<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.7.2 Lagrangian Duality</td>
<td>78</td>
</tr>
<tr>
<td>3.8 Moreau-Yosida envelope and the proximal map</td>
<td>82</td>
</tr>
<tr>
<td>3.9 Orthogonally invariant functions</td>
<td>84</td>
</tr>
</tbody>
</table>
Chapter 1

Review of Fundamentals

1.1 Inner products and linear maps

Throughout, we fix an Euclidean space $E$, meaning that $E$ is a finite-dimensional real vector space endowed with an inner product $\langle \cdot, \cdot \rangle$. Recall that an inner-product on $E$ is an assignment $\langle \cdot, \cdot \rangle : E \times E \to \mathbb{R}$ satisfying the following three properties for all $x, y, z \in E$ and scalars $a, b \in \mathbb{R}$:

(Symmetry) $\langle x, y \rangle = \langle y, x \rangle$

(Bilinearity) $\langle ax + by, z \rangle = a \langle x, z \rangle + b \langle y, z \rangle$

(Positive definiteness) $\langle x, x \rangle \geq 0$ and equality $\langle x, x \rangle = 0$ holds if and only if $x = 0$.

The most familiar example is the Euclidean space of $n$-dimensional column vectors $\mathbb{R}^n$, which unless otherwise stated we always equip with the dot-product $\langle x, y \rangle := \sum_{i=1}^{n} x_i y_i$. One can equivalently write $\langle x, y \rangle = x^T y$. A basic result of linear algebra shows that all Euclidean spaces $E$ can be identified with $\mathbb{R}^n$ for some integer $n$, once an orthonormal basis is chosen. Though such a basis-specific interpretation can be useful, it is often distracting, with the indices hiding the underlying geometry. Consequently, it is often best to think coordinate-free.

The space of real $m \times n$-matrices $\mathbb{R}^{m \times n}$ furnishes another example of an Euclidean space, which we always equip with the trace product $\langle X, Y \rangle := \text{tr} X^T Y$. Some arithmetic shows the equality $\langle X, Y \rangle = \sum_{i,j} X_{ij} Y_{ij}$. Thus the trace product on $\mathbb{R}^{m \times n}$ is nothing but the usual dot-product on the matrices stretched out into long vectors. This viewpoint, however, is typically not very fruitful, and it is best to think of the trace product as a standalone object. An important Euclidean subspace of $\mathbb{R}^{n \times n}$ is the space of real symmetric $n \times n$-matrices $S^n$, along with the trace product $\langle X, Y \rangle := \text{tr} XY$. 

1
CHAPTER 1. REVIEW OF FUNDAMENTALS

For any linear mapping \( A : E \to Y \), there exists a unique linear mapping \( A^* : Y \to E \), called the adjoint, satisfying
\[
\langle Ax, y \rangle = \langle x, A^* y \rangle \quad \text{for all points } x \in E, \ y \in Y.
\]

In the most familiar case of \( E = \mathbb{R}^n \) and \( Y = \mathbb{R}^m \), the matrix representing \( A^* \) is simply the transpose of the matrix representing \( A \).

Exercise 1.1. Given a collection of real \( m \times n \) matrices \( A_1, A_2, \ldots, A_l \), define the linear mapping \( A : \mathbb{R}^{m \times n} \to \mathbb{R}^l \) by setting
\[
A(X) := (\langle A_1, X \rangle, \langle A_2, X \rangle, \ldots, \langle A_l, X \rangle).
\]
Show that the adjoint is the mapping \( A^* y = y_1 A_1 + y_2 A_2 + \ldots + y_l A_l \).

Linear mappings \( A \) between \( E \) and itself are called linear operators, and are said to be self-adjoint if equality \( A = A^* \) holds. Self-adjoint operators on \( \mathbb{R}^n \) are precisely those operators that are representable as symmetric matrices. A self-adjoint operator \( A \) is positive semi-definite, denoted \( A \succeq 0 \), whenever
\[
\langle Ax, x \rangle \geq 0 \quad \text{for all } x \in E.
\]

Similarly, a self-adjoint operator \( A \) is positive definite, denoted \( A \succ 0 \), whenever
\[
\langle Ax, x \rangle > 0 \quad \text{for all } 0 \neq x \in E.
\]

A positive semidefinite linear operator \( A \) is positive definite if and only if \( A \) is invertible.

Consider a self-adjoint operator \( A \). A number \( \lambda \) is an eigenvalue of \( X \) if there exists a vector \( 0 \neq v \in E \) satisfying \( Av = \lambda v \). Any such vector \( v \) is called an eigenvector corresponding to \( \lambda \). The Rayleigh-Ritz theorem shows that the following relation always holds:
\[
\lambda_{\min}(A) \leq \frac{\langle Au, u \rangle}{\langle u, u \rangle} \leq \lambda_{\max}(A) \quad \text{for all } u \in E \setminus \{0\},
\]
where \( \lambda_{\min}(A) \) and \( \lambda_{\max}(A) \) are the minimal and maximal eigenvalues of \( A \), respectively. Consequently, an operator \( A \) is positive semidefinite if and only \( \lambda_{\min}(A) \geq 0 \) and \( A \) is positive definite if and only \( \lambda_{\min}(A) > 0 \).

1.2 Norms

A norm on a vector space \( V \) is a function \( \| \cdot \| : V \to \mathbb{R} \) for which the following three properties hold for all point \( x, y \in V \) and scalars \( a \in \mathbb{R} \):

(Absolute homogeneity) \( \|ax\| = |a| \cdot \|x\| \)

(Triangle inequality) \( \|x + y\| \leq \|x\| + \|y\| \)
1.2. NORMS

(Positivity) Equality $\|x\| = 0$ holds if and only if $x = 0$.

The inner product in the Euclidean space $E$ always induces a norm $\|x\| := \sqrt{\langle x, x \rangle}$. Unless specified otherwise, the symbol $\|x\|$ for $x \in E$ will always denote this induced norm. For example, the dot product on $\mathbb{R}^n$ induces the usual 2-norm $\|x\|_2 = \sqrt{x_1^2 + \ldots + x_n^2}$, while the trace product on $\mathbb{R}^{m \times n}$ induces the Frobenius norm $\|X\|_F = \sqrt{\text{tr}(X^TX)}$.

Other important norms are the $l_p$ norms on $\mathbb{R}^n$: $\|x\|_p = \left\{ \begin{array}{ll} (|x_1|^p + \ldots + |x_n|^p)^{1/p} & \text{for } 1 \leq p < \infty \\ \max\{|x_1|, \ldots, |x_n|\} & \text{for } p = \infty \end{array} \right.$

The most notable of these are the $l_1$, $l_2$, and $l_\infty$ norms. For an arbitrary norm $\| \cdot \|$ on $E$, the dual norm $\| \cdot \|^*$ on $E$ is defined by $\|v\|^* := \max\{\langle v, x \rangle : \|x\| \leq 1\}$.

For $p, q \in [1, \infty]$, the $l_p$ and $l_q$ norms on $\mathbb{R}^n$ are dual to each other whenever $p^{-1} + q^{-1} = 1$. For an arbitrary norm $\| \cdot \|$ on $E$, the Cauchy-Schwarz inequality holds: $|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$.  

Exercise 1.2. Given a positive definite linear operator $A$ on $E$, show that the assignment $\langle v, w \rangle_A := \langle Av, w \rangle$ is an inner product on $E$, with the induced norm $\|v\|_A = \sqrt{\langle Av, v \rangle}$. Show that the dual norm with respect to the original inner product is $\|v\|^*_A = \|v\|_{A^{-1}} = \sqrt{\langle A^{-1}v, v \rangle}$.

All norms on $E$ are “equivalent” in the sense that any two are within a constant factor of each other. More precisely, for any two norms $\rho_1(\cdot)$ and $\rho_2(\cdot)$, there exist constants $\alpha, \beta \geq 0$ satisfying $\alpha \rho_1(x) \leq \rho_2(x) \leq \beta \rho_1(x)$ for all $x \in E$.

Case in point, for any vector $x \in \mathbb{R}^n$, the relations hold:

- $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$
- $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$
- $\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty$.

For our purposes, the term “equivalent” is a misnomer: the proportionality constants $\alpha, \beta$ strongly depend on the (often enormous) dimension of the vector space $E$. Hence measuring quantities in different norms can yield strikingly different conclusions.

Consider a linear map $A : E \to Y$, and norms $\| \cdot \|_a$ on $E$ and $\| \cdot \|_b$ on $Y$. We define the induced matrix norm $\|A\|_{a,b} := \max_{x : \|x\|_a \leq 1} \|Ax\|_b$. 
The reader should verify the inequality
\[ \|Ax\|_b \leq \|A\|_{a,b} \|x\|_a. \]

In particular, if \( \|\cdot\|_a \) and \( \|\cdot\|_b \) are the norms induced by the inner products in \( E \) and \( Y \), then the corresponding matrix norm is called the operator norm of \( A \) and will be denoted simply by \( \|A\| \). In the case \( E = Y \) and \( a = b \), we simply use the notation \( \|A\|_a \) for the induced norm.

**Exercise 1.3.** Equip \( \mathbb{R}^n \) and \( \mathbb{R}^m \) with the \( l_p \)-norms. Then for any matrix \( A \in \mathbb{R}^{m \times n} \), show the equalities
\[
\|A\|_1 = \max_{j=1,\ldots,n} \|A_{\cdot j}\|_1 \\
\|A\|_\infty = \max_{i=1,\ldots,n} \|A_{i \cdot}\|_1
\]
where \( A_{\cdot j} \) and \( A_{i \cdot} \) denote the \( j \)’th column and \( i \)’th row of \( A \), respectively.

### 1.3 Eigenvalue and singular value decompositions of matrices

The symbol \( S^n \) will denote the set of \( n \times n \) real symmetric matrices, while \( O(n) \) will denote the set of \( n \times n \) real orthogonal matrices – those satisfying \( X^T X = X X^T = I \). Any symmetric matrix \( A \in S^n \) admits an eigenvalue decomposition, meaning a factorization of the form \( A = U \Lambda U^T \) with \( U \in O(n) \) and \( \Lambda \in S^n \) a diagonal matrix. The diagonal elements of \( \Lambda \) are precisely the eigenvalues of \( A \) and the columns of \( U \) are corresponding eigenvectors.

More generally, any matrix \( A \in \mathbb{R}^{m \times n} \) admits a singular value decomposition, meaning a factorization of the form \( A = U D V^T \), where \( U \in O(m) \) and \( V \in O(n) \) are orthogonal matrices and \( D \in \mathbb{R}^{m \times n} \) is a diagonal matrix with nonnegative diagonal entries. The diagonal elements of \( D \) are uniquely defined and are called the singular values of \( A \). Supposing without loss of generality \( m \leq n \), the singular values of \( A \) are precisely the square roots of the eigenvalues of \( A A^T \). In particular, the operator norm of any matrix \( A \in \mathbb{R}^{m \times n} \) equals its maximal singular-value.

### 1.4 Point-set topology and differentiability

The symbol \( B_r(x) \) will denote an open ball of radius \( r \) around a point \( x \), namely \( B_r(x) := \{ y \in E : \|y - x\| < r \} \). The closure of a set \( Q \subseteq E \), denoted \( \text{cl}Q \), consists of all points \( x \) such that the ball \( B_\epsilon(x) \) intersects \( Q \) for all \( \epsilon > 0 \); the interior of \( Q \), written as \( \text{int}Q \), is the set of all points \( x \) such that \( Q \) contains some open ball around \( x \). We say that \( Q \) is an open set if it coincides with its interior and a closed set if it coincides with its
1.4. POINT-SET TOPOLOGY AND DIFFERENTIABILITY

Any set $Q$ in $E$ that is closed and bounded is called a **compact set**. The following classical result will be fundamentally used.

**Theorem 1.4** (Bolzano-Weierstrass). *Any sequence in a compact set $Q \subset E$ admits a subsequence converging to a point in $Q$.*

For the rest of the section, we let $E$ and $Y$ be two Euclidean spaces, and $U$ an open subset of $E$. A mapping $F : Q \to Y$, defined on a subset $Q \subset E$, is *continuous* at a point $x \in Q$ if for any sequence $x_i$ in $Q$ converging to $x$, the values $F(x_i)$ converge to $F(x)$. We say that $F$ is *continuous* if it is continuous at every point $x \in Q$. By equivalence of norms, continuity is a property that is independent of the choice of norms on $E$ and $Y$. We say that $F$ is $L$-Lipschitz continuous if

$$
\|F(y) - F(x)\| \leq L\|y - x\| \quad \text{for all } x, y \in Q.
$$

**Theorem 1.5** (Extreme value theorem). *Any continuous function $f : Q \to \mathbb{R}$ on a compact set $Q \subset E$ attains its supremum and infimum values.*

A function $f : U \to \mathbb{R}$ is *differentiable* at a point $x$ in $U$ if there exists a vector, denoted by $\nabla f(x)$, satisfying

$$
\lim_{h \to 0} \frac{f(x + h) - f(x) - \langle \nabla f(x), h \rangle}{\|h\|} = 0.
$$

Rather than carrying such fractions around, it is convenient to introduce the following notation. The symbol $o(r)$ will always stand for a term satisfying $0 = \lim_{r \downarrow 0} o(r)/r$. Then the equation above simply amounts to

$$
f(x + h) = f(x) + \langle \nabla f(x), h \rangle + o(\|h\|).
$$

The vector $\nabla f(x)$ is called the **gradient** of $f$ at $x$. In the most familiar setting $E = \mathbb{R}^n$, the gradient is simply the vector of partial derivatives

$$
\nabla f(x) = \begin{pmatrix}
\frac{\partial f(x)}{\partial x_1} \\
\frac{\partial f(x)}{\partial x_2} \\
\vdots \\
\frac{\partial f(x)}{\partial x_n}
\end{pmatrix}
$$

If the gradient mapping $x \mapsto \nabla f(x)$ is well-defined and continuous on $U$, we say that $f$ is $C^1$-smooth. We say that $f$ is $\beta$-smooth if $f$ is $C^1$-smooth and its gradient mapping $\nabla f$ is $\beta$-Lipschitz continuous.

More generally, consider a mapping $F : U \to Y$. We say that $F$ is *differentiable* at $x \in U$ if there exists a linear mapping taking $E$ to $Y$, denoted by $\nabla F(x)$, satisfying

$$
F(x + h) = F(x) + \nabla F(x)h + o(\|h\|).
$$
The linear mapping \( \nabla F(x) \) is called the Jacobian of \( F \) at \( x \). If the assignment \( x \mapsto \nabla F(x) \) is continuous, we say that \( F \) is \( C^1 \)-smooth. In the most familiar setting \( \mathbf{E} = \mathbb{R}^n \) and \( \mathbf{Y} = \mathbb{R}^m \), we can write \( F \) in terms of coordinate functions \( F(x) = (F_1(x), \ldots, F_m(x)) \), and then the Jacobian is simply

\[
\nabla F(x) = \begin{pmatrix}
\nabla F_1(x)^T \\
\nabla F_2(x)^T \\
\vdots \\
\nabla F_m(x)^T
\end{pmatrix} = \begin{pmatrix}
\frac{\partial F_1(x)}{\partial x_1} & \frac{\partial F_1(x)}{\partial x_2} & \cdots & \frac{\partial F_1(x)}{\partial x_n} \\
\frac{\partial F_2(x)}{\partial x_1} & \frac{\partial F_2(x)}{\partial x_2} & \cdots & \frac{\partial F_2(x)}{\partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial F_m(x)}{\partial x_1} & \frac{\partial F_m(x)}{\partial x_2} & \cdots & \frac{\partial F_m(x)}{\partial x_n}
\end{pmatrix}.
\]

Finally, we introduce second-order derivatives. A \( C^1 \)-smooth function \( f: U \to \mathbb{R} \) is twice differentiable at a point \( x \in U \) if the gradient map \( \nabla f: U \to \mathbf{E} \) is differentiable at \( x \). Then the Jacobian of the gradient \( \nabla(\nabla f)(x) \) is denoted by \( \nabla^2 f(x) \) and is called the Hessian of \( f \) at \( x \). Unraveling notation, the Hessian \( \nabla^2 f(x) \) is characterized by the condition

\[
\nabla f(x + h) = \nabla f(x) + \nabla^2 f(x)h + o(\|h\|).
\]

If the map \( x \mapsto \nabla^2 f(x) \) is continuous, we say that \( f \) is \( C^2 \)-smooth. If \( f \) is indeed \( C^2 \)-smooth, then a basic result of calculus shows that \( \nabla^2 f(x) \) is a self-adjoint operator.

In the standard setting \( \mathbf{E} = \mathbb{R}^n \), the Hessian is the matrix of second-order partial derivatives

\[
\nabla^2 f(x) = \begin{pmatrix}
\frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\
\frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2}
\end{pmatrix}.
\]

The matrix is symmetric, as long as it varies continuously with \( x \) in \( U \).

**Exercise 1.6.** Define the function

\[
f(x) = \frac{1}{2}(\mathcal{A}x, x) + (v, x) + c
\]

where \( \mathcal{A}: \mathbf{E} \to \mathbf{E} \) is a linear operator, \( v \) is lies in \( \mathbf{E} \), and \( c \) is a real number.

1. Show that if \( \mathcal{A} \) is replaced by the self-adjoint operator \( (\mathcal{A} + \mathcal{A}^*)/2 \), the function values \( f(x) \) remain unchanged.

2. Assuming \( \mathcal{A} \) is self-adjoint derive the equations:

\[
\nabla f(x) = \mathcal{A}x + v \quad \text{and} \quad \nabla^2 f(x) = \mathcal{A}.
\]

3. Using parts 1 and 2 describe \( \nabla f(x) \) and \( \nabla^2 f(x) \) when \( \mathcal{A} \) is not necessarily self-adjoint.
Exercise 1.7. Define the function $f(x) = \frac{1}{2}\|F(x)\|^2$, where $F: E \to Y$ is a $C^1$-smooth mapping. Prove the identity $\nabla f(x) = \nabla F(x)^*F(x)$.

Exercise 1.8. Consider a function $f: U \to \mathbb{R}$ and a linear mapping $A: Y \to E$ and define the composition $h(x) = f(Ax)$.

1. Show that if $f$ is differentiable at $Ax$, then $\nabla h(x) = A^*\nabla f(Ax)$.

2. Show that if $f$ is twice differentiable at $Ax$, then $\nabla^2 h(x) = A^*\nabla^2 f(Ax)A$.

Exercise 1.9. Consider a mapping $F(x) = G(H(x))$ where $H$ is differentiable at $x$ and $G$ is differentiable at $H(x)$. Derive the formula $\nabla F(x) = \nabla G(H(x))\nabla H(x)$.

Exercise 1.10. Define the two sets
\[ R^n_{++} := \{ x \in \mathbb{R}^n : x_i > 0 \text{ for all } i = 1, \ldots, n \}, \]
\[ S^n_{++} := \{ X \in S^n : X > 0 \}. \]
Consider the two functions $f: R^n_{++} \to \mathbb{R}$ and $F: S^n_{++} \to \mathbb{R}$ given by
\[ f(x) = -\sum_{i=1}^n \log x_i \quad \text{and} \quad F(X) = -\ln \det(X), \]
respectively. Note, from basic properties of the determinant, the equality $F(X) = f(\lambda(X))$, where we set $\lambda(X) := (\lambda_1(X), \ldots, \lambda_n(X))$.

1. Find the derivatives $\nabla f(x)$ and $\nabla^2 f(x)$ for $x \in R^n_{++}$.

2. Using the property $\text{tr}(AB) = \text{tr}(BA)$, prove $\nabla F(X) = -X^{-1}$ and $\nabla^2 F(X)[V] = X^{-1}VX^{-1}$ for any $X > 0$.

[Hint: To compute $\nabla F(X)$, justify $F(X+tV) - F(X) + t(V) = -\ln \det(I + X^{-1/2}VX^{-1/2}) + \text{tr} (X^{-1/2}VX^{-1/2})$.]

By rewriting the expression in terms of eigenvalues of $X^{-1/2}VX^{-1/2}$, deduce that the right-hand-side is $o(t)$. To compute the Hessian, observe
\[ (X + V)^{-1} = X^{-1/2} \left( I + X^{-1/2}VX^{-1/2} \right)^{-1} X^{-1/2}, \]
and then use the expansion
\[ (I + A)^{-1} = I - A + A^2 - A^3 + \ldots = I - A + O(||A||^2_{op}), \]
whenever $||A||_{op} < 1$. ]
3. Show
\[ \langle \nabla^2 F(X)[V], V \rangle = \| X^{-\frac{1}{2}} VX^{-\frac{1}{2}} \|^2_F \]
for any \( X \succ 0 \) and \( V \in S^n \). Deduce that the operator \( \nabla^2 F(X) : S^n \to S^n \) is positive definite.

1.5 Fundamental theorems of calculus & accuracy in approximation

For any two points \( x, y \in \mathbb{E} \), define the closed segment \((x, y) := \{ \lambda x + (1 - \lambda)y : \lambda \in [0, 1] \}\). The open segment \((x, y)\) is defined analogously. A set \( Q \) in \( \mathbb{E} \) is convex if for any two points \( x, y \in Q \), the entire segment \([x, y]\) is contained in \( Q \). For this entire section, we let \( U \) be an open, convex subset of \( \mathbb{E} \). Consider a \( C^1 \)-smooth function \( f : U \to \mathbb{R} \) and a point \( x \in U \). Classically, the linear function
\[ l(x; y) = f(x) + \langle \nabla f(x), y - x \rangle \]
is a best first-order approximation of \( f \) near \( x \). If \( f \) is \( C^2 \)-smooth, then the quadratic function
\[ Q(x; y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle \]
is a best second-order approximation of \( f \) near \( x \). These two functions play a fundamental role when designing and analyzing algorithms, they furnish simple linear and quadratic local models of \( f \). In this section, we aim to quantify how closely \( l(x; \cdot) \) and \( Q(x; \cdot) \) approximate \( f \). All results will follow quickly by restricting multivariate functions to line segments and then applying the fundamental theorem of calculus for univariate functions. To this end, the following observation plays a basic role.

**Exercise 1.11.** Consider a function \( f : U \to \mathbb{R} \) and two points \( x, y \in U \). Define the univariate function \( \varphi : [0, 1] \to \mathbb{R} \) given by \( \varphi(t) = f(x + t(y - x)) \) and let \( x_t := x + t(y - x) \) for any \( t \).

1. Show that if \( f \) is \( C^1 \)-smooth, then equality
\[ \varphi'(t) = \langle \nabla f(x_t), y - x \rangle \]
holds for any \( t \in (0, 1) \).

2. Show that if \( f \) is \( C^2 \)-smooth, then equality
\[ \varphi''(t) = \langle \nabla^2 f(x_t)(y - x), y - x \rangle \]
holds for any \( t \in (0, 1) \).

The fundamental theorem of calculus now takes the following form.
1.5. FUNDAMENTAL THEOREMS OF CALCULUS & ACCURACY IN APPROXIMATION

**Theorem 1.12** (Fundamental theorem of multivariate calculus). Consider a $C^1$-smooth function $f : U \to \mathbb{R}$ and two points $x, y \in U$. Then equality

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt,$$

holds.

**Proof.** Define the univariate function $\varphi(t) = f(x + t(y - x))$. The fundamental theorem of calculus yields the relation

$$\varphi(1) - \varphi(0) = \int_0^1 \varphi'(t) dt.$$

Taking into account Exercise 1.11, the result follows. □

The following corollary precisely quantifies the gap between $f(y)$ and its linear and quadratic models, $l(x; y)$ and $Q(x; y)$.

**Corollary 1.13** (Accuracy in approximation). Consider a $C^1$-smooth function $f : U \to \mathbb{R}$ and two points $x, y \in U$. Then we have

$$f(y) = l(x; y) + \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt.$$

If $f$ is $C^2$-smooth, then the equation holds:

$$f(y) = Q(x; y) + \int_0^1 \int_0^t \langle (\nabla^2 f(x + s(y - x)) - \nabla^2 f(x))(y - x), y - x \rangle ds dt.$$

**Proof.** The first equation is immediate from Theorem 1.12. To see the second equation, define the function $\varphi(t) = f(x + t(y - x))$. Then applying the fundamental theorem of calculus twice yields

$$\varphi(1) - \varphi(0) = \int_0^1 \varphi'(t) dt = \int_0^1 (\varphi'(0) + \int_0^t \varphi''(s) ds) dt$$

$$= \varphi'(0) + \frac{1}{2} \varphi''(0) + \int_0^1 \int_0^t \varphi''(s) - \varphi''(0) ds dt.$$

Appealing to Exercise 1.11, the result follows. □

Recall that if $f$ is differentiable at $x$, then the relation holds:

$$\lim_{y \to x} \frac{f(y) - l(x; y)}{\|y - x\|} = 0.$$

An immediate consequence of Corollary 1.13 is that if $f$ is $C^1$-smooth then the equation above is stable under perturbations of the base point $x$: for any point $\bar{x} \in U$ we have

$$\lim_{x,y \to \bar{x}} \frac{f(y) - l(x; y)}{\|y - x\|} = 0.$$
Similarly if $f$ is $C^2$-smooth, then
\[
\lim_{x,y \to \bar{x}} \frac{f(y) - Q(x;y)}{\|y - x\|^2} = 0.
\]

When the mappings $\nabla f$ and $\nabla^2 f$ are Lipschitz continuous, one has even greater control on the accuracy of approximation, in essence passing from little-o terms to big-O terms.

**Corollary 1.14** (Accuracy in approximation under Lipschitz conditions). Given any $\beta$-smooth function $f: U \to \mathbb{R}$, for any points $x, y \in U$ the inequality
\[
\left| f(y) - l(x;y) \right| \leq \frac{\beta}{2} \|y - x\|^2
\]
holds.

If $f$ is $C^2$-smooth with $M$-Lipschitz Hessian, then
\[
\left| f(y) - Q(x;y) \right| \leq \frac{M}{6} \|y - x\|^3.
\]

It is now straightforward to extend the results in this section to mappings $F: U \to \mathbb{R}^m$. Given a curve $\gamma: \mathbb{R} \to \mathbb{R}^m$, we define the integral $\int_0^1 \gamma(t) \, dt = \left(\int_0^1 \gamma_1(t) \, dt, \ldots, \int_0^1 \gamma_m(t) \, dt\right)$, where $\gamma_i$ are the coordinate functions of $\gamma$. The main observation is that whenever $\gamma_i$ are integrable, the inequality
\[
\left\| \int_0^1 \gamma(t) \, dt \right\| \leq \int_0^1 \|\gamma(t)\| \, dt
\]
holds.

To see this, define $w = \int_0^1 \gamma(t) \, dt$ and simply observe
\[
\|w\|^2 = \int_0^1 \langle \gamma(t), w \rangle \, dt \leq \|w\| \int_0^1 \|\gamma(t)\| \, dt.
\]

**Exercise 1.15.** Consider a $C^1$-smooth mapping $F: U \to \mathbb{R}^m$ and two points $x, y \in U$. Derive the equations
\[
F(y) - F(x) = \int_0^1 \nabla F(x + t(y - x))(y - x) \, dt.
\]

In particular, consider a $C^1$-smooth mapping $F: U \to \mathbf{Y}$, where $\mathbf{Y}$ is some Euclidean space, and a point $\bar{x} \in U$. Choosing an orthonormal basis for $\mathbf{Y}$ and applying Exercise 1.15, we obtain the relation
\[
\lim_{x,y \to \bar{x}} \frac{F(y) - F(x) - \nabla F(x)(y - x)}{\|y - x\|} = 0.
\]

Supposing that $F$ is $\beta$-smooth, the stronger inequality holds:
\[
\|F(y) - F(x) - \nabla F(x)(y - x)\| \leq \frac{\beta}{2} \|y - x\|^2.
\]
Exercise 1.16. Show that a $C^1$-smooth mapping $F: U \to Y$ is $L$-Lipschitz continuous if and only if $\|\nabla F(x)\| \leq L$ for all $x \in U$. 
Chapter 2

Smooth minimization

In this chapter, we consider the problem of minimizing a smooth function on a Euclidean space $\mathbb{E}$. Such problems are ubiquitous in computation mathematics and applied sciences.

2.1 Optimality conditions: Smooth Unconstrained

We begin the formal development with a classical discussion of optimality conditions. To this end, consider the problem

$$\min_{x \in \mathbb{E}} f(x)$$

where $f : \mathbb{E} \to \mathbb{R}$ is a $C^1$-smooth function. Without any additional assumptions on $f$, finding a global minimizer of the problem is a hopeless task. Instead, we focus on finding a local minimizer: a point $x$ for which there exists a convex neighborhood $U$ of $x$ such that $f(x) \leq f(y)$ for all $y \in U$. After all, gradients and Hessians provide only local information on the function.

When encountering an optimization problem, such as above, one faces two immediate tasks. First, design an algorithm that solves the problem. That is, develop a rule for going from one point $x_k$ to the next $x_{k+1}$ by using computable quantities (e.g. function values, gradients, Hessians) so that the limit points of the iterates solve the problem. The second task is easier: given a test point $x$, either verify that $x$ solves the problem or exhibit a direction along which points with strictly better function value can be found. Though the verification goal seems modest at first, it always serves as the starting point for algorithm design.

Observe that naively checking if $x$ is a local minimizer of $f$ from the very definition requires evaluation of $f$ at every point near $x$, an impossible task. We now derive a verifiable necessary condition for local optimality.

