APPENDIX

Cost Function Optimization

C

In this appendix, we review a number of optimization schemes that have been encountered throughout the book.

Let $\boldsymbol{\theta}$ be an unknown parameter vector and $J(\boldsymbol{\theta})$ the corresponding cost function to be minimized. Function $J(\boldsymbol{\theta})$ is assumed to be differentiable

C.1 GRADIENT DESCENT ALGORITHM

The algorithm starts with an initial estimate $\boldsymbol{\theta}(0)$ of the minimum point and the subsequent algorithmic iterations are of the form

$$\boldsymbol{\theta}(new) = \boldsymbol{\theta}(old) + \Delta\boldsymbol{\theta} \tag{C.1}$$

$$\Delta\boldsymbol{\theta} = -\mu \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(old)} \tag{C.2}$$

where $\mu > 0$. If a maximum is sought, the method is known as *gradient ascent* and the minus sign in (C.2) is neglected.

Figure C.1 shows the geometric interpretation of the scheme. The new estimate $\boldsymbol{\theta}(new)$ is chosen in the direction that decreases $J(\boldsymbol{\theta})$. The parameter μ is very important and it plays a crucial role in the convergence of the algorithm. If it is too small, the corrections $\Delta\boldsymbol{\theta}$ are small and the convergence to the optimum point is very slow. On the other hand, if it is too large, the algorithm may oscillate around the optimum value and convergence is not possible. However, if the parameter is properly chosen, the algorithm converges to a stationary point of $J(\boldsymbol{\theta})$, which can be either, a local minimum ($\boldsymbol{\theta}_1^0$) or a global minimum ($\boldsymbol{\theta}^0$) or a saddle point ($\boldsymbol{\theta}_2^0$). In other words, it converges to a point where the gradient becomes zero (see Figure C.2). To which of the stationary points the algorithm will converge depends on the position of the initial point, relative to the stationary points. Furthermore, the convergence speed depends on the form of the cost $J(\boldsymbol{\theta})$. Figure C.3 shows the constant $J(\boldsymbol{\theta}) = c$ curves, for two cases and for different values of c, in the two-dimensional space, that is, $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$. The optimum $\boldsymbol{\theta}^0$ is located at the center of the curves. Recall that the gradient $\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is always vertical to the tangent to the

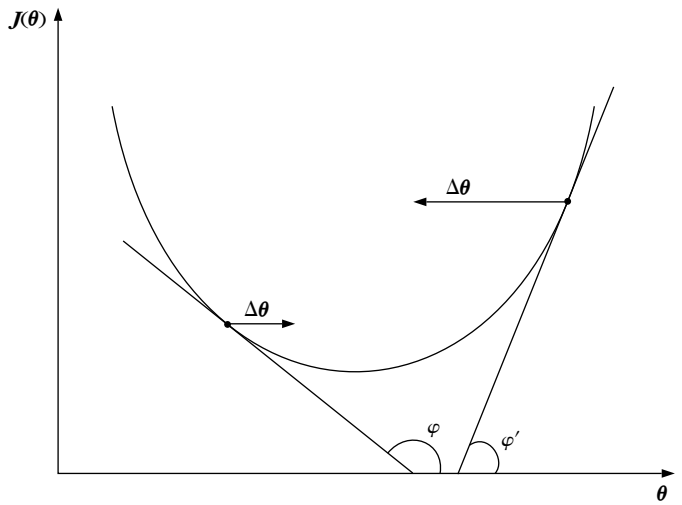

FIGURE C.1

In the gradient descent scheme, the correction of the parameters takes place in the direction that decreases the value of the cost function.

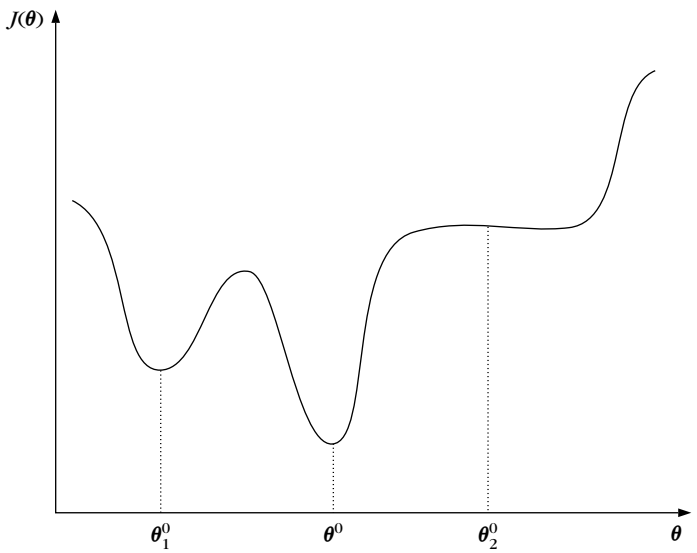

FIGURE C.2

A local minimum, a global minimum, and a saddle point of $J(\theta)$.

