Lagrange Multiplier Algorithms

Instructor: Jin Zhang

Department of Mathematics
Southern University of Science and Technology

Spring 2023

Contents

- Barrier and Interior Point
- Penalty and Augmented Lagrangian Methods
- Exact Penalties-Sequential Quadratic Programming
- Lagrangian and Primal-Dual Interior Point Methods

Barrier and Interior Point

Barrier methods apply to inequality constrained problems of the form

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X, \quad g_j(x) \leq 0, \quad j = 1, \ldots, r \end{array}$$

where f and g_j are continuous real-valued functions, and X is a closed set. The interior (relative to X) of the set defined by the inequality constraints is

$$S = \{x \in X | g_j(x) < 0, j = 1, \ldots, r\}$$

We assume that S is nonempty and that any feasible point that is not in S can be approached arbitrarily closely by a vector from S; that is, given any feasible x and any $\delta > 0$, there exists $\tilde{x} \in S$ such that $\|\tilde{x} - x\| \leq \delta$.

In barrier methods, we add to the barrier function, is continuous and goes to ∞ as any one of the constraints $g_j(x)$ approaches 0 from negative values. The two most common examples of barrier functions are:

$$B(x) = -\sum_{j=1}^{r} \ln\{-g_j(x)\}, \quad \text{logarithmic;} \quad B(x) = -\sum_{j=1}^{r} \frac{1}{g_j(x)}, \quad \text{inverse.}$$

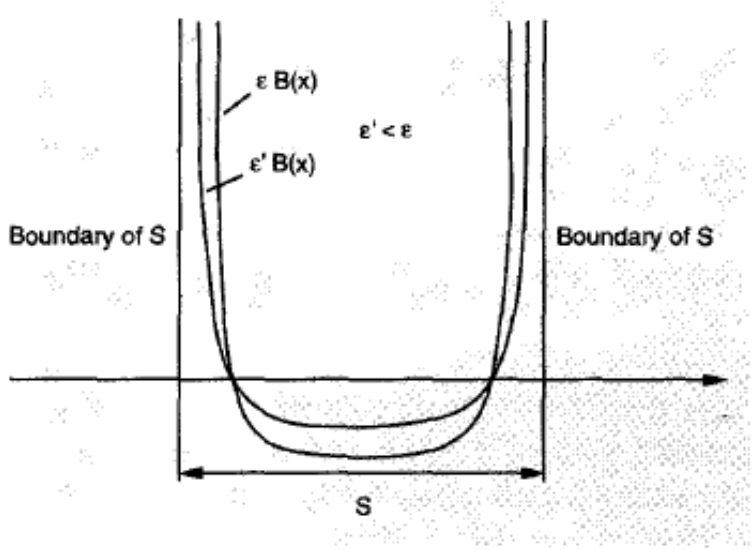

Figure 4.1.1. Form of a barrier function.

The barrier method is defined by introducing a parameter sequence $\{\epsilon^k\}$ with

$$0 < \epsilon^{k+1} < \epsilon^k, \quad k = 0, 1, \ldots, \quad \epsilon^k \to 0$$

It consists of finding

$$x^k = \arg\min_{x \in S} \{f(x) + \epsilon^k B(x)\}, \quad k = 0, 1, \ldots$$

Proposition

Every limit point of a sequence $\{x^k\}$ generated by a barrier method is a global minimum of the original constrained problem.

Proof $\{\overline{x}\}$ be the limit of a subsequence $\{x^k\}_{k\in K}$. If $\overline{x} \in S$, we have $\lim_{k\to\infty, k\in K} \epsilon^k B(x^k) = 0$, while if $\overline{x}$ lies on the boundary of S, we have by assumption $\lim_{k\to\infty, k\in K} B(x^k) = \infty$. In either case we obtain

$$\liminf_{k\to\infty} \epsilon^k B(x^k) \geq 0.$$

which implies that.

$$\liminf_{k\to\infty, k\in K} \left\{f\left(x^k\right) + \epsilon^k B\left(x^k\right)\right\} = f(\overline{x}) + \liminf_{k\to\infty, k\in K} \left\{\epsilon^k B\left(x^k\right)\right\} \geq f(\overline{x}).$$

The vector $\overline{x}$ is a feasible point of the original problem, since $x^k \in S$ and

X is a closed set. If $\overline{x}$ were not a global minimum, there would exist a feasible vector x^* such that $f(x^*) < f(\overline{x})$ and therefore also (using our assumption that x^* can be approached arbitrarily closely through the interior set S an interior point $\tilde{x} \in S$ such that $f(\tilde{x}) < f(\overline{x})$. We now have by the definition of x^k,

$$f\left(x^k\right) + \epsilon^k B\left(x^k\right) \leq f(\tilde{x}) + \epsilon^k B(\tilde{x}), \quad k = 0, 1, \ldots$$

which by taking the limit as $k \to \infty$ and $k \in K$, that $f(\overline{x}) \leq f(\tilde{x})$. This is a contradiction, thereby proving that $\overline{x}$ is a global minimum of the original problem.

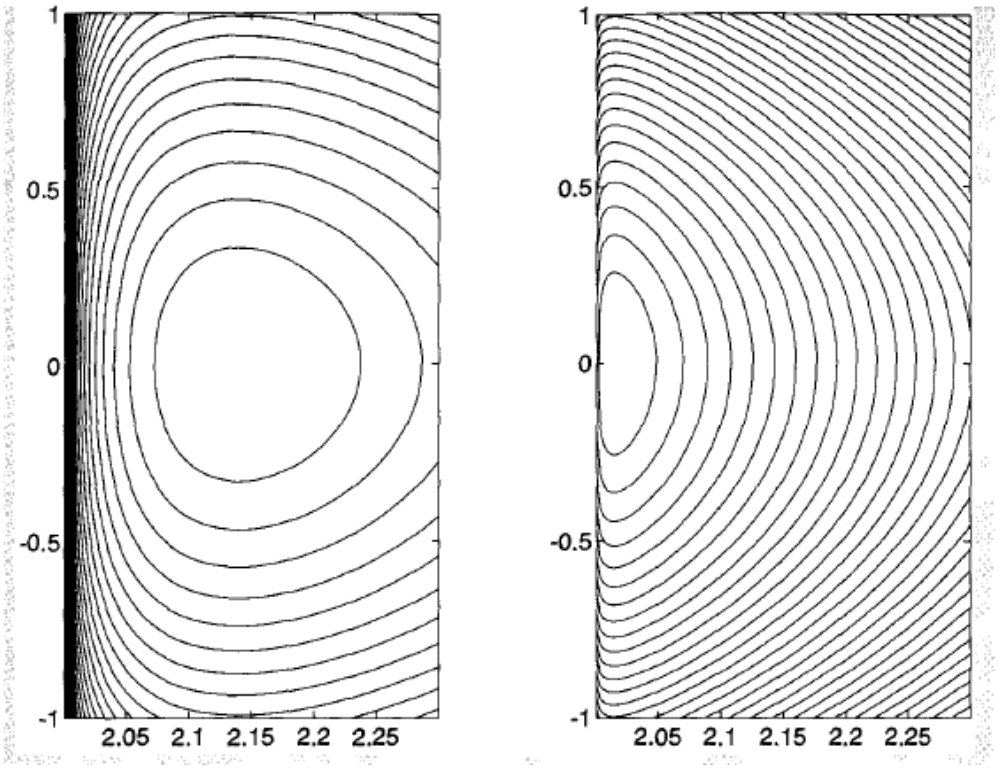

Figure. The convergence process of the barrier method for the two-dimensional problem

$$\text{minimize } f(x) = \frac{1}{2}\left(x_1^2 + x_2^2\right)$$
$$\text{subject to } 2 \leq x_1$$

with optimal solution $x^* = (2, 0)$. For the case of the logarithmic barrier function $B(x) = -\ln(x_1 - 2)$, we have

$$x^k = \arg\min_{x_1 > 2} \left\{ \frac{1}{2}\left(x_1^2 + x_2^2\right) - \epsilon^k \ln\left(x_1 - 2\right) \right\} = \left(1 + \sqrt{1 + \epsilon^k}, 0\right)$$

so as ϵ^k is decreased, the unconstrained minimum x^k approaches the constrained minimum $x^* = (2, 0)$. The figure shows the equal cost surfaces of $f(x) + \epsilon B(x)$ for $\epsilon = 0.3$ left side and $\epsilon = 0.03$ (right side).

Linear Programming and the Logarithmic Barrier

Let us apply the logarithmic barrier method to the linear programming problem

$$\begin{array}{ll} \text{minimize} & c'x \\ \text{subject to} & Ax = b, \quad x \geq 0 \end{array} \qquad \text{(LP)}$$

where $c \in \Re^n$ and $b \in \Re^m$ are given vectors, and A is an $m \times n$ matrix of rank m. We assume that the problem has at least one optimal solution. From the theory, we have that the dual problem, given by

$$\begin{array}{ll} \text{maximize} & b'\lambda \\ \text{subject to} & A'\lambda \leq c \end{array} \qquad \text{(DP)}$$

also has an optimal solution. Furthermore, the optimal values of the primal and the dual problem are equal.

The method involves finding for various $\epsilon > 0$,

$$x(\epsilon) = \arg\min_{x \in S} F_\epsilon(x)$$

where

$$F_\epsilon(x) = c'x - \epsilon \sum_{i=1}^{n} \ln x_i$$

and S is the interior set

$$S = \{x | Ax = b, x > 0\}$$

where $x > 0$ means that all the coordinates of x are strictly positive.

The Central Path

For given A, b, and c, as ϵ is reduced towards 0, $x(\epsilon)$ follows a trajectory that is known as the central path. The figure illustrates the central path for various values of the cost vector c.

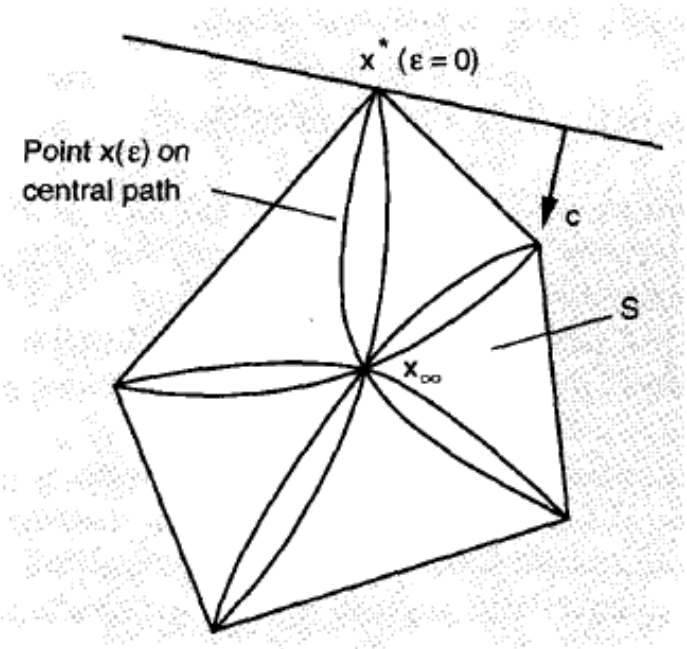

Figure. Central paths corresponding to ten different values of the cost vector c. All central paths start at the same vector, the analytic center x_∞, which corresponds to $\epsilon = \infty$, $x_\infty = \arg\min_{x \in S}\{-\sum_{i=1}^{n} \ln x_i\}$ and end at optimal solutions of (LP).

Note the following:

(a) For fixed A and b, the central paths corresponding to different cost vectors c start at the same vector x_∞. This is the unique minimizing point over S of $-\sum_{i=1}^{n} \ln x_i$ corresponding to $\epsilon = \infty$, and is known as the analytic center of S.

(b) If c is such that (LP) has a unique optimal solution x^*, the central path ends at x^* [that is, $\lim_{\epsilon \to 0} x(\epsilon) = x^*$]. For every sequence $\{\epsilon^k\}$ with $\epsilon^k \to 0$, the corresponding sequence $\{x(\epsilon^k)\}$ converges to x^*.

(c) If c is such that (LP) has multiple optimal solutions, it can be shown that the central path ends at one of the optimal solutions. We will not prove this fact. For a heuristic explanation note that from the definition of $x(\epsilon)$ we have$x(\epsilon) = \arg\min_{x \in S_\epsilon} \{-\sum_{i=1}^{n} \ln x_i\}$ where S_ϵ is the "slice" of the interior set S of points with the same cost value as $x(\epsilon)$, that is, $S_\epsilon = \{x | Ax = b, c'x = c'x(\epsilon), x > 0\}$. Thus $x(\epsilon)$ is the analytic center of S_ϵ.

Following Approximately the Central Path

Motivation: The most straightforward way to implement the logarithmic barrier method is to use some iterative algorithm to minimize the function F_{ϵ_k} for some decreasing sequence $\{\epsilon^k\}$ with $\epsilon^k \to 0$. This is equivalent to finding a sequence $\{x(\epsilon^k)\}$ of points on the central path. However, this approach is inefficient because it requires an infinite number of iterations to compute each point $x(\epsilon^k)$.

For ϵ and a given $x \in S$, replaced x by

$$\tilde{x} = x + \alpha(\overline{x} - x)$$

where $\overline{x}$ is the pure Newton iterate defined as the optimal solution of the quadratic program in the vector $z \in \Re^n$

$$\begin{array}{l} \text{minimize } \nabla F_\epsilon(x)'(z-x) + \frac{1}{2}(z-x)'\nabla^2 F_\epsilon(x)(z-x) \\ \text{subject to } Az = b \end{array}$$

and α is a stepsize selected by some rule. We have

$$\nabla F_\epsilon(x) = c - \epsilon x^{-1}, \quad \nabla^2 F_\epsilon(x) = \epsilon X^{-2}$$

where x^{-1} denotes the vector with coordinates $(x_i)^{-1}$ and X denotes the diagonal matrix with the coordinates $x_i$along the diagonal:

$$X = \begin{pmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & x_n \end{pmatrix}$$

We have that the pure Newton iterate is

$$\overline{x} = x - \epsilon^{-1} X^2 \left(c - \epsilon x^{-1} - A'\lambda \right),$$

where

$$\lambda = \left(AX^2 A' \right)^{-1} AX^2 \left(c - \epsilon x^{-1} \right).$$

These formulas can also be written as

$$\overline{x} = x - Xq(x, \epsilon),$$

where

$$q(x, \epsilon) = \frac{Xz}{\epsilon} - e$$

with e and z being the vectors

$$e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad z = c - A'\lambda,$$

and

$$\lambda = \left(AX^2A'\right)^{-1} AX(Xc - \epsilon e).$$

We can consider $\|q(x,\epsilon)\| = \left\|X^{-1}(x-\bar{x})\right\|$ as a measure of proximity of the current point x to $x(\epsilon)$.
The key result to be shown shortly is that for convergence of the logarithmic barrier method, it is sufficient to stop the minimization of F_{ϵ^k} and decrease ϵ^k to ϵ^{k+1} once the current iterate x^k satisfies $\left\|q\left(x^k,\epsilon^k\right)\right\| < 1$.

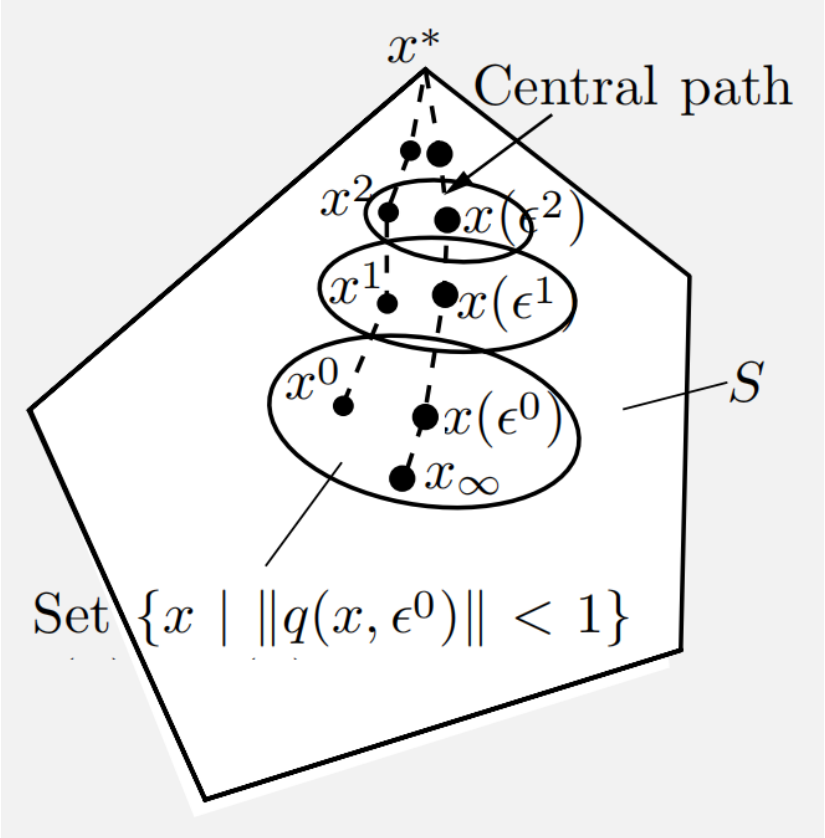

Figure: Following approximately the central path. For each ϵ^k, it is sufficient to carry out the minimization of F_{ϵ^k} up to where $\left\| q\left(x^k, \epsilon^k\right) \right\| < 1$.

Proposition

If $x > 0, Ax = b$, *and* $\|q(x, \epsilon)\| < 1$, *then*

$$c'x - f^* \leq c'x - b'\lambda \leq \epsilon(n + \|q(x, \epsilon)\|\sqrt{n}) \leq \epsilon(n + \sqrt{n}),$$

where $\lambda = (AX^2A')^{-1}AX(Xc - \epsilon e)$, *and* f^* *is the optimal value of* (LP) *that is,*

$$f^* = \min_{Ay=b, y\geq 0} c'y.$$

Proof: Using the definition of q, we can write the hypothesis $\|q(x, \epsilon)\| < 1$ as

$$\left\| \frac{X\left(c - A'\lambda\right)}{\epsilon} - e \right\| < 1$$

Thus the coordinates of $(X(c - A'\lambda)/\epsilon) - e$ must lie between -1 and 1 implying that the coordinates of $X(c - A'\lambda)$ are positive. Since the

diagonal elements of X are positive, it follows that the coordinates of $c - A'\lambda$ are also positive. Hence $c \geq A'\lambda$, and for any optimal solution x^* of (LP), we obtain (using the fact $x^* \geq 0$)

$$f^* = c'x^* \geq \lambda' A x^* = \lambda' b.$$

On the other hand, since $||e|| = y/n$, we have

$$e'\left(\frac{X\left(c - A'\lambda\right)}{\epsilon} - e\right) \leq \|e\| \left\|\frac{X\left(c - A'\lambda\right)}{\epsilon} - e\right\| = \sqrt{n}\|q(x,\epsilon)\| \leq \sqrt{n},$$

and

$$e'\left(\frac{X\left(c - A'\lambda\right)}{\epsilon} - e\right) = \frac{x'\left(c - A'\lambda\right)}{\epsilon} - n = \frac{c'x - b'\lambda}{\epsilon} - n \geq \frac{c'x - f^*}{\epsilon} - n.$$

The result follows.

