Convex Analysis and Nonsmooth Optimization

Aleksandr Y. Aravkin, James V. Burke Dmitriy Drusvyatskiy

January 7, 2019

Contents

Part I

Convex Optimization

Chapter 1

Review of Fundamentals

1.1 Inner products and linear maps

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

(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 $\mathbf{R}^n$, 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 $\mathbf{E}$ can be identified with $\mathbf{R}^n$ 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 $\mathbf{R}^{m \times n}$ furnishes another example of an Euclidean space, which we always equip with the trace product $\langle X, Y \rangle := \operatorname{tr} X^T Y$. Some arithmetic shows the equality $\langle X, Y \rangle = \sum_{i,j} X_{ij} Y_{ij}$. Thus the trace product on $\mathbf{R}^{m \times n}$ 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 $\mathbf{R}^{n \times n}$ is the space of real symmetric $n \times n$-matrices $\mathbf{S}^n$, along with the trace product $\langle X, Y \rangle := \operatorname{tr} XY$.

For any linear mapping $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$, there exists a unique linear mapping $\mathcal{A}^*\colon \mathbf{Y} \to \mathbf{E}$, called the *adjoint*, satisfying

$$\langle \mathcal{A}x, y\rangle = \langle x, \mathcal{A}^* y\rangle \qquad \text{for all points} \qquad x \in \mathbf{E},\, y \in \mathbf{Y}.$$

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

Exercise 1.1. Given a collection of real $m \times n$ matrices $A_1, A_2, \ldots, A_l$, define the linear mapping $\mathcal{A}\colon \mathbf{R}^{m\times n} \to \mathbf{R}^l$ by setting

$$\mathcal{A}(X) := (\langle A_1, X\rangle, \langle A_2, X\rangle, \ldots, \langle A_l, X\rangle).$$

Show that the adjoint is the mapping $\mathcal{A}^* y = y_1 A_1 + y_2 A_2 + \ldots + y_l A_l$.

Linear mappings $\mathcal{A}$ between $\mathbf{E}$ and itself are called *linear operators*, and are said to be *self-adjoint* if equality $\mathcal{A} = \mathcal{A}^*$ holds. Self-adjoint operators on $\mathbf{R}^n$ are precisely those operators that are representable as symmetric matrices. A self-adjoint operator $\mathcal{A}$ is *positive semi-definite*, denoted $\mathcal{A} \succeq 0$, whenever

$$\langle \mathcal{A}x, x\rangle \geq 0 \quad \text{for all } x \in \mathbf{E}.$$

Similarly, a self-adjoint operator $\mathcal{A}$ is *positive definite*, denoted $\mathcal{A} \succ 0$, whenever

$$\langle \mathcal{A}x, x\rangle > 0 \quad \text{for all } 0 \neq x \in \mathbf{E}.$$

A positive semidefinite linear operator $\mathcal{A}$ is positive definite if and only if $\mathcal{A}$ is invertible.

Consider a self-adjoint operator $\mathcal{A}$. A number λ is an *eigenvalue* of X if there exists a vector $0 \neq v \in \mathbf{E}$ satisfying $\mathcal{A}v = \lambda v$. Any such vector v is called an *eigenvector* corresponding to λ. The Rayleigh-Ritz theorem shows that the following relation always holds:

$$\lambda_{\min}(\mathcal{A}) \leq \frac{\langle \mathcal{A}u, u\rangle}{\langle u, u\rangle} \leq \lambda_{\max}(\mathcal{A}) \quad \text{for all } u \in \mathbf{E} \setminus \{0\},$$

where $\lambda_{\min}(\mathcal{A})$ and $\lambda_{\max}(\mathcal{A})$ are the minimal and maximal eigenvalues of $\mathcal{A}$, respectively. Consequently, an operator $\mathcal{A}$ is positive semidefinite if and only $\lambda_{\min}(\mathcal{A}) \geq 0$ and $\mathcal{A}$ is positive definite if and only $\lambda_{\min}(\mathcal{A}) > 0$.

1.2 Norms

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

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

(Triangle inequality) $\|x + y\| \leq \|x\| + \|y\|$

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

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

Other important norms are the l_p*−norms* on $\mathbf{R}^n$:

$$\|x\|_p = \begin{cases} (|x_1|^p + \ldots + |x_n|^p)^{1/p} & \text{for } 1 \le p < \infty \\ \max\{|x_1|, \ldots, |x_n|\} & \text{for } p = \infty \end{cases}.$$

The most notable of these are the l_1, l_2, and l_∞ norms. For an arbitrary norm $\|\cdot\|$ on $\mathbf{E}$, the dual norm $\|\cdot\|^*$ on $\mathbf{E}$ is defined by

$$\|v\|^* := \max\{\langle v, x\rangle : \|x\| \le 1\}.$$

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

$$|\langle x, y\rangle| \le \|x\| \cdot \|y\|^*.$$

Exercise 1.2. Given a positive definite linear operator $\mathcal{A}$ on $\mathbf{E}$, show that the assignment $\langle v, w\rangle_{\mathcal{A}} := \langle \mathcal{A}v, w\rangle$ is an inner product on $\mathbf{E}$, with the induced norm $\|v\|_{\mathcal{A}} = \sqrt{\langle \mathcal{A}v, v\rangle}$, and dual norm $\|v\|_{\mathcal{A}}^* = \|v\|_{\mathcal{A}^{-1}} = \sqrt{\langle \mathcal{A}^{-1}v, v\rangle}$

All norms on $\mathbf{E}$ 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 $\alpha, \beta \ge 0$ satisfying

$$\alpha\rho_1(x) \le \rho_2(x) \le \beta\rho_1(x) \qquad \text{for all } x \in \mathbf{E}.$$

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

$$\begin{aligned} \|x\|_2 &\le \|x\|_1 \le \sqrt{n}\|x\|_2 \\ \|x\|_\infty &\le \|x\|_2 \le \sqrt{n}\|x\|_\infty \\ \|x\|_\infty &\le \|x\|_1 \le n\|x\|_\infty. \end{aligned}$$

For our purposes, the term "equivalent" is a misnomer: the proportionality constants α, β strongly depend on the (often enormous) dimension of the vector space $\mathbf{E}$. Hence measuring quantities in different norms can yield strikingly different conclusions.

Consider a linear map $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$, and norms $\|\cdot\|_a$ on $\mathbf{E}$ and $\|\cdot\|_b$ on $\mathbf{Y}$. We define the *induced matrix norm*

$$\|\mathcal{A}\|_{a,b} := \max_{x\colon \|x\|_a \le 1} \|\mathcal{A}x\|_b.$$

The reader should verify the inequality

$$\|\mathcal{A}x\|_b \le \|\mathcal{A}\|_{a,b}\|x\|_a.$$

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

Exercise 1.3. Equip $\mathbf{R}^n$ and $\mathbf{R}^m$ with the l_p-norms. Then for any matrix $A \in \mathbf{R}^{m\times n}$, show the equalities

$$\begin{aligned}\|A\|_1 &= \max_{j=1,\ldots,n} \|A_{\bullet j}\|_1\\ \|A\|_\infty &= \max_{i=1,\ldots,n} \|A_{i\bullet}\|_1\end{aligned}$$

where $A_{\bullet j}$ and $A_{i\bullet}$ denote the j'th column and i'th row of A, respectively.

1.3 Eigenvalue and singular value decompositions of matrices

The symbol $\mathbf{S}^n$ 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 $X^TX = XX^T = I$. Any symmetric matrix $A \in \mathbf{S}^n$ admits an *eigenvalue decomposition*, meaning a factorization of the form $A = U\Lambda U^T$ with $U \in O(n)$ and $\Lambda \in \mathbf{S}^n$ a diagonal matrix. The diagonal elements of Λ are precisely the eigenvalues of A and the columns of U are corresponding eigenvectors.

More generally, any matrix $A \in \mathbf{R}^{m\times n}$ admits a *singular value decomposition*, meaning a factorization of the form $A = UDV^T$, where $U \in O(m)$ and $V \in O(n)$ are orthogonal matrices and $D \in \mathbf{R}^{m\times n}$ 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 $m \le n$, the singular values of A are precisely the square roots of the eigenvalues of AA^T. In particular, the operator norm of any matrix $A \in \mathbf{R}^{m\times n}$ 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 $B_r(x) := \{y \in \mathbf{E} : \|y - x\| < r\}$. The *closure* of a set $Q \subset \mathbf{E}$, denoted $\operatorname{cl} Q$, consists of all points x such that the ball $B_\epsilon(x)$ intersects Q for all $\epsilon > 0$; the *interior* of Q, written as $\operatorname{int} Q$, 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 closure. Any set Q in $\mathbf{E}$ 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 $Q \subset \mathbf{E}$ admits a subsequence converging to a point in Q.*

For the rest of the section, we let $\mathbf{E}$ and $\mathbf{Y}$ be two Euclidean spaces, and U an open subset of $\mathbf{E}$. A mapping $F\colon Q \to \mathbf{Y}$, defined on a subset $Q \subset \mathbf{E}$, 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 $\mathbf{E}$ and $\mathbf{Y}$. 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 $f\colon Q \to \mathbf{R}$ on a compact set $Q \subset \mathbf{E}$ attains its supremum and infimum values.*

A function $f\colon U \to \mathbf{R}$ 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 $\mathbf{E} = \mathbf{R}^n$, 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 *β-smooth* if f is C^1-smooth and its gradient mapping ∇f is β-Lipschitz continuous.

More generally, consider a mapping $F\colon U \to \mathbf{Y}$. We say that F is *differentiable* at $x \in U$ if there exists a linear mapping taking $\mathbf{E}$ to $\mathbf{Y}$, 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 $\mathbf{E} = \mathbf{R}^n$ and $\mathbf{Y} = \mathbf{R}^m$, 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 $f\colon U \to \mathbf{R}$ is *twice differentiable* at a point $x \in U$ if the gradient map $\nabla f\colon U \to \mathbf{E}$ 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 $\mathbf{E} = \mathbf{R}^n$, the Hessian is the matrix of second-order partial derivatives

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f_1(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix}.$$

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

Exercise 1.6. Define the function

$$f(x) = \tfrac{1}{2}\langle \mathcal{A}x, x\rangle + \langle v, x\rangle + c$$

where $\mathcal{A}\colon \mathbf{E} \to \mathbf{E}$ 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 $F\colon \mathbf{E} \to \mathbf{Y}$ is a C^1-smooth mapping. Prove the identity $\nabla f(x) = \nabla F(x)^* F(x)$.