**Theorem 2.1. (First-order necessary conditions)** Suppose that $x$ is a local minimizer of a function $f : U \to \mathbb{R}$. If $f$ is differentiable at $x$, then equality $\nabla f(x) = 0$ holds.
CHAPTER 2. SMOOTH MINIMIZATION

Proof. Set \( v := -\nabla f(x) \). Then for all small \( t > 0 \), we deduce from the definition of derivative
\[
0 \leq \frac{f(x + tv) - f(x)}{t} = -\|\nabla f(x)\|^2 + o(t) / t.
\]
Letting \( t \) tend to zero, we obtain \( \nabla f(x) = 0 \), as claimed. \( \Box \)

A point \( x \in U \) is a critical point for a \( C^1 \)-smooth function \( f : U \to \mathbb{R} \) if equality \( \nabla f(x) = 0 \) holds. Theorem 2.1 shows that all local minimizers of \( f \) are critical points. In general, even finding local minimizers is too ambitious, and we will for the most part settle for critical points.

To obtain verifiable sufficient conditions for optimality, higher order derivatives are required.

**Theorem 2.2. (Second-order conditions)**
Consider a \( C^2 \)-smooth function \( f : U \to \mathbb{R} \) and fix a point \( x \in U \). Then the following are true.

1. (Necessary conditions) If \( x \in U \) is a local minimizer of \( f \), then
   \[
   \nabla f(x) = 0 \quad \text{and} \quad \nabla^2 f(x) \succeq 0.
   \]
2. (Sufficient conditions) If the relations
   \[
   \nabla f(x) = 0 \quad \text{and} \quad \nabla^2 f(x) \succ 0
   \]
   hold, then \( x \) is a local minimizer of \( f \). More precisely,
   \[
   \liminf_{y \to x} \frac{f(y) - f(x)}{\frac{1}{2} \|y - x\|^2} \geq \lambda_{\text{min}}(\nabla^2 f(x)).
   \]

Proof. Suppose first that \( x \) is a local minimizer of \( f \). Then Theorem 2.1 guarantees \( \nabla f(x) = 0 \). Consider an arbitrary vector \( v \in \mathbb{R} \). Then for all \( t > 0 \), we deduce from a second-order expansion
\[
0 \leq \frac{f(x + tv) - f(x)}{\frac{1}{2} t^2} = \langle \nabla^2 f(x)v, v \rangle + o(t^2) / t^2.
\]
Letting \( t \) tend to zero, we conclude \( \langle \nabla^2 f(x)v, v \rangle \geq 0 \) for all \( v \in \mathbb{R} \), as claimed.

Suppose \( \nabla f(x) = 0 \) and \( \nabla^2 f(x) \succ 0 \). Let \( \epsilon > 0 \) be such that \( B_\epsilon(x) \subset U \). Then for points \( y \to x \), we have from a second-order expansion
\[
\frac{f(y) - f(x)}{\frac{1}{2} \|y - x\|^2} = \left\langle \nabla^2 f(x) \left( \frac{y - x}{\|y - x\|} \right), \frac{y - x}{\|y - x\|} \right\rangle + o(\|y - x\|^2) / \|y - x\|^2
\]
\[
\geq \lambda_{\text{min}}(\nabla^2 f(x)) + o(\|y - x\|^2) / \|y - x\|^2.
\]
Letting \( y \) tend to \( x \), the result follows. \( \Box \)
2.2 CONVEXITY, A FIRST LOOK

The reader may be misled into believing that the role of the necessary conditions and the sufficient conditions for optimality (Theorem 2.2) is merely to determine whether a putative point \( x \) is a local minimizer of a smooth function \( f \). Such a viewpoint is far too limited.

Necessary conditions serve as the basis for algorithm design. If necessary conditions for optimality fail at a point, then there must be some point nearby with a strictly smaller objective value. A method for discovering such a point is a first step for designing algorithms.

Sufficient conditions play an entirely different role. In Section 2.2, we will see that sufficient conditions for optimality at a point \( x \) guarantee that the function \( f \) is strongly convex on a neighborhood of \( x \). Strong convexity, in turn, is essential for establishing rapid convergence of numerical methods.

2.2 Convexity, a first look

Finding a global minimizer of a general smooth function \( f : E \to R \) is a hopeless task, and one must settle for local minimizers or even critical points. This is quite natural since gradients and Hessians only provide local information on the function. However, there is a class of smooth functions, prevalent in applications, whose gradients provide global information. This is the class of convex functions – the main setting for the book. This section provides a short, and limited, introduction to the topic to facilitate algorithmic discussion. Later sections of the book explore convexity in much greater detail.

**Definition 2.3 (Convexity).** A function \( f : U \to (-\infty, +\infty] \) is convex if the inequality

\[
 f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)
\]

holds for all points \( x, y \in U \) and real numbers \( \lambda \in [0, 1] \).

In other words, a function \( f \) is convex if any secant line joining two points in the graph of the function lies above the graph. This is the content of the following exercise.

**Exercise 2.4.** Show that a function \( f : U \to (-\infty, +\infty] \) is convex if and only if the epigraph

\[
 epi \; f := \{(x, r) \in U \times R : f(x) \leq r\}
\]

is a convex subset of \( E \times R \).

**Exercise 2.5.** Show that \( f : U \to (-\infty, +\infty] \) is convex if and only if the inequality

\[
 f \left( \sum_{i=1}^{k} \lambda_i x_i \right) \leq \sum_{i=1}^{k} \lambda_i f(x_i),
\]
CHAPTER 2. SMOOTH MINIMIZATION

holds for all integers \( k \in \mathbb{N} \), all points \( x_1, \ldots, x_k \in U \), and all real \( \lambda_i \geq 0 \) with \( \sum_{i=1}^k \lambda_i = 1 \).

Convexity is preserved under a variety of operations. Point-wise maximum is an important example.

Exercise 2.6. Consider an arbitrary set \( T \) and a family of convex functions \( f_t : U \to (-\infty, +\infty] \) for \( t \in T \). Show that the function \( f(x) := \sup_{t \in T} f_t(x) \) is convex.

Convexity of smooth functions can be characterized entirely in terms of derivatives.

Theorem 2.7 (Differential characterizations of convexity). The following are equivalent for a \( C^1 \)-smooth function \( f : U \to \mathbb{R} \).

(a) (convexity) \( f \) is convex.

(b) (gradient inequality) \( f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \) for all \( x, y \in U \).

(c) (monotonicity) \( \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0 \) for all \( x, y \in U \).

If \( f \) is \( C^2 \)-smooth, then the following property can be added to the list:

(d) The relation \( \nabla^2 f(x) \succeq 0 \) holds for all \( x \in U \).

Proof. Assume \( \text{(a)} \) holds, and fix two points \( x \) and \( y \). For any \( t \in (0,1) \), convexity implies

\[
f(x + t(y - x)) = f(ty + (1 - t)x) \leq tf(y) + (1 - t)f(x),
\]

while the definition of the derivative yields

\[
f(x + t(y - x)) = f(x) + t\langle \nabla f(x), y - x \rangle + o(t).
\]

Combining the two expressions, canceling \( f(x) \) from both sides, and dividing by \( t \) yields the relation

\[
f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle + o(t)/t.
\]

Letting \( t \) tend to zero, we obtain property \( \text{(b)} \).

Suppose now that \( \text{(b)} \) holds. Then for any \( x, y \in U \), appealing to the gradient inequality, we deduce

\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle
\]

and

\[
f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle.
\]

Adding the two inequalities yields \( \text{(c)} \).
Finally, suppose \( (c) \) holds. Define the function \( \varphi(t) := f(x + t(y - x)) \) and set \( x_t := x + t(y - x) \). Then monotonicity shows that for any real numbers \( t, s \in [0, 1] \) with \( t > s \) the inequality holds:

\[
\varphi'(t) - \varphi'(s) = \langle \nabla f(x_t), y - x \rangle - \langle \nabla f(x_s), y - x \rangle = \frac{1}{t - s} \langle \nabla f(x_t) - \nabla f(x_s), x_t - x_s \rangle \geq 0.
\]

Thus the derivative \( \varphi' \) is nondecreasing, and hence for any \( x, y \in U \), we have

\[
f(y) = \varphi(1) = \varphi(0) + \int_0^1 \varphi'(r) \, dr \geq \varphi(0) + \varphi'(0) = f(x) + \langle \nabla f(x), y - x \rangle.
\]

Some thought now shows that \( f \) admits the representation

\[
f(y) = \sup_{x \in U} \{ f(x) + \langle \nabla f(x), y - x \rangle \}
\]

for any \( y \in U \). Since a pointwise supremum of an arbitrary collection of convex functions is convex (Exercise \[2.6\]), we deduce that \( f \) is convex, establishing \( (a) \).

Suppose now that \( f \) is \( C^2 \)-smooth. Then for any fixed \( x \in U \) and \( h \in E \), and all small \( t > 0 \), property \( (b) \) implies

\[
f(x) + t \langle \nabla f(x), h \rangle \leq f(x + th) = f(x) + t \langle \nabla f(x), h \rangle + \frac{t^2}{2} \langle \nabla^2 f(x) h, h \rangle + o(t^2).
\]

Canceling out like terms, dividing by \( t^2 \), and letting \( t \) tend to zero we deduce \( \langle \nabla^2 f(x) h, h \rangle \geq 0 \) for all \( h \in E \). Hence \( (d) \) holds. Conversely, suppose \( (d) \) holds. Then Corollary \[1.13\] immediately implies for all \( x, y \in E \) the inequality

\[
f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \int_0^1 \int_0^t \langle \nabla^2 f(x + s(y - x))(y - x), y - x \rangle \, ds \, dt \geq 0.
\]

Hence \( (b) \) holds, and the proof is complete. \( \square \)

**Exercise 2.8.** Show that the functions \( f \) and \( F \) in Exercise \[1.10\] are convex.

**Exercise 2.9.** Consider a \( C^1 \)-smooth function \( f: \mathbb{R}^n \to \mathbb{R} \). Prove that each condition below holding for all points \( x, y \in \mathbb{R}^n \) is equivalent to \( f \) being \( \beta \)-smooth and convex.

1. \( 0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{\beta}{2} \| x - y \|^2 \)
2. \( f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\beta} \| \nabla f(x) - \nabla f(y) \|^2 \leq f(y) \)
3. \( \frac{1}{\beta} \| \nabla f(x) - \nabla f(y) \|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \)
4. $0 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \beta \|x - y\|^2$

**[Hint]** Suppose first that $f$ is convex and $\beta$-smooth. Then 1 is immediate. Suppose now 1 holds and define the function $\phi(y) = f(y) - \langle \nabla f(x), y \rangle$. Show using 1 that

$$\phi(x) = \min \phi \leq \phi(y) - \frac{1}{\beta} \nabla \phi(y) \leq \phi(y) - \frac{1}{2\beta} \|\nabla \phi(y)\|^2.$$ 

Deduce the property 2. To deduce 3 from 2, add two copies of 2 with $x$ and $y$ reversed. Next applying Cauchy-Schwartz to 3 immediately implies that $f$ is $\beta$-smooth and convex. Finally, show that 1 implies 4 by adding two copies of 1 with $x$ and $y$ reversed. Conversely, rewriting 4 deduce that the gradient of the function $\phi(x) = -f(x) + \frac{\alpha}{2} \|x\|^2$ is monotone and therefore that $\phi$ is convex. Rewriting the gradient inequality for $\phi$ arrive at 1.]

Global minimality, local minimality, and criticality are equivalent notions for smooth convex functions.

**Corollary 2.10** (Minimizers of convex functions). For any $C^1$-smooth convex function $f : U \to \mathbb{R}$ and a point $x \in U$, the following are equivalent.

(a) $x$ is a global minimizer of $f$,

(b) $x$ is a local minimizer of $f$,

(c) $x$ is a critical point of $f$.

**Proof.** The implications $(a) \Rightarrow (b) \Rightarrow (c)$ are immediate. The implication $(c) \Rightarrow (a)$ follows from the gradient inequality in Theorem 2.7. □

**Exercise 2.11.** Consider a $C^1$-smooth convex function $f : E \to \mathbb{R}$. Fix a linear subspace $L \subset E$ and a point $x_0 \in E$. Show that $x \in L$ minimizes the restriction $f_L : L \to \mathbb{R}$ if and only if the gradient $\nabla f(x)$ is orthogonal to $L$.

Strengthening the gradient inequality in Theorem 2.7 in a natural ways yields an important subclass of convex functions. These are the functions for which numerical methods have a chance of converging at least linearly.

**Definition 2.12** (Strong convexity). We say that a $C^1$-smooth function $f : U \to \mathbb{R}$ is $\alpha$-strongly convex (with $\alpha \geq 0$) if the inequality

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|^2$$

holds for all $x, y \in U$.

Figure 2.1 illustrates geometrically a $\beta$-smooth and $\alpha$-convex function.

In particular, a very useful property to remember is that if $x$ is a minimizer of an $\alpha$-strongly convex $C^1$-smooth function $f$, then for all $y$ it holds:

$$f(y) \geq f(x) + \frac{\alpha}{2} \|y - x\|^2.$$
2.3. RATES OF CONVERGENCE

Figure 2.1: Illustration of a $\beta$-smooth and $\alpha$-strongly convex function $f$, where $Q_x(y) := f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|y - x\|^2$ is an upper model based at $x$ and $q_x(y) := f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|^2$ is a lower model based at $x$. The fraction $Q := \beta/\alpha$ is often called the condition number of $f$.

Exercise 2.13. Show that a $C^1$-smooth function $f : U \to \mathbb{R}$ is $\alpha$-strongly convex if and only if the function $g(x) = f(x) - \frac{\alpha}{2} \|x\|^2$ is convex.

The following is an analogue of Theorem 2.7 for strongly convex functions.

**Theorem 2.14 (Characterization of strong convexity).** The following properties are equivalent for any $C^1$-smooth function $f : U \to \mathbb{R}$ and any constant $\alpha \geq 0$.

(a) $f$ is $\alpha$-convex.

(b) The inequality $\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \alpha \|y - x\|^2$ holds for all $x, y \in U$.

If $f$ is $C^2$-smooth, then the following property can be added to the list:

(c) The relation $\nabla^2 f(x) \succeq \alpha I$ holds for all $x \in U$.

**Proof.** By Exercise 2.13, property (a) holds if and only if $f - \frac{\alpha}{2} \| \cdot \|^2$ is convex, which by Theorem 2.7 is equivalent to (b). Suppose now that $f$ is $C^2$-smooth. Theorem 2.7 then shows that $f - \frac{\alpha}{2} \| \cdot \|^2$ is convex if and only if (c) holds.

2.3 Rates of convergence

In the next section, we will begin discussing algorithms. A theoretically sound comparison of numerical methods relies on precise rates of progress in the iterates. For example, we will predominantly be interested in how fast the quantities $f(x_k) - \inf f$, $\nabla f(x_k)$, or $\|x_k - x^*\|$ tend to zero as a function
of the counter $k$. In this section, we review three types of convergence rates that we will encounter.

Fix a sequence of real numbers $a_k > 0$ with $a_k \to 0$.

1. We will say that $a_k$ converges sublinearly if there exist constants $c, q > 0$ satisfying
   \[ a_k \leq \frac{c}{k^q} \quad \text{for all } k. \]
   Larger $q$ and smaller $c$ indicates faster rates of convergence. In particular, given a target precision $\varepsilon > 0$, the inequality $a_k \leq \varepsilon$ holds for every $k \geq \left(\frac{c}{\varepsilon}\right)^{1/q}$. The importance of the value of $c$ should not be discounted; the convergence guarantee depends strongly on this value.

2. The sequence $a_k$ is said to converge linearly if there exist constants $c > 0$ and $q \in (0, 1]$ satisfying
   \[ a_k \leq c \cdot (1 - q)^k \quad \text{for all } k. \]
   In this case, we call $1 - q$ the linear rate of convergence. Fix a target accuracy $\varepsilon > 0$, and let us see how large $k$ needs to be to ensure $a_k \leq \varepsilon$. To this end, taking logs we get
   \[ c \cdot (1 - q)^k \leq \varepsilon \iff k \geq \frac{-1}{\ln(1 - q)} \ln \left(\frac{c}{\varepsilon}\right). \]
   Taking into account the inequality $\ln(1 - q) \leq -q$, we deduce that the inequality $a_k \leq \varepsilon$ holds for every $k \geq \frac{1}{q} \ln\left(\frac{\varepsilon}{c}\right)$. The dependence on $q$ is strong, while the dependence on $c$ is very weak, since the latter appears inside a log.

3. The sequence $a_k$ is said to converge quadratically if there is a constant $c$ satisfying
   \[ a_{k+1} \leq c \cdot a_k^2 \quad \text{for all } k. \]
   Observe then unrolling the recurrence yields
   \[ a_{k+1} \leq \frac{1}{c} (ca_0)^{2k+1}. \]
   The only role of the constant $c$ is to ensure the starting moment of convergence. In particular, if $ca_0 < 1$, then the inequality $a_k \leq \varepsilon$ holds for all $k \geq \log_2 \ln\left(\frac{1}{c}\right) - \log_2(-\ln(ca_0))$. The dependence on $c$ is negligible.

### 2.4 Two basic methods

This section presents two classical minimization algorithms: gradient descent and Newton’s method. It is crucial for the reader to keep in mind how the convergence guarantees are amplified when (strong) convexity is present.
2.4. TWO BASIC METHODS

2.4.1 Majorization view of gradient descent

Consider the optimization problem
\[ \min_{x \in E} f(x), \]
where \( f \) is a \( \beta \)-smooth function. Our goal is to design an iterative algorithm that generates iterates \( x_k \), such that any limit point of the sequence \( \{x_k\} \) is critical for \( f \). It is quite natural, at least at first, to seek an algorithm that is monotone, meaning that the sequence of function values \( \{f(x_k)\} \) is decreasing. Let us see one way this can be achieved, using the idea of majorization. In each iteration, we will define a simple function \( m_k \) (the “upper model”) agreeing with \( f \) at \( x_k \), and majorizing \( f \) globally, meaning that the inequality \( m_k(x) \geq f(x) \) holds for all \( x \in E \). Defining \( x_{k+1} \) to be the global minimizer of \( m_k \), we immediately deduce
\[ f(x_{k+1}) \leq m_k(x_{k+1}) \leq m_k(x_k) = f(x_k). \]
Thus function values decrease along the iterates generated by the scheme, as was desired.

An immediate question now is where such upper models \( m_k \) can come from. Here’s one example of a quadratic upper model:
\[ m_k(x) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{\beta}{2} \|x - x_k\|^2. \]
(C.1)
Clearly \( m_k \) agrees with \( f \) at \( x_k \), while Corollary [1.14] shows that the inequality \( m_k(x) \geq f(x) \) holds for all \( x \in E \), as required. It is precisely this ability to find quadratic upper models of the objective function \( f \) that separates minimization of smooth functions from those that are non-smooth.

Notice that \( m_k \) has a unique critical point, which must therefore equal \( x_{k+1} \) by first-order optimality conditions, and therefore we deduce
\[ x_{k+1} = x_k - \frac{1}{\beta} \nabla f(x_k). \]
This algorithm, likely familiar to the reader, is called gradient descent. Let us now see what can be said about limit points of the iterates \( x_k \). Appealing to Corollary [1.14] we obtain the descent guarantee
\[ f(x_{k+1}) \leq f(x_k) - \langle \nabla f(x_k), \beta^{-1} \nabla f(x_k) \rangle + \frac{\beta}{2} \|\beta^{-1} \nabla f(x_k)\|^2 \]
\[ = f(x_k) - \frac{1}{2\beta} \|\nabla f(x_k)\|^2. \]
(2.2)
Rearranging, and summing over the iterates, we deduce
\[ \sum_{i=0}^{k} \|\nabla f(x_i)\|^2 \leq 2\beta(f(x_0) - f(x_{k+1})). \]
Thus either the function values $f(x_k)$ tend to $-\infty$, or the sequence $\{\|\nabla f(x_i)\|^2\}$ is summable and therefore every limit point of the iterates $x_k$ is a critical point of $f$, as desired. Moreover, setting $f^* := \lim_{k \to \infty} f(x_k)$, we deduce the precise rate at which the gradients tend to zero:

$$\min_{i=0,\ldots,k} \|\nabla f(x_i)\|^2 \leq \frac{1}{k+1} \sum_{i=0}^k \|\nabla f(x_i)\|^2 \leq \frac{2\beta (f(x_0) - f^*)}{k+1}.$$

We have thus established the following result.

**Theorem 2.15 (Gradient descent).** Consider a $\beta$-smooth function $f : \mathbb{E} \to \mathbb{R}$. Then the iterates generated by the gradient descent method satisfy

$$\min_{i=0,\ldots,k} \|\nabla f(x_i)\|^2 \leq \frac{2\beta (f(x_0) - f^*)}{k+1}.$$

Convergence guarantees improve dramatically when $f$ is convex. Henceforth let $x^*$ be a minimizer of $f$ and set $f^* = f(x^*)$.

**Theorem 2.16 (Gradient descent and convexity).** Suppose that $f : \mathbb{E} \to \mathbb{R}$ is convex and $\beta$-smooth. Then the iterates generated by the gradient descent method satisfy

$$f(x_k) - f^* \leq \frac{\beta \|x_0 - x^*\|^2}{2k} \tag{2.3}$$

and

$$\min_{i=0,\ldots,2k} \|\nabla f(x_i)\| \leq \frac{\beta \|x_0 - x^*\|}{k}. \tag{2.4}$$

**Proof.** Since $x_{k+1}$ is the minimizer of the $\beta$-strongly convex quadratic $m_k(\cdot)$ in (2.1), we deduce

$$f(x_{k+1}) \leq m_k(x_{k+1}) \leq m_k(x^*) - \frac{\beta}{2} \|x_{k+1} - x^*\|^2.$$

We conclude

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle + \frac{\beta}{2} (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2)$$

$$\leq f^* + \frac{\beta}{2} (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2).$$

Summing for $i = 1, \ldots, k+1$ yields the inequality

$$\sum_{i=1}^k (f(x_i) - f^*) \leq \frac{\beta}{2} \|x_0 - x^*\|^2,$$

and therefore

$$f(x_k) - f^* \leq \frac{1}{k} \sum_{i=1}^k (f(x_i) - f^*) \leq \frac{\beta \|x_0 - x^*\|^2}{2k},$$
2.4. TWO BASIC METHODS

as claimed. Next, we verify (2.4). To this end, imagine we run gradient descent for \( k \) iterations and then again for \( k \) iterations starting from \( x_k \). Then from Theorem 2.15 we deduce

\[
\min_{i=k,\ldots,2k} \| \nabla f(x_i) \| \leq \sqrt{\frac{2\beta(f(x_k) - f^*)}{k+1}}.
\]

Applying the already proved estimate (2.3) to upper bound \( f(x_k) - f^* \), we arrive at (2.4), as claimed.

Thus when the gradient method is applied to a potentially nonconvex \( \beta \)-smooth function, the gradients \( \| \nabla f(x_k) \| \) decay as \( \frac{\sqrt{\beta} \| x_k - x^* \|}{\sqrt{k}} \), while for convex functions the estimate significantly improves to \( \frac{\beta \| x_k - x^* \|}{k} \).

Better linear rates on gradient, functional, and iterate convergence is possible when the objective function is strongly convex.

**Theorem 2.17 (Gradient descent and strong convexity).**

Suppose that \( f : E \to \mathbb{R} \) is \( \alpha \)-strongly convex and \( \beta \)-smooth. Then the iterates generated by the gradient descent method satisfy

\[
\| x_k - x^* \|^2 \leq \left( \frac{Q - 1}{Q + 1} \right)^k \| x_0 - x^* \|^2,
\]

where \( Q := \beta/\alpha \) is the condition number of \( f \).

**Proof.** Appealing to strong convexity, we have

\[
\| x_{k+1} - x^* \|^2 = \| x_k - x^* - \beta^{-1} \nabla f(x_k) \|^2
\]

\[
= \| x_k - x^* \|^2 + \frac{2}{\beta} \langle \nabla f(x_k), x^* - x_k \rangle + \frac{1}{\beta^2} \| \nabla f(x_k) \|^2
\]

\[
\leq \| x_k - x^* \|^2 + \frac{2}{\beta} \left( f^* - f(x_k) - \frac{\alpha}{2} \| x_k - x^* \|^2 \right) + \frac{1}{\beta^2} \| \nabla f(x_k) \|^2
\]

\[
= \left( 1 - \frac{\alpha}{\beta} \right) \| x_k - x^* \|^2 + \frac{2}{\beta} \left( f^* - f(x_k) + \frac{1}{2\beta} \| \nabla f(x_k) \|^2 \right).
\]

Seeking to bound the second summand, observe the inequalities

\[
f^* + \frac{\alpha}{2} \| x_{k+1} - x^* \|^2 \leq f(x_{k+1}) \leq f(x_k) - \frac{1}{2\beta} \| \nabla f(x_k) \|^2.
\]

Thus we deduce

\[
\| x_{k+1} - x^* \|^2 \leq \left( 1 - \frac{\alpha}{\beta} \right) \| x_k - x^* \|^2 - \frac{\alpha}{\beta} \| x_{k+1} - x^* \|^2.
\]

Rearranging yields

\[
\| x_{k+1} - x^* \|^2 \leq \left( \frac{Q - 1}{Q + 1} \right) \| x_k - x^* \|^2 \leq \left( \frac{Q - 1}{Q + 1} \right)^{k+1} \| x_0 - x^* \|^2,
\]

as claimed. \( \square \)
Thus for gradient descent, the quantities \( \|x_k - x^*\|^2 \) converge to zero at a linear rate \( Q^{-1} = 1 - \frac{2}{Q+1} \). We will often instead use the simple upper bound, \( 1 - \frac{2}{Q+1} \leq 1 - Q^{-1} \), to simplify notation. Analogous linear rates for \( \|\nabla f(x_k)\| \) and \( f(x_k) - f^* \) follow immediately from \( \beta \)-smoothness and strong convexity. In particular, in light of Section 2.3, we can be sure that the inequality \( \|x_k - x^*\|^2 \leq \varepsilon \) holds after \( k \geq \frac{Q+1}{2} \ln \left( \frac{\|x_0 - x^*\|^2}{\varepsilon} \right) \) iterations.

**Exercise 2.18** (Polyak stepsize). Consider a differentiable convex function \( f : E \to \mathbb{R} \) and let \( x^* \) be any of its minimizers. Consider the gradient descent iterates

\[
x_{k+1} = x_k - \gamma_k \nabla f(x_k),
\]

for some sequence \( \alpha_k \geq 0 \).

1. By writing the term \( \|x_{k+1} - x^*\|^2 = \|(x_{k+1} - x_k) + (x_k - x^*)\|^2 \) and expanding the square, deduce the estimate

\[
\frac{1}{2} \|x_{k+1} - x^*\|^2 \leq \frac{1}{2} \|x_k - x^*\|^2 - \gamma_k (f(x_k) - f(x^*)) + \frac{\gamma_k}{2} \|
\nabla f(x_k)\|^2. \tag{2.5}
\]

2. Supposing that you know the minimal value \( f^* \) of \( f \), show that the sequence \( \gamma_k = \frac{f(x_k) - f^*}{\|
\nabla f(x_k)\|^2} \) minimizes the right-hand-side of (2.5) in \( \gamma \), thereby yielding the guarantee

\[
\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \left( \frac{f(x_k) - f^*}{\|
\nabla f(x_k)\|^2} \right)^2.
\]

3. Let \( x_k \) be the sequence generated by the gradient method with \( \alpha_k = \frac{f(x_k) - f^*}{\|
\nabla f(x_k)\|^2} \). Supposing that \( f \) is \( \beta \)-smooth, conclude the estimate

\[
f \left( \frac{1}{k} \sum_{i=0}^{k-1} x_i \right) - f^* \leq \frac{\beta \|x_0 - x^*\|^2}{k}.
\]

If \( f \) is in addition \( \alpha \)-strongly convex, derive the guarantee

\[
\|x_{k+1} - x^*\|^2 \leq \left( 1 - \frac{\alpha^2}{4\beta^2} \right) \|x_k - x^*\|^2.
\]

### 2.4.2 Newton’s method

In this section we consider Newton’s method, an algorithm much different from gradient descent. Consider the problem of minimizing a \( C^2 \)-smooth function \( f : E \to \mathbb{R} \). Finding a critical point \( x \) of \( f \) can always be recast as the problem of solving the nonlinear equation \( \nabla f(x) = 0 \). Let us consider the equation solving question more generally. Let \( G : E \to E \) be a \( C^1 \)-smooth map. We seek a point \( x^* \) satisfying \( G(x^*) = 0 \). Given a current
iterate $x$, Newton’s method simply linearizes $G$ at $x$ and solves the equation $G(x) + \nabla G(x)(y - x) = 0$ for $y$. Thus provided that $\nabla G(x)$ is invertible, the next Newton iterate is given by

$$x_N = x - [\nabla G(x)]^{-1}G(x).$$

Coming back to the case of minimization, with $G = \nabla f$, the Newton iterate $x_N$ is then simply the unique critical point of the best quadratic approximation of $f$ at $x_k$, namely

$$Q(x; y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle,$$

provided that the Hessian $\nabla^2 f(x)$ is invertible. The following theorem establishes the progress made by each iteration of Newton’s method for equation solving.

