In optimization theory, the optimality conditions for interior points are usually much simpler than the optimality conditions for boundary points. In this chapter, we deal with the former, easier case. Boundary points appear more prominently in constrained optimization, when one tries to optimize a function, subject to several functional constraints. For this reason, the optimality conditions for boundary points are generally discussed in constrained optimization, whereas the optimality conditions for interior points are discussed in unconstrained optimization, regardless of whether the optimization problem at hand has constraints.

In this chapter, we first establish some basic results on the existence of global minimizer or maximizers of continuous functions on a metric space. These are the famous Weierstrass theorem and its variants, which are essentially the only general tools available for establishing the existence of optimizers.

The rest of the chapter is devoted to obtaining the fundamental first-order and second-order necessary and sufficient optimality conditions for minimizing or maximizing differentiable functions. Since the tools here are based on differentiation, and differentiation is a local theory, the optimality conditions generally apply to local optimizers. The necessary and sufficient conditions play different, usually complementary, roles. A typical necessary condition for a local minimizer, say, states that certain conditions, usually given as equalities or inequalities, must be satisfied at a local minimizer. A typical sufficient condition for a local minimizer, however, states that if certain conditions are satisfied at a given point, then that point must be a local minimizer.

The nature (local minimum, local maximum, or saddle point) of a critical point $x$ of a twice differentiable function $f$ is deduced from the definiteness properties of the quadratic form $q(d) = \langle D^2 f(x) d, d \rangle$ involving the Hessian matrix $D^2 f(x)$. Thus, there is a need for an efficient recognition of a symmetric matrix. Several tools are developed in Section 2.4 for this purpose. A novel feature of this section is that we give an exposition of a simple tool, Descartes’s
rule of sign, that can be used to count exactly the number of positive and negative eigenvalues of a symmetric matrix, including $D^2 f(x)$.

The inverse function theorem and the closely related implicit function theorem are important tools in many branches of analysis. Another closely related result, Lyusternik’s theorem [191], is an important tool in optimization, where it is used in the derivation of optimality conditions in constrained optimization. We give an elementary proof of the implicit function theorem in finite-dimensional vector spaces in Section 2.5, following Carathéodory [54], and use it to prove the inverse function theorem and Lyusternik’s theorem in finite dimensions. The proof of the same theorems in Banach spaces is given in Chapter 3 using Ekeland’s $\varepsilon$-variational principle. If one is interested only in finite-dimensional versions of these results, it suffices to read only Section 2.5.

The local behavior of a $C^2$ function $f$ around a nondegenerate critical point $x$ ($D^2 f(x)$ is nonsingular) is determined by the Hessian matrix $D^2 f(x)$. This is the content of Morse’s lemma, which is treated in Section 2.6. Morse’s lemma is a basic result in Morse theory, which investigates the relationships between various types of critical points of a function $f$; see, for example, Milnor [197] for an introduction to Morse theory.

2.1 Basic Results on the Existence of Optimizers

We start by defining various types of optimal points.

**Definition 2.1.** Let $f : U \to \mathbb{R}$ be a function on a set $U \subseteq \mathbb{R}^n$. Let $x^* \in U$ be an arbitrary point, and let $B_r(x^*) := \{x \in U : \|x - x^*\| < r\}$ be the open ball of radius $r$ around $x^*$. The point $x^*$ is called

(a) a **local minimizer of $f$** if

$$f(x^*) \leq f(x) \quad \text{for all } x \text{ in some ball } B_r(x^*),$$

and a **strict local minimizer of $f$** if

$$f(x^*) < f(x) \quad \text{for all } x \in B_r(x^*), \ x \neq x^*;$$

(b) a **global minimizer of $f$ on $U$** if

$$f(x^*) \leq f(x) \quad \text{for all } x \in U,$$

and a **strict global minimizer of $f$ on $U$** if

$$f(x^*) < f(x) \quad \text{for all } x \in U, \ x \neq x^*;$$

(c) a **critical point of $f$** if $f$ is Gâteaux differentiable at $x^*$ and $\nabla f(x^*) = 0$;

(d) a **saddle point of $f$** if it is a critical point and there exist points $y, z$ in any ball $B_r(x^*)$ such that $f(y) < f(x^*) < f(z)$.
Parallel definitions apply for a maximizer of \( f \). We call a point \( x^\ast \) an optimizer of \( f \) if \( x^\ast \) is a minimizer or a maximizer in any of the senses above.

The most basic result on the existence of optimizers is the following theorem, due to Weierstrass.

**Theorem 2.2. (Weierstrass)** Let \( f : K \to \mathbb{R} \) be a continuous function defined on a compact metric space \( K \). Then there exists a global minimizer \( x^\ast \in K \) of \( f \) on \( K \), that is,

\[
f(x^\ast) \leq f(x) \quad \text{for all } x \in K.
\]

**Proof.** Let \( \{x_k\} \) in \( K \) be a minimizing sequence for \( f \), that is, \( f(x_k) \to \inf \{f(x) : x \in K\} =: f^\ast \), where we may have \( f^\ast = -\infty \). Since \( K \) is compact, there exists a subsequence \( \{x_{k_i}\} \) converging to \( x^\ast \in K \). Since \( f \) is continuous, we have \( f(x^\ast) = \lim_{i \to \infty} f(x_{k_i}) = f^\ast \in \mathbb{R} \), and thus the point \( x^\ast \) is a global minimizer of \( f \) on \( K \). \( \square \)

An alternative proof runs as follows:

**Proof.** Define \( K_n := \{x \in K : f(x) > n\} \). Then \( K_n \) is open, and \( K = \bigcup_{n=\infty}^{\infty} K_n \), that is, \( \{K_n\}_{n=\infty}^{\infty} \) is an open cover of \( K \). Since \( K \) is compact, a finite subset \( \{K_{n_i}\}_{i=1}^{k} \) also covers \( K \), that is, \( K = \bigcup_{i=1}^{k} K_{n_i} \). Then \( K = K_n \), where \( n := \min \{n_i : i = 1, \ldots, k\} \), and \( f^\ast := \inf \{f(x) : x \in K\} > -\infty \). Thus, \( f \) is bounded from below on \( K \).

Suppose that \( f \) does not have a global minimizer on \( K \). Define \( F_n := \{x \in K : f(x) > f^\ast + 1/n\} \). Then \( F_n \) is an open subset of \( K \) and \( K = \bigcup_{n=1}^{\infty} F_n \). As above, we have \( K = F_n \) for some \( n > 1 \), that is, \( f(x) > f^\ast + 1/n \) for all \( x \in K \), a contradiction to the definition of \( f^\ast \). \( \square \)

We remark that the second proof is more general, since it is valid verbatim on all compact topological spaces, not only compact metric spaces.

The compactness assumption can be relaxed somewhat.

**Theorem 2.3.** Let \( f : E \to \mathbb{R} \) be a continuous function defined on a metric space \( E \). If \( f \) has a nonempty, compact sublevel set \( \{x \in E : f(x) \leq \alpha\} \), then \( f \) achieves a global minimizer on \( E \).

**Proof.** Let \( \{x_n\} \) be a minimizing sequence for \( f \), that is,

\[
f(x_n) \to \inf \{f(x) : x \in E\} = \inf_{E} f =: f^\ast.
\]

Denote by \( D \) the sublevel set above, that is, \( D = \{x \in E : f(x) \leq \alpha\} \). Clearly, there exists \( N \) such that \( x_n \in D \) for all \( n \geq N \). Since \( D \) is compact, \( \{x_n\}_{N}^{\infty} \) has a convergent subsequence \( x_{n_k} \to x^\ast \in D \). Since \( f \) is continuous, we have

\[
f(x^\ast) = \lim_{n \to \infty} f(x_n) = f^\ast.
\]

This means that \( f \) achieves its minimum on \( E \) at the point \( x^\ast \). \( \square \)
**Definition 2.4.** A function \( f : D \to \mathbb{R} \) on a subset \( D \) of a normed vector space \( E \) is called coercive if
\[
f(x) \to \infty \quad \text{as} \quad \|x\| \to \infty.
\]

**Corollary 2.5.** If \( f : D \to \mathbb{R} \) is a continuous coercive function defined on a closed set \( D \subseteq \mathbb{R}^n \), then \( f \) achieves a global minimum on \( D \).

**Proof.** The sublevel sets \( l_\alpha(f) = \{ x \in D : f(x) \leq \alpha \} \) are closed, since \( f \) is continuous, and bounded since \( f \) is coercive. Thus, \( f \) achieves its minimum on \( L \) at a point \( x^* \), which is also a global minimizer of \( f \) on \( D \). \( \square \)

**Example 2.6.** *(The fundamental theorem of algebra)*

This famous theorem states that every polynomial
\[
p(z) := a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0,
\]
with leading coefficient \( a_n \neq 0 \) and where the coefficients \( a_i \) are complex numbers, has a complex root, hence \( n \) complex roots counting multiplicities. The problem has a fascinating history, and it is generally agreed that the first rigorous proof of it was given by the great mathematician Gauss in 1797, when he was just 20 years old, and appeared in his doctoral thesis of 1799. Here, we give an elementary proof of this result. This very short proof from [253] uses optimization techniques, but the essential idea is already in Fefferman [92], and probably in earlier works.

Consider minimizing the function
\[
f(z) = |p(z)|
\]
over the complex numbers. We have
\[
|p(z)| = |z|^n \cdot \left| a_n + \frac{a_{n-1}}{z} + \frac{a_{n-2}}{z^2} + \cdots + \frac{a_1}{z^{n-1}} + \frac{a_0}{z^n} \right|.
\]
As \( |z| \to \infty \), the norm of the sum above converges to \( |a_n| > 0 \). Thus, \( f(z) \) is a coercive function, and so has a minimizer \( z^* \) in \( \mathbb{C} \).

Without loss of any generality, we may assume that \( z^* = 0 \); otherwise, we can consider the polynomial \( q(z) = p(z + z^*) \). We have
\[
|a_0| = f(0) \leq f(z) = \left| \sum_{k=0}^n a_k z^k \right|, \quad z \in \mathbb{C}.
\]
If \( a_0 = 0 \), \( z = 0 \) is a root of \( p \), and we are done. We claim that in fact, \( a_0 = 0 \). Suppose \( a_0 \neq 0 \) and let
\[
p(z) = a_0 + a_k z^k + z^{k+1} q(z),
\]
where \( a_k \neq 0 \) is the first nonzero coefficient after \( a_0 \) and \( q \) is a polynomial. Choose a \( k \)th root \( w \in \mathbb{C} \) of \(-a_0/a_k\). Then
2.2 First-Order Optimality Conditions

\[ p(tw) = a_0 + a_k t^k w^k + t^{k+1} w^{k+1} q(tw) = (1 - t^k)a_0 + t^k [tw^{k+1} q(tw)] . \]

If \( 0 < t < 1 \) is small enough, then \( t |w^{k+1} q(tw)| < |a_0| \), and

\[ |p(tw)| < (1 - t^k)|a_0| + t^k |a_0| = |a_0| , \]
a contradiction.