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

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

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

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

$$\begin{aligned} \mathbf{R}^n_{++} &:= \{x \in \mathbf{R}^n : x_i > 0 \text{ for all } i = 1, \ldots, n\}, \\ \mathbf{S}^n_{++} &:= \{X \in \mathbf{S}^n : X \succ 0\}. \end{aligned}$$

Consider the two functions $f\colon \mathbf{R}^n_{++} \to \mathbf{R}$ and $F\colon \mathbf{S}^n_{++} \to \mathbf{R}$ given by

$$f(x) = -\sum_{i=1}^n \log x_i \qquad \text{and} \qquad F(X) = -\ln\det(X),$$

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 $x \in \mathbf{R}^n_{++}$.

2. Prove $\nabla F(X) = -X^{-1}$ and $\nabla^2 F(X)[V] = X^{-1}VX^{-1}$ for any $X \succ 0$.

3. Using the property $\mathrm{tr}\,(AB) = \mathrm{tr}\,(BA)$, prove
$$\langle \nabla^2 F(X)[V], V\rangle = \|X^{-\frac{1}{2}}VX^{-\frac{1}{2}}\|_F^2$$
for any $X \succ 0$ and $V \in \mathcal{S}^n$. Deduce that the operator $\nabla^2 F(X)\colon \mathbf{S}^n \to \mathbf{S}^n$ is positive definite.

1.5 Fundamental theorems of calculus & accuracy in approximation

For any two points $x, y \in \mathbf{E}$, 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 $\mathbf{E}$ 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 $\mathbf{E}$. Consider a C^1-smooth function $f \colon U \to \mathbf{R}$ 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

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

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 $f \colon U \to \mathbf{R}$ and two points $x, y \in U$. Define the univariate function $\varphi \colon [0, 1] \to \mathbf{R}$ 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

$$\varphi'(t) = \langle \nabla f(x_t), y - x\rangle \quad \text{holds for any } t \in (0, 1).$$

2. Show that if f is C^2-smooth, then equality

$$\varphi''(t) = \langle \nabla^2 f(x_t)(y - x), y - x\rangle \quad \text{holds for any } t \in (0, 1).$$

The fundamental theorem of calculus now takes the following form.

Theorem 1.12 (Fundamental theorem of multivariate calculus)**.** *Consider a C^1-smooth function $f \colon U \to \mathbf{R}$ 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 $f : U \to \mathbf{R}$ 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

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

Appealing to Excercise 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 ∇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 β-smooth function $f\colon U \to \mathbf{R}$, for any points $x, y \in U$ the inequality*

$$\Big|f(y) - l(x; y)\Big| \leq \frac{\beta}{2}\|y-x\|^2 \quad \textit{holds.}$$

If f is C^2-smooth with M-Lipschitz Hessian, then

$$\Big|f(y) - Q(x; y)\Big| \leq \frac{M}{6}\|y-x\|^3.$$

It is now straightforward to extend the results in this section to mappings $F\colon U \to \mathbf{R}^m$. Given a curve $\gamma\colon \mathbf{R} \to \mathbf{R}^m$, we define the intergral $\int_0^1 \gamma(t)\, dt = \left(\int_0^1 \gamma_1(t)\, dt, \ldots, \int_0^1 \gamma_m(t)\, dt\right)$, where γ_i are the coordinate functions of γ. The main observation is that whenever γ_i are integrable, the inequality

$$\left\|\int_0^1 \gamma(t)\, dt\right\| \leq \int_0^1 \|\gamma(t)\|\, dt \quad \text{holds.}$$

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 $F\colon U \to \mathbf{R}^m$ and two points $x, y \in U$. Derive the equations

$$F(y) - F(x) = \int_0^1 \nabla F(x + t(y-x))(y-x)\, dt.$$

$$F(y) = F(x) + \nabla F(x)(y-x) + \int_0^1 (\nabla F(x + t(y-x)) - \nabla F(x))(y-x)\, dt.$$

In particular, consider a C^1-smooth mapping $F\colon U \to \mathbf{Y}$, where $\mathbf{Y}$ is some Euclidean space, and a point $\bar{x} \in U$. Choosing an orthonormal basis for $\mathbf{Y}$ and applying Excercise 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 β-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 $F\colon U \to \mathbf{Y}$ is L-Lipschitz continuous if and only if $\|\nabla F(x)\| \leq L$ for all $x \in \mathbf{E}$.

1.6 Tangent Cones: First-Order Approximation of Sets

Just as a first-order approximation to function is extremely useful in applications, so are “first-order” approximations of sets. This approximation to a set $S \subset \mathbf{E}$ at a point $x \in S$ is called the *tangent cone* to S at x. Our notion of tangency is based on the *distance to a set* given by

$$\operatorname{dist}_S(y) := \inf_{x \in S} \|y - x\|.$$

Definition 1.17 (Tangent Cone)**.** Let $S \subset \mathbf{E}$. We say that the vector v is tangent to S at a point $\bar{x} \in S$ if for $t > 0$

$$\operatorname{dist}_S(\bar{x} + tv) \leq o(t).$$

We call the set of all such tangent vectors the tangent cone to S at $\bar{x}$ and denote it by $T(\bar{x} \,|\, S)$.

Exercise 1.18. Show that $T(\bar{x} \,|\, S)$ is a closed cone, where a set $K \subset \mathbf{E}$ is said to be a *cone* if $\lambda K \subset K$ for all $\lambda > 0$.

Exercise 1.19. Show that

$$\begin{aligned} T(\bar{x} \,|\, S) &= \left\{ v \,\middle|\, \exists \{x^k\} \subset S,\, t_k \downarrow 0, \text{ s.t. } t_k^{-1}(x^k - \bar{x}) \to v \right\} \\ &= \left\{ tu \,\middle|\, t > 0,\ \exists \{x^k\} \subset S \setminus \{\bar{x}\},\, x^k \to \bar{x},\ \text{s.t. } \frac{x^k - \bar{x}}{\|x^k - \bar{x}\|} \to u \right\} \cup \{0\}. \end{aligned}$$

Exercise 1.20. If C is a nonempty convex subset of $\mathbf{E}$, show that

$$T(\bar{x} \,|\, C) = \operatorname{cl}\{t(x - \bar{x}) \,|\, x \in C,\ t \geq 0\}.$$

Exercise 1.21. If C is a convex polyhedron, show that

$$T(\bar{x} \,|\, C) = \{t(x - \bar{x}) \,|\, x \in C,\ t \geq 0\}.$$

Recall that C is a convex polyhedon if there exist $a^i \in E$ and $\beta_i \in \mathbf{R}\, i = 1, \dots, k$ such that $C = \left\{ x \,\middle|\, \langle a^i, x \rangle \leq \beta_i,\ i = 1, \dots, k \right\}$.

Exercise 1.22. Let $f : \mathbf{E} \to \mathbf{R}$ be C^1-smooth and set

$$\operatorname{gph} f := \{(x, f(x)) \,|\, x \in \mathbf{E}\}$$

be the *graph* of f. Show that

$$T((\bar{x}, f(\bar{x})) \,|\, \operatorname{gph} f) = \left\{ (v, \nabla f(\bar{x})^T v) \,|\, v \in \mathbf{E} \right\}.$$

That is, $T((\bar{x}, f(\bar{x})) \,|\, \operatorname{gph} f)$ is the subspace parallel to the graph of the linearization of f at $\bar{x}$.

The great challenge in using tangent cones is the development of a calculus that is as rich as the one available for differentiable functions.

Chapter 2

Smooth minimization

In this chapter, we consider the problem of minimizing a smooth function on a Euclidean space $\mathbf{E}$. Such problems are ubiquitous in computation mathematics and applied sciences. Before we delve into the formal development, it is instructive to look at some specific and typical examples that will motivate much of the discussion. We will often refer to these examples in latter parts of the chapter to illustrate the theory and techniques.

Example 2.1 (Linear Regression)**.** Suppose we wish to predict an output $b \in \mathbf{R}$ of a certain system on an input $a \in \mathbf{R}^n$. Let us also make the following two assumptions: (i) the relationship between the input a and the output b is fairly simple and (ii) we have available examples $a_i \in \mathbf{R}^n$ together with inexactly observed responses $b_i \in \mathbf{R}$ for $i = 1, \ldots, m$. Taken together, $\{(b, a_i)\}_{i=1}^m$ is called the training data.

Linear regression is an important example, where we postulate a linear relationship between the examples a and the response b. Trying to learn such a relationship from the training data amounts to finding a weight vector $x \in \mathbb{R}^{n+1}$ satisfying

$$b_i \approx x_0 + \langle a_i, x \rangle \qquad \text{for each } i = 1, \ldots, m.$$

To simplify notation, we may assume that the examples a_i lie in $\mathbf{R}^{n+1}$ with the first coordinate of a_i equal to one, so that we can simply write $b_i \approx \langle a_i, x \rangle$. The linear regression problem then takes the form

$$\min_x \sum_{i=1}^n \tfrac{1}{2}|\langle a_i, x \rangle - b_i|^2 = \tfrac{1}{2}\|Ax - b\|^2,$$

where $A \in \mathbb{R}^{m \times (n+1)}$ is a matrix whose rows are the examples a_i and b is the vector of responses. The use of the squared l_2-norm as a measure of misfit is a choice here. Other measures of misfit are often more advantageous from a modeling viewpoint – more on this later.

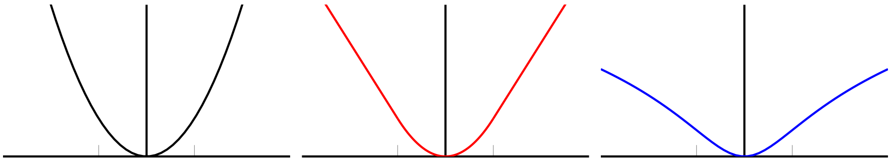

Figure 2.1: Least squares, Huber, and Student's t penalties. All three take (nearly) identical values for small inputs, but Huber and Student's t penalize larger inputs less harshly than least squares.

Example 2.2 (Ridge regularization)**.** The linear regression problem always has a solution, but the solution is only unique if A has a trivial null-space. To obtain unique solutions, as well as to avoid "over-fitting", *regularization* is often incorporated in learning problems. The simplest kind of regularization is called Tikhonov regularization or ridge regression:

$$\min_x \ \tfrac{1}{2}\|Ax-b\|^2+\lambda\|x-x_0\|^2,$$

where $\lambda > 0$ is a regularization parameter that must be chosen, and x_0 is most often taken to be the zero vector.

Example 2.3 (Robust Regression)**.** While least squares is a good criterion in many cases, it is known to be vulnerable to outliers in the data. Therefore other smooth criteria ρ can be used to measure the discrepancy between b_i and $\langle a_i, x\rangle$:

$$\min_x \ \sum_{i=1}^m \rho(\langle a_i, x\rangle - b_i).$$

Two common examples of robust penalties are:

- Huber: $\rho_\kappa(z) = \begin{cases} \frac{1}{2}\|z\|^2 & |z| \le \kappa \\ \kappa|z| - \frac{1}{2}\kappa^2 & |z| > \kappa. \end{cases}$
- Student's t: $\rho_\nu(z) = \log(\nu + z^2)$.