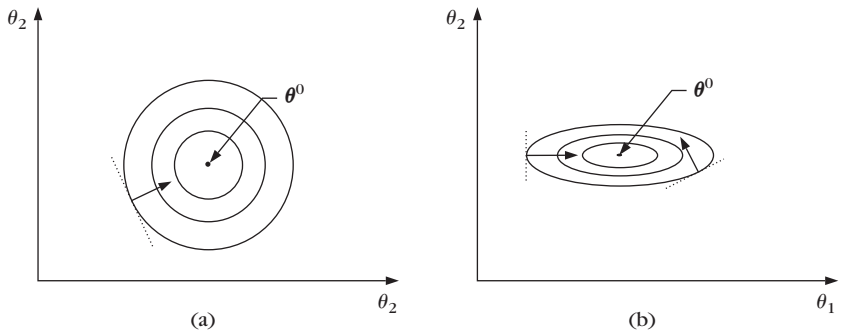

FIGURE C.3

Curves of constant cost values. In (a) the negative gradient always points to the optimum. In (b) it points to the optimum at only a few places, and convergence can be slow. The correction term can follow a zig zag path.

constant J curves. Indeed, if $J(\boldsymbol{\theta}) = c$, then

$$dc = 0 = \frac{\partial J(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} d\boldsymbol{\theta} \Longrightarrow \frac{J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \perp d\boldsymbol{\theta} \tag{C.3}$$

Furthermore, at each point $\boldsymbol{\theta}$ on a curve $J(\boldsymbol{\theta}) = c$, the gradient $\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ points to the direction of the maximum increase of $J(\boldsymbol{\theta})$. This is easily seen by writing

$$dJ = \frac{\partial J(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} d\boldsymbol{\theta} = |\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}||d\boldsymbol{\theta}| \cos\phi$$

where $\cos\phi$ is maximum for $\phi = 0$, that is, when the two involved vectors are parallel. Thus $\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ necessarily points to the direction of the maximum increase of $J(\boldsymbol{\theta})$. Hence, in the case of Figure C.3a the negative gradient, that is, the correction term, always points to the optimum (minimum) point. In principle, in such cases, convergence can be achieved in a single step. The scenario is different for the case of Figure C.3b. There, $\Delta\boldsymbol{\theta}$ points to the center at only very few places. Thus, convergence in this case can be quite slow and $\Delta\boldsymbol{\theta}$ can oscillate back and forth following a zigzag path until it rests at the optimum.

- *Quadratic surface*: Let $J(\boldsymbol{\theta})$ be of a quadratic form, that is,

$$J(\boldsymbol{\theta}) = b - \boldsymbol{p}^T\boldsymbol{\theta} + \frac{1}{2}\boldsymbol{\theta}^T R\boldsymbol{\theta} \tag{C.4}$$

  where R is assumed to be positive definite, in order (C.4) to have a (single) minimum (why?). Then,

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = R\boldsymbol{\theta} - \boldsymbol{p} \tag{C.5}$$

  Thus, the optimum value is given by

$$R\boldsymbol{\theta}^0 = \boldsymbol{p} \tag{C.6}$$

The tth iteration step in (C.1) then becomes

$$\boldsymbol{\theta}(t) = \boldsymbol{\theta}(t-1) - \mu\left(R\boldsymbol{\theta}(t-1) - \boldsymbol{p}\right) \tag{C.7}$$

Subtracting $\boldsymbol{\theta}^0$ from both sides and taking into account (C.6), (C.7) becomes

$$\tilde{\boldsymbol{\theta}}(t) = \tilde{\boldsymbol{\theta}}(t-1) - \mu R\tilde{\boldsymbol{\theta}}(t-1) = (I - \mu R)\tilde{\boldsymbol{\theta}}(t-1) \tag{C.8}$$

where $\tilde{\boldsymbol{\theta}}(t) \equiv \boldsymbol{\theta}(t) - \boldsymbol{\theta}^0$. Now let R be a symmetric matrix. Then, as we know from Appendix B, it can be diagonalized, that is,

$$R = \Phi^T \Lambda \Phi \tag{C.9}$$

where Φ is the orthogonal matrix with columns the orthonormal eigenvectors of R and Λ the diagonal matrix having the corresponding eigenvalues on its diagonal. Incorporating (C.9) into (C.8) we obtain

$$\hat{\boldsymbol{\theta}}(t) = (I - \mu\Lambda)\hat{\boldsymbol{\theta}}(t-1) \tag{C.10}$$

where $\hat{\boldsymbol{\theta}}(t) \equiv \Phi\tilde{\boldsymbol{\theta}}(t)$. Matrix $I - \mu\Lambda$ is now diagonal, and (C.10) is equivalent to

$$\hat{\boldsymbol{\theta}}_i(t) = (1 - \mu\lambda_i)\hat{\boldsymbol{\theta}}_i(t-1) \tag{C.11}$$

where $\hat{\boldsymbol{\theta}} \equiv [\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_l]^T$. Considering (C.11) for successive iteration steps we obtain

$$\hat{\theta}_i(t) = (1 - \mu\lambda_i)^t \hat{\theta}_i(0) \tag{C.12}$$

which converges to

$$\lim_{t\to\infty} \hat{\theta}_i(t) = 0, \Longrightarrow \lim_{t\to\infty} \theta_i(t) = \theta_i^0, \quad i = 1, 2, \ldots, l \tag{C.13}$$

provided that $|1 - \mu\lambda_i| < 1, i = 1, 2, \ldots, l$. Thus, we can conclude that

$$\boldsymbol{\theta} \longrightarrow \boldsymbol{\theta}^0, \quad \text{if } \mu < \frac{2}{\lambda_{\max}} \tag{C.14}$$

where $\lambda_{\max}$ is the maximum eigenvalue of R (which is positive since R is positive definite). Thus, the convergence speed of the gradient descent algorithm is controlled by the ratio $\lambda_{\min}/\lambda_{\max}$ as (C.12) and (C.14) suggest.