**Theorem 2.19** (Progress of Newton’s method). Consider a $C^1$-smooth map $G : E \to E$ with the Jacobian $\nabla G$ that is $\beta$-Lipschitz continuous. Suppose that at some point $x$, the Jacobian $\nabla G(x)$ is invertible. Then the Newton iterate $x_N := x - [\nabla G(x)]^{-1}G(x)$ satisfies

$$\|x_N - x^*\| \leq \frac{\beta}{2} \|\nabla G(x)^{-1}\| \cdot \|x - x^*\|^2,$$

where $x^*$ is any point satisfying $G(x^*) = 0$.

**Proof.** Fixing an orthonormal basis, we can identify $E$ with $\mathbb{R}^m$ for some integer $m$. Then appealing to (1.15), we deduce

$$x_N - x^* = x - x^* - \nabla G(x)^{-1}G(x)
= \nabla G(x)^{-1}(\nabla G(x)(x - x^*) + G(x^*) - G(x))
= \nabla G(x)^{-1} \left( \int_0^1 (\nabla G(x) - \nabla G(x + t(x^* - x)))(x - x^*) \, dt \right).$$

Thus

$$\|x_N - x^*\| \leq \|\nabla G(x)^{-1}\| \cdot \|x - x^*\| \int_0^1 \|\nabla G(x) - \nabla G(x + t(x^* - x))\| \, dt
\leq \frac{\beta}{2} \|\nabla G(x)^{-1}\| \cdot \|x - x^*\|^2,$$

as claimed. □

To see the significance of Theorem 2.19, consider a $\beta$-smooth map $G : E \to E$. Suppose that $x^*$ satisfies $G(x^*) = 0$ and the Jacobian $\nabla G(x^*)$ is invertible. Then there exist constants $\epsilon, R > 0$, so that the inequality $\|\nabla G(x)^{-1}\| \leq \epsilon$ and $\|\nabla^2 G(x)^{-1}\| \leq R$ for all $x$ sufficiently close to $x^*$, establishing superlinear convergence of the Newton method.
$R$ holds for all $x \in B_{\epsilon}(x^*)$. Then provided that Newton’s method is initialized at a point $x_0$ satisfying $\|x_0 - x^*\| < \frac{2}{3R}$, the distance $\|x_{k+1} - x^*\|$ shrinks with each iteration at a quadratic rate.

Notice that guarantees for Newton’s method are local. Moreover it appears impossible from the analysis to determine whether a putative point is in the region of quadratic convergence. The situation becomes much better for a special class of functions, called self-concordant. Such functions form the basis for the so-called interior-point-methods in conic optimization. We will not analyze this class of functions in this text.

2.5 Computational complexity for smooth convex minimization

In the last section, we discussed at great length convergence guarantees of the gradient descent method for smooth convex optimization. Are there algorithms with better convergence guarantees? Before answering this question, it is important to understand the rates of convergence that one can even hope to prove. This section discusses so-called lower complexity bounds, expressing limitations on the convergence guarantees that any algorithm for smooth convex minimization can have.

Lower-complexity bounds become more transparent if we restrict attention to a natural subclass of first-order methods.

**Definition 2.20 (Linearly-expanding first-order method).** An algorithm is called a linearly-expanding first-order method if when applied to any $\beta$-smooth function $f$ on $\mathbb{R}^n$ it generates an iterate sequence $\{x_k\}$ satisfying

$$x_k \in x_0 + \text{span} \{\nabla f(x_0), \ldots, \nabla f(x_{k-1})\} \quad \text{for } k \geq 1.$$ 

Most first-order methods that we will encounter fall within this class. We can now state out first lower-complexity bound.

**Theorem 2.21 (Lower-complexity bound for smooth convex optimization).** For any $k$, with $1 \leq k \leq (n - 1)/2$, and any $x_0 \in \mathbb{R}^n$ there exists a convex $\beta$-smooth function $f : \mathbb{R}^n \to \mathbb{R}$ so that iterates generated by any linearly-expanding first-order method started at $x_0$ satisfy

$$f(x_k) - f^* \geq \frac{3\beta \|x_0 - x^*\|^2}{32(k+1)^2}, \quad (2.6)$$

$$\|x_k - x^*\|^2 \geq \frac{1}{8\beta} \|x_0 - x^*\|^2, \quad (2.7)$$

where $x^*$ is any minimizer of $f$.

For simplicity, we will only prove the bound on functional values (2.6).

Without loss of generality, assume $x_0 = 0$. The argument proceeds by
constructing a uniformly worst function for all linearly-expanding first-order methods. The construction will guarantee that in the \( k \)'th iteration of such a method, the iterate \( x_k \) will lie in the subspace \( \mathbb{R}^k \times \{0\}^{n-k} \). This will cause the function value at the iterates to be far from the optimal value.

Here is the precise construction. Fix a constant \( \beta > 0 \) and define the following family of quadratic functions

\[
f_k(z_1, z_2, \ldots, z_n) = \frac{\beta}{4} \left( \frac{1}{2} (z_1^2 + \sum_{i=1}^{k-1} (z_i - z_{i+1})^2 + z_k^2) - z_1 \right)
\]

indexed by \( k = 1, \ldots, n \). It is easy to check that \( f \) is convex and \( \beta \)-smooth. Indeed, a quick computation shows

\[
\langle \nabla f(x)v, v \rangle = \frac{\beta}{4} \left( \frac{1}{2} (v_1^2 + \sum_{i=1}^{k-1} (v_i - v_{i+1})^2 + v_k^2) \right)
\]

and therefore

\[
0 \leq \langle \nabla f(x)v, v \rangle \leq \frac{\beta}{4} \left( \frac{1}{2} (v_1^2 + \sum_{i=1}^{k-1} 2(v_i^2 + v_{i+1}^2) + v_k^2) \right) \leq \beta \|v\|^2.
\]

**Exercise 2.22.** Establish the following properties of \( f_k \).

1. Appealing to first-order optimality conditions, show that \( f_k \) has a unique minimizer

\[
\bar{x}_k = \begin{cases} 
1 - \frac{i}{k+1}, & \text{if } i = 1, \ldots, k \\
0 & \text{if } i = k + 1, \ldots, n 
\end{cases}
\]

with optimal value

\[
f_k^* = \frac{\beta}{8} \left( -1 + \frac{1}{k+1} \right).
\]

2. Taking into account the standard inequalities,

\[
\sum_{i=1}^{k} i = \frac{k(k+1)}{2} \quad \text{and} \quad \sum_{i=1}^{k} i^2 \leq \frac{(k+1)^3}{3},
\]

show the estimate \( \|\bar{x}_k\|^2 \leq \frac{1}{3} (k+1) \).

3. Fix indices \( 1 < i < j < n \) and a point \( x \in \mathbb{R}^i \times \{0\}^{n-i} \). Show that equality \( f_i(x) = f_j(x) \) holds and that the gradient \( \nabla f_k(x) \) lies in \( \mathbb{R}^{i+1} \times \{0\}^{n-(i+1)} \).
Proving Theorem 2.21 is now easy. Fix \( k \) and apply the linearly-expanding first order method to \( f := f_{2k+1} \) staring at \( x_0 = 0 \). Let \( x^* \) be the minimizer of \( f \) and \( f^* \) the minimum of \( f \). By Exercise 2.22 (part 3), the iterate \( x_k \) lies in \( \mathbb{R}^k \times \{0\}^{n-k} \). Therefore by the same exercise, we have \( f(x_k) = f_k(x_k) \geq \min f_k \).

Taking into account parts 1 and 2 of Exercise 2.22, we deduce

\[
\frac{f(x_k) - f^*}{\|x_0 - x^*\|^2} \geq \frac{\beta}{8} \left( 1 + \frac{1}{k+1} \right) - \frac{\beta}{8} \left( 1 + \frac{1}{2k+1} \right) = \frac{3\beta}{32(k+1)^2}.
\]

This proves the result.

The complexity bounds in Theorem 2.21 do not depend on strong convexity constants. When the target function class consists of \( \beta \)-smooth strongly convex functions, the analogous complexity bounds become

\[
f(x_k) - f^* \geq \left( \sqrt{\frac{Q}{Q + 1}} - 1 \right)^{2k} \|x_0 - x^*\|^2,
\]

(2.8)

\[
\|x_k - x^*\|^2 \geq \frac{\alpha}{2} \left( \sqrt{\frac{Q}{Q + 1}} - 1 \right)^{2k} \|x_0 - x^*\|^2,
\]

(2.9)

where \( x^* \) is any minimizer of \( f \) and \( Q := \beta/\alpha \) is the condition number. These bounds are proven in a similar way as Theorem 2.21, where one modifies the definition of \( f_k \) by adding a multiple of the quadratic \( \| \cdot \|^2 \).

Let us now compare efficiency estimates of gradient descent with the lower-complexity bounds we have just discovered. Consider a \( \beta \)-smooth convex functions \( f \) on \( \mathbb{E} \) and suppose we wish to find a point \( x \) satisfying \( f(x) - f^* \leq \varepsilon \). By Theorem 2.16, gradient descent will require at most \( k \leq O \left( \frac{\beta\|x_0 - x^*\|^2}{\varepsilon} \right) \) iterations. On the other hand, the lower-complexity bound (2.6) shows that no first-order method can be guaranteed to achieve the goal within \( k \leq O \left( \sqrt{\frac{\beta\|x_0 - x^*\|^2}{\varepsilon}} \right) \) iterations. Clearly there is a large gap.

Note that the bound (2.7) in essence says that convergence guarantees based on the distance to the solution set are meaningless for convex minimization in general.

Assume that in addition that \( f \) is \( \alpha \)-strongly convex. Theorem 2.16 shows that gradient descent will find a point \( x \) satisfying \( \|x - x^*\|^2 \leq \varepsilon \) after at most \( k \leq O \left( \frac{\beta\|x_0 - x^*\|^2}{\alpha \ln \left( \frac{\|x_0 - x^*\|^2}{\varepsilon} \right)} \right) \) iterations. Looking at the corresponding lower-complexity bound (2.9), we see that no first-order method can be guaranteed to find a point \( x \) with \( \|x - x^*\|^2 \leq \varepsilon \) after at most \( k \leq O \left( \sqrt{\frac{\beta\|x_0 - x^*\|^2}{\alpha \ln \left( \frac{\|x_0 - x^*\|^2}{\varepsilon} \right)}} \right) \) iterations. Again there is a large gap between convergence guarantees of gradient descent and the lower-complexity bound.

Thus the reader should wonder: are the proved complexity bounds too week or do their exist algorithms that match the lower-complexity
bounds stated above. In the following sections, we will show that the lower-complexity bounds are indeed sharp and there exist algorithms that match the bounds. Such algorithms are said to be “optimal”.

2.6 Conjugate Gradient Method

Before describing optimal first-order methods for general smooth convex minimization, it is instructive to look for inspiration at the primordial subclass of smooth optimization problems. We will consider minimizing strongly convex quadratics. For this class, the conjugate gradient method – well-known in numerical analysis literature – achieves rates that match the worst-case bound (2.8) for smooth strongly convex minimization.

Setting the groundwork, consider the minimization problem:

\[
\min_x f(x) := \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle,
\]

where \( b \in \mathbb{R}^n \) is a vector and \( A \in \mathbb{S}^n \) is a positive definite matrix. Clearly this problem amounts to solving the equation \( Ax = b \). We will be interested in iterative methods that approximately solve this problem, with the cost of each iteration dominated by a matrix vector multiplication. Notice, that if we had available an eigenvector basis, the problem would be trivial. Such a basis is impractical to compute and store for huge problems. Instead, the conjugate gradient method, which we will describe shortly, will cheaply generate partial eigenvector-like bases on the fly.

Throughout we let \( x^* := A^{-1}b \) and \( f^* := f(x^*) \). Recall that \( A \) induces the inner product \( \langle \cdot, \cdot \rangle_A := \langle Av, w \rangle \) and the norm \( \|v\|_A := \sqrt{\langle Av, v \rangle} \) (Exercise 1.2).

**Exercise 2.23.** Verify for any point \( x \in \mathbb{R}^n \) the equality

\[
f(x) - f^* = \frac{1}{2} \|x - x^*\|_A^2.
\]

We say that two vectors \( v \) and \( w \) are \( A \)-orthogonal if they are orthogonal in the inner product \( \langle \cdot, \cdot \rangle_A \). We will see shortly how to compute cheaply (and on the fly) an A-orthogonal basis.

Suppose now that we have available to us (somehow) an \( A \)-orthogonal basis \( \{v_1, v_2, \ldots, v_n\} \), where \( n \) is the dimension of \( \mathbb{R}^n \). Consider now the following iterative scheme: given a point \( x_1 \in \mathbb{R}^n \) define

\[
\begin{align*}
  t_k &= \arg\min_t f(x_k + tv_k) \\
  x_{k+1} &= x_k + t_kv_k
\end{align*}
\]

This procedure is called a conjugate direction method. Determining \( t_k \) is easy from optimality conditions. Henceforth, define the residuals \( r_k := b - Ax_k \). Notice that the residuals are simply the negative gradients \( r_k = -\nabla f(x_k) \).
Exercise 2.24. Prove the formula \( t_k = \frac{\langle r_k, v_k \rangle}{\| v_k \|^2_A} \)

Observe that the residuals \( r_k \) satisfy the equation

\[
 r_{k+1} = r_k - t_k A v_k.
\]

(2.10)

We will use this recursion throughout. The following theorem shows that such iterative schemes are “expanding subspace methods”.

Theorem 2.25 (Expanding subspaces). Fix an arbitrary initial point \( x_1 \in \mathbb{R}^n \). Then the equation

\[
 \langle r_{k+1}, v_i \rangle = 0 \quad \text{holds for all } i = 1, \ldots, k
\]

(2.11)

and \( x_{k+1} \) is the minimizer of \( f \) over the set \( x_1 + \text{span}\{v_1, \ldots, v_k\} \).

Proof. We prove the theorem inductively. Assume that equation (2.11) holds with \( k \) replaced by \( k - 1 \). Taking into account the recursion (2.10) and Exercise 2.24, we obtain

\[
 \langle r_{k+1}, v_k \rangle = \langle r_k, v_k \rangle - t_k \| v_k \|^2_A = 0.
\]

Now for any index \( i = 1, \ldots, k - 1 \), we have

\[
 \langle r_{k+1}, v_i \rangle = \langle r_k, v_i \rangle - t_k \langle v_k, v_i \rangle_A = \langle r_k, v_i \rangle = 0,
\]

where the last equation follows by the inductive assumption. Thus we have established (2.11). Now clearly \( x_{k+1} \) lies in \( x_1 + \text{span}\{v_1, \ldots, v_k\} \). On the other hand, equation (2.11) shows that the gradient \( \nabla f(x_{k+1}) = -r_{k+1} \) is orthogonal to \( \text{span}\{v_1, \ldots, v_k\} \). It follows immediately that \( x_{k+1} \) minimizes \( f \) on \( x_1 + \text{span}\{v_1, \ldots, v_k\} \), as claimed.

Corollary 2.26. The conjugate direction method finds \( x^* \) after at most \( n \) iterations.

Now suppose that we have available a list of nonzero \( A \)-orthogonal vectors \( \{v_1, \ldots, v_{k-1}\} \) and we run the conjugate direction method for as long as we can yielding the iterates \( \{x_1, \ldots, x_k\} \). How can we generate a new \( A \)-orthogonal vector \( v_k \) using only \( v_{k-1} \)? Notice that \( r_k \) is orthogonal to all the vectors \( \{v_1, \ldots, v_{k-1}\} \). Hence it is natural to try to expand in the direction \( r_k \). More precisely, let us try to set \( v_k = r_k + \beta_k v_{k-1} \) for some constant \( \beta_k \). Observe that \( \beta_k \) is uniquely defined by forcing \( v_k \) to be \( A \)-orthogonal with \( v_{k-1} \):

\[
 0 = \langle v_k, v_{k-1} \rangle_A = \langle r_k, v_{k-1} \rangle_A + \beta_k \| v_{k-1} \|^2_A.
\]

What about \( A \)-orthogonality with respect to the rest of the vectors? For all \( i \leq k - 2 \), we have the equality

\[
 \langle v_k, v_i \rangle_A = \langle r_k, v_i \rangle_A + \beta_k \langle v_{k-1}, v_i \rangle_A = \langle r_k, Av_i \rangle_A = t_i^{-1} \langle r_k, r_i - r_{i+1} \rangle_A.
\]
Supposing now that in each previous iteration \( i = 1, \ldots, k - 1 \) we had also set \( v_i := r_i + \beta_i v_{i-1} \), we can deduce the inclusions \( r_i, r_{i+1} \in \text{span} \{ v_i, v_{i-1}, v_{i+1} \} \). Appealing to Theorem 2.25 and the inequality above, we thus conclude that the set \( \{ v_1, \ldots, v_k \} \) is indeed \( \Lambda \)-orthogonal. The scheme just outlined is called the conjugate gradient method.

\[
\text{Algorithm 1: Conjugate gradient (CG)}
\]

1. Given \( x_0 \);
2. Set \( r_0 \leftarrow b - Ax_0, v_0 \leftarrow r_0, k \leftarrow 0 \);
3. while \( r_k \neq 0 \) do
4. \quad \quad \quad \quad \quad \quad t_k \leftarrow \frac{\langle r_k, v_k \rangle}{\|v_k\|_A^2}
5. \quad \quad \quad \quad \quad \quad x_{k+1} \leftarrow x_k + t_k v_k
6. \quad \quad \quad \quad \quad \quad r_{k+1} \leftarrow b - Ax_{k+1}
7. \quad \quad \quad \quad \quad \quad \beta_{k+1} \leftarrow -\frac{\langle r_{k+1}, v_k \rangle_A}{\|v_k\|_A^2}
8. \quad \quad \quad \quad \quad \quad v_{k+1} \leftarrow r_{k+1} + \beta_{k+1} v_k
9. \quad \quad k \leftarrow k + 1
10. end
11. return \( x_k \)

Convergence analysis of the conjugate gradient method relies on the observation that the expanding subspaces generated by the scheme are extremely special. Define the Krylov subspace of order \( k \) by the formula

\[
K_k(y) = \text{span} \{ y, Ay, A^2 y, \ldots, A^k y \}.
\]

**Theorem 2.27.** Consider the iterates \( x_k \) generated by the conjugate gradient method. Supposing \( x_k \neq x^* \), we have

\[
\langle r_k, r_i \rangle = 0 \quad \text{for all} \quad i = 0, 1, \ldots, k - 1, \quad (2.12)
\]

\[
\langle v_k, v_i \rangle_A = 0 \quad \text{for all} \quad i = 0, 1, \ldots, k - 1, \quad (2.13)
\]

and

\[
\text{span} \{ r_0, r_1, \ldots, r_k \} = \text{span} \{ v_0, v_1, \ldots, v_k \} = K_k(r_0). \quad (2.14)
\]

**Proof.** We have already proved equation (2.13), as this was the motivation for the conjugate gradient method. Equation (2.12) follows by observing the inclusion \( r_i \in \text{span} \{ v_i, v_{i-1} \} \) and appealing to Theorem 2.25. We prove the final claim (2.14) by induction. Clearly the equations hold for \( k = 0 \). Suppose now that they hold for some index \( k \). We will show that they continue to hold for \( k + 1 \).

Observe first that the inclusion

\[
\text{span} \{ r_0, r_1, \ldots, r_{k+1} \} \subseteq \text{span} \{ v_0, v_1, \ldots, v_{k+1} \} \quad (2.15)
\]
holds since \( r_i \) lie in \( \text{span}\{v_i, v_{i-1}\} \). Taking into account the induction assumption, we deduce \( v_{k+1} \in \text{span}\{r_{k+1}, v_k\} \subseteq \text{span}\{r_0, r_1, \ldots, r_{k+1}\} \). Hence equality holds in (2.15).

Next note by the induction hypothesis the inclusion
\[
r_{k+1} = r_k - t_kAv_k \in \mathcal{K}_k(r_0) - \mathcal{K}_{k+1}(r_0) \subseteq \mathcal{K}_{k+1}(r_0).
\]
Conversely, by the induction hypothesis, we have
\[
A^{k+1}r_0 = A(A^kr_0) \subseteq \text{span}\{Av_0, \ldots, Av_k\} \subseteq \text{span}\{r_0, \ldots, r_{k+1}\}.
\]
This completes the proof. \( \square \)

Thus as the conjugate gradient method proceeds, it forms minimizers of \( f \) over the expanding subspaces \( x_0 + \mathcal{K}_k(r_0) \). To see convergence implications of this observation, let \( \mathcal{P}_k \) be the set of degree \( k \) univariate polynomials with real coefficients. Observe that a point lies in \( \mathcal{K}_k(r_0) \) if and only if has the form \( p(A)r_0 \) for some polynomial \( p \in \mathcal{P}_k \). Therefore we deduce
\[
2(f(x_{k+1}) - f^*) = \inf_{x \in x_0 + \mathcal{K}_k(r_0)} 2(f(x) - f^*)
\]
\[
= \inf_{x \in x_0 + \mathcal{K}_k(r_0)} \|x - x^*\|^2_A = \min_{p \in \mathcal{P}_k} \|x_0 - p(A)r_0 - x^*\|^2_A
\]
Let \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \) be the eigenvalues of \( A \) and let \( A = U\Lambda U^T \) be an eigenvalue decomposition of \( A \). Define \( z := U^T(x_0 - x^*) \). Plugging in the definition of \( r_0 \) in the equation above, we obtain
\[
2(f(x_{k+1}) - f^*) = \min_{p \in \mathcal{P}_k} \|(x_0 - x^*) + p(A)(x_0 - x^*)\|^2_A
\]
\[
= \min_{p \in \mathcal{P}_k} \| (I + p(\Lambda)\Lambda)z \|^2_A
\]
\[
= \min_{p \in \mathcal{P}_k} \sum_{i=1}^n \lambda_i (1 + p(\lambda_i)\lambda_i)^2 z_i^2
\]
\[
\leq \left( \sum_{i=1}^n \lambda_i z_i^2 \right) \min_{p \in \mathcal{P}_k} \max_{i=1, \ldots, n} (1 + p(\lambda_i)\lambda_i)^2.
\]
Observe now the inequality \( \sum_{i=1}^n \lambda_i z_i^2 = \|z\|^2_A = \|x_0 - x^*\|^2_A \). Moreover, by polynomial factorization, polynomials of the form \( 1 + p(\lambda)\lambda \), with \( p \in \mathcal{P}_k \), are precisely the degree \( k + 1 \) polynomials \( q \in \mathcal{P}_{k+1} \) satisfying \( q(0) = 1 \). We deduce the key inequality
\[
f(x_{k+1}) - f^* \leq \frac{1}{2} \|x_0 - x^*\|^2_A \cdot \max_{i=1, \ldots, n} q(\lambda_i)^2 \quad (2.16)
\]
for any polynomial \( q \in \mathcal{P}_{k+1} \) with \( q(0) = 1 \). Convergence analysis now proceeds by exhibiting polynomials \( q \in \mathcal{P}_{k+1} \), with \( q(0) = 1 \), that evaluate to small numbers on the entire spectrum of \( A \). For example, the following is an immediate consequence.
2.6. CONJUGATE GRADIENT METHOD

Theorem 2.28 (Fast convergence with multiplicities). If $A$ has $m$ distinct eigenvalues, then the conjugate gradient method terminates after at most $m$ iterations.

Proof. Let $\gamma_1, \ldots, \gamma_m$ be the distinct eigenvalues of $A$ and define the degree $m$ polynomial $q(\lambda) := \frac{(-1)^m}{\pi^{1/2} \gamma_1 \cdots \gamma_m} (\lambda - \gamma_1) \cdots (\lambda - \gamma_m)$. Observe $q(0) = 1$. Moreover, clearly equality $0 = q(\gamma_i)$ holds for all indices $i$. Inequality (2.16) then implies $f(x_m) - f^* = 0$, as claimed.

For us, the most interesting convergence guarantee is derived from Chebyshev polynomials. These are the polynomials defined recursively by

$$T_0 = 1,$$

$$T_1(t) = t,$$

$$T_{k+1}(t) = 2tT_k(t) - T_{k-1}(t).$$

Before proceeding, we explain why Chebyshev polynomials appear naturally. Observe that inequality (2.16) implies

$$f(x_{k+1}) - f^* \leq \frac{1}{2}\|x_0 - x^*\|^2_A \cdot \max_{\lambda \in [\lambda_n, \lambda_1]} q(\lambda)^2.$$

It is a remarkable fact that Chebyshev polynomials, after an appropriate rescaling of the domain, minimize the right-hand-side over all polynomials $q \in P_{k+1}$ satisfying $q(0) = 1$. We omit the proof since we will not use this result for deriving convergence estimates. See Figure 2.2 for an illustration.

![Figure 2.2](image-url): $T_5$, $T_{10}$, $T_{40}$ are shown in red, black, and violet, respectively, on the interval $[-1, 1]$.

For any $k \geq 0$, the Chebyshev polynomials $T_k$ satisfy the following two key properties

(i) $|T_k(t)| \leq 1$ for all $t \in [-1, 1]$,

(ii) $T_k(t) := \frac{1}{2} \left( (t + \sqrt{t^2 - 1})^k + (t - \sqrt{t^2 - 1})^k \right)$ whenever $|t| \geq 1$. 
CHAPTER 2. SMOOTH MINIMIZATION

Theorem 2.29 (Linear convergence rate). Letting $Q = \lambda_1/\lambda_n$ be the condition number of $A$, the inequalities

$$f(x_k) - f^* \leq 2 \left( \frac{1}{\sqrt{Q} + 1} \right)^2 \|x_0 - x^*\|^2_A$$

hold for all $k$.

Proof. Define the normalization constant $c := T_k \left( \frac{\lambda_1 + \lambda_n}{\lambda_1 - \lambda_n} \right)$ and consider the degree $k$ polynomial $q(\lambda) = c^{-1} \cdot T_k \left( \frac{\lambda_1 + \lambda_n - 2\lambda}{\lambda_1 - \lambda_n} \right)$. Taking into account $q(0) = 1$, the inequality (2.16), and properties (i) and (ii), we deduce

$$\frac{f(x_k) - f^*}{\frac{1}{2} \|x_0 - x^*\|^2_A} \leq \max_{\lambda \in [\lambda_n, \lambda_1]} q(\lambda)^2 \leq T_k \left( \frac{\lambda_1 + \lambda_n}{\lambda_1 - \lambda_n} \right)^{-2} = 4 \left[ \left( \frac{\sqrt{Q} + 1}{\sqrt{Q} - 1} \right)^k + \left( \frac{\sqrt{Q} - 1}{\sqrt{Q} + 1} \right)^k \right]^{-2} \leq 4 \left( \frac{\sqrt{Q} - 1}{\sqrt{Q} + 1} \right)^{2k}.$$

The result follows. □

Thus linear convergence guarantees of the conjugate gradient method match those given by the lower complexity bounds (2.8).

2.7 Optimal methods for smooth convex minimization

In this section, we discuss optimal first-order methods for minimizing $\beta$-smooth functions. These are the methods whose convergence guarantees match the lower-complexity bounds (2.6) and (2.8).

2.7.1 Fast gradient methods

We begin with the earliest optimal method proposed by Nesterov. Our analysis, however, follows Beck-Teboulle and Tseng. To motivate the scheme, let us return to the conjugate gradient method (Algorithm 1). There are many ways to adapt the method to general convex optimization. Obvious modifications, however, do not yield optimal methods.

With $f$ a strongly convex quadratic, the iterates of the conjugate gradient method satisfy

$$x_{k+1} = x_k + t_k v_k = x_k + t_k (r_k + \beta_k v_{k-1}) = x_k - t_k \nabla f(x_k) + \frac{t_k \beta_{k-1}}{t_{k-1}} (x_k - x_{k-1}).$$

Thus $x_{k+1}$ is obtained by taking a gradient step $x_k - t_k \nabla f(x_k)$ and correcting it by the momentum term $\frac{t_k \beta_{k-1}}{t_{k-1}} (x_k - x_{k-1})$, indicating the direction from which one came. A direct extension of this procedure to minimizing general
smooth convex functions leads to the so-called Heavy Ball Method. Unfortunately, the convergence guarantees of this algorithm do not match the lower complexity bounds we previously established. Instead, let us slightly change the functional form of the recurrence. To this end, let \( f : E \to \mathbb{R} \) be a \( \beta \)-smooth convex function and consider the following recurrence

\[
\begin{align*}
y_k &= x_k + \gamma_k (x_k - x_{k-1}) \\
x_{k+1} &= y_k - \frac{1}{\beta} \nabla f(y_k)
\end{align*}
\]

for an appropriately chosen control sequence \( \gamma_k \geq 0 \). The reader should think of \( \{x_k\} \) as the iterate sequence, while \( \{y_k\} \) – the points at which we take gradient steps – are the corrections to \( x_k \) due to momentum.