**2.2 First-Order Optimality Conditions**

Theorem 2.7. (First-order necessary condition for a local optimizer)

Let \( f : U \to \mathbb{R} \) be a Gâteaux differentiable function on an open set \( U \subseteq \mathbb{R}^n \). A local optimizer is a critical point, that is,

\[ x \text{ a local optimizer } \implies \nabla f(x) = 0 . \]

Clearly, the theorem holds verbatim if \( U \subseteq \mathbb{R}^n \) is an arbitrary set with a nonempty interior, \( f \) is Gâteaux differentiable on \( \text{int} \ U \), and \( x \in \text{int} \ U \). We will not always point out such obvious facts in the interest of not complicating the statements of our theorems.

**Proof.** We first assume that \( x \) is a local minimizer of \( f \). If \( d \in \mathbb{R}^n \), then

\[ f'(x; d) = \lim_{t \to 0} \frac{f(x + td) - f(x)}{t} = \langle \nabla f(x), d \rangle . \]

If \( |t| \) is small, then the numerator above is nonnegative, since \( x \) is a local minimizer. If \( t > 0 \), then the difference quotient is nonnegative, so in the limit as \( t \downarrow 0 \), we have \( f'(x; d) \geq 0 \). However, if \( t < 0 \), the difference quotient is nonpositive, and we have \( f'(x; d) \leq 0 \). Thus, we conclude that \( f'(x; d) = \langle \nabla f(x), d \rangle = 0 \). If \( x \) is a local maximizer of \( f \), then \( \langle \nabla f(x), d \rangle = 0 \), since \( x \) is a local minimizer of \(-f\). Picking \( d = \nabla f(x) \) gives \( f'(x; d) = \|\nabla f(x)\|^2 = 0 \), that is, \( \nabla f(x) = 0 \). \( \square \)

We note that Theorem 2.7 proves the following more general result.

**Corollary 2.8.** Let \( f : U \to \mathbb{R} \) be a function on an open set \( U \subseteq \mathbb{R}^n \). If \( x \in U \) is a local minimizer of \( f \) and the directional derivative \( f'(x; d) \) exists for a direction \( d \in \mathbb{R}^n \), then \( f'(x; d) \geq 0 \).

**Remark 2.9.** Functions that have directional derivatives but are not necessarily differentiable occur naturally in optimization, for example in minimizing a function that is the pointwise maximum of a set of differentiable functions. See Danskin’s theorem, Theorem 1.29, on page 20.

In fact, it is possible use this approach to derive optimality conditions for constrained optimization problems. See Section 12.1 for the derivation of optimality conditions in semi-infinite programming.
Example 2.10. Here is an optimization problem from the theory of orthogonal polynomials; see [250], whose solution is obtained using a novel technique, a differential equation.

We determine the minimizers and the minimum value of the function

\[ f(x_1, \ldots, x_n) = \frac{1}{2} \sum_{1}^{n} x_j^2 - \sum_{1 \leq i < j \leq n} \ln |x_i - x_j|. \]

Differentiate \( f \) with respect to each variable \( x_j \) and set to zero to obtain

\[ \frac{\partial f}{\partial x_j} = x_j - \sum_{i \neq j} \frac{1}{x_j - x_i} = 0. \]

To solve for \( x \), consider the polynomial

\[ g(x) = \prod_{1}^{n} (x - x_j), \]

which has roots at the point \( x = x_1, \ldots, x_n \). Differentiating this function gives

\[ g'(x_j) = \prod_{i \neq j} (x_j - x_i), \quad g''(x_j) = 2 \prod_{i \neq j} \frac{1}{x_j - x_i}, \]

so that \( \partial f / \partial x_j = 0 \) can be written as

\[ g''(x_j) - 2x_j g'(x_j) = 0, \]

meaning that the polynomial

\[ g''(x) - 2x g'(x) \]

of order \( n \) has the same roots as the polynomial \( g(x) \), so must be proportional to \( g(x) \). Comparing the coefficients of \( x^n \) gives

\[ g''(x) - 2x g'(x) + 2ng(x) = 0. \]

The solution to this differential equation is the Hermite polynomial of order \( n \),

\[ H_n(x) = n! \sum_{0}^{[n/2]} \frac{(-1)^k (2x)^{n-2k}}{k!(n-2k)!}. \]

Therefore, the solutions \( x_j \) are the roots of the Hermite polynomial \( H_n(x) \).

The discriminant of \( H_n \) is given by

\[ \prod_{i<j} (x_i - x_j)^2 = 2^{-(n(n-1)/2)} \prod_{1}^{n} j^j, \]
and the above formula for $H_n$ gives
\[
\sum_{j=1}^n x_j^2 = n(n-1)/2.
\]
Thus, the minimum value of $f$ is
\[
\frac{1}{4} n(n-1)(1 + \ln 2) - \frac{1}{2} \sum_{j=1}^n j \ln j.
\]

### 2.3 Second-Order Optimality Conditions

**Definition 2.11.** An $n \times n$ matrix $A$ is called positive semidefinite if
\[
\langle Ad, d \rangle \geq 0 \text{ for all } d \in \mathbb{R}^n.
\]
It is called positive definite if
\[
\langle Ad, d \rangle > 0 \text{ for all } d \in \mathbb{R}^n, \ d \neq 0.
\]

Note that if $A$ is positive semidefinite, then $a_{ii} = \langle Ae_i, e_i \rangle \geq 0$, and if $A$ is positive definite, then $a_{ii} > 0$. Similarly, choosing $d = te_i + e_j$ gives $q(t) := a_{ii}t^2 + 2a_{ij}t + a_{jj} \geq 0$ for all $t \in \mathbb{R}$. Recall that the quadratic function $q(t)$ is nonnegative (positive) if and only if its discriminant $\Delta = 4(a_{ij}^2 - a_{ii}a_{jj})$ is nonpositive (negative). Thus, $a_{ii}a_{jj} - a_{ij}^2 \geq 0$ if $A$ is positive semidefinite, and $a_{ii}a_{jj} - a_{ij}^2 > 0$ if $A$ is positive definite.

**Theorem 2.12. (Second-order necessary condition for a local minimizer)** Let $f : U \to \mathbb{R}$ be twice Gâteaux differentiable on an open set $U \subseteq \mathbb{R}^n$ in the sense that there exist a vector $\nabla f(x)$ and a symmetric matrix $Hf(x)$ such that for all $h \in \mathbb{R}^n$,
\[
f(x + th) = f(x) + t \langle \nabla f(x), h \rangle + \frac{t^2}{2} \langle Hf(x)h, h \rangle + o(t^2).
\]

(This condition is satisfied if $f$ has continuous second-order partial derivatives, that is, if $f \in C^2$.)

If $x \in U$ is a local minimizer of $f$, then the matrix $Hf(x)$ is positive semidefinite.

**Proof.** The first-order necessary condition implies $\nabla f(x) = 0$. Since $x$ is a local minimizer, we have $f(x + th) \geq f(x)$ if $|t|$ is small enough. Then, (2.1) gives
\[
\frac{t^2}{2} \langle Hf(x)h, h \rangle + o(t^2) \geq 0.
\]
Dividing by $t^2$ and letting $t \to 0$ gives
\[
h^T Hf(x)h \geq 0 \text{ for all } h \in \mathbb{R}^n,
\]
proving that $Hf(x)$ is positive semidefinite. \qed
We remark that the converse does not hold; see Exercise 9 on page 56. However, we have the following theorem.

**Theorem 2.13. (Second-order sufficient condition for a local minimizer)** Let \( f : U \to \mathbb{R} \) be \( C^2 \) on an open set \( U \subseteq \mathbb{R}^n \). If \( x \in U \) is a critical point and \( Hf(x) \) is positive definite, then \( x \) is a strict local minimizer of \( f \) on \( U \).

**Proof.** Define \( A := Hf(x) \). Since \( g(d) := \langle Ad, d \rangle > 0 \) for all \( d \) on the unit sphere \( S := \{ d \in \mathbb{R}^n : \|d\| = 1 \} \) and \( S \) is compact, it follows that there exists \( \alpha > 0 \) such that \( g(d) \geq \alpha > 0 \) for all \( d \in S \). Since \( g \) is homogeneous, we have \( g(d) \geq \alpha \|d\|^2 \) for all \( d \in \mathbb{R}^n \).

Let \( \|d\| \) be sufficiently small. It follows from the multivariate Taylor’s formula (Corollary 1.24) and the fact \( \nabla f(x) = 0 \) that

\[
\begin{align*}
  f(x + d) &= f(x) + \langle \nabla f(x), d \rangle + \frac{1}{2}\langle Ad, d \rangle + o(\|d\|^2) \\
  &\geq f(x) + \|d\|^2 \left( \frac{\alpha}{2} + o(\|d\|^2) \right) \\
  &> f(x).
\end{align*}
\]

This proves that \( x \) is a strict local minimizer of \( f \). \( \square \)

The positive definiteness condition on \( A \) is really needed. Exercise 9 describes a problem in which a critical point \( x \) has \( Hf(x) \) positive semidefinite, but \( x \) is actually a saddle point.

However, a global positive semidefiniteness condition on \( Hf(x) \) has strong implications.

**Theorem 2.14. (Second-order sufficient condition for a global minimizer)** Let \( f : U \to \mathbb{R} \) be a function with positive semidefinite Hessian on an open convex set \( U \subseteq \mathbb{R}^n \). If \( x \in U \) is a critical point, then \( x \) is a global minimizer of \( f \) on \( U \).

**Proof.** Let \( y \in U \). It follows from the multivariate Taylor’s formula (Theorem 1.23) that there exists a point \( z \in (x, y) \) such that

\[
f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}(y - x)^T Hf(z)(y - x).
\]

Since \( \nabla f(x) = 0 \) and \( Hf(z) \) is positive semidefinite, we have \( f(y) \geq f(x) \) for all \( y \in D \). Thus, \( x \) is a global minimizer of \( f \) on \( U \). \( \square \)

**Remark 2.15.** We remark that a function with a positive semidefinite Hessian is a convex function. If the Hessian is positive definite at every point, then the function is strictly convex. In this case, the function \( f \) has at most one critical point, which is the unique global minimizer. Chapter 4 treats convex (not necessarily differentiable) functions in detail.
Theorem 2.16. (**Second-order sufficient condition for a saddle point**)

Let \( f : U \to \mathbb{R} \) be twice Gâteaux differentiable on an open set \( U \subseteq \mathbb{R}^n \) in the sense of (2.1). If \( x \in U \) is a critical point and \( Hf(x) \) is indefinite, that is, it has at least one positive and one negative eigenvalue, then \( x \) is a saddle point of \( f \) on \( U \).

**Proof.** Define \( A := Hf(x) \). If \( \lambda > 0 \) is an eigenvalue of \( A \) with a corresponding eigenvector \( d \in \mathbb{R}^n \), \( \|d\| = 1 \), then \( \langle Ad, d \rangle = \langle \lambda d, d \rangle = \lambda \), and it follows from Corollary 1.24 that for sufficiently small \( t > 0 \),

\[
f(x + td) = f(x) + t\langle \nabla f(x), d \rangle + \frac{t^2}{2} \langle Ad, d \rangle + o(t^2)
\]

\[
= f(x) + \frac{t^2}{2} \lambda + o(t^2) > f(x).
\]

Similarly, if \( \lambda < 0 \) is an eigenvalue of \( A \) with a corresponding eigenvector \( d \), \( \|d\| = 1 \), then \( f(x + td) < f(x) \) for small enough \( t > 0 \). This proves that \( x \) is a saddle point. \( \square \)

**Definition 2.17.** Let \( f : U \to \mathbb{R} \) be a \( C^2 \) function on an open set \( U \subseteq \mathbb{R}^n \). A critical point \( x \in U \) is called nondegenerate if the Hessian matrix \( D^2 f(x) \) is nonsingular.