- *Nonquadratic cost functions*: If $J(\boldsymbol{\theta})$ is not quadratic, we can mobilize Taylor's theorem and assume that at some step near a stationary point, $\boldsymbol{\theta}^0$, $J(\boldsymbol{\theta})$ can be written approximately as

$$J(\boldsymbol{\theta}) = J(\boldsymbol{\theta}^0) + (\boldsymbol{\theta} - \boldsymbol{\theta}^0)^T \boldsymbol{g} + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^0)^T H(\boldsymbol{\theta} - \boldsymbol{\theta}^0) \tag{C.15}$$

where $\boldsymbol{g}$ is the gradient at $\boldsymbol{\theta}^0$ and H is the corresponding Hessian matrix, that is,

$$\boldsymbol{g} = \left.\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}, \quad H(i,j) = \left.\frac{\partial^2 J(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \tag{C.16}$$

Thus, in the neighborhood of $\boldsymbol{\theta}^0$, $J(\boldsymbol{\theta})$ is given approximately by a quadratic form and the convergence of the algorithm is controlled by the eigenvalues of the Hessian matrix.

C.2 NEWTON'S ALGORITHM

The problems associated with the dependence of the convergence speed on the eigenvalue spread can be overcome by using Newton's iterative scheme, where the correction in (C.2) is defined by

$$\Delta\boldsymbol{\theta} = -H^{-1}(old)\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(old)} \tag{C.17}$$

where $H(old)$ is the Hessian matrix computed at $\boldsymbol{\theta}(old)$. Newton's algorithm converges much faster than the gradient descent method and, practically, its speed is independent of the eigenvalue spread. Faster convergence can be demonstrated by looking at the approximation in (C.15). Taking the gradient results in

$$\frac{\partial J(\boldsymbol{\theta})}{\partial(\boldsymbol{\theta})} = \frac{\partial J(\boldsymbol{\theta})}{\partial(\boldsymbol{\theta})}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} + H(\boldsymbol{\theta} - \boldsymbol{\theta}^0) \tag{C.18}$$

Thus, the gradient is a linear function of $\boldsymbol{\theta}$ and hence the Hessian is constant, that is H. Having assumed that $\boldsymbol{\theta}^0$ is a stationary point, the first term on the right-hand side becomes zero. Now let $\boldsymbol{\theta} = \boldsymbol{\theta}(old)$. Then, according to Newton's iteration

$$\boldsymbol{\theta}(new) = \boldsymbol{\theta}(old) - H^{-1}(H(\boldsymbol{\theta}(old) - \boldsymbol{\theta}^0)) = \boldsymbol{\theta}^0 \tag{C.19}$$

Thus, the minimum is found in a single iteration. Of course, in practice, this is not true, as the approximations are not exactly valid. It is true, however, for quadratic costs.

Following a more formal proof (e.g., [Luen 84]), it can be shown that the convergence of Newton's algorithm is quadratic (i.e., the error at one step is proportional to the square of the previous step) while that of the gradient descent is linear. This speedup in convergence is achieved at increased computational cost, since Newton's algorithm requires the computation and then inversion of the Hessian matrix. Furthermore, numerical issues concerning the invertibility of H arise.

C.3 CONJUGATE-GRADIENT METHOD

Discussing the gradient descent method, we saw that, in general, a zigzag path is followed from the initial estimate to the optimum. This drawback is overcome by the following scheme, which results in improved convergence speed with respect to the gradient descent method. Compute the correction term according to the following rule:

$$\Delta\boldsymbol{\theta}(t) = \boldsymbol{g}(t) - \beta(t)\Delta\boldsymbol{\theta}(t-1) \tag{C.20}$$

where

$$\boldsymbol{g}(t) = \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)} \tag{C.21}$$

and

$$\beta(t) = \frac{\boldsymbol{g}^T(t)\boldsymbol{g}(t)}{\boldsymbol{g}^T(t-1)\boldsymbol{g}(t-1)} \tag{C.22}$$

or

$$\beta(t) = \frac{\boldsymbol{g}^T(t)\big(\boldsymbol{g}(t) - \boldsymbol{g}(t-1)\big)}{\boldsymbol{g}^T(t-1)\boldsymbol{g}(t-1)} \tag{C.23}$$

The former is known as the Fletcher–Reeves and the latter as the Polak–Ribiere formula.

For a more rigorous treatment of the topic the reader is referred to [Luen 84]. Finally, it must be stated that a number of variants of these schemes have appeared in the literature.

C.4 OPTIMIZATION FOR CONSTRAINED PROBLEMS

C.4.1 Equality Constraints

We will first focus on linear equality constraints and then generalize to the nonlinear case. Although the philosophy for both cases is the same, it is easier to grasp the basics when linear constraints are involved. Thus the problem is cast as

$$\begin{aligned} &\text{minimize} \quad J(\boldsymbol{\theta}) \\ &\text{subject to} \quad A\boldsymbol{\theta} = \boldsymbol{b} \end{aligned}$$

where A is an $m \times l$ matrix and $\boldsymbol{b}, \boldsymbol{\theta}$ are $m \times 1$ and $l \times 1$ vectors, respectively. It is assumed that the cost function $J(\boldsymbol{\theta})$ is twice continuously differentiable and it is, in general, a nonlinear function. Furthermore, we assume that the rows of A are linearly independent, hence A has full row rank. This assumption is known as the *regularity assumption*.

