Part VI

Preliminaries for Optimization Theory

Chapter 40

Extrema of Real-Valued Functions

40.1 Local Extrema, Constrained Local Extrema, and Lagrange Multipliers

Let $J\colon E \to \mathbb{R}$ be a real-valued function defined on a normed vector space E (or more generally, any topological space). Ideally we would like to find where the function J reaches a minimum or a maximum value, at least locally. In this chapter we will usually use the notations $dJ(u)$ or $J'(u)$ (or dJ_u or J'_u) for the derivative of J at u, instead of $\mathrm{D}J(u)$. Our presentation follows very closely that of Ciarlet [41] (Chapter 7), which we find to be one of the clearest.

Definition 40.1. If $J\colon E \to \mathbb{R}$ is a real-valued function defined on a normed vector space E, we say that J has a *local minimum* (or *relative minimum*) at the point $u \in E$ if there is some open subset $W \subseteq E$ containing u such that

$$J(u) \leq J(w) \quad \text{for all } w \in W.$$

Similarly, we say that J has a *local maximum* (or *relative maximum*) at the point $u \in E$ if there is some open subset $W \subseteq E$ containing u such that

$$J(u) \geq J(w) \quad \text{for all } w \in W.$$

In either case, we say that J has a *local extremum* (or *relative extremum*) at u. We say that J has a *strict local minimum* (resp. *strict local maximum*) at the point $u \in E$ if there is some open subset $W \subseteq E$ containing u such that

$$J(u) < J(w) \quad \text{for all } w \in W - \{u\}$$

(resp.

$$J(u) > J(w) \quad \text{for all } w \in W - \{u\}).$$

By abuse of language, we often say that the point u itself “is a local minimum” or a “local maximum,” even though, strictly speaking, this does not make sense.

We begin with a well-known necessary condition for a local extremum.

Proposition 40.1. *Let E be a normed vector space and let $J\colon \Omega \to \mathbb{R}$ be a function, with Ω some open subset of E. If the function J has a local extremum at some point $u \in \Omega$ and if J is differentiable at u, then*

$$dJ_u = J'(u) = 0.$$

Proof. Pick any $v \in E$. Since Ω is open, for t small enough we have $u + tv \in \Omega$, so there is an open interval $I \subseteq \mathbb{R}$ such that the function φ given by

$$\varphi(t) = J(u + tv)$$

for all $t \in I$ is well-defined. By applying the chain rule, we see that φ is differentiable at $t = 0$, and we get

$$\varphi'(0) = dJ_u(v).$$

Without loss of generality, assume that u is a local minimum. Then we have

$$\varphi'(0) = \lim_{t \mapsto 0_-} \frac{\varphi(t) - \varphi(0)}{t} \leq 0$$

and

$$\varphi'(0) = \lim_{t \mapsto 0_+} \frac{\varphi(t) - \varphi(0)}{t} \geq 0,$$

which shows that $\varphi'(0) = dJ_u(v) = 0$. As $v \in E$ is arbitrary, we conclude that $dJ_u = 0$. □

A point $u \in \Omega$ such that $J'(u) = 0$ is called a *critical point* of J.

If $E = \mathbb{R}^n$, then the condition $dJ_u = 0$ is equivalent to the system

$$\begin{aligned} \frac{\partial J}{\partial x_1}(u_1, \ldots, u_n) &= 0 \\ &\vdots \\ \frac{\partial J}{\partial x_n}(u_1, \ldots, u_n) &= 0. \end{aligned}$$

The condition of Proposition 40.1 is only a *necessary* condition for the existences of an extremum, but not a sufficient condition. Here are some counter-examples. If $f\colon \mathbb{R} \to \mathbb{R}$ is the function given by $f(x) = x^3$, since $f'(x) = 3x^2$, we have $f'(0) = 0$, but 0 is neither a minimum nor a maximum of f. If $g\colon \mathbb{R}^2 \to \mathbb{R}$ is the function given by $g(x, y) = x^2 - y^2$, then $g'_{(x,y)} = (2x \;\; -2y)$, so $g'_{(0,0)} = (0 \;\; 0)$, yet near $(0, 0)$ the function g takes negative and positive values.

In many practical situations, we need to look for local extrema of a function J *under additional constraints*. This situation can be formalized conveniently as follows: We have a function $J\colon \Omega \to \mathbb{R}$ defined on some open subset Ω of a normed vector space, but we also have some subset U of Ω, and we are looking for the local extrema of J *with respect to the set* U.

The elements $u \in U$ are often called *feasible solutions* of the optimization problem consisting in finding the local extrema of some objective function J with respect to some subset U of Ω defined by a set of constraints. Note that in most cases, U is *not* open. In fact, U is usually closed.

Definition 40.2. If $J\colon \Omega \to \mathbb{R}$ is a real-valued function defined on some open subset Ω of a normed vector space E and if U is some subset of Ω, we say that J has a *local minimum* (or *relative minimum*) at the point $u \in U$ *with respect to* U if there is some open subset $W \subseteq \Omega$ containing u such that

$$J(u) \leq J(w) \quad \text{for all } w \in U \cap W.$$

Similarly, we say that J has a *local maximum* (or *relative maximum*) at the point $u \in U$ *with respect to* U if there is some open subset $W \subseteq \Omega$ containing u such that

$$J(u) \geq J(w) \quad \text{for all } w \in U \cap W.$$

In either case, we say that J has a *local extremum* at u *with respect to* U.

It is very important to note that the hypothesis that Ω *is open* is crucial for the validity of Proposition 40.1. For example, if J is the identity function on $\mathbb{R}$ and $U = [0,1]$, a closed subset, then $J'(x) = 1$ for all $x \in [0,1]$, even though J has a minimum at $x = 0$ and a maximum at $x = 1$.

Therefore, in order to find necessary conditions for a function $J\colon \Omega \to \mathbb{R}$ to have a local extremum with respect to a subset U of Ω (where Ω is open), we need to somehow incorporate the definition of U into these conditions. This can be done in two cases:

(1) The set U is defined by a set of equations,

$$U = \{x \in \Omega \mid \varphi_i(x) = 0,\ 1 \leq i \leq m\},$$

where the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

(2) The set U is defined by a set of inequalities,

$$U = \{x \in \Omega \mid \varphi_i(x) \leq 0,\ 1 \leq i \leq m\},$$

where the functions $\varphi_i\colon \Omega \to \mathbb{R}$ are continuous (and usually differentiable).

In (1), the equations $\varphi_i(x) = 0$ are called *equality constraints*, and in (2), the inequalities $\varphi_i(x) \leq 0$ are called *inequality constraints*.

An inequality constraint of the form $\varphi_i(x) \geq 0$ is equivalent to the inequality constraint $-\varphi_x(x) \leq 0$. An equality constraint $\varphi_i(x) = 0$ is equivalent to the conjunction of the two inequality constraints $\varphi_i(x) \leq 0$ and $-\varphi_i(x) \leq 0$, so the case of inequality constraints subsumes the case of equality constraints. However, the case of equality constraints is easier to deal with, and in this chapter we will restrict our attention to this case.

If the functions φ_i are convex and Ω is convex, then U is convex. This is a very important case that we will discuss later. In particular, if the functions φ_i are affine, then the equality constraints can be written as $Ax = b$, and the inequality constraints as $Ax \leq b$, for some $m \times n$ matrix A and some vector $b \in \mathbb{R}^m$. We will also discuss the case of affine constraints later.

In the case of equality constraints, a necessary condition for a local extremum with respect to U can be given in terms of *Lagrange multipliers*. In the case of inequality constraints, there is also a necessary condition for a local extremum with respect to U in terms of generalized Lagrange multipliers and the *Karush–Kuhn–Tucker* conditions. This will be discussed in Chapter 50.

We begin by considering the case where $\Omega \subseteq E_1 \times E_2$ is an open subset of a product of normed vector spaces and where U is the zero locus of some continuous function $\varphi\colon \Omega \to E_2$, which means that

$$U = \{(u_1, u_2) \in \Omega \mid \varphi(u_1, u_2) = 0\}.$$

For the sake of brevity, we say that J has a *constrained local extremum* at u instead of saying that J has a *local extremum* at the point $u \in U$ *with respect to* U. Fortunately, there is a necessary condition for constrained local extrema in terms of *Lagrange multipliers*.

Theorem 40.2. *(Necessary condition for a constrained extremum) Let $\Omega \subseteq E_1 \times E_2$ be an open subset of a product of normed vector spaces, with E_1 a Banach space (E_1 is complete), let $\varphi\colon \Omega \to E_2$ be a C^1-function (which means that $d\varphi(\omega)$ exists and is continuous for all $\omega \in \Omega$), and let*

$$U = \{(u_1, u_2) \in \Omega \mid \varphi(u_1, u_2) = 0\}.$$

Moreover, let $u = (u_1, u_2) \in U$ be a point such that

$$\frac{\partial \varphi}{\partial x_2}(u_1, u_2) \in \mathcal{L}(E_2; E_2) \quad \text{and} \quad \left(\frac{\partial \varphi}{\partial x_2}(u_1, u_2)\right)^{-1} \in \mathcal{L}(E_2; E_2),$$

and let $J\colon \Omega \to \mathbb{R}$ be a function which is differentiable at u. If J has a constrained local extremum at u, then there is a continuous linear form $\Lambda(u) \in \mathcal{L}(E_2; \mathbb{R})$ such that

$$dJ(u) + \Lambda(u) \circ d\varphi(u) = 0.$$

Proof. The plan of attack is to use the implicit function theorem; Theorem 39.14. Observe that the assumptions of Theorem 39.14 are indeed met. Therefore, there exist some open subsets $U_1 \subseteq E_1$, $U_2 \subseteq E_2$, and a continuous function $g\colon U_1 \to U_2$ with $(u_1, u_2) \in U_1 \times U_2 \subseteq \Omega$ and such that

$$\varphi(v_1, g(v_1)) = 0$$

for all $v_1 \in U_1$. Moreover, g is differentiable at $u_1 \in U_1$ and

$$dg(u_1) = -\left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_1}(u).$$

It follows that the restriction of J to $(U_1 \times U_2) \cap U$ yields a function G of a single variable, with

$$G(v_1) = J(v_1, g(v_1))$$

for all $v_1 \in U_1$. Now, the function G is differentiable at u_1 and it has a local extremum at u_1 on U_1, so Proposition 40.1 implies that

$$dG(u_1) = 0.$$

By the chain rule,

$$\begin{aligned} dG(u_1) &= \frac{\partial J}{\partial x_1}(u) + \frac{\partial J}{\partial x_2}(u) \circ dg(u_1) \\ &= \frac{\partial J}{\partial x_1}(u) - \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_1}(u). \end{aligned}$$

From $dG(u_1) = 0$, we deduce

$$\frac{\partial J}{\partial x_1}(u) = \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_1}(u),$$

and since we also have

$$\frac{\partial J}{\partial x_2}(u) = \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \frac{\partial \varphi}{\partial x_2}(u),$$

if we let

$$\Lambda(u) = -\frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1},$$

then we get

$$\begin{aligned} dJ(u) &= \frac{\partial J}{\partial x_1}(u) + \frac{\partial J}{\partial x_2}(u) \\ &= \frac{\partial J}{\partial x_2}(u) \circ \left(\frac{\partial \varphi}{\partial x_2}(u)\right)^{-1} \circ \left(\frac{\partial \varphi}{\partial x_1}(u) + \frac{\partial \varphi}{\partial x_2}(u)\right) \\ &= -\Lambda(u) \circ d\varphi(u), \end{aligned}$$

which yields $dJ(u) + \Lambda(u) \circ d\varphi(u) = 0$, as claimed. □

In most applications, we have $E_1 = \mathbb{R}^{n-m}$ and $E_2 = \mathbb{R}^m$ for some integers m, n such that $1 \leq m < n$, Ω is an open subset of $\mathbb{R}^n$, $J\colon \Omega \to \mathbb{R}$, and we have m functions $\varphi_i\colon \Omega \to \mathbb{R}$ defining the subset

$$U = \{v \in \Omega \mid \varphi_i(v) = 0,\ 1 \leq i \leq m\}.$$

Theorem 40.2 yields the following necessary condition:

Theorem 40.3. *(Necessary condition for a constrained extremum in terms of Lagrange multipliers) Let Ω be an open subset of $\mathbb{R}^n$, consider m C^1-functions $\varphi_i\colon \Omega \to \mathbb{R}$ (with $1 \leq m < n$), let*

$$U = \{v \in \Omega \mid \varphi_i(v) = 0,\ 1 \leq i \leq m\},$$

and let $u \in U$ be a point such that the derivatives $d\varphi_i(u) \in \mathcal{L}(\mathbb{R}^n; \mathbb{R})$ are linearly independent; equivalently, assume that the $m \times n$ matrix $\big((\partial\varphi_i/\partial x_j)(u)\big)$ has rank m. If $J\colon \Omega \to \mathbb{R}$ is a function which is differentiable at $u \in U$ and if J has a local constrained extremum at u, then there exist m numbers $\lambda_i(u) \in \mathbb{R}$, uniquely defined, such that

$$dJ(u) + \lambda_1(u)d\varphi_1(u) + \cdots + \lambda_m(u)d\varphi_m(u) = 0;$$

equivalently,

$$\nabla J(u) + \lambda_1(u)\nabla\varphi_1(u) + \cdots + \lambda_1(u)\nabla\varphi_m(u) = 0.$$

Proof. The linear independence of the m linear forms $d\varphi_i(u)$ is equivalent to the fact that the $m \times n$ matrix $A = \big((\partial\varphi_i/\partial x_j)(u)\big)$ has rank m. By reordering the columns, we may assume that the first m columns are linearly independent. If we let $\varphi\colon \Omega \to \mathbb{R}^m$ be the function defined by

$$\varphi(v) = (\varphi_1(v), \ldots, \varphi_m(v))$$

for all $v \in \Omega$, then we see that $\partial\varphi/\partial x_2(u)$ is invertible and both $\partial\varphi/\partial x_2(u)$ and its inverse are continuous, so that Theorem 40.2 applies, and there is some (continuous) linear form $\Lambda(u) \in \mathcal{L}(\mathbb{R}^m; \mathbb{R})$ such that

$$dJ(u) + \Lambda(u) \circ d\varphi(u) = 0.$$

However, $\Lambda(u)$ is defined by some m-tuple $(\lambda_1(u), \ldots, \lambda_m(u)) \in \mathbb{R}^m$, and in view of the definition of φ, the above equation is equivalent to

$$dJ(u) + \lambda_1(u)d\varphi_1(u) + \cdots + \lambda_m(u)d\varphi_m(u) = 0.$$

The uniqueness of the $\lambda_i(u)$ is a consequence of the linear independence of the $d\varphi_i(u)$. $\square$

The numbers $\lambda_i(u)$ involved in Theorem 40.3 are called the *Lagrange multipliers* associated with the constrained extremum u (again, with some minor abuse of language). The linear independence of the linear forms $d\varphi_i(u)$ is equivalent to the fact that the Jacobian matrix $\big((\partial\varphi_i/\partial x_j)(u)\big)$ of $\varphi = (\varphi_1, \ldots, \varphi_m)$ at u has rank m. If $m = 1$, the linear independence of the $d\varphi_i(u)$ reduces to the condition $\nabla\varphi_1(u) \neq 0$.

A fruitful way to reformulate the use of Lagrange multipliers is to introduce the notion of the *Lagrangian* associated with our constrained extremum problem. This is the function $L\colon \Omega \times \mathbb{R}^m \to \mathbb{R}$ given by

$$L(v, \lambda) = J(v) + \lambda_1 \varphi_1(v) + \cdots + \lambda_m \varphi_m(v),$$

with $\lambda = (\lambda_1, \ldots, \lambda_m)$. Then, observe that there exists some $\mu = (\mu_1, \ldots, \mu_m)$ and some $u \in U$ such that

$$dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u) = 0$$

if and only if

$$dL(u, \mu) = 0,$$

or equivalently

$$\nabla L(u, \mu) = 0;$$

that is, iff (u, λ) is a *critical point* of the Lagrangian L.