Note that setting \( \gamma_k = 0 \) reduces to gradient descent. We will now see that the added flexibility of choosing nonzero \( \gamma_k \) leads to faster methods. Define the linearization

\[
l(y; x) = f(x) + \langle \nabla f(x), y - x \rangle.
\]

The analysis begins as gradient descent (Theorem \([2.16]\)). Since \( y \mapsto l(y; y_k) + \frac{\beta}{2} \|y - y_k\|^2 \) is a strongly convex quadratic, we deduce

\[
f(x_{k+1}) \leq l(x_{k+1}; y_k) + \frac{\beta}{2} \|x_{k+1} - y_k\|^2
\]

\[
\leq l(y; y_k) + \frac{\beta}{2} (\|y - y_k\|^2 - \|y - x_{k+1}\|),
\]

for all points \( y \in E \). Let \( x^* \) be the minimizer of \( f \) and \( f^* \) its minimum. In the analysis of gradient descent, we chose the comparison point \( y = x^* \). Instead, let us use the different point \( y = a_k x^* + (1 - a_k) x_k \) for some \( a_k \in (0, 1] \). We will determine \( a_k \) momentarily. We then deduce

\[
f(x_{k+1}) \leq l(a_k x^* + (1 - a_k) x_k; y_k)
\]

\[
+ \frac{\beta}{2} (\|a_k x^* + (1 - a_k) x_k - y_k\|^2 - \|a_k x^* + (1 - a_k) x_k - x_{k+1}\|^2)
\]

\[
= a_k l(x^*; y_k) + (1 - a_k) l(x_k; y_k)
\]

\[
+ \frac{\beta a_k^2}{2} (\|x^* - [x_k - a_k^{-1} (x_k - y_k)]\|^2 - \|x^* - [x_k - a_k^{-1} (x_k - x_{k+1})]\|^2).
\]

Convexity of \( f \) implies the upper bounds \( l(x^*; y_k) \leq f(x^*) \) and \( l(x_k; y_k) \leq f(x_k) \). Subtracting \( f^* \) from both sides and dividing by \( a_k^2 \) then yields

\[
\frac{1}{a_k} (f(x_{k+1}) - f^*) \leq \frac{1 - a_k}{a_k^2} (f(x_k) - f^*)
\]

\[
+ \frac{\beta}{2} (\|x^* - [x_k - a_k^{-1} (x_k - y_k)]\|^2
\]

\[
- \|x^* - [x_k - a_k^{-1} (x_k - x_{k+1})]\|^2).
\]

(2.17)
Naturally, we would like to now force telescoping in the last two lines by carefully choosing $\gamma_k$ and $a_k$. To this end, looking at the last term, define the sequence

$$z_{k+1} := x_k - a_k^{-1}(x_k - x_{k+1}). \quad (2.18)$$

Let us try to choose $\gamma_k$ and $a_k$ to ensure the equality $z_k = x_k - a_k^{-1}(x_k - y_k)$. From the definition (2.18) we get

$$z_k = x_{k-1} - a_{k-1}^{-1}(x_{k-1} - x_k) = x_k + (1 - a_{k-1}^{-1})(x_{k-1} - x_k).$$

Taking into account the definition of $y_k$, we conclude

$$z_k = x_k + (1 - a_{k-1}^{-1})\gamma_k^{-1}(x_k - y_k).$$

Therefore, the necessary equality

$$(1 - a_{k-1}^{-1})\gamma_k^{-1} = -a_k^{-1}$$

holds as long as we set $\gamma_k = a_k(a_k^{-1} - 1)$. Thus the inequality (2.17) becomes

$$\frac{1}{a_k^2}(f(x_{k+1}) - f^*) + \frac{\beta}{2}\|x^* - z_{k+1}\|^2 \leq \frac{1 - a_k}{a_k^2}(f(x_k) - f^*) + \frac{\beta}{2}\|x^* - z_k\|^2. \quad (2.19)$$

Set now $a_0 = 1$ and for each $k \geq 1$, choose $a_k \in (0, 1]$ satisfying

$$\frac{1 - a_k}{a_k^2} \leq \frac{1}{a_{k-1}^2}. \quad (2.20)$$

Then the right-hand-side of (2.19) is upper-bounded by the same term as the left-hand-side with $k$ replaced by $k - 1$. Iterating the recurrence (2.19) yields

$$\frac{1}{a_k^2}(f(x_{k+1}) - f^*) \leq \frac{1 - a_0}{a_0}(f(x_0) - f^*) + \frac{\beta}{2}\|x^* - z_0\|^2.$$

Taking into account $a_0 - 1 = 0$ and $z_0 = x_0 - a_0^{-1}(x_0 - y_0) = y_0$, we finally conclude

$$f(x_{k+1}) - f^* \leq a_k^2 \frac{\beta}{2}\|x^* - y_0\|^2.$$

Looking back at (2.20), the choices $a_k = \frac{2}{k+2}$ are valid, and will yield the efficiency estimate

$$f(x_{k+1}) - f^* \leq \frac{2\beta\|x^* - y_0\|^2}{(k+2)^2}.$$

Thus the scheme is indeed optimal for minimizing $\beta$-smooth convex functions, since this estimate matches the lower complexity bound (2.7). A slightly faster rate will occur when choosing $a_k \in (0, 1]$ to satisfy (2.20) with equality, meaning

$$a_{k+1} = \frac{\sqrt{a_k^4 + 4a_k^2 - a_k^2}}{2}. \quad (2.21)$$
2.7. OPTIMAL METHODS FOR SMOOTH CONVEX MINIMIZATION

Exercise 2.30. Suppose \( a_0 = 1 \) and \( a_k \) is given by (2.21) for each index \( k \geq 1 \). Using induction, establish the bound \( a_k \leq \frac{2}{k+2} \), for each \( k \geq 0 \).

As a side-note, observe that the choice \( a_k = 1 \) for each \( k \) reduces the scheme to gradient descent. Algorithm 2 and Theorem 2.31 summarize our findings.

Algorithm 2: Fast gradient method for smooth convex minimization

Input: Starting point \( x_0 \in \mathbb{E} \).
Set \( k = 0 \) and \( a_0 = a_{-1} = 1 \);
for \( k = 0, \ldots, K \) do
Set
\[
y_k = x_k + a_k (a_{k-1} - 1) (x_k - x_{k-1})
\]
\[
x_{k+1} = y_k - \frac{1}{\beta} \nabla f(y_k)
\]
(2.22)
Choose \( a_{k+1} \in (0, 1) \) satisfying
\[
\frac{1 - a_{k+1}}{a_{k+1}^2} \leq \frac{1}{a_k^2}.
\]
(2.23)
\( k \leftarrow k + 1 \).

Theorem 2.31 (Progress of the fast-gradient method). Suppose that \( f \) is a \( \beta \)-smooth convex function. Then provided we set \( a_k \leq \frac{2}{k+2} \) for all \( k \) in Algorithm 2, the iterates generated by the scheme satisfy
\[
f(x_k) - f^* \leq \frac{2\beta\|x^* - x_0\|^2}{(k + 1)^2}.
\]
(2.24)

Let us next analyze the rate at which Algorithm 2 forces the gradient to tend to zero. One can try to apply the same reasoning as in the proof of Theorem 2.16. One immediately runs into a difficulty, however, namely there is no clear relationship between the values \( f(y_k) \) and \( f(x_k) \). This difficulty can be overcome by introducing an extra gradient step in the scheme. A simpler approach is to take slightly shorter gradient steps in (2.22).

Theorem 2.32 (Gradient convergence of the fast-gradient method).
Suppose that \( f \) is a \( \beta \)-smooth convex function. In Algorithm 2, set \( a_k \leq \frac{2}{k+2} \) for all \( k \) and replace line (2.22) by \( x_{k+1} = y_k - \frac{1}{2\beta} \nabla f(y_k) \). Then the iterates
generated by the algorithm satisfy
\[ f(x_k) - f^* \leq \frac{4\beta\|x^* - x_0\|^2}{(k+1)^2}, \quad (2.25) \]
\[ \min_{i=1, \ldots, k} \|\nabla f(y_i)\| \leq \frac{8\sqrt{3} \cdot \beta\|x^* - x_0\|}{\sqrt{k(k+1)(2k+1)}}. \quad (2.26) \]

Proof. The proof is a slight modification of the argument outlined above of Theorem 2.31. Observe
\[ f(x_{k+1}) \leq l(x_{k+1}; y_k) + \frac{\beta}{2}\|x_{k+1} - y_k\|^2 \]
\[ \leq l(x_{k+1}; y_k) + \frac{2\beta}{2}\|x_{k+1} - y_k\|^2 - \frac{1}{8\beta} \|\nabla f(y_k)\|^2 \]
\[ \leq l(y; y_k) + \frac{2\beta}{2} (\|y - y_k\|^2 - \|y - x_{k+1}\|^2) - \frac{1}{8\beta} \|\nabla f(y_k)\|^2. \]
Continuing as before, we set \( z_k = x_k - a_k^{-1}(x_k - y_k) \) and obtain
\[ \frac{1}{a_k} (f(x_{k+1}) - f^*) + \beta\|x^* - z_{k+1}\|^2 \leq \frac{1 - a_k}{a_k} (f(x_k) - f^*) + \beta\|x^* - z_k\|^2 - \frac{1}{8\beta} \|\nabla f(y_k)\|^2. \]
Recall \( \frac{1 - a_k}{a_k} \leq \frac{1}{a_{k-1}} \), \( a_1 = 1 \), and \( z_0 = x_0 \). Iterating the inequality yields
\[ \frac{1}{a_k^2} (f(x_{k+1}) - f^*) + \beta\|x^* - z_{k+1}\|^2 \leq \beta\|x^* - x_0\|^2 - \frac{1}{8\beta} \sum_{i=1}^{k} \|\nabla f(y_i)\|^2 / a_i^4. \]
Ignoring the second terms on the left and right sides yields (2.25). On the other hand, lower-bounding the left-hand-side by zero and rearranging gives
\[ \min_{i=1, \ldots, k} \|\nabla f(y_i)\|^2 \cdot \sum_{i=1}^{k} \left( \frac{1}{a_i^2} \right) \leq 8\beta^2\|x^* - x_0\|^2. \]
Taking into account the inequality
\[ \sum_{i=1}^{k} \left( \frac{1}{a_i^2} \right) \geq \sum_{i=1}^{k} \frac{(i+2)^2}{4} \geq \frac{1}{4} \sum_{i=1}^{k} i^2 = \frac{k(k+1)(2k+1)}{24}, \]
we conclude
\[ \min_{i=1, \ldots, k} \|\nabla f(y_i)\|^2 \leq \frac{192\beta^2\|x^* - x_0\|^2}{k(k+1)(2k+1)}. \]
Taking a square root of both sides gives (2.26). \( \square \)

Thus the iterate generated by the fast gradient method with a damped step-size satisfy \( \min_{i=1, \ldots, k} \|\nabla f(y_i)\| \leq \mathcal{O}\left( \frac{\beta\|x^* - x_0\|}{k^{3/2}} \right) \). This is in contrast to gradient descent, which has the worse efficiency estimate \( \mathcal{O}\left( \frac{\beta\|x^* - x_0\|}{k} \right) \). We will see momentarily that surprisingly even a better rate is possible by applying a fast gradient method to a small perturbation of \( f \).
2.7. OPTIMAL METHODS FOR SMOOTH CONVEX MINIMIZATION

A restart strategy for strongly convex functions

Recall that gradient descent converges linearly for smooth strongly convex functions. In contrast, to make Algorithm 2 linearly convergent for this class of problems, one must modify the method. Indeed, the only modification that is required is in the definition of $a_k$ in (2.23). The argument behind the resulting scheme relies on a different algebraic technique called estimate sequences. This technique is more intricate and more general than the arguments we outlined for sublinear rates of convergence. We will explain this technique in Section 2.7.2.

There is, however, a different approach to get a fast linearly convergent method simply by periodically restarting Algorithm 2. Let $f : E \rightarrow \mathbb{R}$ be a $\beta$-smooth and $\alpha$-convex function. Imagine that we run the basic fast-gradient method on $f$ for a number of iterations (an epoch) and then restart. Let $x^i_k$ be the $k$’th iterate generated in epoch $i$. Theorem 2.31 along with strong convexity yields the guarantee

$$f(x^i_k) - f^* \leq \frac{2\beta \|x^* - x^i_0\|^2}{(k+1)^2} \leq \frac{4\beta}{\alpha(k+1)^2}(f(x^i_0) - f^*).$$ (2.27)

Suppose that in each epoch, we run a fast gradient method (Algorithm 2) for $N$ iterations. Given an initial point $x_0 \in E$, set $x^0_0 := x_0$ and set $x^i_0 := x^{i-1}_N$ for each $i \geq 1$. Thus we initialize each epoch with the final iterate of the previous epoch.

Then for any $q \in (0, 1)$, as long as we use $N_q \geq \sqrt{\frac{4\beta}{q\alpha}}$ iterations in each epoch we can ensure the contraction:

$$f(x^i_0) - f^* \leq q(f(x^{i-1}_0) - f^*) \leq q^i(f(x_0) - f^*).$$

The total number of iterations to obtain $x^i_0$ is $iN_q$. We deduce

$$f(x^i_0) - f^* \leq (q^{1/N_q})^{iN_q}(f(x_0) - f^*).$$

Let us therefore choose $q$ according to

$$\min_q q^{1/N_q}.$$

Using logarithmic differentiation, the optimal choice is $q = e^{-2}$, yielding $N_q = \left\lceil 2e\sqrt{\frac{\beta}{\alpha}} \right\rceil$. Thus we have a complete algorithm (Algorithm 3).

To see that this is indeed an optimal method, observe the bound

$$q^{1/N_q} \leq e^{-2}\left(2e\sqrt{\frac{\beta}{\alpha}}\right)^{-1} \leq e^{\frac{2}{1+2\sqrt{\beta/\alpha}}}. $$
Algorithm 3: Fast gradient method with restarts

**Input:** Starting point \( x_0 \in \mathbb{E} \).

Set \( i, k = 0, x^0_0 = x_0 \), and \( N = \left\lceil 2e\sqrt{\frac{\beta}{\alpha}} \right\rceil \).

for \( i = 0, \ldots, K \) do

Let \( x^{N}_i \) be the \( N \)’th iterate generated by Algorithm 2, initialized with \( x^0_i \).

Set \( i = i + 1 \) and \( x^0_{i+1} = x^{N}_i \).

end

Simple algebra shows

\[
-\frac{2}{1+2e\sqrt{\beta/\alpha}} \in (-\frac{1}{3}, 0].
\]

Noting for \( x \in (-\frac{1}{3}, 0) \), the inequality

\[
e^x \leq 1 + x + \frac{1}{2}x^2 \leq 1 + \frac{5}{6}x,
\]

we conclude

\[
q^{1/N} \leq 1 - \frac{5/3}{1 + 2e\sqrt{\beta/\alpha}}.
\]

Thus the method will find a point \( x \) satisfying \( f(x) - f^* \leq \varepsilon \) after at most

\[
\frac{1+2e\sqrt{\beta/\alpha}}{5/3} \ln \left( \frac{f(x_0) - f^*}{\varepsilon} \right)
\]

iterations of fast gradient methods. This matches the lower complexity bound \((2.8)\) for smooth strongly convex minimization.

### 2.7.2 Fast gradient methods through estimate sequences

In this section, we describe an algebraic technique for designing fast gradient method for minimizing a \( \beta \)-smooth \( \alpha \)-convex function. In the setting \( \alpha = 0 \), the algorithm will turn out to be identical to Algorithm 2. The entire construction relies on the following gadget.

**Definition 2.33 (Estimate Sequences).** Given real numbers \( \lambda_k \in [0, 1] \) and functions \( \phi_k: \mathbb{E} \to \mathbb{R} \), we say that the sequence \( (\lambda_k, \phi_k(x)) \) is an estimate sequence if \( \lambda_k \searrow 0 \) and the inequality

\[
\phi_k(x) \leq (1 - \lambda_k)f(x) + \lambda_k\phi_0(x) \tag{2.28}
\]

holds for all \( x \in \mathbb{E} \) and \( k \geq 0 \).

This notion may seem abstract at first sight. Its primary use comes from the following observation. Suppose we are given an estimate sequence and we can find a point \( x_k \) satisfying

\[
f(x_k) \leq \phi^*_k := \min_x \phi_k(x).
\]

Then we immediately deduce

\[
f(x_k) \leq (1 - \lambda_k)f^* + \lambda_k\phi^*_0
\]
and hence
\[ f(x_k) - f^* \leq \lambda_k (\phi_0^* - f^*). \] (2.29)
Thus the rate at which \( \lambda_k \) tends to zero directly controls the rate at which the values \( f(x_k) \) tend to \( f^* \).

Thus we have two items to consider when designing an algorithm based on estimate sequences: \((i)\) how to choose an estimate sequence \((\lambda_k, \phi_k(x))\) and \((ii)\) how to choose \( x_k \) satisfying \( f(x_k) \leq \phi_k^* \).

Let us address the first question. Looking at the definition, it is natural to form an estimate sequence by successively averaging quadratic models of \( f \) formed at varying points \( y_k \). Define the lower quadratic models
\[ Q_y(x) := f(y) + \langle \nabla f(y), x - y \rangle + \frac{\alpha}{2} \|x - y\|^2. \]

**Exercise 2.34.** Suppose that \( f : E \rightarrow \mathbb{R} \) is \( C^1 \)-smooth and \( \alpha \)-strongly convex. Fix two sequences \( \{y_k\}_{k \geq 0} \subset E \) and \( \{t_k\}_{k \geq 0} \subset [0,1] \), and consider an arbitrary function \( \phi_0 : E \rightarrow \mathbb{R} \). Define the sequence \((\lambda_k, \phi_k)\) inductively as follows:
\[
\begin{cases}
\lambda_0 = 1 \\
\lambda_{k+1} = (1 - t_k) \lambda_k \\
\phi_{k+1} = (1 - t_k) \phi_k + t_k Q_{y_k}
\end{cases}
\]

1. Show that the sequence \((\lambda_k, \phi_k)\) satisfies (2.28). (Hint: Begin by noting \( \phi_{k+1} \leq (1 - t_k) \phi_k + t_k f \).)
2. Show that provided \( \sum_{k=0}^{\infty} t_k = +\infty \), we have \( \lambda_k \searrow 0 \) and therefore \((\lambda_k, \phi_k)\) is an estimate sequence for \( f \).

It is clear that if we choose \( \phi_0 \) to be a simple quadratic \( \phi_0(x) = \phi_0^* + \frac{\gamma_0}{2} \|x - v_0\|^2 \), then all \( \phi_k \) will be simple quadratics as well, in the sense that their Hessians will be multiples of identity.

**Exercise 2.35.** Let
\[ \phi_0(x) = \phi_0^* + \frac{\gamma_0}{2} \|x - v_0\|^2, \]
where \( \phi_0^* \in \mathbb{R}, \gamma_0 \geq 0, \) and \( v_0 \in E \) are chosen arbitrary. Show by induction that the functions \( \phi_k \) in Exercise 2.34 preserve the same form:
\[ \phi_k(x) = \phi_k^* + \frac{\gamma_k}{2} \|x - v_k\|^2, \]
where
\[ \gamma_{k+1} = (1 - t_k)\gamma_k + t_k\alpha, \]
\[ v_{k+1} = \frac{1}{\gamma_{k+1}} \left[ (1 - t_k)\gamma_kv_k + t_k\alpha y_k - t_k \nabla f(y_k) \right], \]
\[ \phi_{k+1}^* = (1 - t_k)\phi_k^* + t_kf(y_k) - \frac{t_k^2}{2\gamma_{k+1}} \| \nabla f(y_k) \|^2 \]
\[ + \frac{t_k(1 - t_k)\gamma_k}{\gamma_{k+1}} \left( \frac{\alpha}{2} \| y_k - v_k \|^2 + \langle \nabla f(y_k), v_k - y_k \rangle \right). \tag{2.30} \]

Now having available an estimate sequence constructed above, let’s try to find the sequence \( \{x_k\} \) satisfying \( f(x_k) \leq \phi_k^* \). Suppose we already have available a point \( x_k \) satisfying this condition; let us see how to choose \( x_{k+1} \).

Lowerbounding the term \( \| y_k - v_k \| \) in (2.30) by zero, we deduce
\[ \phi_{k+1}^* \geq (1 - t_k)f(x_k) + t_kf(y_k) - \frac{t_k^2}{2\gamma_{k+1}} \| \nabla f(y_k) \|^2 \]
\[ + \frac{t_k(1 - t_k)\gamma_k}{\gamma_{k+1}} \langle \nabla f(y_k), v_k - y_k \rangle. \]

Combining this with \( f(x_k) \geq f(y_k) + \langle \nabla f(y_k), x_k - y_k \rangle \), yields
\[ \phi_{k+1}^* \geq \left( f(y_k) - \frac{t_k^2}{2\gamma_{k+1}} \| \nabla f(y_k) \|^2 \right) + (1 - t_k)\langle \nabla f(y_k), \frac{t_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k \rangle. \]

The term in parenthesis is reminiscent of a descent condition for a gradient step, \( f(y_k) - \frac{1}{2\beta} \| \nabla f(y_k) \|^2 \geq f(y_k - \beta^{-1}\nabla f(y_k)) \). Let us therefore ensure \( \frac{t_k^2}{\gamma_{k+1}} = \frac{1}{2\beta} \), by finding \( t_k \) satisfying
\[ t_k^2\beta = \gamma_{k+1} = (1 - t_k)\gamma_k + t_k\alpha, \]
and set
\[ x_{k+1} = y_k - \frac{1}{\beta} \nabla f(y_k). \]

We then deduce
\[ \phi_{k+1}^* \geq f(x_{k+1}) + (1 - t_k)\langle \nabla f(y_k), \frac{t_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k \rangle. \]

Finally let us ensure
\[ \frac{t_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k = 0, \]
by setting
\[ y_k = \frac{t_k\gamma_k v_k + \gamma_{k+1}x_k}{\gamma_k + t_k\alpha}. \]
Algorithm 4: Fast gradient method based on estimate sequences

**Input:** Starting point \( x_0 \in \mathbf{E} \).

Set \( k = 0, \; v_0 = x_0, \; \) and \( \phi_0^* = f(x_0) \);

**for** \( k = 0, \ldots, K \) **do**

Compute \( t_k \in (0, 1) \) from equation

\[
\beta t_k^2 = (1 - t_k) \gamma_k + t_k \alpha. \tag{2.31}
\]

Set

\[
\gamma_{k+1} = (1 - t_k) \gamma_k + t_k \alpha \tag{2.32}
\]

\[
y_k = \frac{t_k \gamma_k v_k + \gamma_{k+1} x_k}{\gamma_k + t_k \alpha} \tag{2.33}
\]

\[
x_{k+1} = y_k - \frac{1}{\beta} \nabla f(y_k) \tag{2.34}
\]

\[
v_{k+1} = \frac{(1 - t_k) \gamma_k v_k + t_k \alpha y_k - t_k \nabla f(y_k)}{\gamma_{k+1}} \tag{2.35}
\]

Set \( k \leftarrow k + 1 \).

**end**

With this choice, we can be sure \( \phi_{k+1}^* \geq f(x_{k+1}) \) as needed. Algorithm 4 outlines this general scheme.

Appealing to (2.29) and exercise 2.34, we see that the point \( x_k \) generated by Algorithm 4 satisfy

\[
f(x_k) - f^* \leq \lambda_k \left[ f(x_0) - f^* + \frac{\gamma_0}{2} \| x_0 - x^* \|^2 \right], \tag{2.36}
\]

where \( \lambda_0 = 1 \) and \( \lambda_k = \prod_{i=0}^{k-1} (1 - t_i) \). Thus in understanding convergence guarantees of the method, we must estimate the rate at which \( \lambda_k \) decays.

**Theorem 2.36** (Decay of \( \lambda_k \)). Suppose in Algorithm 4 we set \( \gamma_0 \geq \alpha \). Then

\[
\lambda_k \leq \min \left\{ \left( 1 - \sqrt{\frac{\alpha}{\beta}} \right)^k, \frac{4\beta}{(2\sqrt{\beta} + k\sqrt{\gamma_0})^2} \right\}.
\]

**Proof.** Observe that if \( \gamma_k \geq \alpha \), then

\[
\beta t_k^2 = \gamma_{k+1} = (1 - t_k) \gamma_k + t_k \alpha \geq \alpha.
\]

This implies \( t_k \geq \sqrt{\frac{\alpha}{\beta}} \) and hence \( \lambda_k = \prod_{i=0}^{k-1} (1 - t_i) \leq \left( 1 - \sqrt{\frac{\alpha}{\beta}} \right)^k \).

For the other inequality, let \( c_j = \frac{1}{\sqrt{\lambda_j}} \). Taking into account that \( \lambda_j \) are
CHAPTER 2. SMOOTH MINIMIZATION

decreasing, observe
\[ c_{j+1} - c_j = \frac{\sqrt{\lambda_j} - \sqrt{\lambda_{j+1}}}{\sqrt{\lambda_j \lambda_{j+1}}} = \frac{\lambda_j - \lambda_{j+1}}{\sqrt{\lambda_j \lambda_{j+1} (\sqrt{\lambda_j} + \sqrt{\lambda_{j+1}})}} \]
\[ \geq \frac{\lambda_j - \lambda_{j+1}}{2\lambda_j \sqrt{\lambda_{j+1}}} = \frac{\lambda_j - (1-t_j)\lambda_j}{2\lambda_j \sqrt{\lambda_{j+1}}} = \frac{t_j}{2\sqrt{\lambda_{j+1}}}. \]

Notice \( \gamma_0 = \gamma_0 \lambda_0 \). Assuming \( \gamma_j \geq \gamma_0 \lambda_j \) we arrive at the analogous inequality for \( j+1 \), namely
\[ \gamma_{j+1} \geq (1-t_j)\gamma_j \geq \gamma_0 \lambda_j \geq \gamma_0 \lambda_{j+1}. \]

Thus \( \gamma_0 \lambda_{j+1} \leq \gamma_{j+1} = \beta t_j^2 \), which implies that \( \frac{t_j}{2\sqrt{\lambda_{j+1}}} \geq \frac{1}{2} \cdot \frac{\sqrt{\gamma_0}}{\beta} \). So we deduce that
\[ c_{j+1} - c_j \geq \frac{1}{2} \cdot \frac{\sqrt{\gamma_0}}{\beta}. \]

Summing over \( j = 0, \ldots, k-1 \), we get
\[ c_k - c_0 \geq \frac{k}{2} \cdot \frac{\sqrt{\gamma_0}}{\beta} \]
and hence
\[ \frac{1}{\sqrt{\lambda_k}} - 1 \geq \frac{k}{2} \cdot \frac{\sqrt{\gamma_0}}{\beta}. \]

The claimed estimate
\[ \lambda_k \leq \frac{4\beta}{(2\sqrt{\beta} + k\sqrt{\gamma_0})^2} \]
follows. \( \square \)

**Corollary 2.37.** Setting \( \gamma_0 = \beta \) in Algorithm 4 yields iterates satisfying
\[ f(x_k) - f^* \leq \beta \min \left\{ \left( 1 - \sqrt{\frac{\alpha}{\beta}} \right)^k, \frac{4}{(k+2)^2} \right\} \cdot \|x_0 - x^*\|^2. \]

**Proof.** This follows immediately from inequality (2.36), Theorem 2.36 and the inequality \( f(x_0) - f^* \leq \frac{\beta}{2} \|x_0 - x^*\|^2. \) \( \square \)

Let us try to eliminate \( v_k \). Solving for \( v_k \) in (2.33) and plugging in this description into (2.35) and rearranging yields the equality
\[ v_{k+1} = \frac{1}{\gamma_{k+1}} \left( \frac{1-t_k}{t_k} \gamma_k + t_k \alpha \right) y_k - \frac{1-t_k}{t_k} x_k - \frac{1}{\gamma_{k+1}} \nabla f(y_k). \]
Hence we deduce
\[ v_{k+1} = \frac{1}{\gamma_{k+1}} y_k - \frac{1-t_k}{t_k} x_k - \frac{1}{t_k \beta} \nabla f(y_k) \]
\[ = x_k + \frac{1}{t_k} (x_{k+1} - x_k), \]
where the first inequality follows from (2.32) and (2.31), while the last uses (2.34). Plugging in the analogous expression of \( v_{k+1} \) into (2.33) yields
\[ y_{k+1} = x_{k+1} + \frac{t_{k+1} \gamma_{k+1} (1-t_k)}{t_k (\gamma_{k+1} + \gamma_{k+1} + \gamma_{k+1})} (x_{k+1} - x_k) \]
\[ = x_{k+1} + \zeta_k (x_{k+1} - x_k), \]
where we define
\[ \zeta_k := \frac{t_{k+1} \gamma_{k+1} (1-t_k)}{t_k (\gamma_{k+1} + \gamma_{k+1} + \gamma_{k+1})}. \]
Thus \( v_k \) is eliminated from the algorithm. Let us now eliminate \( \gamma_k \). To this end note from (2.31) \( t_{k+1} \alpha = \beta t_{k+1}^2 - (1 - t_{k+1}) \gamma_{k+1} \), and hence
\[ \zeta_k := \frac{t_{k+1} \gamma_{k+1} (1-t_k)}{t_k (\beta t_{k+1} + \gamma_{k+1})} = \frac{t_{k+1} (1-t_k)}{t_{k+1} + t_k}, \]
where the last equality uses \( \gamma_{k+1} = \beta t_k^2 \). Finally plugging in \( \gamma_{k+1} = \beta t_k^2 \) into (2.31) yields
\[ t_{k+1}^2 = (1 - t_{k+1}) t_k^2 + \frac{\alpha}{\beta} t_{k+1}. \]
Thus \( \gamma_k \) is eliminated from the scheme.