Let $\boldsymbol{\theta}_*$ be a local minimizer of $J(\boldsymbol{\theta})$ over the set $\{\boldsymbol{\theta}: A\boldsymbol{\theta} = \boldsymbol{b}\}$. Then it is not difficult to show (e.g., [Nash 96]) that, at this point, the gradient of $J(\boldsymbol{\theta})$ is given by

$$\frac{\partial}{\partial \boldsymbol{\theta}}(J(\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*} = A^T\lambda \tag{C.24}$$

where $\boldsymbol{\lambda} \equiv [\lambda_1, \ldots, \lambda_m]^T$. Taking into account that

$$\frac{\partial}{\partial \boldsymbol{\theta}}(A\boldsymbol{\theta}) = A^T \tag{C.25}$$

Eq. (C.24) states that, at a constrained minimum, the gradient of the cost function is a linear combination of the gradients of the constraints. This is quite natural. Let us take a simple example involving a single linear constraint, that is,

$$\boldsymbol{a}^T\boldsymbol{\theta} = b$$

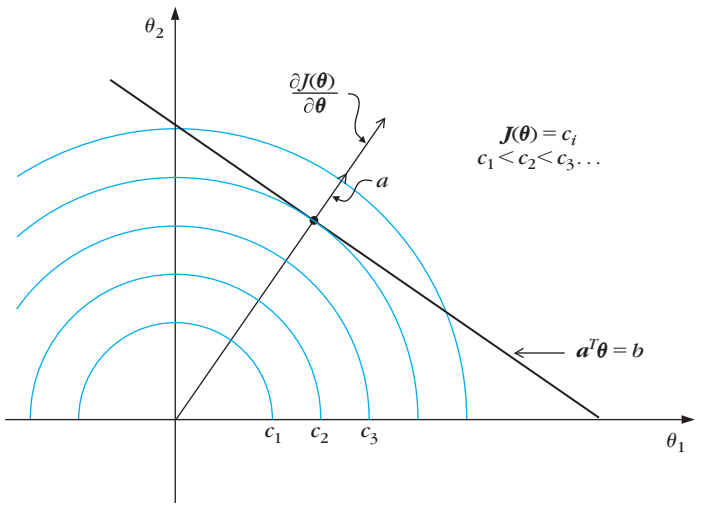

FIGURE C.4

At the minimizer, the gradient of the cost function is in the direction of the gradient of the constraint function.

Equation (C.24) then becomes

$$\frac{\partial}{\partial \boldsymbol{\theta}}(J(\boldsymbol{\theta}_*)) = \lambda \boldsymbol{a}$$

where the parameter λ is now a scalar. Figure C.4 shows an example of isovalue contours of $J(\boldsymbol{\theta}) = c$ in the two-dimensional space ($l = 2$). The constrained minimum coincides with the point where the straight line "meets" the isovalue contours for the first time, as one moves from small to large values of c. This is the point where the line is tangent to an isovalue contour; hence at this point the gradient of the cost function is in the direction of $\boldsymbol{a}$ (see Chapter 3).

Let us now define the function

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = J(\boldsymbol{\theta}) - \boldsymbol{\lambda}^T(A\boldsymbol{\theta} - \boldsymbol{b}) \tag{C.26}$$

$$= J(\boldsymbol{\theta}) - \sum_{i=1}^{m} \lambda_i(\boldsymbol{a}_i^T\boldsymbol{\theta} - b_i) \tag{C.27}$$

where $\boldsymbol{a}_i^T$, $i = 1, 2, \ldots, m$, are the rows of A. $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is known as the *Lagrangian function* and the coefficients, λ_i, $i = 1, 2, \ldots, m$, as the *Lagrange multipliers*. The optimality condition (C.24), together with the constraints, which the minimizer has to satisfy, can now be written in a compact form as

$$\nabla \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbf{0} \tag{C.28}$$

where ∇ denotes the gradient operation with respect to both $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$. Indeed, equating with zero the derivatives of the Lagrangian with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ gives,

respectively,

$$\frac{\partial}{\partial \boldsymbol{\theta}} J(\boldsymbol{\theta}) = A^T \boldsymbol{\lambda}$$
$$A\boldsymbol{\theta} = \boldsymbol{b}$$

The above is a set of $m + l$ unknowns, that is, $(\theta_1, \ldots, \theta_l, \lambda_1, \ldots, \lambda_m)$, with $m + l$ equations, whose solution provides the minimizer $\boldsymbol{\theta}_*$ and the corresponding Lagrange multipliers. Similar arguments hold for nonlinear equation constraints. Let us consider the problem

$$\begin{aligned} &\text{minimize} && J(\boldsymbol{\theta}) \\ &\text{subject to} && f_i(\boldsymbol{\theta}) = 0, \quad i = 1, 2, \ldots, m \end{aligned}$$

The minimizer is again a *stationary point* of the corresponding Lagrangian

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = J(\boldsymbol{\theta}) - \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{\theta})$$

and it results from the solution of the set of $m + l$ equations

$$\nabla \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbf{0}$$

The regularity condition for nonlinear constraints requires the gradients of the constraints $\frac{\partial}{\partial \boldsymbol{\theta}}(f_i(\boldsymbol{\theta}))$ to be linearly independent.

C.4.2 Inequality Constraints

The general problem can be cast as follows:

$$\begin{aligned} &\text{minimize} && J(\boldsymbol{\theta}) \\ &\text{subject to} && f_i(\boldsymbol{\theta}) \geq 0, \quad i = 1, 2, \ldots, m \end{aligned} \tag{C.29}$$

Each one of the constraints defines a region in $\mathcal{R}^l$. The intersection of all these regions defines the area in which the constrained minimum, $\boldsymbol{\theta}_*$, must lie. This is known as the *feasible region* and the points in it (candidate solutions) as *feasible points*. The type of the constraints control the type of the feasible region, that is, whether it is convex or concave. At this point, it will not harm us to recall a few definitions.