Indeed $dL(u, \mu) = 0$ if equivalent to

$$\begin{aligned} \frac{\partial L}{\partial v}(u, \mu) &= 0 \\ \frac{\partial L}{\partial \lambda_1}(u, \mu) &= 0 \\ &\vdots \\ \frac{\partial L}{\partial \lambda_m}(u, \mu) &= 0, \end{aligned}$$

and since

$$\frac{\partial L}{\partial v}(u, \mu) = dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u)$$

and

$$\frac{\partial L}{\partial \lambda_i}(u, \mu) = \varphi_i(u),$$

we get

$$dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u) = 0$$

and

$$\varphi_1(u) = \cdots = \varphi_m(u) = 0,$$

that is, $u \in U$.

If we write out explicitly the condition

$$dJ(u) + \mu_1 d\varphi_1(u) + \cdots + \mu_m d\varphi_m(u) = 0,$$

we get the $n \times m$ system

$$\begin{aligned}
\frac{\partial J}{\partial x_1}(u) + \lambda_1 \frac{\partial \varphi_1}{\partial x_1}(u) + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_1}(u) &= 0 \\
&\vdots \\
\frac{\partial J}{\partial x_n}(u) + \lambda_1 \frac{\partial \varphi_1}{\partial x_n}(u) + \cdots + \lambda_m \frac{\partial \varphi_m}{\partial x_n}(u) &= 0,
\end{aligned}$$

and it is important to note that the matrix of this system is the *transpose* of the Jacobian matrix of φ at u. If we write $\mathrm{Jac}(J)(u) = \big((\partial \varphi_i/\partial x_j)(u)\big)$ for the Jacobian matrix of J (at u), then the above system is written in matrix form as

$$\nabla J(u) + (\mathrm{Jac}(J)(u))^{\top} \lambda = 0,$$

where λ is viewed as a column vector, and the Lagrangian is equal to

$$L(u, \lambda) = J(u) + (\varphi_1(u), \ldots, \varphi_m(u))\lambda.$$

Remark: If the Jacobian matrix $\mathrm{Jac}(J)(v) = \big((\partial \varphi_i/\partial x_j)(v)\big)$ has rank m for all $v \in U$ (which is equivalent to the linear independence of the linear forms $d\varphi_i(v)$), then we say that $0 \in \mathbb{R}^m$ is a *regular value* of φ. In this case, it is known that

$$U = \{v \in \Omega \mid \varphi(v) = 0\}$$

is a *smooth submanifold of dimension* $n - m$ *of* $\mathbb{R}^n$. Furthermore, the set

$$T_v U = \{w \in \mathbb{R}^n \mid d\varphi_i(v)(w) = 0,\ 1 \leq i \leq m\} = \bigcap_{i=1}^{m} \mathrm{Ker}\, d\varphi_i(v)$$

is the *tangent space* to U at v (a vector space of dimension $n - m$). Then, the condition

$$dJ(v) + \mu_1 d\varphi_1(v) + \cdots + \mu_m d\varphi_m(v) = 0$$

implies that $dJ(v)$ vanishes on the tangent space $T_v U$. Conversely, if $dJ(v)(w) = 0$ for all $w \in T_v U$, this means that $dJ(v)$ is orthogonal (in the sense of Definition 11.3 (Vol. I)) to $T_v U$. Since (by Theorem 11.4 (b) (Vol. I)) the orthogonal of $T_v U$ is the space of linear forms spanned by $d\varphi_1(v), \ldots, d\varphi_m(v)$, it follows that $dJ(v)$ must be a linear combination of the $d\varphi_i(v)$. Therefore, when 0 is a regular value of φ, Theorem 40.3 asserts that if $u \in U$ is a local extremum of J, then $dJ(u)$ must vanish on the tangent space $T_u U$. We can say even more. The subset $Z(J)$ of Ω given by

$$Z(J) = \{v \in \Omega \mid J(v) = J(u)\}$$

(the *level set of level* $J(u)$) is a hypersurface in Ω, and if $dJ(u) \neq 0$, the zero locus of $dJ(u)$ is the tangent space $T_uZ(J)$ to $Z(J)$ at u (a vector space of dimension $n-1$), where

$$T_uZ(J) = \{w \in \mathbb{R}^n \mid dJ(u)(w) = 0\}.$$

Consequently, Theorem 40.3 asserts that

$$T_uU \subseteq T_uZ(J);$$

this is a geometric condition.

The beauty of the Lagrangian is that the constraints $\{\varphi_i(v) = 0\}$ have been incorporated into the function $L(v, \lambda)$, and that the necessary condition for the existence of a constrained local extremum of J is reduced to the necessary condition for the existence of a local extremum of the *unconstrained* L.

However, one should be careful to check that the assumptions of Theorem 40.3 are satisfied (in particular, the linear independence of the linear forms $d\varphi_i$). For example, let $J\colon \mathbb{R}^3 \to \mathbb{R}$ be given by

$$J(x, y, z) = x + y + z^2$$

and $g\colon \mathbb{R}^3 \to \mathbb{R}$ by

$$g(x, y, z) = x^2 + y^2.$$

Since $g(x, y, z) = 0$ iff $x = y = 0$, we have $U = \{(0, 0, z) \mid z \in \mathbb{R}\}$ and the restriction of J to U is given by

$$J(0, 0, z) = z^2,$$

which has a minimum for $z = 0$. However, a “blind” use of Lagrange multipliers would require that there is some λ so that

$$\frac{\partial J}{\partial x}(0, 0, z) = \lambda \frac{\partial g}{\partial x}(0, 0, z), \quad \frac{\partial J}{\partial y}(0, 0, z) = \lambda \frac{\partial g}{\partial y}(0, 0, z), \quad \frac{\partial J}{\partial z}(0, 0, z) = \lambda \frac{\partial g}{\partial z}(0, 0, z),$$

and since

$$\frac{\partial g}{\partial x}(x, y, z) = 2x, \quad \frac{\partial g}{\partial y}(x, y, z) = 2y, \quad \frac{\partial g}{\partial z}(0, 0, z) = 0,$$

the partial derivatives above all vanish for $x = y = 0$, so at a local extremum we should also have

$$\frac{\partial J}{\partial x}(0, 0, z) = 0, \quad \frac{\partial J}{\partial y}(0, 0, z) = 0, \quad \frac{\partial J}{\partial z}(0, 0, z) = 0,$$

but this is absurd since

$$\frac{\partial J}{\partial x}(x, y, z) = 1, \quad \frac{\partial J}{\partial y}(x, y, z) = 1, \quad \frac{\partial J}{\partial z}(x, y, z) = 2z.$$

The reader should enjoy finding the reason for the flaw in the argument.

One should also keep in mind that Theorem 40.3 gives only a necessary condition. The (u, λ) may *not* correspond to local extrema! Thus, it is always necessary to analyze the local behavior of J near a critical point u. This is generally difficult, but in the case where J is affine or quadratic and the constraints are affine or quadratic, this is possible (although not always easy).

Let us apply the above method to the following example in which $E_1 = \mathbb{R}$, $E_2 = \mathbb{R}$, $\Omega = \mathbb{R}^2$, and

$$\begin{aligned} J(x_1, x_2) &= -x_2 \\ \varphi(x_1, x_2) &= x_1^2 + x_2^2 - 1. \end{aligned}$$

Observe that

$$U = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1^2 + x_2^2 = 1\}$$

is the unit circle, and since

$$\nabla\varphi(x_1, x_2) = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix},$$

it is clear that $\nabla\varphi(x_1, x_2) \neq 0$ for every point $= (x_1, x_2)$ on the unit circle. If we form the Lagrangian

$$L(x_1, x_2, \lambda) = -x_2 + \lambda(x_1^2 + x_2^2 - 1),$$

Theorem 40.3 says that a necessary condition for J to have a constrained local extremum is that $\nabla L(x_1, x_2, \lambda) = 0$, so the following equations must hold:

$$\begin{aligned} 2\lambda x_1 &= 0 \\ -1 + 2\lambda x_2 &= 0 \\ x_1^2 + x_2^2 &= 1. \end{aligned}$$

The second equation implies that $\lambda \neq 0$, and then the first yields $x_1 = 0$, so the third yields $x_2 = \pm 1$, and we get two solutions:

$$\begin{aligned} \lambda &= \frac{1}{2}, & (x_1, x_2) &= (0, 1) \\ \lambda &= -\frac{1}{2}, & (x_1', x_2') &= (0, -1). \end{aligned}$$

We can check immediately that the first solution is a minimum and the second is a maximum. The reader should look for a geometric interpretation of this problem.

Let us now consider the case in which J is a quadratic function of the form

$$J(v) = \frac{1}{2} v^\top A v - v^\top b,$$

where A is an $n \times n$ symmetric matrix, $b \in \mathbb{R}^n$, and the constraints are given by a linear system of the form

$$Cv = d,$$

where C is an $m \times n$ matrix with $m < n$ and $d \in \mathbb{R}^m$. We also assume that C has rank m. In this case, the function φ is given by

$$\varphi(v) = (Cv - d)^\top,$$

because we view $\varphi(v)$ as a row vector (and v as a column vector), and since

$$d\varphi(v)(w) = C^\top w,$$

the condition that the Jacobian matrix of φ at u have rank m is satisfied. The Lagrangian of this problem is

$$L(v, \lambda) = \frac{1}{2} v^\top A v - v^\top b + (Cv - d)^\top \lambda = \frac{1}{2} v^\top A v - v^\top b + \lambda^\top (Cv - d),$$

where λ is viewed as a column vector. Now, because A is a symmetric matrix, it is easy to show that

$$\nabla L(v, \lambda) = \begin{pmatrix} Av - b + C^\top \lambda \\ Cv - d \end{pmatrix}.$$

Therefore, the necessary condition for contrained local extrema is

$$\begin{aligned} Av + C^\top \lambda &= b \\ Cv &= d, \end{aligned}$$

which can be expressed in matrix form as

$$\begin{pmatrix} A & C^\top \\ C & 0 \end{pmatrix} \begin{pmatrix} v \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix},$$

where the matrix of the system is a symmetric matrix. We should not be surprised to find the system of Section 42, except for some renaming of the matrices and vectors involved. As we know from Section 42.2, the function J has a minimum iff A is positive definite, so in general, if A is only a symmetric matrix, the critical points of the Lagrangian do *not* correspond to extrema of J.

We now investigate conditions for the existence of extrema involving the second derivative of J.

40.2 Using Second Derivatives to Find Extrema

For the sake of brevity, we consider only the case of local minima; analogous results are obtained for local maxima (replace J by $-J$, since $\max_u J(u) = -\min_u -J(u)$). We begin with a necessary condition for an unconstrained local minimum.

Proposition 40.4. *Let E be a normed vector space and let $J\colon \Omega \to \mathbb{R}$ be a function, with Ω some open subset of E. If the function J is differentiable in Ω, if J has a second derivative $\mathrm{D}^2 J(u)$ at some point $u \in \Omega$, and if J has a local minimum at u, then*

$$\mathrm{D}^2 J(u)(w, w) \geq 0 \quad \textit{for all } w \in E.$$

Proof. Pick any nonzero vector $w \in E$. Since Ω is open, for t small enough, $u + tw \in \Omega$ and $J(u + tw) \geq J(u)$, so there is some open interval $I \subseteq \mathbb{R}$ such that

$$u + tw \in \Omega \quad \text{and} \quad J(u + tw) \geq J(u)$$

for all $t \in I$. Using the Taylor–Young formula and the fact that we must have $dJ(u) = 0$ since J has a local minimum at u, we get

$$0 \leq J(u + tw) - J(u) = \frac{t^2}{2}\mathrm{D}^2 J(u)(w, w) + t^2 \left\|w\right\|^2 \epsilon(tw),$$

with $\lim_{t \mapsto 0} \epsilon(tw) = 0$, which implies that

$$\mathrm{D}^2 J(u)(w, w) \geq 0.$$

Since the argument holds for all $w \in E$ (trivially if $w = 0$), the proposition is proved. □

One should be cautioned that there is no converse to the previous proposition. For example, the function $f\colon x \mapsto x^3$ has no local minimum at 0, yet $df(0) = 0$ and $\mathrm{D}^2 f(0)(u, v) = 0$. Similarly, the reader should check that the function $f\colon \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x, y) = x^2 - 3y^3$$

has no local minimum at $(0, 0)$; yet $df(0, 0) = 0$ and $\mathrm{D}^2 f(0, 0)(u, v) = 2u^2 \geq 0$.

When $E = \mathbb{R}^n$, Proposition 40.4 says that a necessary condition for having a local minimum is that the Hessian $\nabla^2 J(u)$ be positive semidefinite (it is always symmetric).

We now give sufficient conditions for the existence of a local minimum.

Theorem 40.5. *Let E be a normed vector space, let $J\colon \Omega \to \mathbb{R}$ be a function with Ω some open subset of E, and assume that J is differentiable in Ω and that $dJ(u) = 0$ at some point $u \in \Omega$. The following properties hold:*

(1) If $\mathrm{D}^2 J(u)$ exists and if there is some number $\alpha \in \mathbb{R}$ such that $\alpha > 0$ and

$$\mathrm{D}^2 J(u)(w, w) \geq \alpha \left\|w\right\|^2 \quad \textit{for all } w \in E,$$

then J has a strict local minimum at u.

(2) If $\mathrm{D}^2 J(v)$ exists for all $v \in \Omega$ and if there is a ball $B \subseteq \Omega$ centered at u such that

$$\mathrm{D}^2 J(v)(w, w) \geq 0 \quad \textit{for all } v \in B \textit{ and all } w \in E,$$

then J has a local minimum at u.

Proof. (1) Using the formula of Taylor–Young, for every vector w small enough, we can write

$$\begin{aligned} J(u+w) - J(u) &= \frac{1}{2}\mathrm{D}^2 J(u)(w,w) + \|w\|^2 \epsilon(w) \\ &\geq \left(\frac{1}{2}\alpha + \epsilon(w)\right) \|w\|^2 \end{aligned}$$

with $\lim_{w \mapsto 0} \epsilon(w) = 0$. Consequently if we pick $r > 0$ small enough that $|\epsilon(w)| < \alpha$ for all w with $\|w\| < r$, then $J(u+w) > J(u)$ for all $u + w \in B$, where B is the open ball of center u and radius r. This proves that J has a local strict minimum at u.

(2) The formula of Taylor–Maclaurin shows that for all $u + w \in B$, we have

$$J(u+w) = J(u) + \frac{1}{2}\mathrm{D}^2 J(v)(w,w) \geq J(u),$$

for some $v \in (u, w+w)$. □

There are no converses of the two assertions of Theorem 40.5. However, there is a condition on $\mathrm{D}^2 J(u)$ that implies the condition of Part (1). Since this condition is easier to state when $E = \mathbb{R}^n$, we begin with this case.

Recall that a $n \times n$ symmetric matrix A is *positive definite* if $x^\top A x > 0$ for all $x \in \mathbb{R}^n - \{0\}$. In particular, A must be invertible.

Proposition 40.6. *For any symmetric matrix A, if A is positive definite, then there is some $\alpha > 0$ such that*

$$x^\top A x \geq \alpha \|x\|^2 \quad \textit{for all } x \in \mathbb{R}^n.$$

Proof. Pick any norm in $\mathbb{R}^n$ (recall that all norms on $\mathbb{R}^n$ are equivalent). Since the unit sphere $S^{n-1} = \{x \in \mathbb{R}^n \mid \|x\| = 1\}$ is compact and since the function $f(x) = x^\top A x$ is never zero on S^{n-1}, the function f has a minimum $\alpha > 0$ on S^{n-1}. Using the usual trick that $x = \|x\| (x/\|x\|)$ for every nonzero vector $x \in \mathbb{R}^n$ and the fact that the inequality of the proposition is trivial for $x = 0$, from

$$x^\top A x \geq \alpha \quad \text{for all } x \text{ with } \|x\| = 1,$$

we get

$$x^\top A x \geq \alpha \|x\|^2 \quad \text{for all } x \in \mathbb{R}^n,$$

as claimed. □

We can combine Theorem 40.5 and Proposition 40.6 to obtain a useful sufficient condition for the existence of a strict local minimum. First let us introduce some terminology.