A well-known result, Morse’s lemma [202], states that if \( x \) is a nondegenerate critical point, then the Hessian \( Df(x_0) \) determines the behavior of \( f \) around \( x_0 \). More precisely, it states that if \( f : U \to \mathbb{R} \) is at least \( C^{2+k} \) (\( k \geq 1 \)) on an open set \( U \subseteq \mathbb{R}^n \), and if \( x_0 \in U \) is a nondegenerate critical point of \( f \), then there exist open neighborhoods \( V \ni x_0 \) and \( W \ni 0 \) in \( \mathbb{R}^n \) and a one-to-one and onto \( C^k \) map \( \varphi : V \to W \) such that

\[
f(x) = f(x_0) + \frac{1}{2} \langle D^2 f(x_0) \varphi(x), \varphi(x) \rangle.
\]

This is the content of Theorem 2.32 on page 49. See also Corollary 2.33.

We end this section by noting that the second-order tests considered above, and especially Morse’s lemma, give conclusive information about a critical point except when the Hessian matrix is degenerate. In these degenerate cases, nothing can be deduced about the critical point in general: it could be a local minimizer, local maximizer, or a saddle point. For example, the origin \((x, y) = (0, 0)\) is a critical point of the function \( f(x, y) = x^3 - 3xy^2 \) (the real part of the complex function \((x + iy)^3\)), with \( D^2 f(0, 0) = 0 \). It is a saddle point, and the graph of this function is called a monkey saddle. A computer plot of the graph of \( f \) will reveal that this saddle is different from the familiar horse saddle in that there is also a third depression for the tail of the monkey.

**Example 2.18.** Consider the family of problems

\[
\min f(x, y) := x^2 + y^2 + \beta xy + x + 2y.
\]
We have
\[ \nabla f(x, y) = \begin{pmatrix} 2x + \beta y + 1 \\ 2y + \beta x + 2 \end{pmatrix}, \quad Hf(x, y) = \begin{bmatrix} 2 & \beta \\ \beta & 2 \end{bmatrix}. \]

We have \( \nabla f(x, y) = 0 \) if and only if
\[ 2x + \beta y = -1, \]
\[ \beta x + 2y = -2. \]

If \( \beta \neq \pm 2 \), then the unique solution to the above equations is \((x^*, y^*) = (2\beta - 2, \beta - 4)/(4 - \beta^2)\). If \( \beta = 2 \), the above equations become \( 2x + 2y = -1 \) and \( 2x + 2y = -2 \), thus inconsistent. Similarly, if \( \beta = -2 \), we also have an inconsistent system of equations. Therefore, no critical points exist for \( \beta = \mp 2 \).

The eigenvalues of \( A := Hf(x, y) \) can be calculated explicitly: the characteristic polynomial of \( A \) is
\[ \det(A - \lambda I) = (2 - \lambda)^2 - \beta^2 = 0, \]
which has solutions \( \lambda = 2 \mp \beta \). These are the eigenvalues of \( A \). Thus, the eigenvalues of \( A \) are positive for \( -2 < \beta < 2 \). In this case, the optimal solution \((x^*, y^*)\) calculated above is a global minimizer of \( f \) by Theorem 2.13 and Corollary 2.20 below. In the case \( |\beta| > 2 \), one eigenvalue of \( A \) is positive and the other negative, so that the corresponding optimal solution \( z^* := (x^*, y^*) \) is a saddle point by Theorem 2.16.

Finally, let us consider the behavior of \( f \) when \( \beta = \mp 2 \), when it has no critical point. If \( \beta = 2 \), then \( f(x, y) = (x + y)^2 + x + 2y \); thus \( f(x, -x) = -x \) and \( f(x, -x) \to \mp \infty \) as \( x \to \pm \infty \). When \( \beta = -2 \), \( f \) has a similar behavior.

### 2.4 Quadratic Forms

We have seen that symmetric positive semidefinite and positive definite matrices are important in the second-order optimality conditions for a local minimizer. In this section, we give characterizations of such matrices.

We recall the spectral decomposition or orthogonal diagonalization of symmetric matrices.

**Theorem 2.19. (Spectral decomposition of a symmetric matrix)** Let \( A \) be an \( n \times n \) real symmetric matrix. There exist a real diagonal matrix \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n) \) and a real orthogonal matrix \( U = [u_1, \ldots, u_n] \) such that
\[ A = U\Lambda U^T. \]

The scalar \( \lambda_i \) is an eigenvalue of \( A \), and \( u_i \) is an eigenvector of \( A \) corresponding to \( \lambda_i \).
Proof. It is well known from linear algebra that \( A \) has \( n \) real eigenvalues \( \{\lambda_i\}_{i=1}^n \) with corresponding eigenvectors \( \{u_i\}_{i=1}^n \), \( \|u_i\| = 1 \), which are mutually orthogonal, that is, \( \langle u_i, u_j \rangle = 0 \) for \( i \neq j \). From \( Au_i = \lambda_i u_i \), we obtain

\[
AU = A[u_1, \ldots, u_n] = [Au_1, \ldots, Au_n] = [\lambda_1 u_1, \ldots, \lambda_n u_n] = U\Lambda,
\]

where \( U = [u_1, \ldots, u_n] \) and \( A = \text{diag}(\lambda_1, \ldots, \lambda_n) \). Since the eigenvalues are orthogonal, we have

\[
U^T U = \begin{bmatrix}
  u_1^T \\
  \vdots \\
  u_n^T
\end{bmatrix} [u_1, \ldots, u_n] = [\langle u_i, u_j \rangle] = I,
\]

that is, \( U \) is an orthogonal matrix with inverse \( U^T \). It follows that \( A = A(UU^T) = (AU)U^T = U\Lambda U^T \).

In Section 10.1 (page 251), we will give an optimization proof of this theorem. This approach provides a variational characterization of the eigenvalues, which has many applications.

**Corollary 2.20.** Let \( A \) be an \( n \times n \) symmetric matrix. Then \( A \) is positive semidefinite if and only if all eigenvalues of \( A \) are nonnegative. Moreover, \( A \) is positive definite if and only if all eigenvalues of \( A \) are positive.

**Proof.** We have

\[
d^T Ad = d^T U \Lambda U^T d = (U^T d)^T \Lambda (U^T d).
\]

Since \( U \) is nonsingular, we see that \( d^T Ad \geq 0 \) for all \( d \in \mathbb{R}^n \) if and only if \( d^T \Lambda d \geq 0 \) for all \( d \in \mathbb{R}^n \). In other words, \( A \) is positive semidefinite if and only if \( \Lambda \) is. Since

\[
d^T \Lambda d = \sum_{i=1}^n \lambda_i d_i^2,
\]

\( A \) is positive semidefinite if and only \( \lambda_i \geq 0 \) for each \( i = 1, \ldots, n \). This proves the first part of the theorem. The proof of the second part is similar. \( \square \)

Although this result characterizes the symmetric positive semidefinite and positive definite matrices, the determination of the eigenvalues of \( A \) is not an easy computational task unless \( n \) is small. However, here we are interested only in the signs of the eigenvalues and not their exact numerical values.

It is also possible to simultaneously “diagonalize” two symmetric matrices, provided one of them is positive definite. This result is frequently useful in optimization. For example, it may be used to give a quick proof of the fact that the function \( F(X) = -\ln \det X \) is convex on the cone of positive definite matrices.
Theorem 2.21. Let $A$ and $B$ be symmetric $n \times n$ matrices such that at least one of the matrices is positive definite. The matrices can be simultaneously diagonalized in the sense that there exists a nonsingular matrix $X \in \mathbb{R}^{n \times n}$ such that

$$X^T AX = \text{diag}\{\lambda_1, \ldots, \lambda_n\}, \quad X^T BX = \text{diag}\{\delta_1, \ldots, \delta_n\}.$$ 

We remark that the notion of diagonalization in this theorem is different from its usual definition in linear algebra where diagonalizing a square matrix $A$ means finding an invertible matrix $X$ such that $X^{-1} AX$ is a diagonal matrix. The diagonalization above is more appropriate for quadratic forms, because substituting $x = Xy$ in the quadratic form $q_1(x) = \langle Ax, x \rangle$ gives the quadratic form $q_2(y) := q_1(Xx) = \langle Cy, y \rangle$ where $C = X^T AX$.

Proof. Suppose that $B$ is positive definite. Then $B$ has the spectral decomposition $U^T BU = D$, where $U \in \mathbb{R}^{n \times n}$ is orthogonal and $D = \text{diag}\{d_1, \ldots, d_n\}$ is a diagonal matrix with all $d_i > 0$. Define the square root of $B$, 

$$C := U \text{diag}\{\sqrt{d_1}, \ldots, \sqrt{d_n}\} U^T.$$ 

Note that $C^{-1} BC^{-1} = I$. Now $A := C^{-1} AC^{-1}$ has the spectral decomposition $V^T AV = \Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$. Setting $X = C^{-1} V$, we see that

$$X^T AX = \Lambda, \quad X^T BX = V^T C^{-1} BC^{-1} V = V^T V = I,$$

completing the proof. \qed

2.4.1 Counting Roots of Polynomials in Intervals

The number of positive (and negative) eigenvalues can be counted by a simple rule dating back to Descartes in seventeenth century.

Definition 2.22. Let $a_0, a_1, \ldots, a_n$ be a sequence of real numbers. If all the numbers in the sequence are nonzero, the total number of variations of sign in the sequence, denoted by $V(a_0, a_1, \ldots, a_n)$, is the number of times consecutive numbers $a_{k-1}$ and $a_k$ differ in sign, that is,

$$V(a_0, a_1, \ldots, a_n) := |\{k : a_{k-1} a_k < 0, k = 1, \ldots, n\}|.$$ 

If the sequence $a_0, a_1, \ldots, a_n$ contains zeros, then $V(a_0, a_1, \ldots, a_n)$ is defined to be the variations of the reduced sequence by ignoring all zero elements in the sequence. Also, we define $V(a_0) = 0$ for any $a_0 \in \mathbb{R}$.

For example, $V(1, 0, 0, -3, 2, 0, 1, -7, 3) = V(1, -3, 2, 1, -7, 3) = 4$.

Theorem 2.23. (Descartes’s rule of sign) Let $p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$ be a polynomial of degree $n$ with real coefficients. Then the number of positive roots $N_p(0, \infty)$ of $p$ is given by
2.4 Quadratic Forms

\[ N_p(0, \infty) = V(a_0, a_1, \ldots, a_n) - 2\kappa \]

for some nonnegative integer \( \kappa \).

Moreover, if the roots of \( p \) are all real, then \( \kappa = 0 \), that is,

\[ N_p(0, \infty) = V(a_0, a_1, \ldots, a_n). \]

A simple proof of the theorem is given in Appendix B.