Convex functions. A function $f(\boldsymbol{\theta})$

$$f: S \subseteq \mathcal{R}^l \longrightarrow \mathcal{R}$$

is called convex in S, if for every $\boldsymbol{b}\theta$ and $\boldsymbol{\theta}' \in S$

$$f(\lambda\boldsymbol{\theta} + (1 - \lambda)\boldsymbol{\theta}') \leq \lambda f(\boldsymbol{\theta}) + (1 - \lambda) f(\boldsymbol{\theta}')$$

for every $\lambda \in [0, 1]$. If strict inequality holds, we say that the function is strict convex.

Concave functions. A function $f(\boldsymbol{\theta})$ is called concave, if for every $\boldsymbol{\theta}, \boldsymbol{\theta}' \in S$

$$f(\lambda\boldsymbol{\theta} + (1 - \lambda)\boldsymbol{\theta}') \geq \lambda f(\boldsymbol{\theta}) + (1 - \lambda) f(\boldsymbol{\theta}')$$

for every $\lambda \in [0, 1]$. For strict inequality, the function is known as strict concave.

Figure C.5 shows three functions, one convex, one concave, and one which is neither convex nor concave.

Convex sets. A set $S \subseteq \mathcal{R}^l$ is called convex, if for every pair of points $\boldsymbol{\theta}, \boldsymbol{\theta}' \in S$, the line segment joining these points also belongs to the set. In other words, all points $\lambda\boldsymbol{\theta} + (1 - \lambda)\boldsymbol{\theta}'$, $\lambda \in [0, 1]$ belong to the set. Figure C.6 shows two sets, one convex and one nonconvex.

Remarks

- If $f(\boldsymbol{\theta})$ is convex then $-f(\boldsymbol{\theta})$ is concave and vice versa. Furthermore, if $f_i(\boldsymbol{\theta})$, $i = 1, 2, \ldots, m$, are convex, so is the sum $\sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{\theta}), \lambda_i \geq 0$. Similarly, if $f_i(\boldsymbol{\theta})$ are concave, so is their summation.

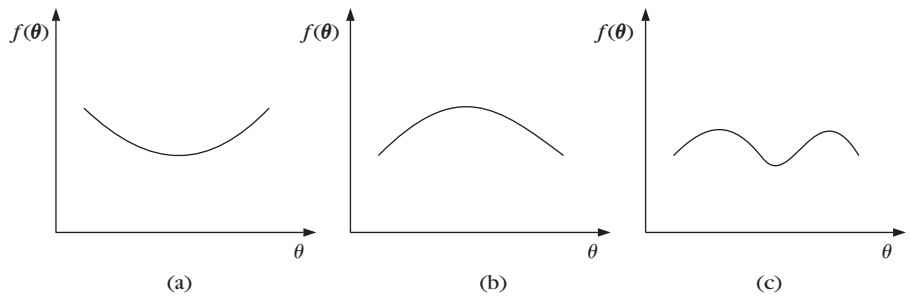

FIGURE C.5

(a) A convex function, (b) a concave function, and (c) a function that is neither convex nor concave.

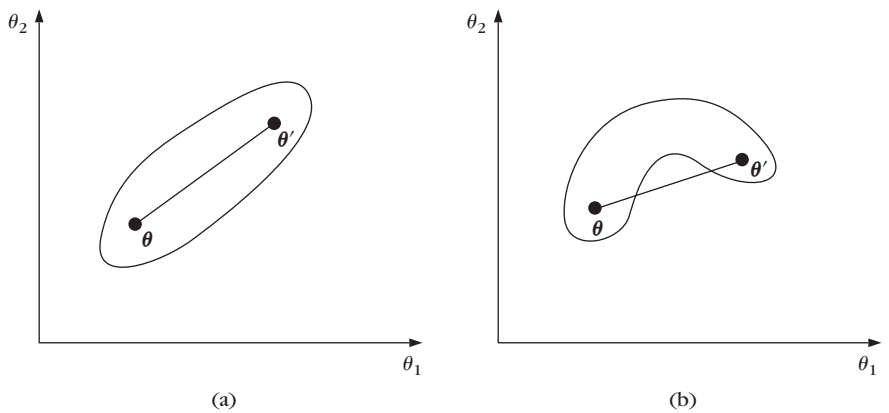

FIGURE C.6

(a) A convex set and (b) a concave set of points.

- If a function $f(\boldsymbol{\theta})$ is convex, it can be shown that *a local minimum is also a global one. This can be easily checked from the graph of Figure C.5. Furthermore, if the function is strict convex then this minimum is unique*. For concave functions, the above also hold true but for points where a maximum occurs.
- A direct consequence of the respective definitions is that if $f(\boldsymbol{\theta})$ is convex then the set

$$X = \{\boldsymbol{\theta} | f(\boldsymbol{\theta}) \leq b,\ b \in \mathcal{R}\}$$

  is convex. Also, if $f(\boldsymbol{\theta})$ is concave then the set

$$X = \{\boldsymbol{\theta} | f(\boldsymbol{\theta}) \geq b,\ b \in \mathcal{R}\}$$

  is also convex.
- The intersection of convex sets is also a convex set.

From the above remarks, one can easily conclude that, if each one of the functions in the constraints in (C.29) is concave, then the feasible region is a convex one. This is also valid if the constraints are linear, since a linear function can be considered either as convex or concave. For more on these issues, the interested reader may refer, for example, to [Baza 79].

