Chapter 5

Constrained Optimization

Engineering design optimization problems are very rarely unconstrained. Moreover, the constraints that appear in these problems are typically nonlinear. This motivates our interest in general *nonlinearly constrained* optimization theory and methods in this chapter.

Recall the statement of a general optimization problem,

$$\begin{aligned} \text{minimize} \quad & f(x) && (5.1)\\ \text{with respect to} \quad & x \in \mathbb{R}^n && (5.2)\\ \text{subject to} \quad & \hat{c}_j(x) = 0, \quad j = 1, \ldots, \hat{m} && (5.3)\\ & c_k(x) \geq 0, \quad k = 1, \ldots, m && (5.4) \end{aligned}$$

Example 5.21. Graphical Solution of a Constrained Optimization Problem

Suppose we want to solve the following optimization problem,

$$\begin{aligned} \text{minimize} \quad & f(x) = 4x_1^2 - x_1 - x_2 - 2.5 && (5.5)\\ \text{with respect to} \quad & x_1, x_2 && (5.6)\\ \text{subject to} \quad & c_1(x) = x_2^2 - 1.5x_1^2 + 2x_1 - 1 \geq 0, && (5.7)\\ & c_2(x) = x_2^2 + 2x_1^2 - 2x_1 - 4.25 \leq 0 && (5.8) \end{aligned}$$

How can we solve this? One intuitive way would be to make a contour plot of the objective function, overlay the constraints and determine the feasible region, as shown in Fig. 5.1. By inspection, it is easy to see where the constrained optimum is. At the optimum, only one of the constraints is active. However, we want to be able to find the minimum numerically, without having to plot the functions, since this is impractical in the general case. We can see that in this case, the feasible space comprises two disconnected regions. In addition, although all the functions involved are smooth, the boundaries of the feasible regions are nonmsooth. These characteristics complicate the numerical solution of constrained optimization problems.

5.1 Optimality Conditions for Constrained Problems

The optimality conditions for nonlinearly constrained problems are important because they form the basis for algorithms for solving such problems.

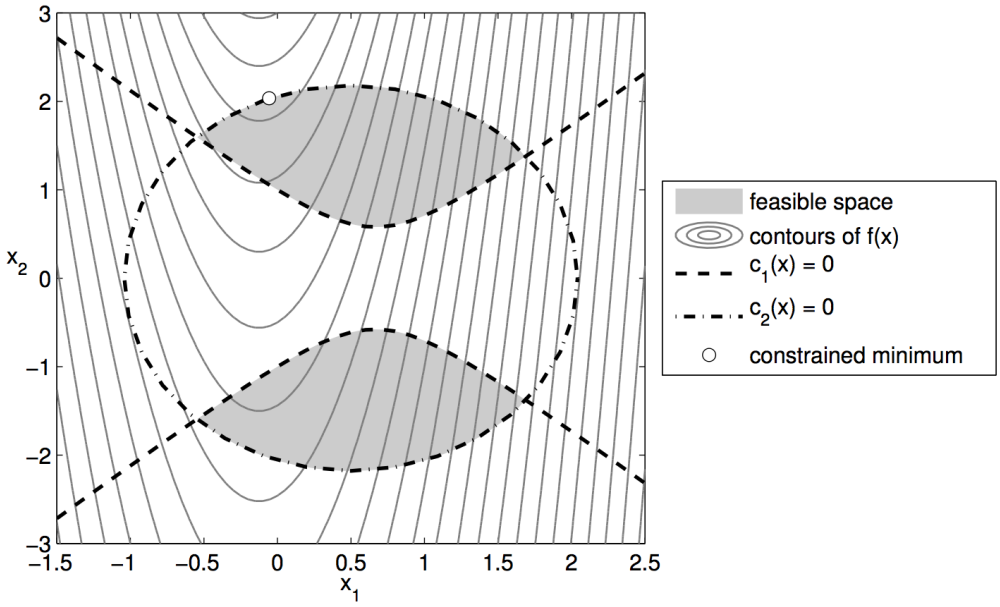


Figure 5.1: Example contours and feasible regions for a simple constrained optimization problem.

5.1.1 Nonlinear Equality Constraints

Suppose we have the following optimization problem with equality constraints,

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{with respect to} \quad & x \in \mathbb{R}^n \\ \text{subject to} \quad & \hat{c}_j(x) = 0, \quad j = 1, \ldots, \hat{m} \end{aligned}$$

To solve this problem, we could solve for $\hat{m}$ components of x by using the equality constraints to express them in terms of the other components. The result would be an unconstrained problem with $n - \hat{m}$ variables. However, this procedure is only feasible for simple explicit functions.

Joseph Louis Lagrange is credited with developing a more general method to solve this problem, which we now review. At a stationary point, the total differential of the objective function has to be equal to zero, i.e.,

$$\mathrm{d}f = \frac{\partial f}{\partial x_1}\,\mathrm{d}x_1 + \frac{\partial f}{\partial x_2}\,\mathrm{d}x_2 + \cdots + \frac{\partial f}{\partial x_n}\,\mathrm{d}x_n = \nabla f^T\,\mathrm{d}x = 0. \tag{5.9}$$

Unlike unconstrained optimization, the infinitesimal vector $\mathrm{d}x = [\,\mathrm{d}x_1,\, \mathrm{d}x_2, \ldots,\, \mathrm{d}x_n]^T$ is not arbitrary here; if x is the sought after minimum, then the perturbation $x + \mathrm{d}x$ must be feasible: $\hat{c}_j(x + \mathrm{d}x) = 0$. Consequently, the above equations *does not* imply that $\nabla f = 0$.

For a feasible point, the total differential of each of the constraints $(\hat{c}_1, \ldots \hat{c}_{\hat{m}})$ must also be zero:

$$\mathrm{d}\hat{c}_j = \frac{\partial \hat{c}_j}{\partial x_1}\,\mathrm{d}x_1 + \cdots + \frac{\partial \hat{c}_j}{\partial x_n}\,\mathrm{d}x_n = \nabla \hat{c}_j^T\,\mathrm{d}x = 0, \quad j = 1, \ldots, \hat{m} \tag{5.10}$$

To interpret the above equation, recall that the gradient of a function is orthogonal to its contours. Thus, since the displacement $\mathrm{d}x$ satisfies $\hat{c}_j(x + dx) = 0$ (the equation for a contour), it follow that $\mathrm{d}x$ is orthogonal to the gradient $\nabla \hat{c}_j$.

Lagrange suggested that one could multiply each constraint variation by a scalar $\hat{\lambda}_j$ and subtract it from the objective function,

$$\mathrm{d}f - \sum_{j=1}^{\hat{m}} \hat{\lambda}_j \, \mathrm{d}\hat{c}_j = 0 \Rightarrow \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} - \sum_{j=1}^{\hat{m}} \hat{\lambda}_j \frac{\partial \hat{c}_j}{\partial x_i} \right) \mathrm{d}x_i = 0. \tag{5.11}$$

Notice what has happened: the components of the infinitesimal vector $\mathrm{d}x$ have become independent and arbitrary, because we have accounted for the constraints. Thus, for this equation to be satisfied, we need a vector $\hat{\lambda}$ such that the expression inside the parenthesis vanishes, i.e.,

$$\frac{\partial f}{\partial x_i} - \sum_{j=1}^{\hat{m}} \hat{\lambda}_j \frac{\partial \hat{c}_j}{\partial x_i} = 0, \qquad (i = 1, 2, \ldots, n) \tag{5.12}$$

which is a system of n equations and $n + m$ unknowns. To close the system, we recognize that the m constraints must also be satisfied.