Definition 40.3. Given a function $J\colon \Omega \to \mathbb{R}$ as before, say that a point $u \in \Omega$ is a *nondegenerate critical point* if $dJ(u) = 0$ and if the Hessian matrix $\nabla^2 J(u)$ is invertible.

Proposition 40.7. *Let $J\colon \Omega \to \mathbb{R}$ be a function defined on some open subset $\Omega \subseteq \mathbb{R}^n$. If J is differentiable in Ω and if some point $u \in \Omega$ is a nondegenerate critical point such that $\nabla^2 J(u)$ is positive definite, then J has a strict local minimum at u.*

Remark: It is possible to generalize Proposition 40.7 to infinite-dimensional spaces by finding a suitable generalization of the notion of a nondegenerate critical point. Firstly, we assume that E is a Banach space (a complete normed vector space). Then, we define the dual E' of E as the set of continuous linear forms on E, so that $E' = \mathcal{L}(E;\mathbb{R})$. Following Lang, we use the notation E' for the space of continuous linear forms to avoid confusion with the space $E^* = \mathrm{Hom}(E,\mathbb{R})$ of all linear maps from E to $\mathbb{R}$. A continuous bilinear map $\varphi\colon E \times E \to \mathbb{R}$ in $\mathcal{L}_2(E,E;\mathbb{R})$ yields a map Φ from E to E' given by

$$\Phi(u) = \varphi_u,$$

where $\varphi_u \in E'$ is the linear form defined by

$$\varphi_u(v) = \varphi(u,v).$$

It is easy to check that φ_u is continuous and that the map Φ is continuous. Then, we say that φ is *nondegenerate* iff $\Phi\colon E \to E'$ is an isomorphism of Banach spaces, which means that Φ is invertible and that both Φ and Φ^{-1} are continuous linear maps. Given a function $J\colon \Omega \to \mathbb{R}$ differentiable on Ω as before (where Ω is an open subset of E), if $\mathrm{D}^2 J(u)$ exists for some $u \in \Omega$, we say that u is a *nondegenerate critical point* if $dJ(u) = 0$ and if $\mathrm{D}^2 J(u)$ is nondegenerate. Of course, $\mathrm{D}^2 J(u)$ is positive definite if $\mathrm{D}^2 J(u)(w,w) > 0$ for all $w \in E - \{0\}$.

Using the above definition, Proposition 40.6 can be generalized to a nondegenerate positive definite bilinear form (on a Banach space) and Theorem 40.7 can also be generalized to the situation where $J\colon \Omega \to \mathbb{R}$ is defined on an open subset of a Banach space. For details and proofs, see Cartan [34] (Part I Chapter 8) and Avez [9] (Chapter 8 and Chapter 10).

In the next section we make use of convexity; both on the domain Ω and on the function J itself.

40.3 Using Convexity to Find Extrema

We begin by reviewing the definition of a convex set and of a convex function.

Definition 40.4. Given any real vector space E, we say that a subset C of E is *convex* if either $C = \emptyset$ or if for every pair of points $u, v \in C$, the line segment connecting u and v is contained in C, i.e.,

$$(1-\lambda)u + \lambda v \in C \quad \text{for all } \lambda \in \mathbb{R} \text{ such that } 0 \leq \lambda \leq 1.$$

Given any two points $u\, v \in E$, the *line segment* $[u,v]$ is the set

$$[u,v] = \{(1-\lambda)u + \lambda v \in E \mid \lambda \in \mathbb{R},\ 0 \leq \lambda \leq 1\}.$$

Clearly, a nonempty set C is convex iff $[u, v] \subseteq C$ whenever $u, v \in C$. See Figure 40.1 for an example of a convex set.

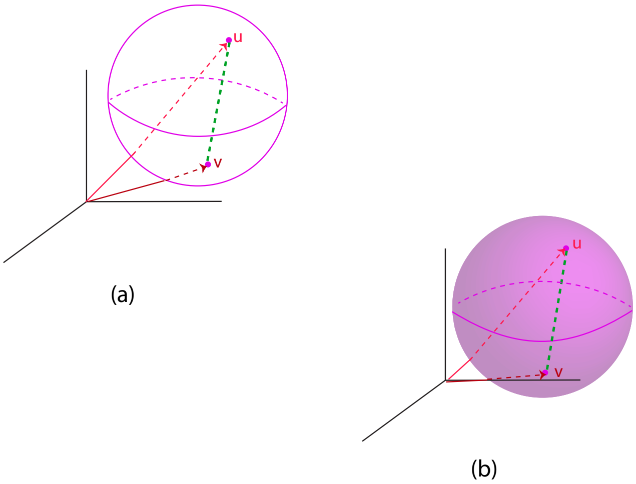

Figure 40.1: Figure (a) shows that a sphere is not convex in $\mathbb{R}^3$ since the dashed green line does not lie on its surface. Figure (b) shows that a solid ball is convex in $\mathbb{R}^3$.

Definition 40.5. If C is a nonempty convex subset of E, a function $f \colon C \to \mathbb{R}$ is *convex* (on C) if for every pair of points $u, v \in C$,

$$f((1-\lambda)u + \lambda v) \leq (1-\lambda)f(u) + \lambda f(v) \quad \text{for all } \lambda \in \mathbb{R} \text{ such that } 0 \leq \lambda \leq 1;$$

the function f is *strictly convex* (on C) if for every pair of distinct points $u, v \in C$ $(u \neq v)$,

$$f((1-\lambda)u + \lambda v) < (1-\lambda)f(u) + \lambda f(v) \quad \text{for all } \lambda \in \mathbb{R} \text{ such that } 0 < \lambda < 1;$$

see Figure 40.2. The *epigraph*[1] **epi**(f) of a function $f \colon A \to \mathbb{R}$ defined on some subset A of $\mathbb{R}^n$ is the subset of $\mathbb{R}^{n+1}$ defined as

$$\mathbf{epi}(f) = \{(x, y) \in \mathbb{R}^{n+1} \mid f(x) \leq y,\ x \in A\}.$$

A function $f \colon C \to \mathbb{R}$ defined on a convex subset C is *concave* (resp. *strictly concave*) if $(-f)$ is convex (resp. strictly convex).

It is obvious that a function f if convex iff its epigraph **epi**(f) is a convex subset of $\mathbb{R}^{n+1}$.

[1] "Epi" means above.

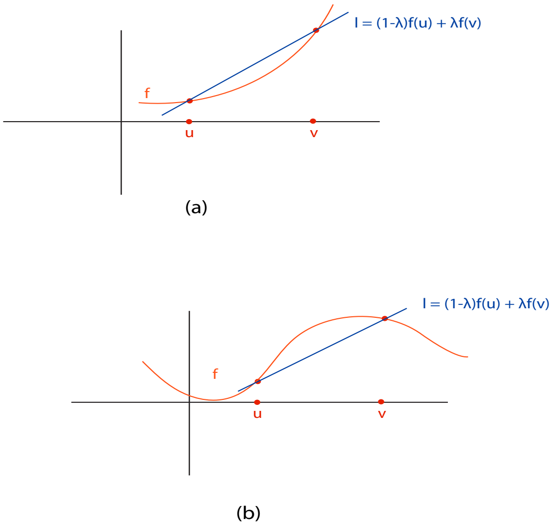

Figure 40.2: Figures (a) and (b) are the graphs of real valued functions. Figure (a) is the graph of convex function since the blue line lies above the graph of f. Figure (b) shows the graph of a function which is not convex.

Subspaces $V \subseteq E$ of a vector space E are convex; *affine subspaces*, that is, sets of the form $u + V$, where V is a subspace of E and $u \in E$, are convex. Balls (open or closed) are convex. Given any linear form $\varphi\colon E \to \mathbb{R}$, for any scalar $c \in \mathbb{R}$, the *closed half–spaces*

$$H^+_{\varphi,c} = \{u \in E \mid \varphi(u) \geq c\}, \qquad H^-_{\varphi,c} = \{u \in E \mid \varphi(u) \leq c\},$$

are convex. Any intersection of half–spaces is convex. More generally, any intersection of convex sets is convex.

Linear forms are convex functions (but not strictly convex). Any norm $\| \, \| \colon E \to \mathbb{R}_+$ is a convex function. The max function,

$$\max(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$$

is convex on $\mathbb{R}^n$. The exponential $x \mapsto e^{cx}$ is strictly convex for any $c \neq 0$ ($c \in \mathbb{R}$). The logarithm function is concave on $\mathbb{R}_+ - \{0\}$, and the *log-determinant function* $\log\det$ is concave on the set of symmetric positive definite matrices. This function plays an important role in convex optimization. An excellent exposition of convexity and its applications to optimization can be found in Boyd [29].

Here is a necessary condition for a function to have a local minimum with respect to a convex subset U.

Theorem 40.8. *(Necessary condition for a local minimum on a convex subset) Let $J\colon \Omega \to \mathbb{R}$ be a function defined on some open subset Ω of a normed vector space E and let $U \subseteq \Omega$ be a nonempty convex subset. Given any $u \in U$, if $dJ(u)$ exists and if J has a local minimum in u with respect to U, then*

$$dJ(u)(v-u) \geq 0 \quad \textit{for all } v \in U.$$

Proof. Let $v = u + w$ be an arbitrary point in U. Since U is convex, we have $u + tw \in U$ for all t such that $0 \leq t \leq 1$. Since $dJ(u)$ exists, we can write

$$J(u+tw) - J(u) = dJ(u)(tw) + \|tw\| \, \epsilon(tw)$$

with $\lim_{t \mapsto 0} \epsilon(tw) = 0$. However, because $0 \leq t \leq 1$,

$$J(u+tw) - J(u) = t(dJ(u)(w) + \|w\| \, \epsilon(tw))$$

and since u is a local minimum with respect to U, we have $J(u+tw) - J(u) \geq 0$, so we get

$$t(dJ(u)(w) + \|w\| \, \epsilon(tw)) \geq 0.$$

The above implies that $dJ(u)(w) \geq 0$, because otherwise we could pick $t > 0$ small enough so that

$$dJ(u)(w) + \|w\| \, \epsilon(tw) < 0,$$

a contradiction. Since the argument holds for all $v = u + w \in U$, the theorem is proved. □

Observe that the convexity of U is a substitute for the use of Lagrange multipliers, but we now have to deal with an *inequality* instead of an equality.

Consider the special case where U is a subspace of E. In this case since $u \in U$ we have $2u \in U$, and for any $u + w \in U$, we must have $2u - (u+w) = u - w \in U$. The previous theorem implies that $dJ(u)(w) \geq 0$ and $dJ(u)(-w) \geq 0$, that is, $dJ(u)(w) \leq 0$, so $dJ(u) = 0$. Since the argument holds for $w \in U$ (because U is a subspace, if $u, w \in U$, then $u + w \in U$), we conclude that

$$dJ(u)(w) = 0 \quad \text{for all } w \in U.$$

We will now characterize convex functions when they have a first derivative or a second derivative.

Proposition 40.9. *(Convexity and first derivative) Let $f\colon \Omega \to \mathbb{R}$ be a function differentiable on some open subset Ω of a normed vector space E and let $U \subseteq \Omega$ be a nonempty convex subset.*

(1) The function f is convex on U iff

$$f(v) \geq f(u) + df(u)(v-u) \quad \textit{for all } u, v \in U.$$

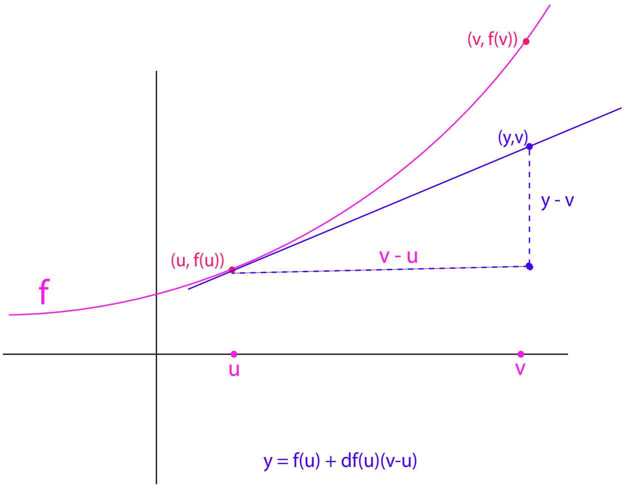

Figure 40.3: An illustration of a convex valued function f. Since f is convex it always lies above its tangent line.

(2) The function f is strictly convex on U iff

$$f(v) > f(u) + df(u)(v-u) \quad \textit{for all } u, v \in U \textit{ with } u \neq v.$$

See Figure 40.3.

Proof. Let $u, v \in U$ be any two distinct points and pick $\lambda \in \mathbb{R}$ with $0 < \lambda < 1$. If the function f is convex, then

$$f((1-\lambda)u + \lambda v) \leq (1-\lambda)f(u) + \lambda f(v),$$

which yields

$$\frac{f((1-\lambda)u + \lambda v) - f(u)}{\lambda} \leq f(v) - f(u).$$

It follows that

$$df(u)(v-u) = \lim_{\lambda \mapsto 0} \frac{f((1-\lambda)u + \lambda v) - f(u)}{\lambda} \leq f(v) - f(u).$$

If f is strictly convex, the above reasoning does not work, because a strict inequality is not necessarily preserved by "passing to the limit." We have recourse to the following trick: For any ω such that $0 < \omega < 1$, observe that

$$(1-\lambda)u + \lambda v = u + \lambda(v-u) = \frac{\omega - \lambda}{\omega} u + \frac{\lambda}{\omega}(u + \omega(v-u)).$$

If we assume that $0 < \lambda \leq \omega$, the convexity of f yields

$$f(u + \lambda(v - u)) \leq \frac{\omega - \lambda}{\omega} f(u) + \frac{\lambda}{\omega} f(u + \omega(v - u)).$$

If we subtract $f(u)$ to both sides, we get

$$\frac{f(u + \lambda(v - u)) - f(u)}{\lambda} \leq \frac{f(u + \omega(v - u)) - f(u)}{\omega}.$$

Now, since $0 < \omega < 1$ and f is strictly convex,

$$f(u + \omega(v - u)) = f((1 - \omega)u + \omega v) < (1 - \omega)f(u) + \omega f(v),$$

which implies that

$$\frac{f(u + \omega(v - u)) - f(u)}{\omega} < f(v) - f(u),$$

and thus we get

$$\frac{f(u + \lambda(v - u)) - f(u)}{\lambda} \leq \frac{f(u + \omega(v - u)) - f(u)}{\omega} < f(v) - f(u).$$

If we let λ go to 0, by passing to the limit we get

$$df(u)(v - u) \leq \frac{f(u + \omega(v - u)) - f(u)}{\omega} < f(v) - f(u),$$

which yields the desired strict inequality.

Let us now consider the converse of (1); that is, assume that

$$f(v) \geq f(u) + df(u)(v - u) \quad \text{for all } u, v \in U.$$

For any two distinct points $u, v \in U$ and for any λ with $0 < \lambda < 1$, we get

$$\begin{aligned} f(v) &\geq f(v + \lambda(v - u)) - \lambda df(v + \lambda(u - v))(u - v) \\ f(u) &\geq f(v + \lambda(u - v)) + (1 - \lambda)df(v + \lambda(u - v))(u - v), \end{aligned}$$

and if we multiply the first inequality by $1 - \lambda$ and the second inequality by λ and them add up the resulting inequalities, we get

$$(1 - \lambda)f(v) + \lambda f(u) \geq f(v + \lambda(u - v)) = f((1 - \lambda)v + \lambda u),$$

which proves that f is convex.