Path-Following by Using Newton's Method

Since in order to implement the termination criterion $\|q(x,\epsilon)\| < 1$, we must calculate the pure Newton iterate $\overline{x} = x - Xq(x,\epsilon)$, it is natural to use a convergent version of Newton's method for approximate minimization of F_ϵ. This method replaces x by

$$\tilde{x} = x + \alpha(\overline{x} - x),$$

where α is a stepsize selected by the minimization rule or the Armijo rule (with unit initial stepsize) over the range of positive stepsizes such that x is an interior point.

Proposition

If $x > 0, Ax = b$, *and* $\|q(x,\epsilon)\| < 1$, *then the pure Newton iterate* $\overline{x} = x - Xq(x,\epsilon)$ *is an interior point, i.e.,* $\vec{x} \in S$. *Furthermore, we have* $\|q(\overline{x},\epsilon)\| < 1$ *and in fact*

$$\|q(\overline{x},\epsilon)\| \leq \|q(x,\epsilon)\|^2.$$

Proof: Let us define $p = Xz/\epsilon = X(c - A'\lambda)/\epsilon$, so that $q(x,\epsilon) = p - e$. Since $\|p - e\| < 1$, we see that the coordinates of p satisfy $0 < p_i < 2$ for all i. We have $\overline{x} = x - X(p - e)$, so that $\overline{x}_i = (2 - p_i)x_i > 0$ for all i, and since also $A\overline{x} = b$, it follows that $\overline{x}$ is an interior point.
It can be shown that the vector $\overline{\lambda}$ corresponding to $\overline{x}$ satisfies

$$\overline{\lambda} = \arg \min_{\xi \in \Re^m} \left\| \frac{\overline{X}\,(c - A'\xi)}{\epsilon} - e \right\|,$$

where $\vec{x}$ is the diagonal matrix with $\overline{x}_i$ along the diagonal. Hence,

$$\|q(\overline{x},\epsilon)\| = \left\| \frac{\overline{X}\left(c - A'\overline{\lambda}\right)}{\epsilon} - e \right\| \leq \left\| \frac{\overline{X}\left(c - A'\lambda\right)}{\epsilon} - e \right\| = \left\| \overline{X}X^{-1}p - e \right\|.$$

Since $\overline{x} = 2x - Xp$, we have

$$\overline{X}X^{-1}p = (2X - XP)X^{-1}p = 2p - Pp,$$

where P is the diagonal matrix with p_i along the diagonal. The last two relations yield

$$\|q(\overline{x},\epsilon)\|^2 \leq \|2p - Pp - e\|^2 \leq \sum_{i=1}^{n} \left(2p_i - p_i^2 - 1\right)^2 = \sum_{i=1}^{n} \left(p_i - 1\right)^4$$

$$\leq \left(\sum_{i=1}^{n} \left(p_i - 1\right)^2 \right)^2 = \|p - e\|^4 = \|q(x,\epsilon)\|^4 .$$

This proves the result.

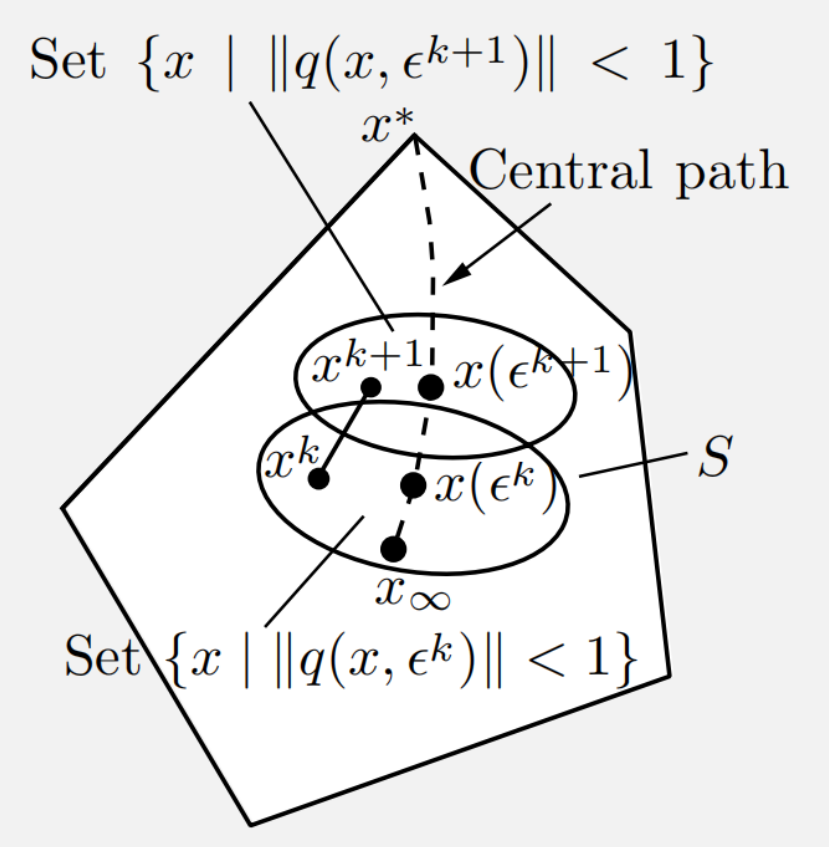

Figure: Following approximately the central path by using a single Newton step for each ϵ^k. If ϵ^k is close to ϵ^{k+1} and x^k is close to the central path, one expects that x^{k+1} obtained from x^k by a single pure Newton step will also be close to the central path.

Proposition

Suppose that $x > 0, Ax = b$, *and that* $\|q(x, \epsilon)\| \leq \gamma$ *for some* $\gamma < 1$. *For any* $\delta \in \left(0, n^{1/2}\right)$, *let* $\bar{\epsilon} = \left(1 - \delta n^{-1/2}\right)\epsilon$. *Then*

$$\|q(\overline{x}, \bar{\epsilon})\| \leq \frac{\gamma^2 + \delta}{1 - \delta n^{-1/2}}.$$

In particular, if

$$\delta \leq \frac{\gamma(1 - \gamma)}{1 + \gamma},$$

we have $\|q(\overline{x}, \bar{\epsilon})\| \leq \gamma$

Proof: Let $\theta = \delta n^{-1/2}$. We have

$$q(\overline{x}, \bar{\epsilon}) = \frac{\overline{X}z}{\bar{\epsilon}} - e = \frac{\overline{X}z}{(1-\theta)\epsilon} - e = \frac{q(\overline{x}, \epsilon) + e}{1 - \theta} - e = \frac{1}{1-\theta}(q(\overline{x}, \epsilon) + \theta e).$$

Thus,

$$\begin{aligned}\|q(\overline{x},\overline{\epsilon})\| &\leq \frac{1}{1-\theta}(\|q(\overline{x},\epsilon)\| + \theta\|e\|) \\ &= \frac{1}{1-\theta}\left(\|q(\overline{x},\epsilon)\| + \theta n^{1/2}\right) \\ &\leq \frac{1}{1-\theta}\left(\|q(x,\epsilon)\|^2 + \delta\right) \\ &\leq \frac{\gamma^2+\delta}{1-\theta}\end{aligned}$$

Finally, Since $\left(\gamma^2+\delta\right)/(1-\delta) \leq \gamma$, which, in combination with the relation just proved, implies that $\|q(\overline{x},\overline{\epsilon})\| \leq \gamma$.

One can maintain x very close to the central path (i.e., $\|q(x,\epsilon)\|$ very small) provided ont takes δ to be very small (or $1-\delta n^{-1/2}$ very close to 1).
Unfortunately, even when γ is close to 1, in order to guarantee the single-step attainment of the tolerance $\|q(x,\epsilon)\| < \gamma$, it is still necessary to decrease ϵ very slowly.
This means that, even though each approximate minimization after the first will require a single Newton step, a very large number of approximate minimizations will be needed to attain an acceptable accuracy.
Thus, it may be more efficient in practice to decrease ϵ^k at a faster rate, while accepting the possibility of multiple Newton steps before switching from ϵ^k to ϵ^{k+1}.

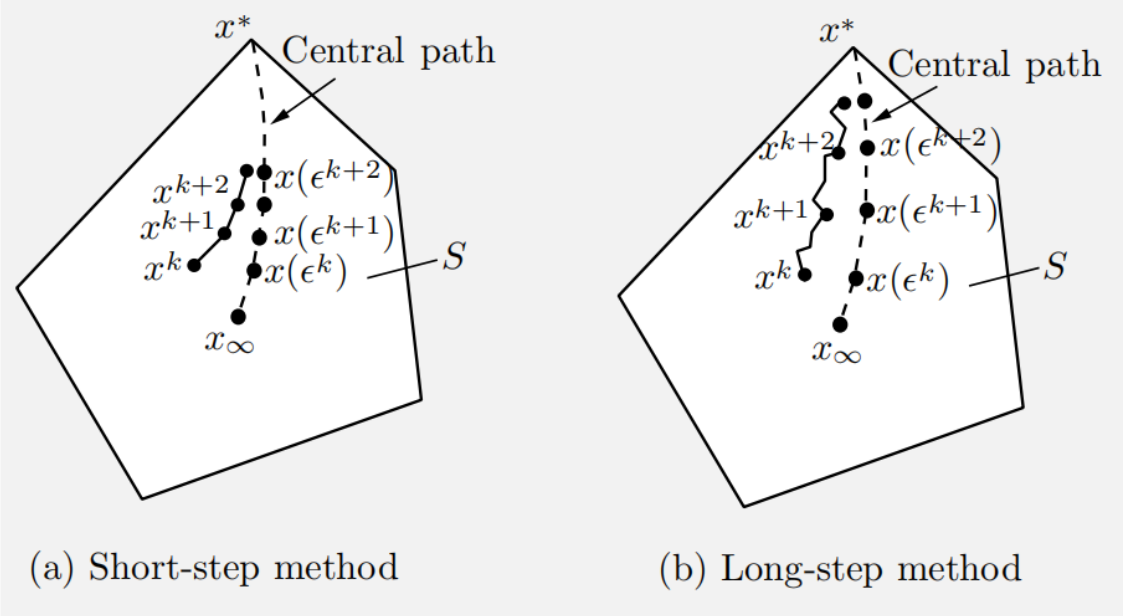

Figure: Following approximately the central path by decreasing ϵ^k slowly as in (a) or quickly as in (b). In (a) a single Newton step is required in each approximate minimization at the expense of a large number of approximate minimizations.

Quadratic and Convex Programming

The logarithmic barrier method in conjunction with Newton's method can also be fruitfully applied to the convex programming problem

$$\min\ f(x), \qquad \text{s.t. } Ax = b,\ x \geq 0,$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is convex.

Primal-Dual Interior Point Methods

We now focus on the linear program

$$\begin{aligned} &\text{minimize } \mathrm{c}'x, \\ &\text{subject to} \qquad Ax = b, \quad x \geq 0, \end{aligned}$$

where $\mathrm{c} \in \Re^n$ and $b \in \Re^m$ are given vectors, and A is an $m \times n$ matrix of rank m. The dual problem is given by

$$\begin{aligned} &\text{maximize} \quad b'\lambda, \\ &\text{subject to} \quad A'\lambda \leq \mathrm{c}. \end{aligned}$$

(LP) has an optimal solution if and only if (DP) has an optimal solution. Furthermore, when optimal solutions to (LP) and (DP) exist, the corresponding optimal values are equal.

Recall that the logarithmic barrier method involves finding for various $\epsilon > 0$,

$$x(\epsilon) = \arg\min_{x \in S} F_\epsilon(x),$$

where

$$F_\epsilon(x) = \mathrm{c}'x - \epsilon \sum_{i=1}^{n} \ln x_i.$$

and S is the interior set

$$S = \{x | Ax = b, x > 0\}.$$

We assume that S is nonempty and bounded.

We now consider a Lagrangian approach, where we apply Newton's method for solving the system of optimality conditions for the problem of minimizing $F_{\epsilon^k}(\cdot)$ over S. The salient features of this approach are:

(a) Only one Newton/Lagrangian iteration is carried out for each value of ϵ^k.

(b) For every k, the pair (x^k, λ^k) is such that x^k is an interior point of the positive orthant, that is, $x^k > 0$, while λ^k is an interior point of the dual feasible region, that is, $\mathrm{c} - A'\lambda^k > 0$.

(c) Global convergence is enforced by using as the merit function the expression

$$P^k = x^{k'} z^k + \left\| Ax^k - b \right\|.$$

where z^k is the vector of slack variables

$$z^k = \mathrm{c} - A'\lambda^k.$$

Let us write the necessary and sufficient conditions for (x, λ) to be a (global) minimum-Lagrange multiplier pair for the problem of minimizing the barrier function $F_\epsilon(x)$ subject to $Ax = b$. They are

$$\mathrm{c} - \epsilon x^{-1} - A'\lambda = 0, \quad Ax = b,$$

where x^{-1} denotes the vector with coordinates $(x_i)^{-1}$. Let z be the vector of slack variables

$$z = \mathrm{c} - A'\lambda.$$

Note that λ is dual feasible if and only if $z \geq 0$. This can be written in the equivalent form

$$XZe = \epsilon e$$
$$Ax = b$$
$$z + A'\lambda = \mathrm{c}.$$

Given (x, λ, z) satisfying $z + A'\lambda = c$, and such that $x > 0$ and $z > 0$, a Newton iteration for solving this system is

$$x(\alpha, \epsilon) = x + \alpha \Delta x,$$
$$\lambda(\alpha, \epsilon) = \lambda + \alpha \Delta \lambda,$$
$$z(\alpha, \epsilon) = z + \alpha \Delta z,$$

where α is a stepsize such that $0 < \alpha \leq 1$ and

$$x(\alpha, \epsilon) > 0, \quad z(\alpha, \epsilon) > 0$$

and the pure Newton step $(\Delta x, \Delta \lambda, \Delta z)$ solves

$$X \Delta z + Z \Delta x = -v,$$
$$A \Delta x = b - Ax,$$
$$\Delta z + A' \Delta \lambda = 0,$$

with v defined by

$$v = XZe - \epsilon e.$$

Merit Function Improvement

Let

$$\begin{aligned} P(\alpha, \epsilon) &= g(\alpha, \epsilon) + \|Ax(\alpha, \epsilon) - b\| \\ &= g - \alpha(g - n\epsilon) + \alpha^2 (Ax - b)' \Delta\lambda + (1 - \alpha)\|Ax - b\| \end{aligned}$$

or

$$P(\alpha, \epsilon) = P - \alpha(g - n\epsilon + \|Ax - b\|) + \alpha^2 (Ax - b)' \Delta\lambda.$$

Thus if ϵ is chosen to satisfy

$$\epsilon < \frac{g}{n}$$

and α is chosen to be small enough so that the second order term $\alpha^2 (Ax - b)' \Delta\lambda$ is dominated by the first order term $\alpha(g - n\epsilon)$, the merit function will be improved as a result of the iteration.

A General Class of Primal-Dual Algorithms

Let us consider now the general class of algorithms of the form

$$x^{k+1} = x(\alpha^k, \epsilon^k), \quad \lambda^{k+1} = \lambda(\alpha^k, \epsilon^k), \quad z^{k+1} = z(\alpha^k, \epsilon^k),$$

where α^k and ϵ^k are positive scalars such that

$$x^{k+1} > 0, \quad z^{k+1} > 0, \quad \epsilon^k < \frac{g^k}{n},$$

where g^k is the inner product

$$g^k = x^{k'} z^k + \left(Ax^k - b\right)' \lambda^k.$$

and α^k is such that the merit function P^k is reduced. Initially we must have $x^0 > 0$, and $z^0 = \mathrm{c} - A'\lambda^0 > 0$. These methods have been called primal-dual.

With properly chosen sequences α^k and ϵ^k, and appropriate implementation, the practical performance of the primal-dual methods has been shown to be excellent. The choice

$$\epsilon^k = \frac{g^k}{n^2},$$

leading to the relation

$$g^{k+1} = \left(1 - \alpha^k + \alpha^k/n\right) g^k$$

for feasible x^k, has been suggested as a good practical rule.
When x^k is feasible, α^k is chosen as $\theta\tilde{\alpha}^k$, where θ is a factor very close to 1(say 0.999), and $\tilde{\alpha}^k$ is the maximum stepsize α that guarantees that $x\left(\alpha, \epsilon^k\right) \geq 0$ and $z\left(\alpha, \epsilon^k\right) \geq 0$

$$\tilde{\alpha}^k = \min\left\{\min_{i=1,\ldots,n}\left\{\frac{x_i^k}{-\Delta x_i} | \Delta x_i < 0\right\}, \min_{i=1,\ldots,n}\left\{\frac{z_i^k}{-\Delta z_i} | \Delta z_i < 0\right\}\right\}.$$

When x^k is not feasible, the choice of α^k must also be such that the merit function is improved.

Predictor-Corrector Variants

Given (x, z, λ) with $x > 0$, and $z = \mathrm{c} - A'\lambda > 0$, the predictor iteration, solves for $(\Delta\hat{x}, \Delta\hat{z}, \Delta\hat{\lambda})$ the system

$$X\Delta\hat{z} + Z\Delta\hat{x} = -\hat{v},$$
$$A\Delta\hat{x} = b - Ax,$$
$$\Delta\hat{z} + A'\Delta\hat{\lambda} = 0,$$

with $\hat{v}$ defined by

$$\hat{v} = XZe - \hat{\epsilon}e.$$

The corrector iteration solves for $(\Delta\overline{x}, \Delta\overline{z}, \Delta\overline{\lambda})$ the system

$$
\begin{aligned}
X\Delta\overline{z} + Z\Delta\overline{x} &= -\overline{v}, \\
A\Delta\overline{x} &= b - A(x + \Delta\hat{x}), \\
\Delta\overline{z} + A'\Delta\overline{\lambda} &= 0,
\end{aligned}
$$

with $\overline{v}$ defined by

$$
\overline{v} = (X + \Delta\hat{X})(Z + \Delta\hat{Z})e - \overline{\epsilon}e,
$$

where ΔX and ΔZ are the diagonal matrices corresponding to $\Delta\hat{x}$ and $\Delta\hat{z}$, respectively. Here $\hat{\epsilon}$ and $\overline{\epsilon}$ are the barrier parameters corresponding to the two iterations.

The composite Newton direction is

$$\begin{aligned}\Delta x &= \Delta\hat{x} + \Delta\overline{x},\\ \Delta z &= \Delta\hat{z} + \Delta\overline{z},\\ Delta\lambda &= \Delta\lambda + \Delta\overline{\lambda},\end{aligned}$$

and the corresponding iteration is

$$\begin{aligned}x(\alpha,\epsilon) &= x + \alpha\Delta x,\\ \lambda(\alpha,\epsilon) &= \lambda + \alpha\Delta\lambda,\\ z(\alpha,\epsilon) &= z + \alpha\Delta z,\end{aligned}$$

where α is a stepsize such that $0 < \alpha \leq 1$ and

$$x(\alpha,\epsilon) > 0, \quad z(\alpha,\epsilon) > 0.$$

Penalty and Augmented Lagrangian Methods

Consider first the equality constrained problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \quad x \in X, \end{array}$$

where $f : \Re^n \mapsto \Re, h : \Re^n \mapsto \Re^m$ are given functions, and X is a given subset of $\Re^n$.