Suppose we define a function as the objective function minus a weighted sum of the constraints,

$$\mathcal{L}(x, \hat{\lambda}) = f(x) - \sum_{j=1}^{\hat{m}} \hat{\lambda}_j \hat{c}_j(x) \Rightarrow$$

$$\boxed{\mathcal{L}(x, \hat{\lambda}) = f(x) - \hat{\lambda}^T \hat{c}(x)} \tag{5.13}$$

We call this function the *Lagrangian* of the constrained problem, and the weights the *Lagrange* multipliers. A stationary point of the Lagrangian with respect to both x and $\hat{\lambda}$ will satisfy

$$\frac{\partial \mathcal{L}}{\partial x_i} = \frac{\partial f}{\partial x_i} - \sum_{j=1}^{\hat{m}} \hat{\lambda}_j \frac{\partial \hat{c}_j}{\partial x_i} = 0, \quad (i = 1, \ldots, n) \tag{5.14}$$

$$\frac{\partial \mathcal{L}}{\partial \hat{\lambda}_j} = \hat{c}_j = 0, \quad (j = 1, \ldots, \hat{m}). \tag{5.15}$$

Thus, at a stationary point of the Lagrangian encapsulates our required conditions: the constraints are satisfied and the gradient conditions (5.12) are satisfied. These first-order conditions are known as the *Karush–Kuhn–Tucker (KKT)* conditions. They are necessary conditions for the optimum of a constrained problem.

As in the unconstrained case, the first-order conditions are not sufficient to guarrantee a local minimum. For this, we turn to the second-order sufficient conditions (which, as in the unconstrained case, are not necessary). For equality constrained problems we are concerned with the behaviour of the Hessian of the Lagrangian, denoted $\nabla^2_{xx}\mathcal{L}(x, \hat{\lambda})$, at locations where the KKT conditions hold. In particular, we look for positive-definiteness *in a subspace defined by the linearized constraints.* Geometrically, if we move away from a stationary point $(x^*, \hat{\lambda}^*)$ along a direction w that satisfies the linearized constraints, the Lagrangian should look like a quadratric along this direction. To be precise, the second-order sufficient conditions are

$$w^T \nabla^2_{xx}\mathcal{L}(x^*, \hat{\lambda}^*) w > 0,$$

for all $w \in \mathbb{R}^n$ such that

$$\nabla \hat{c}_j(x^*)^T w = 0, \qquad j = 1, \ldots, \hat{m}.$$

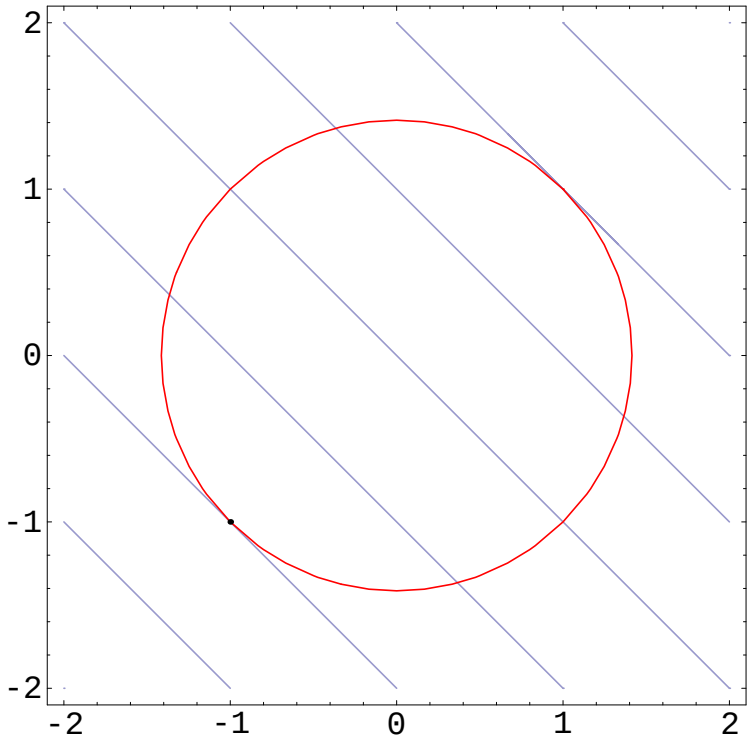


Figure 5.2: Contour plot for constrained problem (5.18).

Example 5.22. Problem with Single Equality Constraint

Consider the following equality constrained problem:

$$\begin{aligned} \text{minimize} \quad & f(x) = x_1 + x_2 && (5.16)\\ \text{weight respect to} \quad & x_1, x_2 && (5.17)\\ \text{subject to} \quad & \hat{c}_1(x) = x_1^2 + x_2^2 - 2 = 0 && (5.18) \end{aligned}$$

The objective function and constraint of the above problem are shown in Fig. 5.2. By inspection we can see that the feasible region for this problem is a circle of radius $\sqrt{2}$. The solution x^* is obviously $(-1,-1)^T$. From any other point in the circle it is easy to find a way to move in the feasible region (the boundary of the circle) while decreasing f.

In this example, the Lagrangian is

$$\mathcal{L} = x_1 + x_2 - \hat{\lambda}_1(x_1^2 + x_2^2 - 2) \tag{5.19}$$

And the optimality conditions are

$$\begin{aligned} \nabla_x \mathcal{L} &= \begin{bmatrix} 1 - 2\hat{\lambda}_1 x_1 \\ 1 - 2\hat{\lambda}_1 x_2 \end{bmatrix} = 0 \Rightarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2\hat{\lambda}_1} \\ \frac{1}{2\hat{\lambda}_1} \end{bmatrix} \\ \nabla_{\hat{\lambda}_1} \mathcal{L} &= x_1^2 + x_2^2 - 2 = 0 \Rightarrow \hat{\lambda}_1 = \pm\frac{1}{2} \end{aligned}$$

To establish which are minima as opposed to other types of stationary points, we need to look at the second-order conditions. Directions $w = (w_1, w_2)^T$ that satisfy the linearized constraints are

given by

$$\nabla \hat{c}_1(x^*)^T w = \frac{1}{\hat{\lambda}_1}(w_1 + w_2) = 0$$
$$\Rightarrow \quad w_2 = -w_1$$

while the Hessian of the Lagrangian at the stationary points is

$$\nabla_x^2 \mathcal{L} = \begin{bmatrix} -2\hat{\lambda}_1 & 0 \\ 0 & -2\hat{\lambda}_1 \end{bmatrix}.$$

Consequently, the Hessian of the Lagrangian in the subspace defined by w is

$$w^T \nabla_{xx}^2 \mathcal{L}(x^*) w = \begin{bmatrix} w_1 & -w_1 \end{bmatrix} \begin{bmatrix} -2\hat{\lambda}_1 & 0 \\ 0 & -2\hat{\lambda}_1 \end{bmatrix} \begin{bmatrix} w_1 \\ -w_1 \end{bmatrix} = -4\hat{\lambda}_1 w_1^2$$

In this case $\hat{\lambda}_1^* = -\frac{1}{2}$ corresponds to a positive-definite Hessian (in the space w) and, therefore, the solution to the problem is $(x_1, x_2)^T = (\frac{1}{2\hat{\lambda}_1}, \frac{1}{2\hat{\lambda}_1})^T = (-1, -1)^T$.

Also note that at the solution the constraint normal $\nabla \hat{c}_1(x^*)$ is parallel to $\nabla f(x^*)$, i.e., there is a scalar $\hat{\lambda}_1^*$ such that

$$\nabla f(x^*) = \hat{\lambda}_1^* \nabla \hat{c}_1(x^*).$$

We can derive this expression by examining the first-order Taylor series approximations to the objective and constraint functions. To retain feasibility with respect to $\hat{c}_1(x) = 0$ we require that

$$\begin{aligned} \hat{c}_1(x+d) &= 0 \Rightarrow \\ \hat{c}_1(x+d) &= \underbrace{\hat{c}_1(x)}_{=0} + \nabla \hat{c}_1^T(x) d + \mathcal{O}(d^T d). \end{aligned}$$

Linearizing this we get,

$$\boxed{\nabla \hat{c}_1^T(x) d = 0}.$$

We also know that a direction of improvement must result in a decrease in f, i.e.,

$$f(x+d) - f(x) < 0.$$

Thus to first order we require that

$$\begin{aligned} f(x) + \nabla f^T(x) d - f(x) &< 0 \Rightarrow \\ \boxed{\nabla f^T(x) d < 0}&. \end{aligned}$$

A necessary condition for optimality is that there be no direction satisfying both of these conditions. The only way that such a direction cannot exist is if $\nabla f(x)$ and $\nabla \hat{c}_1(x)$ are parallel, that is, if $\nabla f(x) = \hat{\lambda}_1 \nabla \hat{c}_1(x)$ holds.