Note that both (nearly) agree with $\frac{1}{2}\|z\|^2$ for small values of z, but penalize larger z less harshly, see Figure 2.1.

Example 2.4 (General Linear Models)**.** The use of the squared l_2-norm in linear regression (Example 2.1) was completely ad hoc. Let us see now how statistical modeling dictates this choice and leads to important extensions. Suppose that the observed response b_i is a realization of a random variable $\mathbf{b}_i$, which is indeed linear in the input vector up to an additive statistical error. That is, assume that there is a vector $x \in \mathbf{R}^{n+1}$ satisfying

$$\mathbf{b}_i = \langle a_i, x\rangle + \boldsymbol{\varepsilon}_i \qquad \text{for } i = 1, \ldots, m,$$

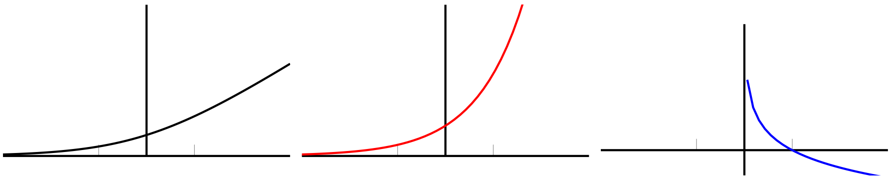

Figure 2.2: Penalties $\log(1+\exp(\cdot))$, $\exp(\cdot)$, and $-\log(\cdot)$ are used to model binary observations, count observations, and non-negative observations.

where ε_i is a normally distributed random variable with zero mean and variance σ^2. Thus $\mathbf{b}_i$ is normally distributed with mean $\mu_i := \langle a_i, x\rangle$ and variance σ^2. Assuming that the responses are independent, the *likelihood* of observing $\mathbf{b}_i = b_i$ for $i = 1, \dots, m$ is given by

$$L(\{b_i|\mu_i, \sigma^2\}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{\sigma^2}\sum_{i=1}^{m}\frac{1}{2}(b_i - \mu_i)^2\right).$$

To find a good choice of x, we maximize this likelihood with respect to x, or equivalently, minimize its negative logarithm:

$$\begin{aligned}\min_x \quad -\log(L\{b_i|\mu_i(x), \sigma^2\}) &= \min_x \; \tfrac{1}{2}\sum_{i=1}^{m}(b_i - \langle a_i, x\rangle)^2 \\ &= \min_x \; \tfrac{1}{2}\|Ax - b\|^2.\end{aligned}$$

This is exactly the linear regression problem in Example 2.1

The assumption that the $\mathbf{b_i}$ are normally distributed limits the systems one can model. More generally, responses $\mathbf{b}_i$ can have special restrictions. They may be count data (number of trees that fall over in a storm), indicate class membership (outcome of a medical study, hand-written digit recognition), or be non-negative (concentration of sugar in the blood). These problems and many others can be modeled using *general linear models (GLMs)*. Suppose the distribution of $\mathbf{b}_i$ is parametrized by (μ_i, σ^2):

$$L(b_i|\mu_i, \sigma^2) = g_1(b_i, \sigma^2)\exp\left(\frac{b_i\mu_i - g_2(\mu_i)}{g_3(\sigma^2)}\right).$$

To obtain the GLM, set $\mu_i := \langle a_i, x\rangle$, and minimize the negative log-likelihood:

$$\min_x \sum_{i=1}^{m} g_2(\langle a_i, x\rangle) - b_i\langle a_i, x\rangle,$$

ignoring g_1 and g_3 as they do not depend on x. This problem is smooth exactly when g_2 is smooth. Important examples are shown in Figure 2.2:

- Linear regression ($b_i \in \mathbf{R}$): $g_2(z) = \frac{1}{2}\|z\|^2$.

- Binary classification ($b_i \in \{0,1\}$): $g_2(z) = \log(1 + \exp(z))$.
- Poisson regression ($b_i \in \mathbb{Z}_+$): $g_2(z) = \exp(z)$.
- Exponential regression ($b_i \geq 0$): $g_2(z) = -\ln(z)$.

Example 2.5 (Nonlinar inverse problems)**.** Suppose that we are given multivariate responses $b_i \in \mathbf{R}^k$, along with functions $f_i : \mathbf{R}^n \to \mathbf{R}^k$. Our task is to find the best weights $x \in \mathbf{R}^n$ to describe b_i. This gives rise to *nonlinear* least squares:

$$\min_x \sum_{i=1}^m \frac{1}{2}\|f_i(x) - b_i\|^2$$

For example, global seismologists image the subsurface of the earth using earthquake data. In this case, x encodes density of subterranean layers and initial conditions at the start the earthquake i (e.g. 'slip' of a tectonic plate), b_i are echograms collected during earthquake i, and f_i is a (smooth) function of layer density and initial conditions that predicts b_i.

Example 2.6 (Low-rank factorization)**.** Suppose that we can observe some entries a_{ij} of a large matrix $A \in \mathbf{R}^{m\times n}$, with ij ranging over some small index set $\mathcal{I}$. The goal in many applications is to recover A (i.e. fill in the missing entries) from the partially observed information and an a priori upper bound k on the rank of A. One approach is to determine a factorization $A = LR^T$, for some matrices $L \in \mathbf{R}^{m\times k}$ and $R \in \mathbf{R}^{n\times k}$. This approach leads to the problem

$$\min_{L,R} \tfrac{1}{2} \sum_{ij\in\mathcal{I}} \|(LR^T)_{ij} - a_{ij}\|^2 + g(L,R),$$

where g is a smooth regularization function. Such formulations were successfully used, for exampe, to 'fill in' the Netflix Prize dataset, where only about 1% of the data (ratings of 15000 movies by 500,000 users) was present.

2.1 Optimality conditions: Smooth Unconstrained

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

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

where $f \colon \mathbf{E} \to \mathbf{R}$ 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 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 2.7. *(First-order necessary conditions) Suppose that x is a local minimizer of a function $f : U \to \mathbf{R}$. If f is differentiable at x, then equality $\nabla f(x) = 0$ holds.*

Proof. Set $v := -\nabla f(x)$. Then for all small $t > 0$, we deduce from the definition of derivative

$$0 \leq \frac{f(x+tv) - f(x)}{t} = -\|\nabla f(x)\|^2 + \frac{o(t)}{t}.$$

Letting t tend to zero, we obtain $\nabla f(x) = 0$, as claimed. □

A point $x \in U$ is a *critical point* for a C^1-smooth function $f \colon U \to \mathbf{R}$ if equality $\nabla f(x) = 0$ holds. Theorem 2.7 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 2.8. *(Second-order conditions)*
Consider a C^2-smooth function $f \colon U \to \mathbf{R}$ 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,

$$\liminf_{y \to x} \frac{f(y) - f(x)}{\frac{1}{2}\|y - x\|^2} \geq \lambda_{\min}(\nabla^2 f(x)).$$

Proof. Suppose first that x is a local minimizer of f. Then Theorem 2.7 guarantees $\nabla f(x) = 0$. Consider an arbitrary vector $v \in \mathbf{E}$. Then for all $t > 0$, we deduce from a second-order expansion

$$0 \leq \frac{f(x+tv) - f(x)}{\frac{1}{2}t^2} = \langle \nabla^2 f(x)v, v\rangle + \frac{o(t^2)}{t^2}.$$

Letting t tend to zero, we conclude $\langle \nabla^2 f(x)v, v\rangle \geq 0$ for all $v \in \mathbf{E}$, 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

$$\begin{aligned}\frac{f(y) - f(x)}{\frac{1}{2}\|y-x\|^2} &= \left\langle \nabla^2 f(x)\left(\frac{y-x}{\|y-x\|}\right), \frac{y-x}{\|y-x\|}\right\rangle + \frac{o(\|y-x\|^2)}{\|y-x\|^2}\\ &\geq \lambda_{\min}(\nabla^2 f(x)) + \frac{o(\|y-x\|^2)}{\|y-x\|^2}.\end{aligned}$$

Letting y tend to x, the result follows. □

The reader may be misled into believing that the role of the necessary conditions and the sufficient conditions for optimality (Theorem 2.8) 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 2.3, 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 Optimality Conditions: Smooth Constrained

By using the tangent cones in Definition 1.17, as well as Exercises 1.20 and 1.21, simple optimality conditions for C^1-smooth convexly constrained problems are easily obtained.

Theorem 2.9. *(First-order necessary conditions) Suppose U is an open set in $\mathbf{E}$ and that $\bar{x}$ is a local minimizer of a function $f : U \to \mathbf{R}$ over the nonempty closed set $\Omega \subset U$. If f is differentiable at $\bar{x}$, then $\langle \nabla f(\bar{x}), v\rangle \geq 0$ for all $v \in T(\bar{x} \,|\, \Omega)$.*

Proof. Let $v \in T(\bar{x} \,|\, S)$. With no loss in generality, we may assume that $v \neq 0$. Then, by Exercise 1.19, there is a $t > 0$ and a sequence $\{x^k\} \subset \Omega \backslash \{\bar{x}\}$

such that $x^k \to \bar{x}$, $\left\|x^k - \bar{x}\right\|^{-1}(x^k - \bar{x}) \to u$, and $v = tu$. Since $\bar{x}$ is a local minimizer of f on Ω, we may assume that $f(\bar{x}) \le f(x^k)$ for all k. Then, for all k,

$$f(\bar{x}) \le f(\bar{x}) + \left\langle \nabla f(\bar{x}),\, x^k - \bar{x} \right\rangle + o\left(\left\|x^k - \bar{x}\right\|\right),$$

and so

$$0 \le \left\langle \nabla f(\bar{x}),\, x^k - \bar{x} \right\rangle + o\left(\left\|x^k - \bar{x}\right\|\right) \quad \forall\, k.$$

Dividing through by $\left\|x^k - \bar{x}\right\|$ and taking the limit in k yields the result. $\square$

Notice that the first-order necessary condition above works for arbitrary nonempty closed sets Ω. However, to obtain a second-order conditions, we assume that Ω is convex, and make use of the tangent cone characterizations in Exercises 1.20 and 1.21.

Theorem 2.10. *(Second-order conditions)*
Consider a C^2-smooth function $f : U \to \mathbf{R}$, where $U \subset \mathbf{E}$ is open. Fix a point $\bar{x} \in \Omega \subset U$, where Ω is a nonempty close convex set. Then the following are true.

