon(\lambda_0\|y - \overline{x}\|)^{-1}\} > 0$ so that $\overline{x} + \lambda d \in \Omega$ and $\|\overline{x} - (\overline{x} + \lambda d)\| \leq \epsilon$ for all $\lambda \in [0, \overline{\lambda}]$. By hypothesis, we now have

$$\begin{aligned} f_0(\overline{x}) &\leq f_0(\overline{x} + \lambda d) \\ &= f_0(\overline{x}) + \lambda \nabla f_0(\overline{x})^T(y - \overline{x}) + \tfrac{\lambda^2}{2} d^T \nabla^2 f_0(\overline{x}) d + o(\lambda^2) \\ &\leq f_0(\overline{x}) + \tfrac{\lambda^2}{2} d^T \nabla^2 f_0(\overline{x}) d + o(\lambda^2), \end{aligned}$$

where the second inequality follows from Theorem 5.4.1. Therefore $d^T \nabla^2 f_0(\overline{x}) d \geq 0$.

(2) We show that $f_0(\overline{x}) \leq f_0(x) - \nu\|x - \overline{x}\|^2$ for some $\nu > 0$ for all $x \in \Omega$ near $\bar{x}$. Indeed, if this were not the case there would exist sequences $\{x_k\} \subset \Omega$, $\{\nu_k\} \subset \mathbb{R}_+$ with $x_k \to \overline{x}$, $\nu_k \downarrow 0$, and

$$f_0(x_k) < f_0(\overline{x}) + \nu_k \|x_k - \overline{x}\|^2$$

for all $k = 1, 2, \ldots$ where, with no loss of generality, $\frac{x_k - \overline{x}}{\|x_k - \overline{x}\|} \to d$. Clearly, $d \in T_\Omega(\bar{x})$. Moreover,

$$\begin{aligned} f_0(\overline{x}) + \nabla f_0(\overline{x})^T(x_k - \overline{x}) \quad & + o(\|x_k - \overline{x}\|) \\ & = f_0(x_k) \\ & \leq f_0(\overline{x}) + \nu_k \|x_k - \overline{x}\|^2 \end{aligned}$$

so that $\nabla f_0(\overline{x})^T d = 0$.

Now, since $\nabla f_0(\overline{x})^T(x_k - \overline{x}) \geq 0$ for all $k = 1, 2, \ldots$,

$$\begin{aligned} & f_0(\overline{x}) + \tfrac{1}{2}(x_k - \overline{x})^T \nabla^2 f_0(\overline{x})(x_k - \overline{x}) + o(\|x_k - \overline{x}\|^2) \\ & \quad \leq f_0(\overline{x}) + \nabla f_0(\overline{x})^T(x_k - \overline{x}) + \tfrac{1}{2}(x_k - \overline{x})^T \nabla^2 f_0(\overline{x})(x_k - \overline{x}) \\ & \qquad + o(\|x_k - \overline{x}\|^2) \\ & \quad = f_0(x_k) \\ & \quad < f_0(\overline{x}) + \nu_k \|x_k - \overline{x}\|^2. \end{aligned}$$

Hence,

$$\left(\frac{x_k - \overline{x}}{\|x_k - \overline{x}\|}\right)^T \nabla^2 f_0(\overline{x}) \left(\frac{x_k - \overline{x}}{\|x_k - \overline{x}\|}\right) \leq \nu_k + \frac{o(\|x_k - \overline{x}\|^2)}{\|x_k - \overline{x}\|^2}$$

Taking the limit in k we obtain the contradiction

$$0 < d^T \nabla^2 f_0(\overline{x}) d \leq 0,$$

whereby the result is established. ■

Although it is possible to weaken the assumption of polyhedrality in Part 1. of the above theorem, such weakenings are somewhat artificial as they essentially imply that $T_\Omega(x) = \bigcup_{\lambda \geq 0}(\Omega - x)$. The following example illustrates what can go wrong when the assumption of polyhedrality is dropped.

EXAMPLE: Consider the problem

$$\begin{array}{ll} \min \frac{1}{2}(x_2 - x_1^2) & \\ \text{subject to} & 0 \leq x_2, x_1^3 \leq x_2^2. \end{array}$$

Observe that the constraint region in this problem can be written as $\Omega := \{(x_1, x_2)^T : |x_1|^{\frac{3}{2}} \leq x_2\}$, therefore

$$\begin{aligned} f_0(x) &= \frac{1}{2}(x_2 - x_1^2) \\ &\geq \frac{1}{2}(|x_1|^{\frac{3}{2}} - |x_1|^2) \\ &= \frac{1}{2}|x_1|^{\frac{3}{2}}(1 - |x_1|^{\frac{1}{2}}) > 0 \end{aligned}$$

whenever $0 < |x_1| \leq 1$. Consequently, the origin is a strict local solution for this problem. Nonetheless,

$$T_\Omega(0) \cap [\nabla f_0(0)]^\perp = \{(\delta, 0)^T : \delta \in \mathbb{R}\},$$

while

$$\nabla^2 f_0(0) = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}.$$

That is, even though the origin is a strict local solution, the hessian of f_0 is not positive semidefinite on $T_\Omega(0)$.

In employing the second-order conditions given above, one needs to be careful about the relationship between the hessian of f_0 and the set $K := T_\Omega(x) \cap [\nabla f_0(x)]^\perp$. In particular, the positive definiteness (or semidefiniteness) of the hessian of f_0 on the cone K does not necessarily imply the positive definiteness (or semidefiniteness) of the hessian of f_0 on the subspace spaned by K. This is illustrated by the following example.

EXAMPLE: Consider the problem

$$\begin{array}{ll} \min(x_1^2 - \frac{1}{2}x_2^2) & \\ \text{subject to} & -x_1 \leq x_2 \leq x_1. \end{array}$$

Clearly, the origin is the unique global solution for this problem. Moreover, the constraint region for this problem, Ω, satisfies

$$T_\Omega(0) \cap [\nabla f(0)]^\perp = T_\Omega(0) = \Omega\ ,$$

with the span of Ω being all of $\mathbb{R}^2$. Now, while the hessian of f_0 is positive definite on Ω, it is not positive definite on all of $\mathbb{R}^2$.

In the polyhedral case it is easy to see that the sufficiency result in Theorem 5.4.2 is equivalent to the sufficiency result of Theorem 5.3.1. However, in the nonpolyhedral case, these results are not comparable. It is easy to see that Theorem 5.4.2 can handle more

general situations than Theorem 5.4.2 even if Ω is given in the form (5.1.1). Just let one of the active constraint functions be nondifferentiable at the solution. On the other hand, Theorem 5.4.2 can provide information when Theorem 5.4.2 does not. This is illustrated by the following example.

EXAMPLE: Consider the problem

$$\begin{array}{ll} \min & x_2 \\ \text{subject to} & x_1^2 \leq x_2. \end{array}$$

Clearly, $\bar{x} = 0$ is the unique global solution to this convex program. Moreover,

$$\begin{aligned} f_0(\bar{x}) + \frac{1}{2}\|x - \bar{x}\| &= \frac{1}{2}(x_1^2 + x_2^2) \\ &\leq \frac{1}{2}(x_2 + x_2^2) \\ &\leq x_2 = f_0(x) \end{aligned}$$

for all x in the constraint region Ω with $\|x - \bar{x}\| \leq 1$. It is easily verified that this growth property is predicted by Theorem 5.4.2.

5.5 Application to Solving Trust-Region Subproblems

Let us recall the form of the basic trust-region subproblem for unconstrained minimization:

$$\text{(TR)} \qquad \min_{\|x\| \leq \Delta} c^T x + \frac{1}{2} x^T Q x$$

where $Q \in \mathbb{R}^{n \times n}_s$, $c \in \mathbb{R}^n$, and $\Delta > 0$. In this discussion, we assume that the norm is the usual Euclidean norm $\|\cdot\|_2$. Although the matrix Q is not assumed to be positive semi-definite, this problem behaves very much like a convex programming problem with respect to its optimality conditions.

THEOREM 5.5.1 *Consider the problem (TR). A point $\bar{x} \in \Delta\mathbb{B}$ solves (TR) if and only if there is a $\overline{\lambda} \geq 0$ such that*

(a) $\overline{\lambda}[\|\overline{x}\| - \Delta] = 0$.

(b) $c + (Q + \overline{\lambda} I)\overline{x} = 0$, *and*

(c) $Q + \overline{\lambda} I$ *is positive semi-definite.*

PROOF: ($\Longrightarrow$) First observe that since $\|0\| < \Delta$ the Slater constraint qualification holds. Therefore there exists $\overline{\lambda} \geq 0$ such that (a) and (b) hold. To see (c) observe that for every

$x \in \Delta\mathbb{B}$ we have

$$\begin{aligned} 0 &\leq c^T x + \frac{1}{2}x^T Q x - c^T \overline{x} - \frac{1}{2}\overline{x}^T Q \overline{x} \\ &= -\overline{x}^T(Q + \overline{\lambda} I)x + \frac{1}{2}x^T Q x + \overline{x}^T(Q + \overline{\lambda} I)\overline{x} - \frac{1}{2}\overline{x}^T Q \overline{x} \\ &= \frac{1}{2}x^T Q x - \overline{x}^T Q x + \frac{1}{2}\overline{x} Q \overline{x} + \overline{\lambda}\left[\frac{1}{2}\overline{x}^T\overline{x} - 2\overline{x}^T x + \frac{1}{2}x^T x\right] + \frac{\overline{\lambda}}{2}\overline{x}^T\overline{x} - \frac{\overline{\lambda}}{2}x^T x \\ &= \frac{1}{2}(x - \overline{x})^T(Q + \overline{\lambda} I)(x - \overline{x}) + \frac{\overline{\lambda}}{2}[\|\overline{x}\|^2 - \|x\|^2] \end{aligned}$$

so

$$\overline{\lambda}[\|x\|^2 - \|\overline{x}\|^2] \leq (x - \overline{x})^T(Q + \lambda I)(x - \overline{x}).$$

If $\overline{\lambda} = 0$, then this inequality implies that Q is positive semi-definite. If $\lambda > 0$, then $\|\overline{x}\|^2 = \delta$. Therefore,

$$0 \leq (x - \overline{x})^T(Q + \lambda I)(x - \overline{x})$$

for all $x \in \Delta\mathbb{B}$ with $\|x\| = \Delta$. Let $d \in \mathbb{R}^n$ be such that $d^T\overline{x} \neq 0$ and $\|d\| = 1$. Setting $x = \overline{x} - 2d^T\overline{x}d$ we get that $\|x\| = \Delta$ so that $(2d^T\overline{x})^2 d^T(Q + \lambda I)d \geq 0$. By continuity, $Q + \lambda I$ is therefore positive semi-definite.

($\Longleftarrow$) Consider the Lagrangian

$$\begin{aligned} L(x, \overline{\lambda}) &= c^T x + \frac{1}{2}x^T Q x + \frac{\overline{\lambda}}{2}[\|x\|^2 - \Delta^2] \\ &= c^T x + \frac{1}{2}x^T[Q + \overline{\lambda} I]x - \frac{\overline{\lambda}}{2}\Delta^2. \end{aligned}$$

Since $Q + \overline{\lambda} I$ is positive semi-definite, $L(x, \overline{\lambda})$ is convex in x. Therefore, $L(\cdot, \overline{\lambda})$ attains a global minimum at any point x for which $\nabla_x L(x, \overline{\lambda}) = 0$. Hence $\overline{x}$ is a solution to $\min\{L(x, \overline{\lambda}) : x \in \mathbb{R}^k\}$. Thus, in particular, for all $x \in \Delta\mathbb{B}$

$$\begin{aligned} c^T x + \frac{1}{2}x^T Q x &\geq c^T x + \frac{1}{2}x^T Q x + \frac{\overline{\lambda}}{2}[\|x\|^2 - \Delta^2] \\ &\geq L(\overline{x}, \overline{\lambda}) \\ &= c^T\overline{x} + \frac{1}{2}\overline{x}^T Q \overline{x}. \end{aligned}$$

Consequently, $\overline{x}$ is a global solution to (TR). ■

This is a remarkable result on the structure of the problem (TR). It also provides the key to a very efficient method for solving (TR). Indeed,either there is a unique solution to (TR) lying in the interior of $\Delta\mathbb{B}$, in which case Q is necessarily positive definite, or there is a solution lying on the boundary of $\Delta\mathbb{B}$. In the latter case, we need only solve the system

$$\begin{aligned} (Q + \lambda I)x &= -c \\ 0 \leq \lambda, \qquad \|x\|^2 &= \Delta^2 \end{aligned}$$

subject to the condition that $Q + \lambda I$ is positive semi-definite. In this regard, observe that if $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ is the spectrum of Q, then $\lambda + \lambda_1 \leq \lambda + \lambda_2 \leq \cdots \leq \lambda + \lambda_n$ is the spectrum of $\lambda I + Q$. Hence the condition that $\lambda \geq 0$ and $Q + \lambda I$ is positive semi-definite boils down to saying that

$$\lambda \geq \lambda_l := \max\{0, -\lambda_1\}.$$

That is, λ_l is a lower bound on λ. In the case that $\lambda \neq 0$, one can also obtain an upper bound on λ. Indeed, if $\lambda \neq 0$, then $\|\overline{x}\| = \Delta$. In this case the condition $\overline{\lambda}\overline{x} = -(c + Q\overline{x})$ implies that

$$\overline{\lambda}\|\overline{x}\| \leq \|c\| + \|Q\| \, \|\overline{x}\|,$$

and since $\|\overline{x}\| = \Delta$ we have

$$\overline{\lambda} \leq \frac{\|c\|}{\Delta} + \|Q\| =: \lambda_u$$

is an upper bound on $\overline{\lambda}$. If it is further known that Q is positive semi-definite then this upper can be refined. In this case we have

$$\overline{\lambda}\overline{x}^T\overline{x} = -c^T\overline{x} - \overline{x}^TQ\overline{x} \leq -c^T\overline{x} \leq \|c\| \, \|\overline{x}\|,$$

and so $\overline{\lambda} \leq \|c\|/\Delta = \lambda_u^*$ is a better upper bound. Also observe that since Q is symmetric

$$\|Q\| = \max\{|\lambda_1|, |\lambda_n|\},$$

and both λ_1 and λ_n can be estimated using the power method.

Having both upper and lower bounds on $\overline{\lambda}$ one can now consider applying Newton's method to the system

$$\begin{aligned} (Q + \lambda I)x &= -c \\ \|x\|^2 &= \Delta^2 \end{aligned}$$

over $[\lambda_l, \lambda_u]$. Although this approach is reasonable, it is far from the most efficient. We now discuss an alternative approach based on well-known facts from the theory of rational functions.