---

**Algorithm 5:** Simplified fast gradient method

**Input:** Starting point \( x_0 \in \mathbb{E} \) and \( t_0 \in (0, 1) \).

Set \( k = 0 \) and \( y_0 = x_0 \);

for \( k = 0, \ldots, K \) do

\[ x_{k+1} = y_k - \frac{1}{\beta} \nabla f(y_k). \]

Compute \( t_{k+1} \in (0, 1) \) from the equation

\[ t_{k+1}^2 = (1 - t_{k+1}) t_k^2 + \frac{\alpha}{\beta} t_{k+1} \quad (2.37) \]

Set

\[ y_{k+1} = x_{k+1} + \frac{t_k (1-t_k)}{t_k^2 + t_{k+1}} (x_{k+1} - x_k). \]

end

Thus we have established the following.

**Corollary 2.38.** Setting \( t_0 = \frac{\alpha}{\beta} \) in Algorithm 5 yields iterates satisfying

\[ f(x_k) - f^* \leq \beta \min \left\{ \left( 1 - \sqrt[2]{\frac{\alpha}{\beta}} \right)^k, \frac{4}{(k+2)^2} \right\} \cdot \|x_0 - x^*\|^2. \]
It is important to note that in the case \( \alpha = 0 \), Algorithm 5 is exactly Algorithm 2 with \( a_k = t_k \). Indeed, equality (2.37) can be rewritten as
\[
\frac{1 - t_{k+1}}{t_{k+1}^2} = \frac{1}{t_k^2},
\]
which is exactly the equality in (2.23). Moreover observe
\[
\frac{t_k(1 - t_k)}{t_k^2 + t_{k+1}} = \left( \frac{t_k^2}{t_k^2 + t_{k+1}} \right) (t_k^{-1} - 1) = t_{k+1}(t_k^{-1} - 1),
\]
where the second equality follows from (2.37). Thus the interpolation coefficients in the definition of \( y_k \) are exactly the same.

### 2.7.3 Optimal quadratic averaging

The disadvantage of the derivation of the fast gradient methods discussed in the previous sections is without a doubt a lack of geometric intuition. Indeed the derivation of the schemes was entirely based on algebraic manipulations. In this section, we present a different method that is better grounded in geometry. The scheme we outline is based on averaging quadratic (lower) models of the functions, and therefore shares some superficial similarity with the approach based on estimate sequence. The way that the quadratics are used, however, is completely different. It is also important to note that the scheme has two disadvantages, when compared with the fast-gradient methods described in the previous sections: (1) it requires being able to compute exact minimizers of the function along lines and (2) the method only applies to minimizing strongly convex functions.

Henceforth, let \( f : E \rightarrow \mathbb{R} \) be a \( \beta \)-smooth and \( \alpha \)-convex function with \( \alpha > 0 \). We denote the unique minimizer of \( f \) by \( x^* \), its minimal value by \( f^* \), and its condition number by \( \kappa := \beta/\alpha \). For any points \( x, y \in E \), we let \( \text{line.search}(x, y) \) be the minimizer of \( f \) on the line between \( x \) and \( y \). We assume throughout this section that \( \text{line.search}(x, y) \) is computable. This is a fairly mild assumption for a number of settings. For example, suppose that \( f \) has the form \( f(x) = h(Ax) + g(x) \) for some smooth convex functions \( h, g \), a linear map \( A \), and a vector \( b \). In many applications, the cost of each iteration of first order methods on this problem is dominated by the cost of the vector matrix multiplication \( Ax \). Consider now the univariate line-search problem
\[
\min_t f(x + tv) = \min_t h(Ax + tAv) + g(x + tv).
\]
Since one can precompute \( Av \), evaluations of \( g(t) \) for varying \( t \) are cheap. Consequently, the univariate problem can be solved by specialized methods.

Given a point \( x \in E \), we define the following two points
\[
x^+ := x - \frac{1}{\beta} \nabla f(x) \quad \text{and} \quad x^{++} := x - \frac{1}{\alpha} \nabla f(x).
\]
The first point $x^+$ is the familiar gradient step, while the role of $x^{++}$ will become apparent shortly.

The starting point for our development is the elementary observation that every point $\bar{x}$ provides a quadratic under-estimator of the objective function, having a canonical form. Indeed, completing the square in the strong convexity inequality

$$f(x) \geq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{\alpha}{2} \|x - \bar{x}\|^2$$

yields

$$f(x) \geq \left( f(\bar{x}) - \frac{\|\nabla f(\bar{x})\|^2}{2\alpha} \right) + \frac{\alpha}{2} \|x - \bar{x}^{++}\|^2. \quad (2.38)$$

Suppose we have now available two quadratic lower-estimators:

$$f(x) \geq Q_A(x) := v_A + \frac{\alpha}{2} \|x - x_A\|^2,$$

$$f(x) \geq Q_B(x) := v_B + \frac{\alpha}{2} \|x - x_B\|^2.$$ 

Clearly, the minimal values of $Q_A$ and of $Q_B$ lower-bound the minimal value of $f$. For any $\lambda \in [0,1]$, the average $Q_\lambda := \lambda Q_A + (1-\lambda)Q_B$ is again a quadratic lower-estimator of $f$. Thus we are led to the question: what choice of $\lambda$ yields the tightest lower-bound on the minimal value of $f$? To answer this question, observe the equality

$$Q_\lambda(x) := \lambda Q_A(x) + (1-\lambda)Q_B(x) = v_\lambda + \frac{\alpha}{2} \|x - c_\lambda\|^2,$$

where

$$c_\lambda = \lambda x_A + (1-\lambda)x_B$$

and

$$v_\lambda = v_B + \left( v_A - v_B + \frac{\alpha}{2} \|x_A - x_B\|^2 \right) \lambda - \left( \frac{\alpha}{2} \|x_A - x_B\|^2 \right) \lambda^2. \quad (2.39)$$

In particular, the average $Q_\lambda$ has the same canonical form as $Q_A$ and $Q_B$. A quick computation now shows that $v_\lambda$ (the minimum of $Q_\lambda$) is maximized by setting

$$\bar{\lambda} := \text{proj}_{[0,1]} \left( \frac{1}{2} + \frac{v_A - v_B}{\alpha \|x_A - x_B\|^2} \right).$$

With this choice of $\lambda$, we call the quadratic function $\bar{Q} = \bar{v} + \frac{\alpha}{2} \|x - \bar{c}\|^2$ the \textit{optimal averaging} of $Q_A$ and $Q_B$. See Figure 2.3 for an illustration.

An algorithmic idea emerges. Given a current iterate $x_k$, form the quadratic lower-model $Q(\cdot)$ in (2.38) with $\bar{x} = x_k$. Then let $Q_k$ be the optimal averaging of $Q$ and the quadratic lower model $Q_{k-1}$ from the previous step. Finally define $x_{k+1}$ to be the minimizer of $Q_k$, and repeat.
CHAPTER 2. SMOOTH MINIMIZATION

Figure 2.3: The optimal averaging of $Q_A(x) = 1 + 0.5(x + 2)^2$ and $Q_B(x) = 3 + 0.5(x - 4)^2$.

Though attractive, the scheme does not converge at an optimal rate. The main idea behind acceleration is a separation of roles: one must maintain two sequences of points $x_k$ and $c_k$. The points $x_k$ will generate quadratic lower models as above, while $c_k$ will be the minimizers of the quadratics. The proposed method is summarized in Algorithm 6.

\begin{algorithm}
\textbf{Algorithm 6: Optimal Quadratic Averaging}
\begin{itemize}
  \item \textbf{Input:} Starting point $x_0$ and strong convexity constant $\alpha > 0$.
  \item \textbf{Output:} Final quadratic $Q_K(x) = v_K + \frac{\alpha}{2} \|x - c_K\|^2$ and $x^+_K$.
  \item Set $Q_0(x) = v_0 + \frac{\alpha}{2} \|x - c_0\|^2$, where $v_0 = f(x_0) - \frac{\|\nabla f(x_0)\|^2}{2\alpha}$ and $c_0 = x_0^+$;
  \item for $k = 1, \ldots, K$ do
    \begin{itemize}
      \item Set $x_k = \text{line search}(c_{k-1}, x_{k-1}^+)$;
      \item Set $Q(x) = \left(f(x_k) - \frac{\|\nabla f(x_k)\|^2}{2\alpha}\right) + \frac{\alpha}{2} \|x - x_k^+\|^2$;
      \item Let $Q_k(x) = v_k + \frac{\alpha}{2} \|x - c_k\|^2$ be the optimal averaging of $Q$ and $Q_{k-1}$;
    \end{itemize}
  \end{itemize}
\end{algorithm}

The analysis of the scheme relies on the following easy observation.

\textbf{Lemma 2.39.} Suppose that $\overline{Q} = \bar{v} + \frac{\alpha}{2} \|\bar{x}\|^2$ is the optimal averaging of the quadratics $Q_A = v_A + \frac{\alpha}{2} \|x_A\|^2$ and $Q_B = v_B + \frac{\alpha}{2} \|x_B\|^2$. Then the quantity $\bar{v}$ is nondecreasing in both $v_A$ and $v_B$. Moreover, whenever the
2.7. OPTIMAL METHODS FOR SMOOTH CONVEX MINIMIZATION

Inequality $|v_A - v_B| \leq \frac{\alpha}{2} \|x_A - x_B\|^2$ holds, we have

$$\bar{v} = \frac{\alpha}{8} \|x_A - x_B\|^2 + \frac{1}{2} (v_A + v_B) + \frac{1}{2\alpha} \left( \frac{v_A - v_B}{\|x_A - x_B\|} \right)^2.$$ 

Proof. Define $\hat{\lambda} := \frac{1}{2} + \frac{v_A - v_B}{\alpha \|x_A - x_B\|^2}$. Notice that we have

$$\hat{\lambda} \in [0, 1] \quad \text{if and only if} \quad |v_A - v_B| \leq \frac{\alpha}{2} \|x_A - x_B\|^2.$$ 

If $\hat{\lambda}$ lies in $[0, 1]$, equality $\bar{\lambda} = \hat{\lambda}$ holds, and then from (2.39) we deduce

$$\bar{v} = v_{\bar{\lambda}} = \frac{\alpha}{8} \|x_A - x_B\|^2 + \frac{1}{2} (v_A + v_B) + \frac{1}{\alpha} \left( \frac{v_A - v_B}{\|x_A - x_B\|} \right)^2.$$ 

If $\hat{\lambda}$ does not lie in $[0, 1]$, then an easy argument shows that $\bar{v}$ is linear in $v_A$ either with slope one or zero. If $\hat{\lambda}$ lies in $(0, 1)$, then we compute

$$\frac{\partial \bar{v}}{\partial v_A} = \frac{1}{2} + \frac{1}{\alpha \|x_A - x_B\|^2} (v_A - v_B),$$

which is nonnegative because $\frac{|v_A - v_B|}{\alpha \|x_A - x_B\|^2} \leq \frac{1}{2}$. Since $\bar{v}$ is clearly continuous, it follows that $\bar{v}$ is nondecreasing in $v_A$, and by symmetry also in $v_B$. \qed

The following theorem shows that Algorithm 6 achieves the optimal linear rate of convergence.

**Theorem 2.40** (Convergence of optimal quadratic averaging). In Algorithm 6, for every index $k \geq 0$, the inequalities $v_k \leq f^* \leq f(x^+_k)$ hold and we have

$$f(x^+_k) - v_k \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k (f(x^+_0) - v_0).$$

Proof. Since in each iteration, the algorithm only averages quadratic minors of $f$, the inequalities $v_k \leq f^* \leq f(x^+_k)$ hold for every index $k$. Set $r_0 = \frac{2}{\alpha} (f(x^+_0) - v_0)$ and define the quantities $r_k := \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k r_0$. We will show by induction that the inequality $v_k \geq f(x^+_k) - \frac{\alpha}{2} r_k$ holds for all $k \geq 0$. The base case $k = 0$ is immediate, and so assume we have

$$v_{k-1} \geq f(x^+_{k-1}) - \frac{\alpha}{2} r_{k-1}$$

for some index $k - 1$. Next set $v_A := f(x_k) - \frac{\|\nabla f(x_k)\|^2}{2\alpha}$ and $v_B := v_{k-1}$. Then the function

$$Q_k(x) = v_k + \frac{\alpha}{2} \|x - c_k\|^2,$$
is the optimal averaging of $Q_A(x) = v_A + \frac{\alpha}{2} \|x - x_k^+\|^2$ and $Q_B(x) = v_B + \frac{\alpha}{2} \|x - c_{k-1}\|^2$. Taking into account the inequality $f(x_k^+) \leq f(x_k) - \frac{1}{2\beta} \|\nabla f(x_k)\|^2$ yields the lower bound $\hat{v}_A$ on $v_A$:

$$v_A = f(x_k) - \frac{\|\nabla f(x_k)\|^2}{2\alpha} \geq f(x_k^+) - \frac{\alpha}{2} \frac{\|\nabla f(x_k)\|^2}{\alpha^2} \left(1 - \frac{1}{\kappa}\right) := \hat{v}_A.$$ 

The induction hypothesis and the choice of $x_k$ yield a lower bound $\hat{v}_B$ on $v_B$:

$$v_B \geq f(x_{k-1}^+) - \frac{\alpha}{2} r_{k-1} \geq f(x_k) - \frac{\alpha}{2} r_{k-1}$$

$$\geq f(x_k^+) + \frac{1}{2\beta} \|\nabla f(x_k)\|^2 - \frac{\alpha}{2} r_{k-1}$$

$$= f(x_k^+) - \frac{\alpha}{2} \left(r_{k-1} - \frac{1}{\alpha^2\kappa} \|\nabla f(x_k)\|^2\right) := \hat{v}_B.$$ 

Define the quantities $d := \|x_k^+-c_{k-1}\|$ and $h := \|\nabla f(x_k)\|$. We now split the proof into two cases. First assume $h^2 \leq \frac{r_{k-1}}{2\alpha}$. Then we deduce

$$v_k \geq v_A \geq \hat{v}_A = f(x_k^+) - \frac{\alpha}{2} h^2 \left(1 - \frac{1}{\kappa}\right)$$

$$\geq f(x_k^+) - \frac{\alpha}{2} r_{k-1} \left(1 - \frac{1}{\kappa}\right)$$

$$\geq f(x_k^+) - \frac{\alpha}{2} r_{k-1} \left(1 - \frac{1}{\sqrt{\kappa}}\right)$$

$$= f(x_k^+) - \frac{\alpha}{2} r_k.$$ 

Hence in this case, the proof is complete.

Next suppose $h^2 > \frac{r_{k-1}}{2\alpha}$ and let $v + \frac{\alpha}{2} \cdot -c\|^2$ be the optimal average of the two quadratics $\hat{v}_A + \frac{\alpha}{2} \cdot -x_{k-1}\|^2$ and $\hat{v}_B + \frac{\alpha}{2} \cdot -c_{k-1}\|^2$. By Lemma 2.39 the inequality $v_k \geq v$ holds. We claim that equality

$$v = \hat{v}_B + \frac{\alpha}{8} \left(d^2 + \frac{2}{\alpha} (\hat{v}_A - \hat{v}_B)^2\right)$$

holds. (2.40)

This follows immediately from Lemma 2.39 once we show $\frac{1}{2} \geq \frac{|\hat{v}_A - \hat{v}_B|}{\alpha d^2}$. To this end, note first the equality $|\hat{v}_A - \hat{v}_B| = \frac{|r_{k-1} - h^2|}{2\alpha d^2}$. The choice $x_k = \text{line_search}(c_{k-1}, x_{k-1}^+)$ ensures:

$$d^2 - h^2 = \|x_k - c_{k-1}\| - \frac{2}{\alpha} (\nabla f(x_k), x_k - c_{k-1}) = \|x_k - c_{k-1}\|^2 \geq 0.$$ 

Thus we have $h^2 - r_{k-1} < h^2 \leq d^2$. Finally, the assumption $h^2 > \frac{r_{k-1}}{2}$ implies

$$r_{k-1} - h^2 < \frac{r_{k-1}}{2} < h^2 \leq d^2.$$ (2.41)
Hence we can be sure that (2.40) holds. Plugging in $\hat{v}_A$ and $\hat{v}_B$ yields
\[ v = f(x_k^+) - \frac{\alpha}{2} \left( r_{k-1} - \frac{1}{\kappa} h^2 - \frac{(d^2 + r_{k-1} - h^2)^2}{4d^2} \right). \]

Hence the proof is complete once we show the inequality
\[ r_{k-1} - \frac{1}{\kappa} h^2 - \frac{(d^2 + r_{k-1} - h^2)^2}{4d^2} \leq \left( 1 - \frac{1}{\sqrt{\kappa}} \right) r_{k-1}. \]

After rearranging, our task simplifies to showing
\[ \frac{r_{k-1}}{\sqrt{\kappa}} \leq \frac{h^2}{\kappa} + \frac{(d^2 + r_{k-1} - h^2)^2}{4d^2}. \]

Taking derivatives and using inequality (2.41), one can readily verify that the right-hand-side is nondecreasing in $d^2$ on the interval $d^2 \in [h^2, +\infty)$. Thus plugging in the endpoint $d^2 = h^2$ we deduce
\[ \frac{h^2}{\kappa} + \frac{(d^2 + r_{k-1} - h^2)^2}{4d^2} \geq \frac{h^2}{\kappa} + \frac{r_{k-1}^2}{4h^2}. \]

Minimizing the right-hand-side over all $h$ satisfying $h^2 \geq \frac{r_{k-1}}{2}$ yields the inequality
\[ \frac{h^2}{\kappa} + \frac{r_{k-1}^2}{4h^2} \geq \frac{r_{k-1}}{\sqrt{\kappa}}. \]

The proof is complete.

A nice feature of the quadratic averaging viewpoint is that one can empirically speed up the algorithm by optimally averaging more than two quadratics each time.

**Exercise 2.41.** Fix $t$ quadratics $Q_i(x) := v_i + \frac{\alpha}{2} \| x - c_i \|^2$, with $i \in \{1, \ldots, t\}$. Define the matrix $C = [c_1 \ c_2 \ \ldots \ c_t]$ and vector $v = [v_1 \ v_2 \ \ldots \ v_t]^T$.

1. For any $\lambda \in \Delta_t$, show that the average quadratic
\[ Q_\lambda(x) := \sum_{i=1}^{t} \lambda_i Q_i(x) \]

maintains the same canonical form as each $Q_i$. More precisely, show the representation
\[ Q_\lambda(x) = v_\lambda + \frac{\alpha}{2} \| x - c_\lambda \|^2, \]

where
\[ c_\lambda = C\lambda \quad \text{and} \quad v_\lambda = \left\langle \frac{\alpha}{2} \text{diag} (C^T C) + v, \lambda \right\rangle - \frac{\alpha}{2} \| C\lambda \|^2. \]
2. Deduce that the optimal quadratic averaging problem

\[
\max_{\lambda \in \Delta_t} \min_x \sum_{i=1}^t \lambda_i Q_i(x)
\]

is equivalent to the convex quadratic optimization problem

\[
\min_{\lambda \in \Delta_t} \frac{\alpha}{2} \|C\lambda\|^2 - \left(\frac{\alpha}{2} \text{diag} (C^T C) + v, \lambda\right).
\]

**Exercise 2.42** (Ridge Regression). In this exercise, you will consider the ridge regression problem:

\[
\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - y\|^2_2 + \frac{\lambda}{2} \|x\|^2_2,
\]

which aims to recover a point \(x_{opt}\) from (noisy) linear measurements \(y\). For \(n = 100\), \(m = 80\), and \(\lambda = 1\), generate data as follows:

- the underlying signal is drawn \(x_{opt} \sim N(0, I)\),
- the measurement matrix \(A \in \mathbb{R}^{m \times n}\) is drawn with independent standard Gaussian rows \(A_i \sim N(0, I)\),
- the observed data \(y \in \mathbb{R}^m\) is\(y = Ax_{opt} + \epsilon, \epsilon_i \sim N(0, 0.25)\) i.i.d.

Note that the objective function is \(\beta\)-smooth with \(\beta = \|A^*A\|_{op} + \lambda\), and \(\alpha\)-strongly convex with \(\alpha = \lambda\). In your experiments, you may set \(\beta = 4(m + n) + \lambda^4\). Write code that generates the problem data as above and implement the following algorithms:

1. gradient descent
2. Algorithm 2 in the course notes
3. Algorithm 5 in the course notes.

For each algorithm, plot the the function value over the first 50 iterations on a semilog plot.

**Exercise 2.43** (Huber regression). Consider the huber-ized version of ridge regression:

\[
\min_{x \in \mathbb{R}^n} \sum_{i=1}^m h_\eta(y_i - a_i^T x) + \frac{\lambda}{2} \|x\|^2_2
\]

\(^1\)By Gaussian concentration, with high probability we have \(\|A\|_{op} \lesssim \sqrt{m} + \sqrt{n}\).
where \( a_i^T \) are the rows of \( A \), and \( h_\eta \) is the huber function with parameter \( \eta \):

\[
h_\eta(w) = \begin{cases} 
\frac{1}{2}w^2, & \text{if } |w| \leq \eta \\
\eta(|w| - \eta/2), & \text{otherwise.}
\end{cases}
\]

The huber function penalizes large deviations less than the quadratic cost function, and is therefore often used when outliers are present in the data. Note that the objective function in this case is \( \beta \)-smooth with \( \beta > m\eta + \lambda \) and \( \alpha \)-strongly convex with \( \alpha = \lambda \).

Generate \( x_{\text{opt}} \) and \( A \) as in Exercise 1, and generate \( y = Ax_{\text{opt}} + \epsilon \) where \( \epsilon \) contains 5 outliers drawn from \( N(0, 25) \) and the rest of its entries are again independent \( N(0, 0.25) \). Implement the same three algorithms as in Exercise 1 on this problem, and plot the function values for the first 100 iterations in a semilog plot.

**Exercise 2.44** (Logistic regression). Consider the regularized logistic regression problem:

\[
\min_{\theta \in \mathbb{R}^n} \sum_{i=1}^m \ln(1 + \exp(-y_i\theta^T x_i)) + \frac{\lambda}{2}\|\theta\|_2^2.
\]

Here \( y_i \in \{\pm 1\} \) are binary labels for the data points \( x_i \in \mathbb{R}^n \), and we are trying to find the best vector \( \theta \) of parameters for the model

\[
P(y = 1|x, \theta) = \frac{1}{1 + e^{-\theta^T x}}.
\]

(We are doing this by minimizing the negative log likelihood function plus a regularization term.)

Implement the same three algorithms as before, with \( n = 50 \) and \( m = 100 \) and \( \lambda = 1 \). Generate data as follows:

1. \( \theta_{\text{opt}} = (1, \ldots, 1)^T \);
2. \( X \in \mathbb{R}^{m \times n} \) has i.i.d. standard normal entries, with rows \( x_i^T \);
3. \( z = X\theta_{\text{opt}} \) and the true vector of probabilities is \( p = 1/(1 + \exp(-z)) \), where the operations are applied entrywise to the vector \( z \);
4. \( y \in \{\pm 1\}^m \) has independent Bernoulli entries with \( P(y_i = 1) = p_i \).
5. \( \theta_0 \) has i.i.d. standard normal entries.

Implement the same three algorithms as in Exercise 1 on this problem, and plot the function values for the first 100 iterations in a semilog plot. You may take the smoothness parameter of the objective function to be \( \beta = 100 \) and the strong convexity parameter to be \( \alpha = 1 \).
CHAPTER 2. SMOOTH MINIMIZATION

References.
The convergence guarantees of gradient descent in Section 2.4.1 are classical and can be found for example in the textbooks [2, 9]. Lower complexity bounds originate in [8]; our discussion in Subsection 2.5 follows the text [9]. Our discussion of the conjugate gradient algorithm in Section 2.6 follows the monograph [13]. The fast gradient method in Section 2.7.1 and the restart strategy originate with Nesterov [11]; the proof of convergence we present is from [1]. The technique of estimate sequences in Section 2.7.2 was introduced in [12]. The quadratic averaging algorithm originates in [5], and is formally equivalent to the earlier geometric descent algorithm [4]. Estimates on the gradient norms (2.4), (2.26) follow from [10].
Chapter 3

Convex geometry and analysis

Algorithms for minimizing smooth convex functions rely heavily on basic results of mathematical analysis, summarized in Section 1.5. Algorithms for minimizing nonsmooth convex functions require a deeper dive into convexity. This chapter is devoted to developing the main guiding principles of convex geometry and analysis, with a focus on building geometric intuition.

3.1 Basic convex geometry

Convex analysis is a study of convex functions. At its core, convex analysis is based on the geometry of convex sets – the content of this section. Recall for any two points $x, y \in E$, the closed line segment joining $x$ and $y$ is

$$[x, y] := \{\lambda x + (1 - \lambda)y : 0 \leq \lambda \leq 1\}.$$ 

A set $Q \subseteq E$ is convex if for any two points $x, y \in Q$, the line segment $[x, y]$ is also contained in $Q$. Define the unit simplex:

$$\Delta_n = \left\{ \lambda \in \mathbb{R}^n : \sum_{i=1}^{n} \lambda_i = 1, \lambda \geq 0 \right\}.$$ 

We say that a point $x$ is a convex combination of points $x_1, \ldots, x_k \in E$ if it can be written as $x = \sum_{i=1}^{k} \lambda_i x_i$ for some $\lambda \in \Delta_k$.

**Exercise 3.1.** Show that a set $Q \subset E$ is convex if and only if any convex combination of points $x_1, \ldots, x_t \in Q$ lies in $Q$ for any integer $t \geq 2$.

The following exercise shows that convexity is a very stable property, being preserved under a variety of operations.

**Exercise 3.2.** Prove the following statements.
1. **(Pointwise sum)** For any two convex sets \( Q_1, Q_2 \subset E \), the sum
\[
Q_1 + Q_2 := \{x + y : x \in Q_1, y \in Q_2\}
\]
is convex.

2. **(Intersection)** The intersection \( \bigcap_{i \in I} Q_i \) of any convex sets \( Q_i \), indexed by an arbitrary set \( I \), is convex.

3. **(Linear image/preimage)** For any convex sets \( Q \subset E \) and \( L \in Y \) and linear maps \( A : E \to Y \) and \( A^{-1}L \) are convex sets.

The **convex hull** of a set \( Q \subseteq E \), denoted \( \text{conv}(Q) \), is the intersection of all convex sets containing \( Q \). The following exercise shows that equivalently \( \text{conv}(Q) \) is the set of all convex combinations of points in \( Q \).

**Exercise 3.3.** For any set \( Q \subset E \), prove the equality:
\[
\text{conv}(Q) = \left\{ \sum_{i=1}^{k} \lambda_i x_i : k \in \mathbb{N}_+, \ x_1, \ldots, x_k \in Q, \ \lambda \in \Delta_k \right\}.
\] (3.1)

The following theorem shows that in the description (3.1), it suffices to take \( k \leq n + 1 \).

**Theorem 3.4 (Carathéodory).** Consider a set \( Q \subset E \), where \( E \) is an \( n \)-dimensional Euclidean space. Then any point \( x \in \text{conv}(Q) \) can be written as a convex combination of some \( n + 1 \) points in \( Q \).

**Proof.** Since \( x \) belongs to \( \text{conv}(Q) \), we may write \( x = \sum_{i=1}^{k} \lambda_i x_i \) for some integer \( k \), points \( x_1, \ldots, x_k \in Q \), and multipliers \( \lambda \in \Delta_k \). If the inequality \( k \leq n + 1 \) holds, then there is nothing to prove. Hence, suppose \( k \geq n + 2 \). Then the vectors
\[
x_2 - x_1, \ldots, x_k - x_1
\]
are linearly dependent. That is there exists numbers \( \mu_i \) for \( i = 2, \ldots, k \) not all zero and satisfying
\[
0 = \sum_{i=2}^{k} \mu_i (x_i - x_1) = \sum_{i=2}^{k} \mu_i x_i - (\sum_{i=2}^{k} \mu_i)x_1.
\]
Defining \( \mu_1 := -\sum_{i=2}^{k} \mu_i \), we deduce \( \sum_{i=1}^{k} \mu_i x_i = 0 \) and \( \sum_{i=1}^{k} \mu_i = 0 \). Then for any real number \( \alpha \) we obtain the equalities
\[
x = \sum_{i=1}^{k} \lambda_i x_i - \alpha \sum_{i=1}^{k} \mu_i x_i = \sum_{i=1}^{k} (\lambda_i - \alpha \mu_i)x_i
\]
and
\[
\sum_{i=1}^{k} (\lambda_i - \alpha \mu_i) = 1.
\]
3.1. BASIC CONVEX GEOMETRY

We will now choose $\alpha$ so that all the coefficients $\lambda_i - \alpha \mu_i$ are nonnegative and at least one of them is zero. Indeed, simply choose an index $i^* \in \text{argmin}_i \{\lambda_i/\mu_i : \mu_i > 0\}$. Hence $x$ is a convex combination of $k - 1$ points, as the coefficient $\lambda_{i^*} - \alpha \mu_{i^*}$ is zero. Continuing this process, we will obtain a description of $x$ as a convex combination of $k \leq n + 1$ points.