1. *(Necessary conditions) Assume that Ω is a convex polyhedron. If $\bar{x}$ is a local minimizer of f on Ω, then*

$$\langle \nabla f(\bar{x}),\, v\rangle \ge 0 \qquad \forall\; v \in T\,(\bar{x}\,|\,\Omega)$$

and

$$v^T \nabla^2 f(x) v \ge 0 \qquad \forall\; v \in T\,(\bar{x}\,|\,\Omega) \cap \operatorname{span}\,(\nabla f(\bar{x}))^{\perp}.$$

2. *(Sufficient conditions) If the relations*

$$\langle \nabla f(\bar{x}),\, v\rangle \ge 0 \qquad \forall\; v \in T\,(\bar{x}\,|\,\Omega)$$

and

$$v^T \nabla^2 f(\bar{x}) v > 0 \qquad \forall\; v \in \left(T\,(\bar{x}\,|\,\Omega) \cap \operatorname{span}\,(\nabla f(\bar{x}))^{\perp}\right) \setminus \{0\}.$$

hold, then there is an $\epsilon > 0$ and $\beta > 0$ such that

$$f(x) \ge f(\bar{x}) + \frac{\beta}{2}\|x - \bar{x}\|^2 \qquad \forall\; x \in B_\epsilon(\bar{x}) \cap \Omega. \tag{2.1}$$

Proof. Theorem 2.9 tells us that $\langle \nabla f(\bar{x}),\, v\rangle \ge 0$ for all $v \in T\,(\bar{x}\,|\,\Omega)$. Next let $v \in T\,(\bar{x}\,|\,\Omega) \cap \operatorname{span}\,(\nabla f(\bar{x}))^{\perp}$. With no loss in generality, we may assume that $\|v\| = 1$. By Exercise 1.21 there exists $\bar{t} > 0$ such that $\bar{x} + tv \in \Omega$ for all $t \in (0, \bar{t})$. Since $\bar{x}$ is a local solution, we may take $\bar{t}$ so small that that $f(\bar{x}) \le f(\bar{x} + tv)$ for all $t \in (0, \bar{t})$. Then, for all $t \in (0, \bar{t})$,

$$f(\bar{x}) \le f(\bar{x}) + t\langle \nabla f(\bar{x}),\, v\rangle + \frac{t^2}{2}\langle \nabla f(\bar{x})v,\, v\rangle + o\left(t^2\right)$$

and so

$$0 \le \frac{1}{2}\langle \nabla f(\bar{x})v,\, v\rangle + \frac{o(t^2)}{t^2} \quad \forall\, t \in (0, \bar{t}).$$

Letting $t \to 0$ yields the second-order necessary condition.

To see the second-order sufficient condition, we suppose that the result is false so that there exists a sequences $\beta_k \downarrow 0$ and $x^k \to \bar{x}$ such that

$$f(x^k) < f(\bar{x}) + \frac{\beta_k}{2}\left\|x^k - \bar{x}\right\|^2 \qquad \forall\, k,$$

or equivalently,

$$\begin{aligned} f(\bar{x}) + \left\langle \nabla f(\bar{x}),\, x^k - \bar{x}\right\rangle + \frac{1}{2}\left\langle \nabla^2 f(\bar{x})(x^k - \bar{x}),\, (x^k - \bar{x})\right\rangle + o(\left\|x^k - \bar{x}\right\|^2) \\ \le f(\bar{x}) + \frac{\beta_k}{2}\left\|x^k - \bar{x}\right\|^2 \qquad \forall\, k. \end{aligned} \tag{2.2}$$

With no loss in generality, we may assume that there is a unit vector u such that $\left\|x^k - \bar{x}\right\|^{-1}(x^k - \bar{x}) \to u \in T\,(\bar{x}\,|\,\Omega)$. Dividing by $\left\|x^k - \bar{x}\right\|$ and letting $k \uparrow \infty$ yields $0 \le \langle \nabla f(\bar{x}),\, u\rangle \le 0$ so that $u \in (T\,(\bar{x}\,|\,\Omega) \cap \text{span}\,(\nabla f(\bar{x}))^{\perp}) \setminus \{0\}$. Further note that by Exercise 1.20, $(x^k - \bar{x}) \in T\,(\bar{x}\,|\,\Omega)$ for all k so that $\left\langle \nabla f(\bar{x}),\, x^k - \bar{x}\right\rangle \ge 0$ for all k. Hence, (2.2) tells us that

$$\frac{1}{2}\left\langle \nabla^2 f(\bar{x})(x^k - \bar{x}),\, (x^k - \bar{x})\right\rangle + o(\left\|x^k - \bar{x}\right\|^2) \le \frac{\beta_k}{2}\left\|x^k - \bar{x}\right\|^2 \qquad \forall\, k.$$

Dividing by $\left\|x^k - \bar{x}\right\|^2$ and taking the limit as $k \uparrow \infty$ gives the contradiction $\left\langle \nabla^2 f(\bar{x})u,\, u\right\rangle \le 0$ which proves the result. □

In later sections we will improve on the second-order conditions in this theorem by delving deeper into the curvature properties of the set Ω. These later results will not only allow us to remove the convexity hypotheses, but will also be stronger even in the convex case. As a first illustration of the limitations of Theorem 2.10, the following example shows that the polyhedrality hypothesis used in the necessary condition cannot be weakened.

Example 2.11. Consider the problem

$$\begin{array}{ll} \min & \frac{1}{2}(x_2 - x_1^2) \\ \text{subject to} & 0 \le x_2,\ x_1^3 \le x_2^2. \end{array}$$

Observe that the constraint region in this problem can be written as $\Omega := \{(x_1, x_2) : |x_1|^{\frac{3}{2}} \le x_2\}$, therefore

$$\begin{aligned} f(x) &= \frac{1}{2}(x_2 - x_1^2) \\ &\ge \frac{1}{2}(|x_1|^{\frac{3}{2}} - |x_1|^2) \\ &= \frac{1}{2}|x_1|^{\frac{3}{2}}(1 - |x_1|^{\frac{1}{2}}) > 0 \end{aligned}$$

whenever $0 < |x_1| \leq 1$. Consequently, the origin is a strict local solution for this problem. Nonetheless,

$$T\,(0\,|\,\Omega) \cap [\nabla f(0)]^{\perp} = \{(\delta, 0)\ :\ \delta \in \mathbf{R}\},$$

while

$$\nabla^2 f(0) = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}.$$

That is, even though the origin is a strict local solution, the Hessian of f is negative definite on $T\,(0\,|\,\Omega) \cap [\nabla f(0)]^{\perp}$.

The second-order sufficiency condition in Theorem 2.10 is also lacking since, as is shown in the next example, the quadratic growth condition (2.1) can be satisfied even if the hessian is not positive definite on the set $(T\,(\bar{x}\,|\,\Omega) \cap \operatorname{span}(\nabla f(\bar{x}))^{\perp}) \setminus \{0\}$.

Example 2.12. Consider the problem

$$\begin{array}{ll} \text{min} & x_2 \\ \text{subject to} & x_1^2 \leq x_2. \end{array}$$

Clearly, $\bar{x} = 0$ is the unique global solution to this convex program. Moreover,

$$\begin{aligned} f(\bar{x}) + \frac{1}{2}\,\|x - \bar{x}\|^2 &= \frac{1}{2}(x_1^2 + x_2^2) \\ &\leq \frac{1}{2}(x_2 + x_2^2) \\ &\leq x_2 = f_0(x) \end{aligned}$$

for all x in the constraint region Ω with $\|x - \bar{x}\| \leq 1$. However, $\nabla^2 f(\bar{x}) = 0$.

2.3 Convexity, a first look

Finding a global minimizer of a general smooth function $f\colon \mathbf{E} \to \mathbf{R}$ 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 2.13 (Convexity)**.** A function $f\colon U \to (-\infty, +\infty]$ 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 point in the graph of the function lies above the graph. This is the content of the following exercise.

Exercise 2.14. Show that a function $f\colon U \to (-\infty, +\infty]$ is convex if and only if the *epigraph*

$$\operatorname{epi} f := \{(x, r) \in U \times \mathbf{R} : f(x) \leq r\}$$

is a convex subset of $\mathbf{E} \times \mathbf{R}$.

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

Exercise 2.15. Consider an arbitrary set T and a family of convex functions $f_t\colon U \to (-\infty, +\infty]$ 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 2.16 (Differential characterizations of convexity)**.** *The following are equivalent for a C^1-smooth function $f\colon U \to \mathbf{R}$.*

(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 (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 (b).

Suppose now that (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 (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:

$$\begin{aligned} \varphi'(t) - \varphi'(s) &= \langle \nabla f(x_t), y - x \rangle - \langle \nabla f(x_s), y - x \rangle \\ &= \frac{1}{t - s} \langle \nabla f(x_t) - \nabla f(x_s), x_t - x_s \rangle \geq 0. \end{aligned}$$

Thus the derivative φ' 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 (Excercise 2.15), we deduce that f is convex, establishing (a).

Suppose now that f is C^2-smooth. Then for any fixed $x \in U$ and $h \in \mathbf{E}$, 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 $h \in \mathbf{E}$. Hence (d) holds. Conversely, suppose (d) holds. Then Corollary 1.13 immediately implies for all $x, y \in \mathbf{E}$ 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. □

Exercise 2.17. Show that the functions f and F in Exercise 1.10 are convex.

Exercise 2.18. Consider a C^1-smooth function $f\colon \mathbf{R}^n \to \mathbf{R}$. Prove that each condition below holding for all points $x, y \in \mathbf{R}^n$ is equivalent to f being β-smooth and convex.

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

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

Corollary 2.19 (Minimizers of convex functions)**.** *For any C^1-smooth convex function $f\colon U \to \mathbf{R}$ 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 2.16. □

Exercise 2.20. Consider a C^1-smooth convex function $f\colon \mathbf{E} \to \mathbf{R}$. Fix a linear subspace $\mathcal{L} \subset \mathbf{E}$ and a point $x_0 \in \mathbf{E}$. Show that $x \in \mathcal{L}$ minimizes the restriction $f_{\mathcal{L}}\colon \mathcal{L} \to \mathbf{R}$ if and only if the gradient $\nabla f(x)$ is orthogonal to $\mathcal{L}$.

Strengthening the gradient inequality in Theorem 2.16 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 2.21 (Strong convexity)**.** We say that a C^1-smooth function $f\colon U \to \mathbf{R}$ is *α-strongly convex* (with $\alpha \ge 0$) if the inequality

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y - x\|^2 \quad \text{holds for all } x, y \in U.$$

Figure 2.3 illustrates geometrically a β-smooth and α-convex function.

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

$$f(y) \ge f(x) + \frac{\alpha}{2}\|y - x\|^2.$$

Exercise 2.22. Show that a C^1-smooth function $f\colon U \to \mathbf{R}$ is α-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 2.16 for strongly convex functions.