Consider the function

$$\varphi(\lambda) = \frac{1}{\Delta} - \frac{1}{\|(\lambda I + Q)^{-1}c\|}$$

on the interval $(\lambda^*, +\infty)$. We claim that φ is a convex function on the interval $(\lambda^*, +\infty)$. In order to see this let $\{v_1, \ldots, v_n\}$ be an orthonormal basis of eigenvectors for Q with associated eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, respectively, and suppose

$$c = \sum_{i=1}^{n} \mu_i v_i \ .$$

Then

$$\varphi(\lambda) = \frac{1}{\Delta} - \left[\sum_{i=1}^{n} \left(\frac{\mu_i}{\lambda + \lambda_i}\right)^2\right]^{-1/2}.$$

Hence

$$\varphi'(\lambda) = -\left[\sum_{i=1}^{n}\left(\frac{\mu_i}{\lambda+\lambda_i}\right)^2\right]^{-3/2}\left[\sum_{i=1}^{n}\mu_i^{-1}\left(\frac{\mu_i}{\lambda+\lambda_i}\right)^3\right]$$

and

$$\varphi''(\lambda) = 3\left[\sum_{i=1}^{n}\left(\frac{\mu_i}{\lambda+\lambda_i}\right)^2\right]^{-5/2}\left[\left(\sum_{i=1}^{n}\mu_i^{-2}\left(\frac{\mu_i}{\lambda+\lambda_i}\right)^4\right)\left(\sum_{i=1}^{n}\left(\frac{\mu_i}{\lambda+\lambda_i}\right)^2\right) - \left(\sum_{i=1}^{n}\mu_i^{-1}\left(\frac{\mu_i}{\lambda+\lambda_i}\right)^3\right)^2\right].$$

Recall that the Cauchy-Schwartz inequality implies that

$$\left[\sum_{i=1}^{n}a_ib_i\right]^2 \le \sum_{i=1}^{n}a_i^2\sum_{i=1}^{n}b_i^2.$$

Setting $a_i = \frac{\mu_i}{\lambda+\lambda_i}$ and $b_i = \mu_i^{-1}\left(\frac{\mu_i}{\lambda+\lambda_i}\right)^2$, we find that

$$\left[\sum_{i=1}^{n}\mu_i^{-1}\left(\frac{\mu_i}{\lambda+\lambda_i}\right)^3\right]^2 \le \left[\sum_{i=1}^{n}\left(\frac{\mu_i}{\lambda+\lambda_i}\right)^2\right]\left[\sum_{i=1}^{n}\mu_i^{-2}\left(\frac{\mu_i}{\lambda+\lambda_i}\right)^4\right].$$

Therefore, $\varphi''(\lambda) \ge 0$ on $(\lambda^*, +\infty)$ so that φ is convex there.

Moreover, φ is nearly linear on $(\lambda^*, +\infty)$ and is essentially flat for all large λ. This is obvious from the expression for $\varphi''(\lambda)$ given above since $\varphi''(\lambda) = O\left(\frac{1}{\lambda}\right)$.

We can now describe a general procedure for solving (TR). First obtain an estimate for λ_1 the smallest eigenvalue of Q. If $\lambda_1 > 0$, then Q is positive definite. In this case we terminate at a solution $\overline{x} = -Q^{-1}c$ if $\|Q^{-1}c\| < \Delta$. If $\lambda_2 \le 0$ or $\|Q^{-1}c\| > \Delta$, then apply Newton's method to locate a zero of the function φ on the interval $\left(\lambda_1, \frac{\|c\|}{\Delta} + \|Q\|\right]$. That is, we are applying Newton's method to a nearly linear convex function on $\mathbb{R}$. Since φ is nearly linear, this procedure converges rapidly if a solution exists. Since φ is convex we also have that

$$\lambda_{k+1} < \lambda_k \text{ if } \overline{\lambda} < \lambda_k$$

and

$$\lambda_{k+1} > \lambda_k \text{ if } \overline{\lambda} > \lambda_k.$$

Moreover, if $\lambda_k < \overline{\lambda}$, then $\lambda_{k+1} < \overline{\lambda}$. All of these facts follow from the subdifferential inequality of convex analysis.

The difficult case occurs when $\overline{\lambda} = \lambda_l$. In this case $Q + \lambda I$ is not invertible if Q is not positive definite. If in measuring the behavior of the iterate, it is thought that $\overline{\lambda} = \lambda_l$, then one can terminate by computing a least-norm solution to the equation $-c = (Q + \lambda_l I)x$.

The best known methods basically work in this way. However, they also incorporate more sophisticated methods for approximating and updating λ_l and λ_u at each iteration. We do not go into this here. Nonetheless, we should mention that if the Newton step λ_{k+1} is such that

$$\lambda_{k+1} < \lambda_l$$

then one should reset $\lambda_{k+1} = \frac{1}{2}(\lambda_k + \lambda_l)$ in order to preserve the inclusion $\lambda_k \in (\lambda_l, \lambda_u]$.

Chapter 6

LP's, QP's, and LCP's

6.1 Introduction

The KKT conditions for linear and quadratic programming yield an instance of a more general class of problems called *linear complementarity problems*. In order to see this connection, consider the quadratic program

$$\begin{array}{ll} \mathcal{Q} & \text{minimize} \quad \frac{1}{2}u^T Q u - c^T u \\ & \text{subject to} \quad Au \leq b,\ 0 \leq u, \end{array}$$

where $Q \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{m\times n}$, $c \in \mathbb{R}^n$, and $b \in \mathbb{R}^m$. If we assume that Q is positive semi-definite, then $\mathcal{Q}$ is a convex programming problem. By setting $Q = 0$, we recover linear programming as a special case.

Define

$$M \ = \ \begin{pmatrix} Q & A^T \\ -A & 0 \end{pmatrix} \quad \text{and} \quad q \ = \ \begin{pmatrix} c \\ b \end{pmatrix}. \tag{6.1.1}$$

Then the KKT conditions for the quadratic program $\mathcal{Q}$ are equivalent to the conditions

$$y = Mx + q, \quad y^T x = 0, \quad 0 \leq x, \quad \text{and} \quad 0 \leq y\ ,$$

where

$$x \ = \ \begin{pmatrix} u \\ v \end{pmatrix}.$$

The classical approach to solving $\mathcal{Q}$ when $Q = 0$ is the simplex algorithm due to George Danzig. If $Q \neq 0$, then the corresponding method is called Lemke's algorithm. Both of these approaches are known as pivoting methods. There is an enormous literature on methods of this type. Nonetheless, we do not consider pivoting strategies in this section. Rather we consider an approach of a much more modern vintage. This approach is defined for a more general class of problems known as linear complementarity problems or LCP's.

(LCP) Find $x \in \mathbb{R}^n$ such that

$$y = Mx + q,\ x^T y = 0,\ 0 \le x,\ 0 \le y$$

where $M \in \mathbb{R}^{n\times n}$ and $q \in \mathbb{R}^n$.

We denote by $\mathcal{S}$ the solution set to (LCP):

$$\mathcal{S} := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : F(x, y) = 0, 0 \le x, 0 \le y\}.$$

Of special interest to us is the case where M is assumed to be positive semi-definite, but **not** necessarily symmetric.

DEFINITION 6.1.1 *The problem (LCP) is said to be a monotone linear complementarity problem if the matrix M is positive semi-definite.*

6.2 Boundedness Properties of LCP

The algorithms we consider are designed to solve monotone LCP's. One of the most important properties of monotone LCP's is that they are naturally associated with the following convex quadratic program:

$$\text{(QP-LCP)} \qquad \begin{array}{l} \min x^T(Mx + q) \\ \text{subject to } 0 \le Mx + q,\ 0 \le x. \end{array}$$

That this is a convex QP follows immediately from the fact that M is positive semi-definite. Moreover, the optimal value of (QP–LCP) is non-negative since $0 \le Mx+q$ and $0 \le x$. The optimal value is zero precisely when $\mathcal{S} \neq \emptyset$ in which case the solution set is given by

$$\{x : \exists\ y \ge 0 \text{ such that } (x, y) \in \mathcal{S}\}.$$

This observation is the key to analyzing the boundedness properties of the set $\mathcal{S}$ and the sets

$$\mathcal{F}(t) = \{(x, y) : 0 \le x,\ 0 \le y,\ Mx + q = y,\ x^T y \le t\}$$

for $t \ge 0$. These sets play an important role in the analysis to follow.

Note that since (QP-LCP) is a convex quadratic program its solution set is a convex set. Thus, in particular, this implies that the set $\mathcal{S}$ is always a closed convex set (although possibly empty). Moreover, since $x^T(Mx + q)$ is a convex function, all of its level sets are convex as well. In particular, we get that the sets

$$\mathcal{F}_1(t) = \{x\ :\ 0 \le x,\ 0 \le Mx + q,\ x^T(Mx + q) \le t\}$$

are closed convex sets for every $t \ge 0$. Therefore, the sets

$$\mathcal{F}(t) = \{(x, y)\ :\ x \in \mathcal{F}_1(t),\ y = Mx + q\}$$

are closed convex sets as well since they are the linear image of a closed convex set.

We now consider the boundedness of the sets $\mathcal{F}(t)$. For this we make use of the following sets:

$$\begin{aligned} \mathcal{F} &:= \{(x,y) \in \mathbb{R}^n \times \mathbb{R}^n : Mx + q = y, 0 \leq x, 0 \leq y\} \\ \mathcal{F}_+ &:= \{(x,y) \in \mathbb{R}^n \times \mathbb{R}^n : Mx + q = y, 0 < x, 0 < y\}. \end{aligned}$$

THEOREM 6.2.1 *If M is positive semi–definite and $\mathcal{F}_+$ is nonempty, then $\mathcal{F}(t)$ is bounded for all $t \geq 0$.*

PROOF: Let $(\bar{x}, \bar{y}) \in \mathcal{F}_+$ and let $(x,y) \in \mathcal{F}(t)$. Then $(x - \bar{x})^T(y - \bar{y}) \geq 0$ since M is positive semi–definite. Therefore,

$$t + \bar{x}^T\bar{y} \geq x^Ty + \bar{x}^T\bar{y} \geq \bar{x}^Ty + \bar{y}^Tx \geq \kappa\|(x,y)\|_1 \ ,$$

where $\kappa := \min_{1=1,2,\ldots,n}\{\bar{x}_i, \bar{y}_i\}$. ■

6.3 The Central Path

Given a vector $x \in \mathbb{R}^n$ we denote by X the diagonal matrix $\mathrm{diag}(x)$. Hence $Y = \mathrm{diag}(y)$, $U = \mathrm{diag}(u)$, $W = \mathrm{diag}(w)$, etc. Consider the function

$$F(x,y) = \begin{bmatrix} Mx - y + q \\ XYe \end{bmatrix}$$

where $e \in \mathbb{R}^n$ is the vector of all cones. Clearly, $(x,y) \in \mathbb{R}^n \times \mathbb{R}^n$ solves (LCP) if and only if $0 \leq x, y$, and $F(x,y) = 0$. The basic idea behind interior point algorithms for solving (LCP) is to apply a damped Newton's method to the function $F(x,y)$ on the interior of the cone $\mathbb{R}^n_+ \times \mathbb{R}^n_+$. In this regard, following result is key.

THEOREM 6.3.1 *If M is positive semi-definite, then $F'(x,y)$ is non-singular whenever $0 < x$, $0 < y$.*

PROOF: Let $(x,y) \in \mathrm{int}(\mathbb{R}^n \times \mathbb{R}^n)$ and suppose that $F'(x,y)\begin{pmatrix} u \\ v \end{pmatrix} = 0$. Then

$$v = Mu \text{ and } v = -X^{-1}Yu.$$

Hence, $0 \geq -u^TX^{-1}Yu = u^TMu \geq 0$, so $u^TX^{-1}Yu = 0$ or $u = 0$. But then $v = 0$ as well. ■

Thus, the Newton step is well defined at points in $\mathrm{int}(\mathbb{R}^n_+ \times \mathbb{R}^n_+)$. Moreover, one can always choose a step length so that a damped Newton step stays in $\mathrm{int}(\mathbb{R}^n_+ \times \mathbb{R}^n_+)$. However, it may happen that the iterates approach the boundary of $\mathbb{R}^n_+ \times \mathbb{R}^n_+$ too quickly and the procedure gets bogged down. For this reason we introduce the notion of a central path.

DEFINITION 6.3.1 *The set*

$$\mathcal{C} := \{(x,y) \in \mathcal{F} : XYe = te \text{ for some } t > 0\}$$

is called the central path for (LCP).

We now proceed to show that if $\mathcal{F}_+ \neq \emptyset$ and M is positive semi-definite, then $\mathcal{C}$ exists. The first step is to establish the following lemma concerning the function

$$u(x,y) = XYe.$$

LEMMA 6.3.1 *Suppose M is positive semi–definite and $\mathcal{F}_+ \neq \emptyset$.*

(1) The system

$$u(x,y) = a \text{ and } (x,y) \in \mathcal{F}_+$$

has a solution for every $a > 0$.

(2) The mapping $u : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is diffeomorphism between $\mathcal{F}_+$ and $int(\mathbb{R}^n_+)$, i.e. u is a one–to–one surjective mapping between $\mathcal{F}_+$ and $int(\mathbb{R}^n_+)$ with $u \in C^\infty$ on $\mathcal{F}_+$ and $u^{-1} \in C^\infty$ on $int(\mathbb{R}^n_+)$.

PROOF:

(1) Let $a > 0$ and $(\overline{x}, \overline{y}) \in \mathcal{F}_+$. Set $\overline{a} = u(\overline{x}, \overline{y})$. Consider the function

$$\widehat{F}(x,y,t) := F(x,y) - \begin{bmatrix} 0 \\ (1-t)\overline{a} + ta \end{bmatrix}.$$

Note that $\widehat{F}(\overline{x}, \overline{y}, 0) = 0$ and

$$\nabla_{(x,y)} \widehat{F}(x,y,t) = \nabla F(x,y) = \begin{bmatrix} M & -1 \\ Y & X \end{bmatrix}.$$

Hence, by the implicit function theorem, there is an open neighborhood $U \subset \mathbb{R}^n \times \mathbb{R}^n$ containing $(\overline{x}, \overline{y})$, $\delta > 0$, and a unique smooth mapping $t \mapsto (x(t), y(t))$ on $[0, \delta)$ such that

$$(x(t), y(t)) \in U \text{ and } \widehat{F}(x(t), y(t)) = 0 \text{ on } [0, \delta).$$

Let $\overline{\delta}$ be the largest such δ in $[0,1]$. We claim that $\overline{\delta} = 1$. First observe that $(x(t), y(t)) \in \mathcal{F}(\overline{t})$ for $\overline{t} := \max\{\overline{a}^T e, a^T e\}$. Moreover, $\mathcal{F}(\overline{t})$ is a compact set by Theorem 6.2.1. Hence, for some sequence $t_i \uparrow \overline{\delta}$ there exists an $(\widehat{x}, \widehat{y})$ such that $(x,(t_i), y(t_i)) \to (\widehat{x}, \widehat{y})$. Clearly, $(\widehat{x}, \widehat{y}) \in \mathcal{F}_{++}$. Applying the implicit function theorem again at $(\widehat{x}, \widehat{y})$ yields a contradiction to the maximality of $\overline{\delta}$. Finally, observe that

$$F(x(1), y(1)) = a$$

which establishes the result.