By defining the Lagrangian function

$$\mathcal{L}(x, \hat{\lambda}_1) = f(x) - \hat{\lambda}_1 \hat{c}_1(x), \tag{5.20}$$

and noting that $\nabla_x \mathcal{L}(x, \hat{\lambda}_1) = \nabla f(x) - \hat{\lambda}_1 \nabla \hat{c}_1(x)$, we can state the necessary optimality condition as follows: At the solution x^* there is a scalar $\hat{\lambda}_1^*$ such that $\nabla_x \mathcal{L}(x^*, \hat{\lambda}_1^*) = 0$.

Thus we can search for solutions of the equality-constrained problem by searching for a stationary point of the Lagrangian function. The scalar $\hat{\lambda}_1$ is the Lagrange multiplier for the constraint $\hat{c}_1(x) = 0$.

5.1.2 Nonlinear Inequality Constraints

Suppose we now have a general problem with equality and inequality constraints.

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{w.r.t} \quad & x \in \mathbb{R}^n \\ \text{subject to} \quad & \hat{c}_j(x) = 0, \quad j = 1, \ldots, \hat{m} \\ & c_k(x) \geq 0, \quad k = 1, \ldots, m \end{aligned}$$

The optimality (KKT) conditions for this problem can also be obtained for this case by modifying the Lagrangian to be

$$\mathcal{L}(x, \hat{\lambda}, \lambda, s) = f(x) - \hat{\lambda}^T \hat{c}(x) - \lambda^T \left(c(x) - s^2 \right), \tag{5.21}$$

where λ are the Lagrange multipliers associated with the inequality constraints and s is a vector of *slack variables*.

The first order KKT conditions are

$$\begin{aligned} \nabla_x \mathcal{L} = 0 \Rightarrow \frac{\partial \mathcal{L}}{\partial x_i} &= \frac{\partial f}{\partial x_i} - \sum_{j=1}^{\hat{m}} \hat{\lambda}_j \frac{\partial \hat{c}_j}{\partial x_i} - \sum_{k=1}^{m} \lambda_k \frac{\partial c_k}{\partial x_i} = 0, \quad i = 1, \ldots, n \\ \nabla_{\hat{\lambda}} \mathcal{L} = 0 \Rightarrow \frac{\partial \mathcal{L}}{\partial \hat{\lambda}_j} &= \hat{c}_j = 0, \quad j = 1, \ldots, \hat{m} \\ \nabla_{\lambda} \mathcal{L} = 0 \Rightarrow \frac{\partial \mathcal{L}}{\partial \lambda_k} &= c_k - s_k^2 = 0 \quad k = 1, \ldots, m \\ \nabla_{s} \mathcal{L} = 0 \Rightarrow \frac{\partial \mathcal{L}}{\partial s_k} &= \lambda_k s_k = 0, \quad k = 1, \ldots, m \\ \lambda_k &\geq 0, \quad k = 1, \ldots, m. \end{aligned}$$

Now we have $n + \hat{m} + 2m$ equations and two main possibilities for each inequality constraint k:

$s_k > 0$: the k-th constraint is inactive, and $\lambda_k = 0$.

$s_k = 0$: the k-th constraint is active, and $\lambda_k \neq 0$. λ_k must then be non-negative, otherwise from the first equations, the gradient of objective and gradient of constraint point in the same direction.

Sufficient conditions are obtained by examining the second-order requirements. The set of sufficient conditions is as follows:

1. KKT necessary conditions must be satisfied at x^*.
2. The Hessian matrix of the Lagrangian,

$$\nabla^2 \mathcal{L} = \nabla^2 f(x^*) - \sum_{j=1}^{\hat{m}} \hat{\lambda}_j \nabla^2 \hat{c}_j - \sum_{k=1}^{m} \lambda_k \nabla^2 c_k \tag{5.22}$$

is positive definite in the feasible space. This is a subspace of n-space and is defined as follows: any direction y that satisfies

$$y \neq 0 \tag{5.23}$$

$$\nabla \hat{c}_j^T(x^*) y = 0, \qquad \text{for all} \quad j = 1, \ldots, \hat{m} \tag{5.24}$$

$$\nabla c_k^T(x^*) y = 0, \qquad \text{for all} \quad k \quad \text{for which} \quad \lambda_k > 0. \tag{5.25}$$

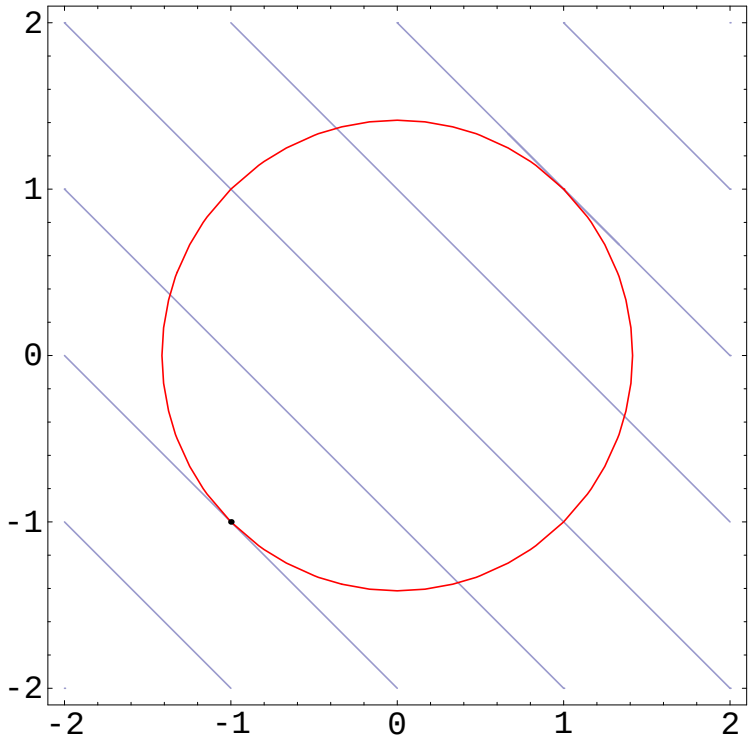


Figure 5.3

Then the Hessian of the Lagrangian in feasible space must be positive definite,

$$y^T \nabla^2 \mathcal{L}(x^*) y > 0. \tag{5.26}$$

Example 5.23. Problem with a Single Inequality Constraint

Suppose we now have the same problem, but with an inequality replacing the equality constraint,

$$\begin{aligned} \text{minimize } f(x) &= x_1 + x_2 \\ \text{s.t.} \quad c_1(x) &= 2 - x_1^2 - x_2^2 \geq 0 \end{aligned}$$

The feasible region is now the circle and its interior. Note that $\nabla c_1(x)$ now points towards the center of the circle.

Graphically, we can see that the solution is still $(-1, -1)^T$ and therefore $\lambda_1^* = 1/2$.