**Corollary 2.24.** Let \( A_{n \times n} \) be a symmetric matrix and let \( p(\lambda) = \det(\lambda I - A) = a_0 + a_1 \lambda + \cdots + a_n \lambda^n \) be the characteristic polynomial of \( A \). The number of positive eigenvalues of \( A \) is given by

\[ N_p(0, \infty) = V(a_0, a_1, \ldots, a_n), \]

and the number of negative eigenvalues by

\[ N_p(-\infty, 0) = V(a_0, -a_1, a_2, \ldots, (-1)^n a_n). \]

**Proof.** The characteristic polynomial has only real roots, these being the eigenvalues of \( A \). This proves the first equality. The second equality follows by considering the polynomial \( q(\lambda) = -p(\lambda) \) and noting that the \( k \) coefficient of \( q \) is \((-1)^k a_k\). \( \square \)

Alternatively, \( N_p(-\infty, 0) \) can be computed by noting that the positive, negative, and zero eigenvalues (counted according to its multiplicity) of \( A \) add up to \( n \).

### 2.4.2 Sylvester’s Theorem

There is also a remarkable determinant test due to Sylvester to recognize a symmetric positive definite matrix. We first need to introduce some concepts.

Let \( A \) be an \( n \times n \) symmetric matrix. The submatrix

\[
A_k := \begin{bmatrix}
a_{11} & \cdots & a_{1k} \\
\vdots & \ddots & \vdots \\
 a_{k1} & \cdots & a_{kk}
\end{bmatrix}
\]

consisting of the first \( k \) rows and columns of \( A \) is called the \( k \)th leading principal submatrix of \( A \), and its determinant \( \det A_k \) is called the \( k \)th leading principal minor of \( A \).

**Theorem 2.25. (Sylvester)** Let \( A \) be an \( n \times n \) symmetric matrix. Then \( A \) is positive definite if and only if all the leading principal minors of \( A \) are positive, that is, \( A \) is positive definite if and only if \( \det A_i > 0, i = 1, \ldots, n \).
Proof. We first prove that if $A$ is positive definite, then all leading principal minors of $A$ are positive. We use induction on $n$, the dimension of $A$. The proof is trivial for $n = 1$. Assuming that the result is true for $n$, we will prove it for $n + 1$. Let $A$ be an $(n+1) \times (n+1)$ symmetric, positive definite matrix. We write

$$A = \begin{bmatrix} B & b \\ b^T & c \end{bmatrix},$$

where $B$ is a symmetric $n \times n$ matrix, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$. Choosing $0 \neq d \in \mathbb{R}^n$, we have

$$0 < (d^T, 0)A \begin{bmatrix} d \\ 0 \end{bmatrix} = (d^T, 0) \begin{bmatrix} B & b \\ b^T & c \end{bmatrix} \begin{bmatrix} d \\ 0 \end{bmatrix} = d^TBd,$$

that is, $B$ is positive definite. By the induction hypothesis, we have $\det A_i > 0$, $i = 1, \ldots, n$. Since $A$ is positive definite, its eigenvalues $\{\lambda_i\}_{i=1}^{n+1}$ are all positive. Thus, we also have $\det A_{n+1} = \det A = \lambda_1 \cdots \lambda_{n+1} > 0$.

Conversely, let us prove that if all $\det A_i > 0$, $i = 1, \ldots, n+1$, then $A$ is positive definite. The proof is again by induction on $n$. The proof is trivial for $n = 1$. Suppose the theorem is true for $n$; we will prove it for $n+1$. Since $\det A_i > 0$ for $i = 1, \ldots, n$ we see by the induction hypothesis that $B$ is positive definite. Suppose $A$ is not positive definite. Then $\lambda_{n+1} < 0$, and since $\det A = \lambda_1 \cdots \lambda_{n+1} > 0$, we must also have $\lambda_n < 0$. Let $u_n$ and $u_{n+1}$ be the eigenvectors of $A$ corresponding to $\lambda_n$ and $\lambda_{n+1}$, respectively. We have $\langle u_n, u_{n+1} \rangle = 0$, so that we can choose scalars $\alpha_n$ and $\alpha_{n+1}$ such that $u = \alpha_n u_n + \alpha_{n+1} u_{n+1}$ is not zero but has the last ($(n+1)$th) component equal to zero, say $u = (v, 0)^T$ where $v \neq 0$. Then $u^T A u = v^T B v > 0$, since $B$ is positive definite. However, we also have

$$0 < u^T A u = \langle \alpha_n u_n + \alpha_{n+1} u_{n+1}, A(\alpha_n u_n + \alpha_{n+1} u_{n+1}) \rangle$$

$$= \langle \alpha_n u_n + \alpha_{n+1} u_{n+1}, \lambda_n \alpha_n u_n + \lambda_{n+1} \alpha_{n+1} u_{n+1} \rangle$$

$$= \lambda_n \alpha_n^2 \langle u_n, u_n \rangle + \lambda_{n+1} \alpha_{n+1}^2 \langle u_{n+1}, u_{n+1} \rangle < 0,$$

where the last inequality follows from the facts $\lambda_i < 0$ and $\|u_i\| = 1$, $i = n, n+1$. This contradiction shows that all eigenvalues of $A$ are positive. Corollary 2.20 implies that $A$ is positive definite.

\[\square\]

This simple proof is taken from Carathéodory [54], p. 187.

Another elegant proof of Sylvester’s theorem, more in the spirit of optimization techniques, is outlined in Exercise 12 at the end of the chapter.

2.5 The Inverse Function, Implicit Function, and Lyusternik Theorems in Finite Dimensions

In this section, we first give an elementary proof of the implicit function theorem in finite-dimensional vector spaces. This proof has a variational flavor,
and is used to prove the inverse function theorem and Lyusternik’s theorem. The implicit function theorem will also be utilized to prove Morse’s lemma in Section 2.6.

**Theorem 2.26. (Implicit function theorem)** Let \( f : U \times V \to \mathbb{R}^m \) be a \( C^1 \) mapping, where \( U \subseteq \mathbb{R}^n \) and \( V \subseteq \mathbb{R}^m \) are open sets. Let \((x_0, y_0) \in U \times V \) be a point such that \( f(x_0, y_0) = 0 \) and \( D_y f(x_0, y_0) : \mathbb{R}^m \to \mathbb{R}^m \), the derivative of \( f \) with respect to \( y \), is nonsingular.

Then there exist neighborhoods \( U_1 \ni x_0 \) and \( V_1 \ni y_0 \) and a \( C^1 \) mapping \( y : U_1 \to V_1 \) such that a point \((x, y) \in U_1 \times V_1 \) satisfies \( f(x, y) = 0 \) if and only if \( y = y(x) \). The derivative of \( y \) at \( x_0 \) is given by

\[
D_y(x_0) = -D_y f(x_0, y_0)^{-1} D_x f(x_0, y_0).
\]

Moreover, if \( f \) is \( k \)-times continuously differentiable, that is, \( f \in C^k \), then \( y(x) \in C^k \).

The linear case should help one to remember the form of the implicit function theorem: if \( f(x, y) = Ax + By \) and \( D_y f = B \) is an invertible matrix, then the equation \( f(x, y) = \alpha \) gives \( Ax + By = \alpha \). This may be solved for \( y \) by premultiplying it by \( B^{-1} \), giving \( y(x) = B^{-1}(\alpha - Ax) \).

**Proof.** Assume without loss of generality that \( x_0 = 0 \) and \( y_0 = 0 \), by considering the function \((x, y) \mapsto f(x + x_0, y + y_0) - f(x_0, y_0)\) if necessary. Let \( f(x) = (f_1(x, y), \ldots, f_m(x, y)) \), where \( f_i \) is the \( i \)th coordinate function of \( f \). Since \( Df \) is continuous, there exist neighborhoods \( U_0 \) and \( V_0 \) of the origin in \( \mathbb{R}^n \) and \( \mathbb{R}^m \), respectively, such that the matrix

\[
\begin{bmatrix}
\nabla_y f_1(x, y_1)^T \\
\nabla_y f_2(x, y_2)^T \\
\vdots \\
\nabla_y f_m(x, y_m)^T
\end{bmatrix}
\tag{2.2}
\]

is invertible for all \((x, y_i) \in U_0 \times V_0\).

We claim that for every \( x \in U_0 \), there exists at most one \( y \in V_0 \) such that \( f(x, y) = 0 \). Otherwise, there would exist \( y, z \in V_0, y \neq z \), such that \( f(x, y) = f(x, z) = 0 \). The mean value theorem (Lemma 1.12) implies that there exists \( y_i \in (y, z) \) such that

\[
f_i(x, z) - f_i(x, y) = \langle \nabla_y f_i(x, y_i), z - y \rangle = 0, \quad i = 1, \ldots, m.
\]

Since the matrix in (2.2) is nonsingular, we obtain \( y = z \), a contradiction that proves our claim.

Let \( B_r(0) \subseteq V_0 \). Since \( f(0, 0) = 0 \), we have \( f(0, y) \neq 0 \) for \( y \in S_r(0) := \{y \in \mathbb{R}^l : \|y\| = r\} \), and since \( f \) is continuous on \( U_0 \times V_0 \), there exists \( \alpha > 0 \) such that \( \|f(0, y)\| \geq \alpha \) for all \( y \in S_r(0) \). It follows that the function

\[
\begin{bmatrix}
\nabla_y f_1(x, y_1)^T \\
\nabla_y f_2(x, y_2)^T \\
\vdots \\
\nabla_y f_m(x, y_m)^T
\end{bmatrix}
\tag{2.2}
\]

is invertible for all \((x, y_i) \in U_0 \times V_0\).

We claim that for every \( x \in U_0 \), there exists at most one \( y \in V_0 \) such that \( f(x, y) = 0 \). Otherwise, there would exist \( y, z \in V_0, y \neq z \), such that \( f(x, y) = f(x, z) = 0 \). The mean value theorem (Lemma 1.12) implies that there exists \( y_i \in (y, z) \) such that

\[
f_i(x, z) - f_i(x, y) = \langle \nabla_y f_i(x, y_i), z - y \rangle = 0, \quad i = 1, \ldots, m.
\]

Since the matrix in (2.2) is nonsingular, we obtain \( y = z \), a contradiction that proves our claim.

Let \( B_r(0) \subseteq V_0 \). Since \( f(0, 0) = 0 \), we have \( f(0, y) \neq 0 \) for \( y \in S_r(0) := \{y \in \mathbb{R}^l : \|y\| = r\} \), and since \( f \) is continuous on \( U_0 \times V_0 \), there exists \( \alpha > 0 \) such that \( \|f(0, y)\| \geq \alpha \) for all \( y \in S_r(0) \). It follows that the function

\[
\begin{bmatrix}
\nabla_y f_1(x, y_1)^T \\
\nabla_y f_2(x, y_2)^T \\
\vdots \\
\nabla_y f_m(x, y_m)^T
\end{bmatrix}
\tag{2.2}
\]

is invertible for all \((x, y_i) \in U_0 \times V_0\).

We claim that for every \( x \in U_0 \), there exists at most one \( y \in V_0 \) such that \( f(x, y) = 0 \). Otherwise, there would exist \( y, z \in V_0, y \neq z \), such that \( f(x, y) = f(x, z) = 0 \). The mean value theorem (Lemma 1.12) implies that there exists \( y_i \in (y, z) \) such that

\[
f_i(x, z) - f_i(x, y) = \langle \nabla_y f_i(x, y_i), z - y \rangle = 0, \quad i = 1, \ldots, m.
\]

Since the matrix in (2.2) is nonsingular, we obtain \( y = z \), a contradiction that proves our claim.