Much of our analysis in this section will focus on the case where $X = \Re^n$, and x^* together with a Lagrange multiplier vector $\lambda*$ satisfies the sufficient optimality conditions. At the center of our development is the augmented Lagrangian function $L_c : \Re^n \times \Re^m \mapsto \Re$ given by

$$L_c(x, \lambda) = f(x) + \lambda' h(x) + \frac{c}{2}\|h(x)\|^2,$$

where c is a positive penalty parameter.

There are two mechanisms by which unconstrained minimization of $L_c(\cdot, \lambda)$ can yield points close to x^* :

(a) By taking λ close to λ^*.

(b) By taking c very large.

Example

Consider the two-dimensional problem

$$\begin{array}{ll} \text{minimize} & f(x) = \frac{1}{2}\left(x_1^2 + x_2^2\right) \\ \text{subject to} & x_1 = 1 \end{array}$$

with optimal solution $x^* = (1, 0)$ and corresponding Lagrange multiplier $\lambda^* = -1$. The augmented Lagrangian is

$$L_c(x, \lambda) = \frac{1}{2}\left(x_1^2 + x_2^2\right) + \lambda\left(x_1 - 1\right) + \frac{c}{2}\left(x_1 - 1\right)^2.$$

and by setting its gradient to zero we can verify that its unique unconstrained minimum $x(\lambda, c)$ has coordinates given by

$$x_1(\lambda, c) = \frac{c - \lambda}{c + 1}, \quad x_2(\lambda, c) = 0.$$

Thus, we have for all $c > 0$,

$$\lim_{\lambda \to \lambda^*} x_1(\lambda, c) = x_1(-1, c) = 1 = x_1^*, \quad \lim_{\lambda \to \lambda^*} x_2\left(\lambda^*, c\right) = 0 = x_2^*,$$

showing that as λ is chosen close to λ^*, the unconstrained minimum of $L_c(x, \lambda)$ approaches the constrained minimum.

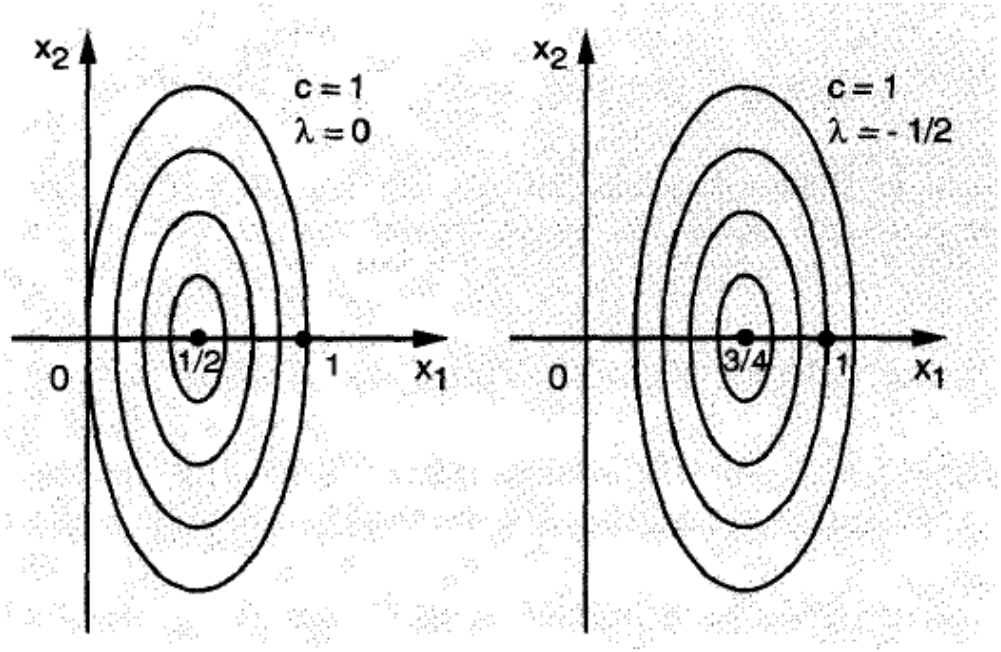

Figure. Equal cost surfaces of the augmented Lagrangian

$$L_c(x,\lambda) = \frac{1}{2}\left(x_1^2 + x_2^2\right) + \lambda\left(x_1 - 1\right) + \frac{c}{2}\left(x_1 - 1\right)^2,$$

of Example for $c = 1$ and two different values of λ. The unconstrained minimum of $L_c(x,\lambda)$ approaches the constrained minimum $x^* = (1,0)$ as $\lambda \to \lambda^* = -1$

We also have for all A,

$$\lim_{c\to\infty} x_1(\lambda, c) = 1 = x_1^*, \quad \lim_{c\to\infty} x_2(\lambda, c) = 0 = x_2^*,$$

showing that as c increases, the unconstrained minimum of $L_c(x, \lambda)$ approaches the constrained minimum.

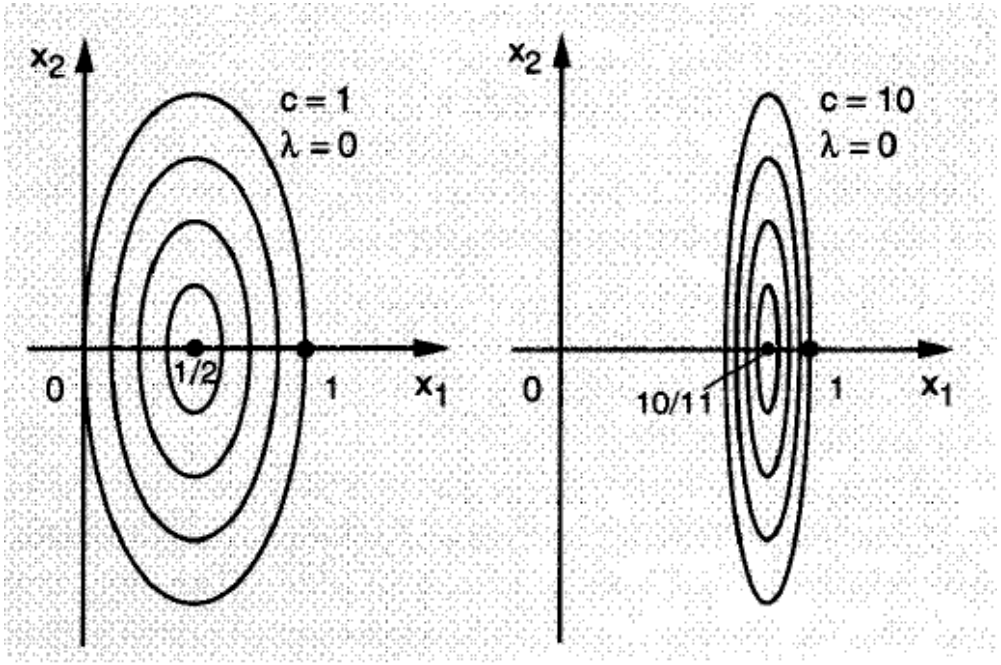

Figure. Equal cost surfaces of the augmented Lagrangian

$$L_c(x,\lambda) = \frac{1}{2}\left(x_1^2 + x_2^2\right) + \lambda\left(x_1 - 1\right) + \frac{c}{2}\left(x_1 - 1\right)^2,$$

of Example for $\lambda = 0$ and two different values of c. The unconstrained minimum of $L_c(x,\lambda)$ approaches the constrained minimum $x^* = (1,0)$ as $c \to \infty$.

The Quadratic Penalty Function Method

The quadratic penalty function method is motivated by the preceding considerations. It consists of solving a sequence of problems of the form

$$\begin{array}{ll} \text{minimize} & L_{c^k}\left(x, \lambda^k\right) \\ \text{subject to} & x \in X \end{array},$$

where $\left\{\lambda^k\right\}$ is a sequence in $\Re^m$ and $\left\{c^k\right\}$ is a positive penalty parameter sequence.

Proposition

Assume that f and h are continuous functions, that X is a closed set, and that the constraint set $\{x \in X | h(x) = 0\}$ is nonempty. For $k = 0, 1, \ldots,$ let x^k be a global minimum of the problem

$$\begin{array}{ll} \textit{minimize} & L_{c^k}\left(x, \lambda^k\right) \\ \textit{subject to} & x \in X \end{array},$$

where $\left\{\lambda^k\right\}$ is bounded, $0 < c^k < c^{k+1}$ for all k, and $c^k \to \infty$. Then every limit point of the sequence $\left\{x^k\right\}$ is a global minimum of the original problem.

Proof: Let $\overline{x}$ be a limit point of $\left\{x^k\right\}$. We have by definition of x^k

$$L_{c^k}\left(x^k, \lambda^k\right) \leq L_{c^k}\left(x, \lambda^k\right), \quad \forall x \in X.$$

Let f^* denote the optimal value of the original problem. We have

$$\begin{aligned} f^* &= \inf_{h(x)=0, x\in X} f(x) \\ &= \inf_{h(x)=0, x\in X} \left\{ f(x) + \lambda^{k'} h(x) + \frac{c^k}{2}\|h(x)\|^2 \right\} \\ &= \inf_{h(x)=0, x\in X} L_{c^k}\left(x, \lambda^k\right) \end{aligned}$$

Hence, by taking the infimum of the right-hand side of above inequality over $x \in X, h(x) = 0$, we obtain

$$L_{c^k}\left(x^k, \lambda^k\right) = f\left(x^k\right) + \lambda^{k'} h\left(x^k\right) + \frac{c^k}{2}\left\|h\left(x^k\right)\right\|^2 \leq f^*.$$

The sequence $\left\{\lambda^k\right\}$ is bounded and hence it has a limit point $\overline{\lambda}$. Without loss of generality, we may assume that $\lambda^k \to \overline{\lambda}$.

By taking the limit superior in the relation above and by using the continuity of f and h, we obtain

$$f(\overline{x}) + \overline{\lambda}' h(\overline{x}) + \limsup_{k \to \infty} \frac{c^k}{2} \left\| h\left(x^k\right) \right\|^2 \leq f^*.$$

Since $\left\| h\left(x^k\right) \right\|^2 \geq 0$ and $c^k \to \infty$, it follows that $h\left(x^k\right) \to 0$ and

$$h(\overline{x}) = 0,$$

for otherwise the left-hand side of above inequality would equal ∞, while $f^* < \infty$ (since the constraint set is assumed nonempty). Since X is a closed set, we also obtain that $\overline{x} \in X$. Hence, $\overline{x}$ is feasible, we have $f(\overline{x}) \leq f^*$, it follows that $\overline{x}$ is optimal.

Lagrange Multiplier Estimates —Inexact Minimization

In particular, when $X = \Re^n$, and f and h are differentiable, the algorithm for solving the unconstrained problem

$$\begin{array}{ll} \text{minimize} & L_{c^k}\left(x, \lambda^k\right) \\ \text{subject to} & x \in \Re^n \end{array}$$

will typically be terminated at a point x^k satisfying

$$\left\|\nabla_x L_{c^k}\left(x^k, \lambda^k\right)\right\| \leq \epsilon^k,$$

where ϵ^k is some small scalar.

Proposition

Assume that $X = \Re^n$, *and* f *and* h *are continuously differentiable. For* $k = 0, 1, \ldots,$ *let* x^k *satisfy*

$$\left\| \nabla_x L_{c^k} \left(x^k, \lambda^k \right) \right\| \leq \epsilon^k,$$

where $\left\{ \lambda^k \right\}$ *is bounded, and* $\left\{ \epsilon^k \right\}$ *and* $\left\{ c^k \right\}$ *satisfy*

$$0 < \mathrm{c}^k < \mathrm{c}^{k+1}, \quad \forall k, \quad c^k \to \infty,$$
$$0 \leq \epsilon^k, \quad \forall k, \quad \epsilon^k \to 0.$$

Assume that a subsequence $\left\{ x^k \right\}_K$ *converges to a vector* x^* *such that* $\nabla h \left(x^* \right)$ *has rank* m. *Then*

$$\left\{ \lambda^k + c^k h \left(x^k \right) \right\}_K \to \lambda^*,$$

where $\lambda*$ *is a vector satisfying, together with* x^*, *the first order necessary conditions*

$$\nabla f \left(x^* \right) + \nabla h \left(x^* \right) \lambda^* = 0, \quad h \left(x^* \right) = 0.$$

Proof: Without loss of generality we assume that the entire sequence $\left\{x^k\right\}$ converges to x^*. Define for all k

$$\tilde{\lambda}^k = \lambda^k + c^k h\left(x^k\right).$$

We have

$$\nabla_x L_{c^k}\left(x^k, \lambda^k\right) = \nabla f\left(x^k\right) + \nabla h\left(x^k\right)\left(\lambda^k + c^k h\left(x^k\right)\right) = \nabla f\left(x^k\right) + \nabla h\left(x^k\right)\tilde{\lambda}^k.$$

Since $\nabla h\left(x^*\right)$ has rank m, $\nabla h\left(x^k\right)$ has rank m for all k that are sufficiently large. Without loss of generality, we assume that $\nabla h\left(x^k\right)$ has rank m for all k . Then, by multiplying above equality with

$$\left(\nabla h\left(x^k\right)' \nabla h\left(x^k\right)\right)^{-1} \nabla h\left(x^k\right)',$$

we obtain

$$\tilde{\lambda}^k = \left(\nabla h\left(x^k\right)' \nabla h\left(x^k\right)\right)^{-1} \nabla h\left(x^k\right)' \left(\nabla_x L_{c^k}\left(x^k, \lambda^k\right) - \nabla f\left(x^k\right)\right).$$

The hypothesis implies that $\nabla_x L_{ck}\left(x^k, \lambda^k\right) \to 0$, so

$$\tilde{\lambda}^k \to \lambda^*,$$

where

$$\lambda^* = -\left(\nabla h\left(x^*\right)' \nabla h\left(x^*\right)\right)^{-1} \nabla h\left(x^*\right)' \nabla f\left(x^*\right).$$

Using again the fact $\nabla_x L_{c^k}\left(x^k, \lambda^k\right) \to 0$, we see that

$$\nabla f\left(x^*\right) + \nabla h\left(x^*\right) \lambda^* = 0,$$

Since $\left\{\lambda^k\right\}$ is bounded and $\lambda^k + c^k h\left(x^k\right) \to \lambda^*$, it follows that $\left\{c^k h\left(x^k\right)\right\}$ is bounded. Since $c^k \to \infty$, we must have $h\left(x^k\right) \to 0$ and we conclude that $h\left(x^*\right) = 0$.

Practical Behavior — Ill-Conditioning

Let us now consider the practical behavior of the quadratic penalty method. Assume that the k th unconstrained minimization of $L_{c^k}\left(x, \lambda^k\right)$ is terminated when

$$\left\|\nabla_x L_c\left(x^k, \lambda^k\right)\right\| \leq \epsilon^k,$$

where $\epsilon^k \to 0$. There are three possibilities:

(a) The method breaks down because an x^k satisfying $\|\nabla_x L_{c^k}(x^k, \lambda^k)\| \leq \epsilon^k$ cannot be found.

(b) A sequence $\{x^k\}$ with $\|\nabla_x L_c(x^k, \lambda^k)\| \leq \epsilon^k$ for all k is obtained but it either has no limit points, or for each of its limit points x^* the matrix $\nabla h(x^*)$ has linearly dependent columns.

(c) A sequence $\{x^k\}$ with $\|\nabla_x L_{c^k}(x^k, \lambda^k)\| \leq \epsilon^k$ for all k is found and it has a limit point x^* such that $\nabla h(x^*)$ has rank m. Then, by Prop. $4.2.2, x*$ together with $\lambda*$ [the corresponding limit point of $\{\lambda^k + c^k h(x^k)\}$] satisfies the first order necessary conditions for optimality.

Example

Consider the problem

$$\begin{array}{l} \text{minimize } f(x) = \frac{1}{2}\left(x_1^2 + x_2^2\right) \\ \text{subject to } x_1 = 1 \end{array} .$$

The augmented Lagrangian is

$$L_c(x, \lambda) = \frac{1}{2}\left(x_1^2 + x_2^2\right) + \lambda\left(x_1 - 1\right) + \frac{c}{2}\left(x_1 - 1\right)^2 ,$$

and its Hessian is

$$\nabla_{xx}^2 L_c(x, \lambda) = I + c \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \end{pmatrix} = \begin{pmatrix} 1+c & 0 \\ 0 & 1 \end{pmatrix} .$$

The ratio of largest to smallest eigenvalue of the Hessian is $1+c$ and tends to ∞ as $c \to \infty$, The associated ill-conditioning can also be observed from the narrow level sets of the augmented Lagrangian for large c.

Inequality Constraints

Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_1(x) = 0, \ldots, h_m(x) = 0, \ g_1(x) \leq 0, \ldots, g_r(x) \leq 0 \end{array}$$

The quadratic penalty method for the inequality constrained problem consists of a sequence of minimizations of the form

$$\begin{array}{ll} \text{minimize} & L_c\left(x, \lambda^k, \mu^k\right) \\ \text{subject to} & x \in X \end{array},$$

where

$$L_c(x, \lambda, \mu) = f(x) + \lambda' h(x) + \frac{c}{2}\|h(x)\|^2 + \sum_{j=1}^{r}\{\mu_j g_j^+(x, \mu, c) + \frac{c}{2}(g_j^+(x, \mu, c))^2\}$$

$\{\lambda^k\}$ and $\{\mu^k\}$ are sequences in $\Re^m$ and $\Re^r$, with coordinates denoted by λ_i^k and μ_j^k, respectively, and $\{c^k\}$ is a positive penalty parameter sequence.

Multiplier Methods - Main Ideas

Let us return to the case where $X = \Re^n$ and the problem has only equality constraints,

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0 \end{array}$$

- We mentioned earlier that optimal solutions of this problem can be well approximated by unconstrained minima of the augmented Lagrangian $L_c(\cdot, \lambda)$ under two types of circumstances:
  (a) The vector λ is close to a Lagrange multiplier.
  (b) The penalty parameter c is large.
- Consider intelligent ways to update λ^k so that it tends to a Lagrange multiplier.