The Karush–Kuhn–Tucker (KKT) Conditions

This is a set of *necessary* conditions, which a local minimizer $\boldsymbol{\theta}_*$ of the problem given in (C.29) has to satisfy. If $\boldsymbol{\theta}_*$ is a point that satisfies the regularity condition, then there exists a vector $\boldsymbol{\lambda}$ of Lagrange multipliers so that the following are valid:

$$\begin{aligned} &(1) \quad \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_*, \boldsymbol{\lambda}) = \mathbf{0} \\ &(2) \quad \lambda_i \geq 0, \quad i = 1, 2, \ldots, m \\ &(3) \quad \lambda_i f_i(\boldsymbol{\theta}_*) = 0, \quad i = 1, 2, \ldots, m \end{aligned} \tag{C.30}$$

Actually, there is a fourth condition concerning the Hessian of the Lagrangian function, which is not of interest to us. The above set of equations is also part of the sufficiency conditions; however, in this case, there are a few subtle points and the interested reader is referred to more specialized textbooks, for example, [Baza 79, Flet 87, Bert 95, Nash 96].

Conditions (3) in (C.30) are known as *complementary slackness conditions*. They state that at least one of the terms in the products is zero. In the case where, in each one of the equations, only one of the two terms is zero, that is, either λ_i or $f_i(\boldsymbol{\theta}_*)$, we talk about *strict complementarity*.

Remarks

- The first condition is most natural. It states that the minimum must be a stationary point of the Lagrangian, with respect to $\boldsymbol{\theta}$.

- A constraint, ($f_i(\boldsymbol{\theta}_*)$), is called *inactive* if the corresponding Lagrange multiplier is zero. This is because this constraint does not affect the problem. A constrained minimizer $\boldsymbol{\theta}_*$ can lie either in the interior of the feasible region or on its boundary. In the former case, the problem is equivalent to an unconstrained one. Indeed, if it happens that a minimum is located within the feasible region, then the value of the cost function in a region around this point will increase (or remain the same) as one moves away from this point. Hence, this point will be a stationary point of the cost function $J(\boldsymbol{\theta})$. Thus in this case, the constraints are redundant and do not affect the problem. In words, the constraints are inactive and this is equivalent to setting the Lagrange multipliers equal to zero. The nontrivial constrained optimization task is when the (unconstrained) minimum of the cost function is located outside the feasible region. In this case, the constrained minimum will be located on the boundary of the feasible region. In other words, in this nontrivial case, there will be one or more of the constraints for which $f_i(\boldsymbol{\theta}_*) = 0$. These constitute the *active constraints*. The rest of the constraints will be inactive with the corresponding Lagrange multipliers being zero.

  Figure C.7 illustrates a simple case with the following constraints:

$$f_1(\boldsymbol{\theta}) = \theta_1 + 2\theta_2 - 2 \geq 0$$
$$f_2(\boldsymbol{\theta}) = \theta_1 - \theta_2 + 2 \geq 0$$
$$f_3(\boldsymbol{\theta}) = -\theta_1 + 2 \geq 0$$

  The (unconstrained) minimum of the cost function is located outside the feasible region. The dotted lines are the isovalue curves $J(\boldsymbol{\theta}) = c$, with $c_1 < c_2 < c_3$. The constrained minimum coincides with the point where an isovalue curve "touches" the boundary of the feasible region for the first time (smallest value of c). This point may belong to more than one of the constraints, for example, it may be a corner point of the boundary.

- The Lagrange multipliers of the active constraints are *nonnegative*. To understand why this is so, let us consider for simplicity the case of linear constraints $A\boldsymbol{\theta} \geq \boldsymbol{b}$, where A includes the active constraints only. If $\boldsymbol{\theta}_*$ is a minimizer lying on the active constraints, then any other feasible point can be written as

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_* + \boldsymbol{p}$$
$$A\boldsymbol{p} \geq \mathbf{0}$$

  since this guarantees that $A\hat{\boldsymbol{\theta}} \geq \boldsymbol{b}$. If the direction $\boldsymbol{p}$ points into the feasible region (Figure C.7) then $A\boldsymbol{p} \neq \mathbf{0}$, that is, some of its components are strictly positive. Since $\boldsymbol{\theta}_*$ is a minimizer, from condition (1) in (C.30) we have that

$$\frac{\partial}{\partial \boldsymbol{\theta}} J(\boldsymbol{\theta}_*) = A^T \boldsymbol{\lambda}$$

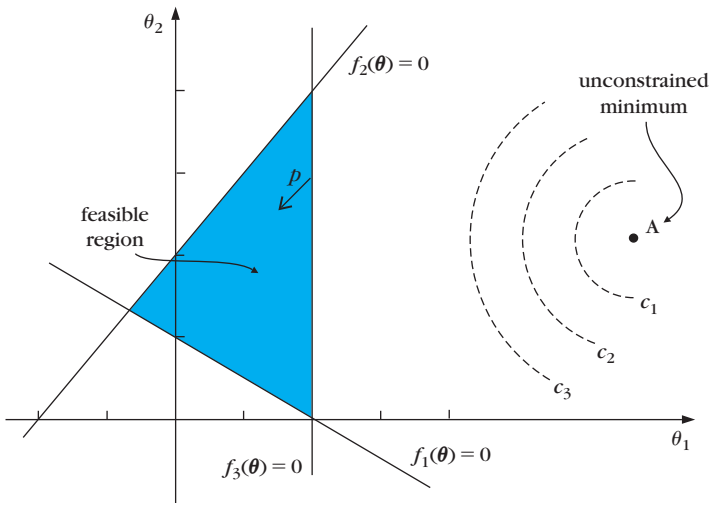

FIGURE C.7

An example of the nontrivial case, where the unconstrained minimum lies outside the feasible region.