Let \( B_r(0) \subseteq V_0 \). Since \( f(0, 0) = 0 \), we have \( f(0, y) \neq 0 \) for \( y \in S_r(0) := \{y \in \mathbb{R}^l : \|y\| = r\} \), and since \( f \) is continuous on \( U_0 \times V_0 \), there exists \( \alpha > 0 \) such that \( \|f(0, y)\| \geq \alpha \) for all \( y \in S_r(0) \). It follows that the function

\[
\begin{bmatrix}
\nabla_y f_1(x, y_1)^T \\
\nabla_y f_2(x, y_2)^T \\
\vdots \\
\nabla_y f_m(x, y_m)^T
\end{bmatrix}
\tag{2.2}
\]
Unconstrained Optimization

\[ F(x, y) := \|f(x, y)\|^2 = \sum_{i=1}^{m} f_i(x, y)^2 \]

satisfies the properties

\[ F(0, y) \geq \alpha > 0 \quad \text{for} \quad y \in S_r(0) \quad \text{and} \quad F(0, 0) = 0. \]

Since \( F \) is continuous, there exists an open neighborhood \( U_1 \subseteq U_0 \) of \( 0 \in \mathbb{R}^n \) such that

\[ F(x, y) \geq \frac{\alpha}{2}, \quad F(x, 0) \leq \frac{\alpha}{2} \quad \text{for all} \quad x \in U_1, \ y \in S_r(0). \]

Thus, for a fixed \( x \in U_1 \), the function \( y \mapsto F(x, y(x)) \) achieves its minimum on \( B_r(0) \) at a point \( y(x) \) in the interior of \( B_r(0) \), and we have

\[ D_y F(x, y(x)) = 2D_y f(x, y(x))f(x, y(x)) = 0, \]

and since the matrix \( D_y f(x, y(x)) \) is nonsingular, we conclude that

\[ f(x, y(x)) = 0. \]

Writing \( \Delta y := y(x + \Delta x) - y(x) \), we have by the mean value theorem

\[ 0 = D_x f(\tilde{x}, \tilde{y}) \Delta x + D_y f(\tilde{x}, \tilde{y}) \Delta y \]

for some point \((\tilde{x}, \tilde{y})\) on the line segment between \((x, y(x))\) and \((x + \Delta x, y(x + \Delta x))\). This implies that as \( \|\Delta x\| \) goes to zero, so does \( \|\Delta y\| \), proving that \( y(x) \) is a continuous function.

The function \( y(x) \) is actually \( C^1 \), since by Taylor’s formula

\[
0 = f(x + \Delta x, y(x + \Delta x)) - f(x, y(x)) \\
= D_x f(x, y(x)) \Delta x + D_y f(x, y(x)) \Delta y + o((\Delta x, \Delta y)),
\]

and since \( o((\Delta x, \Delta y)) = o(\Delta x) \) by the continuity of \( y(x) \), we have

\[ \Delta y = -D_y^{-1} f(x, y(x))D_x f(x, y(x)) \Delta x + o(\Delta x). \]

This proves that \( y(x) \) is Fréchet differentiable at \( x \) with

\[ D_y(y(x)) = -D_y^{-1} f(x, y(x))D_x f(x, y(x)). \]

If \( f \in C^2 \), then \( D_y^{-1} f(x, y(x)) = \text{Adj} D_y f(x, y(x))/\det D_y f(x, y(x)) \) and \( D_x f(x, y(x)) \) are \( C^1 \), and the above formula shows that the function \( y(x) \) is \( C^2 \). In general, if \( C^k \), we prove by induction on \( k \) that \( y(x) \) is \( C^k \).

This elementary proof is taken from Carathéodory [54], pp. 10–13. A similar kind of proof, using penalty functions, will used in Chapter 9 to obtain optimality conditions for constrained optimization problems.
Corollary 2.27. (Inverse function theorem) Let $f$ be a $C^1$ map from a neighborhood of $x_0 \in \mathbb{R}^n$ into $\mathbb{R}^n$.

If $Df(x_0)$ is nonsingular, then there exist neighborhoods $U \ni x_0$ and $V \supseteq y_0 = f(x_0)$ such that $f : U \to V$ is a $C^1$ diffeomorphism, and

$$Df^{-1}(y) = Df(x)^{-1} \text{ for all } (x, y) \in U \times V, \ y = f(x).$$

Moreover, if $f$ is $C^k$, then $f$ is a $C^k$ diffeomorphism on $U$.

Proof. Define the function $F(x, y) = f(x) - y$, and note that $D_x F(x_0, y) = Df(x_0)$ is nonsingular. Apply Theorem 2.26 to $F$. \hfill \Box

The map $f : \mathbb{R}^2 \to \mathbb{R}^2$ given by $f(x, y) = (e^x \cos y, e^x \sin y)$ has the Jacobian $\det Df(x, y) = e^x \neq 0$, hence locally one-to-one around every point $(x, y) \in \mathbb{R}^2$. However, $f$ is clearly not one-to-one globally.

Definition 2.28. Let $M$ be a nonempty subset of $\mathbb{R}^n$ and $x \in M$. A vector $d \in \mathbb{R}^n$ is called a tangent direction of $M$ at $x$ if there exist a sequence $x_n \in M$ converging to $x$ and a nonnegative sequence $\alpha_n$ such that

$$\lim_{n \to \infty} \alpha_n(x_n - x) = d.$$ 

The tangent cone of $M$ at $x$, denoted by $T_M(x)$, is the set of all tangent directions of $M$ at $x$.

This definition is sufficient for our purposes. We remark that the same definition is valid in a topological vector space. A detailed study of this and several related concepts is needed in nonsmooth analysis; see [230] and [199, 200].

Theorem 2.29. (Lyusternik) Let $f : U \to \mathbb{R}^m$ be a $C^1$ map, where $U \subset \mathbb{R}^n$ is an open set. Let $M = f^{-1}(f(x_0))$ be the level set of a point $x_0 \in U$.

If the derivative $Df(x_0)$ is a linear map onto $\mathbb{R}^m$, then the tangent cone of $M$ at $x_0$ is the null space of the linear map $Df(x_0)$, that is,

$$T_M(x_0) = \{d \in \mathbb{R}^n : Df(x_0)d = 0\}.$$ 

Remark 2.30. Let $f = (f_1, \ldots, f_m)$, where $\{f_i\}$ are the components functions of $f$. It is easy to verify that

$$\text{Ker } Df(x_0) = \{d \in \mathbb{R}^n : \langle \nabla f_i(x_0), d \rangle = 0, i = 1, \ldots, m\},$$

and that the surjectivity of $Df(x_0)$ is equivalent to the linear independence of the gradient vectors $\{\nabla f_i(x_0)\}_i^n$.

Proof. We may assume that $x_0 = 0$ and $f(x_0) = 0$, by considering the function $x \mapsto f(x + x_0) - f(x_0)$ if necessary. Define $A := Df(0)$. The proof of the inclusion $T_M(0) \subseteq \text{Ker } A$ is easy: if $d \in T_M(0)$, then there exist points $x(t) = td + o(t) \in M$, and we have
0 = f(0 + td + o(t)) = f(0) + tDf(0)(d) + o(t) = tDf(0)(d) + o(t).

Dividing both sides by $t$ and letting $t \to 0$, we obtain $Df(0)(d) = 0$.

The proof of the reverse inclusion $\ker A \subseteq T_M(0)$ is based on the idea that the equation $f(x) = 0$ can be written as $f(y, z) = 0$ in a form that is suitable for applying the implicit function theorem.

Define $K := \ker A$ and $L := K^\perp$. Since $A$ is onto $\mathbb{R}^m$, we can identify $K$ and $L$ with $\mathbb{R}^{n-m}$ and $\mathbb{R}^m$, respectively, by introducing a suitable basis in $\mathbb{R}^n$. We write a point $x \in \mathbb{R}^n$ in the form $x = (y, z) \in K \times L$. We have

$$A = [D_y f(0), D_z f(0)],$$

so that $D_y f(0) = 0$. Since $A$ has rank $m$, it follows that $D_z f(0)$ is nonsingular.

Theorem 2.26 implies that there exist neighborhoods $U_1 \subseteq \mathbb{R}^m$ and $U_2 \subseteq \mathbb{R}^{n-m}$ around the origin and a $C^1$ map $\alpha : U_1 \to U_2$, $\alpha(0) = 0$, such that $x = (y, z) \in U_1 \times U_2$ satisfies $f(x) = 0$ if and only if $z = \alpha(y)$. The equation $f(x) = 0$ can then be written as $f(y, \alpha(y)) = 0$. Differentiating this equation and using the chain rule, we obtain

$$0 = D_y f(y, \alpha(y)) + D_z f(y, \alpha(y)) D\alpha(y).$$

At the origin $x = 0$, $D_y f(0) = 0$, and $D_z f(0)$ nonsingular, so that $D\alpha(0) = 0$. If $|y|$ is small, we have

$$\alpha(y) = \alpha(0) + D\alpha(0)y + o(y) = o(y).$$

Let $d = (d_1, 0) \in K$. As $t \to 0$, the point $x(t) := (td_1, \alpha(td_1)) = (td_1, o(t))$ lies in $M$, that is, $f(x(t)) = 0$, and satisfies $(x(t) - td)/t = (0, o(t))/t \to 0$. This implies that $K \subseteq T_M(0)$, and the theorem is proved. \hfill \Box

2.6 Morse’s Lemma

Let $f : U \to \mathbb{R}$ be a $C^{2+k}$ ($k \geq 0$) function on an open set $U \subseteq \mathbb{R}^n$. Recall that a critical point $x \in U$ is called nondegenerate if the Hessian matrix $D^2 f(x)$ is nonsingular. Morse’s lemma, due originally to Morse [202], states that after a local, possibly nonlinear, change of coordinates, the function $f$ is identical to its quadratic form $q(x) := f(x_0) + \langle D^2 f(x_0)(x - x_0), x - x_0 \rangle$. Thus, the quadratic function $q(x)$ determines the behavior of the function $f$ around $x_0$.

Morse’s original proof uses the Gram–Schmidt process. A modern version of the proof can be found in Milnor [197]. The simple proof below is from [6]. It has the virtue that the same proof, with obvious modifications, works in Banach spaces.

The following technical result is needed in the proof of Morse’s lemma.
Lemma 2.31. Let \( S^n \) be the space of \( n \times n \) symmetric matrices, \( A \in S^n \) nonsingular, and let \( S^n_A \) be the vector space of \( n \times n \) matrices \( X \) such that \( AX \) is symmetric. The quadratic map

\[
q_A : S^n_A \to S^n \quad \text{defined by} \quad q_A(X) = X^T AX
\]

is locally one-to-one around \( I \in S^n_A \). Consequently, there exist open neighborhoods \( U \ni I \) and \( V \ni A \) such that \( q_A^{-1} : V \to U \) is a well-defined, infinitely differentiable map.

Proof. We have

\[
q(I + tH) := q_A(I + tH) = (I + tH^T)A(I + tH) = A + t(H^T A + AH) + t^2 H^T AH = A + 2tAH + t^2 AH^2,
\]

so that \( Dq(I)(H) = 2AH \). The mapping \( Dq(I) \) is one-to-one, since \( Dq(I)(H) = AH = 0 \) implies \( H = 0 \), due to the fact that \( A \) is nonsingular.