The Method of Multipliers

A first update formula for λ^k in the quadratic penalty method is

$$\lambda^{k+1} = \lambda^k + c^k h\left(x^k\right).$$

Example: Consider the problem

$$\begin{array}{l} \text{minimize } f(x) = \frac{1}{2}\left(x_1^2 + x_2^2\right) \\ \text{subject to } x_1 = 1 \end{array}$$

with optimal solution $x^* = (1, 0)$ and Lagrange multiplier $\lambda^* = -1$. The augmented Lagrangian is

$$L_c(x, \lambda) = \frac{1}{2}\left(x_1^2 + x_2^2\right) + \lambda\left(x_1 - 1\right) + \frac{c}{2}\left(x_1 - 1\right)^2.$$

The vectors x^k generated by the method of multipliers minimize $L_{ck}\left(\cdot, \lambda^k\right)$ and are given by

$$x^k = \left(\frac{c^k - \lambda^k}{c^k + 1}, 0\right).$$

Using this expression, then

$$\lambda^{k+1} = \lambda^k + c^k \left(\frac{c^k - \lambda^k}{c^k + 1} - 1 \right) = \frac{\lambda^k}{c^k + 1} - \frac{c^k}{c^k + 1},$$

or by introducing the Lagrange multiplier $\lambda^* = -1$

$$\lambda^{k+1} - \lambda^* = \frac{\lambda^k - \lambda^*}{c^k + 1}.$$

From this formula, it can be seen that

(a) $\lambda^k \to \lambda^* = -1$ and $x^k \to x^* = (1, 0)$ for every nondecreasing sequence $\{c^k\}$[since the scalar $1/(c^k + 1)$ multiplying $\lambda^k - \lambda^*$ in the above formula is always less than one].

(b) The convergence rate becomes faster as c^k becomes larger; in fact $\{|\lambda^k - \lambda^*|\}$ converges superlinearly if $c^k \to \infty$

Note that it is not necessary to increase c^k to ∞, although doing so results in a better convergence rate.

Example: Consider the problem

$$\begin{array}{l} \text{minimize } \frac{1}{2}\left(-x_1^2 + x_2^2\right) \\ \text{subject to } x_1 = 1 \end{array}$$

with optimal solution $x^* = (1, 0)$ and Lagrange multiplier $\lambda^* = 1$, The augmented Lagrangian is given by

$$L_c(x, \lambda) = \frac{1}{2}\left(-x_1^2 + x_2^2\right) + \lambda\left(x_1 - 1\right) + \frac{c}{2}\left(x_1 - 1\right)^2.$$

The vector x^k minimizing $L_{c^k}\left(x, \lambda^k\right)$ is given by

$$x^k = \left(\frac{c^k - \lambda^k}{c^k - 1}, 0\right).$$

For this formula to be correct, however, it is necessary that $c^k > 1$; for $c^k < 1$ the augmented Lagrangian has no minum, and the same is true for $c^k = 1$ unless $\lambda^k = 1$. Then

$$\lambda^{k+1} = \lambda^k + c^k \left(\frac{c^k - \lambda^k}{c^k - 1} - 1 \right) = -\frac{\lambda^k}{c^k - 1} + \frac{c^k}{c^k - 1},$$

or by introducing the Lagrange multiplier $\lambda^* = 1$,

$$\lambda^{k+1} - \lambda^* = -\frac{\lambda^k - \lambda^*}{c^k - 1}.$$

Geometric Interpretation of the Method of Multipliers

Assume that f and h are twice differentiable and let x^* be a local minimum of f over $h(x) = 0$. Assume also that x^* is regular and together with its associated Lagrange multiplier vector λ^* satisfies the second order sufficiency conditions. Then, the assumptions of the sensitivity theorem are satisfied and we can consider the primal function

$$p(u) = \min_{h(x)=u} f(x),$$

defined for u in an open sphere centered at $u = 0$. Note that we have

$$p(0) = f(x^*), \quad \nabla p(0) = -\lambda^*.$$

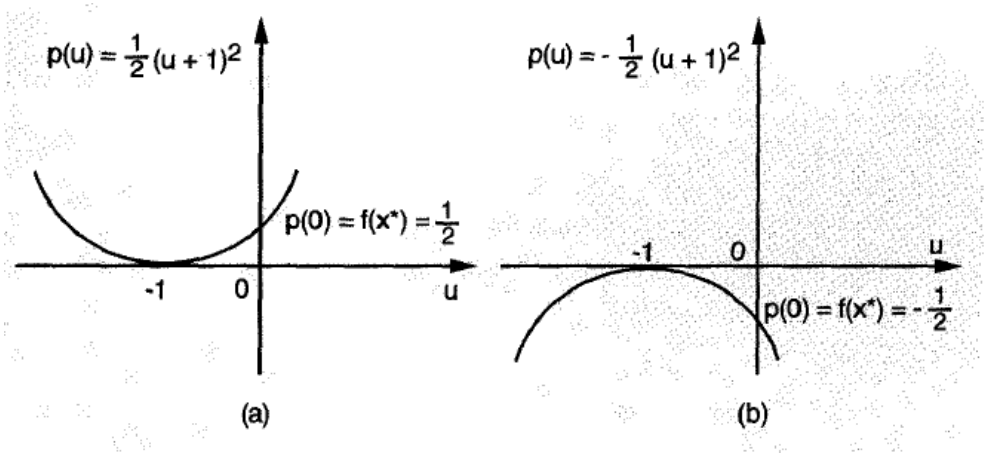

Figure. Illustration of the primal function. In (a) we show the primal function

$$p(u) = \min_{x_1 - 1 = u} \frac{1}{2} \left(x_1^2 + x_2^2 \right).$$

In (b) we show the primal function

$$p(u) = \min_{x_1 - 1 = u} \frac{1}{2} \left(-x_1^2 + x_2^2 \right)$$

The latter primal function is not convex because the cost function is not convex on the subspace that is orthogonal to the constraint set.

We can break down the minimization of $L_c(\cdot, \lambda)$ into two stages, first minimizing over all x such that $h(x) = u$ with u fixed, and then minimizing over all u. Thus,

$$\begin{aligned}\min_x L_c(x, \lambda) &= \min_u \min_{h(x)=u} \left\{ f(x) + \lambda' h(x) + \frac{c}{2}\|h(x)\|^2 \right\} \\ &= \min_u \left\{ p(u) + \lambda' u + \frac{c}{2}\|u\|^2 \right\}.\end{aligned}$$

where the minimization above is understood to be local in a neighborhood of $u = 0$. The minimum is attained at the point $u(\lambda, c)$ for which the gradient of $p(u) + \lambda' u + \frac{c}{2}\|u\|^2$ is zero, or, equivalently,

$$\nabla \left\{ p(u) + \frac{c}{2}\|u\|^2 \right\} \Big|_{u=u(\lambda,c)} = -\lambda$$

Thus,

$$\min_x L_c(x,\lambda) - \lambda' u(\lambda,c) = p(u(\lambda,c)) + \frac{c}{2}\|u(\lambda,c)\|^2$$

so the tangent hyperplane to the graph of $p(u) + \frac{c}{2}\|u\|^2$ at $u(\lambda,c)$ (which has "slope"-λ) intersects the vertical axis at the value $\min_x L_c(x,\lambda)$. It can be seen that if c is sufficiently large, then the function

$$p(u) + \lambda' u + \frac{c}{2}\|u\|^2$$

is convex in a neighborhood of the origin. Furthermore, for λ close to λ^* and large c, the value min $_x L_c(x,\lambda)$ is close to $p(0) = f(x^*)$

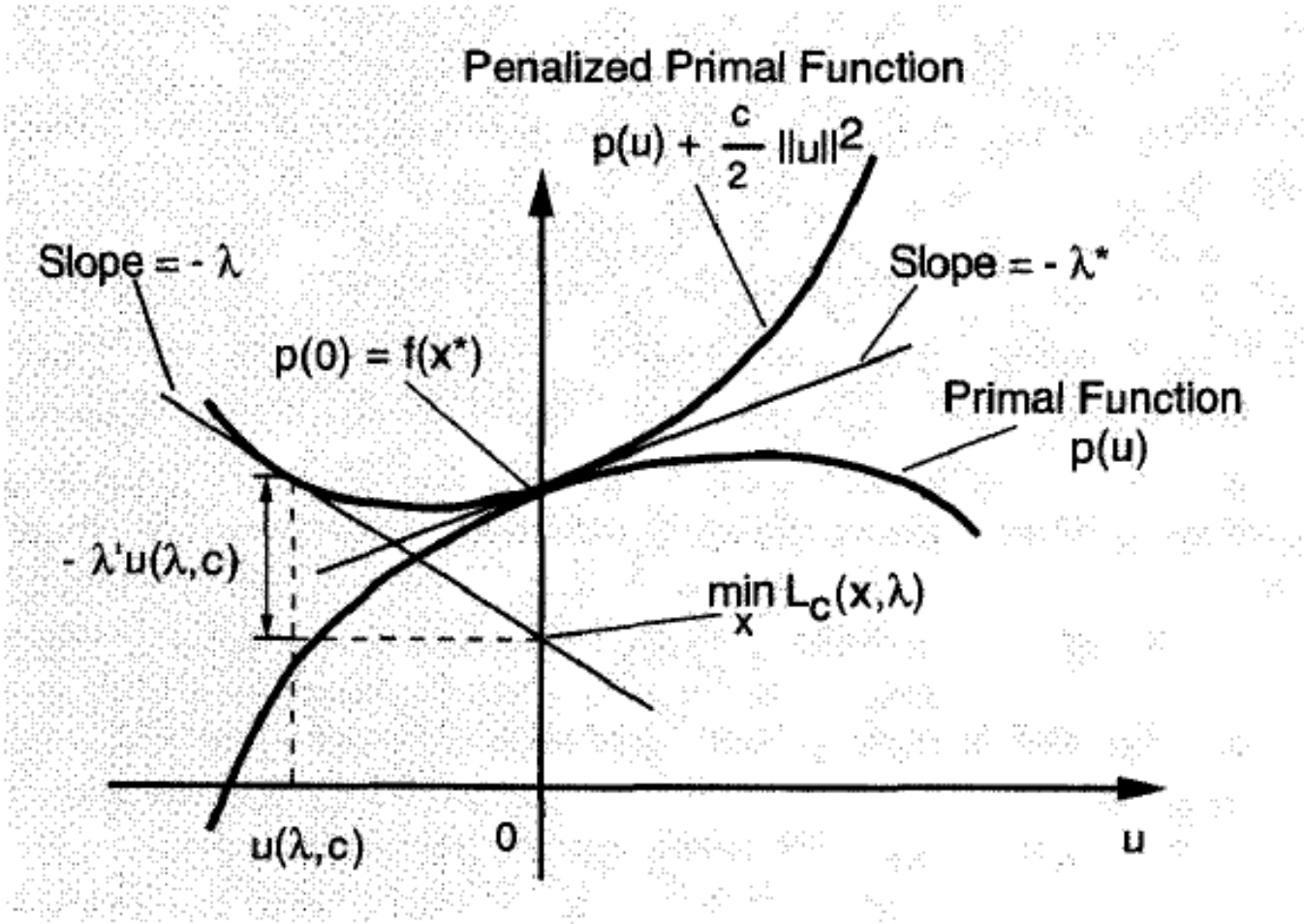

Figure 4.2.5. Geometric interpretation of minimization of the augmented Lagrangian.

The multiplier iteration

$$\lambda^{k+1} = \lambda^k + c^k h\left(x^k\right).$$

To understand this figure, note that if x^k minimizes $L_{ck}\left(\cdot, \lambda^k\right)$, then by the preceding analysis the vector u^k given by $u^k = h\left(x^k\right)$ minimizes $p(u) + \lambda^{k\prime} u + \frac{c^k}{2}\|u\|^2$. Hence,

$$\nabla\left\{p(u) + \frac{c^k}{2}\|u\|^2\right\}\Big|_{u=u^k} = -\lambda^k,$$

and

$$\nabla p\left(u^k\right) = -\left(\lambda^k + c^k u^k\right) = -\left(\lambda^k + c^k h\left(x^k\right)\right).$$

It follows that the next multiplier λ^{k+1} is

$$\lambda^{k+1} = \lambda^k + c^k h\left(x^k\right) = -\nabla p\left(u^k\right).$$

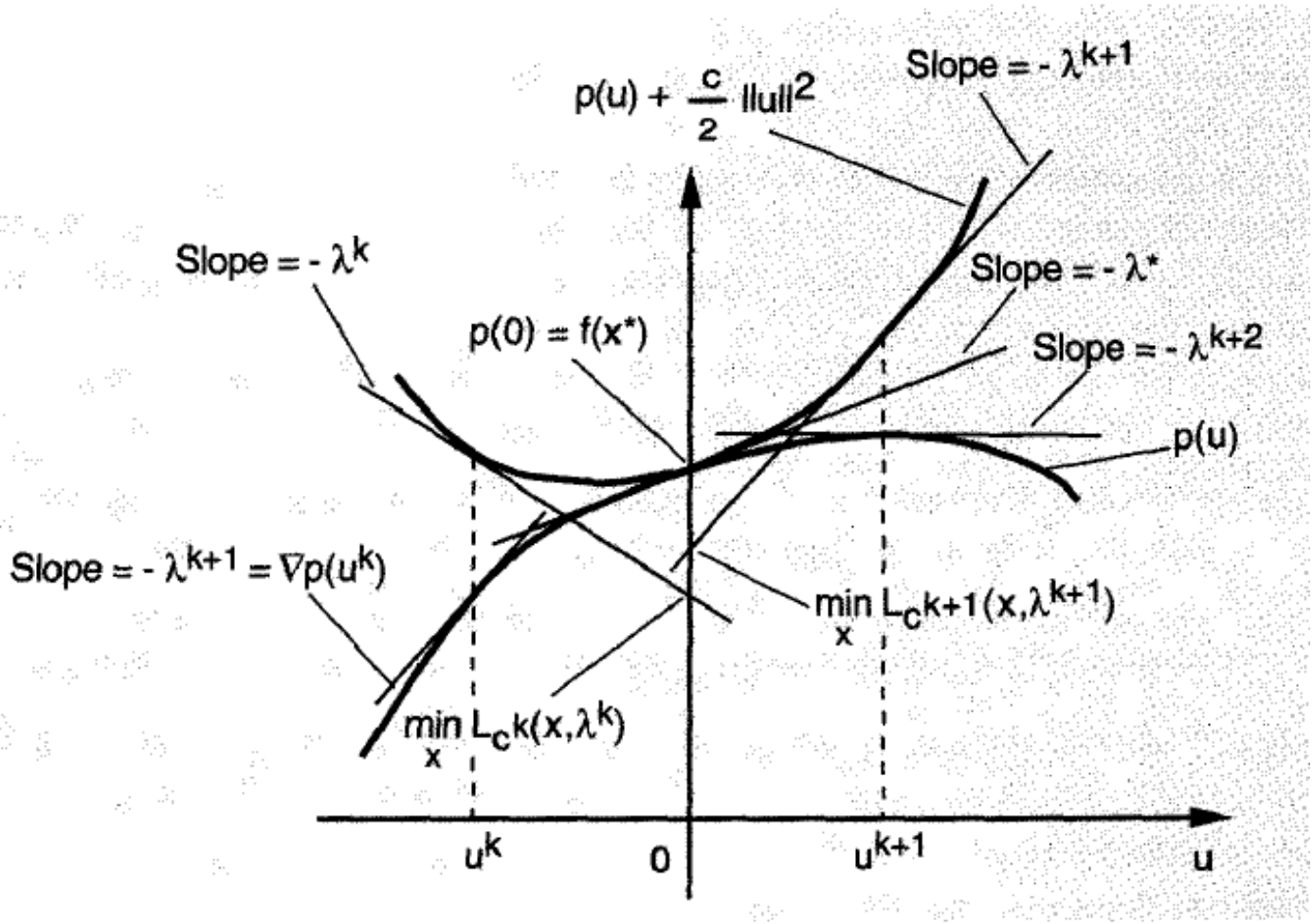

Figure 4.2.6. Geometric interpretation of the first order multiplier iteration.

Computational Aspects —Choice of Parameters

The main considerations to be kept in mind for selecting the penalty parameter sequence are the following:

(a) c^k should eventually become larger than the threshold level necessary to bring to bear the positive features of the multiplier iteration.

(b) The initial parameter c^0 should not be too large to the point where it causes ill-conditioning at the first unconstrained minimization.

(c) c^k should not be increased too fast to the point where too much ill-conditioning is forced upon the unconstrained minimization routine too early.

(d) c^k should not be increased too slowly, at least in the early minimizations, to the extent that the multiplier iteration has poor convergence rate.

Inexact Minimization of the Augmented Lagrangian

In practice the minimization of $L_{ck}(x, \lambda^k)$ is typically terminated early. For example, it may be terminated at a point x^k satisfying

$$\|\nabla_x L_{c^k}(x^k, \lambda^k)\| \leq \epsilon^k,$$

where $\{\epsilon^k\}$ is a positive sequence converging to zero. Then it is still appropriate to use the multiplier update

$$\lambda^{k+1} = \lambda^k + c^k h(x^k),$$

although in theory, some of the linear convergence rate results to be given shortly will not hold any more. This deficiency does not seem to be important in practice, but can also be corrected by using the alternative termination criterion

$$\|\nabla_x L_{\mathrm{c}}(x^k, \lambda^k)\| \leq \min\{\epsilon^k, \gamma^k \|h(x^k)\|\},$$

where $\{\epsilon^k\}$ and $\{\gamma^k\}$ are positive sequences converging to zero.

Inequality Constraints

To treat inequality constraints $g_j(x) \leq 0$ in the context of the method of multipliers, we convert them into equality constraints $g_j(x)+z_j^2 = 0$, using the additional variables z_j. In particular, the multiplier update formulas are

$$\begin{aligned} \lambda^{k+1} &= \lambda^k + c^k h(x^k), \\ \mu_j^{k+1} &= \max\{0, \mu_j^k + c^k g_j(x^k)\} \end{aligned} \quad ,$$

where x^k minimizes the augmented Lagrangian

$$\begin{aligned} L_{c^k}(x, \lambda^k, \mu^k) =& f(x) + \lambda^{k'} h(x) + \frac{c}{2}\|h(x)\|^2 \\ &+ \frac{1}{2c^k} \sum_{j=1}^{r} \{(\max\{0, \mu_j^k + c^k g_j(x)\})^2 - (\mu_j^k)^2\}. \end{aligned}$$

Partial Elimination of Constraints

The method of multipliers with partial elimination of constraints for the problem

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } h(x) = 0, \quad g(x) \leq 0 \end{aligned}$$

consists of finding x^k that solves the problem

$$\begin{aligned} &\text{minimize } f(x) + \lambda^{k'} h(x) + \tfrac{c}{2}\|h(x)\|^2 \\ &\text{subject to } g(x) \leq 0 \end{aligned}$$

followed by the multiplier iteration

$$\lambda^{k+1} = \lambda^k + c^k h(x^k).$$

Convergence Analysis of Multiplier Methods

We now discuss the convergence properties of multiplier methods and substantiate the conclusions derived informally earlier. We focus attention throughout on the equality constrained problem

$$\begin{array}{l}\text{minimize } f(x) \\ \text{subject to } h(x) = 0\end{array} \quad ,$$

and on a particular local minimum x^*, We assume that x^* is regular and together with a Lagrange multiplier vector λ^* satisfies the second order sufficiency conditions.