The change of the cost function along the direction of $\boldsymbol{p}$ is proportional to

$$\boldsymbol{p}^T \frac{\partial}{\partial \boldsymbol{\theta}}(J(\boldsymbol{\theta})) = \boldsymbol{p}^T A^T \boldsymbol{\lambda}$$

and since $\boldsymbol{\theta}_*$ is a minimizer, this must be a direction of ascent at $\boldsymbol{\theta}_*$. Thus $\boldsymbol{\lambda}$ must be nonnegative to guarantee that $\boldsymbol{p}^T A^T \boldsymbol{\lambda} \geq 0$ for any $\boldsymbol{p}$ pointing into the feasible region. An active constraint whose corresponding Lagrange multiplier is zero is known as *degenerate*.

- It can be shown that, if the cost function is convex and the feasible region is also convex, then a local minimum is also a global one. A little thought (and a look at Figure C.7) suffices to see why this is so.

Having now discussed all these nice properties, the major question arises: how can one compute a constrained (local) minimum? Unfortunately, this is not always an easy task. A straightforward approach would be to assume that some of the constraints are active and some inactive, and check if the resulting Lagrange multipliers of the active constraints are nonnegative. If not, then choose another combination of constraints and repeat the procedure until one ends up with nonnegative multipliers. However, in practice, this may require a prohibitive amount of computation. Instead, a number of alternative approaches have been proposed. In the sequel, we will review some basics from Game Theory and use these to reformulate the KKT conditions. This new setup can be useful in a number of cases in practice.

Min-Max Duality

Let us consider two players, namely X and Y, playing a game. Player X will choose a strategy, say, x and simultaneously player Y will choose a strategy y. As a result, X will pay to Y the amount $\mathcal{F}(x, y)$, which can also be negative, that is, X wins. Let us now follow their thinking, prior to their final choice of strategy, assuming that the players are good professionals.

X: If Y knew that I was going to choose x, then, since he/she is a clever player, he/she would choose y to make his/her profit maximum, that is,

$$\mathcal{F}^*(x) = \max_y \mathcal{F}(x, y)$$

Thus, in order to make my *worst-case payoff* to Y minimum, I have to choose x so as to minimize $\mathcal{F}^*(x)$, that is,

$$\min_x \mathcal{F}^*(x)$$

This problem is known as the *min-max* problem since it seeks the value

$$\min_x \max_y \mathcal{F}(x, y)$$

Y: X is a good player, so if he/she knew that I am going to play y, he/she would choose x so that to make his/her payoff minimum, that is,

$$\mathcal{F}_*(y) = \min_x \mathcal{F}(x, y)$$

Thus, in order to make my *worst-case profit* maximum I must choose y that maximizes $\mathcal{F}_*(y)$, that is,

$$\max_y \mathcal{F}_*(y)$$

This is known as the max-min problem, since it seeks the value

$$\max_y \min_x \mathcal{F}(x, y)$$

The two problems are said to be *dual to each other*. The first is known to be the *primal*, whose objective is to minimize $\mathcal{F}^*(x)$ and the second is the *dual* problem with the objective to maximize $\mathcal{F}_*(y)$.

For any x and y, the following is valid:

$$\mathcal{F}_*(y) \equiv \min_x \mathcal{F}(x, y) \leq \mathcal{F}(x, y) \leq \max_y \mathcal{F}(x, y) \equiv \mathcal{F}^*(x) \tag{C.31}$$

which easily leads to

$$\max_y \min_x \mathcal{F}(x, y) \leq \min_x \max_y \mathcal{F}(x, y) \tag{C.32}$$

Saddle Point Condition

Let $\mathcal{F}(\boldsymbol{x}, \boldsymbol{y})$ be a function of two vector variables with $\boldsymbol{x} \in X \subseteq \mathcal{R}^l$ and $\boldsymbol{y} \in Y \subseteq \mathcal{R}^l$. If a pair of points $(\boldsymbol{x}_*, \boldsymbol{y}_*)$, with $\boldsymbol{x}_* \in X, \boldsymbol{y}_* \in Y$ satisfies the condition

$$\mathcal{F}(\boldsymbol{x}_*, \boldsymbol{y}) \leq \mathcal{F}(\boldsymbol{x}_*, \boldsymbol{y}_*) \leq \mathcal{F}(\boldsymbol{x}, \boldsymbol{y}_*) \tag{C.33}$$

for every $\boldsymbol{x} \in X$ and $\boldsymbol{y} \in Y$, we say that it satisfies the *saddle point condition*. It is not difficult to show (e.g., [Nash 96]) that a pair $(\boldsymbol{x}_*, \boldsymbol{y}_*)$ satisfies the saddle point conditions *if and only if*

$$\max_{\boldsymbol{y}} \min_{\boldsymbol{x}} \mathcal{F}(\boldsymbol{x}, \boldsymbol{y}) = \min_{\boldsymbol{x}} \max_{\boldsymbol{y}} \mathcal{F}(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{F}(\boldsymbol{x}_*, \boldsymbol{y}_*) \tag{C.34}$$