The proof of the converse of (2) is similar, except that the inequalities are replaced by strict inequalities. $\square$

We now establish a convexity criterion using the second derivative of f. This criterion is often easier to check than the previous one.

Proposition 40.10. *(Convexity and second derivative) Let $f\colon \Omega \to \mathbb{R}$ be a function twice differentiable on some open subset Ω of a normed vector space E and let $U \subseteq \Omega$ be a nonempty convex subset.*

(1) The function f is convex on U iff

$$\mathrm{D}^2 f(u)(v-u, v-u) \geq 0 \quad \textit{for all } u, v \in U.$$

(2) If

$$\mathrm{D}^2 f(u)(v-u, v-u) > 0 \quad \textit{for all } u, v \in U \textit{ with } u \neq v,$$

then f is strictly convex.

Proof. First,assume that the inequality in Condition (1) is satisfied. For any two distinct points $u, v \in U$, the formula of Taylor–Maclaurin yields

$$\begin{aligned} f(v) - f(u) - df(u)(v-u) &= \frac{1}{2}\mathrm{D}^2 f(w)(v-u, v-u) \\ &= \frac{\rho^2}{2}\mathrm{D}^2 f(w)(v-w, v-w), \end{aligned}$$

for some $w = (1-\lambda)u + \lambda v = u + \lambda(v-u)$ with $0 < \lambda < 1$, and with $\rho = 1/(1-\lambda) > 0$, so that $v - u = \rho(v-w)$. Since $\mathrm{D}^2 f(u)(v-w, v-w) \geq 0$ for all $u, w \in U$, we conclude by applying Proposition 40.9(1).

Similarly, if (2) holds, the above reasoning and Proposition 40.9(2) imply that f is strictly convex.

To prove the necessary condition in (1), define $g\colon \Omega \to \mathbb{R}$ by

$$g(v) = f(v) - df(u)(v),$$

where $u \in U$ is any point considered fixed. If f is convex, since

$$g(v) - g(u) = f(v) - f(u) - df(u)(v-u),$$

Proposition 40.9 implies that $f(v) - f(u) - df(u)(v-u) \geq 0$, which implies that g has a local minimum at u with respect to all $v \in U$. Therefore, we have $dg(u) = 0$. Observe that g is twice differentiable in Ω and $\mathrm{D}^2 g(u) = \mathrm{D}^2 f(u)$, so the formula of Taylor–Young yields for every $v = u + w \in U$ and all t with $0 \leq t \leq 1$,

$$\begin{aligned} 0 \leq g(u+tw) - g(u) &= \frac{t^2}{2}\mathrm{D}^2 f(u)(tw, tw) + \left\|tw\right\|^2 \epsilon(tw) \\ &= \frac{t^2}{2}(\mathrm{D}^2 f(u)(w, w) + 2\left\|w\right\|^2 \epsilon(wt)), \end{aligned}$$

with $\lim_{t \mapsto 0} \epsilon(wt) = 0$, and for t small enough, we must have $\mathrm{D}^2 f(u)(w, w) \geq 0$, as claimed. □

The converse of Proposition 40.10 (2) is false as we see by considering the function f given by $f(x) = x^4$.

Example 40.1. On the other hand, if f is a quadratic function of the form

$$f(u) = \frac{1}{2}u^\top Au - u^\top b$$

where A is a symmetric matrix, we know that

$$df(u)(v) = v^\top (Au - b),$$

so

$$\begin{aligned} f(v) - f(u) - df(u)(v-u) &= \frac{1}{2}v^\top Av - v^\top b - \frac{1}{2}u^\top Au + u^\top b - (v-u)^\top (Au - b) \\ &= \frac{1}{2}v^\top Av - \frac{1}{2}u^\top Au - (v-u)^\top Au \\ &= \frac{1}{2}v^\top Av + \frac{1}{2}u^\top Au - v^\top Au \\ &= \frac{1}{2}(v-u)^\top A(v-u). \end{aligned}$$

Therefore, Proposition 40.9 implies that if A is positive semidefinite, then f is convex and if A is positive definite, then f is strictly convex. The converse follows by Proposition 40.10.

We conclude this section by applying our previous theorems to convex functions defined on convex subsets. In this case, local minima (resp. local maxima) are global minima (resp. global maxima).

Definition 40.6. Let $f\colon E \to \mathbb{R}$ be any function defined on some normed vector space (or more generally, any set). For any $u \in E$, we say that f has a *minimum* in u (resp. *maximum* in u) if

$$f(u) \leq f(v) \ \ (\text{resp.}\ f(u) \geq f(v)) \quad \text{for all } v \in E.$$

We say that f has a *strict minimum* in u (resp. *strict maximum* in u) if

$$f(u) < f(v) \ \ (\text{resp.}\ f(u) > f(v)) \quad \text{for all } v \in E - \{u\}.$$

If $U \subseteq E$ is a subset of E and $u \in U$, we say that f has a *minimum* in u (resp. *strict minimum* in u) *with respect to U* if

$$f(u) \leq f(v) \quad \text{for all } v \in U \quad (\text{resp.}\ f(u) < f(v) \quad \text{for all } v \in U - \{u\}),$$

and similarly for a *maximum* in u (resp. *strict maximum* in u) *with respect to U* with $\leq$ changed to $\geq$ and $<$ to $>$.

Sometimes, we say *global* maximum (or minimum) to stress that a maximum (or a minimum) is not simply a local maximum (or minimum).

Theorem 40.11. *Given any normed vector space E, let U be any nonempty convex subset of E.*

(1) For any convex function $J\colon U \to \mathbb{R}$, for any $u \in U$, if J has a local minimum at u in U, then J has a (global) minimum at u in U.

(2) Any strict convex function $J\colon U \to \mathbb{R}$ has at most one minimum (in U), and if it does, then it is a strict minimum (in U).

(3) Let $J\colon \Omega \to \mathbb{R}$ be any function defined on some open subset Ω of E with $U \subseteq \Omega$ and assume that J is convex on U. For any point $u \in U$, if $dJ(u)$ exists, then J has a minimum in u with respect to U iff

$$dJ(u)(v-u) \geq 0 \quad \textit{for all } v \in U.$$

(4) If the convex subset U in (3) is open, then the above condition is equivalent to

$$dJ(u) = 0.$$

Proof. (1) Let $v = u + w$ be any arbitrary point in U. Since J is convex, for all t with $0 \leq t \leq 1$, we have

$$J(u + tw) = J(u + t(v-u)) \leq (1-t)J(u) + tJ(v),$$

which yields

$$J(u + tw) - J(u) \leq t(J(v) - J(u)).$$

Because J has a local minimum in u, there is some t_0 with $0 < t_0 < 1$ such that

$$0 \leq J(u + t_0 w) - J(u),$$

which implies that $J(v) - J(u) \geq 0$.

(2) If J is strictly convex, the above reasoning with $w \neq 0$ shows that there is some t_0 with $0 < t_0 < 1$ such that

$$0 \leq J(u + t_0 w) - J(u) < t_0(J(v) - J(u)),$$

which shows that u is a strict global minimum (in U), and thus that it is unique.

(3) We already know from Theorem 40.8 that the condition $dJ(u)(v-u) \geq 0$ for all $v \in U$ is necessary (even if J is not convex). Conversely, because J is convex, careful inspection of the proof of part (1) of Proposition 40.9 shows that only the fact that $dJ(u)$ exists in needed to prove that

$$J(v) - J(u) \geq dJ(u)(v-u) \quad \text{for all } v \in U,$$

and if

$$dJ(u)(v-u) \geq 0 \quad \text{for all } v \in U,$$

then

$$J(v) - J(u) \geq 0 \quad \text{for all } v \in U,$$

as claimed.

(4) If U is open, then for every $u \in U$ we can find an open ball B centered at u of radius ϵ small enough so that $B \subseteq U$. Then, for any $w \neq 0$ such that $\|w\| < \epsilon$, we have both $v = u + w \in B$ and $v' = u - w \in B$, so condition (3) implies that

$$dJ(u)(w) \geq 0 \quad \text{and} \quad dJ(u)(-w) \geq 0,$$

which yields

$$dJ(u)(w) = 0.$$

Since the above holds for all $w \neq 0$ such such that $\|w\| < \epsilon$ and since $dJ(u)$ is linear, we leave it to the reader to fill in the details of the proof that $dJ(u) = 0$. □

Theorem 40.11 can be used to rederive the fact that the least squares solutions of a linear system $Ax = b$ (where A is an $m \times n$ matrix) are given by the normal equation

$$A^{\top} A x = A^{\top} b.$$

For this, we consider the quadratic function

$$J(v) = \frac{1}{2} \|Av - b\|_2^2 - \frac{1}{2} \|b\|_2^2,$$

and our least squares problem is equivalent to finding the minima of J on $\mathbb{R}^n$. A computation reveals that

$$\begin{aligned} J(v) &= \frac{1}{2} \|Av - b\|_2^2 - \frac{1}{2} \|b\|_2^2 \\ &= \frac{1}{2}(Av - b)^{\top}(Av - b) - \frac{1}{2} b^{\top} b \\ &= \frac{1}{2}(v^{\top} A^{\top} - b^{\top})(Av - b) - \frac{1}{2} b^{\top} b \\ &= \frac{1}{2} v^{\top} A^{\top} A v - v^{\top} A^{\top} b, \end{aligned}$$

and so

$$dJ(u) = A^{\top} A u - A^{\top} b.$$

Since $A^{\top} A$ is positive semidefinite, the function J is convex, and Theorem 40.11(4) implies that the minima of J are the solutions of the equation

$$A^{\top} A u - A^{\top} b = 0.$$

The considerations in this chapter reveal the need to find methods for finding the zeros of the derivative map

$$dJ \colon \Omega \to E',$$

where Ω is some open subset of a normed vector space E and E' is the space of all continuous linear forms on E (a subspace of E^*). Generalizations of *Newton's method* yield such methods and they are the objet of the next chapter.

40.4 Summary

The main concepts and results of this chapter are listed below:

- *Local minimum*, *local maximum*, *local extremum*, *strict local minimum*, *strict local maximum*.
- Necessary condition for a local extremum involving the derivative; *critical point*.
- *Local minimum with respect to a subset U*, *local maximum with respect to a subset U*, *local extremum with respect to a subset U*.
- *Constrained local extremum*.
- Necessary condition for a constrained extremum.
- Necessary condition for a constrained extremum in terms of *Lagrange multipliers*.
- *Lagrangian*.
- *Critical points of a Lagrangian*.
- Necessary condition of an unconstrained local minimum involving the second-order derivative.
- Sufficient condition for a local minimum involving the second-order derivative.
- A sufficient condition involving *nondegenerate critical points*.
- *Convex sets*, *convex functions*, *concave functions*, *strictly convex functions*, *strictly concave functions*,
- Necessary condition for a local minimum on a convex set involving the derivative.
- Convexity of a function involving a condition on its first derivative.
- Convexity of a function involving a condition on its second derivative.
- Minima of convex functions on convex sets.

Chapter 41

Newton's Method and Its Generalizations

41.1 Newton's Method for Real Functions of a Real Argument

In Chapter 40 we investigated the problem of determining when a function $J\colon \Omega \to \mathbb{R}$ defined on some open subset Ω of a normed vector space E has a local extremum. Proposition 40.1 gives a necessary condition when J is differentiable: if J has a local extremum at $u \in \Omega$, then we must have

$$J'(u) = 0.$$

Thus we are led to the problem of finding the zeros of the derivative

$$J'\colon \Omega \to E',$$

where $E' = \mathcal{L}(E;\mathbb{R})$ is the set of linear continuous functions from E to $\mathbb{R}$; that is, the *dual* of E, as defined in the remark after Proposition 40.7.

This leads us to consider the problem in a more general form, namely: Given a function $f\colon \Omega \to Y$ from an open subset Ω of a normed vector space X to a normed vector space Y, find

(i) Sufficient conditions which guarantee the *existence of a zero* of the function f; that is, an element $a \in \Omega$ such that $f(a) = 0$.

(ii) An *algorithm* for approximating such an a, that is, a sequence (x_k) of points of Ω whose limit is a.

When $X = Y = \mathbb{R}$, we can use *Newton's method*. We pick some initial element $x_0 \in \mathbb{R}$ "close enough" to a zero a of f, and we define the sequence (x_k) by

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},$$

for all $k \geq 0$, provided that $f'(x_k) \neq 0$. The idea is to define x_{k+1} as the intersection of the x-axis with the tangent line to the graph of the function $x \mapsto f(x)$ at the point $(x_k, f(x_k))$. Indeed, the equation of this tangent line is

$$y - f(x_k) = f'(x_k)(x - x_k),$$

and its intersection with the x-axis is obtained for $y = 0$, which yields

$$x = x_k - \frac{f(x_k)}{f'(x_k)},$$

as claimed.

For example, if $\alpha > 0$ and $f(x) = x^2 - \alpha$, Newton's method yields the sequence

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{\alpha}{x_k}\right)$$

to compute the square root $\sqrt{\alpha}$ of α. It can be shown that the method converges to $\sqrt{\alpha}$ for any $x_0 > 0$. Actually, the method also converges when $x_0 < 0$! Find out what is the limit.

The case of a real function suggests the following method for finding the zeros of a function $f\colon \Omega \to Y$, with $\Omega \subseteq X$: given a starting point $x_0 \in \Omega$, the sequence (x_k) is defined by

$$x_{k+1} = x_k - (f'(x_k))^{-1}(f(x_k))$$

for all $k \geq 0$.

For the above to make sense, it must be ensured that

(1) All the points x_k remain within Ω.

(2) The function f is differentiable within Ω.

(3) The derivative $f'(x)$ is a bijection from X to Y for all $x \in \Omega$.

These are rather demanding conditions but there are sufficient conditions that guarantee that they are met. Another practical issue is that it may be very costly to compute $(f'(x_k))^{-1}$ at every iteration step. In the next section, we investigate generalizations of Newton's method which address the issues that we just discussed.

41.2 Generalizations of Newton's Method

Suppose that $f\colon \Omega \to \mathbb{R}^n$ is given by n functions $f_i\colon \Omega \to \mathbb{R}$, where $\Omega \subseteq \mathbb{R}^n$. In this case, finding a zero a of f is equivalent to solving the system

$$\begin{aligned}
f_1(a_1 \ldots, a_n) &= 0 \\
f_2(a_1 \ldots, a_n) &= 0 \\
&\vdots \\
f_n(a_1 \ldots, a_n) &= 0.
\end{aligned}$$

A single iteration of Newton's method consists in solving the linear system

$$(J(f)(x_k))\epsilon_k = -f(x_k),$$

and then setting

$$x_{k+1} = x_k + \epsilon_k,$$

where $J(f)(x_k) = (\frac{\partial f_i}{\partial x_j}(x_k))$ is the Jacobian matrix of f at x_k.

In general, it is very costly to compute $J(f)(x_k)$ at each iteration and then to solve the corresponding linear system. If the method converges, the consecutive vectors x_k should differ only a little, as also the corresponding matrices $J(f)(x_k)$. Thus, we are led to a variant of Newton's method which consists in keeping the same matrix for p consecutive steps (where p is some fixed integer ≥ 2):

$$\begin{aligned}
x_{k+1} &= x_k - (f'(x_0))^{-1}(f(x_k)), && 0 \leq k \leq p-1\\
x_{k+1} &= x_k - (f'(x_p))^{-1}(f(x_k)), && p \leq k \leq 2p-1\\
&\vdots\\
x_{k+1} &= x_k - (f'(x_{rp}))^{-1}(f(x_k)), && rp \leq k \leq (r+1)p-1\\
&\vdots
\end{aligned}$$

It is also possible to set $p = \infty$, that is, to use the same matrix $f'(x_0)$ for all iterations, which leads to iterations of the form

$$x_{k+1} = x_k - (f'(x_0))^{-1}(f(x_k)), \quad k \geq 0,$$

or even to replace $f'(x_0)$ by a particular matrix A_0 which is easy to invert:

$$x_{k+1} = x_k - A_0^{-1}f(x_k), \quad k \geq 0.$$

In the last two cases, if possible, we use an LU factorization of $f'(x_0)$ or A_0 to speed up the method. In some cases, it may even possible to set $A_0 = I$.