Given a point x that is not optimal, we can find a step d that both stays feasible and decreases the objective function f, to first order. As in the equality constrained case, the latter condition is expressed as

$$\boxed{\nabla f^T(x) d < 0}. \tag{5.27}$$

The first condition, however is slightly different, since the constraint is not necessarily zero, i.e.

$$c_1(x + d) \geq 0 \tag{5.28}$$

Performing a Taylor series expansion we have,

$$\underbrace{c_1(x + d)}_{\geq 0} \approx c_1(x) + \nabla c_1^T(x) d. \tag{5.29}$$

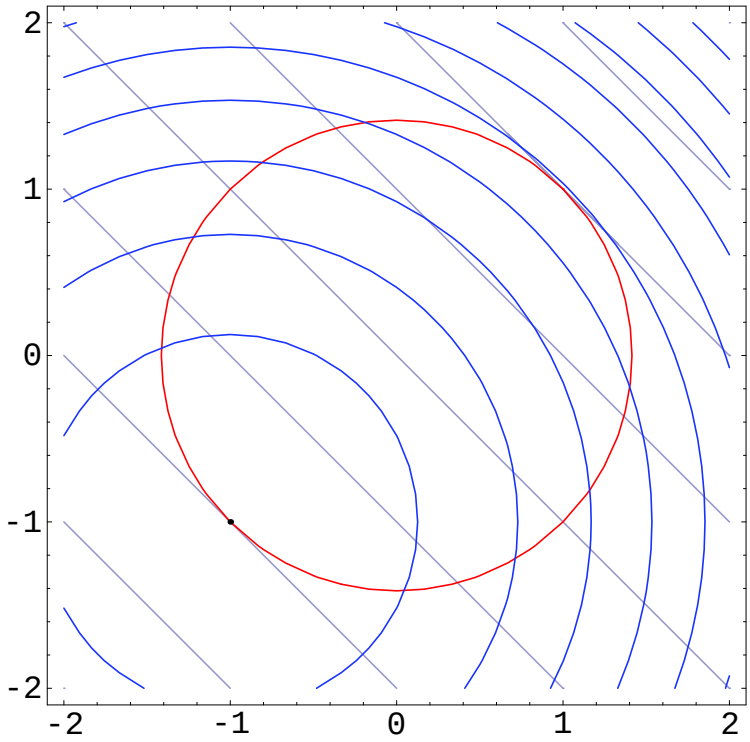


Figure 5.4: Contour plots of $f(x)$, $c_1 = 0$ and $\mathcal{L}(x, \lambda_1^*, s_1^*)$

Thus feasibility is retained to a first order if

$$\boxed{c_1(x) + \nabla c_1^T(x) d \geq 0}. \tag{5.30}$$

In order to find valid steps d it helps two consider two possibilities.

1. Suppose x lies strictly inside the circle ($c_1(x) > 0$). In this case, any vector d satisfies the condition (5.30), provided that its length is sufficiently small. The only situation that will prevent us from finding a descent direction is if $\nabla f(x) = 0$.

2. Consider now the case in which x lies on the boundary, i.e., $c_1(x) = 0$. The conditions thus become $\nabla f^T(x) d < 0$ and $\nabla c_1^T(x) d \geq 0$. The two regions defined by these conditions fail to intersect only when $\nabla f(x)$ and $\nabla c_1(x)$ point in the same direction, that is, when

$$\nabla f(x)^T d = \lambda_1 c_1(x), \qquad \text{for some } \lambda_1 \geq 0. \tag{5.31}$$

The optimality conditions for these two cases can again be summarized by using the Lagrangian function, that is,

$$\nabla_x \mathcal{L}(x^*, \lambda_1^*) = 0, \qquad \text{for some } \lambda_1^* \geq 0 \qquad \text{and } \lambda_1^* s_1^* = 0. \tag{5.32}$$

The last condition is known as a *complementarity condition* and implies that the Lagrange multiplier can be strictly positive only when the constraint is active.

Example 5.24. Lagrangian Whose Hessian is Not Positive Definite

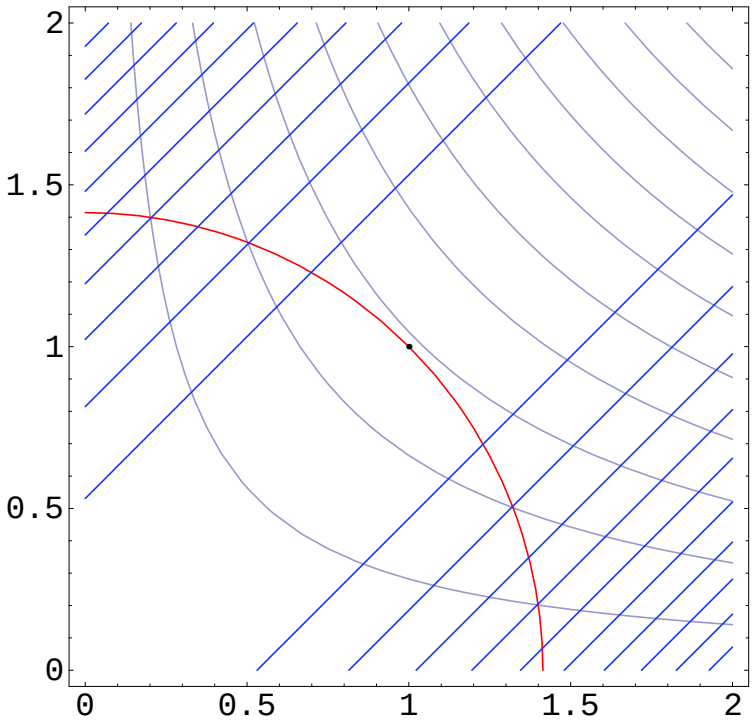


Figure 5.5: Contour plots of $f(x)$, $c_1 = 0$ and $\mathcal{L}(x, \lambda_1^*)$

$$\begin{aligned} \text{minimize} \quad & f(x) = -x_1 x_2 && (5.33)\\ \text{s.t.} \quad & \hat{c}_1(x) = 2 - x_1^2 - x_2^2 = 0 && (5.34)\\ & x_1 \geq 0, \quad x_2 \geq 0 && (5.35) \end{aligned}$$

If we solve this problem, we will find the Hessian of the Lagrangian shown in Fig. 5.5, which is positive semi-definite. However, it is positive definite in the feasible directions.

Example 5.25. Problem with Two Inequality Constraints

Suppose we now add another inequality constraint,

$$\begin{aligned} \text{minimize} \quad & f(x) = x_1 + x_2 && (5.36)\\ \text{s.t.} \quad & c_1(x) = 2 - x_1^2 - x_2^2 \geq 0, \qquad c_2(x) = x_2 \geq 0. && (5.37) \end{aligned}$$

The feasible region is now a half disk. Graphically, we can see that the solution is now $(-\sqrt{2}, 0)^T$ and that both constraints are active at this point.

The Lagrangian for this problem is

$$\mathcal{L}(x, \lambda, s) = f(x) - \lambda_1 \left(c_1(x) - s_1^2\right) - \lambda_2 \left(c_2(x) - s_2^2\right), \tag{5.38}$$

where $\lambda = (\lambda_1, \lambda_2)^T$ is the vector of Lagrange multipliers. The first order optimality conditions are thus,

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = 0, \qquad \text{for some } \lambda^* \geq 0. \tag{5.39}$$