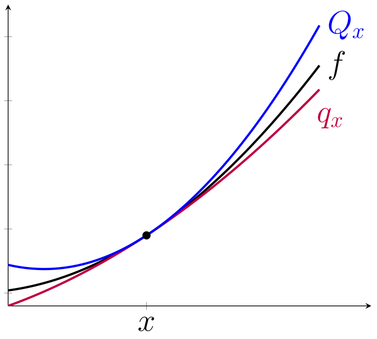

Figure 2.3: Illustration of a β-smooth and α-strongly convex function f, where $Q_x(y) := f(x) + \langle \nabla f(x), y - x\rangle + \frac{\beta}{2}\|y-x\|^2$ is an upper models based at x and $q_x(y) := f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y-x\|^2$ is a lower model based at x. The fraction $Q := \beta/\alpha$ is often called the *condition number* of f.

Theorem 2.23 (Characterization of strong convexity)**.** *The following properties are equivalent for any C^1-smooth function $f\colon U \to \mathbf{R}$ and any constant $\alpha \geq 0$.*

(a) f is α-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 Excercise 2.22, property (a) holds if and only if $f - \frac{\alpha}{2}\|\cdot\|^2$ is convex, which by Theorem 2.16, is equivalent to (b). Suppose now that f is C^2-smooth. Theorem 2.16 then shows that $f - \frac{\alpha}{2}\|\cdot\|^2$ is convex if and only if (c) holds. □

2.4 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
$$a_k \leq \frac{c}{k^q} \qquad \text{for all } k.$$

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 $k \geq (\frac{c}{\varepsilon})^{1/q}$. 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

$$a_k \leq c \cdot (1-q)^k \qquad \text{for all } k.$$

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

$$c \cdot (1-q)^k \leq \varepsilon \quad \Longleftrightarrow \quad k \geq \frac{-1}{\ln(1-q)} \ln\left(\frac{c}{\varepsilon}\right).$$

Taking into account the inequality $\ln(1 - q) \leq -q$, we deduce that the inequality $a_k \leq \varepsilon$ holds for every $k \geq \frac{1}{q} \ln(\frac{c}{\varepsilon})$. 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

$$a_{k+1} \leq c \cdot a_k^2 \qquad \text{for all } k.$$

Observe then unrolling the recurrence yields

$$a_{k+1} \leq \frac{1}{c}(ca_0)^{2^{k+1}}.$$

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 $k \geq \log_2 \ln(\frac{1}{c\varepsilon}) - \log_2(-\ln(ca_0))$. The dependence on c is negligible.

2.5 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.

2.5.1 Majorization view of gradient descent

Consider the optimization problem

$$\min_{x \in \mathbf{E}} f(x),$$

where f is a β-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 $x \in \mathbf{E}$. 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:

$$m_k(x) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{\beta}{2}\|x - x_k\|^2. \tag{2.3}$$

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 $x \in \mathbf{E}$, 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

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \langle \nabla f(x_k), \beta^{-1}\nabla f(x_k)\rangle + \frac{\beta}{2}\|\beta^{-1}\nabla f(x_k)\|^2 \\ &= f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|^2. \end{aligned} \tag{2.4}$$

Rearranging, and summing over the iterates, we deduce

$$\sum_{i=1}^{k} \|\nabla f(x_i)\|^2 \leq 2\beta\big(f(x_1) - f(x_{k+1})\big).$$

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 points 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:

$$\min_{i=1,\ldots,k} \|\nabla f(x_i)\|^2 \leq \frac{1}{k}\sum_{i=1}^{k} \|\nabla f(x_i)\|^2 \leq \frac{2\beta\big(f(x_1) - f^*\big)}{k}.$$

We have thus established the following result.

Theorem 2.24 (Gradient descent)**.** *Consider a β-smooth function $f \colon \mathbf{E} \to \mathbf{R}$. Then the iterates generated by the gradient descent method satisfy*

$$\min_{i=1,\ldots,k} \|\nabla f(x_i)\|^2 \leq \frac{2\beta\big(f(x_1) - f^*\big)}{k}.$$

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

Theorem 2.25 (Gradient descent and convexity)**.** *Suppose that $f \colon \mathbf{E} \to \mathbf{R}$ is convex and β-smooth. Then the iterates generated by the gradient descent method satisfy*

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

and

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

Proof. Since x_{k+1} is the minimizer of the β-strongly convex quadratic $m_k(\cdot)$ in (2.3), 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

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

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

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

as claimed. Next, summing the basic descent inequality

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

for $k = m, \ldots, 2m-1$, we obtain

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

Taking into account the inequality

$$\frac{1}{2\beta}\sum_{i=m}^{2m-1}\|\nabla f(x_i)\|^2 \geq \frac{m}{2\beta}\cdot \min_{i=1,\ldots 2m}\|\nabla f(x_i)\|^2,$$

we deduce

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

as claimed. □

Thus when the gradient method is applied to a potentially nonconvex β-smooth function, the gradients $\|\nabla f(x_k)\|$ decay as $\frac{\beta\|x_1-x^*\|}{\sqrt{k}}$, while for convex functions the estimate significantly improves to $\frac{\beta\|x_1-x^*\|}{k}$.

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

Theorem 2.26 (Gradient descent and strong convexity)**.**
Suppose that $f\colon \mathbf{E} \to \mathbf{R}$ is α-strongly convex and β-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

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \beta^{-1}\nabla f(x_k)\|^2 \\
&= \|x_k - x^*\|^2 + \frac{2}{\beta}\langle \nabla f(x_k), x^* - x_k\rangle + \frac{1}{\beta^2}\|\nabla f(x_k)\|^2 \\
&\leq \|x_k - x^*\|^2 + \frac{2}{\beta}\Big(f^* - f(x_k) - \frac{\alpha}{2}\|x_k - x^*\|^2\Big) + \frac{1}{\beta^2}\|\nabla f(x_k)\|^2 \\
&= \left(1 - \frac{\alpha}{\beta}\right)\|x_k - x^*\|^2 + \frac{2}{\beta}\Big(f^* - f(x_k) + \frac{1}{2\beta}\|\nabla f(x_k)\|^2\Big).
\end{aligned}$$

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. □

Thus for gradient descent, the quantities $\|x_k - x^*\|^2$ converge to zero at a linear rate $\frac{Q-1}{Q+1} = 1 - \frac{2}{Q+1}$. 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 β-smoothness and strong convexity. In particular, in light of Section 2.4, 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.

Example 2.27 (Linear Regression)**.** Consider a linear regression problem as in Example 2.1:

$$\min_x \frac{1}{2}\|Ax - b\|^2. \tag{2.5}$$

This problem has a unique solution only if A is injective, and in this case, the solution is

$$\bar{x} = (A^T A)^{-1} A^T b.$$

When the solution is not unique, for example if $m < n$, it is common to *regularize* the problem by adding a strongly-convex quadratic perturbation:

$$\min_x \frac{1}{2}\|Ax - b\|^2 + \frac{\eta}{2}\|x\|^2. \tag{2.6}$$

This strategy is called *ridge regression* (Example 2.2). This problem always has a closed form solution, regardless of properties of A:

$$\bar{x}_\eta = (A^T A + \eta I)^{-1} A^T b,$$

In this example, we apply steepest descent with constant step length to the ridge regression problem. Despite the availability of closed form solutions, iterative approaches are essential for large-scale applications, where forming $A^T A$ is not feasible. Indeed, for many real-world applications, practitioners may have access to A and A^T only through the action of these operators on vectors.

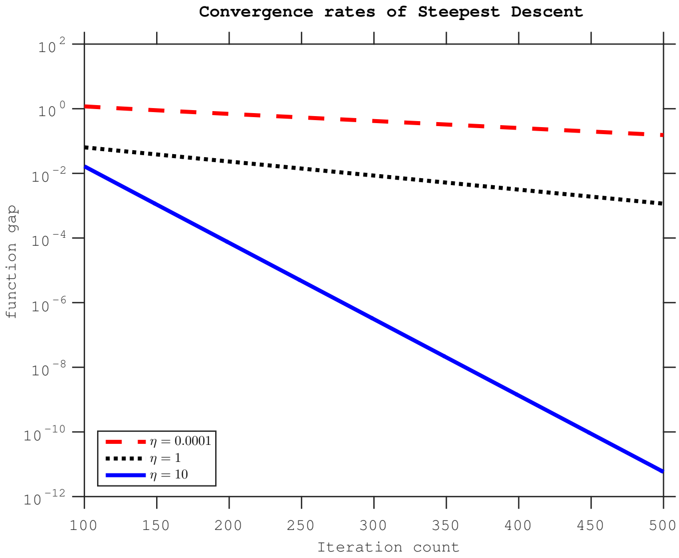

Figure 2.4: Convergence rate of steepest descent for Ridge Regression. In this example, the condition number of A is set to 10, and we show convergence of functional iterates $f(x_k) - f(x^*)$ for several values of η.

Since the eigenvalues of $A^TA + \eta I$ are simply eigenvalues of A^TA shifted by η, it is clear that the Lipschitz constant of the gradient of (2.6) is $\beta = \lambda_{\max}(A^TA) + \eta$. Each iteration of steepest descent is therefore given by

$$x_{k+1} = x_k - \frac{1}{\lambda_{\max}(A^TA) + \eta}\left(A^T(Ax_k - b) + \eta x_k\right). \tag{2.7}$$

The strong convexity constant α of the objective functions is

$$\alpha = \lambda_{\min}(A^TA) + \eta.$$

Therefore, Theorem 3.15 guarantees

$$\|x_k - x^*\|^2 \leq \left(1 - \frac{\eta + \lambda_{\min}(A^TA)}{\eta + \lambda_{\max}(A^TA)}\right)^k \|x_0 - x^*\|^2.$$

The convergence rates of the steepest descent algorithm (for both iterates and function values) for ridge regression is shown in Figure 2.4. The linear rate is evident.

2.5.2 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 $f\colon \mathbf{E} \to \mathbf{R}$. 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 $G\colon \mathbf{E} \to \mathbf{E}$ 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 2.28 (Progress of Newton's method)**.** *Consider a C^1-smooth map $G\colon \mathbf{E} \to \mathbf{E}$ with the Jacobian ∇G that is β-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 $\mathbf{E}$ with $\mathbf{R}^m$ for some integer m. Then appealing to (1.15), we deduce

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

Thus

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

as claimed. □

To see the significance of Theorem 2.28, consider a β-smooth map $G\colon \mathbf{E} \to \mathbf{E}$. 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 $\|\nabla G(x)^{-1}\| \leq R$ holds for all $x \in B_\epsilon(x^*)$. Then provided that Newton's method is initialized at a point x_0 satisfying $\|x_0 - x^*\| < \frac{2\epsilon}{\beta R}$, 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.