(2) In Part (1) above, we have already shown that u is a surjective map from $\mathcal{F}_+$ to $\mathbb{R}^n_+$. We now show that it is one–to–one. Assume to the contrary, that $u(x^1, y^1) = u(x^2, y^2)$ for distinct points (x^1, y^1) and (x^2, y^2) in $\mathcal{F}_+$. Then

$$M(x^1 - x^2) = y^1 - y^2 \text{ and } x^1_i y^1_i = x^2_i y^2_i > 0 \; \forall \; i = 1, \ldots, n.$$

Since $(x^1 - x^2)^T M(x^1 - x^2) \geq 0$, we have

$$(x^1 - x^2)^T (y^1 - y^2) \geq 0.$$

Hence for some i with $x^1_i \neq x^2_i$ we must have $(x^1_i - x^2_i)(y^1_i - y^2_i) \geq 0$. If $x^1_i > x^2_i$, then $y^1_i \geq y^2_i > 0$. But then $x^1_i y^1_i \neq x^2_i y^2_i$. Similarly, if $x^1_i < x^2_i$, then $0 < y^1_i \leq y^2_i$, so again $x^1_i y^1_i \neq x^2_i y^2_i$. This contradiction establishes that u is one–to–one.

Finally, it is clear that u is C^∞. To see that u^{-1} is C^∞ simply note that $(u^{-1})'(a) = [XM + Y]^{-1}$ where $u^{-1}(a) = (x, y)$. To see that $[XM + Y]^{-1}$ exists write $[XM + Y] = X[M + X^{-1}Y]$ where both X and $[M + X^{-1}Y]$ are positive definite matrices.

■

An immediate consequence of this Lemma is the following existence theorem for (LCP).

THEOREM 6.3.2 *If M is positive semi-definite and $\mathcal{F}_+ \neq \emptyset$, then there exists a solution to (LCP).*

PROOF: Let $(\overline{x}, \overline{y}) \in \mathcal{F}_+$. Then $\mathcal{F}(\overline{x}^T \overline{y})$ is compact by Theorem **??**. Moreover, the system $F(x, y) = \begin{bmatrix} 0 \\ \mu \overline{x}^T \overline{y} e \end{bmatrix}$ is solvable for all $\mu \in (0, 1]$. Hence there exist $\{(x_i, y_i)\} \subset \mathcal{F}_+$, $\mu_i \downarrow 0$, and $(\widehat{x}, \widehat{y}) \in \mathcal{F}$ such that $(x_i, y_i) \to (\widehat{x}, \widehat{y})$ and $F(x_i, y_i) = \begin{bmatrix} 0 \\ \mu_i \overline{x}^T \overline{y} e \end{bmatrix}$. But then $F(\widehat{x}, \widehat{y}) = 0$ so that $(\widehat{x}, \widehat{y}) \in \mathcal{S}$. ■

The existence of the central path can now also be established. The proof is similar to the proof given for Part 2 of Lemma 6.3.1.

THEOREM 6.3.3 *If M is positive semi-definite and $\mathcal{F}_+ \neq \emptyset$, then the central path $\mathcal{C}$ exists as a smooth curve in $\mathcal{F}_+$.*

PROOF: Just compose the smooth trajectory $\{te : t > 0\} \subset \text{int}\,(\mathbb{R}^n_+)$ with the diffeomorphism u^{-1} in Lemma 6.3.1 to obtain the result. ■

6.3.1 Asymptotic behavior of the central path

In this section we study the limiting behavior of the central path as $t \downarrow 0$. In particular, we show that this limit exists and is a solution of (LCP). The key to this analysis is the potential function

$$P(x, y, t) = x^T y - t \sum_{i=1}^{n} \ln(x_i y_i)$$

defined over the set $\mathcal{F}_+ \times \{t > 0\}$. Let us first observe that for fixed $t > 0$ the function $P(\cdot, \cdot, t)$ is strictly convex on $\mathcal{F}_+$. In order to see this observe that

$$\nabla^2_{(x,y)} P(x, y, t) = \begin{bmatrix} tX^{-2} & I \\ I & tY^{-2} \end{bmatrix}.$$

Hence if (x_1, y_1), $(x_2, y_2) \in \mathcal{F}_+$, then

$$\begin{bmatrix} x_1 - x_2 \\ y_1 - y_2 \end{bmatrix}^T \nabla^2_{(x,y)} P(x,y,t) \begin{bmatrix} x_1 - x_2 \\ y_1 - y_2 \end{bmatrix} = t[(x_1-x_2)^T X^{-2}(x_1-x_2) + (y_1-y_2)^T Y^{-2}(y_1-y_2)] + 2(x_1-x_2)^T(y_1-y_2) > 0.$$

Therefore, for each $t > 0$, the solution to the problem

$$\begin{array}{ll} (\mathrm{P}_t) & \min P(x, y, t) \\ & \text{subject to } (x, y) \in \mathcal{F}_+, \end{array}$$

if it exists, is unique. With this in mind we give the following theorem.

THEOREM 6.3.4 *If M is positive semi-definite and $\mathcal{F}_+ \neq \emptyset$, then the unique solution to the problem (P_t) exists and corresponds to the unique solution of the equation $F(x, y) = \begin{bmatrix} 0 \\ te \end{bmatrix}$, i.e., it lies on the central path.*

PROOF: Due to our observation concerning the strict convexity of $P(x, y, t)$, we need only show that the unique solution, $(x(t), y(t))$, to $F(x, y) = \begin{bmatrix} 0 \\ te \end{bmatrix}$ satisfies the first-order optimality conditions for (P_t). The first-order conditions for (P_t) are

$$\nabla_{(x,y)} P(x, y, t) \in \ker \begin{bmatrix} M & -I \end{bmatrix}^{\perp} = \mathrm{Ran} \begin{bmatrix} M \\ -I \end{bmatrix}.$$

Since $\nabla_{(x,y)} P(x, y, t) = \begin{bmatrix} y - tx^{-1} \\ x - ty^{-1} \end{bmatrix}$, where $(x^{-1})_i := (x_i)^{-1}$, these conditions imply the existence of a vector $v \in \mathbb{R}^n$ such that

$$\begin{aligned} y - tx^{-1} &= Mv \\ x - ty^{-1} &= -v. \end{aligned}$$

Multiplying the first of these equations by X and the second by Y, we get the system

$$\begin{aligned} Xy - te &= XMv \\ Yx - te &= -Yv. \end{aligned}$$

Therefore, $[XM+Y]v = 0$, or equivalently, $[M+X^{-1}Y]v = 0$. But $[M+X^{-1}Y]$ is a positive definite matrix so we must have $v = 0$. Consequently, the conditions $Xy = te$, $0 < x$, $0 < y$, and $Mx + q = y$, are equivalent to the first-order necessary and sufficient conditions in (P_t). The unique solution of this system is $(x(t), y(t))$ so this is the unique solution to (P_t). ■

Next set

$$\begin{aligned} E &= \{i : x_i = 0 = y_i \text{ for all } (x,y) \in \mathcal{S}\}, \\ B &= \{i : x_i \neq 0 \text{ for some } (x,y) \in \mathcal{S}\}, \text{ and} \\ N &= \{i : y_i \neq 0 \text{ for some } (x,y) \in \mathcal{S}\}. \end{aligned}$$

We make the following observations about these index sets:

1. Since $\mathcal{S}$ is convex, there exists $(\widehat{x}, \widehat{y}) \in \mathcal{S}$ with

   $$\begin{aligned} \widehat{x}_i > 0 \quad \forall \quad i \in B, \text{ and} \\ \widehat{y}_i > 0 \quad \forall \quad i \in N. \end{aligned}$$

   To obtain $(\widehat{x}, \widehat{y})$ just take a convex combination of points (x, y) for which $x_i > 0$ $i \in B$ and $y_i > 0$ for $i \in N$.

2. Due to the above observation we have $B \cap N = \emptyset$. This implies that the sets B, E, and N form a partition of the integers from 1 to n, i.e. $\{1, 2, \ldots, n\} = B \cup N \cup E$ with $B \cap N = \emptyset$, $B \cap E = \emptyset$, and $N \cap E = \emptyset$.

3. For all $(x, y) \in \mathcal{S}$ we have $x_i = 0$ for all $i \in \{1, 2, \ldots, n\} \backslash \bar{B}$ and $y_i = 0$ for all $i \in \{1, 2, \ldots, n\} \backslash \bar{N}$, where $\bar{B} = B \cup E$ and $\bar{N} = N \cup E$.

4. The solution set $\mathcal{S}$ has the representation

   $$\mathcal{S} = \left\{ (x,y) \;\middle|\; \begin{array}{ll} 0 \leq x_B, & 0 \leq y_N, \\ 0 = x_{\overline{N}}, & 0 = y_{\overline{B}}, \end{array} \; M_B x_B + q = y_N \right\}, \tag{6.3.2}$$

We claim that the limit as $t \searrow 0$ in the central path is the unique solution to the problem

$$\begin{array}{ll} (\mathrm{P}_0) & \min - [\sum_B \ln x_i + \sum_N \ln y_i] \\ & \text{subject to } (x,y) \in \mathcal{S}. \end{array}$$

One can view the problem (P_0) as the limit of the problems (P_t) as $t \searrow 0$. Observe that

$$-\sum_B \ln x_i - \sum_N \ln y_i = -\ln \left[\left(\prod_B x_i \right) \left(\prod_N y_i \right) \right].$$

Therefore, since $-\ln(\mu)$ is strictly decreasing for $\mu > 0$, minimizing $-\ln\left[\left(\prod_B x_i\right)\left(\prod_N y_i\right)\right]$ over $\mathcal{S}$ is the same as maximizing $\left(\prod_B x_i\right)\left(\prod_N y_i\right)$ over $\mathcal{S}$. That is, (P_0) is equivalent to the problem

$$(\hat{P}_0) \quad \begin{array}{l} \max \left(\prod_B x_i\right)\left(\prod_N y_i\right) \\ \text{subject to } (x,y) \in \mathcal{S}. \end{array} \tag{6.3.3}$$

Using this fact we can show that the problem (P_0) has a solution and that it is unique.

LEMMA 6.3.2 *If M is positive semi-definite and $\mathcal{F}_+ \neq \emptyset$, then the solution (x^*, y^*) to (P_0) exists, is unique, and satisfies $x_B^* > 0$ and $y_N^* > 0$.*

PROOF: By Theorem **??**, $\mathcal{S}$ is a compact set. Hence the solution to $(\hat{P}_0)$, or equivalently (P_0), exists since problem $(\hat{P}_0)$ is the maximization of a continuous function over a compact set. The fact that the solution is unique is the consequence of the fact that the objective function in (P_0) is strictly convex on $\mathcal{S}$ as seen by considering the representation (6.3.2). The condition that the solution (x^*, y^*) satisfies $x_B^* > 0$ and $y_N^* > 0$ follows from the finiteness of the optimal value. ■

Before proving the main result, we first establish the following technical lemma.

LEMMA 6.3.3 *Let $(x^*, y^*) \in \mathcal{S}$ be the unique solution to (P_0), let $(x, y) \in \mathcal{C}$, and set $\mu = x^T y/n$. Then $XYe = \mu e$, $\sum_B \frac{x_i^*}{x_i} + \sum_N \frac{y_i^*}{y_i} \leq n$, $x_B \geq \frac{1}{n} x_B^* > 0$, and $y_N \geq \frac{1}{n} y_N^* > 0$.*

PROOF: As usual,

$$\begin{array}{rcl} 0 & \leq & (x - x^*)^T (y - y^*) \\ & = & x^T y - x^{*T} y - x^T y^* + x^{*T} y^*, \end{array}$$

so

$$x^{*T} y + x^T y^* \leq x^T y = n\mu.$$

Since $(x, y) \in \mathcal{C}$, we have $XYe = \mu e$ so

$$x = \mu y^{-1} \text{ and } y = \mu x^{-1}.$$

But then

$$\begin{array}{rcl} \mu(x^{*^T} x^{-1} + y^{*^T} y^{-1}) & \leq & x^{*T} y + y^{*T} x \\ & = & \mu n, \end{array}$$

or equivalently,

$$\sum_B \frac{x_i^*}{x_i} + \sum_N \frac{y_i^*}{y_i} \leq n.$$

Due to the positivity of each term in the sum, we get that

$$\frac{x_i^*}{x_i} \leq n \quad \text{for } i \in B \quad \text{and} \quad \frac{y_i^*}{y_i} \leq n \quad \text{for } i \in N,$$

or equivalently,

$$\frac{1}{n} x_B^* \leq x_B \quad \text{and} \quad \frac{1}{n} y_N^* \leq y_N.$$

■

THEOREM 6.3.5 *Let M be positive semi-definite, $\mathcal{F}_+ \neq \emptyset$, and assume that $E = \emptyset$. Then the limit of the central path $\mathcal{C}$ exists as $t \downarrow 0$ and is the solution to the problem (P_0).*

PROOF: Let $(\widehat{x}, \widehat{y})$ be any cluster point of $\mathcal{C}$ as $t \downarrow 0$. Since $(\widehat{x}, \widehat{y}) \in \mathcal{S}$, we have that $\widehat{x}_{\overline{N}} = 0$ and $\widehat{y}_{\overline{B}} = 0$. Since this is true for every cluster point, we obtain that $x_{\overline{N}}(t) \to 0$ and $y_{\overline{B}}(t) \to 0$.

Letting (x^*, y^*) be the unique solution to (P_0) and taking the limit as $t \searrow 0$, we obtain from Lemma 6.3.3 that

$$\sum_B \frac{x_i^*}{\hat{x}_i} + \sum_N \frac{y_i^*}{\hat{y}_i} \leq n, \tag{6.3.4}$$

$\hat{x}_B \geq \frac{1-\beta}{n} x_B^* > 0$, and $\hat{y}_N \geq \frac{1-\beta}{n} y_N^* > 0$. Thus, in particular, $(\hat{x}, \hat{y})$ is feasible for (P_0).