The above considerations lead us to the definition of a *generalized Newton method*, as in Ciarlet [41] (Chapter 7). Recall that a linear map $f \in \mathcal{L}(E; F)$ is called an *isomorphism* iff f is continuous, bijective, and f^{-1} is also continuous.

Definition 41.1. If X and Y are two normed vector spaces and if $f\colon \Omega \to Y$ is a function from some open subset Ω of X, a *generalized Newton method* for finding zeros of f consists of

(1) A sequence of families $(A_k(x))$ of linear isomorphisms from X to Y, for all $x \in \Omega$ and all integers $k \geq 0$;

(2) Some starting point $x_0 \in \Omega$;

(3) A sequence (x_k) of points of Ω defined by

$$x_{k+1} = x_k - (A_k(x_\ell))^{-1}(f(x_k)), \quad k \geq 0,$$

where for every integer $k \geq 0$, the integer ℓ satisfies the condition

$$0 \leq \ell \leq k.$$

The function $A_k(x)$ usually depends on f'.

Definition 41.1 gives us enough flexibility to capture all the situations that we have previously discussed:

$$\begin{aligned}
&A_k(x) = f'(x), && \ell = k \\
&A_k(x) = f'(x), && \ell = \min\{rp, k\},\ \text{if}\ rp \leq k \leq (r+1)p - 1, r \geq 0 \\
&A_k(x) = f'(x), && \ell = 0 \\
&A_k(x) = A_0, &&
\end{aligned}$$

where A_0 is a linear isomorphism from X to Y. The first case corresponds to Newton's orginal method and the others to the variants that we just discussed. We could also have $A_k(x) = A_k$, a fixed linear isomorphism independent of $x \in \Omega$.

The following theorem inspired by the *Newton–Kantorovich theorem* gives sufficient conditions that guarantee that the sequence (x_k) constructed by a generalized Newton method converges to a zero of f close to x_0. Althoug quite technical, these conditions are not very surprising.

Theorem 41.1. *Let X be a Banach space, let $f\colon \Omega \to Y$ be differentiable on the open subset $\Omega \subseteq X$, and assume that there are constants $r, M, \beta > 0$ such that if we let*

$$B = \{x \in X \mid \|x - x_0\| \leq r\} \subseteq \Omega,$$

then

(1)

$$\sup_{k \geq 0} \sup_{x \in B} \left\| A_k^{-1}(x) \right\|_{\mathcal{L}(Y;X)} \leq M,$$

(2) $\beta < 1$ and

$$\sup_{k \geq 0} \sup_{x, x' \in B} \|f'(x) - A_k(x')\|_{\mathcal{L}(X;Y)} \leq \frac{\beta}{M}$$

(3)

$$\|f(x_0)\| \leq \frac{r}{M}(1 - \beta).$$

Then, the sequence (x_k) *defined by*

$$x_{k+1} = x_k - A_k^{-1}(x_\ell)(f(x_k)), \quad 0 \leq \ell \leq k$$

is entirely contained within B *and converges to a zero* a *of* f*, which is the only zero of* f *in* B*. Furthermore, the convergence is geometric, which means that*

$$\|x_k - a\| \leq \frac{\|x_1 - x_0\|}{1 - \beta} \beta^k.$$

A proof of Theorem 41.1 can be found in Ciarlet [41] (Section 7.5). It is not really difficult but quite technical.

If we assume that we already know that some element $a \in \Omega$ is a zero of f, the next theorem gives sufficient conditions for a special version of a generalized Newton method to converge. For this special method, the linear isomorphisms $A_k(x)$ are independent of $x \in \Omega$.

Theorem 41.2. *Let* X *be a Banach space, and let* $f \colon \Omega \to Y$ *be differentiable on the open subset* $\Omega \subseteq X$*. If* $a \in \Omega$ *is a point such that* $f(a) = 0$*, if* $f'(a)$ *is a linear isomorphism, and if there is some* λ *with* $0 < \lambda < 1/2$ *such that*

$$\sup_{k \geq 0} \|A_k - f'(a)\|_{\mathcal{L}(X;Y)} \leq \frac{\lambda}{\|(f'(a))^{-1}\|_{\mathcal{L}(Y;X)}},$$

then there is a closed ball B *of center* a *such that for every* $x_0 \in B$*, the sequence* (x_k) *defined by*

$$x_{k+1} = x_k - A_k^{-1}(f(x_k)), \quad k \geq 0,$$

is entirely contained within B *and converges to* a*, which is the only zero of* f *in* B*. Furthermore, the convergence is geometric, which means that*

$$\|x_k - a\| \leq \beta^k \|x_0 - a\|,$$

for some $\beta < 1$*.*

A proof of Theorem 41.2 can be also found in Ciarlet [41] (Section 7.5).

For the sake of completeness, we state a version of the Newton–Kantorovich theorem, which corresponds to the case where $A_k(x) = f'(x)$. In this instance, a stronger result can be obtained especially regarding upper bounds, and we state a version due to Gragg and Tapia which appears in Problem 7.5-4 of Ciarlet [41].

Theorem 41.3. *(Newton–Kantorovich) Let* X *be a Banach space, and let* $f \colon \Omega \to Y$ *be differentiable on the open subset* $\Omega \subseteq X$*. Assume that there exist three positive constants* λ, μ, ν *and a point* $x_0 \in \Omega$ *such that*

$$0 < \lambda\mu\nu \leq \frac{1}{2},$$

and if we let

$$\begin{aligned}\rho^- &= \frac{1-\sqrt{1-2\lambda\mu\nu}}{\mu\nu}\\ \rho^+ &= \frac{1+\sqrt{1-2\lambda\mu\nu}}{\mu\nu}\\ B &= \{x \in X \mid \|x - x_0\| < \rho^-\}\\ \Omega^+ &= \{x \in \Omega \mid \|x - x_0\| < \rho^+\},\end{aligned}$$

then $\overline{B} \subseteq \Omega$, $f'(x_0)$ *is an isomorphism of* $\mathcal{L}(X;Y)$, *and*

$$\begin{aligned}\left\|(f'(x_0))^{-1}\right\| &\leq \mu,\\ \left\|(f'(x_0))^{-1}f(x_0)\right\| &\leq \lambda,\\ \sup_{x,y\in\Omega^+} \|f'(x) - f'(y)\| &\leq \nu \|x - y\|.\end{aligned}$$

Then, $f'(x)$ *is isomorphism of* $\mathcal{L}(X;Y)$ *for all* $x \in B$, *and the sequence defined by*

$$x_{k+1} = x_k - (f'(x_k))^{-1}(f(x_k)), \quad k \geq 0$$

is entirely contained within the ball B *and converges to a zero* a *of* f *which is the only zero of* f *in* Ω^+. *Finally, if we write* $\theta = \rho^-/\rho^+$, *then we have the following bounds:*

$$\begin{aligned}\|x_k - a\| &\leq \frac{2\sqrt{1-2\lambda\mu\nu}}{\lambda\mu\nu}\frac{\theta^{2k}}{1-\theta^{2k}}\|x_1 - x_0\| && \text{if } \lambda\mu\nu < \frac{1}{2}\\ \|x_k - a\| &\leq \frac{\|x_1 - x_0\|}{2^{k-1}} && \text{if } \lambda\mu\nu = \frac{1}{2},\end{aligned}$$

and

$$\frac{2\|x_{k+1} - x_k\|}{1 + \sqrt{(1 + 4\theta^{2k}(1+\theta^{2k})^{-2})}} \leq \|x_k - a\| \leq \theta^{2k-1}\|x_k - x_{k-1}\|.$$

We can now specialize Theorems 41.1 and 41.2 to the search of zeros of the derivative $J'\colon \Omega \to E'$, of a function $J\colon \Omega \to \mathbb{R}$, with $\Omega \subseteq E$. The second derivative J'' of J is a continuous bilinear form $J''\colon E \times E \to \mathbb{R}$, but is is convenient to view it as a linear map in $\mathcal{L}(E, E')$; the continuous linear form $J''(u)$ is given by $J''(u)(v) = J''(u, v)$. In our next theorem, we assume that the $A_k(x)$ are isomorphisms in $\mathcal{L}(E, E')$.

Theorem 41.4. *Let* E *be a Banach space, let* $J\colon \Omega \to \mathbb{R}$ *be twice differentiable on the open subset* $\Omega \subseteq E$, *and assume that there are constants* $r, M, \beta > 0$ *such that if we let*

$$B = \{x \in E \mid \|x - x_0\| \leq r\} \subseteq \Omega,$$

then

(1)
$$\sup_{k\geq 0}\sup_{x\in B}\left\|A_k^{-1}(x)\right\|_{\mathcal{L}(E';E)}\leq M,$$

(2) $\beta < 1$ *and*
$$\sup_{k\geq 0}\sup_{x,x'\in B}\|J''(x)-A_k(x')\|_{\mathcal{L}(E;E')}\leq\frac{\beta}{M}$$

(3)
$$\|J'(x_0)\|\leq\frac{r}{M}(1-\beta).$$

Then, the sequence (x_k) *defined by*
$$x_{k+1}=x_k-A_k^{-1}(x_\ell)(J'(x_k)),\quad 0\leq\ell\leq k$$
is entirely contained within B *and converges to a zero* a *of* J'*, which is the only zero of* J' *in* B*. Furthermore, the convergence is geometric, which means that*
$$\|x_k-a\|\leq\frac{\|x_1-x_0\|}{1-\beta}\beta^k.$$

In the next theorem, we assume that the $A_k(x)$ are isomorphisms in $\mathcal{L}(E,E')$ that are independent of $x\in\Omega$.

Theorem 41.5. *Let* E *be a Banach space, and let* $J\colon\Omega\to\mathbb{R}$ *be twice differentiable on the open subset* $\Omega\subseteq E$*. If* $a\in\Omega$ *is a point such that* $J'(a)=0$*, if* $J''(a)$ *is a linear isomorphism, and if there is some* λ *with* $0<\lambda<1/2$ *such that*
$$\sup_{k\geq 0}\|A_k-J''(a)\|_{\mathcal{L}(E;E')}\leq\frac{\lambda}{\|(J''(a))^{-1}\|_{\mathcal{L}(E';E)}},$$
then there is a closed ball B *of center* a *such that for every* $x_0\in B$*, the sequence* (x_k) *defined by*
$$x_{k+1}=x_k-A_k^{-1}(J'(x_k)),\quad k\geq 0,$$
is entirely contained within B *and converges to* a*, which is the only zero of* J' *in* B*. Furthermore, the convergence is geometric, which means that*
$$\|x_k-a\|\leq\beta^k\|x_0-a\|,$$
for some $\beta<1$*.*

When $E=\mathbb{R}^n$, the Newton method given by Theorem 41.4 yield an iteration step of the form
$$x_{k+1}=x_k-A_k^{-1}(x_\ell)\nabla J(x_k),\quad 0\leq\ell\leq k,$$

where $\nabla J(x_k)$ is the gradient of J at x_k (here, we identify E' with $\mathbb{R}^n$). In particular, Newton's original method picks $A_k = J''$, and the iteration step is of the form

$$x_{k+1} = x_k - (\nabla^2 J(x_k))^{-1} \nabla J(x_k), \quad k \geq 0,$$

where $\nabla^2 J(x_k)$ is the Hessian of J at x_k.

As remarked in Ciarlet [41] (Section 7.5), generalized Newton methods have a very wide range of applicability. For example, various versions of gradient descent methods can be viewed as instances of Newton method. See Section 49.9 for an example.

Newton's method also plays an important role in convex optimization, in particular, interior-point methods. A variant of Newton's method dealing with equality constraints has been developed. We refer the reader to Boyd and Vandenberghe [29], Chapters 10 and 11, for a comprehensive exposition of these topics.

41.3 Summary

The main concepts and results of this chapter are listed below:

- Newton's method for functions $f\colon \mathbb{R} \to \mathbb{R}$.
- Generalized Newton methods.
- The *Newton-Kantorovich* theorem.

Chapter 42

Quadratic Optimization Problems

42.1 Quadratic Optimization: The Positive Definite Case

In this chapter, we consider two classes of quadratic optimization problems that appear frequently in engineering and in computer science (especially in computer vision):

1. Minimizing
$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b$$
over all $x \in \mathbb{R}^n$, or subject to linear or affine constraints.

2. Minimizing
$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b$$
over the unit sphere.

In both cases, A is a symmetric matrix. We also seek necessary and sufficient conditions for f to have a global minimum.

Many problems in physics and engineering can be stated as the minimization of some energy function, with or without constraints. Indeed, it is a fundamental principle of mechanics that nature acts so as to minimize energy. Furthermore, if a physical system is in a stable state of equilibrium, then the energy in that state should be minimal. For example, a small ball placed on top of a sphere is in an unstable equilibrium position. A small motion causes the ball to roll down. On the other hand, a ball placed inside and at the bottom of a sphere is in a stable equilibrium position, because the potential energy is minimal.

The simplest kind of energy function is a quadratic function. Such functions can be conveniently defined in the form

$$Q(x) = x^\top Ax - x^\top b,$$

where A is a symmetric $n \times n$ matrix, and x, b, are vectors in $\mathbb{R}^n$, viewed as column vectors. Actually, for reasons that will be clear shortly, it is preferable to put a factor $\frac{1}{2}$ in front of the quadratic term, so that

$$Q(x) = \frac{1}{2} x^\top A x - x^\top b.$$

The question is, under what conditions (on A) does $Q(x)$ have a global minimum, preferably unique?

We give a complete answer to the above question in two stages:

1. In this section, we show that if A is symmetric positive definite, then $Q(x)$ has a unique global minimum precisely when
$$Ax = b.$$

2. In Section 42.2, we give necessary and sufficient conditions in the general case, in terms of the pseudo-inverse of A.

We begin with the matrix version of Definition 22.2 (Vol. I).

Definition 42.1. A symmetric *positive definite matrix* is a matrix whose eigenvalues are strictly positive, and a symmetric *positive semidefinite matrix* is a matrix whose eigenvalues are nonnegative.

Equivalent criteria are given in the following proposition.

Proposition 42.1. *Given any Euclidean space E of dimension n, the following properties hold:*

(1) Every self-adjoint linear map $f \colon E \to E$ is positive definite iff

$$\langle f(x), x \rangle > 0$$

for all $x \in E$ with $x \neq 0$.

(2) Every self-adjoint linear map $f \colon E \to E$ is positive semidefinite iff

$$\langle f(x), x \rangle \geq 0$$

for all $x \in E$.

Proof. (1) First, assume that f is positive definite. Recall that every self-adjoint linear map has an orthonormal basis $(e_1, \ldots, e_n)$ of eigenvectors, and let $\lambda_1, \ldots, \lambda_n$ be the corresponding eigenvalues. With respect to this basis, for every $x = x_1 e_1 + \cdots + x_n e_n \neq 0$, we have

$$\langle f(x), x \rangle = \Big\langle f\Big(\sum_{i=1}^n x_i e_i\Big), \sum_{i=1}^n x_i e_i \Big\rangle = \Big\langle \sum_{i=1}^n \lambda_i x_i e_i, \sum_{i=1}^n x_i e_i \Big\rangle = \sum_{i=1}^n \lambda_i x_i^2,$$

which is strictly positive, since $\lambda_i > 0$ for $i = 1, \ldots, n$, and $x_i^2 > 0$ for some i, since $x \neq 0$.