The map \( Dq(I) \) is also onto, since given \( Y \in S^n \), the matrix \( X := A^{-1}Y/2 \) is in \( S^n_A \) and satisfies \( Dq(I)(X) = Y \). The rest of the lemma follows from the inverse function theorem (Corollary 2.27). \( \Box \)

Theorem 2.32. (Morse’s lemma) Let \( k \geq 1 \) and \( f : U \to \mathbb{R} \) be a \( C^{2+k} \) function on an open set \( U \subseteq \mathbb{R}^n \). If \( x_0 \in U \) is a nondegenerate critical point of \( f \), then there exist open neighborhoods \( V \ni x_0 \) and \( W \ni 0 \) in \( \mathbb{R}^n \) and a \( C^k \) diffeomorphism \( \varphi : V \to W \) such that

\[
f(x) = f(x_0) + \frac{1}{2} \langle D^2 f(x_0) \varphi(x), \varphi(x) \rangle.
\]

Proof. We may assume without any loss of generality that \( U \) is a convex set, \( x_0 = 0 \), and \( f(0) = 0 \). Let \( 0 \neq x \in U \), and define \( \alpha(t) := f(tx) \). We have

\[
\alpha(1) = \alpha(0) + \alpha'(0) + \int_0^1 (1 - t) \alpha''(t) \, dt
\]

by Theorem 1.5, and since \( \alpha'(t) = \langle \nabla f(tx), x \rangle \), \( \nabla f(0) = 0 \) and \( \alpha''(t) = \langle D^2 f(tx)x, x \rangle \), we obtain

\[
f(x) = \frac{1}{2} \langle A(x)x, x \rangle, \quad \text{where} \quad A(x) := 2 \int_0^1 (1 - t)D^2 f(tx) \, dt.
\]

Note that \( A : U \to S^n \) is a \( C^k \) map, and \( A(0) = 2(\int_0^1 (1 - t) \, dt)D^2 f(0) = D^2 f(0) \). Consequently, the map

\[
H : V_0 \to Z \quad \text{defined by} \quad H = q_{A(0)}^{-1} \circ A,
\]

where \( V_0 \) is a neighborhood of \( 0 \in \mathbb{R}^n \) and \( Z \) is a neighborhood of \( I \in S^n_{A(0)} \) as in Lemma 2.31, is also \( C^k \).
We have $A = q_{A(0)} \circ H$, that is, $A(x) = q_{A(0)}(H(x)) = H(x)^T A(0) H(x)$ for $x \in V_0$, and
\[
f(x) = \langle A(x)x, x \rangle = \langle H(x)^T A(0) H(x) x, x \rangle = \langle A(0) H(x) x, H(x) x \rangle = \frac{1}{2} \langle A(0) \varphi(x), \varphi(x) \rangle,
\]
where
\[
\varphi(x) := H(x)x, \quad \varphi : V_0 \to \mathbb{R}^n,
\]
is a $C^k$ map. Since $H(0) = I$, we have
\[
x + o(\|x\|) = H(0)x + o(\|x\|) = (H(0) + DH(0)x + o(\|x\|))x = H(x)x = \varphi(x) = \varphi(0) + D\varphi(0)x + o(\|x\|),
\]
where the third and fifth equalities follow from Taylor’s formula. This proves that $D\varphi(0) = I$, and hence is nonsingular. Thus, the inverse function theorem implies that there exist neighborhoods $V, W$ of $0 \in \mathbb{R}^n$, $V \subseteq V_0$, such that $\varphi : V \to W$ is a $C^k$ diffeomorphism. □

**Corollary 2.33.** Let $f : U \to \mathbb{R}$ be a $C^{2+k}$ function as in Theorem 2.32, and let $x_0 \in U$ be a nondegenerate critical point of $f$ such that the Hessian matrix $A = Df(x_0)$ has $k \ (0 \leq k \leq n)$ positive and $n - k$ negative eigenvalues. Then there exists a local, nonlinear coordinate transformation $y = \psi(x)$ ($\psi : W \to V$ is a $C^k$ diffeomorphism between some neighborhoods $W \ni 0$ and $V \ni x_0$) such that
\[
f(\psi(y)) = f(x_0) + y_1^2 + \cdots + y_k^2 - y_{k+1}^2 - \cdots - y_n^2.
\]

**Proof.** Let $A := Df(x_0)$ have the spectral decomposition $A = U^T \Sigma U$, where $A = \text{diag}(\lambda_1, \ldots, \lambda_k, \ldots, \lambda_n)$ with $\lambda_i > 0$ for $i \leq k$ and $\lambda_i < 0$ for $i > k$. Let $\varphi : V \to W$ be the $C^k$ mapping in Theorem 2.32, where $V$ and $W$ are open neighborhoods of $x_0$ and 0, respectively. Define
\[
y = \psi^{-1}(x) := \frac{1}{\sqrt{2}} |\Sigma|^{1/2} U \varphi(x), \quad x \in V,
\]
where $|\Sigma|^{1/2}$ is the diagonal matrix with diagonal entries $\sqrt{|\lambda_i|}$, $i = 1, \ldots, n$. Theorem 2.32 and a straightforward computation give the representation (2.3). □

The proofs in this section work for functions $f$ that are at least $C^3$. However, appropriate versions of Morse’s lemma exist for $C^2$ functions; see, for example, [254]. There also exist higher-order versions of Morse’s lemma for critical points $x_0$ such that there exists $k \geq 2$ such that $D^i f(x_0) = 0$ for $i = 1, \ldots, k - 1$ and $D^k f(x_0)$ is nondegenerate in a certain sense; see [51].
2.7 Semicontinuous Functions

Semicontinuous functions are of independent interest in analysis. They also play an important role in optimization, since they appear in Ekeland’s variational principle and in the theory of convex functions.

The concept of semicontinuous functions can be defined on a topological space. For our purposes, it will be sufficient to consider metric spaces. In this section, $E$ will denote a metric space with a distance function $d$.

We start with some notions of limits. In optimization theory, various operations converge to $\pm \infty$, thus making it convenient to consider extended real numbers by adding $\infty$ and/or $-\infty$ to real numbers.

**Definition 2.34.** Let $\{x_n\}$ be a sequence of extended real numbers, that is, $x_n \in \mathbb{R} \cup \{\pm \infty\}$. The limit inferior of $\{x_n\}$ is

$$
\lim_{n \to \infty} x_n := \lim_{n \to \infty} \inf \{x_n, x_{n+1}, \ldots\} = \sup_n \inf_{k \geq n} x_k,
$$

where the second equality follows since $\{\inf_{k \geq n} x_k\}$ is an increasing sequence in $n$. Similarly, the limit superior of $\{x_n\}$ is

$$
\liminf_{n \to \infty} x_n := \lim_{n \to \infty} \sup x_k = \inf_n \sup_{k \geq n} x_k.
$$

Let $f : E \to \mathbb{R} \cup \{\pm \infty\}$ be an extended real-valued function. The limit inferior of $f$ as $x \in E$ converges to $x_0 \in E$ is defined by

$$
\lim_{x \to x_0} f(x) := \lim_{\delta \to 0} \inf_{d(x, x_0) < \delta} f(x) = \sup_{\delta \to 0} \inf_{d(x, x_0) < \delta} f(x),
$$

and its limit superior by

$$
\limsup_{x \to x_0} f(x) := \lim_{\delta \to 0} \sup_{d(x, x_0) < \delta} f(x) = \inf_{\delta \to 0} \sup_{d(x, x_0) < \delta} f(x).
$$

**Lemma 2.35.** Let $f : E \to \mathbb{R} \cup \{\pm \infty\}$. We have

$$
\lim_{x \to x_0} f(x) = \inf_{\{x_n\} \to \infty} \lim_{n \to \infty} f(x_n),
$$

where the infimum on the right-hand side is taken over all sequences $x_n \to x_0$.

Similarly,

$$
\limsup_{x \to x_0} f(x) = \sup_{\{x_n\} \to \infty} \lim_{n \to \infty} f(x_n).
$$

**Proof.** We prove only the first equality, since the second one follows immediately from it. Define

$$
M := \lim_{x \to x_0} f(x), \quad L := \inf_{\{x_n\} \to \infty} \lim_{n \to \infty} f(x_n), \quad N_\delta := \{x \in E : d(x, x_0) < \delta\}.
$$
First, we consider the case \( M = -\infty \). Note that it is enough to show the existence of a sequence \( x_n \to x_0 \) such that \( f(x_n) \to -\infty \). Since \( M = -\infty \), it follows from the definition of \( \lim_{x \to x_0} f(x) \) above that \( \inf x \in N_{1/n} f(x) = -\infty \) for all \( n > 0 \). Thus, we can find \( x_n \in N_{1/n} \) satisfying \( f(x_n) < -n \), proving the lemma.

Next, consider the case \( M = \infty \). Let \( \{x_n\} \) be an arbitrary sequence converging to \( x_0 \). We claim that \( f(x_n) \to \infty \), from which the lemma follows immediately. Since \( M = \infty \), for a given \( \alpha > 0 \), there exists \( \delta > 0 \) such that \( \alpha < \inf_{N_\delta} f(x) \). Since \( x_n \to x_0 \), there exists \( N \) such that \( x_n \in N_\delta \) for all \( n \geq N \). Thus, \( f(x_n) > \alpha \) for all \( n \geq N \), and the claim is proved.

Finally, consider the case \( -\infty < M < \infty \). On the one hand, given \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that \( \inf_{N_\delta} f(x) > M - \varepsilon \). Thus, \( f(x) > M - \varepsilon \) for all \( x \in N_\delta \). Let \( \{x_n\} \) be an arbitrary sequence converging to \( x_0 \). Since \( x_n \in N_\delta \) for all large enough \( n \), we have \( \lim_{n \to \infty} f(x_n) \geq M - \varepsilon \), for any \( \varepsilon > 0 \). Thus, \( \lim_{n \to \infty} f(x_n) \geq M \) for any sequence converging to \( x_0 \), proving the inequality \( L \geq M \). On the other hand, since \( \inf_{x \in N_\delta} f(x) \not\to M \) as \( \delta \searrow 0 \), we have \( \inf_{x \in N_{1/n}} f(x) \not\to M \) as \( n \to \infty \). If we pick \( x_n \in N_{1/n} \) such that \( f(x_n) \leq (\inf_{x \in N_{1/n}} f(x)) + 1/n \), then

\[
L \leq \lim_{n \to \infty} f(x_n) \leq \lim_{n \to \infty} \left( \inf_{x \in N_{1/n}} f(x) + 1/n \right) = \lim_{n \to \infty} \inf_{x \in N_{1/n}} f(x) = M.
\]

This proves the reverse inequality \( L \leq M \). \( \square \)

We are now ready to define semicontinuous functions.

**Definition 2.36.** Let \( f : E \to \mathbb{R} \cup \{\pm \infty\} \). The function \( f \) is called **lower semicontinuous at a point** \( x_0 \in E \) if

\[
f(x_0) \leq \lim_{x \to x_0} f(x).
\]

Equivalently, by virtue of Lemma 2.35, \( f \) is lower semicontinuous at \( x_0 \) if

\[
f(x_0) \leq \lim_{n \to \infty} f(x_n),
\]

for every sequence \( x_n \to x_0 \).