Next recall that the arithmetic–geometric mean inequality says that for any collection $\{\gamma_1, \gamma_2, \ldots, \gamma_N\}$ of non–negative real numbers we have that

$$\left(\prod_{i=1}^N \gamma_i \right)^{1/N} \leq \frac{1}{N} \sum_{i=1}^N \gamma_i \ .$$

Therefore, by (6.3.4) and the fact that $B \cup N = \{1, 2, \ldots, n\}$, we have

$$\left(\prod_B \frac{x_i^*}{\hat{x}_i} \prod_N \frac{y_i^*}{\hat{y}_i} \right) \leq \left(\frac{1}{n} \sum_B \frac{x_i^*}{\hat{x}_i} + \sum_N \frac{y_i^*}{\hat{y}_i} \right)^n \leq 1^n = 1 \ .$$

Consequently,

$$\begin{aligned} \left(\prod_B x_i^* \prod_N y_i^* \right) &= \left(\prod_B \hat{x}_i \prod_N \hat{y}_i \right) \left(\prod_B \frac{x_i^*}{\hat{x}_i} \prod_N \frac{y_i^*}{\hat{y}_i} \right) \\ &\leq \left(\prod_B \hat{x}_i \prod_N \hat{y}_i \right). \end{aligned}$$

But then $(\widehat{x}, \widehat{y})$ must also be a solution to $(\hat{P}_0)$ in which case $(\widehat{x}, \widehat{y}) = (x^*, y^*)$ by uniqueness. Since (x^*, y^*) is the only possible cluster point, it must be the case that the limit of the central path is (x^*, y^*). ■

6.4 A Theoretical Infeasible Interior Point Algorithm

In this section we consider a specific algorithm for solving the equation

$$F(x,y) = 0$$

where, as before,

$$F(x,y) = \begin{bmatrix} Mx+q-y \\ XYe \end{bmatrix}.$$

The procedure is initiated in the region $\text{int}(\mathbb{R}^n_+ \times \mathbb{R}^n_+)$ and generates iterates that stay in this region. For this reason such methods are called *interior point methods*. Given a point $(x^0, y^0) \in \text{int}(\mathbb{R}^n_+ \times \mathbb{R}^n_+)$ one needs to compute an iterate that reduces the value of both $\|Mx^0+q-y^0\|$ and ${x^0}^T y^0$. If it so happens that $Mx^0+q-y^0 = 0$ from the outset, then this quality is preserved. Indeed, if all of the iterates must satisfy the equation $Mx+q-y=0$, then the method is called a *feasible* interior point method (FIP). If the iterates do not necessarily satisfy this equation, then the method is called an *infeasible* interior point method (IIP). From the practical point of view, the (IIP) methods are more tractable since it is often very difficult to obtain an initial $(x^0, y^0) > 0$ with $Mx^0 + q = y^0$. Indeed, it may be that no such (x^0, y^0) exists. Nonetheless, if we set

$$\mathcal{F}_+(q) = \{(x,y) > 0 : Mx+q = u\},$$

and

$$\mathcal{F}(q) = \{(x,y) \geq 0 : Mx+q = y\},$$

then if $\mathcal{F}(q) \neq \emptyset$ and $\epsilon > 0$, there exists $\widehat{q}$ with $\|q-\widehat{q}\| \leq \epsilon$ such that $\mathcal{F}_+(q) \neq \emptyset$. Simply take $(\overline{x}, \overline{y}) \in \mathcal{F}(q)$ and $(u,v) > 0$ and choose $\delta > 0$ such that $\delta\|Mu - v\| \leq \epsilon$, then set $(\widehat{x}, \widehat{y}) = (\overline{x} + \delta u, \overline{y} + \delta v) > 0$ and $\widehat{q} = q - \delta(Mu - v)$ so that $\|q - \widehat{q}\| \leq \epsilon$ while

$$\begin{aligned} M\widehat{x} + \widehat{q} &= M\overline{x} + \delta Mu + q - \delta Mu + \delta v \\ &= \overline{y} + \delta v = \widehat{y}. \end{aligned}$$

The algorithm that we consider has two basic features:

(1) The quantities $x^T y$ and $\|Mx+q-y\|$ are reduced at the same rate at each iteration, and

(2) all iterates stay in a fixed neighborhood of the central path;

$$N(t) = \{(x,y) \geq 0 \,:\, \phi_t(x,y) \leq \beta\,\}$$

where $\beta > 0$ and the function $\phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is given by

$$\phi_t(x,y) = \|te - XYe\|_2 / t\ .$$

The basic idea of the iteration is as follows: At each iteration we take a damped Newton stet (the predictor step) for the equations

$$XYe = 0, \quad Mx + q - y = 0;$$

This step is followed by a Newton step (the corrector step) for the equations

$$XYe = te, \quad Mx + q - y = -s$$

for a suitably chosen vector s. The purpose of the *corrector step* is to return the iterates to a position closer to the central path.

Infeasible Interior Point Algorithm:

Initialization: Choose $(\beta_0, \beta_1, \beta_2) \in \mathbb{R}^3$ satisfying

$$0 < \beta_0 < \beta_1 < \beta_2 < \sqrt{\beta_0} < 1$$

and set

$$\eta + 1 = \frac{\beta_0 - \beta_2^2}{\beta_0 + \sqrt{n}}. \tag{6.4.5}$$

Find $(x^0, y^0, t_0) \in \mathbb{R}^{2n+1}_+$ satisfying

$$\varphi_{t_0}(x^0, y^0) \leq \beta_0.$$

Having (x^ν, y^ν, t_ν) obtain $(x^{\nu+1}, y^{\nu+1}, t_{\nu+1})$ as follows:

Predictor Step: Set

$$\widehat{x}^\nu = x^\nu + \theta_\nu \Delta x^\nu, \quad \widehat{y}^\nu = y^\nu + \theta_\nu \Delta y^\nu, \quad \widehat{t}^\nu = (1 - \theta_n)\widehat{t}^\nu \tag{6.4.6}$$

where $(\Delta x^\nu, \Delta y^\nu)$ is the unique vector in $\mathbb{R}^{2n}$ satisfying

$$\begin{aligned} Y^\nu \Delta x^\nu + X^\nu \Delta y^\nu &= -X^\nu y^\nu, \\ M\Delta x^\nu - \Delta y^\nu &= -(Mx^\nu + q - y^\nu), \end{aligned} \tag{6.4.7}$$

and θ_ν is the largest $\theta \in (0, 1]$ satisfying $(x^\nu + \theta \Delta x^\nu, y^\nu + \theta \Delta y^\nu) > 0$ and

$$\varphi_{(1-\theta)t_\nu}(x^\nu + \theta \Delta x^\nu, y^\nu + \theta \Delta y^\nu) \leq \beta_1. \tag{6.4.8}$$

Corrector Step: Set

$$x^{\nu+1} = \widehat{x}^\nu + \Delta \widehat{x}^\nu, \quad y^{\nu+1} = \widehat{y}^\nu + \Delta \widehat{y}^\nu, \quad t_{\nu+1} = (1 - \gamma_\nu)\widehat{t}^\nu, \tag{6.4.9}$$

where γ_ν is the largest $\gamma \in (0, \eta_1]$ satisfying

$$\|\widehat{t}_\nu e - \widehat{X}^\nu \widehat{y}^\nu - \gamma \widehat{X}^\nu (M\widehat{x}^\nu + q - \widehat{y}^\nu)\| / \widehat{t}_\nu \leq \beta_2, \tag{6.4.10}$$

and $(\Delta \widehat{x}^\nu, \Delta \widehat{y}^\nu)$ is the unique vector satisfying

$$\begin{aligned} \widehat{t}^\nu [\widehat{X}^\nu]^{-1} \Delta \widehat{x}^\nu + \widehat{X}^\nu \Delta \widehat{y}^\nu &= \widehat{t}^\nu e - \widehat{X}^\nu \widehat{y}^\nu \\ M\Delta \widehat{x}^\nu - \Delta \widehat{y}^\nu &= -\gamma_\nu (M\widehat{x}^\nu + q - \widehat{y}^\nu). \end{aligned} \tag{6.4.11}$$

Remarks:

1. To obtain θ_ν in the predictor step, first compute the largest value, $\overline{\theta}$, of $\theta \in (0,1]$ for which
$$(x^\nu + \theta\Delta x^\nu, y^\nu + \theta\Delta y^\nu) \geq 0.$$
Then solve the equation
$$\varphi_{(1-\theta)t_\nu}(x^\nu + \theta\Delta x^\nu, y^\nu + \theta\Delta y^\nu)^2 = \beta_1^2.$$
This is a fourth degree polynomial in θ. Take θ_ν to be the largest root of this polynomial that is less than $\overline{\theta}$.

2. To obtain γ_ν in the corrector step, just find the roots of the quadratic polynomial
$$\|\widehat{t}_\nu e - \widehat{X}^\nu\widehat{y}^\nu - \gamma\widehat{X}^\nu(M\widehat{x}^\nu + q - \widehat{y}^\nu)\|^2 = (\beta_0\widehat{t}^\nu)^2$$
and take γ_ν to be the largest root less than η_1.

3. The expression "$\widehat{t}^\nu[\widehat{X}^\nu]^{-1}$" in the computation of the update in the Corrector step can be replaced by "$\widehat{Y}^\nu$" without effecting the convergence of the iterates. We choose the so-called *primal scaling* in order to simplify the analysis.

4. One choice for β_0, β_1, and β_2 is $\beta_0 = .2, \beta_1 = .201$, and $\beta_2 = .3$. In this case one can take $\eta = \frac{1.1}{2+10\sqrt{n}}$.

5. The initial values for (x^0, y^0, t_0) can be taken to be $x^0 = e$, $y^0 = e$, $t_0 = \beta_0$. However, this ignores the feasibility condition $Mx + q = y$. This can slow down convergence and inhibit rapid local convergence. To compensate for this set $\mu = .9\|Me + q\|_\infty^{-1}$ if $Me + q \neq 0$; otherwise set $\mu = 0$. Then take $(x^0, y^0, t_0) = (e, e + \mu(Me + q), .9)$ and $\beta_0 = 1$. This choice is allowed since $x^0 > 0$, $y^0 > 0$, and $\|X^0y^0 - e\|_2 = \mu\|Me + q\|_2 = .9\frac{\|Me+q\|_2}{\|Me+q\|_\infty} \leq .9$.

Our first objective is to show that the iterates defined in this way all satisfy
$$(x^\nu, y^\nu) > 0 \text{ and } \varphi_{t_\nu}(x^\nu, y^\nu) \leq \beta_0.$$

LEMMA 6.4.1 *Fix any $(\beta_0, \beta_1, \beta_2) \in \mathbb{R}^3$ so that $0 < \beta_0 < \beta_1 < \beta_2 < \sqrt{\beta_0} < 1$, and let η_1 be given by (6.4.5). If $(\widehat{x}^\nu, \widehat{y}^\nu, \widehat{t}_\nu) \in \mathbb{R}^{2n+1}$ satisfies $\varphi_{\widehat{t}_\nu}(\widehat{x}^\nu, \widehat{y}^\nu) \leq \beta_1$, then the quantity $(\Delta\widehat{x}^\nu, \Delta\widehat{y}^\nu, \gamma_\nu)$ computed by the corrector step satisfies*
$$(\widehat{x}^\nu + \Delta\widehat{x}^\nu, \widehat{y}^\nu + \Delta\widehat{y}^\nu) > 0, \quad \varphi_{(1-\gamma_\nu)\widehat{t}_\nu}(\widehat{x}^\nu + \Delta\widehat{x}^\nu, \widehat{y}^\nu + \Delta\widehat{y}^\nu) \leq \beta_0.$$

PROOF: For simplicity, set $(x, y, t) = (\widehat{x}^\nu, \widehat{y}^\nu, \widehat{t}^\nu)$ and $(\Delta x, \Delta y) = (\Delta\widehat{x}^\nu, \Delta\widehat{y}^\nu)$. Let
$$r = te - Xy, \quad s = Mx + q - y, \quad \text{and } d = X^{-1}\Delta x.$$

Then (6.4.11) can be written as

$$td + X\Delta y = r, \quad MXd - \Delta y = -\gamma s. \tag{6.4.12}$$

Since $(XMX + tI)$ is positive definite we can solve for d to obtain

$$d = (XMX + tI)^{-1}(r - \gamma Xs).$$

Since XMX is positive semi-definite, we get

$$\begin{aligned} t\|d\|^2 &\leq d^T(XMX + tI)d \\ &= d^T(r - \gamma Xs) \\ &\leq \|d\| \, \|r - \gamma Xs\|, \end{aligned} \tag{6.4.13}$$

so that

$$\|d\| \leq \|r - \gamma Xs\|/t. \tag{6.4.14}$$

Let $x' = x + \Delta x$ and $y' = y + \Delta y$. Equations (6.4.10) and (6.4.14) yield $\|d\| \leq \beta_2 < 1$ so that $e + d > 0$. Also $x > 0$, so $x' = x + Xd = X(e + d) > 0$. Therefore,

$$\begin{aligned} te - X'y' &= te - (I + D)X(y + \Delta y) \\ &= td - DX(y + \Delta y) \\ &= D[te - Xy - X\Delta y] \\ &= tDd, \end{aligned}$$

where the last three equations follow from (6.4.12). Hence

$$\begin{aligned} \|te - X'y'\|_2 &= t\|Dd\|_2 \\ &\leq t\|Dd\|_1 \\ &= t\|d\|_2^2 \\ &\leq \|r - \gamma Xs\|^2/t \\ &\leq (\beta_2)^2 t. \end{aligned}$$

Since $\beta_2 < 1$ and $x' > 0$, this relation implies that $y' > 0$. In conjunction with the triangle inequality, the relation also implies that

$$\begin{aligned} \varphi_{(1-\gamma)t}(x', y') &= \|(1 - \gamma)te - X'y'\|/((1 - \gamma)t) \\ &\leq \|te - X'y'\|/((1 - \gamma)t) + \gamma\sqrt{n}/(1 - \gamma) \\ &\leq (\beta_2)^2/(1 - \gamma) + \gamma\sqrt{n}/(1 - \gamma) \\ &= [(\beta_2)^2 + \gamma\sqrt{n}](1 - \gamma)^{-1} \\ &\leq [(\beta_2)^2 + \eta_1\sqrt{n}](1 - \eta_1)^{-1} \\ &= \beta_0. \end{aligned}$$

■

The following lemma allows us to bound $\|(x^\nu, y^\nu)\|$ from above and certain components of (x^ν, y^ν) from below.

LEMMA 6.4.2 *Let* $(x^*, y^*) \in S$, $\mu \in [0,1]$, *and* (x^0, y^0), $(x, y) \in \mathbb{R}^n_+ \times \mathbb{R}^n_+$ *with and* $Mx + q - y = \mu(Mx^0 + q - y^0)$. *Then*

$$\mu(x^T y^0 + y^T x^0) \leq x^T y + \mu(x^{0T} y^0 + x^{*T} y^0 + x^{0T} y^*)$$

and

$$(1-\mu)(x^T y^* + y^T x^*) \leq x^T y + \mu(x^{0T} y^0 + x^{*T} y^0 + x^{0T} y^*).$$

PROOF: Since $Mx^* + q = y^*$, we have

$$M(x - \mu x^0 - (1-\mu)x^*) = (y - \mu y^0 - (1-\mu)y^*).$$

Multiplying both sides by $x - \mu x^0 - (1-\mu)x^*$ and using the fact that M is positive semi-definite yields

$$0 \leq (x - \mu x^0 - (1-\mu)x^*)^T (y - \mu y^0 - (1-\mu)y^*).$$

Rearranging this inequality yields the inequality

$$\begin{array}{rcl} \mu(x^T y^0 + y^T x^0) & + & (1-\mu)(x^T y^* + y^T x^*) \\ \leq x^T y & + & \mu^2 x^{0T} y^0 + \mu(1-\mu)(x^{*T} y^0 + x^{0T} y^*) \end{array}$$

since $x^{*T} y^* = 0$. This yields the result since $\mu \in [0,1]$. ■

We now have the following global convergence result.