Conversely, assume that

$$\langle f(x), x \rangle > 0$$

for all $x \neq 0$. Then for $x = e_i$, we get

$$\langle f(e_i), e_i \rangle = \langle \lambda_i e_i, e_i \rangle = \lambda_i,$$

and thus $\lambda_i > 0$ for all $i = 1, \ldots, n$.

(2) As in (1), we have

$$\langle f(x), x \rangle = \sum_{i=1}^{n} \lambda_i x_i^2,$$

and since $\lambda_i \geq 0$ for $i = 1, \ldots, n$ because f is positive semidefinite, we have $\langle f(x), x \rangle \geq 0$, as claimed. The converse is as in (1) except that we get only $\lambda_i \geq 0$ since $\langle f(e_i), e_i \rangle \geq 0$. □

Some special notation is customary (especially in the field of convex optimization) to express that a symmetric matrix is positive definite or positive semidefinite.

Definition 42.2. Given any $n \times n$ symmetric matrix A we write $A \succeq 0$ if A is positive semidefinite and we write $A \succ 0$ if A is positive definite.

It should be noted that we can define the relation

$$A \succeq B$$

between any two $n \times n$ matrices (symmetric or not) iff $A - B$ is symmetric positive semidefinite. It is easy to check that this relation is actually a partial order on matrices, called the *positive semidefinite cone ordering*; for details, see Boyd and Vandenberghe [29], Section 2.4.

If A is symmetric positive definite, it is easily checked that A^{-1} is also symmetric positive definite. Also, if C is a symmetric positive definite $m \times m$ matrix and A is an $m \times n$ matrix of rank n (and so $m \geq n$ and the map $x \mapsto Ax$ is surjective onto $\mathbb{R}^m$), then $A^\top CA$ is symmetric positive definite.

We can now prove that

$$Q(x) = \frac{1}{2} x^\top Ax - x^\top b$$

has a global minimum when A is symmetric positive definite.

Proposition 42.2. *Given a quadratic function*

$$Q(x) = \frac{1}{2} x^\top Ax - x^\top b,$$

if A is symmetric positive definite, then $Q(x)$ has a unique global minimum for the solution of the linear system $Ax = b$. The minimum value of $Q(x)$ is

$$Q(A^{-1}b) = -\frac{1}{2} b^\top A^{-1} b.$$

Proof. Since A is positive definite, it is invertible, since its eigenvalues are all strictly positive. Let $x = A^{-1}b$, and compute $Q(y) - Q(x)$ for any $y \in \mathbb{R}^n$. Since $Ax = b$, we get

$$\begin{aligned}
Q(y) - Q(x) &= \frac{1}{2}y^\top Ay - y^\top b - \frac{1}{2}x^\top Ax + x^\top b \\
&= \frac{1}{2}y^\top Ay - y^\top Ax + \frac{1}{2}x^\top Ax \\
&= \frac{1}{2}(y - x)^\top A(y - x).
\end{aligned}$$

Since A is positive definite, the last expression is nonnegative, and thus

$$Q(y) \geq Q(x)$$

for all $y \in \mathbb{R}^n$, which proves that $x = A^{-1}b$ is a global minimum of $Q(x)$. A simple computation yields

$$Q(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b.$$

□

Remarks:

(1) The quadratic function $Q(x)$ is also given by

$$Q(x) = \frac{1}{2}x^\top Ax - b^\top x,$$

but the definition using $x^\top b$ is more convenient for the proof of Proposition 42.2.

(2) If $Q(x)$ contains a constant term $c \in \mathbb{R}$, so that

$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b + c,$$

the proof of Proposition 42.2 still shows that $Q(x)$ has a unique global minimum for $x = A^{-1}b$, but the minimal value is

$$Q(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b + c.$$

Thus, when the energy function $Q(x)$ of a system is given by a quadratic function

$$Q(x) = \frac{1}{2}x^\top Ax - x^\top b,$$

where A is symmetric positive definite, finding the global minimum of $Q(x)$ is equivalent to solving the linear system $Ax = b$. Sometimes, it is useful to recast a linear problem $Ax = b$

as a variational problem (finding the minimum of some energy function). However, very often, a minimization problem comes with extra constraints that must be satisfied for all admissible solutions. For instance, we may want to minimize the quadratic function

$$Q(x_1, x_2) = \frac{1}{2}\big(x_1^2 + x_2^2\big)$$

subject to the constraint

$$2x_1 - x_2 = 5.$$

The solution for which $Q(x_1, x_2)$ is minimum is no longer $(x_1, x_2) = (0, 0)$, but instead, $(x_1, x_2) = (2, -1)$, as will be shown later.

Geometrically, the graph of the function defined by $z = Q(x_1, x_2)$ in $\mathbb{R}^3$ is a paraboloid of revolution P with axis of revolution Oz. The constraint

$$2x_1 - x_2 = 5$$

corresponds to the vertical plane H parallel to the z-axis and containing the line of equation $2x_1 - x_2 = 5$ in the xy-plane. Thus, the constrained minimum of Q is located on the parabola that is the intersection of the paraboloid P with the plane H.

A nice way to solve constrained minimization problems of the above kind is to use the method of *Lagrange multipliers* discussed in Section 40.1. But first, let us define precisely what kind of minimization problems we intend to solve.

Definition 42.3. The *quadratic constrained minimization problem* consists in minimizing a quadratic function

$$Q(x) = \frac{1}{2}x^\top A^{-1}x - b^\top x$$

subject to the linear constraints

$$B^\top x = f,$$

where A^{-1} is an $m \times m$ symmetric positive definite matrix, B is an $m \times n$ matrix of rank n (so that $m \geq n$), and where $b, x \in \mathbb{R}^m$ (viewed as column vectors), and $f \in \mathbb{R}^n$ (viewed as a column vector).

The reason for using A^{-1} instead of A is that the constrained minimization problem has an interpretation as a set of equilibrium equations in which the matrix that arises naturally is A (see Strang [167]). Since A and A^{-1} are both symmetric positive definite, this doesn't make any difference, but it seems preferable to stick to Strang's notation.

As explained in Section 40.1, the method of Lagrange multipliers consists in incorporating the n constraints $B^\top x = f$ into the quadratic function $Q(x)$, by introducing extra variables $\lambda = (\lambda_1, \ldots, \lambda_n)$ called *Lagrange multipliers*, one for each constraint. We form the *Lagrangian*

$$L(x, \lambda) = Q(x) + \lambda^\top(B^\top x - f) = \frac{1}{2}x^\top A^{-1}x - (b - B\lambda)^\top x - \lambda^\top f.$$

We know from Theorem 40.3 that a necessary condition for our constrained optimization problem to have a solution is that $\nabla L(x,\lambda) = 0$. Since

$$\begin{aligned}\frac{\partial L}{\partial x}(x,\lambda) &= A^{-1}x - (b - B\lambda)\\ \frac{\partial L}{\partial \lambda}(x,\lambda) &= B^\top x - f,\end{aligned}$$

we obtain the system of linear equations

$$\begin{aligned}A^{-1}x + B\lambda &= b,\\ B^\top x &= f,\end{aligned}$$

which can be written in matrix form as

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix}\begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}.$$

We shall prove in Proposition 42.3 below that our constrained minimization problem has a unique solution actually given by the above system.

Note that the matrix of this system is symmetric. We solve it as follows. Eliminating x from the first equation

$$A^{-1}x + B\lambda = b,$$

we get

$$x = A(b - B\lambda),$$

and substituting into the second equation, we get

$$B^\top A(b - B\lambda) = f,$$

that is,

$$B^\top AB\lambda = B^\top Ab - f.$$

However, by a previous remark, since A is symmetric positive definite and the columns of B are linearly independent, $B^\top AB$ is symmetric positive definite, and thus invertible. Thus we obtain the solution

$$\lambda = (B^\top AB)^{-1}(B^\top Ab - f), \qquad x = A(b - B\lambda).$$

Note that this way of solving the system requires solving for the Lagrange multipliers first.

Letting $e = b - B\lambda$, we also note that the system

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix}\begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}$$

is equivalent to the system

$$\begin{aligned} e &= b - B\lambda, \\ x &= Ae, \\ B^\top x &= f. \end{aligned}$$

The latter system is called the *equilibrium equations* by Strang [167]. Indeed, Strang shows that the equilibrium equations of many physical systems can be put in the above form. This includes spring-mass systems, electrical networks, and trusses, which are structures built from elastic bars. In each case, x, e, b, A, λ, f, and $K = B^\top AB$ have a physical interpretation. The matrix $K = B^\top AB$ is usually called the *stiffness matrix*. Again, the reader is referred to Strang [167].

In order to prove that our constrained minimization problem has a unique solution, we proceed to prove that the constrained minimization of $Q(x)$ subject to $B^\top x = f$ is equivalent to the unconstrained maximization of another function $-G(\lambda)$. We get $G(\lambda)$ by minimizing the Lagrangian $L(x, \lambda)$ treated as a function of x alone. The function $-G(\lambda)$ is the *dual function* of the Lagrangian $L(x, \lambda)$. Here we are encountering a special case of the notion of dual function defined in Section 50.7.

Since A^{-1} is symmetric positive definite and

$$L(x, \lambda) = \frac{1}{2} x^\top A^{-1} x - (b - B\lambda)^\top x - \lambda^\top f,$$

by Proposition 42.2 the global minimum (with respect to x) of $L(x, \lambda)$ is obtained for the solution x of

$$A^{-1} x = b - B\lambda,$$

that is, when

$$x = A(b - B\lambda),$$

and the minimum of $L(x, \lambda)$ is

$$\min_x L(x, \lambda) = -\frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) - \lambda^\top f.$$

Letting

$$G(\lambda) = \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f,$$

we will show in Proposition 42.3 that the solution of the constrained minimization of $Q(x)$ subject to $B^\top x = f$ is equivalent to the unconstrained maximization of $-G(\lambda)$. This is a special case of the duality discussed in Section 50.7.

Of course, since we minimized $L(x, \lambda)$ with respect to x, we have

$$L(x, \lambda) \geq -G(\lambda)$$

for all x and all λ. However, when the constraint $B^\top x = f$ holds, $L(x, \lambda) = Q(x)$, and thus for any admissible x, which means that $B^\top x = f$, we have

$$\min_x Q(x) \geq \max_\lambda -G(\lambda).$$

In order to prove that the unique minimum of the constrained problem $Q(x)$ subject to $B^\top x = f$ is the unique maximum of $-G(\lambda)$, we compute $Q(x) + G(\lambda)$.

Proposition 42.3. *The quadratic constrained minimization problem of Definition 42.3 has a unique solution* (x, λ) *given by the system*

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}.$$

Furthermore, the component λ *of the above solution is the unique value for which* $-G(\lambda)$ *is maximum.*

Proof. As we suggested earlier, let us compute $Q(x) + G(\lambda)$, assuming that the constraint $B^\top x = f$ holds. Eliminating f, since $b^\top x = x^\top b$ and $\lambda^\top B^\top x = x^\top B\lambda$, we get

$$\begin{aligned} Q(x) + G(\lambda) &= \frac{1}{2} x^\top A^{-1} x - b^\top x + \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f \\ &= \frac{1}{2}(A^{-1}x + B\lambda - b)^\top A(A^{-1}x + B\lambda - b). \end{aligned}$$

Since A is positive definite, the last expression is nonnegative. In fact, it is null iff

$$A^{-1}x + B\lambda - b = 0,$$

that is,

$$A^{-1}x + B\lambda = b.$$

But then the unique constrained minimum of $Q(x)$ subject to $B^\top x = f$ is equal to the unique maximum of $-G(\lambda)$ exactly when $B^\top x = f$ and $A^{-1}x + B\lambda = b$, which proves the proposition. □

We can confirm that the maximum of $-G(\lambda)$, or equivalently the minimum of

$$G(\lambda) = \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f,$$

corresponds to value of λ obtained by solving the system

$$\begin{pmatrix} A^{-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ f \end{pmatrix}.$$

Indeed, since

$$G(\lambda) = \frac{1}{2}\lambda^\top B^\top AB\lambda - \lambda^\top B^\top Ab + \lambda^\top f + \frac{1}{2} b^\top b,$$

and $B^\top AB$ is symmetric positive definite, by Proposition 42.2, the global minimum of $G(\lambda)$ is obtained when

$$B^\top AB\lambda - B^\top Ab + f = 0,$$

that is, $\lambda = (B^\top AB)^{-1}(B^\top Ab - f)$, as we found earlier.

Remarks:

(1) There is a form of duality going on in this situation. The constrained minimization of $Q(x)$ subject to $B^\top x = f$ is called the *primal problem*, and the unconstrained maximization of $-G(\lambda)$ is called the *dual problem*. Duality is the fact stated slightly loosely as

$$\min_x Q(x) = \max_\lambda -G(\lambda).$$

A general treatment of duality in constrained minimization problems is given in Section 50.7.

Recalling that $e = b - B\lambda$, since

$$G(\lambda) = \frac{1}{2}(B\lambda - b)^\top A(B\lambda - b) + \lambda^\top f,$$

we can also write

$$G(\lambda) = \frac{1}{2}e^\top Ae + \lambda^\top f.$$

This expression often represents the total potential energy of a system. Again, the optimal solution is the one that minimizes the potential energy (and thus maximizes $-G(\lambda)$).

(2) It is immediately verified that the equations of Proposition 42.3 are equivalent to the equations stating that the partial derivatives of the Lagrangian $L(x, \lambda)$ are null:

$$\begin{aligned} \frac{\partial L}{\partial x_i} &= 0, \quad i = 1, \ldots, m, \\ \frac{\partial L}{\partial \lambda_j} &= 0, \quad j = 1, \ldots, n. \end{aligned}$$

Thus, the constrained minimum of $Q(x)$ subject to $B^\top x = f$ is an extremum of the Lagrangian $L(x, \lambda)$. As we showed in Proposition 42.3, this extremum corresponds to simultaneously minimizing $L(x, \lambda)$ with respect to x and maximizing $L(x, \lambda)$ with respect to λ. Geometrically, such a point is a *saddle point* for $L(x, \lambda)$. Saddle points are discussed in Section 50.7.

(3) The Lagrange multipliers sometimes have a natural physical meaning. For example, in the spring-mass system they correspond to node displacements. In some general sense, Lagrange multipliers are correction terms needed to satisfy equilibrium equations and the price paid for the constraints. For more details, see Strang [167].

Going back to the constrained minimization of $Q(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2)$ subject to

$$2x_1 - x_2 = 5,$$

the Lagrangian is

$$L(x_1, x_2, \lambda) = \frac{1}{2}\big(x_1^2 + x_2^2\big) + \lambda(2x_1 - x_2 - 5),$$

and the equations stating that the Lagrangian has a saddle point are

$$\begin{aligned} x_1 + 2\lambda &= 0, \\ x_2 - \lambda &= 0, \\ 2x_1 - x_2 - 5 &= 0. \end{aligned}$$

We obtain the solution $(x_1, x_2, \lambda) = (2, -1, -1)$.

The use of Lagrange multipliers in optimization and variational problems is discussed extensively in Chapter 50.

Least squares methods and Lagrange multipliers are used to tackle many problems in computer graphics and computer vision; see Trucco and Verri [176], Metaxas [123], Jain, Katsuri, and Schunck [98], Faugeras [60], and Foley, van Dam, Feiner, and Hughes [64].