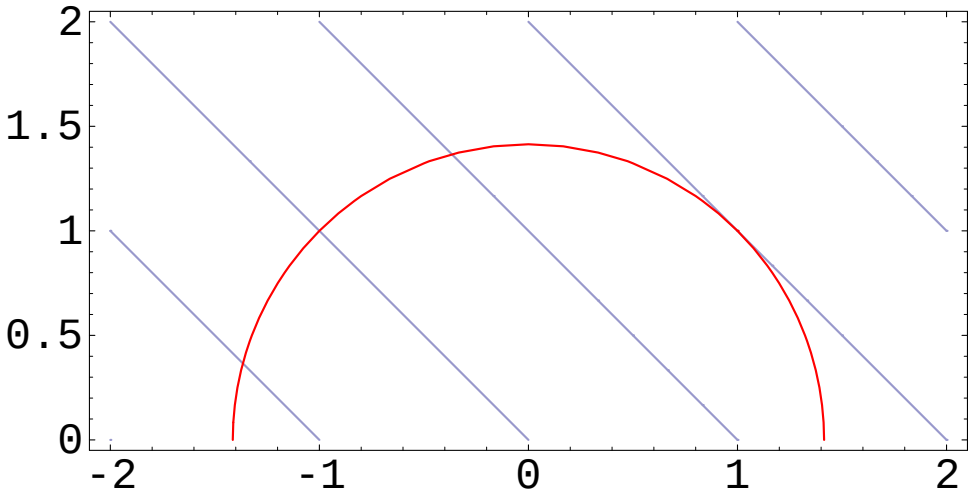


Figure 5.6

Applying the complementarity conditions to both inequality constraints,

$$\lambda_1^* s_1^* = 0, \qquad \text{and } \lambda_2^* s_2^* = 0. \tag{5.40}$$

For $x^* = (-\sqrt{2}, 0)^T$ we have,

$$\nabla f(x^*) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \nabla c_1(x^*) = \begin{bmatrix} 2\sqrt{2} \\ 0 \end{bmatrix}, \qquad \nabla c_2(x^*) = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

and $\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$ when

$$\lambda^* = \begin{bmatrix} \frac{1}{2\sqrt{2}} \\ 1 \end{bmatrix}.$$

Now lets consider other feasible points that are not optimal and examine the Lagrangian and its gradients at these points.

For point $x = (\sqrt{2}, 0)^T$, both constraints are again active. However, $\nabla f(x)$ no longer lies in the quadrant defined by $\nabla c_i(x)^T d \geq 0, \quad i = 1, 2$ and therefore there are descent directions that are feasible, like for example $d = (-1, 0)^T$.

Note that $\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$ at this point for $\lambda = (-\frac{1}{2\sqrt{2}}, 1)^T$. However, since λ_1 is negative, the first order conditions are not satisfied at this point.

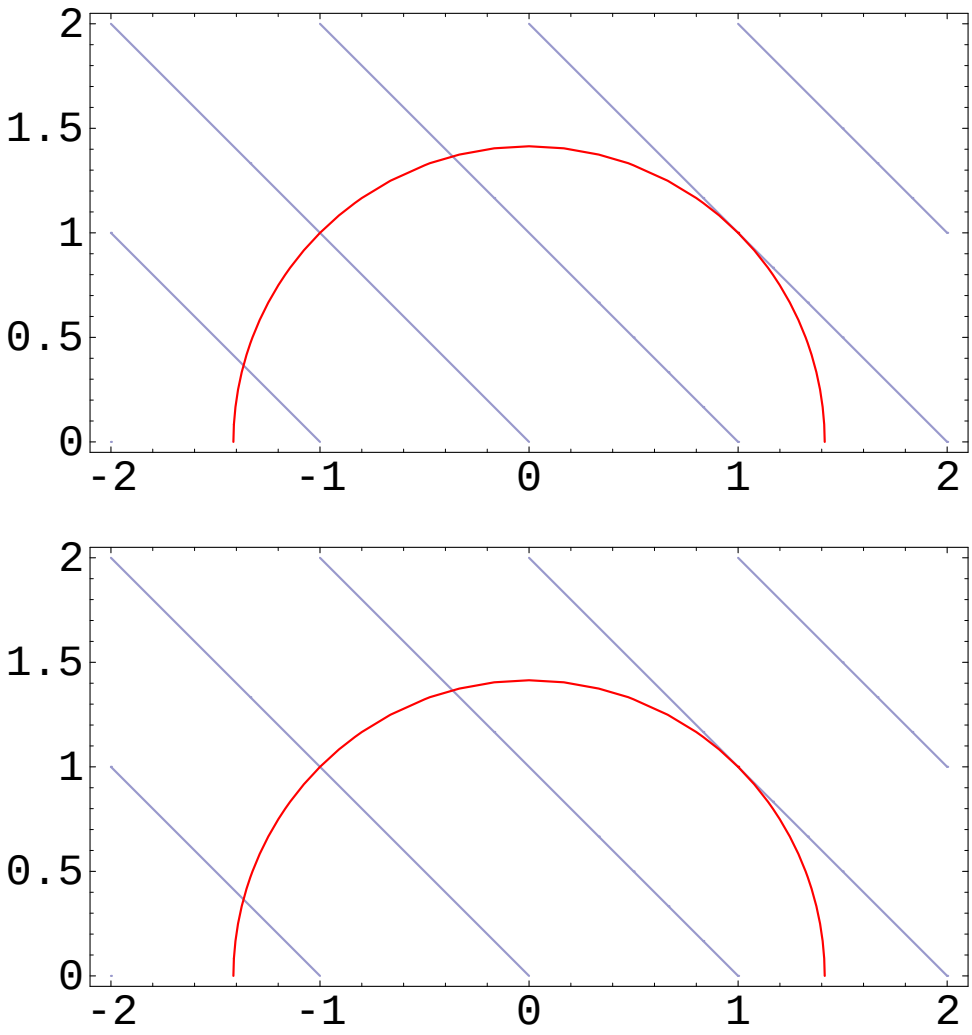
Now consider the point $x = (1, 0)^T$, for which only the second constraint is active. Linearizing f and c as before, d must satisfy the following to be a feasible descent direction,

$$\begin{aligned} c_1(x + d) &\geq 0 \Rightarrow 1 + \nabla c_1(x)^T d \geq 0, & (5.41) \\ c_2(x + d) &\geq 0 \Rightarrow \nabla c_2(x)^T d \geq 0, & (5.42) \\ f(x + d) - f(x) &< 0 \Rightarrow 1 + \nabla f(x)^T d < 0. & (5.43) \end{aligned}$$

We only need to worry about the last two conditions, since the first is always satisfied for a small enough step.

By noting that

$$\nabla f(x^*) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \nabla c_2(x^*) = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$



we can see that the vector $d = \left(-\frac{1}{2}, \frac{1}{4}\right)$, for example satisfies the two conditions.

Since $c_1(x) > 0$, we must have $\lambda_1 = 0$. In order to satisfy $\nabla_x \mathcal{L}(x, \lambda) = 0$ we would have to find λ_2 such that $\nabla f(x) = \lambda_2 \nabla c_2(x)$. No such λ_2 exists and this point is therefore not an optimum.

5.1.3 Constraint Qualification

The KKT conditions are derived using certain assumptions and depending on the problem, these assumptions might not hold.

A point x satisfying a set of constraints is a *regular point* if the gradient vectors of the *active* constraints, $\nabla c_j(x)$ are linearly independent.

To illustrate this, suppose we replaced the $\hat{c}_1(x)$ in the first example in this chapter by the equivalent condition

$$\hat{c}_1(x) = \left(x_1^2 + x_2^2 - 2\right)^2 = 0. \tag{5.44}$$

Then we have

$$\nabla \hat{c}_1(x) = \begin{bmatrix} 4(x_1^2 + x_2^2 - 2)x_1 \\ 4(x_1^2 + x_2^2 - 2)x_2 \end{bmatrix}, \tag{5.45}$$

so $\nabla \hat{c}_1(x) = 0$ for all feasible points and $\nabla f(x) = \hat{\lambda}_1 \nabla \hat{c}_1(x)$ cannot be satisfied. In other words,

there is no (finite) Lagrange multiplier that makes the objective gradient parallel to the constraint gradient, so we cannot solve the optimality conditions. This does not imply there is no solution; on the contrary, the solution remains unchanged for the earlier example. Instead, what it means is that most algorithms will fail, because they assume the constraints are linearly independent.

5.2 Penalty Function Methods

One of the ways of solving constrained optimization problems, at least approximately, is by adding a penalty function to the objective function that depends — in some logical way — on the value of the constraints.

The idea is to minimize a sequence of unconstrained minimization problems where the infeasibility of the constraints is minimized together with the objective function.

There two main types of penalization methods: *exterior penalty functions*, which impose a penalty for violation of constraints, and *interior penalty functions*, which impose a penalty for approaching the boundary of an inequality constraint.