Existence of Local Minima of the Augmented Lagrangian

A first basic issue is whether local minima x^k of the augmented Lagrangian $L_{c^k}(\cdot, \lambda^k)$ exist, so that the method itself is well-define. We have shown that for the local minimum-Lagrange multiplier pair (x^*, λ^*) there exist scalars $\bar{c} > 0, \gamma > 0$, and $\epsilon > 0$, such that

$$L_c(x, \lambda^*) \geq L_c(x^*, \lambda^*) + \frac{\gamma}{2}\|x - x^*\|^2, \quad \forall x \text{ with } \|x - x^*\| < \epsilon, \text{ and } c \geq \bar{c}.$$

It is thus reasonable to infer that if λ is close to λ^*, there should exist a local minimum of $L_c(\cdot, \lambda)$ close to x^* for every $c \geq \bar{c}$. More precisely, for a fixed $c \geq \bar{c}$, we consider the system of equations

$$\nabla_x L_c(x, \lambda) = \nabla f(x) + \nabla h(x)(\lambda + ch(x)) = 0,$$

and we use the implicit function theorem in a neighborhood of (x^*, λ^*). Thus, for λ sufficiently close to λ^*, there is an unconstrained local minimum $x(\lambda, c)$ of $L_c(\cdot, \lambda)$, which is defined via the equation

$$\nabla f(x(\lambda, c)) + \nabla h(x(\lambda, c))(\lambda + ch(x(\lambda, c))) = 0.$$

Proposition

Let $\overline{c}$ be a positive scalar such that

$$\nabla^2_{xx} L_{\overline{c}}\left(x^*, \lambda^*\right) > 0.$$

There exist positive scalars δ, ϵ, and M such that:

(a) *For all (λ, c) in the set $D \subset \Re^{m+1}$ defined by*

$$D = \left\{(\lambda, \mathrm{c}) \middle| \, \|\lambda - \lambda^*\| < \delta \mathrm{c}, \overline{c} \leq \mathrm{c}\right\},$$

the problem

$$\begin{array}{l} \text{minimize } L_c(x, \lambda) \\ \text{subject to } \|x - x^*\| < \epsilon \end{array},$$

has a unique solution denoted $x(\lambda, c)$. The function $x(\cdot, \cdot)$ is continuously differentiable in the interior of D, and for all $(\lambda, c) \in D$, we have

$$\|x(\lambda, c) - x^*\| \leq M \frac{\|\lambda - \lambda^*\|}{\mathrm{c}}.$$

Proposition

(b) *For all* $(\lambda, c) \in D$, *we have*

$$\|\tilde{\lambda}(\lambda, c) - \lambda^*\| \leq M \frac{\|\lambda - \lambda^*\|}{c},$$

where

$$\tilde{\lambda}(\lambda, c) = \lambda + \mathrm{ch}(x(\lambda, c)).$$

(c) *For all* $(\lambda, c) \in D$, *the matrix* $\nabla^2_{xx} L_c(x(\lambda, c), \lambda)$ *is positive definite and the matrix* $\nabla h(x(\lambda, c))$ *has rank* m.

Convergence and Rate of Convergence

Above proposition yields both a convergence and a convergence rate result for the multiplier iteration

$$\lambda^{k+1} = \lambda^k + c^k h(x^k).$$

It shows that if the generated sequence $\{\lambda^k\}$ is bounded [this can be en- forced if necessary by leaving λ^k unchanged if $\lambda^k + c^k h\left(x^k\right)$ does not belong to a prespecified bounded open set known to contain λ^* , the penalty parameter c^k is sufficiently large after a certain index [so that $\left(\lambda^k, c^k\right) \in D$], and after that index, minimization of $L_{ck}\left(\cdot, \lambda^k\right)$ yields the local minimum $x^k = x\left(\lambda^k, c^k\right)$ closest to x^*, then we obtain $x^k \to x^*, \lambda^k \to \lambda^*$. Furthermore, the rate of convergence of the error sequences $\{\left\|x^k - x^*\right\|\}$and $\{\left\|\lambda^k - \lambda^*\right\|\}$ is linear, and it is superlinear if $c^k \to \infty$.

Let $\overline{c}, \delta$, and ϵ be as in above proposition, and define for (λ, c) in the set

$$D = \{(\lambda, c) | \, \|\lambda - \lambda^*\| < \delta c, \overline{c} \leq c\}$$

the dual function q_c by

$$q_c(\lambda) = \min_{\|x - x^*\| < \epsilon} L_c(x, \lambda) = L_c(x(\lambda, c), \lambda).$$

Since $x(\cdot, c)$ is continuously differentiable, the same is true forq_c. We compute the gradient of q_c with respect to λ

$$\nabla q_c(\lambda) = h(x(\lambda, c)),$$

and the Hessian

$$\nabla^2 q_c(\lambda) = -\nabla h(x(\lambda, c))' \left\{\nabla^2_{xx} L_c(x(\lambda, c), \lambda)\right\}^{-1} \nabla h(x(\lambda, c)).$$

When $c^k = c$ for all k, then

$$\lambda^{k+1} = \lambda^k + c \nabla q_c \left(\lambda^k\right).$$

The Second Order Method of Multipliers

In view of the interpretation of the multiplier iteration as a steepest ascent method, it is natural to consider the Newton-like iteration

$$\lambda^{k+1} = \lambda^k - (\nabla^2 q_{c^k}(\lambda^k))^{-1} \nabla q_{c^k}(\lambda^k),$$

for maximizing the dual function. Then,

$$\lambda^{k+1} = \lambda^k + (B^k)^{-1} h(x^k),$$

where

$$B^k = \nabla h(x^k)' \{\nabla^2_{xx} L_{c^k}(x^k, \lambda^k)\}^{-1} \nabla h(x^k)$$

and x^k minimizes $L_c(\cdot, \lambda^k)$.

An alternative form, which turns out to be more appropriate when the minimization of the augmented Lagrangian is inexact is given by

$$\lambda^{k+1} = \lambda^k + (B^k)^{-1}(h(x^k) - \nabla h(x^k)'(\nabla^2_{xx} L_{c^k}(x^k, \lambda^k))^{-1} \nabla_x L_{c^k}(x^k, \lambda^k)).$$

When $\nabla_x L_c k \left(x^k, \lambda^k\right) = 0$, the two forms are equivalent.

$$x^{k+1} = x^k - \left(\nabla^2_{xx} L_{c^k}\left(x^k, \lambda^k\right)\right)^{-1} \nabla_x L_{c^k}\left(x^k, \lambda^{k+1}\right)$$

The Exponential Method of Multipliers

Consider the problem

$$\begin{array}{l}\text{minimize } f(x)\\ \text{subject to } g_1(x) \leq 0, \ldots, g_r(x) \leq 0\end{array}.$$

We introduce a method of multipliers characterized by a twice differentiable penalty function $\psi : \Re$ with the following properties:

(i) $\nabla^2\psi(t) > 0$ for all $t \in \Re$

(ii) $\psi(0) = 0, \nabla\psi(0) = 1$

(iii) $\lim_{t\to-\infty} \psi(t) > -\infty$

(iv) $\lim_{t\to-\infty} \nabla\psi(t) = 0$ and $\lim_{t\to\infty} \nabla\psi(t) = \infty$

A simple and interesting special case is the exponential penalty function

$$\psi(t) = e^t - 1$$

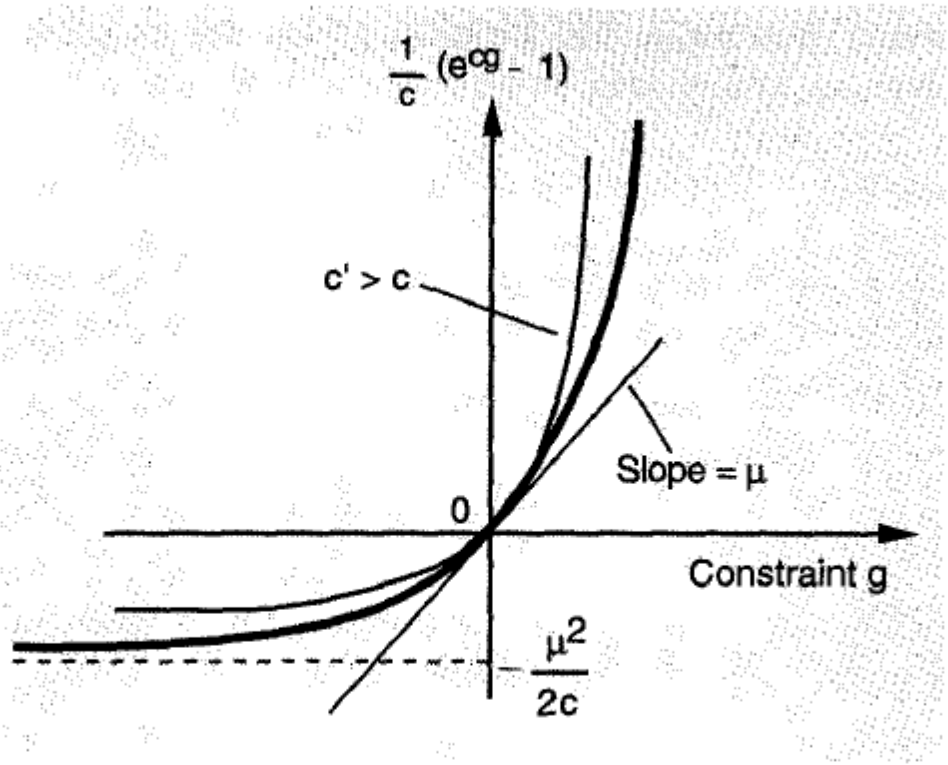

Figure 4.2.8. The penalty term of the exponential method of multipliers.

The method consists of the sequence of unconstrained minimizations

$$x^k \in \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \sum_{j=1}^{m} \frac{\mu_j^k}{c_j^k} \psi\left(c_j^k g_j(x)\right) \right\},$$

followed by the multiplier iterations

$$\mu_j^{k+1} = \mu_j^k \nabla \psi\left(c_j^k g_j\left(x^k\right)\right), \quad j = 1, \ldots, r.$$

Here $\left\{c_j^k\right\}$ is a positive penalty parameter sequence for each j, and the initial multipliers μ_j^0 are arbitrary positive numbers.

Another interesting method, known as the modified barrier method, is based on the following modified version of the logarithmic barrier function

$$\psi(t) = -\ln(1-t)$$

$$\mu_j^{k+1} = \frac{\mu_j^k}{1 - c_j^k g_j\left(x^k\right)}, \quad j = 1, \ldots, r.$$

This method is not really a special case of the generic method because the penalty function ψ is defined only on the set $(-\infty, 1)$, but it shares the same qualitative characteristics as the generic method.

Exact Penalties-Sequential Quadratic Programming

Consider the equality constrained problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_i(x) = 0, \quad i = 1, \ldots, m, \end{array}$$

where f and h_i are continuously differentiable, and let

$$L(x, \lambda) = f(x) + \lambda' h(x)$$

be the corresponding Lagrangian function. Then by minimizing the function

$$P(x, \lambda) = \|\nabla_x L(x, \lambda)\|^2 + \|h(x)\|^2$$

over $(x, \lambda) \in \Re^{n+m}$ we can obtain local minima-Lagrange multiplier pairs (x^*, λ^*) satisfying the first order necessary conditions

$$\nabla_x L\left(x^*, \lambda^*\right) = 0, \quad h\left(x^*\right) = 0.$$

We may view $P(x, \lambda)$ as an exact penalty function, that is, a function whose unconstrained minima are (or strongly relate to) optimal solutions and/or Lagrange multipliers of a constrained problem.

Nondifferentiable Exact Penalty Functions

We will show that solutions of the equality constrained problem are related to solutions of the (nondifferentiable) unconstrained problem

$$\begin{array}{l} \text{minimize } f(x) + cP(x) \\ \text{subject to } x \in \Re^n \end{array},$$

where $c > 0$ and P is the nondifferentiable penalty function defined by

$$P(x) = \max_{i=1\ldots m} |h_i(x)| .$$

Consider also the primal function p defined in a neighborhood of the origin by

$$p(u) = \min \{ f(x) | h(x) = u, \|x - x^*\| < \epsilon \},$$

where $\epsilon > 0$ is some scalar; see the sensitivity theorem. Then if we locally minimize $f + \mathrm{cP}$ around x^*, we can split the minimization in two: first minimize over all x satisfying $h(x) = u$ and then minimize over all possible u. We have

$$\begin{aligned}
&\inf_{\|x-x^*\|<\epsilon} \left\{ f(x) + c \max_{i=1,\ldots,m} |h_i(x)| \right\} \\
&= \inf_{u \in U_\epsilon} \sum_{\{x | h(x)=u, \|x-x^*\|<\epsilon\}} \{ f(x) + c_{i=1,\ldots,m} |h_i(x)| \} \\
&= \inf_{u \in U_\epsilon} p_c(u),
\end{aligned}$$

where

$$U_\epsilon = \{u | h(x) = u \text{ for some } x \text{ with } \|x - x^*\| < \epsilon\},$$

and

$$p_c(u) = p(u) + c \max_{i=1,\ldots,m} |u_i|.$$

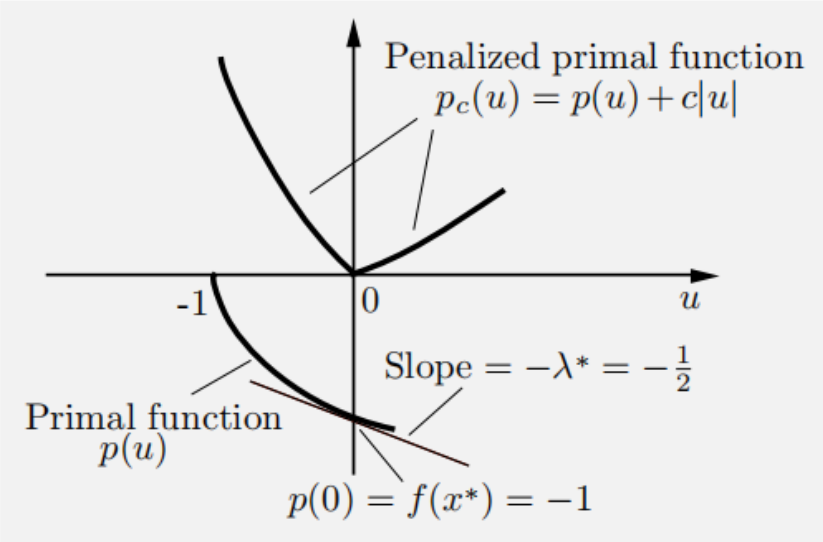

Figure 5.3.2. Illustration of how for c large enough, $u = 0$ is a strict local minimum of $p_c(u) = p(u) + c \max_{i=1,\ldots,m} |u_i|$. The figure corresponds to the two-dimensional problem where $f(x) = x_1$ and $h(x) = x_1^2 + x_2^2 - 1$. The optimal solution and Lagrange multiplier are $x^* = (-1, 0)$ and $\lambda^* = 1/2$, respectively. The primal function is defined for $u \geq -1$ and is given by

$$p(u) = \min_{x_1^2 + x_2^2 - 1 = u} x_1 = -\sqrt{1+u}.$$

Note that $\nabla p(0) = \lambda^*$ and that we must have $c > \lambda^*$ in order for the nondifferentiable penalty function to be exact.

The above argument can be extended to the case where there are additional inequality constraints of the form $g_j(x) \leq 0, j = 1, \ldots, r$.

$$f(x) + c \max \{0, g_1(x), \ldots, g_r(x), |h_1(x)|, \ldots, |h_m(x)|\}.$$

Proposition

Let x^* be a local minimum of the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to } h_i(x) = 0, & i = 1, \ldots, m, \quad g_j(x) \leq 0, \quad j = 1, \ldots, r \end{array},$$

which is regular and satisties together with corresponding Lagrange multiplier vectors λ^* and μ^*, the second order sufficiency conditions. Then, if

$$c > \sum_{i=1}^{m} |\lambda_i^*| + \sum_{j=1}^{r} \mu_j^*,$$

the vector x^* is a strict unconstrained local minimum of $f + cP$, where

$$P(x) = \max\{0, g_1(x), \ldots, g_r(x), |h_1(x)|, \ldots, |h_m(x)|\}.$$

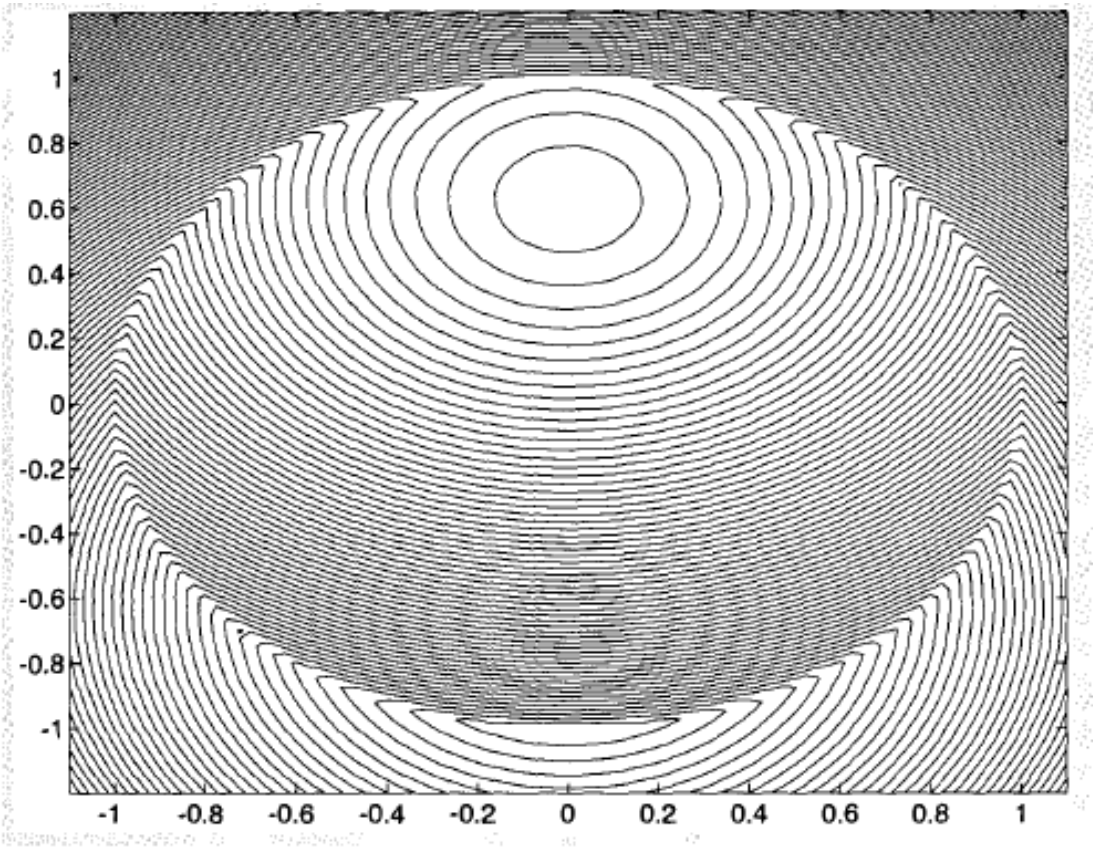

Figure 4.3.2. Contours of the function $f + cP$ for the two-dimensional problem where

$$f(x) = x_1, \qquad h(x) = x_1^2 + x_2^2 - 1$$

(cf. Fig. 4.3.1). For c greater than the Lagrange multiplier $\lambda^* = 1/2$, the optimal solution $x^* = (-1, 0)$ is a local minimum of $f + cP$. This is not so for $c < \lambda^*$. The figure corresponds to $c = 0.8$.