THEOREM 6.4.1 *Let* $\{(x^\nu, y^\nu, t_\nu, \widehat{x}^\nu, \widehat{y}^\nu, \widehat{t}_\nu, \theta_\nu, \gamma_\nu)\}$ *be generated by Algorithm 6.4. Then*

$$(x^\nu, y^\nu) > 0 \quad , \quad \varphi_{t_\nu}(x^\nu, y^\nu) \leq \beta_0 \tag{6.4.15}$$

$$t_\nu = \mu_\nu t_0 \quad , \quad Mx^\nu + q - y^\nu = \mu_\nu(Mx^0 + q - y^0) \tag{6.4.16}$$

$$\widehat{t}_\nu = (1-\theta_\nu)\mu_\nu t_0 \quad , \quad M\widehat{x}^\nu + q - \widehat{y}^\nu = (1-\theta_\nu)\mu_\nu(Mx^0 + q - y^0) \tag{6.4.17}$$

for all ν, *where*

$$\mu_\nu = (1-\gamma_{\nu-1})(1-\theta_{\nu-1})\cdots(1-\gamma_0)(1-\theta_0).$$

If $\{(\widehat{x}^\nu, \widehat{y}^\nu)\}$ *is bounded, then* $S \neq \emptyset$. *If* $S \neq \emptyset$, *then* $\{(\widehat{x}^\nu, \widehat{y}^\nu)\}$ *and* $\{(x^\nu, y^\nu)\}$ *are bounded and for any* $(x^*, y^*) \in S$ *we have* $\gamma_\nu \geq \min\{\eta_1, \eta_2\}$ *for all* ν *where* $\eta_2 = \eta_1$ *if* $Mx^0 + q - y^0 = 0$; *otherwise*

$$\eta_2 = \frac{(\beta_2 - \beta_1)t_0(\min_i y_i^0)/\|Mx^0 + q - y^0\|_\infty}{(1+\beta_1)nt_0 + x^{0T}y^0 + x^{*T}y^0 + x^{0T}y^*}.$$

PROOF:The relations (6.4.16), (6.4.17), and (6.4.17) are easily verfiied by induction on ν. Now for every ν we have $\varphi_{\widehat{t}_\nu}(\widehat{x}^\nu, \widehat{y}^\nu) \leq \beta_1$ which implies that

$$\widehat{X}^\nu \widehat{y}^\nu \leq (1+\beta_1)\widehat{t}_\nu e.$$

This, in turn, yields

$$\widehat{x}^{\nu T}\widehat{y}^\nu \leq (1+\beta_1)n(1-\theta_\nu)\mu_\nu t_0 \tag{6.4.18}$$

by (6.4.17). The inequality $\varphi_{\widehat{t}_\nu}(\widehat{x}^\nu, \widehat{y}^\nu) \leq \beta_1$ also yields the inequality

$$\begin{aligned}
&\|\widehat{t}_\nu e - \widehat{X}^\nu \widehat{y}^\nu - \gamma \widehat{X}^\nu (M\widehat{x}^\nu + q - \widehat{y}^\nu)\|/\widehat{t}_\nu \\
&\quad \leq \varphi_{\widehat{t}_\nu}(\widehat{x}^\nu, \widehat{y}^\nu) + \gamma \|\widehat{X}(M\widehat{x}^\nu + q - \widehat{y}^\nu)\|/\widehat{t}_\nu \\
&\quad \leq \beta_1 + \gamma \|\widehat{X}^\nu (M\widehat{x}^\nu + q - \widehat{y}^\nu)\|/\widehat{t}_\nu
\end{aligned}$$

for all $\gamma \geq 0$. Let $\overline{\gamma}^\nu$ be the largest γ for which the lefthand side of this inequality is less than or equal to β_2. This designation for $\overline{\gamma}^\nu$ implies that $\gamma^\nu = \min\{\eta_1, \overline{\gamma}^\nu\}$. Note in particular, that this value for $\overline{\gamma}^\nu$ must exceed the value of γ for which the righthand side of the inequality is equal to β_2, i.e., $\overline{\gamma}^\nu = +\infty$ if $Mx^* + q - y^0 = 0$; otherwise

$$\begin{aligned}
\overline{\gamma}^\nu &\geq \frac{(\beta_2 - \beta_1)\widehat{t}_\nu}{\|\widehat{X}^\nu (M\widehat{x}^\nu + q - \widehat{y}^\nu)\|} \geq \frac{(\beta_2 - \beta_1)\widehat{t}_\nu}{\|\widehat{x}^\nu\|, \|Mx^\nu + q - y^\nu\|_\infty} \\
&= \frac{(\beta_2 - \beta_1)t_0}{\|\widehat{x}^\nu\|, \|Mx^0 + q - y^0\|_\infty}
\end{aligned} \tag{6.4.19}$$

where the equality follows from the relations

$$\widehat{t}_\nu = (1 - \theta_\nu)\mu_\nu t_0 \quad \text{and} \quad M\widehat{x}^\nu + q - \widehat{y}^\nu = (1 - \theta_\nu)\mu_\nu (Mx^0 + q - y^0)$$

(see relations (6.4.17).

Let us now assume that the sequence $\{(\widehat{x}^\nu, \widehat{y}^\nu)\}$ is bounded. In this case (6.4.19) implies that $\overline{\gamma}_\nu \geq \eta$ for some $\eta > 0$. Hence $\gamma_\nu = \min\{\eta_1, \overline{\gamma}_\nu\} \geq \min\{\eta_1, \eta\}$ for all ν and so $\mu_\nu \to 0$. Since $(\widehat{x}^\nu, \widehat{y}^\nu) > 0$, relations (6.4.18) and (6.4.17) imply that any cluster point of the sequence $\{(\widehat{x}^\nu, \widehat{y}^\nu)\}$ is in S.

Next assume that $S \neq \emptyset$ and fix $(x^*, y^*) \in S$. The relations (6.4.16) and (6.4.17) imply that the conclusion of Lemma 6.4.2 holds with $(x, y) = (\widehat{x}^\nu, \widehat{y}^\nu)$ and $\mu = (1 - \theta_\nu)\mu_\nu$. Therefore,

$$\begin{aligned}
&\|\widehat{x}^\nu\|_1 (\min_i y_i^0) + \|\widehat{y}^\nu\|_1 (\min_i x_i^0) \\
\leq\ & \widehat{x}^{\nu T} y^0 + \widehat{y}^{\nu T} x^0 \\
\leq\ & \widehat{x}^{\nu T} \widehat{y}^\nu / \mu_\nu + x^{0T} y^0 + x^{*T} y^0 + x^{0T} y^* \\
\leq\ & (1 + \beta_1) n t_0 + x^{0T} y^0 + x^{*T} y^0 + x^{0T} y^*
\end{aligned}$$

where the last inequality follows from (6.4.18). Since $(x^0, y^0) > 0$ this shows that $\{(\widehat{x}^\nu, \widehat{y}^\nu)\}$ is bounded. A similar argument using (6.4.17)shows that the sequence $\{(x^\nu, y^\nu)\}$ is bounded. Combining the above inequality with (6.4.19) yields $\overline{\gamma}_\nu \geq \eta_1$ if $Mx^0 + g - y^0 = 0$ and

$$\widehat{\gamma}_\nu \geq \frac{(\beta_2 - \beta_1) t_0 (\min_i y_i^0) / \|Mx_q^0 - y^0\|_\infty}{(1 + \beta_1) n t_0 + x^{0T} y^0 + x^{*T} y^0 + x^{0T} y^*}$$

if $Mx^0 + q - y^0 \neq 0$, for all ν. Hence $\gamma^\nu = \min\{\eta_1, \overline{\gamma}_\nu\} \geq \min\{\eta_1, \eta_2\}$. ■

6.5 A Practical Infeasible Interior Point Algorithm

A major drawback of the algorithm presented in the previous section is the tortuous care with which the parameters defining the algorithm must be chosen. The second drawback is that the algorithm does not perform in practise nearly as well as similar algorithms that are much more easily designed and implemented, but for which a complete convergence theory does not yet exist. We present just such an algorithm in this section. Again, the basic idea is to try to follow the central path to the solution. In order to do this the algorithm must be constructed so that it stays close to the central path while reducing the *homotopy* parameter t at each iteration. Then as t is reduced to zero we hopefully converge to a solution. There are several obstacles that must be overcome for this strategy to succeed. The most obvious and significant of these is that it is very difficult to locate points in the set $\mathcal{F}_+$ let alone points on the central path. For this reason we consider algorithms that initialize at points satisfying $0 < x$ and $0 < y$ but for which the affine constraint $Mx + q = y$ may be violated. Algorithms of this type are called *infeasible* interior point algorithms.

Infeasible interior point algorithms must balance reduction in the homotopy parameter t with reduction in the residual of the affine constraints $Mx+q = y$. Indeed, the overall success of the procedure depends on how this balance in achieved. In general, one must reduce these two quantities at roughly the same rate while simultaneously staying sufficiently close to the central path. An algorithm that attempts to achieve this balance is given below.

Infeasible Interior Point Algorithm for LCP

Initialization

$$\epsilon = 10^{-8} \qquad \begin{pmatrix} \text{stopping} \\ \text{tolerance} \end{pmatrix}$$

$$\sigma = 0.3 \qquad \begin{pmatrix} \text{homotopy} \\ \text{scaling parameter} \end{pmatrix}$$

$$x^0 = 2e \qquad \text{(initial } x\text{)}$$

$$(y^0)_i = \max\{\, (Mx^0 + q)_i,\, 2\},\ i = 1, 2, \ldots, n \qquad \text{(initial } y\text{)}$$

$$\tau = (x^0)^T y^0/n \qquad \begin{pmatrix} \text{homotopy} \\ \text{parameter} \end{pmatrix}$$

$$\rho = \|Mx^0 - y^0 + q\|_\infty \qquad \text{(residual)}$$

Iteration While $n\tau > \epsilon$ or $\rho > \epsilon$,

Step 1 (Compute the Newton Step)

Solve the linear equation

$$F(x^k, y^k) + F'(x^k, y^k) \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} 0 \\ \sigma\tau e \end{pmatrix},$$

or equivalently, solve the equation

$$\begin{bmatrix} M & -I \\ Y & X \end{bmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} -Mx^k + y^k - q \\ \sigma\tau e - X_k Y_k e \end{pmatrix},$$

for Δx and Δy.

Step 2 (Compute a Feasible Steplength)

$$\begin{aligned} t_x &= \min\left\{ \tfrac{-(x^k)_i}{(\Delta x)_i} \ : \ (\Delta x)_i < 0 \right\} \\ t_x &= \min\{\, 1,\, 0.999 t_x \} \end{aligned}$$

$$\begin{aligned} t_y &= \min\left\{ \tfrac{-(y^k)_i}{(\Delta y)_i} \ : \ (\Delta y)_i < 0 \right\} \\ t_y &= \min\{\, 1,\, 0.999 t_y \} \end{aligned}$$

Step 3 (Update Iterates)

$$\begin{aligned} x^{k+1} &= x^k + t_x \Delta x \\ y^{k+1} &= y^k + t_y \Delta y \\ k &= k + 1 \\ \tau &= (x^k)^T y^k / n \\ \rho &= \|Mx^k - y^k + q\|_\infty \end{aligned}$$

Step 4 (Update Scaling Parameter)

$$\sigma = \begin{cases} 1 & \text{, if } n\tau \le \epsilon \text{ and } \rho > \epsilon, \\ \min\left\{ .3,\, (1-t_x)^2,\, (1-t_y)^2,\, \tfrac{|\rho - n\tau|}{\rho + 10n\tau} \right\} & \text{, otherwise.} \end{cases}$$

Chapter 7

The Gradient Projection Algorithm

7.1 Projections and Optimality Conditions

In this section we study the problem

$$\begin{aligned} \mathcal{P}: \quad & \min f(x) \\ & \text{subject to } x \in \Omega \end{aligned}$$

where $\Omega \subset \mathbb{R}^n$ is assumed to be a nonempty closed convex set and f is C^1. The solution method that we will study is known as the gradient projection algorithm and was pioneered by Allen Goldstein of the University of Washington in 1964. In Theorem 5.4.1 we found that if $\overline{x}$ is a local minimum for $\mathcal{P}$ then

$$\nabla f(\overline{x})^T(y - \overline{x}) \leq 0 \tag{7.1.1}$$

for all $y \in \Omega$. Moreover, if f is convex, then condition rfeqgp1 implies that $\overline{x}$ is a local minimum for $\mathcal{P}$. An instance of the function f that is of particular significance is

$$f(x) := \frac{1}{2}\|x - x_0\|_2^2.$$

In this case problem $\mathcal{P}$ becomes one of finding the closest point $\overline{x}$ in Ω to x_0. By applying Theorem 5.4.1 one obtains the celebrated projection theorem for convex sets.

THEOREM 7.1.1 *Let $x_0 \in \mathbb{R}^n$ and let $\Omega \subset \mathbb{R}^n$ be a nonempty closed convex set. Then $\overline{x} \in \Omega$ solves the problem*

$$\min\{\frac{1}{2}\|x - x_0\|_2^2 : x \in \Omega\}$$

if and only if

$$(\overline{x} - x_0)^T(y - \overline{x}) \geq 0 \tag{7.1.2}$$

for all $y \in \Omega$. Moreover, the solution $\overline{x}$ always exists and is unique.

Proof. Existence follows from the compactness of the set

$$\{x \in \Omega : \|x - x_0\|_2 \leq \|\hat{x} - x_0\|_2\}$$

where $\hat{x}$ is any element of Ω. Uniqueness follows from the strong convexity of the 2-norm squared. The remainder of the theorem follows immediately from Theorem 5.4.1 once it is observed that if

$$f(x) = \frac{1}{2}\|x - x_0\|_2^2$$

then

$$\nabla f(x) = x - x_0. \qquad \blacksquare$$

DEFINITION 7.1.1 *Let $\Omega \subset \mathbb{R}^n$ be nonempty closed convex. We define the projection into Ω to be the mapping $P_\Omega : \mathbb{R}^n \to \Omega$ given by*

$$\frac{1}{2}\|P_\Omega(x) - x\|_2^2 = \min\{\frac{1}{2}\|y - x\|_2^2 : y \in \Omega\}.$$

Observe that P_Ω is well-defined by Theorem 7.1.1.

We now introduce two geometric concepts that aid in interpreting the optimality condition given in Theorem 7.1.1. Recall that the tangent cone to Ω at a point $x_0 \in \Omega$ is given by

$$T_\Omega(x_0) = \overline{\bigcup_{\lambda > 0} \lambda(\Omega - x_0)}.$$

Dually, we call the set

$$N_\Omega(x) := \{z : \langle z, y - z\rangle \leq 0 \quad \text{for all } y \in \Omega\}$$

the normal cone to Ω at x.