5.2.1 Exterior Penalty Functions

The modified objective function is defined as the original one plus a term for each constraint, which is positive when the current point violates the constraint and zero otherwise.

Consider the equality-constrained problem:

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

where $\hat{c}(x)$ is an $\hat{m}$-dimensional vector whose j-th component is $\hat{c}_j(x)$. We assume that all functions are twice continuously differentiable.

We require a penalty for constraint violation to be a continuous function ϕ with the following properties

$$\begin{aligned} \phi(x) = 0 \quad \text{if} \quad & x \quad \text{is feasible} \\ \phi(x) > 0 \qquad & \text{otherwise,} \end{aligned}$$

The new objective function is

$$\pi(x, \rho) = f(x) + \rho\phi(x), \tag{5.46}$$

were ρ is positive and is called the penalty parameter.

The penalty method consists of solving a sequence of unconstrained minimization problems of the form

$$\begin{aligned} \text{minimize} \quad & \pi\left(x, \rho_k\right) \\ \text{w.r.t.} \quad & x \end{aligned}$$

for an increasing sequence of positive values of ρ_k tending to infinity. In general, for finite values of ρ_k, the minimizer of the penalty function violate the equality constraints. The increasing penalty forces the minimizer toward the feasible region.

The steps for this method are listed in Algorithm 10

Algorithm 10 General algorithm

Check termination conditions. if x_k satisfies the optimality conditions, he algorithm terminates successfully.

Minimize the penalty function. With x_k as the starting point, execute an algorithm to solve the unconstrained subproblem

$$\begin{aligned} \text{minimize} \quad & \pi(x, \rho_k) \\ \text{w.r.t.} \quad & x \end{aligned}$$

and let the solution of this subproblem be x_{k+1}.

Increase the penalty parameter. Set ρ_{k+1} to a larger value than ρ_k, set $k = k+1$ and return to 1.

The increase in the penalty parameter for each iteration can range from modest ($\rho_{k+1} = 1.4\rho_k$), to ambitious ($\rho_{k+1} = 10\rho_k$), depending on the problem.

The Quadratic Penalty Method

The quadratic penalty function is defined as

$$\pi(x, \rho) = f(x) + \frac{\rho}{2} \sum_{i=1}^{\hat{m}} \hat{c}_i(x)^2 = f(x) + \frac{\rho}{2} \hat{c}(x)^T \hat{c}(x). \tag{5.47}$$

The penalty is equal to the sum of the square of all the constraints and is therefore greater than zero when any constraint is violated and is zero when the point is feasible.

We can modify this method to handle inequality constraints by defining the penalty for these constraints as

$$\phi(x, \rho) = \rho \sum_{i=1}^{m} \left(\max \left[0, -c_i(x) \right] \right)^2 . \tag{5.48}$$

Penalty functions suffer from problems of ill conditioning. The solution of the modified problem approaches the true solution as $\lim_{\rho \to +\infty} x^*(\rho) = x^*$, but, as the penalty parameter increases, the condition number of the Hessian matrix of $\pi(x, \rho)$ increases and tends to ∞. This makes the problem increasingly difficult to solve numerically.

5.2.2 Interior Penalty Methods

Exterior penalty methods generate infeasible points and are therefore not suitable when feasibility has to be strictly maintained. This might be the case if the objective function is undefined or ill-defined outside the feasible region.

The method is analogous to the external penalty method: it creates a sequence of unconstrained modified differentiable functions whose unconstrained minima converge to the optimum solution of the constrained problem in the limit.

Consider the inequality-constrained problem:

$$\begin{aligned} \text{minimize} \quad & f(x) && (5.49) \\ \text{subject to} \quad & c(x) \geq 0 && (5.50) \end{aligned}$$

where $c(x)$ is an m-dimensional vector whose j-th component is $c_j(x)$. Again, we assume that all functions are twice continuously differentiable.

The Logarithmic Barrier Method

The logarithmic barrier function adds a penalty that tends to infinity as x approaches infeasibility. The function is defined as

$$\pi(x, \mu) = f(x) - \mu \sum_{j=1}^{m} \log\left(c_j(x)\right), \tag{5.51}$$

where the positive scalar μ is called the *barrier parameter*.

The Inverse Barrier Function

The inverse barrier function is defined as

$$\pi(x, \mu) = f(x) + \mu \sum_{j=1}^{m} \frac{1}{c_j(x)}, \tag{5.52}$$

and shares many of the same characteristics of the logarithmic barrier.

The solution of the modified problem for both functions approach the real solution as $\lim_{\mu \to 0} x^*(\mu) = x^*$. Again, the Hessian matrix becomes increasingly ill conditioned as μ approaches zero.

Similarly to the an exterior point method, an algorithm using these barrier functions finds the minimum of $\pi(x, \mu_k)$, for a given starting (feasible) point and terminates when norm of gradient is close to zero.

The algorithm then chooses a new barrier parameter μ_{k+1} and a new starting point, finds the minimum of the new problem and so on. A value of 0.1 for the ratio μ_{k+1}/μ_k is usually considered ambitious.

Example 5.26. Constrained Minimization Using a Quadratic Penalty Function

Consider the inequality constrained problem 5.8, introduced in the beginning of this chapter. Fig. 5.7 shows four different subproblems corresponding to increasing penalty parameters ρ. Recall, that this problem has two disjoint feasible regions, and has two local minima. For the lowest value of ρ, we obtain a function whose minimum is close to the global minimum. Although the optimization starts in the feasible region that has the local minimum, the solution of this first subproblem takes the solution past the local minimum towards the global minimum. As ρ increases, the other minimum appears, but by then we already are near the global minimum. As ρ increases further, the solution of each subproblem approaches the correct constrained minimum, but the penalized function becomes highly nonlinear, as can be seen from the abrupt change in the contour spacing.

5.3 Sequential Quadratic Programming (SQP)

To understand the use of SQP in problems with inequality constraints, we begin by considering the equality-constrained problem,

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && \hat{c}_j(x) = 0, \quad j = 1, \ldots, \hat{m} \end{aligned}$$

The idea of SQP is to model this problem at the current point x_k by a quadratic subproblem and to use the solution of this subproblem to find the new point x_{k+1}. SQP represents, in a way, the application of Newton's method to the KKT optimality conditions.

The Lagrangian function for this problem is $\mathcal{L}(x, \hat{\lambda}) = f(x) - \hat{\lambda}^T \hat{c}(x)$. We define the Jacobian of the constraints by

$$A(x)^T = \nabla \hat{c}(x)^T = [\nabla \hat{c}_1(x), \ldots, \nabla \hat{c}_{\hat{m}}(x)] \tag{5.53}$$