2.6 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 2.29 (Linearly-expanding first-order method)**.** An algorithm is called a *linearly-expanding first-order method* if when applied to any β-smooth function f on $\mathbf{R}^n$ it generates an iterate sequence $\{x_k\}$ satisfying

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

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

Theorem 2.30 (Lower-complexity bound for smooth convex optimization)**.** *For any k, with $1 \leq k \leq (n-1)/2$, and any $x_0 \in \mathbf{R}^n$ there exists a convex β-smooth function $f\colon \mathbf{R}^n \to \mathbf{R}$ so that iterates generated by any linearly-expanding first-order method started at x_0 satisfy*

$$f(x_k) - f^* \geq \frac{3\beta\|x_0 - x^*\|^2}{32(k+1)^2}, \tag{2.8}$$

$$\|x_k - x^*\|^2 \geq \tfrac{1}{8}\|x_0 - x^*\|^2, \tag{2.9}$$

where x^* is any minimizer of f.

For simplicity, we will only prove the bound on functional values (2.8). 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 $\mathbf{R}^k \times \{0\}^{n-k}$. 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

$$f_k(z_1, z_2, \ldots, z_n) = \tfrac{\beta}{4}\Big(\tfrac{1}{2}(z_1^2 + \sum_{i=1}^{k-1}(z_i - z_{i+1})^2 + z_k^2) - z_1\Big)$$

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

$$\langle \nabla f(x)v, v\rangle = \tfrac{\beta}{4}\Big((v_1^2 + \sum_{i=1}^{k-1}(v_i - v_{i+1})^2 + v_k^2)\Big)$$

and therefore

$$0 \le \langle \nabla f(x)v, v\rangle \le \tfrac{\beta}{4}\Big((v_1^2 + \sum_{i=1}^{k-1} 2(v_i^2 + v_{i+1}^2) + v_k^2)\Big) \le \beta\|v\|^2.$$

Exercise 2.31. 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

$$f_k^* = \tfrac{\beta}{8}\left(-1 + \tfrac{1}{k+1}\right).$$

2. Taking into account the standard inequalities,

$$\sum_{i=1}^{k} i = \frac{k(k+1)}{2} \qquad \text{and} \qquad \sum_{i=1}^{k} i^2 \le \frac{(k+1)^3}{3},$$

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

3. Fix indices $1 < i < j < n$ and a point $x \in \mathbf{R}^i \times \{0\}^{n-i}$. Show that equality $f_i(x) = f_j(x)$ holds and that the gradient $\nabla f_k(x)$ lies in $\mathbf{R}^{i+1} \times \{0\}^{n-(i+1)}$.

Proving Theorem 2.30 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 2.31 (part 3), the iterate x_k lies in $\mathbf{R}^k \times \{0\}^{n-k}$. 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 2.31, we deduce

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

This proves the result.

The complexity bounds in Theorem 2.30 do not depend on strong convexity constants. When the target function class consists of β-smooth strongly convex functions, the analogous complexity bounds become

$$f(x_k) - f^* \geq \left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{2k} \|x_0 - x^*\|^2, \tag{2.10}$$

$$\|x_k - x^*\|^2 \geq \frac{\alpha}{2}\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{2k} \|x_0 - x^*\|^2, \tag{2.11}$$

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 2.30, 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 β-smooth convex functions f on $\mathbf{E}$ and suppose we wish to find a point x satisfying $f(x) - f^* \leq \varepsilon$. By Theorem 2.25, gradient descent will require at most $k \leq \mathcal{O}\left(\frac{\beta\|x_0 - x^*\|^2}{\varepsilon}\right)$ iterations. On the other hand, the lower-complexity bound (2.8) shows that no first-order method can be guaranteed to achieve the goal within $k \leq \mathcal{O}\left(\sqrt{\frac{\beta\|x_0-x^*\|^2}{\varepsilon}}\right)$ iterations. Clearly there is a large gap. Note that the bound (2.9) 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 α-strongly convex. Theorem 2.25 shows that gradient descent will find a point x satisfying $\|x - x^*\|^2 \leq \varepsilon$ after at most $k \leq \mathcal{O}\left(\frac{\beta}{\alpha}\ln\left(\frac{\|x_0-x^*\|^2}{\varepsilon}\right)\right)$ iterations. Looking at the corresponding lower-complexity bound (2.11), we see that no first-order method can be guaranteed to find a point x with $\|x - x^*\|^2 \leq \varepsilon$ after at most $k \leq \mathcal{O}\left(\sqrt{\frac{\beta}{\alpha}}\ln\left(\frac{\alpha\|x_0-x^*\|^2}{\varepsilon}\right)\right)$ 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 their exist algorithms that match the bounds. Such algorithms are said to be "optimal".

2.7 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 (2.10) for smooth strongly convex minimization.

Setting the groundwork, consider the minimization problem:

$$\min_x f(x) := \tfrac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle,$$

where $b \in \mathbf{R}^n$ is a vector and $A \in \mathbf{S}^n$ 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 $\langle v, w\rangle_A := \langle Av, w\rangle$ and the norm $\|v\|_A := \sqrt{\langle Av, v\rangle}$ (Exercise 1.2).

Exercise 2.32. Verify for any point $x \in \mathbf{R}^n$ 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 $\mathbf{R}^n$. Consider now the following iterative scheme: given a point $x_1 \in \mathbf{R}^n$ define

$$\begin{cases} t_k = \operatorname{argmin}_t \, f(x_k + tv_k) \\ x_{k+1} = x_k + t_k v_k \end{cases}$$

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 2.33. Prove the formula $t_k = \frac{\langle r_k, v_k \rangle}{\|v_k\|_A^2}$.

Observe that the residuals r_k satisfy the equation

$$r_{k+1} = r_k - t_k A v_k. \tag{2.12}$$

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

Theorem 2.34 (Expanding subspaces)**.** *Fix an arbitrary initial point $x_1 \in \mathbf{R}^n$. Then the equation*

$$\langle r_{k+1}, v_i \rangle = 0 \qquad \textit{holds for all } i = 1, \ldots, k \tag{2.13}$$

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

Proof. We prove the theorem inductively. Assume that equation (2.13) holds with k replaced by $k-1$. Taking into account the recursion (2.12) and Exercise 2.33, we obtain

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

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 (2.13). Now clearly x_{k+1} lies in $x_1 + \operatorname{span}\{v_1, \ldots v_k\}$. On the other hand, equation (2.13) shows that the gradient $\nabla f(x_{k+1}) = -r_{k+1}$ is orthogonal to $\operatorname{span}\{v_1, \ldots v_k\}$. It follows immediately that x_{k+1} minimizes f on $x_1 + \operatorname{span}\{v_1, \ldots v_k\}$, as claimed. □

Corollary 2.35. *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 β_k. Observe that β_k is uniquely defined by forcing v_k to be A-orthogonal with v_{k-1}:

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

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

$$\langle v_k, v_i \rangle_A = \langle r_k, v_i \rangle_A + \beta_k \langle v_{k-1}, v_i \rangle_A = \langle r_k, A v_i \rangle = t_i^{-1} \langle r_k, r_i - r_{i+1} \rangle.$$

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 $r_i, r_{i+1} \in \operatorname{span}\{v_i, v_{i-1}, v_{i+1}\}$. Appealing to Theorem 2.34 and the inequality above, we thus conclude that the set $\{v_1, \ldots, v_k\}$ is indeed A-orthogonal. The scheme just outlined is called the *conjugate gradient method.*