Using the notions of a normal cone and a tangent cone we obtain the following restatements of Theorems 5.4.1 and 7.1.1.

THEOREM 7.1.2 *Let $\overline{x}$ be a solution to problem $\mathcal{P}$ and suppose that f is differentiable at $\overline{x}$, then*

$$-\nabla f(\overline{x}) \in N_\Omega(\overline{x}). \tag{7.1.3}$$

Moreover, if f is convex then (7.1.3) is sufficient for $\overline{x}$ to be a global minimizer of f on Ω.

Proof. We need only show that condition (7.1.3) is equivalent to the statement that

$$\nabla f(\overline{x})^T(y - \overline{x}) \geq 0 \quad \text{for all } y \in \Omega.$$

But this is clear from the definition of the normal cone. $\blacksquare$

THEOREM 7.1.3 *Let Ω be a non-empty closed convex subset of $\mathbb{R}^n$ and let P_Ω denote the projector into Ω. Then given $x \in \mathbb{R}^n$ we have $z = P_\Omega(x)$ if and only if*

$$(x - z) \in N_\Omega(z). \tag{7.1.4}$$

Proof. We need only show that (7.1.4) is equivalent to (7.1.2), but again this follows immediately from the definition of the normal cone. ■

We have the following interesting corollary.

COROLLARY 7.1.3.1 *Let $x \in \Omega$, $z \in N_\Omega(x)$, and $t \geq 0$, then*

$$P_\Omega(x + tz) = x.$$

Proof. Simply observe that

$$(x + tz) - P_\Omega(x + tz) = tz \in N_\Omega(x),$$

so that the result follows from the theorem. ■

This yields the following corollary to Theorem 7.1.1 in the context of $\mathcal{P}$.

COROLLARY 7.1.3.2 *Let $\overline{x}$ be a solution to $\mathcal{P}$, then*

$$P_\Omega(\overline{x} - t\nabla f(\overline{x})) = \overline{x} \tag{7.1.5}$$

for all $t \geq 0$.

Proof. Just apply Theorem 7.1.1 and Corollary 7.1.3.1. ■

We now show how (7.1.5) can be used both as a stopping criteria for our algorithm and as a method for generating search directions.

PROPOSITION 7.1.2 *Let $x \in \Omega$ and set $d = P_\Omega(x - t\nabla f(x)) - x$. Then*

$$\nabla f(x)^T d \leq \frac{-\|P_\Omega(x - t\nabla f(x)) - x\|^2}{t}.$$

Proof. Simply observe that

$$\begin{aligned}
\|P_\Omega(x - t\nabla f(x)) - x\|^2 &= \langle P_\Omega(x - t\nabla f(x)) - x, P_\Omega(x - t\nabla f(x)) - x\rangle \\
&= -t\nabla f(x)^T d + \langle P_\Omega(x - t\nabla f(x)) - (x - t\nabla f(x)), P_\Omega(x - t\nabla f(x)) - x\rangle \\
&\leq -t\nabla f(x)^T d
\end{aligned}$$

where the last inequality follows Theorem 7.1.1 equation rfeqgp2. ■

Based on these observations we have the following algorithm.

7.2 The Basic Gradient Projection Method

Initialization: $x \in \Omega$, $\gamma \in (0, 1)$, $c \in (0, 1)$

Having x_k obtain x_{k+1} as follows

1. Set $d_k := P_\Omega(x_k - \nabla f(x_k)) - x_k$

2. Set

$$\begin{aligned} \lambda_k &:= \max \gamma^s \\ &\quad \text{subject to } s \in \{0, 1, 2, \ldots\} \\ &\quad f(x_k) + \gamma^s d_k) - f(x_k) \leq c\gamma^s \nabla f(x_k)^T d_k. \end{aligned}$$

3. Set $x_{k+1} := x_k + \lambda_k d_k$.

We now apply Theorem 2.1.1 to yield a convergence theorem for this method.

THEOREM 7.2.1 *Let $f : \mathbb{R}^n \to R$ be C^1 and let $\Omega \subset \mathbb{R}^n$ be a nonempty closed convex set. Let $x_0 \in \Omega$ be such that f' is uniformly continuous on the set $\overline{co}\{x \in \Omega : f(x) \leq f(x_0)\}$. If $\{x_k\}$ is the sequence generated by gradient projection algorithm given above with starting point x_0, then one of the following must occur.*

1. *There is a k_0 such that $-\nabla f(x_{k_0}) \in N_\Omega(x_{k_0})$.*
2. *$f(x_k) \downarrow -\infty$.*
3. *The sequence $\{\|d_k\|\}$ diverges to $+\infty$,*
4. *For every subsequence $J \subset \mathbb{N}$ for which $\{d_k\}_J$ is bounded, we have that $d_k \underset{J}{\to} 0$, or equivalently*

$$\|P_\Omega(x_k - \nabla f(x_k)) - x_k\| \underset{J}{\to} 0.$$

COROLLARY 7.2.1.1 *Let the hypotheses of Theorem 7.2.1 hold. Furthermore assume that the sequence $\{d_k\}$ is bounded. Then every cluster point $\overline{x}$ of the sequence $\{x_k\}$ satisfies $-\nabla f(\overline{x}) \in N_\Omega(\overline{x})$.*

7.3 The Computation of Projections

We now address the question of implementation. Specifically, how does one compute the projection onto the convex set Ω. In general this is not a finite process. Nonetheless, for certain important convex sets Ω it can be done quite efficiently.

Projection onto box constraints

Let us suppose that Ω is given by $\Omega := \{x \in \mathbb{R}^n : \ell \leq x \leq u\}$, where $\ell, u \in \overline{\mathbb{R}}^n$ with $\overline{\mathbb{R}} = \Omega \cup \{+\infty, -\infty\}$ and $\ell :\leq u$, $i = 1, \ldots, n, \ell_i \neq +\infty$ $i = 1, \ldots, n$ and $u_i \neq -\infty$ $i = 1, \ldots, n$. Then P_Ω can be expressed componentwise as

$$[P_\Omega(x)]_i := \begin{cases} \ell_i & \text{if } x_i \leq \ell_i \\ x_i & \text{if } \ell_i < x_i < u_i \\ u_i & \text{if } u_i \leq x_i \end{cases}$$

Thus, for example, if $\Omega = \mathbb{R}^n_+$, then

$$P_\Omega(x) = x_+.$$

Projection onto a Polyhedron

Let Ω be the polyhedron given by

$$\Omega := \{x \in \mathbb{R}^n : a_i^T x \leq \alpha_i, i = 1, \ldots, 3, a_i^T x = \alpha_i, i = s+1, \ldots, m\}.$$

Then P_Ω is determined by solving the quadratic program

$$\begin{array}{lll} \min \frac{1}{2}\|x - y\|_2^2 & & \\ \text{subject to} & a_i^T x \leq \alpha_i & i = 1, \ldots, s \\ & a_i^T x = \alpha_i & i = s+1, \ldots, m. \end{array}$$

Chapter 8

Exterior Penalty Methods

8.1 Basic Theory

We now return to the constrained optimization problem

$$\begin{array}{lll} \mathcal{P}: & \min f_0(x) & \\ & \text{subject to} & f_i(x) \le 0 \quad i = 1, \ldots, q \\ & & f_i(x) = 0 \quad i = q+1, \ldots, m. \end{array}$$

Observe that the problem $\mathcal{P}$ is really composed of two problems. The first is the problem of feasibility, that is, we need to identify points $x \in \mathbb{R}^n$ such that

$$x \in \Omega := \{x : f_i(x) \le 0 \quad i = 1, \ldots, q, f_i(x) = 0 \quad i = q, \ldots, m\}.$$

This problem is quite difficult in its own right as is evidence by the effort devoted to its solution in the previous chapter. In particular, given the problem $\mathcal{P}$ one cannot be positive that Ω is non-empty. In $\mathcal{P}$ the feasibility problem is complicated by the secondary problem of trying to minimize f_0 over Ω. In all methods designed to solve $\mathcal{P}$ a balance must be struck between trying to attain feasibility and trying to minimize f_0.

In the methods of this section we replace $\mathcal{P}$ by an unconstrained optimization problem of the form

$$\min P_\alpha(x), \tag{8.1.1}$$

where

$$P_\alpha(x) := f_0(x) + \alpha\beta(x).$$

The function β appearing in the definition of P_α is called a penalty term, α the penalty parameter, and P_α an exterior penalty function. The role of the function β is to penalize non-inclusion in Ω, β must satisfy the following three conditions:

(i) $\beta : \mathbb{R}^n \to \mathbb{R}$ is continuous.
(ii) $\beta(x) \ge 0$ for all $x \in \mathbb{R}^n$.
(iii) $\beta(x) = 0$ if and only if $x \in \Omega$.

Several examples of functions satisfying (8.1) were examined in the previous chapters, for example

$$\begin{aligned}
\hat{\beta}_2(x) &= \frac{1}{2}[\sum_{i=1}^{q}(\max\{0, f_i(x)\})^2 + \sum_{i=q+1}^{m}(f_i(x)^2] \\
\beta_1(x) &= \sum_{i=1}^{q}\max\{0, f_i(x)\} + \sum_{i=q+1}^{m}|f_i(x)| \\
\beta_\infty(x) &= \max\{0; f_i(x), i = 1, \ldots, q; |f_i(x)| i = q+1, \ldots, m\} \\
\beta_2(x) &= (2\hat{\beta}_2(x))^{1/2} \ .
\end{aligned}$$

The function $\hat{\beta}_2$ has an advantage over β_i $i = 1, 2, \infty$ since it is differentiable on $\mathbb{R}$ whereas the others are not. However, as we will see, a price must be paid for this differentiability. Given a way to construct the penalty function P_α, consider the following algorithm for solving $\mathcal{P}$.

Exterior Penalty (E-P) Algorithm

Initialization: Let $\{\alpha_i\} < \mathbb{R}_+$ be such that $\alpha_i \uparrow \infty$ with $\alpha_i < \alpha_{i+1}$ for all $i = 1, 2, \ldots$.
For $k = 1, 2, \ldots$, let x_k solve $\min\{P_{\alpha_k}(x) : x \in \mathbb{R}^n\}$.

Although the above algorithm appears to be somewhat unwieldy we will show that it has several practical refinements. In order to visualize how the method behaves we consider a two dimensional example:

$$\begin{aligned}
&\min(x_1^2 - x_2) \\
&\text{subject to } x_1 + x_2 - 1 = 0, \ 0 \leq x \ .
\end{aligned}$$

The solution is clearly the point $(x_1, x_2) = (0, 1)$. Set

$$P_\alpha(x) = x_1^2 - x_2 + \frac{\alpha}{2}((x_1 + x_2 - 1)^2 + (-x_1)_+^2)$$

Then

(a) $\frac{\partial P_\alpha(x)}{\partial x_1} = 2x_1 + \alpha((x_1 + x_2 - 1) - (-x_1)_+)$

(b) $\frac{\partial P_\alpha(x)}{\partial x_2} = -1 + \alpha(x_1 + x_2 - 1)$.

Let us set these to zero and solve.

First assume that $x_1 \geq 0$, then from (b) $\alpha(x_1 + x_2 - 1) = 1$ so (a) implies $x_1 = -1/2$, a contradiction. Thus it must be the case that $x_1 < 0$. In this case we use $\alpha(x_1 + x_2 - 1) = 1$ in (a) to see that

$$x_1 = \frac{-1}{2 + \alpha}$$

Hence $x_2 = 1 + \frac{1}{\alpha} + \frac{1}{2+\alpha}$. Therefore, as $\alpha \to \infty$ $(x_1, x_2) \to (0, 1)$. We now establish the convergence of the method. We begin with the following lemma.

LEMMA 8.1.1 *Let $\{\alpha_k\}$ be as in the E-P Algorithm and assume that*

$$\arg\min\{P_{\alpha_k}(x)\} \neq \emptyset$$

for all $k = 1, 2, \ldots$ (for example if $\lim_{\|x\|\to\infty} f_0(x) = +\infty$). If $\{x_k\} \subset \mathbb{R}^n$ is generated by the E-P Algorithm , then

1. *$P_{\alpha_k}(x_k) \leq P_{\alpha_{k+1}}(x_{k+1})$,*
2. *$\beta(x_k) \geq \beta(x_{k+1})$ and*
3. *$f(x_k) \leq f(x_{k+1})$*

for all $k = 1, 2, \ldots$.

PROOF: Let us first observe that

$$P_{\alpha_k}(x_k) \leq P_{\alpha_k}(x_{k+1}) \leq P_{\alpha_{k+1}}(x_{k+1}),$$

which establishes (1). In order to see (2), we use (1) to write

$$\begin{aligned} f(x_k) + \alpha_k\beta(x_k) &\leq f(x_{k+1}) + \alpha_k\beta(x_{k+1}) \text{ and} \\ f(x_{k+1}) + \alpha_{k+1}\beta(x_{k+1}) &\leq f(x_k) + \alpha_{k+1}\beta(x_k). \end{aligned}$$

Adding these we get

$$(\alpha_{k+1} - \alpha_k)\beta(x_{k+1}) \leq (\alpha_{k+1} - \alpha_k)\beta(x_k).$$

Hence $\beta(x_{k+1}) \leq \beta(x_k)$.

To obtain (3), we use (1) and (2) to write

$$\begin{aligned} f(x_k) + \alpha_k\beta(x_k) &\leq f(x_{k+1}) + \alpha_k\beta(x_{k+1}) \\ &\leq f(x_{k+1}) + \alpha_k\beta(x_k). \end{aligned}$$

Hence $f(x_k) \leq f(x_{k+1})$. ■

THEOREM 8.1.1 *Let $\{\alpha_k\}$ be as in the E-P Algorithm and suppose that*

$$\arg\min\{P_{\alpha_k}(x)\} \neq \emptyset$$

for all k. Further assume that $\Omega \neq \emptyset$. Then every cluster point of $\{x_k\}$ is a solution to $\mathcal{P}$. In particular, if a cluster point exists, then a solution to $\mathcal{P}$ exists.