Often, convex sets have empty interior. On the other hand, we will now see that any nonempty convex set has nonempty interior relative to the smallest affine subspace containing the convex set. To make this observation precise, let us introduce the following definitions. The affine hull of a convex set $Q$, denoted $\text{aff}(Q)$, is the intersection of all affine sets containing $Q$. Clearly, $\text{aff}(Q)$ is itself an affine set. The relative interior of $Q$, denoted $\text{ri} Q$, is the interior of $Q$ relative to $\text{aff}(Q)$, that is

$$\text{ri} Q := \{x \in Q : \exists \epsilon > 0 \text{ s.t. } B_{\epsilon}(x) \cap \text{aff}(Q) \subseteq Q\}.$$  

The relative boundary of $Q$, denoted $\text{rb} Q$, is defined by $\text{rb} Q := Q \setminus (\text{ri} Q)$.

**Theorem 3.5** (Relative interior is nonempty). For any nonempty convex set $Q \subset E$, the relative interior $\text{ri} Q$ is nonempty.

**Proof.** Without loss of generality, we may translate $Q$ to contain the origin. Let $d$ be the dimension of the linear subspace $\text{aff} Q$. Observe that $Q$ must contain some $d$ linearly independent vectors $x_1, \ldots, x_d$, since otherwise $\text{aff} Q$ would have a smaller dimension than $d$. Consider the linear map $A : \mathbb{R}^d \to \text{aff} Q$, given by $A(\lambda_1, \ldots, \lambda_d) = \sum_{i=1}^d \lambda_i x_i$. Since the range of $A$ contains $x_1, \ldots, x_d$, the map $A$ is surjective. Hence $A$ is a linear isomorphism. Consequently $A$ maps the open set

$$\Omega := \left\{ \lambda \in \mathbb{R}^d : \lambda_i > 0 \text{ for all } i, \sum_{i=1}^d \lambda_i < 1 \right\}$$

to an open subset of $\text{aff} Q$. Note for any $\lambda \in \Omega$, we can write $A\lambda = \sum_{i=1}^d \lambda_i x_i + (1 - \sum_{i=1}^d \lambda_i) \cdot 0$. Hence, convexity of $Q$ implies $A(\Omega) \subseteq Q$, thereby proving the claim.

The following is a useful topological property of convex sets.

**Theorem 3.6** (Accessibility). Consider a convex set $Q$ and two points $x \in \text{ri} Q$ and $y \in \text{cl} Q$. Then the line segment $[x, y]$ is contained in $\text{ri} Q$.

**Proof.** Without loss of generality, suppose that the affine hull of $Q$ is all of $E$. Then since $x$ lies in the interior of $Q$, there is $\epsilon > 0$ satisfying $B_{\epsilon}(x) \subseteq Q$. Define the set $\Lambda := \{\lambda z + (1 - \lambda) y : z \in B_\epsilon(x), \lambda \in (0, 1)\}$. Since $Q$ is convex, $\Lambda$ is an open set satisfying $[x, y] \subset Q$. The result follows.

**Corollary 3.7.** For any nonempty convex set $Q$ in $E$, we have $\text{cl} (\text{ri} Q) = \text{cl} Q$. 

Proof. The inclusion \( \text{ri} \ Q \subseteq \overline{Q} \) immediately implies \( \text{cl} (\text{ri} \ Q) \subseteq \text{cl} Q \). Conversely, fix a point \( y \in \text{cl} Q \). Since \( \text{ri} \ Q \) is nonempty by Theorem 3.5, we may also choose a point \( x \in \text{ri} Q \). Theorem 3.6 then immediately implies \( y \in \text{cl} [x, y] \subseteq \text{cl} (\text{ri} \ Q) \). Since \( y \in \text{cl} Q \) is arbitrary, we have established the equality \( \text{cl} (\text{ri} \ Q) = \text{cl} Q \).

\[ \square \]

**Theorem 3.8.** (Calculus of relative interiors) For any two convex sets \( Q \in E, P \in Y \) and a linear map \( A : E \to Y \), equality holds:

\[ \text{ri} (A(Q) + P) = A(\text{ri} Q) + \text{ri} P. \]

### 3.1.1 Separation theorems

A foundational result of convex geometry shows that there are two ways to think about a closed convex set \( Q \). Tautologically \( Q \) is simply a collection of points. On the other hand, we will show in this section that \( Q \) coincides with the intersection of all half-spaces containing \( Q \). Such a description of \( Q \) is often called a dual representation.

We begin with the following basic definitions. Along with any set \( Q \subset E \) we define the distance function

\[ \text{dist}_Q(y) := \inf_{x \in Q} \|x - y\|; \]

and the projection

\[ \text{proj}_Q(y) := \{ x \in Q : \text{dist}_Q(y) = \|x - y\| \}. \]

Thus \( \text{proj}_Q(y) \) consists of all the nearest points of \( Q \) to \( y \).

**Exercise 3.9.** Show that for any nonempty set \( Q \subseteq E \), the function \( \text{dist}_Q : E \to \mathbb{R} \) is 1-Lipschitz.

If \( Q \) is closed, then the nearest-point set \( \text{proj}_Q(y) \) is nonempty for any \( y \in E \). To see this, fix a point \( \bar{x} \in Q \) and set \( r := \|y - \bar{x}\| \). Then by the extreme value theorem, the function \( x \mapsto \|x - y\| \) attains its minimum over the nonempty compact set \( Q \cap B_r(y) \). A bit of thought shows that this minimizer must lie in \( \text{proj}_Q(y) \). When \( Q \) is convex, the set \( \text{proj}_Q(y) \) is not only nonempty, but is also a singleton.

**Theorem 3.10** (Properties of the projection). For any nonempty, closed, convex set \( Q \subset E \), the set \( \text{proj}_Q(y) \) is a singleton. Moreover, the closest point \( z \in Q \) to \( y \) is characterized by the the property:

\[ \langle y - z, x - z \rangle \leq 0 \quad \text{for all } x \in Q. \]

(3.2)
3.1. BASIC CONVEX GEOMETRY

Proof. Let $Q$ be a nonempty, closed, convex set. Fix $y \in \mathbf{E}$ and set $r := \text{dist}_Q(y)$. If $r = 0$, the theorem holds trivially; hence, we suppose $y \notin Q$.

The claim that any point $z$ satisfying (3.2) lies in $\text{proj}_Q(y)$ is an easy exercise (verify it!). We therefore prove the converse. Since $Q$ is closed, the set $\text{proj}_Q(y)$ is nonempty. Fix a point $z \in \text{proj}_Q(y)$ and define

$$H := \{ x \in \mathbf{E} : \langle y - z, x - z \rangle > 0 \}.$$ 

We will show $H \cap Q = \emptyset$. Indeed, for the sake of contradiction, suppose there is a point $x \in H \cap Q$. Then convexity of $Q$ implies $[x, z] \subset Q$, while the definition of $H$ shows that the segment $[x, z]$ intersects the open ball $B_r(y)$, thereby contradicting the inclusion $z \in \text{proj}_Q(y)$. We conclude that (3.2) holds. To see that $\text{proj}_Q(y)$ is a singleton, consider another point $z' \in \text{proj}_Q(y)$. Then clearly $z'$ lies in the intersection $(\text{cl} B_r(y)) \cap (\mathbf{E} \setminus H)$. The definition of $H$ on the other hand, implies that this intersection is the singleton $\{z\}$. 

Exercise 3.11. Show that for any closed convex set $Q \subset \mathbf{E}$, the nearest point map $x \mapsto \text{proj}_Q(x)$ is 1-Lipschitz.

The following is a fundamental property of convex sets, which we will often use.

**Theorem 3.12** (Strict separation). Consider a closed convex set $Q \subset \mathbf{E}$ and a point $y \notin Q$. Then there is nonzero vector $a \in \mathbf{E}$ and a number $b \in \mathbb{R}$ satisfying

$$\langle a, x \rangle \leq b < \langle a, y \rangle$$

for any $x \in Q$.

Proof. Define the nonzero vector $a := y - \text{proj}_Q(y)$. Then for any $x \in Q$, the condition (3.2) yields the inequalities

$$\langle a, x \rangle \leq \langle a, \text{proj}_Q(y) \rangle = \langle a, y \rangle - \|a\|^2 < \langle a, y \rangle,$$

as claimed. 

In particular, one can now establish the following “dual description” of convex sets, alluded to in the beginning of the section.

**Exercise 3.13.** Given a nonempty set $Q \subset \mathbf{E}$, define

$$\mathcal{F}_Q := \{ (a, b) \in \mathbf{E} \times \mathbb{R} : \langle a, x \rangle \leq b \text{ for all } x \in Q \}.$$ 

Prove that the equality holds:

$$\text{cl conv}(Q) = \bigcap_{(a, b) \in \mathcal{F}_Q} \{ x \in \mathbf{E} : \langle a, x \rangle \leq b \}.$$ 

[Hint: Use the separation theorem.]
3.1.2 Cones and polarity

A particularly nice class of convex sets consists of those that are positively homogeneous. A set $K \subseteq E$ is called a cone if the inclusion $\lambda K \subseteq K$ holds for any $\lambda \geq 0$. For example, the nonnegative orthant $\mathbb{R}_{+}^{n}$ and the set of positive semidefinite matrices $S_{+}^{n}$ are closed convex cones.

**Exercise 3.14.** Show that a set $K \subseteq E$ is a convex cone if and only if for any two points $x, y \in K$ and numbers $\lambda, \mu \geq 0$ the point $\lambda x + \mu y$ lies in $K$.

**Exercise 3.15.** Prove for any convex cone $K \subseteq E$ the equality $\text{aff}(K) = K - K$.

Convex cones behave similarly to linear subspaces. In particular, the following operation is an analogue for cones of taking the orthogonal complement of a linear subspace. For any cone $K \subset E$, the *polar cone* is the set

$$K^o := \{ v \in E : \langle v, x \rangle \leq 0 \text{ for all } x \in K \}.$$  

Thus $K^o$ consists of all vectors $v$ that make an obtuse angle with every vector $x \in K$. For example, the reader should convince themselves of the equalities, $(\mathbb{R}_{+}^{n})^o = \mathbb{R}^{n}$ and $(S_{+}^{n})^o = S^{n}$.

**Exercise 3.16** (Double-polar theorem). For any cone $K$, prove the equality

$$(K^o)^o = \text{cl conv}(K).$$

[Hint: use the separation Theorem 3.12]

Classically, the orthogonal complement to a sum of linear subspaces is the intersection of the orthogonal complements. In much the same way, the polarity operation satisfies “calculus rules”.

**Theorem 3.17** (Polarity calculus). For any linear mapping $A : E \rightarrow Y$ and a cone $K \subset Y$, the chain rule holds

$$(AK)^o = (A^*)^{-1}K^o.$$  

In particular, for any two cones $K_1, K_2 \subset E$, the sum rule holds:

$$(K_1 + K_2)^o = K_1^o \cap K_2^o.$$  

**Proof.** Observe the equivalence

$$y \in (AK)^o \iff \langle Ax, y \rangle \leq 0 \text{ for all } x \in K$$

$$\iff \langle x, A^*y \rangle \leq 0 \text{ for all } x \in K$$

$$\iff y \in (A^*)^{-1}K^o.$$  

This establishes the first equality. The sum rule follows by applying the chain rule to the expression $A(K_1 \times K_2)$ with the mapping $A(x, y) := x + y$.  □
A natural question is how to define a useful notion of polarity for general sets, i.e. those that are not cones. The answer is based on “homogenizing” the set and then applying the polarity operation for cones. Consider a set \( Q \subset E \) and let \( K \) be the cone generated by \( Q \times \{1\} \subset E \times \mathbb{R} \). That is
\[
K = \{(\lambda x, \lambda) \in E \times \mathbb{R} : x \in E, \lambda \geq 0\}.
\]
It is then natural to define the polar set as
\[
Q^o := \{x \in E : (x, -1) \in K^o\}.
\]
Unraveling the definitions, the following algebraic description of the polar appears.

**Exercise 3.18.** Show for any set \( Q \subset E \), the equality
\[
Q^o = \{v \in E : \langle v, x \rangle \leq 1 \text{ for all } x \in Q\}.
\]

Notice that if \( Q \) is a cone, than the above definition of the polar coincides with the definition of the polar we have given for cones. The following is a direct analogue of Theorem 3.16

**Exercise 3.19 (Double polar).** For any set \( Q \subset E \) containing the origin, we have
\[
(Q^o)^o = \text{cl conv}(Q).
\]

### 3.1.3 Tangents and normals

As we have seen, a principal technique of smooth minimization is to form first-order approximations of the underlying function. Let us look at this idea more broadly by constructing first-order approximations of sets.

Consider a set \( Q \subset E \) and a point \( \bar{x} \in Q \). Intuitively, we should think of a first order approximation to \( Q \) at \( \bar{x} \) as the set of all limits of rays \( R_+(x_i - \bar{x}) \) over all possible sequences \( x_i \in Q \) tending to \( \bar{x} \). With this in mind, define the tangent cone to \( Q \) at \( \bar{x} \) by
\[
T_Q(\bar{x}) := \left\{ \lim_{i \to \infty} \frac{x_i - \bar{x}}{\tau_i} : x_i \to \bar{x} \text{ in } Q, \tau_i \searrow 0 \right\}.
\]
The reader should convince themselves that \( T_Q(\bar{x}) \) is a closed cone. Whenever \( Q \) is convex, this definition simplifies drastically.

**Exercise 3.20.** Show for any convex set \( Q \subset E \) and a point \( \bar{x} \in Q \) the equality:
\[
T_Q(\bar{x}) = \text{cl } R_+(Q - \bar{x}) := \text{cl} \{\lambda(x - \bar{x}) : \lambda \geq 0, x \in Q\}.
\]
Tangency has to do with directions pointing into the set. Alternatively, we can also think dually of outward normal vectors to a set $Q$ at $\bar{x} \in Q$. Geometrically, it is intuitive to call a vector $v$ an (outward) normal to $Q$ at $\bar{x}$ if $Q$ is fully contained in the half-space \( \{ x \in \mathbb{E} : \langle v, x - \bar{x} \rangle \leq 0 \} \) up to a first-order error. More precisely, the normal cone to a set $Q \subset \mathbb{E}$ at a point $\bar{x} \in Q$ is defined by

\[
N_Q(\bar{x}) := \{ v \in \mathbb{E} : \langle v, x - \bar{x} \rangle \leq o(\|x - \bar{x}\|) \text{ as } x \to \bar{x} \text{ in } Q \}.
\]

The reader should convince themselves that $N_Q(\bar{x})$ is a closed convex cone.

Again, when $Q$ is convex, the definition simplifies.

**Exercise 3.21.** Show for any convex set $Q \subset \mathbb{E}$ and a point $\bar{x} \in Q$ the equality,

\[
N_Q(\bar{x}) = \{ v \in \mathbb{E} : \langle v, x - \bar{x} \rangle \leq 0 \text{ for all } x \in Q \}
\]

and the polarity correspondence

\[
N_Q(\bar{x}) = (T_Q(\bar{x}))^\circ.
\]

Thus the $o(\|x - \bar{x}\|)$ error in the definition of the normal cone is irrelevant for convex set. That is, every vector $v \in N_Q(\bar{x})$ truly makes an obtuse angle with any direction $x - \bar{x}$ for $x \in Q$.

**Exercise 3.22.** Prove that the following are equivalent for any convex set $Q$ and a point $\bar{x} \in Q$:

1. $v$ lies in $N_Q(\bar{x})$,
2. $v$ lies in $(T_Q(\bar{x}))^\circ$,
3. $\bar{x}$ lies in $\arg\max_{x \in Q} \langle v, x \rangle$.
4. equality $\text{proj}_Q(\bar{x} + \lambda v) = \bar{x}$ holds for all $\lambda \geq 0$,
5. equality $\text{proj}_Q(\bar{x} + \lambda v) = \bar{x}$ holds for some $\lambda > 0$.

**Exercise 3.23.** Show for any convex cone $K$ and a point $x \in K$, the equality

\[
N_K(x) = K^\circ \cap x^\perp.
\]

**Exercise 3.24.** Show for any convex set $Q$ and a point $x \in Q$, the equivalence

\[
x \in \text{int } Q \iff N_Q(x) = \{0\}.
\]
3.2 Convex functions: basic operations and continuity

We next move on to convex analysis – the study of convex functions. We will consider functions $f$ mapping $E$ to the extended-real-line $\mathbb{R} := \mathbb{R} \cup \{\pm \infty\}$. To be completely precise, some care must be taken when working with $\pm \infty$. In particular, we set $0 \cdot \pm \infty = 0$ and avoid expressions $(+\infty) + (-\infty)$. A function $f : E \to \mathbb{R}$ is called proper if it never takes the value $-\infty$ and is finite at some point.

Given a function $f : E \to \mathbb{R}$, the domain of $f$ and the epigraph of $f$ are
\[
dom f := \{x \in E : f(x) < +\infty\},
\]
\[
epi f := \{(x, r) \in E \times \mathbb{R} : f(x) \leq r\},
\]
respectively. Thus $\dom f$ consists of all points $x$ at which $f$ is finite or evaluates to $-\infty$. The epigraph $\epi f$ is simply the set above the graph of the function. Much of convex analysis proceeds by studying convex geometric properties of epigraphs.

Recall that a function $f : E \to \mathbb{R}$ is convex if $\epi f$ is a convex set in $E \times \mathbb{R}$. Equivalently, a proper function $f$ is convex if and only if the inequality
\[
f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)
\]
holds for all $x, y \in E$ and $\lambda \in (0, 1)$.

**Exercise 3.25** (Jensen’s Inequality). Show that a proper function $f : E \to \mathbb{R}$ is convex if $\epi f$ is a convex set in $E \times \mathbb{R}$. Equivalently, a proper function $f$ is convex if and only if the inequality
\[
f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)
\]
holds for all $x, y \in E$ and $\lambda \in (0, 1)$.

**Exercise 3.26.** Let $f : E \to \mathbb{R}$ be a convex function. Show that if there exists a point $x \in \text{ri}(\dom f)$ with $f(x)$ finite, then $f$ must be proper.

We will call a function $f : E \to \mathbb{R}$ closed or lower-semi-continuous if
\[
epi f
\]
is itself a closed set.

**Exercise 3.27.** Show that $f : E \to \mathbb{R}$ is closed if and only if the inequality
\[
\liminf_{y \to x} f(y) \geq f(x)
\]
holds for any $x \in E$.

Consider a function $f : E \to \mathbb{R}$. It is easy to check that the set $\text{cl}(\epi f)$ is itself an epigraph of some closed function. We call this function the closed envelope of $f$ and denote it by $\text{cl} f$. Similarly, $\text{conv}(\epi f)$ is itself an epigraph of some convex function. We call this function the convex envelope of $f$ and denote it by $\text{co} f$. Combining the two operations, yields the closed convex envelope of $f$, which we denote by $\text{co}(f) = \text{cl}(\text{co}(f))$. Though this description is geometrically pleasing, it is not convenient for computation. A better description arises from considering minorants. Given two functions $f$ and $g$ on $E$, we say that $g$ is a minorant of $f$ if it satisfies $g(y) \leq f(y)$ for all $y \in E$. 
Exercise 3.28. Given a proper function \( f : E \to \mathbb{R} \), show the equalities
\[
(co f)(x) = \sup \{g(x) : g : E \to \mathbb{R} \text{ is a closed convex minorant of } f\}
= \sup \{g(x) : g : E \to \mathbb{R} \text{ is an affine minorant of } f\}.
\]

Just like it is often easier to work with convex cones than with convex sets, it is often easier to work with function whose epigraphs are convex cones. We say that \( f : E \to \mathbb{R} \) is sublinear if \( \text{epi } f \) is a convex cone.

Exercise 3.29. Show that a proper function \( f : E \to \mathbb{R} \) is sublinear if and only if \( f(\lambda x + \mu y) \leq \lambda f(x) + \mu f(y) \) for all \( x, y \in E \) and \( \lambda, \mu \geq 0 \).

There are a number of convex functions that naturally arise from convex sets. Given a set \( Q \subseteq E \), define its indicator function
\[
\delta_Q(x) = \begin{cases} 
0, & x \in Q \\
\infty, & x \notin Q
\end{cases}
\]
its support function
\[
\delta^*_Q(v) = \sup_{x \in Q} \langle v, x \rangle,
\]
and its gauge function
\[
\gamma_Q(x) = \inf \{\lambda \geq 0 : x \in \lambda Q\}.
\]
Notice that support functions and gauges are sublinear. Conversely, the following exercise shows that closed sublinear functions are support functions.

Exercise 3.30. Show that if \( g : E \to \mathbb{R} \) is a proper sublinear function, then \( \text{cl } g \) is the support function of the set
\[
Q = \{x : \langle x, y \rangle \leq g(y) \ \forall y \in E\}.
\]

The notation \( \delta^*_Q(v) \) may seem strange at first, since it is not clear what the support function \( \delta^*_Q(v) \) has to do with the indication function \( \delta_Q(x) \). The notation will make sense shortly, in light of Fenchel conjugacy (Section 3.3).

Exercise 3.31. Show that if \( Q \) is convex, then \( \delta_Q, \delta^*_Q, \text{dist}_Q \) and \( \gamma_Q \) are all convex.

Exercise 3.32. Show that for any closed, convex set \( Q \) containing the origin, we have \( \gamma_Q(x) = \delta^*_Q(x) \).

Convexity is preserved under a variety of operations, most notably the following.

1. (Monotone convex composition) If \( f : E \to \mathbb{R} \) is convex, and \( \varphi : \mathbb{R} \to \mathbb{R} \) is convex and nondecreasing, then the composition \( \varphi \circ f \) is convex.
3.2. CONVEX FUNCTIONS: BASIC OPERATIONS AND CONTINUITY

2. **(Finite sums)** If \( f_1, f_2 : E \to \mathbb{R} \) are proper and convex, then the sum \( f_1 + f_2 \) is convex.

3. **(Affine composition)** More generally, if \( A : E \to Y \) is a linear map and \( f : E \to \mathbb{R} \) is a proper convex function, then the composition \( g(x) := f(Ax) \) is convex.

4. **(Pointwise max)** If \( f_i(x) \) is convex, for each \( i \) in an arbitrary index \( I \), then \( f(x) := \max_{i \in I} f_i(x) \) is also convex. Indeed, the reader should verify the relationship \( \text{epi } f = \bigcap_{i \in I} \text{epi } f_i \).

5. **(Lower envelope)** Consider a convex set \( Q \subset E \times \mathbb{R} \) and define the lower envelope
   \[
   f(x) := \inf\{r : (x, r) \in Q\}.
   \]
   To see that \( f \) is convex, it suffices to observe \( \text{epi } f = Q + (\{0\} \times \mathbb{R}_+) \).

6. **(Infimal Convolution)** The infimal convolution of two functions \( f, g : E \to \mathbb{R} \) is the function
   \[
   (f \Box g)(x) = \inf_y \{f(x - y) + g(y)\}. \tag{3.3}
   \]
   Equivalently, we may write
   \[
   (f \Box g)(x) = \inf\{r : (x, r) \in \text{epi } f + \text{epi } g\}.
   \]
   Hence infimal convolution is an example of a lower envelope with \( Q := \text{epi } f + \text{epi } g \). We deduce that if \( f \) and \( g \) are convex, then so is the convolution \( f \Box g \).

7. **(Infimal Projection)** Consider a convex function \( g : E \times Y \to \mathbb{R} \). The function
   \[
   f(x) := \inf_y g(x, y)
   \]
   is called the infimal projection of \( g \). To see that this function is convex, write
   \[
   f(x) = \inf_{y, r} \{r : g(x, y) \leq r\}
   = \inf\{r : \exists y \text{ with } (x, y, r) \in \text{epi } g\}
   = \inf\{r : (x, r) \in \pi_{1,3}(\text{epi } g)\}. \tag{3.4}
   \]
   Here \( \pi_{1,3} \) is the canonical projection \( \pi_{1,3}(x, y, r) = (x, r) \). Thus \( f \) is the lower envelope generated by the convex set \( Q := \pi_{1,3}(\text{epi } g) \). More concretely, we may write
   \[
   \text{epi } f = \pi_{1,3}(\text{epi } g).
   \]
We end this section with a remarkable property: convex functions are always locally Lipschitz continuous on the relative interior of their domains.

**Theorem 3.33.** Let \( f \) be a proper convex function and \( Q \) a compact subset of \( \text{ri} (\text{dom} f) \). Then \( f \) is Lipschitz continuous on \( Q \).

**Proof.** Without loss of generality, by restricting to the affine hull, \( \text{aff}(\text{dom} f) \), we can assume that \( \text{dom} f \) has nonempty interior. Choose \( \epsilon > 0 \) satisfying \( \text{cl}(Q + \epsilon B) \subset \text{int}(\text{dom} f) \), where \( B \) is the unit ball.

Let us first establish a seemingly mild conclusion that \( f \) is bounded on \( Q + \epsilon B \). For the sake of contradiction, suppose there is a sequence \( x_i \in Q + \epsilon B \) with \( |f(x_i)| \to \infty \). Appealing to compactness, we can restrict to a subsequence and assume \( x_i \) converges to some point \( \bar{x} \in \text{int}(\text{dom} f) \). The points \((x_i, f(x_i))\) all lie in the boundary of \( \text{epi} f \).

Now there are two cases: \( f(x_i) \to -\infty \) and \( f(x_i) \to +\infty \). Let’s suppose first \( f(x_i) \to -\infty \). Fix a nonzero vector \((\bar{v}, \bar{\alpha}) \in N_{\text{epi}}(\bar{x}, f(\bar{x}))\), guaranteed to exist by Exercise 3.24. By the nature of epigraphs, the inequality \( \bar{\alpha} \leq 0 \) holds, and hence we deduce

\[
0 \geq \langle (\bar{v}, \bar{\alpha}), (x_i, f(x_i)) - (\bar{x}, f(\bar{x})) \rangle = \langle \bar{v}, x_i - \bar{x} \rangle + \bar{\alpha}(f(x_i) - f(\bar{x})).
\]

Letting \( i \to \infty \) we deduce \( \bar{\alpha} = 0 \). The very definition of the normal cone then implies \( \bar{v} \in N_{\text{dom} f}(\bar{x}) \). By Exercise 3.24 this is a contradiction since \( \bar{x} \) lies in the interior of \( \text{dom} f \).

Suppose now we are in the second case, \( f(x_i) \to +\infty \). Choose nonzero vectors \((v_i, \alpha_i) \in N_{\text{epi}}(x_i, f(x_i))\). Then by definition of the normal, we have

\[
0 \geq \langle (v_i, \alpha_i), (x, f(x)) - (x_i, f(x_i)) \rangle = \langle v_i, x - x_i \rangle + \alpha_i(f(x) - f(x_i))
\]

for all \( x \in \text{dom} f \). Note if \( v_i \) is zero, then \( x_i \) is a global minimizer of \( f \), which is impossible for all large \( i \), since \( f(x_i) \to +\infty \). Therefore, restricting to a subsequence, we may assume \( v_i \neq 0 \) for all \( i \). Moreover, rescaling \((v_i, \alpha_i)\) we may assume \( \|v_i\| = 1 \) and that \( v_i \) converge to some nonzero vector \( \bar{v} \). Letting \( i \) tend to infinity in the inequality above yields \( \alpha_i \to 0 \) and the inequality becomes

\[
\langle \bar{v}, x - \bar{x} \rangle \leq \limsup_{i \to \infty} \alpha_i f(x_i).
\]

Setting \( x = \bar{x} \), we deduce \( \limsup_{i \to \infty} \alpha_i f(x_i) = 0 \). Hence \( \bar{v} \in N_{\text{dom} f}(\bar{x}) \), but this is impossible since \( \bar{x} \) is in the interior of \( \text{dom} f \), yielding a contradiction. Thus, we have proved that \( f \) is bounded on \( Q + \epsilon B \).

Let \( \alpha_1 \) and \( \alpha_2 \) be the lower and upper bounds on \( f \) in \( Q + \epsilon B \). Fix arbitrary points \( x, y \in Q \) and define \( z := y + \frac{\epsilon}{\|y-x\|}(y-x) \). By definition, \( z \) lies in \( Q + \epsilon B \) and we have \( y = (1-\lambda)x + \lambda z \) for \( \lambda := \frac{\|y-x\|}{\epsilon + \|y-x\|} \). Since \( f \) is convex, we deduce

\[
f(y) \leq (1-\lambda)f(x) + \lambda f(z) = f(x) + \lambda(f(z) - f(x)).
\]
3.3. THE FENCHEL CONJUGATE

and therefore

\[ f(y) - f(x) \leq \lambda(\alpha_2 - \alpha_1) \leq \frac{(\alpha_2 - \alpha_1)}{\epsilon} \|y - x\|. \]

Since \( x \) and \( y \) are arbitrary points in \( Q \), we have shown that \( f \) is Lipschitz continuous on \( Q \), as claimed.

In contrast, convex functions can behave very poorly on the relative boundary of their domains. For example the function \( f : \mathbb{R}^2 \to \mathbb{R} \) given by

\[
f(x, y) = \begin{cases} \frac{y^2}{x} & \text{if } x > 0 \\ 0 & \text{if } (x, y) = (0, 0) \\ +\infty & \text{otherwise} \end{cases}
\]

is closed and convex, but is not continuous at the origin relative to its domain. See the graph below.

Figure 3.1: Plot of the function \( f(x, y) \) in equation (3.5).

3.3 The Fenchel conjugate

In convex geometry, one could associate with any convex cone its polar. Convex analysis takes this idea much further through a new operation on functions, called Fenchel conjugacy.