Algorithm 1: Conjugate gradient (CG)

```
Given $x_0$;
Set $r_0 \leftarrow b - Ax_0$, $v_0 \leftarrow r_0$, $k \leftarrow 0$.
while $r_k \neq 0$ do
    $t_k \leftarrow \frac{\langle r_k, v_k \rangle}{\|v_k\|_A^2}$
    $x_{k+1} \leftarrow x_k + t_k v_k$
    $r_{k+1} \leftarrow b - Ax_{k+1}$
    $\beta_{k+1} \leftarrow -\frac{\langle r_{k+1}, v_k \rangle_A}{\|v_k\|_A^2}$
    $v_{k+1} \leftarrow r_{k+1} + \beta_{k+1} v_k$
    $k \leftarrow k + 1$
end
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

$$\mathcal{K}_k(y) = \operatorname{span}\{y, Ay, A^2y, \ldots, A^ky\}.$$

Theorem 2.36. *Consider the iterates x_k generated by the conjugate gradient method. Supposing $x_k \neq x^*$, we have*

$$\langle r_k, r_i \rangle = 0 \qquad \textit{for all} \quad i = 0, 1, \ldots, k-1, \tag{2.14}$$

$$\langle v_k, v_i \rangle_A = 0 \qquad \textit{for all} \quad i = 0, 1, \ldots, k-1, \tag{2.15}$$

and

$$\operatorname{span}\{r_0, r_1, \ldots, r_k\} = \operatorname{span}\{v_0, v_1, \ldots, v_k\} = \mathcal{K}_k(r_0). \tag{2.16}$$

Proof. We have already proved equation (2.15), as this was the motivation for the conjugate gradient method. Equation (2.14) follows by observing the inclusion $r_i \in \operatorname{span}\{v_i, v_{i-1}\}$ and appealing to Theorem 2.34. We prove the final claim (2.16) 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

$$\operatorname{span}\{r_0, r_1, \ldots, r_{k+1}\} \subseteq \operatorname{span}\{v_0, v_1, \ldots, v_{k+1}\} \tag{2.17}$$

holds since r_i lie in span $\{v_i, v_{i-1}\}$. Taking into account the induction assumption, we deduce $v_{k+1} \in$ span $\{r_{k+1}, v_k\} \subseteq$ span $\{r_0, r_1, \ldots, r_{k+1}\}$. Hence equality holds in (2.17).

Next note by the induction hypothesis the inclusion

$$r_{k+1} = r_k - t_k A v_k \in \mathcal{K}_k(r_0) - \mathcal{K}_{k+1}(r_0) \subseteq \mathcal{K}_{k+1}(r_0).$$

Conversely, by the induction hypothesis, we have

$$A^{k+1} r_0 = A(A^k r_0) \subseteq \text{span}\,\{Av_0, \ldots, Av_k\} \subseteq \text{span}\,\{r_0, \ldots, r_{k+1}\}.$$

This completes the proof. □

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

$$\begin{aligned} 2(f(x_{k+1}) - f^*) &= \inf_{x \in x_0 + \mathcal{K}_k(r_0)} 2(f(x) - f^*) \\ &= \inf_{x \in x_0 + \mathcal{K}_k(r_0)} \|x - x^*\|_A^2 = \min_{p \in \mathcal{P}_k} \|x_0 - p(A)r_0 - x^*\|_A^2 \end{aligned}$$

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

$$\begin{aligned} 2(f(x_{k+1}) - f^*) &= \min_{p \in \mathcal{P}_k} \|(x_0 - x^*) + p(A)A(x_0 - x^*)\|_A^2 \\ &= \min_{p \in \mathcal{P}_k} \|(I + p(\Lambda)\Lambda)z\|_\Lambda^2 \\ &= \min_{p \in \mathcal{P}_k} \sum_{i=1}^n \lambda_i (1 + p(\lambda_i)\lambda_i)^2 z_i^2 \\ &\leq \left(\sum_{i=1}^n \lambda_i z_i^2\right) \min_{p \in \mathcal{P}_k} \max_{i=1,\ldots,n} (1 + p(\lambda_i)\lambda_i)^2. \end{aligned}$$

Observe now the inequality $\sum_{i=1}^n \lambda_i z_i^2 = \|z\|_\Lambda^2 = \|x_0 - x^*\|_A^2$. 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

$$f(x_{k+1}) - f^* \leq \frac{1}{2}\|x_0 - x^*\|_A^2 \cdot \max_{i=1,\ldots,n} q(\lambda_i)^2 \tag{2.18}$$

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.

Theorem 2.37 (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 $q(\lambda) := \frac{(-1)^m}{\gamma_1 \cdots \gamma_m}(\lambda - \gamma_1) \cdots (\lambda - \gamma_m)$. Observe $q(0) = 1$. Moreover, clearly equality $0 = q(\gamma_i)$ holds for all indices i. Inequality (2.18) 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

$$\begin{aligned} T_0 &= 1, \\ T_1(t) &= t, \\ T_{k+1}(t) &= 2tT_k(t) - T_{k-1}(t). \end{aligned}$$

Before proceeding, we explain why Chebyshev polynomials appear naturally. Observe that inequality (2.18) implies

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

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 2.5 for an illustration.

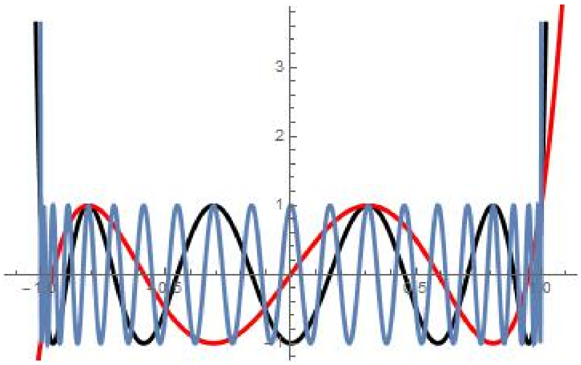

Figure 2.5: T_5, T_{10}, T_{40} are shown in red, black, and violet, respectively, on the interval $[-1, 1]$.

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 $|t| \geq 1$.

Theorem 2.38 (Linear convergence rate)**.** *Letting $Q = \lambda_1/\lambda_n$ be the condition number of A, the inequalities*

$$f(x_k) - f^* \leq 2\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{2k} \|x_0 - x^*\|_A^2 \qquad \textit{hold for all } k.$$

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 (2.18), and properties (i) and (ii), we deduce

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

The result follows. □

Thus linear convergence guarantees of the conjugate gradient method match those given by the lower complexity bounds (2.10).

2.8 Optimal methods for smooth convex minimization

In this section, we discuss *optimal first-order methods* for minimizing β-smooth functions. These are the methods whose convergence guarantees match the lower-complexity bounds (2.8) and (2.10).

2.8.1 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

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

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* $\frac{t_k\beta_k}{t_{k-1}}(x_k - x_{k-1})$, indicating the direction from which one came. Let us emulate this idea on a β-smooth convex function $f\colon \mathbf{E} \to \mathbf{R}$. Consider the following recurrence

$$\left\{ \begin{aligned} y_k &= x_k + \gamma_k(x_k - x_{k-1}) \\ x_{k+1} &= y_k - \frac{1}{\beta}\nabla f(y_k) \end{aligned} \right\},$$

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 γ_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 2.25). Since $y \mapsto l(y; y_k) + \frac{\beta}{2}\|y - y_k\|^2$ is a strongly convex quadratic, we deduce

$$\begin{aligned} f(x_{k+1}) &\leq l(x_{k+1}; y_k) + \frac{\beta}{2}\|x_{k+1} - y_k\|^2 \\ &\leq l(y; y_k) + \frac{\beta}{2}(\|y - y_k\|^2 - \|y - x_{k+1}\|), \end{aligned}$$

for all points $y \in \mathbf{E}$. 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

$$\begin{aligned} f(x_{k+1}) &\leq l(a_k x^* + (1 - a_k)x_k; y_k) \\ &\quad + \frac{\beta}{2}\left(\|a_k x^* + (1 - a_k)x_k - y_k\|^2 - \|a_k x^* + (1 - a_k)x_k - x_{k+1}\|^2\right) \\ &= a_k l(x^*; y_k) + (1 - a_k)l(x_k; y_k) \\ &\quad + \frac{\beta a_k^2}{2}\left(\|x^* - [x_k - a_k^{-1}(x_k - y_k)]\|^2 - \|x^* - [x_k - a_k^{-1}(x_k - x_{k+1})]\|^2\right). \end{aligned}$$

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

$$\begin{aligned} \frac{1}{a_k^2}(f(x_{k+1}) - f^*) \leq \frac{1 - a_k}{a_k^2}&(f(x_k) - f^*) \\ + \frac{\beta}{2}\Big(&\|x^* - [x_k - a_k^{-1}(x_k - y_k)]\|^2 \\ &- \|x^* - [x_k - a_k^{-1}(x_k - x_{k+1})]\|^2\Big). \end{aligned} \tag{2.19}$$

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

$$z_{k+1} := x_k - a_k^{-1}(x_k - x_{k+1}). \tag{2.20}$$

Let us try to choose γ_k and a_k to ensure the equality $z_k = x_k - a_k^{-1}(x_k - y_k)$. From the definition (2.20) 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 $\gamma_k = a_k(a_{k-1}^{-1} - 1)$. Thus the inequality (2.19) becomes

$$\frac{1}{a_k^2}(f(x_{k+1}) - f^*) + \frac{\beta}{2}\|x^* - z_{k+1}\|^2 \le \frac{1 - a_k}{a_k^2}(f(x_k) - f^*) + \frac{\beta}{2}\|x^* - z_k\|^2. \tag{2.21}$$

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

$$\frac{1 - a_k}{a_k^2} \le \frac{1}{a_{k-1}^2}. \tag{2.22}$$

Then the right-hand-side of (2.21) is upper-bounded by the same term as the left-hand-side with k replaced by $k - 1$. Iterating the recurrence (2.21) yields

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

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

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

Looking back at (2.22), the choices $a_k = \frac{2}{k+2}$ are valid, and will yield the efficiency estimate

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

Thus the scheme is indeed optimal for minimizing β-smooth convex functions, since this estimate matches the lower complexity bound (2.9). A slightly faster rate will occur when choosing $a_k \in (0, 1]$ to satisfy (2.22) with equality, meaning

$$a_{k+1} = \frac{\sqrt{a_k^4 + 4a_k^2} - a_k^2}{2}. \tag{2.23}$$

Exercise 2.39. Suppose $a_0 = 1$ and a_k is given by (2.23) 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 2.40 summarize our findings.

Algorithm 2: Fast gradient method for smooth convex minimization

```
Input: Starting point $x_0 \in \mathbf{E}$.
Set $k = 0$ and $a_0 = a_{-1} = 1$;
for $k = 0, \ldots, K$ do
    Set
        $$y_k = x_k + a_k(a_{k-1}^{-1} - 1)(x_k - x_{k-1})$$
        $$x_{k+1} = y_k - \frac{1}{\beta}\nabla f(y_k) \qquad (2.24)$$
    Choose $a_{k+1} \in (0,1)$ satisfying
        $$\frac{1 - a_{k+1}}{a_{k+1}^2} \leq \frac{1}{a_k^2}. \qquad (2.25)$$
    $k \leftarrow k + 1$.
end
```

Theorem 2.40 (Progress of the fast-gradient method)**.** *Suppose that f is a β-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*

$$f(x_k) - f^* \leq \frac{2\beta\|x^* - x_0\|^2}{(k+1)^2}. \tag{2.26}$$

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 2.25. 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 (2.24).

Theorem 2.41 (Gradient convergence of the fast-gradient method)**.** *Suppose that f is a β-smooth convex function. In Algorithm 2, set $a_k \leq \frac{2}{k+2}$ for all k and replace line* (2.24) *by $x_{k+1} = y_k - \frac{1}{2\beta}\nabla f(y_k)$. Then the iterates* *generated by the algorithm satisfy*

$$f(x_k) - f^* \leq \frac{4\beta\|x^* - x_0\|^2}{(k+1)^2}, \tag{2.27}$$

$$\min_{i=1,\ldots,k} \|\nabla f(y_i)\| \leq \frac{8\sqrt{3}\cdot\beta\|x^* - x_0\|}{\sqrt{k(k+1)(2k+1)}}. \tag{2.28}$$

Proof. The proof is a slight modification of the argument outlined above of Theorem 2.40. Observe

$$\begin{aligned} f(x_{k+1}) &\leq l(x_{k+1}; y_k) + \frac{\beta}{2}\|x_{k+1} - y_k\|^2 \\ &\leq l(x_{k+1}; y_k) + \frac{2\beta}{2}\|x_{k+1} - y_k\|^2 - \frac{1}{8\beta}\left\|\nabla f(y_k)\right\|^2 \\ &\leq l(y; y_k) + \frac{2\beta}{2}(\|y - y_k\|^2 - \|y - x_{k+1}\|^2) - \frac{1}{8\beta}\left\|\nabla f(y_k)\right\|^2. \end{aligned}$$

Continuing as before, we set $z_k = x_k - a_k^{-1}(x_k - y_k)$ and obtain

$$\begin{aligned} &\tfrac{1}{a_k^2}(f(x_{k+1}) - f^*) + \beta\|x^* - z_{k+1}\|^2 \leq \\ &\qquad\leq \tfrac{1-a_k}{a_k^2}(f(x_k) - f^*) + \beta\|x^* - z_k\|^2 - \tfrac{1}{8\beta a_k^2}\left\|\nabla f(y_k)\right\|^2. \end{aligned}$$

Recall $\frac{1-a_k}{a_k^2} \leq \frac{1}{a_{k-1}^2}$, $a_1 = 1$, and $z_0 = x_0$. Iterating the inequality yields

$$\frac{1}{a_k^2}(f(x_{k+1}) - f^*) + \beta\|x^* - z_{k+1}\|^2 \leq \beta\|x^* - x_0\|^2 - \frac{1}{8\beta}\sum_{i=1}^{k}\frac{\|\nabla f(y_i)\|^2}{a_i^2}.$$

Ignoring the second terms on the left and right sides yields (2.27). 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 (2.28). □

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.

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 (2.25). 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 2.8.2.

There is, however, a different approach to get a fast linearly convergent method simply by periodically restarting Algorithm 2. Let $f\colon \mathbf{E} \to \mathbf{R}$ be a β-smooth and α-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_k^i be the k'th iterate generated in epoch i. Theorem 2.40 along with strong convexity yields the guarantee

$$f(x_k^i) - f^* \leq \frac{2\beta\|x^* - x_0^i\|^2}{(k+1)^2} \leq \frac{4\beta}{\alpha(k+1)^2}(f(x_0^i) - f^*). \tag{2.29}$$

Suppose that in each epoch, we run a fast gradient method (Algorithm 2) for N iterations. Given an initial point $x_0 \in \mathbf{E}$, set $x_0^0 := x_0$ and set $x_0^i := x_N^{i-1}$ 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:

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

The total number of iterations to obtain x_0^i is iN_q. We deduce

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

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

$$q^{1/N_q} \leq e^{-2\left\lceil 2e\sqrt{\frac{\beta}{\alpha}}\right\rceil^{-1}} \leq e^{\frac{-2}{1+2e\sqrt{\beta/\alpha}}}.$$

Algorithm 3: Fast gradient method with restarts

```
**Input**: Starting point $x_0 \in \mathbf{E}$.
Set $i, k = 0$, $x_0^0 = x_0$, and $N = \left\lceil 2e\sqrt{\frac{\beta}{\alpha}} \right\rceil$.
**for** $i$ = 0, ..., $K$ **do**
    Let $x_i^N$ be the $N$'th iterate generated by Algorithm 2, initialized
    with $x_0^i$.
    Set $i = i + 1$ and $x_{i+1}^0 = x_N^i$.
**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

$$q^{1/N_q} \leq 1 - \frac{5/3}{1 + 2e\sqrt{\beta/\alpha}}.$$

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 (2.10) for smooth strongly convex minimization.