42.2 Quadratic Optimization: The General Case

In this section we complete the study initiated in Section 42.1 and give necessary and sufficient conditions for the quadratic function $\frac{1}{2}x^\top Ax - x^\top b$ to have a global minimum. We begin with the following simple fact:

Proposition 42.4. *If A is an invertible symmetric matrix, then the function*

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b$$

has a minimum value iff $A \succeq 0$, in which case this optimal value is obtained for a unique value of x, namely $x^* = A^{-1}b$, and with

$$f(A^{-1}b) = -\frac{1}{2}b^\top A^{-1}b.$$

Proof. Observe that

$$\frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) = \frac{1}{2}x^\top Ax - x^\top b + \frac{1}{2}b^\top A^{-1}b.$$

Thus,

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b = \frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) - \frac{1}{2}b^\top A^{-1}b.$$

If A has some negative eigenvalue, say $-\lambda$ (with $\lambda > 0$), if we pick any eigenvector u of A associated with λ, then for any $\alpha \in \mathbb{R}$ with $\alpha \neq 0$, if we let $x = \alpha u + A^{-1}b$, then since $Au = -\lambda u$, we get

$$\begin{aligned} f(x) &= \frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) - \frac{1}{2}b^\top A^{-1}b \\ &= \frac{1}{2}\alpha u^\top A \alpha u - \frac{1}{2}b^\top A^{-1}b \\ &= -\frac{1}{2}\alpha^2 \lambda \left\|u\right\|_2^2 - \frac{1}{2}b^\top A^{-1}b, \end{aligned}$$

and since α can be made as large as we want and $\lambda > 0$, we see that f has no minimum. Consequently, in order for f to have a minimum, we must have $A \succeq 0$. If $A \succeq 0$, since A is invertible, it is positive definite, so $(x - A^{-1}b)^\top A(x - A^{-1}b) > 0$ iff $x - A^{-1}b \neq 0$, and it is clear that the minimum value of f is achieved when $x - A^{-1}b = 0$, that is, $x = A^{-1}b$. □

Let us now consider the case of an arbitrary symmetric matrix A.

Proposition 42.5. *If A is a $n \times n$ symmetric matrix, then the function*

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b$$

has a minimum value iff $A \succeq 0$ and $(I - AA^+)b = 0$, in which case this minimum value is

$$p^* = -\frac{1}{2}b^\top A^+ b.$$

Furthermore, if A is diagonlized as $A = U^\top \Sigma U$ (with U orthogonal), then the optimal value is achieved by all $x \in \mathbb{R}^n$ of the form

$$x = A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix},$$

for any $z \in \mathbb{R}^{n-r}$, where r is the rank of A.

Proof. The case that A is invertible is taken care of by Proposition 42.4, so we may assume that A is singular. If A has rank $r < n$, then we can diagonalize A as

$$A = U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U,$$

where U is an orthogonal matrix and where Σ_r is an $r \times r$ diagonal invertible matrix. Then we have

$$\begin{aligned} f(x) &= \frac{1}{2}x^\top U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} Ux - x^\top U^\top Ub \\ &= \frac{1}{2}(Ux)^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} Ux - (Ux)^\top Ub. \end{aligned}$$

If we write

$$Ux = \begin{pmatrix} y \\ z \end{pmatrix} \quad \text{and} \quad Ub = \begin{pmatrix} c \\ d \end{pmatrix},$$

with $y, c \in \mathbb{R}^r$ and $z, d \in \mathbb{R}^{n-r}$, we get

$$\begin{aligned} f(x) &= \frac{1}{2}(Ux)^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} Ux - (Ux)^\top Ub \\ &= \frac{1}{2}(y^\top \; z^\top) \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} - (y^\top \; z^\top) \begin{pmatrix} c \\ d \end{pmatrix} \\ &= \frac{1}{2} y^\top \Sigma_r y - y^\top c - z^\top d. \end{aligned}$$

For $y = 0$, we get

$$f(x) = -z^\top d,$$

so if $d \neq 0$, the function f has no minimum. Therefore, if f has a minimum, then $d = 0$. However, $d = 0$ means that

$$Ub = \begin{pmatrix} c \\ 0 \end{pmatrix},$$

and we know from Proposition 23.5 (Vol. I) that b is in the range of A (here, U is $V^\top$), which is equivalent to $(I - AA^+)b = 0$. If $d = 0$, then

$$f(x) = \frac{1}{2} y^\top \Sigma_r y - y^\top c,$$

and since Σ_r is invertible, by Proposition 42.4, the function f has a minimum iff $\Sigma_r \succeq 0$, which is equivalent to $A \succeq 0$.

Therefore, we have proved that if f has a minimum, then $(I - AA^+)b = 0$ and $A \succeq 0$. Conversely, if $(I - AA^+)b = 0$ and $A \succeq 0$, what we just did proves that f does have a minimum.

When the above conditions hold, since

$$A = U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U$$

is positive semidefinite, the pseudo-inverse A^+ of A is given by

$$A^+ = U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U,$$

and by Proposition 42.4 the minimum is achieved if $y = \Sigma_r^{-1} c$, $z = 0$ and $d = 0$, that is, for x^* given by

$$Ux^* = \begin{pmatrix} \Sigma_r^{-1} c \\ 0 \end{pmatrix} \quad \text{and} \quad Ub = \begin{pmatrix} c \\ 0 \end{pmatrix},$$

from which we deduce that

$$x^* = U^\top \begin{pmatrix} \Sigma_r^{-1} c \\ 0 \end{pmatrix} = U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} c \\ 0 \end{pmatrix} = U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} Ub = A^+ b$$

and the minimum value of f is

$$f(x^*) = -\frac{1}{2} b^\top A^+ b.$$

For any $x \in \mathbb{R}^n$ of the form

$$x = A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix},$$

for any $z \in \mathbb{R}^{n-r}$, we have

$$\begin{aligned} f(x) &= \frac{1}{2}\left(A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix}\right)^\top A \left(A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix}\right) - \left(A^+ b + U^\top \begin{pmatrix} 0 \\ z \end{pmatrix}\right)^\top b \\ &= \frac{1}{2}(A^+ b)^\top A A^+ b + (0\; z^\top) U A A^+ b + \frac{1}{2}(0\; z^\top) U A U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} - (A^+ b)^\top b - (0\; z^\top) U b \\ &= -\frac{1}{2} b^\top A^+ b + (0\; z^\top) U A A^+ b + \frac{1}{2}(0\; z^\top) U A U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} - (0\; z^\top) U b. \end{aligned}$$

We have

$$\begin{aligned} (0\; z^\top) U A A^+ b &= (0\; z^\top) U U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} Ub \\ &= (0\; z^\top) \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} Ub = 0, \end{aligned}$$

$$\begin{aligned} (0\; z^\top) U A U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} &= (0\; z^\top) U U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U U^\top \begin{pmatrix} 0 \\ z \end{pmatrix} \\ &= (0\; z^\top) \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ z \end{pmatrix} = 0, \end{aligned}$$

and

$$(0\; z^\top) Ub = (0\; z^\top) \begin{pmatrix} c \\ 0 \end{pmatrix} = 0,$$

because $(I - AA^+)b = 0$, that is,

$$\begin{aligned} \left(\begin{pmatrix} I_r & 0 \\ 0 & I_{n-r} \end{pmatrix} - U^\top \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} U U^\top \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U\right) b &= \left(\begin{pmatrix} I_r & 0 \\ 0 & I_{n-r} \end{pmatrix} - U^\top \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} U\right) b \\ &= U^\top \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix} Ub = 0, \end{aligned}$$

so if

$$Ub = \begin{pmatrix} c \\ d \end{pmatrix},$$

then $d = 0$. Therefore, $f(x) = -\frac{1}{2} b^\top A^+ b$. □

The problem of minimizing the function

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b$$

in the case where we add either linear constraints of the form $C^\top x = 0$ or affine constraints of the form $C^\top x = t$ (where $t \in \mathbb{R}^m$ and $t \neq 0$) where C is an $n \times m$ matrix can be reduced to the unconstrained case using a QR-decomposition of C. Let us show how to do this for linear constraints of the form $C^\top x = 0$.

If we use a QR decomposition of C, by permuting the columns of C to make sure that the first r columns of C are linearly independent (where $r = \mathrm{rank}(C)$), we may assume that

$$C = Q^\top \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix} \Pi,$$

where Q is an $n \times n$ orthogonal matrix, R is an $r \times r$ invertible upper triangular matrix, S is an $r \times (m - r)$ matrix, and Π is a permutation matrix (C has rank r). Then if we let

$$x = Q^\top \begin{pmatrix} y \\ z \end{pmatrix},$$

where $y \in \mathbb{R}^r$ and $z \in \mathbb{R}^{n-r}$, then $C^\top x = 0$ becomes

$$C^\top x = \Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} Qx = \Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = 0,$$

which implies $y = 0$, and every solution of $C^\top x = 0$ is of the form

$$x = Q^\top \begin{pmatrix} 0 \\ z \end{pmatrix}.$$

Our original problem becomes

$$\begin{aligned} &\text{minimize} && \frac{1}{2}(y^\top \; z^\top) QAQ^\top \begin{pmatrix} y \\ z \end{pmatrix} + (y^\top \; z^\top)Qb \\ &\text{subject to} && y = 0, \; y \in \mathbb{R}^r, \; z \in \mathbb{R}^{n-r}. \end{aligned}$$

Thus, the constraint $C^\top x = 0$ has been simplifed to $y = 0$, and if we write

$$QAQ^\top = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix},$$

where G_{11} is an $r \times r$ matrix and G_{22} is an $(n - r) \times (n - r)$ matrix, and

$$Qb = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad b_1 \in \mathbb{R}^r, \; b_2 \in \mathbb{R}^{n-r},$$

our problem becomes

$$\text{minimize } \frac{1}{2} z^\top G_{22} z + z^\top b_2, \quad z \in \mathbb{R}^{n-r},$$

the problem solved in Proposition 42.5.

Constraints of the form $C^\top x = t$ (where $t \neq 0$) can be handled in a similar fashion. In this case, we may assume that C is an $n \times m$ matrix with full rank (so that $m \leq n$) and $t \in \mathbb{R}^m$. Then we use a QR-decomposition of the form

$$C = P \begin{pmatrix} R \\ 0 \end{pmatrix},$$

where P is an orthogonal $n \times n$ matrix and R is an $m \times m$ invertible upper triangular matrix. If we write

$$x = P \begin{pmatrix} y \\ z \end{pmatrix},$$

where $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^{n-m}$, the equation $C^\top x = t$ becomes

$$(R^\top \; 0) P^\top x = t,$$

that is,

$$(R^\top \; 0) \begin{pmatrix} y \\ z \end{pmatrix} = t,$$

which yields

$$R^\top y = t.$$

Since R is invertible, we get $y = (R^\top)^{-1} t$, and then it is easy to see that our original problem reduces to an unconstrained problem in terms of the matrix $P^\top A P$; the details are left as an exercise.

42.3 Maximizing a Quadratic Function on the Unit Sphere

In this section we discuss various quadratic optimization problems mostly arising from computer vision (image segmentation and contour grouping). These problems can be reduced to the following basic optimization problem: Given an $n \times n$ real symmetric matrix A

$$\begin{aligned} &\text{maximize} && x^\top A x \\ &\text{subject to} && x^\top x = 1, \; x \in \mathbb{R}^n. \end{aligned}$$

In view of Proposition 23.10 (Vol. I), the maximum value of $x^\top A x$ on the unit sphere is equal to the largest eigenvalue λ_1 of the matrix A, and it is achieved for any unit eigenvector u_1 associated with λ_1.

A variant of the above problem often encountered in computer vision consists in minimizing $x^\top Ax$ on the ellipsoid given by an equation of the form

$$x^\top Bx = 1,$$

where B is a symmetric positive definite matrix. Since B is positive definite, it can be diagonalized as

$$B = QDQ^\top,$$

where Q is an orthogonal matrix and D is a diagonal matrix,

$$D = \mathrm{diag}(d_1, \ldots, d_n),$$

with $d_i > 0$, for $i = 1, \ldots, n$. If we define the matrices $B^{1/2}$ and $B^{-1/2}$ by

$$B^{1/2} = Q\,\mathrm{diag}\left(\sqrt{d_1}, \ldots, \sqrt{d_n}\right) Q^\top$$

and

$$B^{-1/2} = Q\,\mathrm{diag}\left(1/\sqrt{d_1}, \ldots, 1/\sqrt{d_n}\right) Q^\top,$$

it is clear that these matrices are symmetric, that $B^{-1/2}BB^{-1/2} = I$, and that $B^{1/2}$ and $B^{-1/2}$ are mutual inverses. Then, if we make the change of variable

$$x = B^{-1/2}y,$$

the equation $x^\top Bx = 1$ becomes $y^\top y = 1$, and the optimization problem

$$\begin{array}{ll} \text{maximize} & x^\top Ax \\ \text{subject to} & x^\top Bx = 1,\ x \in \mathbb{R}^n, \end{array}$$

is equivalent to the problem

$$\begin{array}{ll} \text{maximize} & y^\top B^{-1/2}AB^{-1/2}y \\ \text{subject to} & y^\top y = 1,\ y \in \mathbb{R}^n, \end{array}$$

where $y = B^{1/2}x$ and where $B^{-1/2}AB^{-1/2}$ is symmetric.

The complex version of our basic optimization problem in which A is a Hermitian matrix also arises in computer vision. Namely, given an $n \times n$ complex Hermitian matrix A,

$$\begin{array}{ll} \text{maximize} & x^*Ax \\ \text{subject to} & x^*x = 1,\ x \in \mathbb{C}^n. \end{array}$$

Again by Proposition 23.10 (Vol. I), the maximum value of x^*Ax on the unit sphere is equal to the largest eigenvalue λ_1 of the matrix A and it is achieved for any unit eigenvector u_1 associated with λ_1.

Remark: It is worth pointing out that if A is a *skew-Hermitian* matrix, that is, if $A^* = -A$, then x^*Ax is *pure imaginary or zero*.

Indeed, since $z = x^*Ax$ is a scalar, we have $z^* = \overline{z}$ (the conjugate of z), so we have

$$\overline{x^*Ax} = (x^*Ax)^* = x^*A^*x = -x^*Ax,$$

so $\overline{x^*Ax} + x^*Ax = 2\mathrm{Re}(x^*Ax) = 0$, which means that x^*Ax is pure imaginary or zero.

In particular, if A is a real matrix and if A is *skew-symmetric*, then

$$x^\top Ax = 0.$$

Thus, for any real matrix (symmetric or not),

$$x^\top Ax = x^\top H(A)x,$$

where $H(A) = (A + A^\top)/2$, the symmetric part of A.

There are situations in which it is necessary to add linear constraints to the problem of maximizing a quadratic function on the sphere. This problem was completely solved by Golub [78] (1973). The problem is the following: Given an $n \times n$ real symmetric matrix A and an $n \times p$ matrix C,

$$\begin{aligned} &\text{minimize} && x^\top Ax \\ &\text{subject to} && x^\top x = 1,\ C^\top x = 0,\ x \in \mathbb{R}^n. \end{aligned}$$

As in Section 42.2, Golub shows that the linear constraint $C^\top x = 0$ can be eliminated as follows: If we use a QR decomposition of C, by permuting the columns, we may assume that

$$C = Q^\top \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix} \Pi,$$

where Q is an orthogonal $n \times n$ matrix, R is an $r \times r$ invertible upper triangular matrix, and S is an $r \times (p - r)$ matrix (assuming C has rank r). Then if we let

$$x = Q^\top \begin{pmatrix} y \\ z \end{pmatrix},$$

where $y \in \mathbb{R}^r$ and $z \in \mathbb{R}^{n-r}$, then $C^\top x = 0$ becomes

$$\Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} Qx = \Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = 0,$$

which implies $y = 0$, and every solution of $C^\top x = 0$ is of the form

$$x = Q^\top \begin{pmatrix} 0 \\ z \end{pmatrix}.$$

Our original problem becomes

$$\begin{aligned} \text{minimize} \quad & (y^\top\, z^\top) QAQ^\top \begin{pmatrix} y \\ z \end{pmatrix} \\ \text{subject to} \quad & z^\top z = 1,\ z \in \mathbb{R}^{n-r}, \\ & y = 0,\ y \in \mathbb{R}^r. \end{aligned}$$

Thus, the constraint $C^\top x = 0$ has been simplified to $y = 0$, and if we write

$$QAQ^\top = \begin{pmatrix} G_{11} & G_{12} \\ G_{12}^\top & G_{22} \end{pmatrix},$$

our problem becomes

$$\begin{aligned} \text{minimize} \quad & z^\top G_{22} z \\ \text{subject to} \quad & z^\top z = 1,\ z \in \mathbb{R}^{n-r}, \end{aligned}$$

a standard eigenvalue problem.