Lagrangian Duality

We will now use all the above in order to formulate our original cost function minimization problem as a min-max task of the corresponding Lagrangian function. Under certain conditions, this formulation can lead to computational savings when computing the constrained minimum. The optimization task of our interest is

$$\begin{aligned} &\text{minimize} \quad J(\boldsymbol{\theta}) \\ &\text{subject to} \quad f_i(\boldsymbol{\theta}) \geq 0, \quad i = 1, 2, \ldots, m \end{aligned}$$

The Lagrangian function is

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = J(\boldsymbol{\theta}) - \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{\theta}) \tag{C.35}$$

Let

$$\mathcal{L}^*(\boldsymbol{\theta}) = \max_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{C.36}$$

However, since $\boldsymbol{\lambda} \geq \mathbf{0}$ and $f_i(\boldsymbol{\theta}) \geq 0$, the maximum value of the Lagrangian occurs if the summation in (C.35) is zero (either $\lambda_i = 0$ or $f_i(\boldsymbol{\theta}) = 0$ or both) and

$$\mathcal{L}^*(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) \tag{C.37}$$

Therefore our original problem is equivalent with

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{C.38}$$

As we already know, the dual problem of the above is

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}} \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{C.39}$$

Convex Programming

A large class of practical problems obeys the following two conditions:

$$(1) \quad J(\boldsymbol{\theta}) \text{ is convex} \tag{C.40}$$

$$(2) \quad f_i(\boldsymbol{\theta}) \text{ are concave} \tag{C.41}$$

This class of problems turns out to have a very useful and mathematically tractable property.

Theorem *Let $\boldsymbol{\theta}_*$ be a minimizer of such a problem, which is also assumed to satisfy the regularity condition. Let $\boldsymbol{\lambda}_*$ be the corresponding vector of Lagrange multipliers. Then $(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*)$ is a saddle point of the Lagrangian function, and as we know this is equivalent to*

$$\mathcal{L}(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*) = \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \min_{\boldsymbol{\theta}} \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{C.42}$$

Proof. Since $f_i(\boldsymbol{\theta})$ are concave, $-f_i(\boldsymbol{\theta})$ are convex, so the Lagrangian function

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = J(\boldsymbol{\theta}) - \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{\theta})$$

for $\lambda_i \geq 0$, is also convex. Note, now, that for concave function constraints of the form $f_i(\boldsymbol{\theta}) \geq 0$, the feasible region is convex (see remarks above). The function $J(\boldsymbol{\theta})$ is also convex. Hence, as already stated in the remarks, every local minimum is also a global one; thus for any $\boldsymbol{\theta}$

$$\mathcal{L}(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*) \leq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}_*) \tag{C.43}$$

Furthermore, the complementary slackness conditions suggest that

$$\mathcal{L}(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*) = J(\boldsymbol{\theta}_*) \tag{C.44}$$

and for any $\boldsymbol{\lambda} \geq \mathbf{0}$

$$\mathcal{L}(\boldsymbol{\theta}_*, \boldsymbol{\lambda}) \equiv J(\boldsymbol{\theta}_*) - \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{\theta}_*) \leq J(\boldsymbol{\theta}_*) = \mathcal{L}(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*) \tag{C.45}$$

Combining (C.43) and (C.45) we obtain

$$\mathcal{L}(\boldsymbol{\theta}_*, \boldsymbol{\lambda}) \leq \mathcal{L}(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*) \leq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}_*) \tag{C.46}$$

In other words, *the solution $(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*)$ is a saddle point*. □

This is a very important theorem and it states that the constrained minimum of a convex programming problem can also be obtained as a maximization task applied on the Lagrangian. This leads us to the following very useful formulation of the optimization task.

Wolfe Dual Representation

A convex programming problem is equivalent to

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{C.47}$$

$$\text{subject to } \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbf{0} \tag{C.48}$$

The last equation guarantees that $\boldsymbol{\theta}$ is a minimum of the Lagrangian.

Example C.1

Consider the quadratic problem

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} \\ \text{subject to} \quad & A\boldsymbol{\theta} \geq \boldsymbol{b} \end{aligned}$$

This is a convex programming problem; hence the Wolfe dual representation is valid:

$$\begin{aligned} \text{maximize} \quad & \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} - \boldsymbol{\lambda}^T(A\boldsymbol{\theta} - \boldsymbol{b}) \\ \text{subject to} \quad & \boldsymbol{\theta} - A^T\boldsymbol{\lambda} = \mathbf{0} \end{aligned}$$

For this example, the equality constraint has an analytic solution (this is not, however, always possible). Solving with respect to $\boldsymbol{\theta}$, we can eliminate it from the maximizing function and the resulting dual problem involves only the Lagrange multipliers,

$$\max_{\boldsymbol{\lambda}} \left\{ -\frac{1}{2}\boldsymbol{\lambda}^T AA^T\boldsymbol{\lambda} + \boldsymbol{\lambda}^T\boldsymbol{b} \right\}$$

$$\text{subject to} \quad \boldsymbol{\lambda} \geq \mathbf{0}$$

This is also a quadratic problem but the set of constraints is now simpler.

REFERENCES

[Baza 79] Bazaraa M.S., Shetty C.M. *Nonlinear Programming: Theory and Algorithms*, John Wiley, 1979.

[Bert 95] Bertsekas, D.P., Belmont, M.A. *Nonlinear Programming*, Athenas Scientific, 1995.

[Flet 87] Fletcher, R. *Practical Methods of Optimization*, 2nd ed., John Wiley, 1987.

[Luen 84] Luenberger D.G. *Linear and Nonlinear Programming*, Addison Wesley, 1984.

[Nash 96] Nash S.G., Sofer A. *Linear and Nonlinear Programming*, McGraw-Hill, 1996.