**Definition 3.34.** For a function \( f : E \to \mathbb{R} \), define the **Fenchel conjugate** function \( f^* : E \to \mathbb{R} \) by

\[
f^*(y) = \sup_{x \in E} \{ \langle y, x \rangle - f(x) \}.
\]

This operation arises naturally from epigraphical geometry. Indeed, from the very definition of the Fenchel conjugate, observe that the epigraph, \( \text{epi } f^* \), consists of all pairs \((y, r)\) satisfying \( f(x) \geq \langle y, x \rangle - r \) for all points \( x \). Thus \( \text{epi } f^* \) encodes all affine minorants \( x \mapsto \langle y, x \rangle - r \) of \( f \). An alternate
insightful interpretation is through the support function to the epigraph. Observe

\[
f^\star(y) = \sup_{x \in \mathbf{E}} \{(y, -1), (x, f(x))\}
\]

\[
= \sup_{(x, r) \in \text{epi } f} \{(y, -1), (x, r)\}
\]

\[
= \varphi_{\text{epi } f}(y, -1).
\]

Thus the conjugate \(f^\star(y)\) is exactly the support function of \(\text{epi } f\) evaluated at \((y, -1)\). Since the support function is sublinear, the appearance of \(-1\) in the last coordinate simply serves as a normalization constant.

Let us look at some examples. First, it is clear that the Fenchel conjugate of the indicator function \(\delta_Q\) is exactly the support function of \(Q\), thereby explaining the notation \(\delta_Q^\star\) for the latter. For the function \(f(x) = \frac{1}{2}\|x\|^2\), we have \(f^\star(y) = \frac{1}{2}\|y\|^2\). Thus \(\frac{1}{2}\|\cdot\|^2\) is self-conjugate. For the exponential function \(f(x) = e^x\), the reader should verify the formula

\[
f^\star(y) = \begin{cases} 
y \log(y) - y, & \text{if } y > 0 \\
0, & \text{if } y = 0 \\
\infty, & \text{if } y < 0
\end{cases}
\]

If \(f\) is the quadratic \(f(x) = \frac{1}{2}\langle Ax, x \rangle\) with \(A \succ 0\), then \(f^\star(y) = \frac{1}{2}\|y\|^2\). Let us next see what happens when the conjugacy operation is applied twice \(f^{**} := (f^\star)^\star\). Let us look first at the simplest example of an affine function.

**Exercise 3.35.** Show that for any affine function \(f(x) = \langle a, x \rangle + b\), we have \(f^\star(y) = -b + \delta_{\{a\}}(y)\). Deduce the equality \(f^{**} = f\).

We will also use the following elementary observation.

**Exercise 3.36.** For any function \(g : \mathbf{E} \times \mathbf{Y} \to \mathbb{R}\), we have

\[
\sup_y \inf_x g(x, y) \leq \inf_x \sup_y g(x, y).
\]

We can now prove the main theorem of this section.

**Theorem 3.37** (Biconjugacy). For any proper convex function \(f : \mathbf{E} \to \overline{\mathbb{R}}\), equality \(f^{**} = \overline{\partial} f\) holds.
3.3. The Fenchel Conjugate

Proof. We begin by successively deducing:

\[(f^*)^*(x) = \sup_y \{ (x, y) - f^*(y) \}\]
\[= \sup_y \{ (x, y) - \sup_z \{ (z, y) - f(z) \} \}\]
\[= \sup_y \inf_z \{ (y, x - z) + f(z) \} \]
\[\leq \inf_z \{ +\infty \quad x \neq z \}
\quad \text{and} 
\quad 0 \leq f(z) \quad x = z \]
\[= f(x).\]

The inequality in the fourth line is immediate from Exercise 3.36.

Thus we have established \(f^{**} \leq f\). Notice that \(f^{**}\) is by definition closed and convex. Hence we deduce from Exercise 3.28 the inequality \(f^{**} \subseteq \overline{co}f\). To complete the proof, let \(g(x) = \langle a, x \rangle + b\) be any lower affine minorant of \(f\). By the definition of the conjugate, we see that conjugacy is order reversing and hence \(g^* \geq f^*\). Taking into account Exercise 3.35 then yields \(g = (g^*)^* \leq (f^*)^* \leq f\). Taking the supremum over all affine minorants \(g\) of \(f\) yields \(\overline{co}f \leq f^{**}\), thereby completing the proof.

The biconjugacy theorem incorporates many duality ideas we have already seen in convex geometry. For example, let \(K\) be a nonempty cone. It is immediate from the definition of conjugacy that \(\delta^*_K = \delta_{K^{**}}\). Consequently, Theorem 3.37 shows \(\delta_{\text{cl conv}K} = \overline{co}(\delta_K) = (\delta_K)^{**} = \delta_{K^{**}} = \delta_{K^{***}}\).

Hence we deduce \(K^{***} = \text{cl conv}K\). This is exactly the conclusion of Exercise 3.16.

Exercise 3.38. Show the following.

1. If \(f, g : E \to \overline{R}\) are closed proper convex functions, then equalities hold:
\[(f \Box g)^* = f^* + g^* \quad \text{and} \quad (f + g)^* = \text{cl}(f^* \Box g^*).\]

2. Let \(f : Y \to \overline{R}\) be a proper closed convex function and \(A : E \to Y\) a linear map. Define the composition \(g(x) = f(Ax)\). Then assuming \(\text{dom} g \neq \emptyset\), the conjugate \(g^*\) is the closed envelope of the function \(y \mapsto \inf_x \{ f^*(x) : A^*x = y \}\).

3. Fix a function \(f : E \times Y \to \overline{R}\) and define the function \(g(y) := \inf_x f(x, y)\).
Prove the equality \(g^*(w) = f^*(0, w)\).

4. Given any function \(f : E \to \overline{R}\) and a point \(b \in E\), the conjugate of the function \(g(x) = f(x + b)\) is given by \(g^*(y) = f^*(y) - \langle b, y \rangle\).
3.4 Differential properties

We next turn to differential properties of convex functions.

**Definition 3.39 (Subgradients and the Subdifferential).** Consider a convex function $f: E \to \mathbb{R}$ and a point $x \in E$, with $f(x)$ finite. Then $v \in E$ is called a subgradient of $f$ at $x$ if the inequality

$$f(y) \geq f(x) + \langle v, y - x \rangle$$

holds for all $y \in E$.

The set of all subgradients $v$ of $f$ at $x$ is called the subdifferential and is denoted by $\partial f(x)$.

In words, for fixed $x$, a subgradient $v \in \partial f(x)$ has the property that the linear functional $y \mapsto f(x) + \langle v, y - x \rangle$ globally minorizes $f$. The connection of subdifferentials to epigraphical geometry becomes clear by noting

$$v \in \partial f(x) \iff (v, -1) \in N_{\text{epi} f}(x, f(x)).$$

The subdifferential $\partial f(x)$ is always a closed convex set. Given a convex set $Q \subset E$, observe the equality $\partial \delta Q(x) = N_Q(x)$. Hence the normal cone is an example of a subdifferential.

**Exercise 3.40 (Existence of subgradients).** Consider a proper convex function $f: E \to \mathbb{R}$. Use Theorem 3.33 to show that for any point $x \in \text{ri dom } f$, the subdifferential $\partial f(x)$ is nonempty.

Just like for smooth convex functions, the gradient characterizes global minima, so does the subdifferential for nonsmooth convex functions.

**Proposition 3.41.** Consider a convex function $f: E \to \mathbb{R}$ and a point $x$ with $f(x)$ finite. Then the following are equivalent:

1. $x$ is a global minimizer of $f$
2. $x$ is a local minimizer of $f$
3. $0 \in \partial f(x)$

**Proof.** The implication $\exists \Rightarrow 1 \Rightarrow 2$ is immediate. We argue next the remaining implication $2 \Rightarrow 3$. Suppose $x$ is a local minimizer and fix an arbitrary point $y$. It is easy to see that $f$ must be proper. For any $\lambda \in [0, 1]$ convexity implies

$$f(\lambda y + (1 - \lambda)x) \leq \lambda f(y) + (1 - \lambda)f(x).$$

For $\lambda$ sufficiently small, the left-hand-side is lower bounded by $f(x)$. Rearranging, we deduce $f(y) \geq f(x)$ and the result follows.

The following is a very useful property relating conjugates and subdifferentials.
3.4. DIFFERENTIAL PROPERTIES

Theorem 3.42 (Fenchel-Young Inequality). Consider a convex function $f : E \to \mathbb{R}$. Then for any points $x, y \in E$, the inequality

$$f(x) + f^*(y) \geq \langle x, y \rangle$$

holds, while equality holds if and only if $y \in \partial f(x)$.

Proof. Observe

$$f^*(y) = \sup_z \{\langle z, y \rangle - f(z)\} \geq \langle x, y \rangle - f(x),$$

establishing the claimed inequality. Next observe the inclusion $y \in \partial f(x)$ holds if and only if $f(z) \geq f(x) + \langle y, z - x \rangle$ for all $z$, or equivalently $\langle y, x \rangle - f(x) \geq \langle y, z \rangle - f(z)$ for all $z$. Taking supremum over $z$, this amounts to

$$\langle y, x \rangle - f(x) \geq \sup_z \{\langle y, z \rangle - f(z)\} \equiv f^*(y).$$

This is the reverse direction in the inequality. 

A crucial consequence of the Fenchel-Young inequality is that the conjugacy operation acts as an inverse on the level of subdifferentials.

Corollary 3.43. Suppose $f : E \to \mathbb{R}$ is proper, closed, and convex. Then

$$y \in \partial f(x) \iff x \in \partial f^*(y).$$

Proof. From Theorem 3.42 we deduce $y \in \partial f(x)$ if and only if

$$\langle x, y \rangle = f(x) + f^*(y).$$

On the other hand by Theorem 3.37 we have $f(x) + f^*(y) = (f^*)^*(x) + f^*(y)$. Applying Theorem 3.42 again we deduce $y \in \partial f(x)$ if and only if $x \in \partial f^*(y)$. 

It is natural to ask how differentiability of a convex function relates to the subdifferential. The answer turns to be simple: a proper, closed, convex function is differentiable at a point if and only if the subdifferential there is a singleton. To prove this result, we will need the following observation.

Exercise 3.44. For the any proper, closed, convex function $f : \mathbb{R} \to \mathbb{R}$, the subdifferential graph

$$\text{gph} \partial f := \{(x, v) \in E \times E : v \in \partial f(x)\}.$$

is a closed set.

Theorem 3.45. A proper, closed, convex function $f : E \to \mathbb{R}$ is differentiable at $x$ if and only if $\partial f(x)$ is a singleton, in which case it must be that $\partial f(x) = \{\nabla f(x)\}$. 

Proof. Suppose that $f$ is differentiable at $x$. Then by Exercise [8.40] the subdifferential $\partial f(x)$ is nonempty. Fix any subgradient $v \in \partial f(x)$ and a vector $h \in \mathcal{E}$. Then for all small $t > 0$, the subgradient inequality implies

$$f(x + th) \geq f(x) + t\langle v, h \rangle$$

while differentiability yields

$$f(x + th) = f(x) + t\langle \nabla f(x), h \rangle + o(t).$$

Combining the two estimates we deduce

$$\langle \nabla f(x) - v, h \rangle \geq o(t) / t.$$

Letting $t$ tend to zero, we learn $\langle \nabla f(x) - v, h \rangle \geq 0$ for all $h$. In particular, setting $h = v - \nabla f(x)$, we deduce $\nabla f(x) = v$, as we had to show.

Conversely, suppose that $\partial f(x)$ consists of a single vector $v \in \mathcal{E}$. We first claim that $x$ must lie in $\text{int dom } f$. Indeed, if this were not the case, then we would deduce $v + N_{\text{dom } f}(x) \subset \partial f(x)$, which is a contradiction.

Define the function $g(y) = f(x + y) - f(x) - \langle v, y \rangle$. We aim to show that $g$ is differentiable at $x$ with $\nabla g(x) = 0$, which by standard calculus will directly imply $v = \nabla f(x)$ as claimed. Observe equality $0 = \nabla g(x)$ amounts to the condition $\lim_{z \to 0} g(z) / \|z\| = 0$, which we will now verify. To this end, consider any sequence $z_k \to 0$. Choose now any vectors $v_k \in \partial g(z_k)$, which exist by Exercise [8.40]. The subgradient inequality then implies

$$g(z_k) = g(z_k) - g(0) \leq \langle v_k, z_k \rangle \leq \|v_k\| \cdot \|z_k\|$$

for all indices $k$. Notice that the sequence $\{v_k\}$ is bounded by [8.33]. Let $\bar{v}$ be any limit point of $v_k$. Exercise [8.44] implies that $\bar{v} \in \partial g(0)$. Since $\partial g(0)$ is a singleton, we deduce $\bar{v} = 0$. Thus we deduce

$$\lim_{k \to \infty} g(z_k) / \|z_k\| \leq \lim_{k \to \infty} \|v_k\| = 0,$$

as we had to show.

3.5 Directional derivative

When $f$ is a smooth function, then the directional derivative of $f$ at $x$ in direction $y$, is simply the inner product $\langle \nabla f(x), y \rangle$. We next investigate the relationship between the directional derivative and the subdifferential for nonsmooth convex functions.
3.5. DIRECTIONAL DERIVATIVE

Definition 3.46 (Directional derivative). Let $f : E \to \mathbb{R}$ be a convex function and fix a point $x$ with $f(x)$ finite. The directional derivative of $f$ at $x$ in direction $y$ is defined by

$$f'(x, y) := \lim_{t \downarrow 0} \frac{f(x + ty) - f(x)}{t},$$

provided the limit exists.

The limit in the definition of the directional derivative always exists.

Theorem 3.47. Suppose $f : E \to \mathbb{R}$ is convex and fix a point $x$ with $f(x)$ finite. Then for any point $y$, the quotients $\frac{f(x + ty) - f(x)}{t}$ are nondecreasing in $t$, and therefore $f'(x, y)$ exists.

Proof. Fix any reals $\hat{\lambda}, \lambda$ satisfying $0 < \hat{\lambda} < \lambda$. Observe

$$\frac{f(x + \hat{\lambda}y) - f(x)}{\lambda} = \frac{f \left( \left( \frac{\lambda - \hat{\lambda}}{\lambda} \right) x + \hat{\lambda} \left( x + \lambda y \right) \right) - f(x)}{\lambda} \leq \frac{\frac{\lambda - \hat{\lambda}}{\lambda} f(x) + \hat{\lambda} f(x + \lambda y) - f(x)}{\hat{\lambda}} = \frac{f(x + \lambda y) - f(x)}{\lambda}.$$ 

The result follows.

The function $f'(x, \cdot) : \mathbb{R}^n \to \mathbb{R}$ with $y \mapsto f'(x, y)$ is convex and positively homogeneous (hence sublinear), but $f'(x, \cdot)$ may fail to be closed. Think for example of the direction derivative of the indicator $\delta_{B(0, 1)}$ at $x = (0, 1)$. The following theorem shows that the directional derivative $f'(x, \cdot)$ (up to closure) is precisely the support function of the subdifferential $\partial f(x)$.

Theorem 3.48. Consider a proper convex function $f : E \to \mathbb{R}$ and fix a point $\bar{x} \in \text{dom } f$. Then $\text{cl } f'(x, \cdot)$ is precisely the support function of the subdifferential $\partial f(x)$.

Proof. Observe for for all $v \in \partial f(x)$ and $y \in \mathbb{R}^n$ the inequality

$$f(x + ty) \geq f(x) + t\langle v, y \rangle.$$ 

It follows immediately that

$$f'(x, y) \geq \langle v, y \rangle \quad \text{for all } v \in \partial f(x) \text{ and } y \in E.$$ 

Conversely suppose $v \notin \partial f(x)$. Hence, there exists a vector $z$ satisfying $f(x + z) < f(x) + \langle v, z \rangle$. Theorem 3.47 then implies

$$f'(x, z) = \lim_{t \downarrow 0} \frac{f(x + tz) - f(x)}{t} \leq \frac{f(x + z) - f(x)}{1} < \langle v, z \rangle.$$
We thus deduce the representation

$$\partial f(x) = \{v : \langle v, y \rangle \leq f'(x, y) \text{ for all } y\}.$$ 

Appealing to Exercise 3.30, the result follows. \hfill \Box

**Exercise 3.49.** Let \( f : E \to \mathbb{R} \) be a proper convex function and let \( Q \) be any open convex subset of \( \text{dom } f \). Prove the identity

$$\sup_{x, y \in Q} \frac{|f(x) - f(y)|}{\|x - y\|} = \sup_{x \in Q, v \in \partial f(x)} \|v\|.$$ 

### 3.6 The value function

When solving an optimization problem, one is often interested not only in the optimal value, but also in how this value changes under small perturbation to the problem data. More formally, consider an arbitrary convex function \( F : E \times Y \to \mathbb{R} \) and the parametric optimization problem:

$$p(y) := \inf_x F(x, y)$$

The reader should think of \( y \) as a perturbation parameter and the problem corresponding to \( p(0) \) as the original “primal” problem. The assignment \( y \mapsto p(y) \) is called the *value function*. The reader should recognize \( p \) as the infimal projection of \( F \).

A worthwhile goal is to study how \( p(y) \) varies as \( y \) is perturbed around the origin. Thus with the machinery we have developed, we aim to compute the subdifferential \( \partial p(0) \). Remarkably, we will see that \( \partial p(0) \) coincides with the set of maximizers of an auxiliary convex optimization problem. To make this idea precise, define the new parametric family of problems

$$q(x) := \sup_y -F^*(x, y).$$

Let us call the problem corresponding to \( q(0) \) the *parametric dual*. The terminology is easy to explain. Indeed, invoking Exercise 3.38 (part 3) yields the equality \( p^*(y) = F^*(0, y) \) and therefore

$$p^{**}(0) = \sup_y \{ (0, y) - p^*(y) \} = \sup_y -F^*(0, y) = q(0).$$

Thus \( q(0) \) is the biconjugate of \( p \) evaluated at zero. Therefore we expect that under mild conditions, equality \( p(0) = q(0) \) should hold. This is indeed the case, and moreover, the subdifferential \( \partial p(0) \) consists precisely of the optimal solutions \( y \) to the parametric dual problem.

**Theorem 3.50** (Parametric optimization). *Suppose that \( F : E \times Y \to \mathbb{R} \) is proper, closed, and convex. Then the following are true.*
(a) The inequality \( p(0) \geq q(0) \) always holds.

(b) If \( p(0) \) is finite, then
\[
\partial p(0) = \operatorname{argmax}_y \, - F^*(0, y).
\]

(c) If the subdifferential \( \partial p(0) \) is nonempty, then equality \( p(0) = q(0) \) holds and the supremum \( q(0) \) is attained, if finite.

Proof. Part (a) is immediate from the inequality \( p(0) \geq p^{**}(0) \). Next suppose \( p(0) \) is finite. Observe that a vector \( \phi \) satisfies \( \phi \in \partial p(0) \) if and only if for all \( y \) it holds:
\[
p(0) \leq \inf_x \left\{ F(x, y) - \langle \begin{pmatrix} 0 \\ x \end{pmatrix}, \begin{pmatrix} 0 \\ y \end{pmatrix} \rangle \right\}.
\]

Taking the infimum over \( y \), we deduce \( \phi \in \partial p(0) \) if and only if \( p(0) \leq -F^*(0, \phi) \), which in light of (a) happens if and only if \( \phi \) is dual optimal. Note moreover that the existence of a single subgradient \( \phi \in \partial p(0) \) implies \( p(0) = q(0) \). The proof is complete.

3.7 Duality and subdifferential calculus

The idea of duality has appeared throughout the previous sections, culminating in the definition of the Fenchel conjugate. In this section, we will use these ideas to investigate the so-called primal-dual pairs of convex optimization problems. Roughly speaking, we will see that for a number of well-structured convex minimization problems, there are natural ways to obtain lower bounds on their optimal value. The task of finding the largest such lower bound is itself another optimization, called the dual. A central question is therefore to determine conditions ensuring that the best lower-bound matches the primal optimal value. As a bonus, we would also like to know that optimality in the primal can indeed be certified by a dual feasible solution, or in other words that the dual optimal value is attained.

3.7.1 Fenchel-Rockafellar duality

Consider a general class of structured optimization problems
\[(P) \quad \inf_{x \in E} h(Ax) + g(x),\]
where \( h: Y \to \overline{\mathbb{R}} \), \( g: E \to \overline{\mathbb{R}} \) are some proper functions and \( A: E \to Y \) is a linear map. Let us call this problem the primal. Define now a new convex optimization problem called the dual:
\[(D) \quad \sup_{y \in Y} - h^*(y) - g^*(-A^*y),\]
CHAPTER 3. CONVEX GEOMETRY AND ANALYSIS

The dual problem (D) arises naturally from a lower-bounding viewpoint. Let us try to find simple lower bounds for \( \text{val}(P) \), the optimal value of (P). From the definition of the conjugate, any \( \bar{y} \in \text{dom} \ h^* \) yields the lower bound

\[
\text{val}(P) \geq \inf_{x \in E} \sup_{y \in Y} \langle y, Ax \rangle - h^*(y) + g(x)
\]

\[
= \inf_{x \in E} \langle \bar{y}, Ax \rangle - h^*(\bar{y}) + g(x)
\]

\[
= -h^*(\bar{y}) - \sup_{x \in E} \{(\langle -A^*\bar{y}, x \rangle - g(x)\}
\]

\[
= -h^*(\bar{y}) - g^*(-A^*\bar{y}).
\]

The right-hand-side is exactly the evaluation of the dual objective function at \( \bar{y} \). Thus \( \text{val}(D) \) is the supremum over all lower-bounds on \( \text{val}(P) \) that can be obtained in this way. In particular, we have deduced the weak-duality inequality

\[
\text{val}(P) \geq \text{val}(D).
\]

The goal of this section is to show that when \( h \) and \( g \) are proper, closed, and convex and a mild compatibility holds, we can be sure that strong duality holds: \( \text{val}(P) = \text{val}(D) \) and the dual optimal value is attained, if finite. The argument proceeds by interpreting Fenchel-Rockafellar duality within the parametric framework of Theorem 3.50.

Theorem 3.51 (Fenchel-Rockafellar duality). Consider the problems:

\[(P) \quad \min_x h(Ax) + g(x)\]

\[(D) \quad \max_y -g^*(-A^*y) - h^*(y).\]

where \( g: E \to \mathbb{R} \) and \( h: Y \to \mathbb{R} \) are proper, closed convex functions, and \( A: E \to Y \) is a linear map. Then the two problems fit the perturbation framework of Theorem 3.50 with

\[
F(x, y) = h(Ax + y) + g(x).
\]

In particular, the regularity condition

\[
0 \in \text{ri} (\text{dom} \ h) - A(\text{ri} \text{ dom} \ g) \tag{3.6}
\]

guarantees that the primal and dual optimal values are equal, and the dual optimal value is attained, if finite.

Proof. Let us verify the assumptions of Theorem 3.50 To this end, define the function

\[
F(x, y) = h(Ax + y) + g(x).
\]
3.7. DUALITY AND SUBDIFFERENTIAL CALCULUS

Observe

\[ F^\star(x, y) = \sup_{z, w} \{ \langle (z, w), (x, y) \rangle - h(Az + w) - g(z) \}. \]

Making the substitution \( v := Az + w \), we get

\[ F^\star(x, y) = \sup_{z, v} \{ \langle z, x \rangle + \langle v - Az, y \rangle - h(v) - g(z) \} \]
\[ = \sup_{z} \{ \langle z, x - A^\star y \rangle - g(z) \} + \sup_{v} \{ \langle v, y \rangle - h(v) \} \]
\[ = g^\star(x - A^\star y) + h^\star(y). \]

Thus the Fenchel dual problem \((D)\) is exactly the problem \( q(0) = -\sup_{y} F^\star(0, y) \).

A quick computation shows

\[ \text{dom } p = \text{dom } h - A(\text{dom } g). \]

Hence the assumption [3.6] along with Theorem 3.8 implies \( 0 \in \text{ri (dom } p) \). If \( p(0) = -\infty \), there is nothing to prove. Hence we may suppose that \( p(0) \) is finite. Theorem 3.26 then implies that \( p \) is proper, while Theorem 3.40 in turn guarantees that the subdifferential \( \partial p(0) \) is nonempty. An application of Theorem 3.51 completes the proof.

In many applications, the primal problem looks slightly different:

\[(P) \quad \min_{x} \langle c, x \rangle + h(b - Ax) + g(x), \]

for some vectors \( b \in Y \) and \( c \in E \). Some thought shows that this formulation can be put into the standard form of Theorem 3.51 simply by replacing \( g \) with \( g + \langle c, \cdot \rangle \) and replacing \( h \) with \( h(b - \cdot) \). Then the dual reads:

\[(D) \quad \max_{y} \langle b, y \rangle - g^\star(A^\star y - c) - h^\star(y), \]

and the regularity condition (3.6) becomes

\[ b \in A(\text{ri dom } g) + \text{ri (dom } h). \]

Among the most important consequences of Theorem 3.51 are a sum and a chain rule for subdifferentials.

**Theorem 3.52** (Subdifferential calculus). *Let \( g: E \to \mathbb{R} \) and \( h: Y \to \mathbb{R} \) be proper, closed convex functions and \( A: E \to Y \) a linear map. Then for any point \( x \), the inclusion holds:

\[ \partial(g + h \circ A)(x) \supset \partial g(x) + A^\star \partial h(Ax). \] (3.7)

Moreover, equality holds under the regularity condition

\[ 0 \in \text{ri (dom } h) - A(\text{ri dom } g). \] (3.8)*
Proof. The inclusion 3.7 follows immediately by adding the subgradient inequalities for $g$ at $x$ and for $h$ at $Ax$. Assume now (3.8) holds. Fix a vector $v \in \partial (g + h \circ A)(x)$. Without loss of generality we may assume $v = 0$, since otherwise, we may replace $g$ by $g - \langle v, \cdot \rangle$. Thus $x$ minimizes

$$(P) \quad \min_z \ h(Ax) + g(x).$$

Consider now the dual problem

$$(D) \quad \max_y \ - g^*(-A^*y) - h^*(y).$$

Theorem 3.51 guarantees that the primal and dual optimal values are equal and the dual optimal value is attained. Letting $y$ be any dual optimal solution, we deduce

$$0 = (g(x) + h(Ax)) + (g^*(-A^*y) + h^*(y))
= (g(x) + g^*(-A^*y)) + (h(Ax) + h^*(y))
\geq \langle x, -A^*y \rangle + \langle Ax, y \rangle = 0,$$

where the last line follows from the Fenchel-Young inequality (Theorem 3.42). Hence equality holds throughout and we learn

$$g(x) + g^*(-A^*y) = \langle x, -A^*y \rangle \quad \text{and} \quad h(Ax) + h^*(y) = \langle Ax, y \rangle.$$

Using the characterization of equality in Theorem 3.42 we obtain the decomposition

$$0 = -A^*y + A^*y \in \partial g(x) + A^* \partial h(Ax),$$

as claimed.

Exercise 3.53. Consider a proper, closed, convex function $f : E \to \mathbb{R}$ and a point $\bar{x} \in \text{int} (\text{dom} f)$ such that $f(\bar{x})$ is not the minimum of $f$. Define the sublevel set $Q := \{x : f(x) \leq f(\bar{x})\}$. Prove the formula

$$N_Q(\bar{x}) = \mathbb{R}_+ \partial f(\bar{x}).$$

[Hint: Focus on the epigraphs.]

3.7.2 Lagrangian Duality

Let us next look at a problem class that does not directly fit in the framework of Fenchel-Rockafellar duality. Namely, consider the constrained optimization problem

$$(P) \quad \min f(x) \quad \text{s.t.} \ G(x) \in K, \quad x \in Q$$


where \( f : E \to \mathbb{R} \) is some function, \( G : E \to Y \) is a map, \( K \subset Y \) is a cone, and \( Q \in E \) is an arbitrary set. Notice that we may equivalently write \( (P) \) as \( \min_x h(G(x)) + g(x) \), where we set \( h = \delta_K \) and \( g = f + \delta_Q \). Since \( G \) is nonlinear, we can not appeal to Fenchel-Rockefeller duality. Instead, we will use a direct lower-bounding procedure.

Define the Lagrangian function

\[
L(x, y) := f(x) + \langle y, G(x) \rangle.
\]

Observe now

\[
\sup_{y \in Q^\circ} L(x, y) = \begin{cases} f(x) & \text{if } G(x) \in K \\ +\infty & \text{otherwise} \end{cases}.
\]

Consequently taking the infimum over \( x \in Q \), we deduce

\[
\inf_{x \in Q} \sup_{y \in K^\circ} L(x, y) = \text{val}(P).
\]

Hence any \( \bar{y} \in K^\circ \) certifies a lower bound on the primal optimal value:

\[
\text{val}(P) \geq \inf_{x \in Q} L(x, \bar{y}).
\]

Finding the best lower bound that is achievable in this way is the Lagrange dual optimization problem:

\[
(D) \quad \sup_{y \in K^\circ} \Phi(y) \quad \text{where} \quad \Phi(y) := \inf_{x \in Q} L(x, y).
\]

We will see that under reasonable convexity and compatibility assumptions, we can be sure that \textit{strong duality} holds: \( \text{val}(P) = \text{val}(D) \) and the dual optimal value is attained, if finite.