The function \( f \) is called **upper semicontinuous at** \( x_0 \) if \( -f \) is lower semicontinuous at \( x_0 \), that is,

\[
f(x_0) \geq \lim_{x \to x_0} f(x),
\]

or

\[
f(x_0) \geq \lim_{n \to \infty} f(x_n),
\]

for every sequence \( x_n \to x_0 \).

The function \( f \) is called **lower semicontinuous or closed on** \( E \) if it is lower semicontinuous at every point in \( E \). Similarly, \( f \) is called upper semicontinuous on \( E \) if it is upper semicontinuous at every point in \( E \).
Remark 2.37. We always have \( \lim_{x \to x_0} f(x) \leq f(x_0) \), since \( x_0 \) lies in every neighborhood \( N_\delta \), so that \( f \) is lower semicontinuous at \( x_0 \) if and only if
\[
f(x_0) = \lim_{x \to x_0} f(x).
\]
Similarly, \( f \) is upper semicontinuous at \( x_0 \) if and only if
\[
f(x_0) = \lim_{x \to x_0} f(x).
\]
Also, note that any function is lower semicontinuous at a point \( x \) with \( f(x) = -\infty \), and similarly any function is upper semicontinuous at a point \( x_0 \) with \( f(x) = \infty \).

It will be seen shortly that the semicontinuity properties of \( f \) are tied up with the closedness of its epigraph.

**Definition 2.38.** If \( f : E \to \mathbb{R} \cup \{\pm \mathbb{R}\} \) is a function, the set
\[
\text{epi}(f) := \{(x, t) \in E \times \mathbb{R} : f(x) \leq t\}
\]
is called the epigraph of \( f \). Similarly, the set
\[
\text{hypo}(f) := \{(x, t) \in E \times \mathbb{R} : f(x) \geq t\}
\]
is called the hypograph of \( f \).

**Theorem 2.39.** Let \( f : E \to \mathbb{R} \cup \{\pm \infty\} \). The following are equivalent:

(a) \( f \) is lower semicontinuous (upper semicontinuous) on \( E \),
(b) \( \text{epi}(f) \) (\( \text{hypo}(f) \)) is a closed subset of \( E \times \mathbb{R} \),
(c) The sublevel set \( \{x \in E : f(x) \leq \alpha\} \) (\( \{x \in E : f(x) \geq \alpha\} \)) is closed for all \( \alpha \in \mathbb{R} \).

**Proof.** We prove the theorem only for a lower semicontinuous function, since the upper semicontinuous case follows immediately.

(a) implies (b): Let \((x_n, y_n)\) be a sequence in \( \text{epi}(f) \) converging to a point \((x, y)\). Since \( f \) is lower semicontinuous at \( x \), \( f(x) \leq \lim f(x_n) \leq \lim y_n = y \), proving that \((x, y) \in \text{epi}(f)\).

(b) implies (c): Let \( x_n \) be a sequence in \( L := \{z : f(z) \leq \alpha\} \) converging to a point \( x \in E \). We have \((x_n, \alpha) \in \text{epi}(f)\) converging to \((x, \alpha) \in \text{epi}(f)\), meaning that \( x \in L \). Thus, \( L \) is closed.

(c) implies (a): Let \( f(x) \in \mathbb{R} \). We claim that \( f \) is lower semicontinuous at \( x \). Otherwise, there exists \( \varepsilon > 0 \) such that \( \sup_{\delta > 0} \inf_{N_\delta} f(x) = f(x) - 2\varepsilon \). Thus, for any \( \delta > 0 \), we have \( \inf_{N_\delta} f(x) \leq f(x) - 2\varepsilon \), meaning that we can find a sequence \( x_n \to x \) such that \( f(x_n) \leq f(x) - \varepsilon \), meaning that \( f(x) \leq f(x) - \varepsilon \), a contradiction that proves our claim.
Since $f$ is automatically lower semicontinuous at a point where $f(x) = -\infty$, it remains to consider the case $f(x) = \infty$. If $f$ is not lower semicontinuous at such a point $x$, we have $\sup_{\delta > 0} \inf_{N_\delta} f(x) = \alpha \in \mathbb{R}$. Then $\inf_{N_\delta} f(x) \leq \alpha$ for any $\delta > 0$. Let $\beta \in \mathbb{R}$, $\beta > \alpha$. We can find a sequence $x_n \to x$ such that $f(x_n) \leq \beta$. Since $S = \{z : f(z) \leq \beta\}$ is closed and $x_n \in S$, we have $x \in S$, that is, $f(x) \leq \beta < \infty$, a contradiction. 

Figure 2.1 illustrates the epigraph of a function whose function value jumps up at the point $x$, making the function not lower semicontinuous there. If we had $f(x) = \lim_{y \to x} f(y)$ instead, the function $f$ would be lower semicontinuous at $x$, although it would still be discontinuous at $x$.

![Epigraph of a function](image)

**Fig. 2.1.** Epigraph of a function.

**Corollary 2.40.** If the functions $f, g : E \to \mathbb{R} \cup \{+\infty\}$ are lower semicontinuous, then so is $f + g$.

**Proof.** We claim that

$$\{x : f(x) + g(x) > t\} = \bigcup_{\alpha \in \mathbb{R}} \left( \{x : f(x) > t - \alpha\} \cap \{x : g(x) > \alpha\} \right).$$

If $f(x) + g(x) = t + 2\varepsilon > t$ and $g(x) = \alpha + \varepsilon > \alpha$, then $f(x) = t - \alpha + \varepsilon > t - \alpha$. This proves that the set on the left-hand side is a subset of the one on the right-hand side. The reverse inclusion is trivial, and the claim is proved. Hence the set $\{x : f(x) + g(x) > t\}$ is open, since it is a union of open sets. \qed

**Theorem 2.41.** Let $f : E \to \mathbb{R} \cup \{\infty\}$ be a lower semicontinuous function defined on a metric space $E$. If $f$ has a nonempty compact sublevel set,

$$l_\alpha(f) := \{x \in E : f(x) \leq \alpha\},$$

then $f$ achieves its global minimum on $E$. 
Proof. Let \( \{x_n\} \) be a minimizing sequence for \( f \), that is,
\[
f(x_n) \downarrow \inf \{f(x) : x \in E\} =: \inf_{E} f.
\]
Clearly, there exists an integer \( N \) such that \( x_n \in l_{\alpha}(f) \) for all \( n \geq N \). Since \( l_{\alpha}(f) \) is compact, \( \{x_n\}_{N}^{\infty} \) has a convergent subsequence \( x_{n_k} \to x^* \in l_{\alpha}(f) \).
Since \( f \) is lower semicontinuous at \( x^* \), we have
\[
f(x^*) \leq \lim_{n \to \infty} f(x_n) = \inf_{E} f(x).
\]
This means that \( f(x^*) = \inf_{E} f \), that is, \( f \) achieves its minimum on \( E \) at the point \( x^* \). \( \square \)

We remark that the second proof of Theorem 2.2 can be extended without any changes to give an alternative proof of this theorem.

The following extension of Theorem 2.2 follows immediately.

**Corollary 2.42.** A lower semicontinuous function \( f : K \to \mathbb{R} \) on a compact metric space \( K \) achieves its global minimum on \( K \).

**Corollary 2.43.** Let \( f : D \to \mathbb{R} \) be a lower semicontinuous function defined on a topological space \( D \).

(a) If \( D \) is compact, or
(b) \( D \) is a subset of a finite-dimensional normed vector space \( E \) and \( f \) is coercive,

then \( f \) achieves a global minimum on \( D \).

**Proof.** In either case, all sublevel sets of \( f \) are compact. In (ii), this follows from the fact that the sublevel sets of \( f \) are closed and bounded, hence compact. \( \square \)

We note that Theorem 2.2 follows immediately from part (i) of this corollary.

### 2.8 Exercises

1. (a) Show that for all values of \( a \), the function \( f(x, y) = x^3 - 3axy + y^3 \) has no global minimizers or global maximizers.

(b) For each value of \( a \), find all the critical point(s) of \( f \) and determine their nature, that is, determine whether each critical point is a local minimum, local maximum, or saddle point.

2. Consider the function \( f(x, y) = e^{x^2+y^2} - x^2 - 2y^2 \).

(a) Find the critical points of \( f \).

(b) Find the local (and global) minima and maxima of \( f \) as well as its saddle points.
3. Find the critical points of the function \( f(x, y, z) = xyz e^{-x-y-z} \), and determine their nature.

4. Solve the geometric programming problem

\[
\min_{t_1 > 0, t_2 > 0} \frac{1}{t_1 t_2} + t_1 + t_2.
\]

5. Consider the function

\[
f(x, y, z) = 2x^2 + xy + y^2 + yz + z^2 - 6x - 7y - 8z + 9.
\]

(a) Using the first-order necessary conditions, find a critical point of \( f \).
(b) Verify that the point found in (a) is a local minimum of \( f \) by verifying the second-order sufficient conditions.
(c) Prove that the point is a global minimum of \( f \).

6. Let \( f : \mathbb{R}^n \to \mathbb{R} \) be a Gâteaux differentiable function. If \( \lim_{\|x\| \to \infty} \frac{f(x)}{\|x\|} = \infty \), then show that the gradient function \( \nabla f(x) \) is onto, that is, given \( u \in \mathbb{R}^n \), there exists a point \( x \) such that \( \nabla f(x) = u \).

Hint: Consider functions \( f(x) - \langle c, x \rangle \) that are bounded from below by choosing suitable \( c \in \mathbb{R}^n \).

7. A strong minimizer of a function \( f \) is a point \( x_0 \) satisfying the condition \( f(x_0) = \inf f > -\infty \) and \( x_n \to x_0 \) whenever \( f(x_n) \to \inf f \).

(a) Show that a strong minimizer is a strict minimizer.
(b) Show that a strict minimizer is not necessarily a strong minimizer.
(c) Show that \( x_0 \) is a strong minimizer if and only if \( \text{diam}(S(f, \epsilon)) \searrow 0 \) as \( \epsilon \to 0 \), where \( S(f, \epsilon) = \{x : f(x) \leq \inf f + \epsilon\} \).

8. Consider the following problems.

(a) Let \( f : \mathbb{R} \to \mathbb{R} \) be a function with a continuous derivative. Show that if \( f \) has a local minimizer that is not a global minimizer, then it must have another critical point.
(b) Contrast this with the function \( f : \mathbb{R}^2 \to \mathbb{R} \) given by \( f(x, y) = e^{3y} - 3xe^y + x^3 \). Show that \( f \) has a unique critical point that is a local minimizer but not a global one. Show that the polynomial \( g(x, y) = x^2(1 + y^3) + y^2 \) also has a unique local minimizer that is not a global one.
(c) Show that the polynomial \( g(x, y) = (xy - x - 1)^2 + (x^2 - 1)^2 \) has two local minimizers.

Parts (b) and (c) seem counterintuitive; plotting their graphs using computer software should be helpful in revealing their unusual properties.

9. Consider the function \( f(x, y) = x^2 - \alpha xy^2 + 2y^4 \). Show that for all parameter values except two, the origin \((0, 0)\) is the only critical point of \( f \).