Descent Directions of Exact Penalties

We will develop an algorithm for minimizing $f + \mathrm{cP}$, where $\mathrm{c} > 0$ and

$$P(x) = \max\{g_0(x), g_1(x), \ldots, g_r(x)\}, \quad \forall x \in \Re^n$$
$$g_0(x) = 0, \quad \forall x \in \Re^n,$$

We first introduce some definitions and develop some preliminary results. For $x \in \Re^n, d \in \Re^n$, and $c > 0$, we introduce the index set

$$J(x) = \{j | g_j(x) = P(x), j = 0, 1, \ldots, r\},$$

and we denote

$$\theta_c(x; d) = \max\{\nabla f(x)'d + \mathrm{c}\nabla g_j(x)'d | j \in J(x)\}.$$

The function θ_c plays the role that the gradient would play if the function $f + cP$ were differentiable. In particular,

$$f(x) + \mathrm{c}P(x) + \theta_c(x; d)$$

may be viewed as a linear approximation of $f + \mathrm{cP}$ for variations d around x.

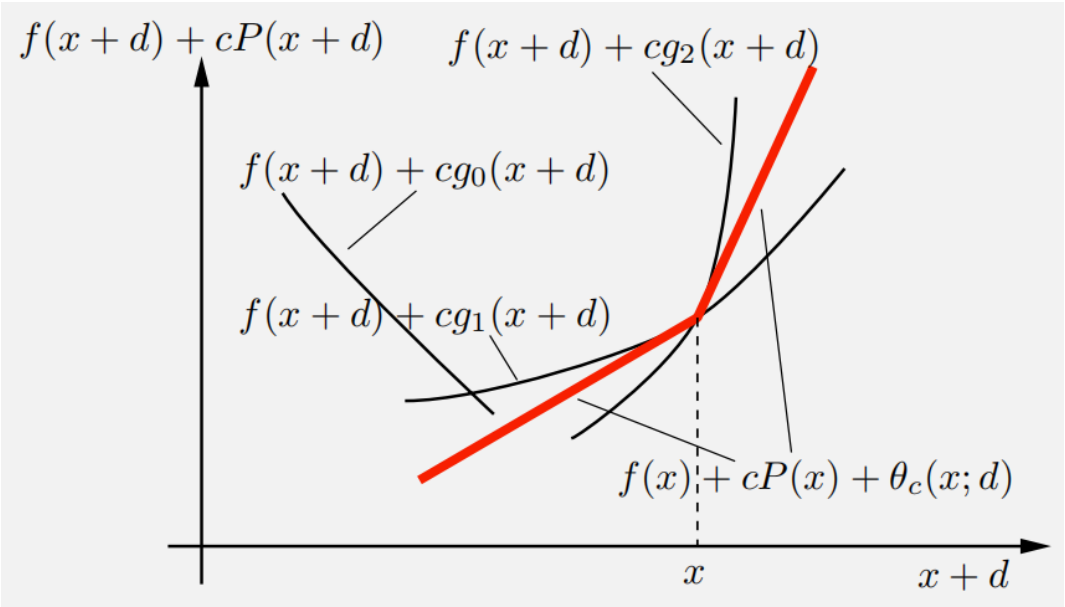

Figure: Illustration of $\theta_c(x;d)$ at x. It is the first order estimate of the variation $f(x+d)+cP(x+d)-f(x)-cP(x)$ of $f+cP$ around x. Here $J(x)$ is $\{1,2\}$.

Since at an unconstrained local minimum x^*, $f+cP$ cannot decrease along any direction, the preceding interpretation of θ_c motivates us to call a vector $x*$ a stationary point of $f+cP$ if for all $d \in \Re^n$ there holds

$$\theta_c\left(x^*; d\right) \geq 0.$$

Such directions can be obtained from the fllowing (convex) quadratic program, in $(d, \xi) \in \Re^{n+1}$,

$$\begin{aligned} &\text{minimize } \nabla f(x)'d + \tfrac{1}{2}d'Hd + c\xi \\ &\text{subject to } (d,\xi) \in \Re^{n+1}, \quad g_j(x) + \nabla g_j(x)'d \leq \xi, \quad j = 0, 1, \ldots, r, \end{aligned}$$

where $c > 0$ and H is a positive definite symmetric matrix. Note that for a fixed d, the minimum with respect to ξ is obtained at

$$\xi = \max_{j=0,1,\ldots,r} \left\{g_j(x) + \nabla g_j(x)'d\right\},$$

alternative form:

$$\min_{d\in\mathbb{R}^n} \quad \max_{j=0,1,\cdots,r} \{f(x) + cg_j(x) + \nabla f(x)'d + c\nabla g_j(x)'d\} + \frac{1}{2}d'Hd,$$

or

$$\min_{d\in\mathbb{R}^n} \quad f(x) + cP(x) + \theta_c(x;d) + \frac{1}{2}d'Hd,$$

for small $\|d\|$.

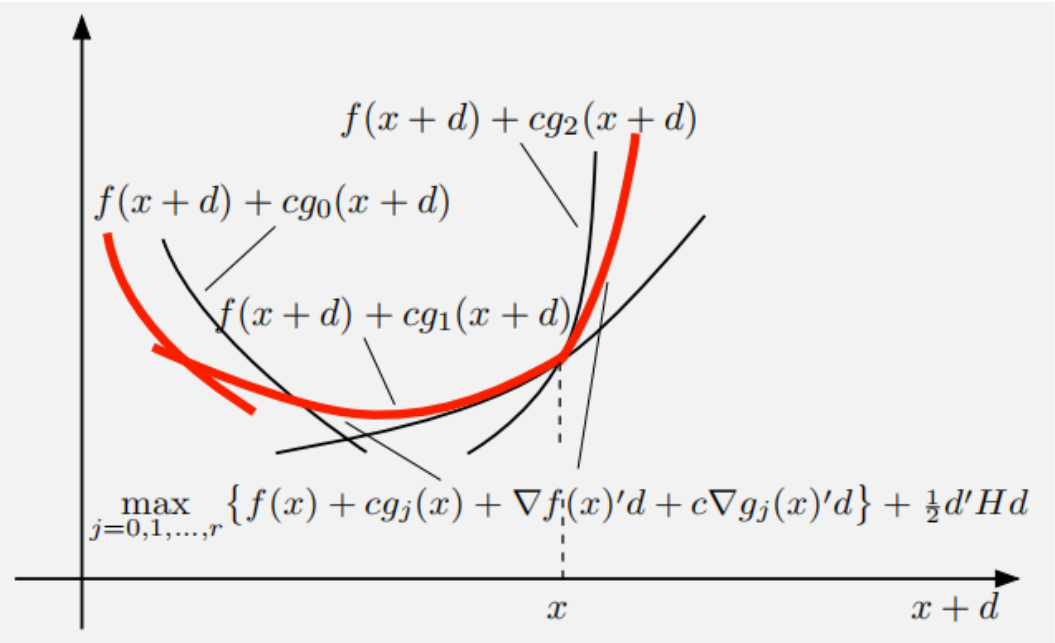

Figure 5.3.5. Illustration of the cost function

$$\max_{j=0,1,\ldots,r}\left\{f(x)+cg_j(x)+\nabla f(x)'d+c\nabla g_j(x)'d\right\}+\tfrac{1}{2}d'Hd$$

of the quadratic program (5.82). For small $\|d\|$ this function takes the form

$$f(x)+cP(x)+\theta_c(x;d)+\tfrac{1}{2}d'Hd,$$

and is a quadratic approximation of $f+cP$ around x. It can be seen that by minimizing this function over d we obtain a direction of descent of $f+cP$ at x.

Proposition

(a) *A local minimum of* $f + cP$ *is a stationary point. Furthermore, for all* $x \in \Re^n, d \in \Re^n$, *and* $\alpha > 0$,

$$f(x + \alpha d) + cP(x + \alpha d) - f(x) - cP(x) = \alpha\theta_c(x; d) + o(\alpha),$$

where $\lim_{\alpha \to 0+} o(\alpha)/\alpha = 0$. *As a result, if* $\theta_c(x; d) < 0$, *then* d *is a descent direction; that is, there exists* $\overline{\alpha} > 0$ *such that*

$$f(x + \alpha d) + cP(x + \alpha d) < f(x) + cP(x), \quad \forall \alpha \in (0, \overline{\alpha}).$$

(b) *If* f *and* g_j *are convex functions, then a stationary point* x^* *of* $f + cP$ *is also a global minimum of* $f + cP$.

(c) *For any* $x \in \Re^n$ *and positive definite symmetric* H, *if* (d, ξ) *is the optimal solution of the quadratic program* (4.53), *then*

$$\theta_c(x; d) \leq -d' H d.$$

(d) *A vector* x *is a stationary point of* $f + cP$ *if and only if the quadratic program* (4.53) *has* $\{d = 0, \xi = P(x)\}$ *as its optimal solution.*

Proof: (a) We have for all $\alpha > 0$ and $j \in J(x)$

$$f(x+\alpha d) + cg_j(x+\alpha d) = f(x) + \alpha \nabla f(x)'d + c\left(g_j(x) + \alpha \nabla g_i(x)'d\right) + o_j(\alpha),$$

where $\lim_{\alpha \to 0+} o_j(\alpha)/\alpha = 0$. Hence, by using the fact $g_j(x) = P(x)$ for all $j \in J(x)$,

$$\begin{aligned} & f(x+\alpha d) + c \max\{g_j(x+\alpha d) | j \in J(x)\} \\ & \quad = f(x) + \alpha \nabla f(x)'d + c \max\{g_j(x) + \alpha \nabla g_j(x)'d | j \in J(x)\} + o(\alpha) \quad , \\ & \quad = f(x) + cP(x) + \alpha \theta_c(x;d) + o(\alpha) \end{aligned}$$

where $\lim_{\alpha \to 0+} o(\alpha)/\alpha = 0$. We have, for all α that are sufficiently small,

$$\begin{aligned} \max\{g_j(x+\alpha d) | j \in J(x)\} &= \max\{g_j(x+\alpha d) | j = 0, 1, \ldots, r\} \\ &= P(x+\alpha d). \end{aligned}$$

Combining the two above relations, we obtain

$$f(x+\alpha d) + cP(x+\alpha d) = f(x) + cP(x) + \alpha \theta_c(x;d) + o(\alpha).$$

If x^* is a local minimum of $f + cP$, then, for all d and $\alpha > 0$ such that $\|d\|$ and α are sufficiently small,

$$\alpha\theta_c\left(x^*; d\right) + o(\alpha) \geq 0.$$

Dividing by α and taking the limit as $\alpha \to 0$, we obtain $\theta_c\left(x^*; d\right) \geq 0$, so x^* is stationary.

(b) By convexity, we have for all j and $x \in \Re^n$,

$$f(x) + cg_j(x) \geq f\left(x^*\right) + cg_j\left(x^*\right) + \left(\nabla f\left(x^*\right) + c\nabla g_j\left(x^*\right)\right)'\left(x - x^*\right).$$

Taking the maximum over j, we obtain

$$f(x)+cP(x) \geq \max_{j=0,1,\ldots,r}\left\{f\left(x^*\right) + cg_j\left(x^*\right) + \left(\nabla f\left(x^*\right) + c\nabla g_j\left(x^*\right)\right)'\left(x - x^*\right)\right\}$$

For a sufficiently small scalar ϵ and for all x with $\|x - x^*\| < \epsilon$, the maximum above is attained for some $j \in J\left(x^*\right)$. since $g_j\left(x^*\right) = P\left(x^*\right)$ for all $j \in J\left(x^*\right)$, we obtain for all x with $\|x - x^*\| < \epsilon$,

$$f(x) + cP(x) \geq f\left(x^*\right) + cP\left(x^*\right) + \theta_c\left(x^*; x - x^*\right) \geq f\left(x^*\right) + cP\left(x^*\right)$$

where the last inequality holds because x^* is a stationary point of $f+cP$. Hence $x*$ is a local minimum of the function $f+cP$, and in view of the convexity of $f+cP, x^*$ is a global minimum.

(c) We have $g_j(x)+\nabla g_j(x)'d \leq \xi$ for all j. since $g_j(x) = P(x)$ for all $j \in J(x)$, it follows that $\nabla g_j(x)'d \leq \xi - P(x)$ for all $j \in J(x)$ and therefore using the definition of θ_c we have

$$\theta_c(x;d) \leq \nabla f(x)'d + c(\xi - P(x)).$$

Let $\{\mu_j\}$ be a set of Lagrange multipliers for the quadratic program. The optimality conditions yield

$$\nabla f(x) + Hd + \sum_{j=0}^{r} \mu_j \nabla g_j(x) = 0$$

$$c - \sum_{j=0}^{r} \mu_j = 0$$

$$g_j(x) + \nabla g_j(x)'d \leq \xi, \quad \mu_j \geq 0, \quad j = 0, 1, \ldots, r$$

$$\mu_j \left(g_j(x) + \nabla g_j(x)'d - \xi\right) = 0, \quad j = 0, 1, \ldots, r.$$

By adding the last equation over all j, we have

$$\begin{aligned}\sum_{j=0}^{r}\mu_j\nabla g_j(x)'d &= \sum_{j=0}^{r}\mu_j\xi - \sum_{j=0}^{r}\mu_j g_j(x)\\ &\geq \sum_{j=0}^{r}\mu_j\left(\xi - \max_{m=0,1,\ldots,r} g_m(x)\right)\\ &= \sum_{j=0}^{r}\mu_j(\xi - P(x))\\ &= \mathrm{c}(\xi - P(x))\end{aligned}$$

We obtain

$$\nabla f(x)'d + d'Hd + c(\xi - P(x)) \leq 0,$$

then,

$$\theta_c(x;d) + d'Hd \leq 0.$$

(d) x is stationary if and only if $\theta_c(x;d) \geq 0$ for all d, which is true if and only if $\{d = 0, \xi = P(x)\}$ is the optimal solution of the quadratic program.

The Linearization Algorithm

We now introduce an iterative descent algorithm for minimizing the exact penalty function $f + cP$. It is called the linearization algorithm or sequential quadratic programming, and like the gradient projection method, it calculates the descent direction by solving a quadratic programming subproblem. It is given by

$$x^{k+1} = x^k + \alpha^k d^k,$$

where α^k is a nonnegative scalar stepsize, and d^k is a direction obtained by solving the quadratic program in (d, ξ)

$$\begin{aligned} &\text{minimize } \nabla f\left(x^k\right)' d + \tfrac{1}{2} d' H^k d + c\xi \\ &\text{subject to } g_j\left(x^k\right) + \nabla g_j\left(x^k\right)' d \leq \xi, \quad j = 0, 1, \ldots, r \end{aligned}$$

The initial vector x^0 is arbitrary and the stepsize α^k is chosen by any one of the stepsize rules listed below:

(a) Minimization rule: Here α^k is chosen so that

$$f\left(x^k+\alpha^k d^k\right)+cP\left(x^k+\alpha^k d^k\right)=\min_{\alpha\geq 0}\left\{f\left(x^k+\alpha d^k\right)+cP\left(x^k+\alpha d^k\right)\right\}$$

(b) Limited minimization rule: Here a fixed scalar $s>0$ is selected and α^k is chosen so that

$$f\left(x^k+\alpha^k d^k\right)=\min_{\alpha\in[0,s]} f\left(x^k+\alpha d^k\right)$$

(c) Armijo rule: Here fixed scalars s,β, and σ with $s>0,\beta\in(0,1)$, and $\sigma\in\left(0,\frac{1}{2}\right)$, are selected, and we set $\alpha^k=\beta^{m_k}s$, where m_k is the first nonnegative integer m for which

$$f\left(x^k\right)+cP\left(x^k\right)-f\left(x^k+\beta^m s d^k\right)-cP\left(x^k+\beta^m s d^k\right)\geq\sigma\beta^m s d^{k\prime}H^k d^k.$$

Proposition

Let $\{x^k\}$ be a sequence generated by the linearization algorithm, where the stepsize α^k is chosen by the minimization rule, the limited minimization rule, or the Armijo rule. Assume that there exist positive scalars γ and Γ such that

$$\gamma\|z\|^2 \leq z'H^k z \leq \Gamma\|z\|^2, \quad \forall z \in \Re^n, \quad k = 0, 1, \ldots,$$

(this condition corresponds to the assumption of a gradient-related direction sequence in unconstrained optimization). Then every limit point of $\{x^k\}$ is a stationary point of $f + cP$.

Proof: We argue by contradiction. Assume that a subsequence $\{x^k\}_K$ generated by the algorithm using the Armijo rule converges to a vector $\overline{x}$ that is not a stationary point of $f + cP$. since $f\left(x^k\right) + cP\left(x^k\right)$ is monotonically decreasing, we have

$$f\left(x^k\right) + cP\left(x^k\right) \to f(\overline{x}) + cP(\overline{x})$$

and hence also

$$f\left(x^k\right) + cP\left(x^k\right) - f\left(x^{k+1}\right) - cP\left(x^{k+1}\right) \to 0.$$

By the definition of the Armijo rule, we have

$$f\left(x^k\right)+cP\left(x^k\right)-f\left(x^{k+1}\right)-cP\left(x^{k+1}\right)\geq\sigma\alpha^k d^{k'}H^k d^k.$$

Hence,

$$\alpha^k d^{k'}H^k d^k\to 0.$$

Since for $k\in K, d^k$ is the optimal solution of the quadratic program, we must have for some set of Lagrange multipliers $\left\{\mu_j^k\right\}$ and all $k\in K$,

$$\nabla f\left(x^k\right)+\sum_{j=0}^{r}\mu_j^k\nabla g_j\left(x^k\right)+H^k d^k=0,\quad c=\sum_{j=0}^{r}\mu_j^k,$$

$$\mu_j^k\geq 0,\quad \mu_j^k\left(g_j\left(x^k\right)+\nabla g_j\left(x^k\right)'d^k-\xi^k\right)=0,\quad j=0,1,\ldots,r,$$

where

$$\xi^k=\max_{j=0,1,\ldots,r}\left\{g_j\left(x^k\right)+\nabla g_j\left(x^k\right)'d^k\right\}.$$

The relations $c=\sum_{j=0}^{r}\mu_j^k$ and $\mu_j^k\geq 0$ imply that the subsequences $\left\{\mu_j^k\right\}$ are bounded. Hence, without loss of generality, we may assume that for some $\mu_j, j=0,1,\ldots,r$, we have

$$\left\{\mu_j^k\right\}_K \to \overline{\mu}_j, \quad j = 0, 1, \ldots, r.$$

Using the assumption $\gamma\|z\|^2 \leq z' H^k z \leq \Gamma\|z\|^2$, we may also assume without loss of generality that

$$\left\{H^k\right\}_K \to \overline{H}.$$

for some positive definite matrix $\overline{H}$.