Remark: There is a way of finding the eigenvalues of G_{22} which does not require the QR-factorization of C. Observe that if we let

$$J = \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix},$$

then

$$JQAQ^\top J = \begin{pmatrix} 0 & 0 \\ 0 & G_{22} \end{pmatrix},$$

and if we set

$$P = Q^\top JQ,$$

then

$$PAP = Q^\top JQAQ^\top JQ.$$

Now, $Q^\top JQAQ^\top JQ$ and $JQAQ^\top J$ have the same eigenvalues, so PAP and $JQAQ^\top J$ also have the same eigenvalues. It follows that the solutions of our optimization problem are among the eigenvalues of $K = PAP$, and at least r of those are 0. Using the fact that CC^+ is the projection onto the range of C, where C^+ is the pseudo-inverse of C, it can also be shown that

$$P = I - CC^+,$$

the projection onto the kernel of $C^\top$. So P can be computed directly in terms of C. In particular, when $n \geq p$ and C has full rank (the columns of C are linearly independent), then we know that $C^+ = (C^\top C)^{-1} C^\top$ and

$$P = I - C(C^\top C)^{-1} C^\top.$$

This fact is used by Cour and Shi [42] and implicitly by Yu and Shi [190].

The problem of adding affine constraints of the form $N^\top x = t$, where $t \neq 0$, also comes up in practice. At first glance, this problem may not seem harder than the linear problem in which $t = 0$, but it is. This problem was extensively studied in a paper by Gander, Golub, and von Matt [75] (1989).

Gander, Golub, and von Matt consider the following problem: Given an $(n+m)\times(n+m)$ real symmetric matrix A (with $n > 0$), an $(n+m)\times m$ matrix N with full rank, and a nonzero vector $t \in \mathbb{R}^m$ with $\left\|(N^\top)^+ t\right\| < 1$ (where $(N^\top)^+$ denotes the pseudo-inverse of $N^\top$),

$$\begin{array}{ll} \text{minimize} & x^\top A x \\ \text{subject to} & x^\top x = 1,\ N^\top x = t,\ x \in \mathbb{R}^{n+m}. \end{array}$$

The condition $\left\|(N^\top)^+ t\right\| < 1$ ensures that the problem has a solution and is not trivial. The authors begin by proving that the affine constraint $N^\top x = t$ can be eliminated. One way to do so is to use a QR decomposition of N. If

$$N = P\begin{pmatrix} R \\ 0 \end{pmatrix},$$

where P is an orthogonal $(n+m)\times(n+m)$ matrix and R is an $m \times m$ invertible upper triangular matrix, then if we observe that

$$\begin{aligned} x^\top A x &= x^\top P P^\top A P P^\top x, \\ N^\top x &= (R^\top\ 0)P^\top x = t, \\ x^\top x &= x^\top P P^\top x = 1, \end{aligned}$$

and if we write

$$P^\top A P = \begin{pmatrix} B & \Gamma^\top \\ \Gamma & C \end{pmatrix},$$

where B is an $m \times m$ symmetric matrix, C is an $n \times n$ symmetric matrix, Γ is an $m \times n$ matrix, and

$$P^\top x = \begin{pmatrix} y \\ z \end{pmatrix},$$

with $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^n$, then we get

$$\begin{aligned} x^\top A x &= y^\top B y + 2 z^\top \Gamma y + z^\top C z, \\ R^\top y &= t, \\ y^\top y + z^\top z &= 1. \end{aligned}$$

Thus

$$y = (R^\top)^{-1} t,$$

and if we write

$$s^2 = 1 - y^\top y > 0$$

and

$$b = \Gamma y,$$

we get the simplified problem

$$\begin{array}{ll} \text{minimize} & z^\top C z + 2z^\top b \\ \text{subject to} & z^\top z = s^2, \; z \in \mathbb{R}^m. \end{array}$$

Unfortunately, if $b \neq 0$, Proposition 23.10 (Vol. I) is no longer applicable. It is still possible to find the minimum of the function $z^\top C z + 2z^\top b$ using Lagrange multipliers, but such a solution is too involved to be presented here. Interested readers will find a thorough discussion in Gander, Golub, and von Matt [75].

42.4 Summary

The main concepts and results of this chapter are listed below:

- Quadratic optimization problems; *quadratic functions*.
- Symmetric *positive definite* and *positive semidefinite* matrices.
- The *positive semidefinite cone ordering*.
- Existence of a global minimum when A is symmetric positive definite.
- Constrained quadratic optimization problems.
- *Lagrange multipliers*; *Lagrangian*.
- *Primal* and *dual* problems.
- Quadratic optimization problems: the case of a symmetric invertible matrix A.
- Quadratic optimization problems: the general case of a symmetric matrix A.
- Adding linear constraints of the form $C^\top x = 0$.
- Adding affine constraints of the form $C^\top x = t$, with $t \neq 0$.
- Maximizing a quadratic function over the unit sphere.
- Maximizing a quadratic function over an ellipsoid.
- Maximizing a Hermitian quadratic form.
- Adding linear constraints of the form $C^\top x = 0$.
- Adding affine constraints of the form $N^\top x = t$, with $t \neq 0$.

Chapter 43

Schur Complements and Applications

43.1 Schur Complements

Schur complements arise naturally in the process of inverting block matrices of the form

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

and in characterizing when symmetric versions of these matrices are positive definite or positive semidefinite. These characterizations come up in various quadratic optimization problems; see Boyd and Vandenberghe [29], especially Appendix B. In the most general case, pseudo-inverses are also needed.

In this chapter we introduce Schur complements and describe several interesting ways in which they are used. Along the way we provide some details and proofs of some results from Appendix A.5 (especially Section A.5.5) of Boyd and Vandenberghe [29].

Let M be an $n \times n$ matrix written as a 2×2 block matrix

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where A is a $p \times p$ matrix and D is a $q \times q$ matrix, with $n = p + q$ (so B is a $p \times q$ matrix and C is a $q \times p$ matrix). We can try to solve the linear system

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c \\ d \end{pmatrix},$$

that is,

$$\begin{aligned} Ax + By &= c, \\ Cx + Dy &= d, \end{aligned}$$

by mimicking Gaussian elimination. If we assume that D is invertible, then we first solve for y, getting

$$y = D^{-1}(d - Cx),$$

and after substituting this expression for y in the first equation, we get

$$Ax + B(D^{-1}(d - Cx)) = c,$$

that is,

$$(A - BD^{-1}C)x = c - BD^{-1}d.$$

If the matrix $A - BD^{-1}C$ is invertible, then we obtain the solution to our system

$$\begin{aligned} x &= (A - BD^{-1}C)^{-1}(c - BD^{-1}d), \\ y &= D^{-1}(d - C(A - BD^{-1}C)^{-1}(c - BD^{-1}d)). \end{aligned}$$

If A is invertible, then by eliminating x first using the first equation, we obtain analogous formulas involving the matrix $D - CA^{-1}B$. The above formulas suggest that the matrices $A - BD^{-1}C$ and $D - CA^{-1}B$ play a special role and suggest the following definition:

Definition 43.1. Given any $n \times n$ block matrix of the form

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where A is a $p \times p$ matrix and D is a $q \times q$ matrix, with $n = p + q$ (so B is a $p \times q$ matrix and C is a $q \times p$ matrix), if D is invertible, then the matrix $A - BD^{-1}C$ is called the *Schur complement* of D in M. If A is invertible, then the matrix $D - CA^{-1}B$ is called the *Schur complement* of A in M.

The above equations written as

$$\begin{aligned} x &= (A - BD^{-1}C)^{-1}c - (A - BD^{-1}C)^{-1}BD^{-1}d, \\ y &= -D^{-1}C(A - BD^{-1}C)^{-1}c \\ &\quad + (D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1})d, \end{aligned}$$

yield a formula for the inverse of M in terms of the Schur complement of D in M, namely

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}.$$

A moment of reflection reveals that

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & 0 \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} \end{pmatrix} \begin{pmatrix} I & -BD^{-1} \\ 0 & I \end{pmatrix},$$

and then

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ -D^{-1}C & I \end{pmatrix} \begin{pmatrix} (A - BD^{-1}C)^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix} \begin{pmatrix} I & -BD^{-1} \\ 0 & I \end{pmatrix}.$$

By taking inverses, we obtain the following result.

Proposition 43.1. *If the matrix D is invertibke, then*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & 0 \\ D^{-1}C & I \end{pmatrix}.$$

The above expression can be checked directly and has the advantage of requiring only the invertibility of D.

Remark: If A is invertible, then we can use the Schur complement $D - CA^{-1}B$ of A to obtain the following factorization of M:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix}.$$

If $D - CA^{-1}B$ is invertible, we can invert all three matrices above, and we get another formula for the inverse of M in terms of $(D - CA^{-1}B)$, namely,

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}.$$

If A, D and both Schur complements $A - BD^{-1}C$ and $D - CA^{-1}B$ are all invertible, by comparing the two expressions for M^{-1}, we get the (nonobvious) formula

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}.$$

Using this formula, we obtain another expression for the inverse of M involving the Schur complements of A and D (see Horn and Johnson [93]):

Proposition 43.2. *If A, D and both Schur complements $A - BD^{-1}C$ and $D - CA^{-1}B$ are all invertible, then*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}.$$

If we set $D = I$ and change B to $-B$, we get

$$(A + BC)^{-1} = A^{-1} - A^{-1}B(I - CA^{-1}B)^{-1}CA^{-1},$$

a formula known as the *matrix inversion lemma* (see Boyd and Vandenberghe [29], Appendix C.4, especially C.4.3).

43.2 Symmetric Positive Definite Matrices and Schur Complements

If we assume that our block matrix M is symmetric, so that A, D are symmetric and $C = B^\top$, then we see that M is expressed as

$$M = \begin{pmatrix} A & B \\ B^\top & D \end{pmatrix} = \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BD^{-1}B^\top & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix}^\top,$$

which shows that M is similar to a block diagonal matrix (obviously, the Schur complement, $A - BD^{-1}B^\top$, is symmetric). As a consequence, we have the following version of "Schur's trick" to check whether $M \succ 0$ for a symmetric matrix.

Proposition 43.3. *For any symmetric matrix M of the form*

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix},$$

if C is invertible, then the following properties hold:

(1) $M \succ 0$ iff $C \succ 0$ and $A - BC^{-1}B^\top \succ 0$.

(2) If $C \succ 0$, then $M \succeq 0$ iff $A - BC^{-1}B^\top \succeq 0$.

Proof. (1) Observe that

$$\begin{pmatrix} I & BC^{-1} \\ 0 & I \end{pmatrix}^{-1} = \begin{pmatrix} I & -BC^{-1} \\ 0 & I \end{pmatrix},$$

and we know that for any symmetric matrix T and any invertible matrix N, the matrix T is positive definite ($T \succ 0$) iff $NTN^\top$ (which is obviously symmetric) is positive definite ($NTN^\top \succ 0$). But a block diagonal matrix is positive definite iff each diagonal block is positive definite, which concludes the proof.

(2) This is because for any symmetric matrix T and any invertible matrix N, we have $T \succeq 0$ iff $NTN^\top \succeq 0$. □

Another version of Proposition 43.3 using the Schur complement of A instead of the Schur complement of C also holds. The proof uses the factorization of M using the Schur complement of A (see Section 43.1).

Proposition 43.4. *For any symmetric matrix M of the form*

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix},$$

if A is invertible then the following properties hold:

(1) $M \succ 0$ *iff* $A \succ 0$ *and* $C - B^\top A^{-1}B \succ 0$.

(2) If $A \succ 0$, *then* $M \succeq 0$ *iff* $C - B^\top A^{-1}B \succeq 0$.

Here is an illustration of Proposition 43.4(2). Consider the nonlinear quadratic constraint

$$(Ax+b)^\top(Ax+b) \leq c^\top x + d,$$

were $A \in \mathrm{M}_n(\mathbb{R}), x, b, c \in \mathbb{R}^n$ and $d \in \mathbb{R}$. Since obviously $I = I_n$ is invertible and $I \succ 0$, we have

$$\begin{pmatrix} I & Ax+b \\ (Ax+b)^\top & c^\top x + d \end{pmatrix} \succeq 0$$

iff $c^\top x + d - (Ax+b)^\top(Ax+b) \succeq 0$ iff $(Ax+b)^\top(Ax+b) \leq c^\top x + d$, since the matrix (a scalar) $c^\top x + d - (Ax+b)^\top(Ax+b)$ is the Schur complement of I in the above matrix.

The trick of using Schur complements to convert nonlinear inequality constraints into linear constraints on symmetric matrices involving the semidefinire ordering $\succeq$ is used extensively to convert nonlinear problems into semidefinite programs; see Boyd and Vandenberghe [29].

When C is singular (or A is singular), it is still possible to characterize when a symmetric matrix M as above is positive semidefinite, but this requires using a version of the Schur complement involving the pseudo-inverse of C, namely $A - BC^+B^\top$ (or the Schur complement, $C - B^\top A^+ B$, of A). We use the criterion of Proposition 42.5, which tells us when a quadratic function of the form $\frac{1}{2}x^\top Px - x^\top b$ has a minimum and what this optimum value is (where P is a symmetric matrix).

43.3 Symmetric Positive Semidefinite Matrices and Schur Complements

We now return to our original problem, characterizing when a symmetric matrix

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}$$

is positive semidefinite. Thus, we want to know when the function

$$f(x,y) = (x^\top, y^\top)\begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = x^\top Ax + 2x^\top By + y^\top Cy$$

has a minimum with respect to both x and y. If we hold y constant, Proposition 42.5 implies that $f(x,y)$ has a minimum iff $A \succeq 0$ and $(I - AA^+)By = 0$, and then the minimum value is

$$f(x^*, y) = -y^\top B^\top A^+ By + y^\top Cy = y^\top(C - B^\top A^+ B)y.$$

Since we want $f(x, y)$ to be uniformly bounded from below for all x, y, we must have $(I - AA^+)B = 0$. Now, $f(x^*, y)$ has a minimum iff $C - B^\top A^+ B \succeq 0$. Therefore, we have established that $f(x, y)$ has a minimum over all x, y iff

$$A \succeq 0, \quad (I - AA^+)B = 0, \quad C - B^\top A^+ B \succeq 0.$$

Similar reasoning applies if we first minimize with respect to y and then with respect to x, but this time, the Schur complement $A - BC^+ B^\top$ of C is involved. Putting all these facts together, we get our main result:

Theorem 43.5. *Given any symmetric matrix*

$$M = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}$$

the following conditions are equivalent:

(1) $M \succeq 0$ *(M is positive semidefinite).*

(2) $A \succeq 0, \quad (I - AA^+)B = 0, \quad C - B^\top A^+ B \succeq 0.$

(3) $C \succeq 0, \quad (I - CC^+)B^\top = 0, \quad A - BC^+ B^\top \succeq 0.$

If $M \succeq 0$ as in Theorem 43.5, then it is easy to check that we have the following factorizations (using the fact that $A^+ A A^+ = A^+$ and $C^+ C C^+ = C^+$):

$$\begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} = \begin{pmatrix} I & BC^+ \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BC^+ B^\top & 0 \\ 0 & C \end{pmatrix} \begin{pmatrix} I & 0 \\ C^+ B^\top & I \end{pmatrix}$$

and

$$\begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} = \begin{pmatrix} I & 0 \\ B^\top A^+ & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & C - B^\top A^+ B \end{pmatrix} \begin{pmatrix} I & A^+ B \\ 0 & I \end{pmatrix}.$$