PROOF: Let $x \in \Omega$ and observe that for each $k = 1, 2, \ldots$,

$$f(x) = P_{\alpha_k}(x) \geq P_{\alpha_k}(x_k) \geq f(x_k). \tag{8.1.2}$$

Let x^* be a cluster point of $\{x_k\}$. Since $\{f(x_k)\}$ is an increasing sequence and $\{\beta(x_k)\}$ is a decreasing sequence, we know that

$$f(x_k) \uparrow f(x^*) \text{ and } \beta(x_k) \downarrow \beta(x^*).$$

Also, by (8.1.2) $\{P_{\alpha_k}(x_k)\}$ is an increasing sequence that is bounded above. Hence there is a P^* such that $P_{\alpha_k}(x_k) \uparrow P^*$. Consequently,

$$\begin{aligned} \lim \beta(x_k) &= \lim(P_\alpha(x_k) - f(x_k))\alpha_k^{-1} \\ &= 0, \end{aligned}$$

so that $\beta(x^*) = 0$ or $x^* \in \Omega$. Also, by (8.1.2), $f(x^*) \leq f(x)$ for all $x \in \Omega$. Hence x^* solves $\mathcal{P}$. ■

COROLLARY 8.1.1.1 *Let $\alpha_0 > 0$, $x_0 \in \arg\min P_{\alpha_0}$, and $\epsilon > 0$ be given. Choose $\delta \in [0, 1)$ so that*

$$(1 - \delta)\gamma > 1 \qquad \textit{and} \qquad \delta\epsilon < \beta(x_0).$$

Select $\bar{x} \in \{x : \beta(x) < \delta\epsilon\}$ and take

$$\alpha > \max\{\gamma\epsilon^{-1}|f(x_0) - f(\bar{x})|, \alpha_0\} \ .$$

Then, either x_0 or $\bar{x}$ solves $\mathcal{P}$, or

$$\beta(x_\alpha) \leq \epsilon \quad \textit{and} \quad f(x_\alpha) \leq \min_{x \in \Omega} f(x) \ ,$$

where $x_\alpha \in \arg\min P_\alpha$.

PROOF: Assume that neither x_0 or $\bar{x}$ solves $\mathcal{P}$ and that the result is false. Then $f(x_0) < f(\overline{x})$ and $\beta(x_\alpha) > \epsilon$. Hence

$$\begin{aligned} 0 &\geq P_\alpha(x_\alpha) - P_\alpha(\overline{x}) \\ &= f(x_\alpha) - f(\overline{x}) + \alpha\beta(x_\alpha) - \alpha\beta(\bar{x}) \\ &\geq f(x_0) - f(\bar{x}) + \alpha(1 - \delta)\epsilon \\ &\geq -|f(x_0) - f(\overline{x})| + (1 - \delta)\gamma|f(x_0) - f(\overline{x})| \\ &= ((1 - \delta)\gamma - 1)|f(x_0) - f(\overline{x})| > 0, \end{aligned}$$

which is a contradiction. ■

Corollary 8.1.1.1 implies that good approximate solutions can be obtained using the Exterior Penalty method without sending α to $+\infty$. In situations where the constraints have a "soft" character to them this is quite acceptable .

8.2 Exact Penalization

Clearly the most unpleasant feature of exterior penalty function methods as we have discussed them thus far is the requirement that the penalty parameters diverge to $+\infty$. For obvious reasons this requirement could instill serious numerical instabilities in any method proposed to solve the subproblems. In this section we will study a class of penalty terms $\beta(x)$ that do not necessarily require the divergence of the penalty parameters. Specifically, we will show that for certain choices of β there is a finite $\overline{\alpha} > 0$ such that if $\overline{x}$ is a local solution to $\mathcal{P}$, then $\overline{x}$ is also a local solution to P_α for all $\alpha \geq \overline{\alpha}$. Such a function P_α is called an exact penalty function. By the example of the previous section, it is clear that β_2 does not in general yield an exact penalty function. As an alternative let us apply $\hat{\beta}_2$ and β_1 to the problem

$$\min_{0 \leq x} x. \tag{8.2.3}$$

The solution to this problem is $x = 0$. If we minimize

$$P_\alpha(x) = x + \frac{\alpha}{2}[(-x)_+]^2$$

we get $x = -\alpha^{-1}$ as the solution, and as $\alpha \uparrow \infty$ $x \to 0$. Next, if we minimize

$$\begin{aligned} P_\alpha(x) &= x + \alpha(-x)_+ \\ &= \begin{cases} x & \text{if } x \geq 0 \\ (1-\alpha)x & \text{if } x < 0 \end{cases} \end{aligned}$$

we get no solution for $\alpha < 1$, infinitely many solutions ($\{x : x \leq 0\}$) for $\alpha = 1$, and the unique solution $x = 0$ for $\alpha > 1$. Therefore, (8.2.3) is an exact penalty function for this problem.

EXERCISE: Show that β_1 is also an exact penalty function for the example in the previous section. Choose $\alpha > 1$.

In the general case, determining whether β_1 is an exact penalty function or not is substantially more difficult due to the nondifferentiability of β_1. Nonetheless, it is possible to approach the problem in the general setting via convex composite functions. We do not do this here. Instead we will simply state the relevant results.

THEOREM 8.2.1 *Let $\overline{x} \in \Omega$, then $\overline{x}$ is a Kuhn-Tucker point for $\mathcal{P}$ if and only if $\overline{x}$ is a stationary point for the penalty function*

$$P_\alpha(x) := f_0(x) + \alpha\, dist(f(x)|K)$$

for all $\alpha \geq \overline{\alpha}$ for some $\overline{\alpha} > 0$, where

$$K := \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$$

and

$$dist(y|K) := \inf\{\|y - z\| : z \in K\}$$

where $\|\cdot\|$ is any given norm on $\mathbb{R}^n$. Moreover, the parameter $\overline{\alpha}$ can be chosen to equal $\|\overline{u}\|_0$ where $\overline{u}$ is any Lagrange multiplier vector at $\overline{x}$ and $\|\cdot\|_0$ is the norm dual to the norm employed in the definition of dist$(\cdot|K)$.

THEOREM 8.2.2 *Let $(\overline{x},\overline{u})$ be a Kuhn-Tucker point for $\mathcal{P}$ and suppose that*

$$d^T\nabla^2_{xx}L(\overline{x},\overline{u})d > 0$$

for every $d \in \mathbb{R}^n\backslash\{0\}$ such that

$$\begin{aligned}\nabla f_0(\overline{x})^T d &= 0\\ \nabla f_i(\overline{x})^T d &\leq 0 \quad i \in A(\overline{x})\\ \nabla f_i(\overline{x})^T d &= 0 \quad i \in \{s+1,\ldots,m\}.\end{aligned}$$

Then $\overline{x}$ is a such that there exist $\epsilon > 0$ and $\nu > 0$ for which

$$f_0(x) \geq f(\overline{x}) + \nu\|x-\overline{x}\|_2^2$$

for all $x \in \Omega \cap (\overline{x}+\epsilon\mathbb{B})$ and

$$P_\alpha(x) \geq P_\alpha(\overline{x}) + \nu\|x-\overline{x}\|_2^2$$

for all $x \in (\overline{x}+\epsilon\mathbb{B})$ for all $\alpha > \|\overline{u}\|_0$ where

$$P_\alpha(x) := f_0(x) + \alpha dist(f(x)|K)$$

and $\|\cdot\|_0$ is the norm dual to that used in the definition of dist$(\cdot|K)$.

Theorems 8.2.1 and 8.2.2 indicate that any exterior penalty function of the form

$$P_\alpha(x) := f_0(x) + \alpha\text{dist}(f(x)|K)$$

is an exact penalty function for $\mathcal{P}$. Unfortunately such functions are not differentiable. Thus locating points at which they attain they're global minimum value may be a difficult if not impossible task. The situation is of course complicated by the need to compute an estimate for an appropriate value of the penalty parameter α. Nonetheless, even in the face of such difficulties these methods can be quite successful. We now show how this can be done in the special cases where the norm is chosen to be either the ℓ_∞ or ℓ_1 norm.

Observe that for the ℓ_1 norm

$$\text{dist}(y|K) = \sum_{i=1}^{s}(y_i)_+ + \sum_{i=s+1}^{m}|y_i|,$$

while for the ℓ_∞ norm

$$\text{dist}(y|K) = \max\{0; y_i,\ i=1,\ldots,s; |y_i| i = s+1,\ldots,m\}.$$

Thus the penalty functions associated with these norms are easily computed. Let us now consider an algorithm for minimizing

$$P_\alpha(x) := f_0(x) + \alpha[\sum_{i=1}^{m} f_i(x)_+ + \sum_{i=s+1}^{m}|f_i(x)|].$$

The $S\ell_1QP$ Algorithm

Initialization: Let $x_0 \in \mathbb{R}^n$, $H_0 \in \mathbb{R}^{n\times n}$ with H_0 symmetric and positive definite, $\delta > 0$, $\gamma \in (0,1)$, $c \in (0,1)$, and let the norm on $\mathbb{R}^m$ be the ℓ_1 norm.

Having (x_i, H_i) obtain (x_{i+1}, H_{i+1}) as follows:

1. Let d_i be the solution to

$$\min_{\|d\|\le\delta} \nabla f_0(x_i)^T d + \frac{1}{2} d_i^T H_i d_i + \alpha \text{dist}(f(x_i) + f'(x_i)d|K)$$

   where $K := \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$.

2. Stop if $\Delta(x_i, d_i) = 0$; otherwise set

$$\begin{aligned} \lambda_i : \quad &= \max \gamma^s \\ &\text{subject to } s \in \{0,1,2,\ldots\}, \text{ and} \\ &P_\alpha(x_i + \gamma^s d_i) - P_\alpha(x_i) \le c\gamma^s \Delta(x_i, d_i), \end{aligned}$$

   where

$$\Delta(x;d) := \nabla f_0(x)^T d + \alpha[\text{dist}(f(x) + f'(x)d|K) - \text{dist}(f(x)|K)].$$

3. Update $x_{i+1} := x_i + \lambda_i d_i$, $H_{i+1} \in \mathbb{R}^{n\times n}$ symmetric positive definite.

Remark: The use of the 1-norm is not crucial to the algorithm.

The convergence theory for the above procedure rests on the following two facts about the function $\Delta(x,d)$.

THEOREM 8.2.3 *Let $x \in \mathbb{R}^n$, $\delta > 0$, and $H \in \mathbb{R}^{n\times n}$ be symmetric positive definite.*

1. *For all $d \in \mathbb{R}^n$*

$$P'_\alpha(x;d) \le \Delta(x,d).$$

2. *If $\overline{d}$ solves*

$$\min_{\|d\|\le\delta} \nabla f_0(x)^T d + \frac{1}{2} d^T H d + \alpha \, dist(f(x) + f'(x)d|K),$$

   then $\Delta(x,\overline{d}) \le 0$ with equality if and only if

$$P'_\alpha(x;d) \ge 0$$

   for all $d \in \mathbb{R}^n$.

Theorem 8.2.3 in conjunction with Theorem 2.1.1 yield the following convergence result.

THEOREM 8.2.4 *Let $f_i : \mathbb{R}^n \to \mathbb{R}$ $i = 0, 1, \ldots, m$ be C' with ∇f_i uniformly continuous on $\{x : P_\alpha(x) \leq P_\alpha(x_0)\}$ for some $x_0 \in \mathbb{R}^n$, where $P_\alpha(x) := f_0(x) + \alpha dist(f(x)|K)$. If $\{x_k\}$ is the sequence generate by the $S\ell_1QP$ Algorithm with initial point x_0, then one of the following must occur:*

1. *There is an i_0 such that*
$$\Delta(x_{i_0}, d_{i_0}) = 0.$$
2. $P_\alpha(x_i) \downarrow -\infty$.
3. $\lim_{i\to\infty} \Delta(x_i, d_i) = 0$.

Thus, every cluster point of the sequence is a stationary point for P_α. If this stationary point is feasible for $\mathcal{P}$, then it is a Kuhn-Tucker point for $\mathcal{P}$.

We now address the question of how one can solve direction finding subproblems of the form

$$\begin{array}{l} \min \nabla f_0(x)^T d + \frac{1}{2} d^T H d\alpha + \text{dist}(f(x) + f'(x)^T d|K) \\ \text{subject to } \|d\| \leq \delta. \end{array} \tag{8.2.4}$$

In this regard one can simply employ the same tricks described in the final section of Chapter 7. For example, if we use the ℓ_∞-norm on $\mathbb{R}^m$, then (8.2.4) can be written as the quadratic program

$$\begin{array}{lll} \mathcal{QP}_1 & \min_{(z,d)} & \nabla f_0(x)^T d + \alpha e^T z + \frac{1}{2} d^T H d \\ & \text{subject to} & f_i(x) + f_i'(x)^T d \leq z_i \quad i = 1, \ldots, s \\ & & 0 \leq z_i \quad i = 1, \ldots, s \\ & & -z_i \leq f_i(x) + \nabla f(x)^T d \leq z_i \quad i = s+1, \ldots, m \\ & & -\delta e \leq d \leq \delta e \end{array}$$

Similarly, if both $\mathbb{R}^n$ and $\mathbb{R}^m$ are equipped with the ℓ_∞-norm, then (8.2.4) becomes the quadratic program

$$\begin{array}{lll} \mathcal{QP}_\infty & \min_{(\gamma,d)} & \nabla f_0(x)^T d + \alpha\gamma + \frac{1}{2} d^T H d \\ & \text{subject to} & f_i(x) + \nabla f_i(x)^T d \leq \gamma \quad i = 1, \ldots, s \\ & & -\gamma \leq f_i(x) + \nabla f_i(x)^T d \leq \gamma \quad i = s+1, \ldots, m \\ & & -\delta e \leq d \leq \delta e \quad . \end{array}$$

However, it should be noted that the penalty functions based on the $1-$ or $2-$norms are vastly superior in practise to one based on the $\infty-$norm. The reason is that the $\infty-$norm penalty function only works on the most violated constraints at a particular point. Whereas the $1-$ and $2-$norm penalty functions work on all of the constraints simultaneously.

Finally how does one update the matrices H_i? For reasons that will be made clear in subsequent sections, one should construct H_i so that it approximates the hessian of the Lagrangian at x_i. In order to do this, we first need approximations to the Kuhn–Tucker multipliers. There are two standard methods for obtaining these multiplier estimates. In the

first method the estimates come directly from the subproblem used to compute the search direction d_i. For example, if the Sl_1QP algorithm is used, then the subproblems are of the form $\mathcal{QP}_1$. One then simply uses the Kuhn–Tucker multipliers for the constraints in this subproblem as estimates of the multipliers for $\mathcal{P}$. In the second approach, one computes least squares estimates for the Kuhn–Tucker multipliers. That is, let the multiplier estimates be a solution to the unconstrained subproblem

$$\min_{u\in\mathbb{R}^m} \frac{1}{2}\|\nabla f_0(x_i) + \sum_{j=1}^{m} u_j \nabla f_j(x_i)\|_2^2$$

or, perhaps, the constrained subproblem

$$\min_{\substack{u\in\mathbb{R}^m \\ u_j\geq 0,\ j=1,2,\cdots,s}} \frac{1}{2}\|\nabla f_0(x_i) + \sum_{j=1}^{m} u_j \nabla f_j(x_i)\|_2^2 \ .$$

Once the multiplier estimates are obtained, we update the H_i's using the BFGS formula with $s_i := x_{i+1} - x_i$ and

$$y_i = \nabla_x L(x_{i+1}, u_{i+1}) - \nabla_x L_\alpha(x_i, u_i).$$

Chapter 9

The Method of Multipliers

9.1 Introduction

In our study of exterior penalty functions in the previous section we found that there was a compromise between differentiability and exactness. That is given a penalty function

$$P_\alpha(x) = f_0(x) + \alpha\beta(x)$$

either the penalty term is differentiable in which case the penalty parameter α must tend to $+\infty$ or $\beta(x)$ is nondifferentiable in which case α need not tend to $+\infty$. In this section we consider a modification to the quadratic differentiable penalty term $\hat{\beta}_2$ which avoids the need to send α to $+\infty$.