Now from the fact $\alpha^k d^{k'} H^k d^k \to 0$, it follows that there are two possibilities. Either

$$\liminf_{k\to\infty, k\in K} \left\|d^k\right\| = 0,$$

or else

$$\liminf_{k\to\infty, k\in K} \alpha^k = 0, \quad \liminf_{k\to\infty, k\in K} \left\|d^k\right\| > 0.$$

If first equality holds, then we may assume without loss of generality that $\left\{d^k\right\}_K \to 0$, we have

$$\nabla f(\overline{x}) + \sum_{j=0}^{r} \overline{\mu}_j \nabla g_j(\overline{x}) = 0, \quad c = \sum_{j=0}^{r} \overline{\mu}_j$$

$$\overline{\mu}_j \geq 0, \quad \overline{\mu}_j \left(g_j(\overline{x}) - \xi\right) = 0, \quad j = 0, 1, \ldots, r.$$

where $\xi = \max_{j=0,1,\ldots,r} g_j(\overline{x})$. Hence the quadratic program corresponding to $\overline{x}$ has $\{d = 0, \xi = P(\overline{x})\}$ as its optimal solution. It follows that $\overline{x}$ is a stationary point of $f + cP$, thus contradicting the hypothesis made earlier.

It will thus suffice to arrive at a contradiction assuming that second inequality holds. We may assume without loss of generality that

$$\left\{\alpha^k\right\}_K \to 0.$$

Then, $\left\{d^k\right\}_K$ is a bounded sequence, we may also assume without loss of generality that

$$\left\{d^k\right\}_K \to \overline{d},$$

where $\overline{d}$ is some vector which cannot be zero.

Since $\left\{\alpha^k\right\}_K \to 0$, it follows, in view of the definition of the Armijo rule, that the initial stepsize s will be reduced at least once for all $k \in K$ after some index $\overline{k}$. This means that for all $k \in K, k \geq \overline{k}$,

$$f\left(x^k\right) + cP\left(x^k\right) - f\left(x^k + \overline{\alpha}^k d^k\right) - cP\left(x^k + \overline{\alpha}^k d^k\right) < \sigma \overline{\alpha}^k d^{k'} H^k d^k,$$

where $\overline{\alpha}^k = \alpha^k / \beta$.
Define for all k and d

$$\zeta^k(d) = \nabla f\left(x^k\right)' d + c \max_{j \in J\left(x^k\right)} \left\{g_j\left(x^k\right) + \nabla g_j\left(x^k\right)' d\right\} - cP\left(x^k\right),$$

and restrict attention to $k \in K, k \geq \overline{k}$, that are sufficiently large so that $\overline{\alpha}^k \leq 1, J\left(x^k\right) \subset J(\overline{x})$, and $J\left(x^k + \overline{\alpha}^k d^k\right) \subset J(\overline{x})$. It will be shown that

$$f\left(x^k\right) + cP\left(x^k\right) - f\left(x^k + \overline{\alpha}^k d^k\right) - cP\left(x^k + \overline{\alpha}^k d^k\right) = -\zeta^k\left(\overline{\alpha}^k d^k\right) + o\left(\overline{\alpha}^k\right),$$

where

$$\lim_{k \to \infty, k \in K} \frac{o\left(\overline{\alpha}^k\right)}{\overline{\alpha}^k} = 0.$$

Indeed, we have

$$f\left(x^k+\overline{\alpha}^k d^k\right)=f\left(x^k\right)+\overline{\alpha}^k\nabla f\left(x^k\right)' d^k+o_0\left(\overline{\alpha}^k\left\|d^k\right\|\right)$$
$$g_j\left(x^k+\overline{\alpha}^k d^k\right)=g_j\left(x^k\right)+\overline{\alpha}^k\nabla g_j\left(x^k\right)' d^k+o_j\left(\overline{\alpha}^k\left\|d^k\right\|\right),\quad j\in J\left(x^k\right)$$

where $o_j(\cdot)$ are functions satisfying $\lim_{k\to\infty} o_j\left(\overline{\alpha}^k\left\|d^k\right\|\right)/\overline{\alpha}^k=0$. Adding and taking the maximum over $j\in J(x)$, and using the fact $J\left(x^k+\overline{\alpha}^k d^k\right)\subset J(\overline{x})$ [implying that $P\left(x^k+\overline{\alpha}^k d^k\right)=\max_j g_j\left(x^k+\overline{\alpha}^k d^k\right)$],] we obtain for sufficiently large k,

$$\begin{aligned}f\left(x^k+\overline{\alpha}^k d^k\right)+cP\left(x^k+\overline{\alpha}^k d^k\right)&=f\left(x^k\right)+\overline{\alpha}^k\nabla f\left(x^k\right)' d^k\\&+c\max_{j\in J\left(x^k\right)}\left\{g_j\left(x^k\right)+\overline{\alpha}^k\nabla g_j\left(x^k\right)' d^k\right\}+o\left(\overline{\alpha}^k\left\|d^k\right\|\right)\\&=f\left(x^k\right)+cP\left(x^k\right)+\zeta^k\left(\overline{\alpha}^k d^k\right)+o\left(\overline{\alpha}^k\right).\end{aligned}$$

We also claim that

$$-\frac{\zeta^k\left(\overline{\alpha}^k d^k\right)}{\tilde{\alpha}^k}\geq-\zeta^k\left(d^k\right)\geq d^{k\prime}H^k d^k.$$

Indeed, let $\left(d^k,\xi^k\right)$ be the optimal solution of the quadratic program

$$\begin{aligned}&\text{minimize } \nabla f\left(x^k\right)' d + \tfrac{1}{2} d' H^k d + c\xi \\ &\text{subject to } g_j\left(x^k\right) + \nabla g_j\left(x^k\right)' d \leq \xi, \quad j = 0, 1, \ldots, r\end{aligned}$$

We have

$$\xi^k = \max_{j=0,1,\ldots,r} \left\{ g_j\left(x^k\right) + \nabla g_j\left(x^k\right)' d^k \right\} \geq \max_{j \in J(x)} \left\{ g_j\left(x^k\right) + \nabla g_j\left(x^k\right)' d^k \right\}$$
$$= \frac{\zeta^k\left(d^k\right) - \nabla f\left(x^k\right)' d^k}{c} + P\left(x^k\right).$$

On the other hand,

$$c\left(\xi^k - P\left(x^k\right)\right) - \nabla f\left(x^k\right)' d^k \geq d^{k'} H^k d^k.$$

The last two equations, together with the relation $\zeta^k\left(\overline{\alpha}^k d\right) \leq \overline{\alpha}^k \zeta^k(d)$ which follows from the convexity of $\zeta^k(\cdot)$.

We obtain

$$\sigma d^{k'} H^k d^k > -\frac{\zeta^k\left(\overline{\alpha}^k d\right)}{\overline{\alpha}^k} + \frac{o\left(\overline{\alpha}^k\right)}{\overline{\alpha}^k},$$

yields

$$(1-\sigma)d^{k'}H^k d^k + \frac{o\left(\overline{\alpha}^k\right)}{\overline{\alpha}^k} < 0.$$

Since $\left\{H^k\right\}_K \to \overline{H}, \left\{d^k\right\}_K \to \overline{d}, \vec{H}$ is positive definite, $\overline{d} \neq 0$, and $o\left(\overline{\alpha}^k\right)/\overline{\alpha}^k \to 0[\mathrm{cf}$. We obtain a contradiction. This completes the proof of the proposition for the case of the Armijo rule.

Consider now the minimization rule and let $\left\{x^k\right\}_K$ converge to a vector $\overline{x}$, which is not a stationary point of $f+cP$. Let $\tilde{x}^{k+1}$ be the point that would be generated from x^k via the Armijo rule and let $\tilde{\alpha}^k$ be the corresponding stepsize. We have

$$f\left(x^k\right) - f\left(x^{k+1}\right) \geq f\left(x^k\right) - f\left(\tilde{x}^{k+1}\right) \geq \sigma\tilde{\alpha}^k d^{k'}H^k d^k.$$

By replacing α^k by $\tilde{\alpha}^k$ in the arguments of the earlier proof, we obtain a contradiction. This line of argument establishes that any stepsize rule that gives a larger reduction in the value of $f+cP$ at each iteration than the Armijo rule inherits its convergence properties, so it also proves the proposition for the limited minimization rule.

Application to Constrained Optimization Problems

Given the inequality constrained problem

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } g_j(x) \leq 0, \quad j = 1, \ldots, r \end{aligned}$$

one can attempt its solution by using the linearization algorithm to minimize the corresponding exact penalty function $f + cP$ for a value of c that exceeds the threshold $\sum_{j=1}^{r} \mu_j^*$.
The most common approach here is based on trying to solve the quadratic program

$$\begin{aligned} &\text{minimize } \nabla f\left(x^k\right)' d + \tfrac{1}{2} d' H^k d \\ &\text{subject to } g_j\left(x^k\right) + \nabla g_j\left(x^k\right)' d \leq 0, \quad j = 1, \ldots, r \end{aligned}$$

which differs from the direction finding quadratic program

$$\begin{aligned} &\text{minimize } \nabla f\left(x^k\right)' d + \tfrac{1}{2} d' H^k d + c\xi \\ &\text{subject to } g_j\left(x^k\right) + \nabla g_j\left(x^k\right)' d \leq \xi, \quad j = 0, 1, \ldots, r, \end{aligned}$$

of the linearization method in that ξ has been set to zero.

Extension to Equality Constraints

The development given earlier for inequality constraints can be extended to the case of additional equality constraints simply by converting each equality constraint $h_i(x) = 0$ to the two inequalities

$$h_i(x) \leq 0, \quad -h_i(x) \leq 0.$$

For example, the direction finding quadratic program of the linearization method is

$$\begin{aligned} &\text{minimize } \nabla f\left(x^k\right)' d + \frac{1}{2} d' H^k d + c\xi \\ &\text{subject to } g_j\left(x^k\right) + \nabla g_j\left(x^k\right)' d \leq \xi, \quad j = 0, 1, \ldots, r \\ &\left| h_i\left(x^k\right) + \nabla h_i\left(x^k\right)' d \right| \leq \xi, \quad i = 1, \ldots, m. \end{aligned}$$

Differentiable Exact Penalty Functions

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & h_i(x) = 0, \quad i = 1, \ldots, m, \end{aligned}$$

where f and h_i are twice continuously differentiable. Assume that the matrix $\nabla h(x)$ has rank m for all x, although much of the following analysis can be conducted assuming $\nabla h(x)$ has rank m in a suitable open subset of $\Re^n$. Motivated by the exact penalty function

$$\|\nabla_x L(x,\lambda)\|^2 + \|h(x)\|^2$$

discussed earlier, we consider the function

$$P_c(x,\lambda) = L(x,\lambda) + \frac{1}{2}\|W(x)\nabla_x L(x,\lambda)\|^2 + \frac{c}{2}\|h(x)\|^2,$$

where

$$L(x,\lambda) = f(x) + \lambda' h(x),$$

and $W(x)$ is any continuously differentiable $m \times n$ matrix such the $m \times m$ matrix $W(x)\nabla h(x)$ is nonsingular for all x.

The use of the matrix function $W(x)$ cannot be motivated easily, but will be justified by subsequent developments. Two examples of choices of $W(x)$ that turn out to be useful are

$$W(x) = \nabla h(x)'$$
$$W(x) = \left(\nabla h(x)' \nabla h(x)\right)^{-1} \nabla h(x)'$$

Proposition

For every compact subset $X \times \Lambda$ of $\Re^{n+m}$ there exists a $\bar{c} > 0$ such that for all $c \geq \bar{c}$, every stationary point (x^*, λ^*) of P_c that belongs to $X \times \Lambda$ satisfies the first order necessary conditions

$$\nabla_x L\left(x^*, \lambda^*\right) = 0, \quad \nabla_\lambda L\left(x^*, \lambda^*\right) = 0.$$

Differentiable Exact Penalty Functions Depending Only on x

One approach to minimizing $P_c(x,\lambda)$ is to first minimize it with respect to λ and then minimize it with respect to x. To simplify the subsequent formulas, let us focus on the function

$$W(x) = \left(\nabla h(x)' \nabla h(x)\right)^{-1} \nabla h(x)'.$$

For this function, $W(x)\nabla h(x)$ is equal to the identity matrix and we have

$$P_c(x,\lambda) = f(x) + \lambda' h(x) + \frac{1}{2}\|W(x)\nabla f(x) + \lambda\|^2 + \frac{c}{2}\|h(x)\|^2.$$

We can minimize explicitly this function with respect to λ by setting

$$\nabla_\lambda P_c = h(x) + W(x)\nabla f(x) + \lambda = 0.$$

Then,

$$\min_\lambda P_c(x,\lambda) = f(x) + \hat{\lambda}(x)' h(x) + \frac{c-1}{2}\|h(x)\|^2,$$

where

$$\hat{\lambda}(x) = -W(x)\nabla f(x).$$

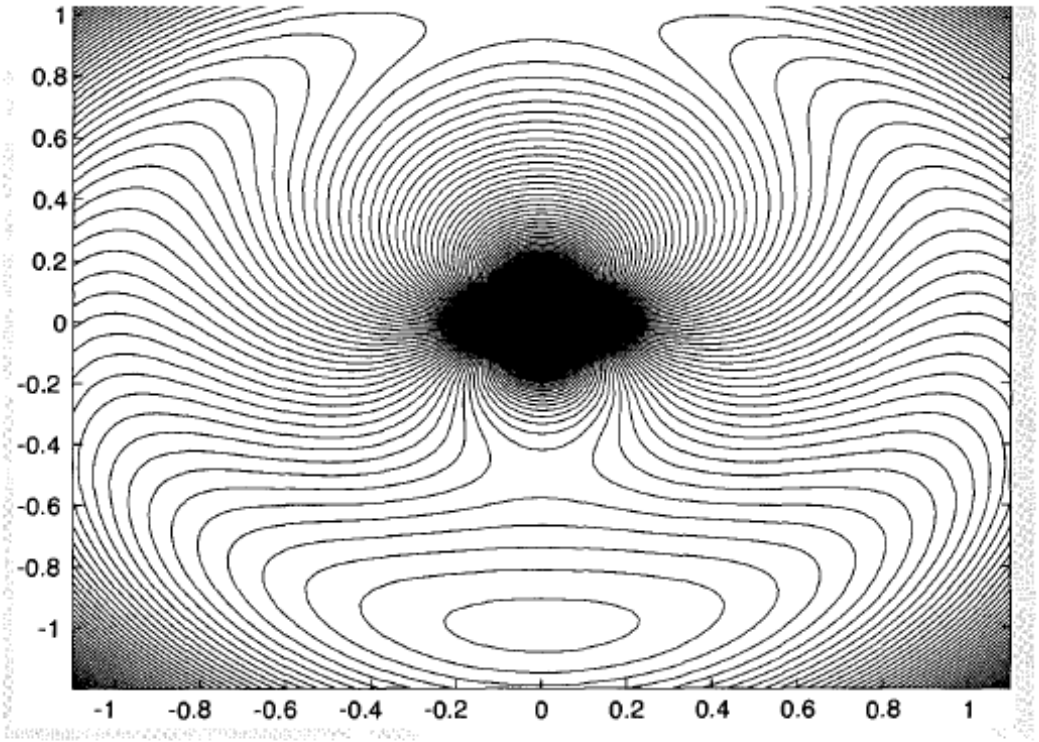

Figure 4.3.5. Contours of the differentiable exact penalty function $\hat{P}_c(x)$ for the two-dimensional problem where

$$f(x) = x_1, \qquad h(x) = x_1^2 + x_2^2 - 1$$

(cf. Figs. 4.3.1 and 4.3.2). The figure corresponds to $c = 2$. Note that there is a singularity at $(0, 0)$, which is a nonregular point at which $\hat{\lambda}(x)$ is undefined. The function $\hat{P}_c(x)$ takes arbitrarily large and arbitrarily small values sufficiently close to $(0, 0)$. This type of singularity can be avoided by using a modification of the exact penalty function (see Exercise 4.3.8).

The algorithms of this section may be viewed as methods for solving the system of necessary optimality conditions of the equality constrained minimization problem

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } h(x) = 0 \end{aligned}$$

Thus the necessary optimality conditions

$$\nabla f(x) + \nabla h(x)\lambda = 0, \quad h(x) = 0,$$

are treated as a system of $(n + m)$ nonlinear equations with $(n + m)$ unknowns, the vectors x and λ.

First Order Methods

We will consider algorithms for solving the Lagrangian system

$$\nabla f(x) + \nabla h(x)\lambda = 0, \quad h(x) = 0.$$

These algorithms, called Lagrangian methods, have the generic form

$$x^{k+1} = G\left(x^k, \lambda^k\right), \quad \lambda^{k+1} = H\left(x^k, \lambda^k\right),$$

where $G : \Re^{n+m} \mapsto \Re^n$ and $H : \Re^{n+m} \mapsto \Re^n$ are continuously differentiable functions. Since the above iteration can only converge to a pair (x^*, λ^*) such that

$$x^* = G\left(x^*, \lambda^*\right), \quad \lambda^* = H\left(x^*, \lambda^*\right),$$

the functions G and H must be chosen so that local minima-Lagrange multiplier pairs satisfy the above equations.

The simplest Lagrangian method (also known as the first order Lagrangian method) is given by

$$x^{k+1} = x^k - \alpha \nabla_x L\left(x^k, \lambda^k\right),$$
$$\lambda^{k+1} = \lambda^k + \alpha h\left(x^k\right),$$

where L is the Lagrangian function

$$L(x, \lambda) = f(x) + \lambda' h(x)$$

and $\alpha > 0$ is a scalar stepsize.

Proposition

Let $G : \Re^{n+m} \mapsto \Re^n$ *and* $H : \Re^{n+m} \mapsto \Re^n$ *be continuously differentiable functions. Assume that* (x^*, λ^*) *satisfies*

$$x^* = G\left(x^*, \lambda^*\right), \quad \lambda^* = H\left(x^*, \lambda^*\right),$$

and that all eigenvalues of the $(n+m) \times (n+m)$ *matrix*

$$R^* = \begin{pmatrix} \nabla_x G\left(x^*, \lambda^*\right) & \nabla_x H\left(x^*, \lambda^*\right) \\ \nabla_\lambda G\left(x^*, \lambda^*\right) & \nabla_\lambda H\left(x^*, \lambda^*\right) \end{pmatrix},$$

lie strictly within the unit circle of the complex plane. Then (x^*, λ^*) *is a point of attraction of the iteration*

$$x^{k+1} = G\left(x^k, \lambda^k\right), \quad \lambda^{k+1} = H\left(x^k, \lambda^k\right),$$

and when the generated sequence $\left\{\left(x^k, \lambda^k\right)\right\}$ *converges to* (x^*, λ^*)*, the rate of convergence of* $\left\|x^k - x^*\right\|$ *and* $\left\|\lambda^k - \lambda^*\right\|$ *is linear.*

Proof: Denote $y = (x, \lambda), y^k = \left(x^k, \lambda^k\right), y^* = (x^*, \lambda^*)$, and consider the function $M : \Re^{n+m} \mapsto \Re^{n+m}$ given by $M(y) = (G(x, \lambda), H(x, \lambda))$. By the mean value theorem, we have for any two vectors y and $\tilde{y}$

$$M(\tilde{y}) - M(y) = R'(\tilde{y} - y),$$

where R is the matrix having as ith column the gradient $\nabla M_i\left(\hat{y}^i\right)$ of the ith component of M evaluated at some vector $\hat{y}^i$ on the line segment connecting y and $\tilde{y}$. By taking $\tilde{y}$ and y sufficiently close to y^*, we can make R as close to the matrix R^*, and therefore we can make the eigenvalues of the transpose R' lie within the unit circle. There exists a norm $\|\cdot\|$ and an open sphere S with respect to that norm centered at (x^*, λ^*) such that, within S , the induced matrix norm of R' is less than $1 - \epsilon$ where ϵ is some positive scalar. Since

$$\|M(\tilde{y}) - M(y)\| \leq \|R'\| \, \|\tilde{y} - y\|,$$

it follows that within the sphere S, M is a contraction mapping. The result then follows from the contraction mapping theorem.