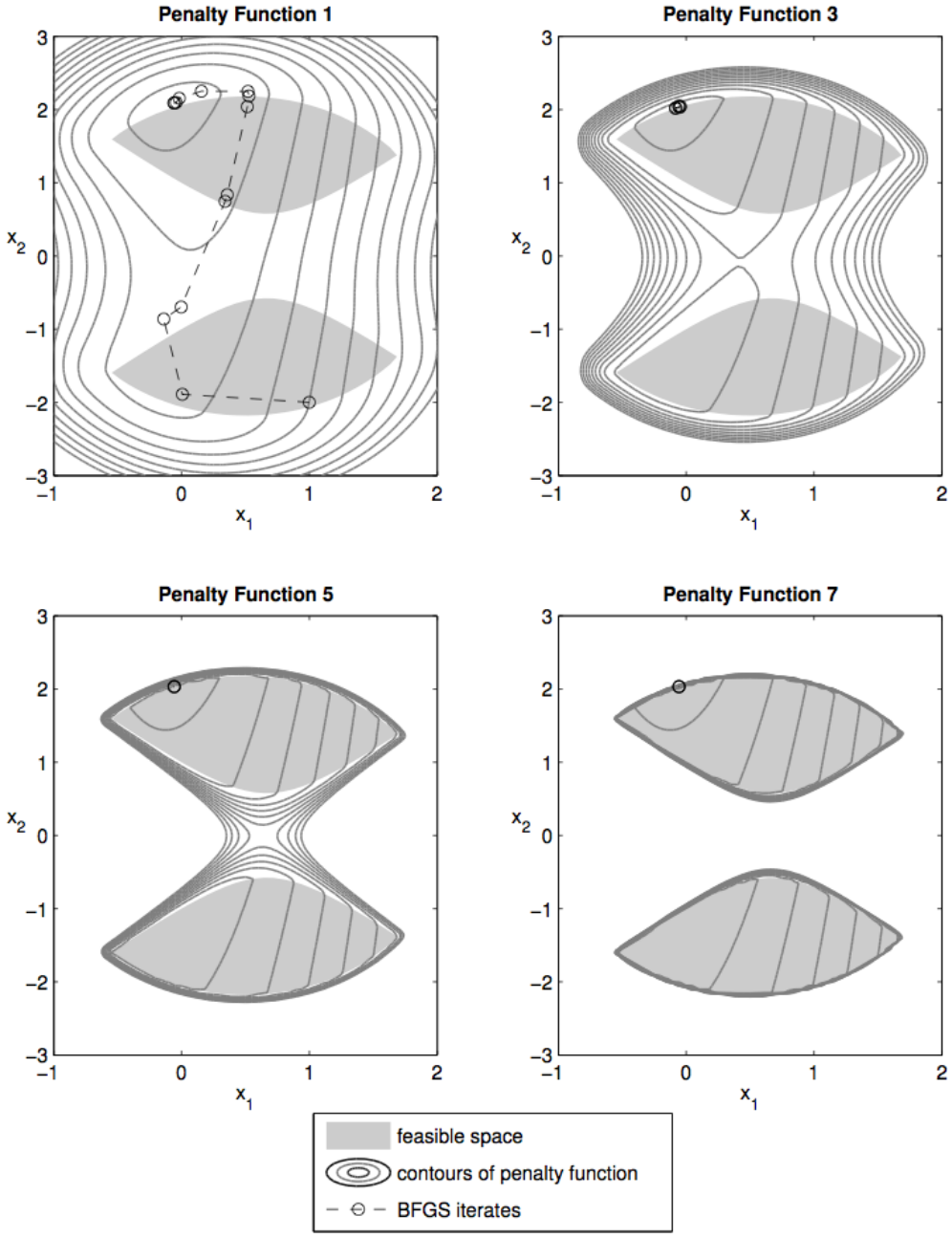


Figure 5.7: Solution of a constrained problem using a sequence of problems with quadratic penalty. Note that by starting with a low penalty parameter, the local optimum was avoided.

which is an $n \times m$ matrix and $g(x) \equiv \nabla f(x)$ is an n-vector as before. Note that A is generally not symmetric.

Applying the first order KKT conditions to this problem we obtain

$$\nabla \mathcal{L}(x, \hat{\lambda}) = 0 \Rightarrow \begin{bmatrix} g(x) - A(x)^T \hat{\lambda} \\ \hat{c}(x) \end{bmatrix} = 0 \tag{5.54}$$

This set of nonlinear equations can be solved using Newton's method,

$$\begin{bmatrix} W(x_k, \hat{\lambda}_k) & -A(x_k)^T \\ A(x_k) & 0 \end{bmatrix} \begin{bmatrix} p_k \\ p_{\hat{\lambda}} \end{bmatrix} = \begin{bmatrix} -g_k + A_k^T \hat{\lambda}_k \\ -\hat{c}_k \end{bmatrix} \tag{5.55}$$

where the Hessian of the Lagrangian is denoted by $W(x, \hat{\lambda}) = \nabla^2_{xx} \mathcal{L}(x, \hat{\lambda})$ and the Newton step from the current point is given by

$$\begin{bmatrix} x_{k+1} \\ \hat{\lambda}_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ \hat{\lambda}_k \end{bmatrix} + \begin{bmatrix} p_k \\ p_{\hat{\lambda}} \end{bmatrix}. \tag{5.56}$$

An alternative way of looking at this formulation of the SQP is to define the following quadratic problem at $(x_k, \hat{\lambda}_k)$

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} p^T W_k p + g_k^T p \\ \text{subject to} \quad & A_k p + \hat{c}_k = 0 \end{aligned}$$

This problem has a unique solution that satisfies

$$\begin{aligned} W_k p + g_k - A_k^T \hat{\lambda}_k = 0 \\ A_k p + \hat{c}_k = 0 \end{aligned}$$

By writing this in matrix form, we see that p_k and $\hat{\lambda}_k$ can be identified as the solution of the Newton equations we derived previously.

$$\begin{bmatrix} W_k & -A_k^T \\ A_k & 0 \end{bmatrix} \begin{bmatrix} p_k \\ \hat{\lambda}_{k+1} \end{bmatrix} = \begin{bmatrix} -g_k \\ -\hat{c}_k. \end{bmatrix} \tag{5.57}$$

This problem is equivalent to (5.55), but the second set of variables, is now the actual vector of Lagrange multipliers $\hat{\lambda}_{k+1}$ instead of the Lagrange multiplier step, $p_{\hat{\lambda}}$.

5.3.1 Quasi-Newton Approximations

Any SQP method relies on a choice of W_k (an approximation of the Hessian of the Lagrangian) in the quadratic model. When W_k is exact, then the SQP becomes the Newton method applied to the optimality conditions.

One way to approximate the Hessian of the Lagrangian would be to use a quasi-Newton approximation, such as the BFGS update formula. We could define,

$$s_k = x_{k+1} - x_k, \qquad y_k = \nabla_x \mathcal{L}(x_{k+1}, \lambda_{k+1}) - \nabla_x \mathcal{L}(x_k, \lambda_{k+1}), \tag{5.58}$$

and then compute the new approximation B_{k+1} using the same formula used in the unconstrained case.

If $\nabla^2_{xx}\mathcal{L}$ is positive definite at the sequence of points x_k, the method will converge rapidly, just as in the unconstrained case. If, however, $\nabla^2_{xx}\mathcal{L}$ is not positive definite, then using the BFGS update may not work well.

To ensure that the update is always well-defined the *damped BFGS updating for SQP* was devised. Using this scheme, we set

$$r_k = \theta_k y_k + (1-\theta_k) B_k s_k, \tag{5.59}$$

where the scalar θ_k is defined as

$$\theta_k = \begin{cases} 1 & \text{if} \quad s_k^T y_k \geq 0.2 s_k^T B_k s_k, \\ \frac{0.8 s_k^T B_k s_k}{s_k^T B_k s_k - s_k^T y_k} & \text{if} \quad s_k^T y_k < 0.2 s_k^T B_k s_k. \end{cases} \tag{5.60}$$

Then we can update B_{k+1} using,

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{r_k r_k^T}{s_k^T r_k}, \tag{5.61}$$

which is the standard BFGS update formula with y_k replaced by r_k. This guarantees that the Hessian approximation is positive definite.

Note that when $\theta_k = 0$ we have $B_{k+1} = B_k$, and that $\theta_k = 1$ yields an unmodified BFGS update. The modified method thus produces an interpolation between the current B_k and the one corresponding to BFGS. The choice of θ_k ensures that the new approximation stays close enough to the current approximation to guarantee positive definiteness.

In addition to using a different quasi-Newton update, SQP algorithms also need modifications to the line search criteria in order to ensure that the method converges from remote starting points. It is common to use a *merit function*, ϕ to control the size of the steps in the line search. The following is one of the possibilities for such a function:

$$\phi(x_k; \mu) = f(x) + \frac{1}{\mu} ||\hat{c}||_1 \tag{5.62}$$

The penalty parameter μ is positive and the L_1 norm of the equality constraints is

$$||\hat{c}||_1 = \sum_{j=1}^{\hat{m}} |\hat{c}_j|. \tag{5.63}$$

To determine the sequence of penalty parameters, the following strategy is often used

$$\mu_k = \begin{cases} \mu_{k-1} & \text{if} \quad \mu_{k-1}^{-1} \geq \gamma + \delta \\ (\gamma + 2\delta)^{-1} & \text{otherwise,} \end{cases} \tag{5.64}$$

where γ is set to $\max(\lambda_{k+1})$ and δ is a small tolerance that should be larger that the expected relative precision of your function evaluations.