We will need the following definition. A map \( G : E \to Y \) is \textit{convex relative to a convex cone} \( K \subset Y \) if the inclusion

\[
G(\lambda y + (1 - \lambda)x) \in \lambda G(y) + (1 - \lambda)G(x) + K
\]

holds for all \( x, y \in E \) and \( \lambda \in [0, 1] \).

**Exercise 3.54.** Let \( g_1, g_2, \ldots, g_n : E \to \mathbb{R} \) be convex functions \( g_{n+1}, \ldots, g_m : E \to \mathbb{R} \) be affine functions. Then the map \( G(x) = (g_1(x), g_2(x), \ldots, g_m(x)) \) is convex relative to the convex cone \( K = \mathbb{R}^n \times \{0\}^{m-n} \).

**Exercise 3.55.** Show that if \( G : \mathbb{R} \to Y \) is convex relative to a convex cone \( K \subset Y \), then the function \( (x, y) \mapsto \delta_K(G(x) + y) \) is convex.

We now apply Theorem 3.50 to the Lagrange primal-dual pair.
Theorem 3.56. Consider the constrained optimization problem
\[
(P) \quad \min_x f(x) \quad \text{subject to } G(x) \in K, \ x \in Q.
\]
where \( f: E \to \mathbb{R} \) is a closed convex function, \( G: E \to Y \) is a continuous map that is convex relative to a closed convex cone \( K \subset Y \), and \( Q \subset E \) is a closed convex set. Define the dual problem
\[
(D) \sup_{y \in K^*} \Phi(y) \quad \text{where} \quad \Phi(y) := \inf_{x \in Q} f(x) + \langle y, G(x) \rangle.
\]
Then the two problems fit the perturbation framework of Theorem 3.50 with
\[
F(x, y) = \delta_Q(x) + f(x) + \delta_K(G(x) + y).
\]
In particular, the regularity condition
\[
\exists x \in Q \cap \text{dom} \ f \text{ with } G(x) \in \text{int} K \quad (3.9)
\]
guarantees that the primal and dual optimal values are equal, and the dual optimal value is attained, if finite.

Proof. Let us verify the assumptions of Theorem 3.50. To this end, define the function
\[
F(x, y) = \delta_Q(x) + f(x) + \delta_K(G(x) + y).
\]
Note that \( F \) is clearly proper and closed, and is convex by Exercise 3.55.

We successively compute
\[
F^*(0, v) = \sup_{x, y} \langle (0, v), (x, y) \rangle - F(x, y)
\]
\[
= \sup_{x \in Q, y} \langle v, y \rangle - f(x) - \delta_K(G(x) + y)
\]
\[
= \sup_{x \in Q, z} \langle v, z - G(x) \rangle - f(x) - \delta_K(z)
\]
\[
= \sup_{z \in K} \langle v, z \rangle - \inf_{x \in Q} \{ \langle v, G(x) \rangle + f(x) \}
\]
\[
= \delta_{Q^*}(v) - \Phi(v).
\]
Thus the Lagrange dual \((D)\) is precisely the problem \( q(0) = \max_v -F^*(0, v) \).

Next observe that \((3.9)\) directly implies \( 0 \in \text{int} (\text{dom} \ p) \). If \( p(0) = -\infty \), there is nothing to prove. Hence we may suppose that \( p(0) \) is finite. Theorem 3.26 then implies that \( p \) is proper, while Theorem 3.40 in turn guarantees that the subdifferential \( \partial p(0) \) is nonempty. An application of Theorem 3.51 completes the proof.

Exercise 3.57. Compute the Fenchel duals of the following problems.

1. \( \min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|^2 + \|x\|_1 \)
2. \( \min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \|Ax - b\|_2^2 : \|x\|_1 \leq \tau \right\} \)

3. \( \min_{x \in \mathbb{R}^n} \|Ax - b\|_\infty + \|x\|_1 \)

4. \( \min_{x \in \mathbb{R}^n} \left\{ \|Ax - b\|_\infty : \|x\|_1 \leq \tau \right\} \)

5. \( \min_{x} \{ \langle c, x \rangle : Ax = b, x \in K \} \), where \( K \) is a closed convex cone.

**Exercise 3.58.** Compute the Lagrangian duals of the following problems

1. \( \min_{x} \{ \langle c, x \rangle : Ax = b, x \in K \} \), where \( K \) is a closed convex cone.

2. \( \min_{X \in \mathbb{S}^n} \{ \langle C, X \rangle : \langle A_i, X \rangle = b_i \text{ for } i = 1, \ldots, m, \ X \succeq 0 \} \), where \( A_i \) are some \( n \times n \) symmetric matrices.

3. \( \min_{x \in \mathbb{R}^n} \{ \frac{1}{2} x^T Q x : Ax \leq b \} \), where \( Q > 0 \).

4. \( \min_{x \in \mathbb{R}^n} \{ \sum_{i=1}^k f(x_i) : Ax \leq b \} \), for a univariate function \( f : \mathbb{R} \to \mathbb{R} \).

**Exercise 3.59.** Suppose \( C \subset \mathbb{E} \) and \( D \subset \mathbb{Y} \) are compact, convex sets and \( \mathcal{A} : \mathbb{E} \to \mathbb{Y} \) is a linear mapping. Use Fenchel duality to prove

\[
\min_{x \in C} \max_{y \in D} \langle Ax, y \rangle = \max_{y \in D} \min_{x \in C} \langle Ax, y \rangle
\]

**Exercise 3.60.** Consider the problem

\[
(P) \quad \min \quad \langle Q_0, x \rangle \\
\text{s.t.} \quad \langle Q_j, x \rangle = b_j \quad \text{for } j = 1, \ldots, m
\]

where \( Q_j \) (for \( j = 0, 1, \ldots, m \)) are \( n \times n \) symmetric matrices and \( b \in \mathbb{R}^m \) is a vector.

1. Prove that the Lagrangian dual of this problem is the *Semi-definite Program*

\[
(D) \quad \max_{y} \quad y^T b \\
\text{s.t.} \quad Q_0 - \sum_{j=1}^m y_j Q_j \succeq 0 \quad (3.10)
\]

2. Use Lagrangian (or Fenchel) duality to derive the dual problem of (D):

\[
(\hat{P}) \quad \min \quad \langle Q_0, X \rangle \\
\text{s.t.} \quad \langle Q_j, X \rangle = b_j \quad \text{for } j = 1, \ldots, m \\
X \succeq 0.
\]

3. When there exists a feasible *positive* definite matrix for (\( \hat{P} \)), the optimal values of (\( \hat{P} \)) and (D) are equal. What is then the relationship between the optimal value of (\( P \)) and that of (\( \hat{P} \))? The problem (\( \hat{P} \)) is called a *convex relaxation* of \( P \). Why?
3.8 Moreau-Yosida envelope and the proximal map

In this section, we consider an important process that associates to any convex function, a smooth convex approximation. This operation is simply the infimal convolution with the convex quadratic \( \frac{1}{2\alpha} \| \cdot \|^2 \) for any \( \alpha > 0 \).

**Definition 3.61.** For any function \( f : E \to \mathbb{R} \) and real \( \alpha > 0 \), define the Moreau-Yosida envelope and the proximal map, respectively:

\[
\begin{align*}
  f_\alpha(x) &:= \min_y f(y) + \frac{1}{2\alpha} \| x - y \|^2 \\
  \text{prox}_\alpha f(x) &:= \arg\min_y f(y) + \frac{1}{2\alpha} \| x - y \|^2.
\end{align*}
\]

**Exercise 3.62.** Compute the proximal operator \( \text{prox}_\alpha f \) of the following functions: \( f(x) = \| x \|_1 \), \( f(x) = \| x \|_2 \), \( f(x) = \| x \|_\infty \).

In the simplest case that \( f \) is the indicator function of a closed convex set \( Q \), the moreau envelope is the squared distance \( f_\alpha(x) = \frac{1}{2\alpha} \text{dist}_Q^2(x) \), while the proximal map is the nearest point projection \( \text{prox}_\alpha f(x) = \text{proj}_Q(x) \). In analogy to the projection operator (Exercise 3.11), the following theorem shows that the proximal map of a proper, closed, convex function is single-valued and 1-Lipschitz continuous.

**Theorem 3.63.** Consider a proper closed convex function \( f : E \to \mathbb{R} \). Then the set \( \text{prox}_f(x) \) is a singleton for every point \( x \in E \). Moreover, for any \( x,y \in E \) the estimate holds:

\[
\| \text{prox}_f(x) - \text{prox}_f(y) \|^2 \leq (\text{prox}_f(x) - \text{prox}_f(y), x - y).
\]

In particular, the proximal map \( x \mapsto \text{prox}_f(x) \) is 1-Lipschitz continuous.

**Proof.** Notice that for every point \( x \), the function \( z \mapsto f(z) + \frac{1}{2} \| z - x \|^2 \) is closed and 1-strongly convex, and hence has a unique minimizer. Thus the proximal set \( \text{prox}_f \) evaluates to a singleton everywhere on \( E \), as claimed. Next, consider any two points \( x,y \in E \) and define \( x^+ = \text{prox}_f(x) \) and \( y^+ = \text{prox}_f(y) \). Then using strong convexity, we deduce

\[
\begin{align*}
  f(x^+) + \frac{1}{2} \| x^+ - x \|^2 &\leq \left( f(y^+) + \frac{1}{2} \| y^+ - x \|^2 \right) - \frac{1}{2} \| y^+ - x^+ \|^2 \\
  &\leq f(y^+) + \frac{1}{2} \| y^+ - y \|^2 - \frac{1}{2} \| y^+ - x^+ \|^2 \\
  &\quad + \frac{1}{2} \| y^+ - x \|^2 - \frac{1}{2} \| y^+ - y \|^2 \\
  &\leq f(x^+) + \frac{1}{2} \| x^+ - y \|^2 - \| y^+ - x^+ \|^2 \\
  &\quad + \frac{1}{2} \| y^+ - x \|^2 - \frac{1}{2} \| y^+ - y \|^2.
\end{align*}
\]
Rearranging yields the claimed estimate.

\[
\|y^+ - x^+\|^2 \leq \frac{1}{2} \left( \|x^+ - y\|^2 - \|y^+ - y\|^2 + \|y^+ - x\|^2 - \|x^+ - x\|^2 \right) = (x^+ - y^+, x - y).
\]

The Cauchy-Schwarz inequality then implies \(\|y^+ - x^+\| \leq \|x - y\|\) as claimed.

The proximal map of a convex function and that of its conjugate are closely related, as the following theorem shows.

**Theorem 3.64.** For any proper closed convex function \(f: E \to \mathbb{R}\), equality holds

\[
\text{prox}_f(x) + \text{prox}_{f^*}(x) = x \quad \forall x \in E.
\]

**Proof.** Using the Corollary 3.43, we successively deduce

\[
z = \text{prox}_f(x) \iff 0 \in \partial \left( f + \frac{1}{2}\|\cdot - x\|^2 \right)(z) \\
\iff x - z \in \partial f(z) \\
\iff z \in \partial f^*(x - z) \\
\iff 0 \in \partial f^*(x - z) - z \\
\iff 0 \in \partial \left( f^* + \frac{1}{2}\|\cdot - x\|^2 \right)(x - z) \iff x - z = \text{prox}_{f^*}(x).
\]

This completes the proof.

The following is the main result of this section. It shows that under standard convexity assumptions, the Moreau-Yosida envelope is continuously differentiable. Indeed, the gradient of \(f_\alpha\) is even \(\alpha^{-1}\)-Lipschitz continuous.

**Theorem 3.65.** For any proper, closed, convex function \(f: E \to \bar{\mathbb{R}}\), the envelope \(f_\alpha\) is continuously differentiable on \(E\) with gradient

\[
\nabla f_\alpha(x) = \alpha^{-1}(x - \text{prox}_\alpha f(x)). \quad (3.11)
\]

**Proof.** Suppose first \(\alpha = 1\) and fix a point \(x \in E\). We aim to show that the subdifferential of \(f_\alpha\) is a singleton at every point. To this end, using Corollary 3.43 we deduce

\[
z \in \partial f_\alpha(x) \iff x \in \partial(f \square \frac{1}{2}\|\cdot\|^2)^*(z) \\
\iff x \in \partial \left( f^* + \frac{1}{2}\|\cdot\|^2 \right)^*(z) \\
\iff x \in \partial f^*(z) + z \\
\iff 0 \in \partial f^*(z) + z \\
\iff z = \text{prox}_{f^*}(x) \\
\iff z = x - \text{prox}_f(x),
\]
where the first equivalence uses Corollary 3.43 and the definition of the Moreau-Yosida regularization, the second equivalence uses Exercise 3.38 and the last equivalence follows from Theorem 3.64. Thus we deduce that the subdifferential \( \partial f_\alpha(x) \) is a singleton. Hence by Theorem 3.45, the envelope \( f_\alpha \) is differentiable with \( \nabla f_\alpha(x) = x - \text{prox}_f(x) \). Finally, in the general setting \( \alpha \neq 1 \), simply observe that \( \alpha \cdot f_\alpha \) is the Moreau-Yosida envelope of \( \alpha f \) with parameter 1. Applying what we have already proved in this setting immediately yields the expression (3.11).

3.9 Orthogonally invariant functions

In this section, we study functions on the space of symmetric matrices that depend on the matrix only through its eigenvalues. Remarkably, we will be able to explicitly compute conjugates, subdifferentials, and proximal maps of such functions.

We begin with some notation. The symbol \( S^n \) will denote the Euclidean space of symmetric matrices, while \( O(n) \) will denote the set of \( n \times n \) orthogonal matrices. The symbol \( \Pi(n) \) will denote set of all permutation of \( \{1, \ldots, n\} \). In particular, for any vector \( x \in \mathbb{R}^n \) and permutation \( \pi \in \Pi(n) \), the symbol \( \pi x \) will denote the same vector as \( x \) but with coordinates permuted according to \( \pi \). Let us record the following two key definitions, whose close relationship will become clear momentarily.

**Definition 3.66.** A function \( f: \mathbb{R}^n \to \mathbb{R} \) is called symmetric if it satisfies

\[
f(\pi x) = f(x), \quad \text{for all } x \in \mathbb{R}^n, \pi \in \Pi(n).
\]

**Definition 3.67.** A function \( F: S^n \to \mathbb{R} \) is called spectral if it satisfies

\[
F(UXU^T) = F(X), \quad \text{for all } X \in S^n, U \in O(n).
\]

Typical examples of spectral functions are the negative log-determinant \( F(X) = -\ln \det(X) + \delta_{S^n_{++}} \) (see Exercise 1.10), the nuclear norm \( F(X) = \sum_i |\lambda_i(X)| \), the operator norm \( F(X) = \max |\lambda_i(X)| \), and the Frobenius norm \( F(X) = \sqrt{\sum_{i=1}^n \lambda_i^2(X)} \). Notice that all of these examples have an intriguing property: the function \( F(X) = f(\lambda(X)) \) for some simple symmetric function \( f: \mathbb{R}^n \to \mathbb{R} \). Indeed, all spectral functions factor in this way.

**Exercise 3.68.** A function \( F: S^n \to \mathbb{R} \) is spectral if and only if we can write \( F = f \circ \lambda \) for some symmetric function \( f: \mathbb{R}^n \to \mathbb{R} \).

[Hint: Explicitly, set \( f(x) = F(\text{Diag}(x)) \).]

In typical circumstances, the spectral function \( F \) may appear highly complicated, whereas \( f \) is simple (e.g. polyhedral). In this section,
we will see that numerous convex analytic properties of $F$ can be described purely in terms of the analogous properties of $f$. The key tool we will employ is Theorem 3.71, which establishes that the eigenvalue map on the space of symmetric matrices is Lipschitz continuous. The proof of the theorem will use the following two exercises. For a vector $x$ in $\mathbb{R}^n$, we denote by $x^\uparrow$ the vector with the same components permuted into nonincreasing order.

**Exercise 3.69.** For any two vectors $x, y \in \mathbb{R}^n$, the inequalities hold:

$$\langle x, y \rangle \leq \langle x^\uparrow, y^\uparrow \rangle.$$

Moreover, equality holds if and only if there exists a permutation $\pi$ satisfying $\pi(x) = x^\uparrow$ and $\pi(y) = y^\uparrow$.

**Exercise 3.70.** Consider a convex symmetric function $f : \mathbb{R}^n \to \mathbb{R}$. Then whenever the inclusion $v \in \partial f(x^\uparrow)$ holds, so does the inclusion $v^\uparrow \in \partial f(x^\uparrow)$.

**Theorem 3.71.** (Trace Inequality) For all $X, Y \in \mathbb{S}^n$, we have

$$\|\lambda(X) - \lambda(Y)\| \leq \|X - Y\|_F,$$

with equality if and only if $X$ and $Y$ admit a simultaneous ordered spectral decomposition, meaning that $\exists U \in O(n)$ satisfying

$$X = U(\text{Diag} (\lambda(X)))U^T, \quad Y = U(\text{Diag} (\lambda(Y)))U^T.$$

**Proof.** Let $X, Y \in \mathbb{S}^n$ and define the set

$$\mathcal{L} = \{UXU^T : U \in O(n)\}.$$

Since $\mathcal{L}$ is closed and compact, we may choose an arbitrary closest point $Z \in \mathcal{L}$ to $Y$. Thus we have

$$\|X - Y\|_F^2 \geq \|Z - Y\|_F^2 = \text{dist}_F^2(Y).$$

Fix now an arbitrary skew-symmetric matrix $W$ and notice that for any $t \in \mathbb{R}$, the matrix exponential $e^{tW}$ is an orthogonal matrix. Define the function

$$\varphi(t) = \frac{1}{2}\|e^{tW}Ze^{-tW} - Y\|_F^2.$$

It is straightforward to check that

$$\varphi'(t) = \langle YZ - ZY, W \rangle.$$

Setting $W := YZ - ZY$ and taking into account $\varphi'(t) = 0$, we deduce $YZ = ZY$. Since $Z$ and $Y$ commute, they must be simultaneously diagonalizable. Thus there exist $U \in O(n)$ and a permutation $\pi$ satisfying

$$Y = U\text{Diag} (\pi \cdot \lambda(Y))U^T, \quad Z = U\text{Diag} (\lambda(Z))U^T.$$
Recall moreover, $\lambda(Z) = \lambda(X)$ since $Z$ and $X$ both lie in $L$. Thus
\[
\|X - Y\|_F \geq \|Z - Y\|_F
= \|\text{Diag}(\pi \cdot \lambda(Y)) - \text{Diag}(\lambda(X))\|_F \geq \|\lambda(Y) - \lambda(X)\|.
\tag{3.14}
\]
This establishes the inequality (3.12). Moreover, if equality holds in (3.12), then we could set $Z = X$ in the first place. Thus $Z$ and $X$ are simultaneously diagonalizable and the estimate (3.14) guarantees the equality
\[
\|\pi \cdot \lambda(Y) - \lambda(X)\| \geq \|\lambda(Y) - \lambda(X)\|.
\]
Squaring and expanding, we deduce
\[
\langle \pi \cdot \lambda(Y), \lambda(X) \rangle = \langle \lambda(Y), \lambda(X) \rangle.
\]
Applying Exercise 3.69, we deduce that there exists a permutation $\hat{\pi}$ satisfying $\lambda(X) = \hat{\pi} \lambda(X)$ and $\pi \lambda(Y) = \hat{\pi} \lambda(Y)$. Consequently, using (3.13) we deduce
\[
Y = U \text{Diag}(\hat{\pi} \lambda(Y)) U^T = U \text{Diag}(\hat{\pi} \lambda(Y)) U^T.
\]
Permuting the columns of $U$ according to $\hat{\pi}^{-1}$ completes the proof.

We can prove a precise relationship between the Fenchel conjugate and the Moreau envelope of a spectral function $f \circ \lambda$ and those of $f$.

**Theorem 3.72.** Consider a symmetric function $f: \mathbb{R}^n \to \mathbb{R}$. Then the inequalities
\[
(f \circ \lambda)^*(Y) = (f^* \circ \lambda)(Y) \quad \text{and} \quad (f \circ \lambda)_\alpha(X) = f_\alpha(\lambda(X)),
\]
hold for all $X, Y \in S^n$.

**Proof.** We begin with the computation of the Fenchel conjugate. To this end, we successively compute
\[
(f \circ \lambda)^*(Y) = \sup_{X \in S^n} \{\langle X, Y \rangle - f(\lambda(X))\}
\leq \sup_{X \in S^n} \{\langle X, Y \rangle - f(\lambda(X))\}
\leq \sup_{z \in \mathbb{R}^n} \{\langle z, \lambda(Y) \rangle - f(z)\} = f^*(\lambda(Y)),
\tag{3.15}
\]
where (3.15) follows from Theorem 3.71. To see the reverse inequality, fix an eigenvalue decomposition $Y = U \text{Diag}(\lambda(Y)) U^T$ with $U \in O(n)$. Observe
\[
f^*(\lambda(Y)) = \sup_{z \in \mathbb{R}^n} \{\langle z, \lambda(Y) \rangle - f(z)\}
\leq \sup_{X \in S^n} \{\langle X, \lambda(Y) \rangle - f(\lambda(X))\}
= \sup_{X \in S^n} \{\langle U \text{Diag}(\lambda(X)) U^T, Y \rangle - f(\lambda(X))\}
\leq \sup_{Z \in S^n} \{\langle Z, Y \rangle - f(\lambda(Z))\} = (f \circ \lambda)^*(Y),
\tag{3.16}
\]
3.9. ORTHOGONALLY INVARIANT FUNCTIONS

where (3.16) uses Exercise 3.69. We conclude \((f \circ \lambda)(X) = (f^* \circ \lambda)(X)\), as claimed.

Next, we verify the analogous expression for the Moreau envelope. Fix a matrix \(Y\) and an eigenvalue decomposition \(X = U\text{Diag}(\lambda(X))U^T\) with \(U \in O(n)\). We then deduce

\[
(f \circ \lambda)(X) = \inf_{Z \in S^n} \left\{ f(\lambda(Z)) + \frac{1}{2\alpha} \|Z - X\|_F^2 \right\}
\]

\[
\leq \inf_{z \in \mathbb{R}^n} \left\{ f(z) + \frac{1}{2\alpha} \|U\text{Diag}(z)U^T - U\text{Diag}(\lambda(X))U^T\|_F^2 \right\}
\]

\[
\leq \inf_{z \in \mathbb{R}^n} \left\{ f(z) + \frac{1}{2\alpha} \|z - \lambda(X)\|^2 \right\} = f_\alpha(\lambda(X)).
\]

Conversely, observe

\[
f_\alpha(\lambda(X)) = \inf_{z \in \mathbb{R}^n} \left\{ f(z) + \frac{1}{2\alpha} \|z - \lambda(X)\|^2 \right\}
\]

\[
\leq \inf_{Z \in S^n} \left\{ f(\lambda(Z)) + \frac{1}{2\alpha} \|\lambda(Z) - \lambda(X)\|^2 \right\}
\]

\[
\leq \inf_{Z \in S^n} \left\{ f(\lambda(Z)) + \frac{1}{2\alpha} \|Z - X\|^2 \right\} = (f \circ \lambda)_\alpha(X) \tag{3.17}
\]

where (3.17) follows from Theorem 3.71. This completes the proof.

With Theorem 3.72 at hand, we can now obtain an explicit representation of the subdifferential \(\partial(f \circ \lambda)(X)\) in terms of \(\partial f(\lambda(X))\).

**Corollary 3.73.** Consider a proper, closed, symmetric function \(f : \mathbb{R}^n \to \mathbb{R}\). Then \(f\) is convex if and only if the spectral function \(f \circ \lambda\) is convex. Moreover, if \(f\) is convex, then the subdifferential \(\partial(f \circ \lambda)(X)\) consists of all matrices \(Y \in S^n\) that admit a simultaneous ordered spectral decomposition with \(X\) and satisfy \(\lambda(Y) \in \partial f(\lambda(X))\). In particular, the expression holds:

\[
\partial(f \circ \lambda)(X) = \left\{ U\text{Diag}(w)U^T : w \in \partial f(\lambda(X)), \ U \in O_X \right\},
\]

where

\[
O_X := \left\{ U \in O(n) : X = U\text{Diag}(\lambda(X))U^T \right\}.
\]

**Proof.** Suppose first \(f \circ \lambda\) is convex. Since \(f\) coincides with the restriction of \(f \circ \lambda\) to diagonal matrices, we immediately deduce that \(f\) is convex. Conversely, suppose that \(f\) is convex. Note that since \(f\) is symmetric, so is its conjugate \(f^*\). Consequently applying Theorem 3.72 twice and appealing to Theorem 3.37, we deduce

\[
((f \circ \lambda)^*)^* = (f^* \circ \lambda)^* = (f^{**} \circ \lambda) = f \circ \lambda.
\]

Since \(f \circ \lambda\) coincides with its double conjugate, we conclude that \(f \circ \lambda\) is convex.
Now suppose that \( f \) is convex. Suppose first \( Y \in \partial(f \circ \lambda)(X) \). Then Theorem 3.42 guarantees the equality
\[
(f \circ \lambda)^*(Y) + (f \circ \lambda)(X) = \langle Y, X \rangle.
\]
We therefore deduce
\[
\langle \lambda(Y), \lambda(X) \rangle \leq f^*(\lambda(Y)) + f(\lambda(X)) = (f \circ \lambda)^*(Y) + (f \circ \lambda)(X) \leq \langle \lambda(Y), \lambda(X) \rangle.
\] (3.18)

Hence equality holds throughout. Therefore from equality in (3.19), the matrices \( X \) and \( Y \) must admit a simultaneously ordered decomposition, while equality in (3.18) guarantees the inclusion \( \lambda(Y) \in \partial f(\lambda(X)) \).

Conversely, suppose \( X \) and \( Y \) admit a simultaneously ordered eigenvalue decomposition and the inclusion \( \lambda(Y) \in \partial f(\lambda(X)) \) holds. Then by similar reasoning as before, we have
\[
\langle \lambda(Y), \lambda(X) \rangle = f^*(\lambda(Y)) + f(\lambda(X)) = (f \circ \lambda)^*(Y) + (f \circ \lambda)(X) \geq \langle Y, X \rangle = \langle \lambda(Y), \lambda(X) \rangle.
\] (3.20)

Thus equality holds throughout. In particular, from equality in (3.20) we conclude \( Y \in \partial(f \circ \lambda)(X) \) as claimed. In particular, the displayed expression for the subdifferential \( f \circ \lambda \) now follow immediately from Exercise 3.70.

\textbf{Exercise 3.74.} Define the function \( F : \mathbb{S}^n \to \mathbb{R} \) by \( F(X) = \|X\|_{op} \). Prove the expression
\[
\partial(f \circ \lambda)(I) = \{Y \succeq 0 : \text{tr}(Y) = 1\}.
\]

\textbf{Exercise 3.75.} Consider a symmetric function \( f : \mathbb{R}^n \to \mathbb{R} \). Establish the formula for the proximal map:
\[
\text{prox}_{f \circ \lambda}(X) = \{UDiag(w)U^T : w \in \text{prox}_f(\lambda(X)), U \in O_X\},
\]
where
\[
O_X := \{U \in O(n) : X = UDiag(\lambda(X))U^T\}.
\]

It is possible to develop a completely parallel theory for “orthogonally invariant” functions on the Euclidean space of rectangular matrices. We now outline such results, leaving the details for the interested reader to fill in.

We will call a function \( F : \mathbb{R}^{m \times n} \to \mathbb{R} \) \emph{orthogonally invariant} if we have
\[
F(UXY^T) = F(X) \quad \text{for all } X \in \mathbb{R}^{m \times n}, U \in O(m), V \in O(n).
\]
Without loss of generality, suppose \( m \leq n \). We will say that a function \( f : \mathbb{R}^m \to \mathbb{R} \) is absolutely symmetric if \( f(\pi x) = x \) for all \( x \in \mathbb{R}^m \) and all signed permutations \( \pi \). It is straightforward to see that a function \( F \) is orthogonally invariant if and only if it factors as \( F = f \circ \sigma \), where \( f \) is some absolutely symmetric function on \( \mathbb{R}^m \) and \( \sigma \) is the singular value map. Similar arguments as in the symmetric case then show the following theorem.

**Theorem 3.76.** Consider an absolutely symmetric, proper, closed function \( f : \mathbb{R}^m \to \mathbb{R} \). Then the expressions hold

\[
(f \circ \sigma)^*(Y) = (f^* \circ \sigma)(Y) \quad \text{and} \quad (f \circ \sigma)_\alpha(X) = f_\alpha(\sigma(X)),
\]

and

\[
\text{prox}_{f \circ \sigma}(X) = \left\{ UD \text{Diag}(w)V^T : w \in \text{prox}_f(\sigma(X)), \ (U,V) \in O_X \right\},
\]

where we define

\[
O_X := \{(U,V) \in O(m) \times O(n) : X = U \text{Diag}(\sigma(X))V^T \}.
\]

Moreover, \( f \circ \sigma \) is convex if and only if \( f \) is convex, in which case the subdifferential admits the form

\[
\partial (f \circ \sigma)(X) = \left\{ UD \text{Diag}(w)V^T : w \in \partial f(\lambda(X)), \ (U,V) \in O_X \right\}.
\]

**References.**

All of the results in this section can be found in the standard monographs on convex analysis, such as [3, 14, 15]. The material in Section 3.9 follows the discussion in the papers [6, 7].
Bibliography