Recall that in each step of the exterior penalty method applied to

$$P_\alpha(x) := f_0(x) + \alpha\left[\sum_{i=1}^{s}(f_i(x)_+)^2 + \sum_{i=s+1}^{m}(f_i(x))^2\right]$$

one solves the unconstrained minimization problem

$$\min_{x\in\mathbb{R}^n} P_\alpha(x).$$

The solution x_α satisfies

$$0 = \nabla P_\alpha(x_\alpha) = \nabla f_0(x_\alpha) + \sum_{i=1}^{s}\nabla f_i(x_\alpha)(\alpha f_i(x_\alpha)_+) + \sum_{i=s+1}^{m}\nabla f(x_\alpha)(\alpha f_i(x)). \tag{9.1.1}$$

Setting

$$(u_\alpha)_i := \begin{cases} \alpha f_i(x_\alpha)_+ & \text{if } i = 1,\dots,s \\ \alpha f_i(x_\alpha) & \text{if } i = s+1,\dots,m \end{cases}$$

we can write (9.1.1) as

$$0 = \nabla_x L(x_\alpha, u_\alpha)$$

where $L(x,u) = f_0(x) + u^T f(x)$ is the Lagrangian for the problem $\mathcal{P}$.

$$\begin{array}{lll} \min & f_0(x) & \\ \text{subject to} & f_0(x) \leq 0 & i = 1, \ldots, s \\ & f_0(x) = 0 & i = s+1, \ldots, m \end{array}$$

Consequently, if $x_\alpha \to \overline{x}$ a local solution to $\mathcal{P}$, then every cluster point of $\{u_\alpha\}$ is a Kuhn-Tucker multiplier for $\overline{x}$. This indicates that we should think of the vectors u_α as multiplier approximates that are to be updated at each iteration. By doing so, we avoid the need to send the penalty parameter α to $+\infty$. The strategy is as follows: given $\alpha > 0$ and an estimate of the multipliers $u \in \mathbb{R}^m$, let $x_{\alpha,u}$ be the solution to the equation

$$\begin{array}{rcll} 0 & = & \nabla f_0(x) & + \sum_{i=1}^{s} \nabla f_i(x)(\alpha f_i(x) + u_i)_+ \\ & & & + \sum_{i=s+1}^{m} \nabla f_i(x)(\alpha f_i(x) + u_i) \ . \end{array}$$

Then update the multiplier estimates u_i via the equations

$$u_i = (\alpha f_i(x_{\alpha,u}) + u_i)_+ \quad \text{for } i = 1, \ldots, s$$

and

$$u_i = (\alpha f_i(x_{\alpha,u}) + u_i) \quad \text{for } i = s+1, \ldots, m.$$

This procedure describes the basic structure of an algorithm known as *the method of multipliers*. Before we provide a precise description of this algorithm let us first examine expression (9.1) more carefully.

9.2 The Augmented Lagrangian

Observe that expression (9.1) is a first order optimality condition for some function. In order to recover this function we can integrate the right hand side of (9.1) in the variable x. By adding in the appropriate constant term this integration yields the function

$$\begin{array}{rcll} L(x,u,\alpha) & := & f_0(x) & + \frac{1}{2\alpha}[\text{dist}_2^2[\alpha f(x) + u | K] - \|u\|_2^2] \\ & := & f_0(x) & + \frac{1}{2\alpha} \sum_{i=1}^{s} ((\alpha f_i(x) + u_i)_+)^2 - u_i^2 \\ & & & + \frac{\alpha}{2} \sum_{i=s+1}^{m} f_i(x)(f_i(x) + u_i). \end{array}$$

where $K := \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$. The function $L(x,u,\alpha)$ is called the augmented Lagrangian for $\mathcal{P}$. The name is derived from the fact that if $s = 0$, that is there are only equality constraints, then $L(x,u,\alpha)$ takes the form

$$L(x,u,\alpha) = L(x,u) + \frac{\alpha}{2}\|f(x)\|_2^2$$

where $L(x,u) = f_0(x) + u^T f(x)$ is the usual Lagrangian. Thus $L(x,u,\alpha)$ can be thought of as arising from the usual Lagrangian after one has incorporated a vehicle for penalizing constraint violation. The augmented Lagrangian possesses the following remarkable property.

THEOREM 9.2.1 *Let $\alpha > 0$, f_i, $i = 0, \ldots, m$ differentiable at $x \in \mathbb{R}^n$. Then*

$$0 = \nabla_{x,u} L(x, u, \alpha)$$

if and only if (x, u) is a Kuhn-Tucker pair for $\mathcal{P}$.

PROOF: Note that $0 = \nabla_{x,u} L(x, u, \alpha)$ if and only if

$$\begin{aligned} 0 &= \nabla f_0(x) + \sum_{i=1}^{s} (\alpha f_i(x) + u_i)_+ \nabla f_i(x) + \sum_{i=s+1}^{m} (\alpha f_i(x) + u_i) \nabla f_i(x) \\ u_i &= (\alpha f_i(x) + u_i)_+ \quad i = 1, \ldots, s \\ 0 &= f_i(x) \quad i = s+1, \ldots, m \ . \end{aligned}$$

Hence the result will be established once we have shown that

$$[(a-b)_+ - a = 0 \Longleftrightarrow a \geq 0, b \geq 0, ab = 0].$$

Case 1: $a - b \geq 0$

If $a - b \geq 0$, then $(a-b)_+ = a - b$ so that $b = 0$. Consequently, $a \geq 0$, $b \geq 0$, $ab = 0$.

Case 2: $a - b \leq 0$

If $(a - b) \leq 0$, then $(a-b)_+ = 0$ so that $a = 0$. Consequently, $a \geq 0$, $b \geq 0$ and $ab = 0$. The converse is trivial. ■

Thus it would seem that we need only find the roots of the equation $0 = \nabla_{x,u} L(x, u, \alpha)$ in order to locate Kuhn-Tucker points for the problem $\mathcal{P}$. This is precisely what the method of multipliers attempts to do.

In order to investigate the rate of convergence for these methods we require the nonsingularity of the hessian

$$\nabla^2_{x,u} L(x, u, \alpha).$$

Unfortunately, $\nabla^2_{x,u} L(x, u, \alpha)$ does not always exists since $(\alpha f_i(x) + u_i)_+$ is not everywhere differentiable. A sufficient condition under which $\nabla^2_{x,u} L(x, u, \alpha)$ does exist near a Kuhn-Tucker point $(\overline{x}, \overline{u})$ for $\mathcal{P}$ is *strict complementary slackness.*

DEFINITION 9.2.1 *Let $(\overline{x}, \overline{u})$ be a Kuhn-Tucker pair for $\mathcal{P}$. We say that the strict complementary slackness condition (SCSC) is satisfied at $(\overline{x}, \overline{u})$ if $\overline{u}_i > 0$ whenever $f_i(\overline{x}) < 0$ $i = 1, \ldots, s$.*

Observe that if the SCSC is satisfied at the K-T pair $(\overline{x}, \overline{u})$ then

$$(\alpha f_i(x) + u_i)_+ = 0$$

for all (x, u) near $(\overline{x}, \overline{u})$ for each $i \notin A(\overline{x}) = \{i : f_i(\overline{x}) = 0\}$, and

$$(\alpha f_i(x) + u_i)_+ = (\alpha f_i(x) + u_i)$$

for all (x,u) near $(\overline{x},\overline{u})$ for each $i \in A(\overline{x})$. Consequently, $\nabla^2_{x,u}L(x,u,\alpha)$ exists near $(\overline{x},\overline{u})$ and is given by

$$\nabla^2_{x,u}L(x,u,\alpha) = \begin{bmatrix} \nabla_{x,x}L(x,u,\alpha) & f_E'(x)^\tau & f_A'(x) & 0 \\ f_E'(x) & 0 & 0 & 0 \\ f_A'(x) & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{\alpha}I_N \end{bmatrix}.$$

Here we have reordered the components of the vector u into the multipliers associated with the equality constraints u_E with $E = \{s+1,\ldots,m\}$, the multipliers associated with the active inequality constraints u_A with $A = A(\overline{x})$, and the multipliers associated with the inactive inequality constraints u_N with $N = \{1,\ldots,s\}\backslash A(x)$. Also for any matrix $M \subset \mathbb{R}^{m\times n}$ and index set $J \subset \{1,\ldots,m\}$, M_J represents that matrix whose rows are those of M with index in J. Finally,

$$\begin{aligned} \nabla_{xx}L(x,u,\alpha) &= \nabla^2 f_0(x) + \sum_{i\in A} \alpha\nabla f_i(x)\nabla f_i(x)^T + (\alpha f_i(x)+u_i)\nabla^2 f_i(x) \\ &\quad + \sum_{i=s+1}^{m} \alpha\nabla f_i(x)\nabla f_i(x)^T + (\alpha f_i(x)+u_i)\nabla^2 f_i(x) \\ &= \nabla^2_{xx}L(x,u) + \alpha \sum_{i\in A\cup\{s+1,\ldots,m\}} \nabla f_i(x)\nabla f_i(x)^T + f_i(x)\nabla^2 f_i(x) \end{aligned}$$

In order to establish the nonsingularity of $\nabla^2 L(x,u,\alpha)$ we need the following three facts from linear algebra whose proof are left as an exercise.

LEMMA 9.2.1 *Let $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times n}$ and $D \in \mathbb{R}^{n\times n}$. If*

1. *the rows of D are linearly independent,*
2. *$Dx = 0$, $x \neq 0 \Longrightarrow x^TBx > 0$, and*
3. *$\mu \geq 0$,*

then the matrix $\begin{bmatrix} B+\mu A^TA & D^T \\ D & 0 \end{bmatrix}$ is nonsingular

THEOREM 9.2.2 *[Finsler's Theorem] Let $B,C \in \mathbb{R}^{n\times n}$ with C positive semi-definite. Then $x^TBx > 0$ for every $x \in \mathbb{R}^n$, $x \neq 0$ such that $x^TCx = 0$ if and only if $B+\mu C$ is positive definite for all $\mu \geq \overline{\mu}$ for some $\overline{\mu}$.*

THEOREM 9.2.3 *[Debreu's Theorem] Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times n}$. Then $x^TBx > 0$ for every $x \in \mathbb{R}^n$ with $x \neq 0$ such that $Ax = 0$ if and only if $B+\mu A^TA$ is positive definite for all $\mu \geq \overline{\mu}$ for some $\overline{\mu}$.*

The following result is an easy consequence of these linear algebraic results.

THEOREM 9.2.4 *(The positive definiteness of $\nabla^2_{xx}L(\overline{x},\overline{u},\alpha)$ and the nonsingularity of $\nabla^2 L(\overline{x},\overline{u},\alpha)$)*

1. *Let $(\overline{x},\overline{u})$ be a Kuhn-Tucker point for $\mathcal{P}$ and suppose that*

   (a) (Strict Complementary Slackness)

   $$\overline{u}_i > 0 \text{ whenever } f_i(x) < 0 \quad i = 1,\ldots,s.$$

   and

   (b) (Second–Order Sufficiency)

   $$\nabla f_i(\overline{x})^T d = 0 \quad i \in A(\overline{x}) \cup \{s+1,\ldots,m\} \Longrightarrow d^T \nabla_{11} L(\overline{x},\overline{u}) d > 0.$$

   Then $\nabla_{xx}L(\overline{x},\overline{u},\alpha)$ is positive definite for all $\alpha \geq \overline{\alpha}$ for some $\overline{\alpha} > 0$.

2. *If in addition to the hypotheses in (1) we assume that*

   (c) (The LI Condition) the gradients $\{\nabla f_i(\overline{x}) : i \in A(\overline{x}) \cup \{s+1,\ldots,m\}\}$ are linearly independent ,

   Then $\nabla^2 L(\overline{x},\overline{u},\alpha)$ is non-singular for all $\alpha > 0$.

PROOF: (i) We have that

$$\nabla^2_{xx}L(\overline{x},\overline{u},\alpha) = \nabla^2_{xx}L(\overline{x},\overline{u}) + \alpha f'_I(\overline{x}) f'_I(\overline{x})^T$$

where $I = A(\overline{x}) \cup \{s+1,\ldots,m)$. Consequently, the result follows from Debreu's Theorem 9.2.3.

(ii) This just follows from Lemma 9.2.1. ■

We now formally state the method of multipliers.

9.2.1 Algorithm: The Method of Multipliers

Let $(x^0,u^0,\alpha^0) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_+$ stop if $|\nabla L(x^i,u^i,\alpha^i)| < \epsilon$. Having (x^i,u^i,α^i) determine $(x^{i+1},u^{i+1},\alpha^{i+1})$ as follows

1. Let x^{i+1} solve

   $$L(x^{i+1},u^i,\alpha_i) = \min_{x\in\mathbb{R}^n} L(x,u^i,\alpha^i)$$

   or

   $$\nabla_x L(x^{i+1},u^i,\alpha^i) = 0.$$

2. Set $u^{i+1} = u^i + \alpha_i \nabla_u L(x^{i+1},u^i,\alpha^i)$ or equivalently

   $$u^{i+1}_j = (\alpha^i f_j(x^{i+1}) + u^i)_+ \quad \text{for } j = 1,\ldots,s$$

   and

   $$u^{i+1}_j = (\alpha^i f_j(x^{i+1}) + u^i) \quad \text{for } j = s+1,\ldots,m.$$

3. Set

$$\alpha^{i+1} := \begin{cases} \alpha^i & \text{if } \|u^{i+1} - u^i\|_\infty \leq \frac{1}{4}\|u^i - u^{i-1}\|_\infty \\ 10\alpha^i & \text{else} \end{cases}$$

In the following theorem we provide a sample of the type of convergence result that can be obtained for this method.

THEOREM 9.2.5 *Let the assumptions (a), (b), and (c) of Theorem 9.2.4 hold and let* $\alpha \geq \overline{\alpha}$. *Let* f_0 *and* f *be* C^2 *near the Kuhn-Tucker point* $\overline{x}$. *Then for* α *sufficiently large, but finite, there is an open neighborhood* V_α *of* $\overline{u}$ *such that for* $u^0 \in V_\alpha$ *there is an* x^0 *such that* $\nabla_x L(x^0, u^0, \alpha) = 0$ *and the iterates* (x^i, u^i) *generated by algorithm 9.2.1 exist and converge to* $(\overline{x}, \overline{u})$ *at the linear root rate*

$$\|x^i - \overline{x}, u^i - \overline{u}\| \leq \delta(\epsilon/\alpha)^i$$

for some positive constants ϵ *and* δ.