The full procedure an SQP with line search is described in Algorithm 11, where D denotes the directional derivative in the p_k direction.

Algorithm 11 SQP algorithm

Input: Initial guess (x_0, λ_0), parameters $0 < \eta < 0.5$
Output: Optimum, x^*
  $k \leftarrow 0$
  Initialize the Hessian estimate, $B_0 \leftarrow I$
  repeat
    Compute p_k and $p_{\hat{\lambda}}$ by solving (5.55), with B_k in place of W_k
    Choose μ_k such that p_k is a descent direction for ϕ at x_k
    $\alpha_k \leftarrow 1$
    while $\phi(x_k + \alpha_k p_k, \mu_k) > \phi(x_k, \mu_k) + \eta \alpha_k D\left[\phi(x_k, p_k)\right]$ **do**
      $\alpha_k \leftarrow \tau_\alpha \alpha_k$ for some $0 < \tau_\alpha < 1$
    end while
    $x_{k+1} \leftarrow x_k + \alpha_k p_k$
    $\hat{\lambda}_{k+1} \leftarrow \hat{\lambda}_k + p_{\hat{\lambda}}$
    Evaluate f_{k+1}, g_{k+1}, c_{k+1} and A_{k+1}
    $s_k \leftarrow \alpha_k p_k$, $y_k \leftarrow \nabla_x \mathcal{L}(x_{k+1}, \lambda_{k+1}) - \nabla_x \mathcal{L}(x_k, \lambda_{k+1})$
    Obtain B_{k+1} by using a quasi-Newton update to B_k
    $k \leftarrow k + 1$
  until Convergence

5.3.2 Inequality Constraints

The SQP method can be extended to handle inequality constraints. Consider general nonlinear optimization problem

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & \hat{c}_j(x) = 0, \quad j = 1, \ldots, \hat{m} \\ & c_k(x) \geq 0, \quad k = 1, \ldots, m \end{aligned}$$

To define the subproblem we now linearize both the inequality and equality constraints and obtain,

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} p^T W_k p + g_k^T p \\ \text{subject to} \quad & \nabla \hat{c}_j(x)^T p + \hat{c}_j(x) = 0, \quad j = 1, \ldots, \hat{m} \\ & \nabla c_k(x)^T p + c_k(x) \geq 0, \quad k = 1, \ldots, m \end{aligned}$$

One of the most common type of strategy to solve this problem, the *active-set method*, is to consider only the active constraints at a given iteration and treat those as equality constraints. This is a significantly more difficult problem because we do not know *a priori* which inequality constraints are active at the solution. If we did, we could just solve the equality constrained problem considering only the active constraints.

The most commonly used active-set methods are feasible-point methods. These start with a feasible solution a never let the new point be infeasible.

Example 5.27. Constrained Minimization Using SQP

We consider the same inequality constrained problem 5.8 that we used to demonstrated the quadratic penalty approach. The sequence of subproblems determined by the SQP method is shown Fig. 5.8. The feasible regions are shown shaded, and the contours are those of the Lagrangian.

Note that these contours change for each subproblem due to a change in the Lagrange multipliers. The intermediate Lagrangian functions are not necessarily convex, but the final one is, with an unconstrained minimum at the constrained optimum. In this case, the algorithm converges to the global optimum, but it could easily have converged to a the local one.

Bibliography

[1] Stephen G. Nash and Ariela Sofer. *Linear and Nonlinear Programming*, chapter 15. McGrawHill, 1996.

[2] Stephen G. Nash and Ariela Sofer. *Linear and Nonlinear Programming*, chapter 16. McGrawHill, 1996.

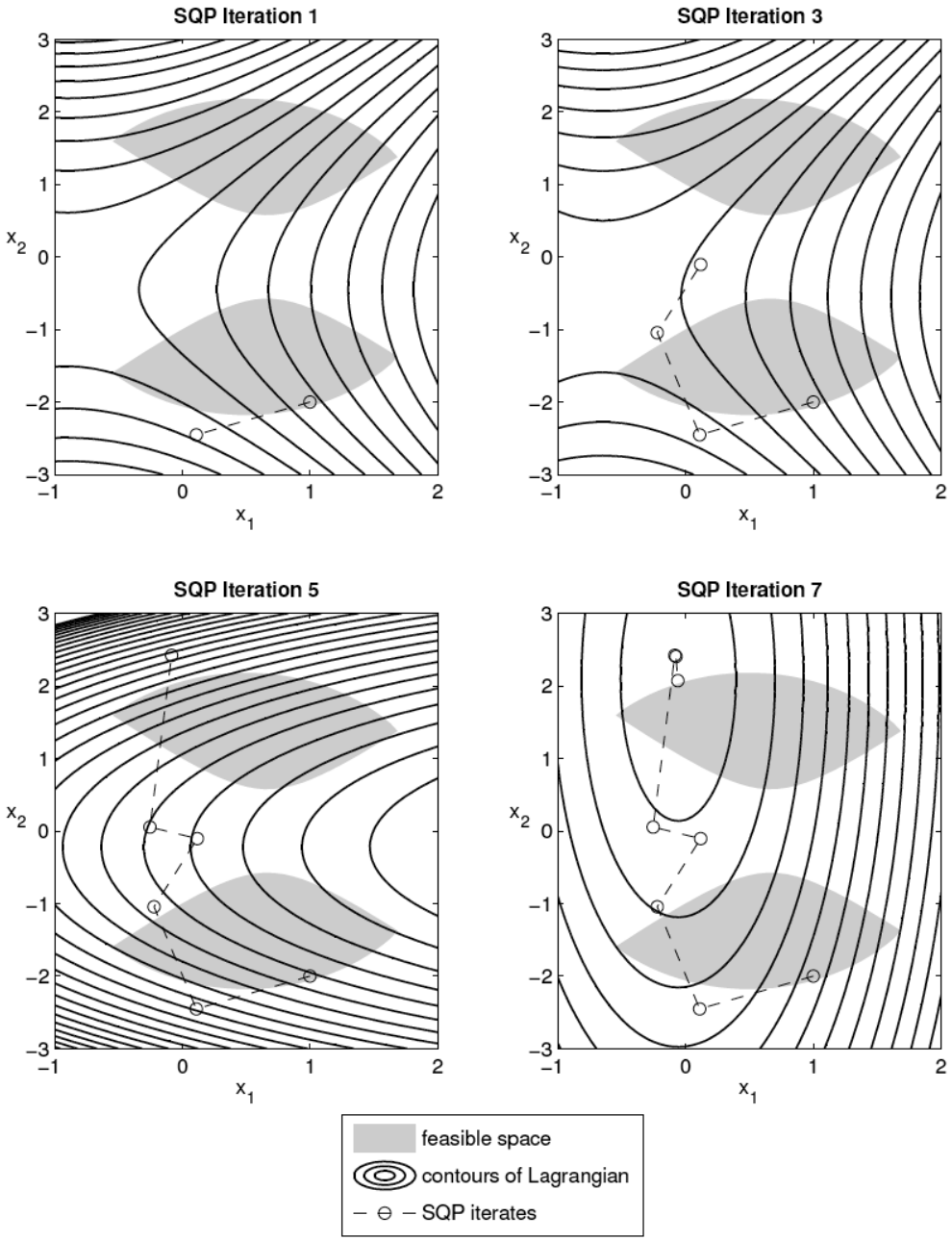


Figure 5.8: Solution of a constrained problem using SQP. Note that the Lagrangian is not always convex.