(a) Find the exceptional \( \alpha \)'s and show that \( f \) has infinitely many critical points for these \( \alpha \) values. Determine the nature of these critical points.
(b) Consider the values of \( \alpha \)'s for which the origin is the only critical point. For each \( \alpha \), determine the nature of the critical point. Show
that in some cases, the origin is a local minimum, but in other cases, it is a saddle point.

(c) Show that even when the origin is a saddle point, \((0,0)\) is a local strict minimizer of \(f\) on every line passing through the origin. In fact, show that, except for one line, the function \(g(t) := f(td)\) satisfies \(g'(0) = 0\) and \(g''(0) > 0\).

10. Consider the quadratic function \(f(x) = \frac{1}{2} \langle Ax, x \rangle + \langle c, x \rangle + a\), where \(A\) is a symmetric \(n \times n\) matrix. If \(f\) is bounded from below on \(\mathbb{R}^n\), show that \(A\) is positive semidefinite, and that \(f\) achieves its minimum on \(\mathbb{R}^n\).

\(\text{Hint: Diagonalize } A\).

11. Let \(f : C \to \mathbb{R}\) be a twice Fréchet differentiable function on an open set in \(\mathbb{R}^n\). Suppose that \(\Delta f(x) := \sum_{i=1}^n \frac{\partial^2 f(x)}{\partial x_i^2} = 0\) for all \(x \in C\), that is, \(f\) is a harmonic function. If \(p\) is a critical point of \(f\), that is, \(\nabla f(p) = 0\), and the Hessian \(Hf(p)\) is not identically zero, then \(p\) must be a saddle point of \(f\).

12. Sylvester’s theorem, Theorem 2.25 on page 43, states that a symmetric matrix \(A\) is positive definite if and only if all the leading principal minors of \(A\) are positive. The purpose of this problem is to give an elegant proof of this result using optimization techniques.

Let \(A\) be an \((n+1) \times (n+1)\) symmetric matrix in the form \(A = \begin{bmatrix} B & b \\ b^T & c \end{bmatrix}\), where \(B\) is a positive definite \(n \times n\) matrix, \(b \in \mathbb{R}^n\), and \(c \in \mathbb{R}\).

(a) Consider the quadratic function \(p(x) := \langle x^T, 1 \rangle \begin{bmatrix} B & b \\ b^T & c \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} = \langle Bx, x \rangle + 2\langle b, x \rangle + c\) on \(\mathbb{R}^n\). Show that the point \(x^* = -B^{-1}b\) is the unique global minimizer of \(p\) on \(\mathbb{R}^n\), and \(p(x^*) = c - \langle B^{-1}b, b \rangle\). Thus, \(p\) is positive on \(\mathbb{R}^n\) if and only if \(c - \langle B^{-1}b, b \rangle > 0\).

(b) Show that \(\det A = \det B : (c - \langle B^{-1}b, b \rangle)\).

\(\text{Hint: Find a suitable vector } d \in \mathbb{R}^n\) such that

\[
\begin{bmatrix}
I & 0 \\
-d^T & 1
\end{bmatrix}
\begin{bmatrix}
B & b \\
b^T & c
\end{bmatrix}
\begin{bmatrix}
I \\
d^T
\end{bmatrix}
= \begin{bmatrix}
B & 0 \\
0 & c - b^T B^{-1}b
\end{bmatrix}.
\]

This is related to the notion of Schur complement in linear algebra.

(c) Prove Sylvester’s theorem by induction on the dimension of \(A\) using parts (a) and (b).

13. The purpose of this problem is to demonstrate that the behavior of a smooth function around a regular point \(x\) is determined by its derivative \(Df(x) \neq 0\). Thus, it is a first-order version of Morse’s lemma.

Let \(f\) be \(C^k\) in a neighborhood of the origin in \(\mathbb{R}^n\), and \(f(0) = 0\). Suppose that 0 is a regular point, that is, \(l := \nabla f(0) \neq 0\). Let \(x_0 \in \mathbb{R}^n\) such \(l(x_0) = 1\).

(a) Show that the linear map \(T : \mathbb{R}^n \to (\operatorname{Ker} l) \times \mathbb{R}\) defined by
Let \( T(x) := (x - l(x)x_0, l(x)) \) is one-to-one and onto, thus an isomorphism between \( \mathbb{R}^n \) and \((\text{Ker } l) \times \mathbb{R}\).

(b) Show that the \( C^k \) map

\[
\psi(x) := (x - l(x)x_0, f(x))
\]

has derivative \( D\psi(0) = T \). Conclude using the inverse function theorem that \( \psi \) is a \( C^k \) diffeomorphism between a neighborhood \( U \) of 0 and its image \( \psi(U) \).

(c) Define the \( C^k \) diffeomorphism

\[
\varphi := T^{-1} \circ \psi,
\]

so that \( \psi = T \circ \varphi \). Show that the equation \( \psi(x) = T(\varphi(x)) \) gives the sought-after formula

\[
f(x) = l(\varphi(x)) \quad \text{for all } x \in U.
\]

We remark that this result is valid in Banach spaces, since the inverse function theorem holds in that setting.

14. Let \( f : U \to \mathbb{R} \) be a \( C^2 \) function on an open set \( U \subseteq \mathbb{R}^n \). Show that the nondegenerate critical points of \( f \) are isolated: if \( x_0 \in U \) is a nondegenerate critical point of \( f \), then there exists an open neighborhood \( V \ni x_0 \) such that \( x_0 \) is the only critical point of \( f \) on \( V \).

**Hint:** The inverse function theorem may be helpful.

15. (Jordan and von Neumann [149]) If \( f \) is a quadratic form, that is, \( f(x) = \langle Ax, x \rangle \), where \( A \) is a symmetric \( n \times n \) matrix, then \( f \) satisfies the properties

(i) \( f(x + y) + f(x - y) = 2f(x) + 2f(y) \) for all \( x, y \in \mathbb{R}^n \),

(ii) \( f \) is continuous.

The property (i) is called the parallelogram law. These two properties characterize quadratic forms, even in more general spaces than \( \mathbb{R}^n \) [149, 104] and [240], pp. 275–276.

For a function \( f : \mathbb{R}^n \to \mathbb{R} \), define

\[
B(x, y) := \frac{1}{4} (f(x + y) - f(x - y)).
\]

(Note that if \( f \) is a quadratic form, then \( B(x, y) = \langle Ax, y \rangle \).)

Parts (a)–(c) below prove that if \( f \) satisfies only (i), then \( B \) is symmetric, \( B(x, y) = B(y, x) \), and additive in each variable, \( B(x + y, z) = B(x, z) + B(y, z) \).

(a) Show that \( f(0) = 0, f(-x) = x \), and \( f(2x) = 4f(x) \). Use these to show that \( B \) is symmetric and \( B(x, x) = f(x) \).
(b) Show that
\[
8B(x, z) + 8B(y, z) = 2f(x + z) + 2f(y + z) - 2f(x - z) - 2f(y - z)
= f(x + y + 2z) + f(x - y) - f(x + y - 2z)
- f(x - y)
= 4B(x + y, 2z).
\]

Consequently,
\[
B(x, z) + B(y, z) = \frac{1}{2} B(x + y, 2z). \tag{2.4}
\]

(c) Show directly from the definition of \(B\) that \(B(0, z) = B(z, 0) = 0\), and use it and (2.4) to prove that \(B(x, z) = B(x, 2z)/2\). Then show that this and (2.4) give
\[
B(x + y, z) = B(x, z) + B(y, z),
\]
that is, the function \(B(x, y)\) is additive in the first variable, and similarly in the second variable.

Now, use property (ii) to prove that

(d) \(B\) is homogeneous in each variable, that is,
\[
B(tx, y) = tB(x, y), \quad B(x, ty) = tB(x, y),
\]
for all \(x, y \in \mathbb{R}^n\), and for all \(t \in \mathbb{R}\).

\textit{Hint:} Use (c) to show that \(B(nx, y) = nB(x, y)\) for all integers \(n\). Next, if \(t = m/n\) is a rational number, define \(z := tx = mx/n\). Then \(nz = mx\), and \(nB(z, y) = mB(x, y)\) or \(B(tx, y) = tB(x, y)\). Finally, use continuity of \(f\) to show that \(B(tx, y) = tB(x, y)\) for all \(t \in \mathbb{R}\).

(e) For each fixed \(y\), the function \(x \mapsto B(x, y)\) is linear. Show that there exists \(l(y) \in \mathbb{R}^n\) such that \(B(x, y) = \langle x, l(y) \rangle\). Show that \(l\) is a linear function of \(y\), and that \(l(y) = Ay\) for some \(n \times n\) matrix \(A\). Show that, without losing any generality, \(A\) may be assumed to be a symmetric matrix.

(f) Prove that a norm \(\|\cdot\|\) is Euclidean, that is, it comes from an inner product, if and only if the function \(f(x) = \|x\|^2\) satisfies the parallelogram law. This is the motivation of the paper of Jordan and von Neumann [149].

16. Let \(f : U \to \mathbb{R}^m\) be a \(C^1\) mapping on an open set \(U \subseteq \mathbb{R}^n\). Suppose that at a point \(x_0 \in U\), \(Df(x_0) : \mathbb{R}^n \to \mathbb{R}^m\) is one-to-one, so that it is an isomorphism between \(\mathbb{R}^n\) and \(L := Df(x_0)(\mathbb{R}^n)\). Assume without loss of generality that \(x_0 = 0\) and \(f(0) = 0\). The purpose of this problem is to prove that \(f(U)\) is \(C^1\) diffeomorphic to a neighborhood of the origin in \(L\), that is, there exists a local \(C^1\) diffeomorphism of \(\mathbb{R}^m\) around the origin such that \(g \circ f\) is a \(C^1\) diffeomorphism between a neighborhood of \(0 \in \mathbb{R}^n\) and a neighborhood of the origin in \(L\). (Thus, \(g\) “straightens out” the image of \(f\) around 0.)
(a) Let $M$ be a subspace of $\mathbb{R}^m$ complementary to $L$, that is, $\mathbb{R}^m = L + M$ and $L \cap M = \{0\}$. Show that the linear map $T : \mathbb{R}^m \to L \times M$, given by $T(u) = Df(0)^{-1}v + w$, where $u = v + w$, $v \in L$, $w \in M$, is a linear isomorphism, and that the $C^1$ map $\bar{f} := T \circ f : U \to L \times M$ satisfies $D\bar{f}(0)(h) = (h, 0)$. Conclude that it is enough to prove our claim for $\bar{f}$, or equivalently, to assume that $f : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^k = \mathbb{R}^{n+k}$, $f(0) = 0$, and $Df(0)(h) = (h, 0)$.

(b) Consider the map $\varphi : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n \times \mathbb{R}^k$ given by $\varphi(x, y) = f(x) + (0, y)$. Show that $D\varphi(0, 0)$ is the identity mapping on $\mathbb{R}^{n+k}$, so that $\varphi$ must be a local diffeomorphism between two neighborhoods of the origin in $\mathbb{R}^{n+k}$. Call its inverse $g$.

(c) Show that

$$(g \circ f)(x) = g(f(x)) = g(\varphi(x, 0)) = (x, 0).$$

Show that this implies that $g$ satisfies the required properties.
Foundations of Optimization
Güler, O.
2010, XVIII, 442 p., Hardcover