Proposition

Assume that f and h are twice continuously differentiable, and let (x^*, λ^*) be a local minimum-Lagrange multiplier pair. Assume also that x^* is regular and that the matrix $\nabla^2_{xx} L(x^*, \lambda^*)$ is positive definite. Then there exists $\overline{\alpha} > 0$, such that for all $\alpha \in (0, \overline{\alpha}]$, (x^*, λ^*) is a point of attraction, and if the generated sequence $\{(x^k, \lambda^k)\}$ converges to (x^*, λ^*), then the rate of convergence of $\|x^k - x^*\|$ and $\|\lambda^k - \lambda^*\|$ is linear.

Proof: The proof consists of showing that, for α sufficiently small, the hypothesis of above Prop. is satisfied. Indeed for $\alpha > 0$, consider the mapping $M_\alpha : \Re^{n+m} \mapsto \Re^{n+m}$ defined by

$$M_\alpha(x, \lambda) = \begin{pmatrix} x - \alpha \nabla_x L(x, \lambda) \\ \lambda + \alpha \nabla_\lambda L(x, \lambda) \end{pmatrix}.$$

Clearly $(x^*, \lambda^*) = M_\alpha(x^*, \lambda^*)$, and we have

$$\nabla M_\alpha(x^*, \lambda^*)' = I - \alpha B,$$

where

where

$$B = \begin{pmatrix} \nabla^2_{xx} L(x^*, \lambda^*) & \nabla h(x^*) \\ -\nabla h(x^*)' & 0 \end{pmatrix}.$$

We will show that the real part of each eigenvalue of B is strictly positive. For any complex vector y, denote by $\hat{y}$ its complex conjugate, and for any complex number γ, denote by $Re(\gamma)$ its real part. Let β be an eigenvalue of B, and let $(z, w) \neq 0$ be a corresponding eigenvector where z and w are complex vectors of dimension n and m, respectively. We have

$$\mathrm{Re}\left\{ (\hat{z}', \hat{w}') B \begin{pmatrix} z \\ w \end{pmatrix} \right\} = Re\left\{ \beta \begin{pmatrix} \hat{z}' & \hat{w}' \end{pmatrix} \begin{pmatrix} z \\ w \end{pmatrix} \right\} = \mathrm{Re}(\beta) \left(\|z\|^2 + \|w\|^2 \right),$$

while at the same time,

$$\mathrm{Re}\left\{ \begin{pmatrix} \hat{z}' & \hat{w}' \end{pmatrix} B \begin{pmatrix} z \\ w \end{pmatrix} \right\} = \mathrm{Re}\left\{ \hat{z}' \nabla^2_{xx} L(x^*, \lambda^*) z + \hat{z}' \nabla h(x^*) w - \hat{w}' \nabla h(x^*)' z \right\}.$$

Since for any real $n \times m$ matrix M, we have

$$Re\left\{ \hat{z}' M' w \right\} = Re\left\{ \hat{w}' M z \right\},$$

Then,

$$\mathrm{Re}\left\{\hat{z}'\nabla_{xx}^2 L\left(x^*,\lambda^*\right)z\right\} = \mathrm{Re}\left\{\begin{pmatrix}\hat{z}' & \hat{w}'\end{pmatrix} B\begin{pmatrix} z \\ w\end{pmatrix}\right\} = \mathrm{Re}(\beta)\left(\|z\|^2+\|w\|^2\right).$$

Since for any positive definite matrix A, we have

$$\mathrm{Re}\left\{\hat{z}'Az\right\} > 0, \quad \forall z \neq 0,$$

it follows from the positive definiteness assumption on $\nabla_{xx}^2 L\left(x^*,\lambda^*\right)$ that either $Re(\beta) > 0$ or else $z = 0$. But if $z = 0$, the equation

$$B\begin{pmatrix} z \\ w\end{pmatrix} = \beta\begin{pmatrix} z \\ w\end{pmatrix}.$$

yields

$$\nabla h\left(x^*\right)w = 0.$$

Since $\nabla h\left(x^*\right)$ has rank m, it follows that $w = 0$. This contradicts our earlier assumption that $(z,w) \neq 0$. Consequently, we must have $Re(\beta) > 0$.

Newton-Like Methods for Equality Constraints

Let us write the Lagrangian system

$$\nabla f(x) + \nabla h(x)\lambda = 0, \quad h(x) = 0,$$

as

$$\nabla L(x, \lambda) = 0.$$

Newton's method for solving this system is given by

$$x^{k+1} = x^k + \Delta x^k, \quad \lambda^{k+1} = \lambda^k + \Delta \lambda^k,$$

where $\left(\Delta x^k, \Delta \lambda^k\right) \in \Re^{n+m}$ is obtained by solving the system of equations

$$\nabla^2 L\left(x^k, \lambda^k\right) \begin{pmatrix} \Delta x^k \\ \Delta \lambda^k \end{pmatrix} = -\nabla L\left(x^k, \lambda^k\right).$$

We say that $\left(x^{k+1}, \lambda^{k+1}\right)$ is well-defined by the Newton iteration, if the matrix $\nabla^2 L\left(x^k, \lambda^k\right)$ is invertible.

Proposition

Let x^* be a strict local minimum that is regular and satisfies together with a corresponding Lagrange multiplier vector λ^* the second order sufficiency conditions. Then (x^*, λ^*) is a point of attraction of the Newton iteration. Furthermore, if the generated sequence converges to (x^*, λ^*), the rate of convergence of $\left\{\left\|\left(x^k, \lambda^k\right) - (x^*, \lambda^*)\right\|\right\}$ is superlinear (at least order two if $\nabla^2 f$ and $\nabla^2 h_i, i = 1, \ldots, m$, are Lipschitz continuous in a neighborhood of x^*).

A First Implementation of Newton's Method

Let us write the Hessian of the Lagrangian function as

$$\nabla^2 L\left(x^k, \lambda^k\right) = \begin{pmatrix} H^k & N^k \\ N^{k'} & 0 \end{pmatrix}, \quad \nabla L\left(x^k, \lambda^k\right) = \begin{pmatrix} \nabla_x L\left(x^k, \lambda^k\right) \\ h\left(x^k\right) \end{pmatrix},$$

where

$$H^k = \nabla^2_{xx} L\left(x^k, \lambda^k\right), \quad N^k = \nabla h\left(x^k\right).$$

Thus,

$$\begin{pmatrix} H^k & N^k \\ N^k & 0 \end{pmatrix} \begin{pmatrix} \Delta x^k \\ \Delta \lambda^k \end{pmatrix} = - \begin{pmatrix} \nabla_x L\left(x^k, \lambda^k\right) \\ h\left(x^k\right) \end{pmatrix}.$$

Let us assume that H^k is invertible and N^k has rank m. Then,

$$H^k \Delta x^k + N^k \Delta \lambda^k = -\nabla_x L\left(x^k, \lambda^k\right),$$
$$N^{k'} \Delta x^k = -h\left(x^k\right).$$

We obtain

$$\lambda^{k+1} = \left(N^{k'}\left(H^k\right)^{-1} N^k\right)^{-1} \left(h\left(x^k\right) - N^{k'}\left(H^k\right)^{-1} \nabla f\left(x^k\right)\right),$$
$$x^{k+1} = x^k - \left(H^k\right)^{-1} \nabla_x L\left(x^k, \lambda^{k+1}\right).$$

A Second Implementation of Newton's Method

If $\left(H^k + cN^k N^k\right)^{-1}$ exists, then

$$x^{k+1} = x^k - \left(H^k + cN^k N^{k'}\right)^{-1} \nabla_x L\left(x^k, \hat{\lambda}^{k+1}\right),$$
$$\lambda^{k+1} = \hat{\lambda}^{k+1} - \mathrm{ch}\left(x^k\right),$$

where

$$\hat{\lambda}^{k+1} = \left(N^{k'}\left(H^k + cN^k N^{k'}\right)^{-1} N^k\right)^{-1} \left(h\left(x^k\right) - N^{k'}\left(H^k + cN^k N^{k'}\right)^{-1} \nabla$$

By observing that

$$\nabla_x L\left(x^k, \lambda^{k+1} + ch\left(x^k\right)\right) = \nabla_x L_c\left(x^k, \lambda^{k+1}\right),$$

we see that an alternative way to write the update equation for x^k is

$$x^{k+1} = x^k - \left(H^k + cN^k N^{k'}\right)^{-1} \nabla_x L_c\left(x^k, \lambda^{k+1}\right).$$

An Implementation of Newton's Method Based on Quadratic Programming

Since

$$\nabla f\left(x^k\right) + H^k \Delta x^k + N^k \lambda^{k+1} = 0, \quad h\left(x^k\right) + N^{k'} \Delta x^k = 0,$$

which are the necessary optimality conditions for $\left(\Delta x^k, \lambda^{k+1}\right)$ to be a global minimum-Lagrange multiplier pair of the quadratic program

$$\begin{array}{l} \text{minimize } \nabla f\left(x^k\right)' \Delta x + \frac{1}{2} \Delta x' H^k \Delta x \\ \text{subject to } h\left(x^k\right) + N^k \Delta x = 0 \end{array}$$

Thus we can obtain $\left(\Delta x^k, \lambda^{k+1}\right)$ by solving this problem.

Merit Functions and Descent Properties of Newton's Method

Since we would like to improve the global convergence properties of Newton's method, it is interesting to look for appropriate merit functions, that is, functions for which $\left(x^{k+1}-x^{k}\right)$ is a descent direction at x^{k} or $\left(x^{k+1}-x^{k}, \lambda^{k+1}-\lambda^{k}\right)$ is a descent direction at $\left(x^{k}, \lambda^{k}\right)$. By this, we mean functions F such that for a sufficiently small positive scalar $\overline{\alpha}$, we have

$$F\left(x^{k}+\alpha\left(x^{k+1}-x^{k}\right)\right)<F\left(x^{k}\right), \quad \forall \alpha \in(0, \overline{\alpha}],$$

or

$$F\left(x^{k}+\alpha\left(x^{k+1}-x^{k}\right), \lambda^{k}+\alpha\left(\lambda^{k+1}-\lambda^{k}\right)\right)<F\left(x^{k}, \lambda^{k}\right), \quad \forall \alpha \in(0, \overline{\alpha}].$$

Proposition

Let x^* be a local minimum that is a regular point and satisfies together with a corresponding Lagrange multiplier vector λ^* the second order sufficiency conditions. There exists a neighborhood S of (x^*, λ^*) such that if $\left(x^k, \lambda^k\right) \in S$ and $x^k \neq x^*$, then $\left(x^{k+1}, \lambda^{k+1}\right)$ is well-defined by the Newton iteration and the following hold:

(a) *There exists a scalar $\bar{c}$ such that for all $c \geq \bar{c}$, the vector $(x^{k+1} - x^k)$ is a descent direction at x^k for the exact penalty function $f(x) + c \max_{i=1,\ldots,m} |h_i(x)|$.*

(b) *The vector $(x^{k+1} - x^k, \lambda^{k+1} - \lambda^k)$ is a descent direction at (x^k, λ^k) for the exact penalty function $P(x, \lambda) = \frac{1}{2}\|\nabla L(x, \lambda)\|^2$. Furthermore, given any scalar $r > 0$, there exists a $\delta > 0$ such that if $\left\|\left(x^k - x^*, \lambda^k - \lambda^*\right)\right\| < \delta$, we have*

$$P\left(x^{k+1}, \lambda^{k+1}\right) \leq rP\left(x^k, \lambda^k\right).$$

(c) *For every scalar c such that $H^k + cN^k N^{k'}$ is positive definite, the vector $\left(x^{k+1} - x^k\right)$ is a descent direction at x^k of the augmented Lagrangian function $L_c\left(\cdot, \lambda^{k+1}\right)$.*

Proof: (a) Take $\overline{c} > 0$ sufficiently large and a neighborhood S of (x^*, λ^*) which is sufficiently small, so that for $\left(x^k, \lambda^k\right) \in S$, the matrix $H^k + \overline{c} N^k N^k$ is positive definite. Since Δx^k is the solution of the quadratic program, it follows that if $x^k \neq x^*$, then Δx^k is a descent direction of the exact penalty function (4.109) for all $c \geq \overline{c}$.

(b) We have

$$\begin{pmatrix} x^{k+1} - x^k \\ \lambda^{k+1} - \lambda^k \end{pmatrix} = -\nabla^2 L\left(x^k, \lambda^k\right)^{-1} \nabla L\left(x^k, \lambda^k\right),$$

and

$$\nabla P\left(x^k, \lambda^k\right) = \nabla^2 L\left(x^k, \lambda^k\right) \nabla L\left(x^k, \lambda^k\right).$$

So

$$\left(\left(x^{k+1} - x^k\right)', \left(\lambda^{k+1} - \lambda^k\right)'\right) \nabla P\left(x^k, \lambda^k\right) = -\left\|\nabla L\left(x^k, \lambda^k\right)\right\|^2 < 0,$$

and the descent property follows.

We have that, given any $\overline{r} > 0$, there exists a $\overline{\delta} > 0$ such that for $\left\|\left(x^k - x^*, \lambda^k - \lambda^*\right)\right\| < \overline{\delta}$, we have

$$\left\| \left(x^{k+1} - x^*, \lambda^{k+1} - \lambda^* \right) \right\| \leq \overline{r} \left\| \left(x^k - x^*, \lambda^k - \lambda^* \right) \right\|.$$

For every (x, λ), we have, by the mean value theorem,

$$\nabla L(x, \lambda) = B \begin{pmatrix} x - x^* \\ \lambda - \lambda^* \end{pmatrix}.$$

where each row of B is the corresponding row of $\nabla^2 L$ evaluated at a point between (x, λ) and (x^*, λ^*). Since $\nabla^2 L(x^*, \lambda^*)$ is invertible, it follows that there is an $\epsilon > 0$ and scalars $\mu > 0$ and $M > 0$ such that for $\|(x - x^*, \lambda - \lambda^*)\| < \epsilon$, we have

$$\mu \|(x - x^*, \lambda - \lambda^*)\| \leq \|\nabla L(x, \lambda)\| \leq M \|(x - x^*, \lambda - \lambda^*)\|.$$

It follows that for each $\overline{r} > 0$ there exists $\delta > 0$ such that, for $\|(x^k - x^*, \lambda^k - \lambda^*)\| < \delta$,

$$\left\| \nabla L \left(x^{k+1}, \lambda^{k+1} \right) \right\| \leq (M\overline{r}/\mu) \left\| \nabla L \left(x^k, \lambda^k \right) \right\|.$$

or, equivalently,

$$P \left(x^{k+1}, \lambda^{k+1} \right) \leq \left(M^2 \overline{r}^2 / \mu^2 \right) P \left(x^k, \lambda^k \right).$$

Given $r > 0$, we take $\overline{r} = (\mu/M)\sqrt{r}$ in the relation above.

(c) We have shown that $x^{k+1} - x^k = -(H^k + cN^k N^{k'})^{-1} \nabla_x L_c(x^k, \lambda^{k+1})$, which implies the conclusion.

Variations of Newton's Method

There are a number of variations of Newton's method, which are obtained by introducing some extra terms in the left-hand side of the Newton system. These variations have the general form

$$x^{k+1} = x^k + \Delta x^k, \quad \lambda^{k+1} = \lambda^k + \Delta \lambda^k,$$

where

$$\left(\nabla^2 L\left(x^k, \lambda^k\right) + V^k\left(x^k, \lambda^k\right)\right) \begin{pmatrix} \Delta x^k \\ \Delta \lambda^k \end{pmatrix} = -\nabla L\left(x^k, \lambda^k\right),$$

where the extra term $V^k\left(x^k, \lambda^k\right)$ is "small" enough relative to $\nabla^2 L\left(x^k, \lambda^k\right)$, so that the eigenvalues of the matrix

$$I - \left(\nabla^2 L\left(x^k, \lambda^k\right) + V^k\left(x^k, \lambda^k\right)\right)^{-1} \nabla^2 L\left(x^k, \lambda^k\right)$$

are within the unit circle.

An interesting approximation of Newton's method is obtained by adding a term $-\left(1/c^k\right)\Delta\lambda^k$ in the left-hand side of the equation $N^k\Delta x^k = -h\left(x^k\right)$, where c^k is a positive parameter, so that Δx^k and $\Delta\lambda^k$ are obtained by solving the system

$$H^k\Delta x^k + N^k\Delta\lambda^k = -\nabla_x L\left(x^k,\lambda^k\right),$$
$$N^{k'}\Delta x^k - \left(1/c^k\right)\Delta\lambda^k = -h\left(x^k\right).$$

As $c^k \to \infty$, the system asymptotically becomes identical to the one corresponding to Newton's method.

Connection with the First Order Method of Multipliers

If $H^k + c^k N^k N^k$ is positive definite, h_i are linear, and f is quadratic, then Newton's method is equivalent to the first order method of multipliers.

Extension to Inequality Constraints

Let us consider the inequality constrained problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_j(x) \leq 0, \quad j = 1, \ldots, r, \end{array}$$

and focus on a local minimum x^* that is regular and together with a Lagrange multiplier μ^*, satisfies the second order sufficiency conditions. Given $\left(x^k, \mu^k\right)$, we obtain $\left(x^{k+1}, \mu^{k+1}\right)$ as an optimal solution-Lagrange multiplier pair of the quadratic program

$$\begin{array}{l} \text{minimize } \nabla f\left(x^k\right)'\left(x - x^k\right) + \frac{1}{2}\left(x - x^k\right)' \nabla^2_{xx} L\left(x^k, \mu^k\right)\left(x - x^k\right) \\ \text{subject to } g_j\left(x^k\right) + \nabla g_j\left(x^k\right)'\left(x - x^k\right) \leq 0, \quad j = 1, \ldots, r. \end{array}$$

It is possible to show that there exists a neighborhood S of (x^*, μ^*) such that if $\left(x^x, \mu^k\right)$ is within S, then $\left(x^{k+1}, \mu^{k+1}\right)$ is uniquely defined as an optimal solution-Lagrange multiplier pair within S (an application of the implicit function theorem is needed to formalize this statement).