2.8.2 Fast gradient methods through estimate sequences

In this section, we describe an algebraic technique for designing fast gradient method for minimizing a β-smooth α-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 2.42 (Estimate Sequences)**.** Given real numbers $\lambda_k \in [0, 1]$ and functions $\phi_k \colon \mathbf{E} \to \mathbf{R}$, we say that the sequence $(\lambda_k, \phi_k(x))$ is an *estimate sequence* if $\lambda_k \searrow 0$ and the inequality

$$\phi_k(x) \leq (1 - \lambda_k)f(x) + \lambda_k\phi_0(x) \tag{2.30}$$

holds for all $x \in \mathbf{E}$ 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

$$f(x_k) - f^* \leq \lambda_k(\phi_0^* - f^*). \tag{2.31}$$

Thus the rate at which λ_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 2.43. Suppose that $f\colon \mathbf{E} \to \mathbf{R}$ is C^1-smooth and α-strongly convex. Fix two sequences $\{y_k\}_{k\geq 0} \subset \mathbf{E}$ and $\{t_k\}_{k\geq 0} \subset [0,1]$, and consider an arbitrary function $\phi_0\colon \mathbf{E} \to \mathbf{R}$. Define the sequence (λ_k, ϕ_k) inductively as follows:

$$\left\{\begin{aligned} \lambda_0 &= 1 \\ \lambda_{k+1} &= (1-t_k)\lambda_k \\ \phi_{k+1} &= (1-t_k)\phi_k + t_k Q_{y_k} \end{aligned}\right\}.$$

1. Show that the sequence (λ_k, ϕ_k) satisfies (2.30). (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 (λ_k, ϕ_k) is an estimate sequence for f.

It is clear that if we choose ϕ_0 to be a simple quadratic $\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2}\|x - v_0\|^2$, then all ϕ_k will be simple quadratics as well, in the sense that their Hessians will be multiples of identity.

Exercise 2.44. Let

$$\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2}\|x - v_0\|^2,$$

where $\phi_0^* \in \mathbf{R}$, $\gamma_0 \geq 0$, and $v_0 \in \mathbf{E}$ are chosen arbitrary. Show by induction that the functions ϕ_k in Exercise 2.43 preserve the same form:

$$\phi_k(x) = \phi_k^* + \frac{\gamma_k}{2}\|x - v_k\|^2,$$

where

$$\begin{aligned}
\gamma_{k+1} &= (1-t_k)\gamma_k + t_k\alpha,\\
v_{k+1} &= \frac{1}{\gamma_{k+1}}\left[(1-t_k)\gamma_k v_k + t_k\alpha y_k - t_k\nabla f(y_k)\right],\\
\phi^*_{k+1} &= (1-t_k)\phi^*_k + t_k f(y_k) - \frac{t_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|^2\\
&\quad + \frac{t_k(1-t_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\alpha}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k\rangle\right). \qquad (2.32)
\end{aligned}$$

Now having available an estimate sequence constructed above, let's try to find the sequence $\{x_k\}$ satisfying $f(x_k) \le \phi^*_k$. 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 (2.32) by zero, we deduce

$$\begin{aligned}
\phi^*_{k+1} \ge (1-t_k)f(x_k) + t_k f(y_k) - \frac{t_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|^2\\
+ \frac{t_k(1-t_k)\gamma_k}{\gamma_{k+1}}\langle \nabla f(y_k), v_k - y_k\rangle.
\end{aligned}$$

Combining this with $f(x_k) \ge f(y_k) + \langle \nabla f(y_k), x_k - y_k\rangle$, yields

$$\phi^*_{k+1} \ge \left(f(y_k) - \frac{t_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|^2\right) + (1-t_k)\langle \nabla f(y_k), \frac{t_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k\rangle.$$

The term in parenthesis is reminiscent of a descent condition for a gradient step, $f(y_k) - \frac{1}{2\beta}\|\nabla f(y_k)\|^2 \ge f(y_k - \beta^{-1}\nabla f(y_k))$. Let us therefore ensure $\frac{t_k^2}{2\gamma_{k+1}} = \frac{1}{2\beta}$, 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

$$\phi^*_{k+1} \ge f(x_{k+1}) + (1-t_k)\langle \nabla f(y_k), \frac{t_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k\rangle.$$

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 seqeunces
  Input: Starting point $x_0 \in \mathbf{E}$.
  Set $k = 0$, $v_0 = x_0$, and $\phi_0^* = f(x_0)$;
  for $k$ = 0, ..., $K$ do
      Compute $t_k \in (0,1)$ from equation
          $\beta t_k^2 = (1-t_k)\gamma_k + t_k\alpha.$    (2.33)
      Set
          $\gamma_{k+1} = (1-t_k)\gamma_k + t_k\alpha$    (2.34)
          $y_k = \frac{t_k\gamma_k v_k + \gamma_{k+1}x_k}{\gamma_k + t_k\alpha}$    (2.35)
          $x_{k+1} = y_k - \frac{1}{\beta}\nabla f(y_k)$    (2.36)
          $v_{k+1} = \frac{(1-t_k)\gamma_k v_k + t_k\alpha y_k - t_k\nabla f(y_k)}{\gamma_{k+1}}$    (2.37)
      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 (2.31) and exercise 2.43, we see that the point x_k generated by Algorithm 4 satisfy

$$f(x_k) - f^* \leq \lambda_k \left[f(x_0) - f^* + \frac{\gamma_0}{2}\|x_0 - x^*\|^2\right], \tag{2.38}$$

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

Theorem 2.45 (Decay of λ_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 $\lambda_k = \Pi_{i=0}^{k-1}(1-t_i) \leq \left(1 - \sqrt{\frac{\alpha}{\beta}}\right)^k$.

For the other inequality, let $c_j = \frac{1}{\sqrt{\lambda_j}}$. Taking into account that λ_j are decreasing, observe

$$\begin{aligned} c_{j+1} - c_j &= \frac{\sqrt{\lambda_j} - \sqrt{\lambda_{j+1}}}{\sqrt{\lambda_j}\sqrt{\lambda_{j+1}}} = \frac{\lambda_j - \lambda_{j+1}}{\sqrt{\lambda_j \lambda_{j+1}}(\sqrt{\lambda_j} + \sqrt{\lambda_{j+1}})} \\ &\geq \frac{\lambda_j - \lambda_{j+1}}{2\lambda_j\sqrt{\lambda_{j+1}}} = \frac{\lambda_j - (1-t_j)\lambda_j}{2\lambda_j\sqrt{\lambda_{j+1}}} = \frac{t_j}{2\sqrt{\lambda_{j+1}}}. \end{aligned}$$

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

$$\gamma_{j+1} \geq (1-t_j)\gamma_j \geq (1-t_j)\gamma_0\lambda_j \geq \gamma_0\lambda_{j+1}.$$

Thus $\gamma_0\lambda_{j+1} \leq \gamma_{j+1} = \beta t_j^2$, which implies that $\frac{t_j}{2\sqrt{\lambda_{j+1}}} \geq \frac{1}{2} \cdot \sqrt{\frac{\gamma_0}{\beta}}$. So we deduce that

$$c_{j+1} - c_j \geq \frac{1}{2} \cdot \sqrt{\frac{\gamma_0}{\beta}}.$$

Summing over $j = 0, \ldots, k-1$, we get

$$c_k - c_0 \geq \frac{k}{2} \cdot \sqrt{\frac{\gamma_0}{\beta}}$$

and hence

$$\frac{1}{\sqrt{\lambda_k}} - 1 \geq \frac{k}{2} \cdot \sqrt{\frac{\gamma_0}{\beta}}.$$

The claimed estimate

$$\lambda_k \leq \frac{4\beta}{\left(2\sqrt{\beta} + k\sqrt{\gamma_0}\right)^2}$$

follows. □

Corollary 2.46. *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 (2.38), Theorem 2.45, and the inequality $f(x_0) - f^* \leq \frac{\beta}{2}\|x_0 - x^*\|^2$. □

Let us try to eliminate v_k. Solving for v_k in (2.35) and plugging in this description into (2.37) and rearranging yields the equality

$$v_{k+1} = \tfrac{1}{\gamma_{k+1}}\tfrac{(1-t_k)\gamma_k + t_k\alpha}{t_k} y_k - \tfrac{1-t_k}{t_k} x_k - \tfrac{t_k}{\gamma_{k+1}}\nabla f(y_k).$$

Hence we deduce

$$\begin{aligned} v_{k+1} &= \tfrac{1}{\gamma_{k+1}} \tfrac{\gamma_{k+1}}{t_k} y_k - \tfrac{1-t_k}{t_k} x_k - \tfrac{1}{t_k \beta} \nabla f(y_k) \\ &= x_k + \tfrac{1}{t_k}(x_{k+1} - x_k). \end{aligned}$$

where the first inequality follows from (2.34) and (2.33), while the last uses (2.36). Plugging in the analogous expression of v_{k+1} into (2.35) yields

$$\begin{aligned} y_{k+1} &= x_{k+1} + \tfrac{t_{k+1}\gamma_{k+1}(1-t_k)}{t_k(\gamma_{k+1}+t_{k+1}\alpha)}(x_{k+1} - x_k) \\ &= x_{k+1} + \zeta_k (x_{k+1} - x_k), \end{aligned}$$

where we define

$$\zeta_k := \tfrac{t_{k+1}\gamma_{k+1}(1-t_k)}{t_k(\gamma_{k+1}+t_{k+1}\alpha)}.$$

Thus v_k is eliminated from the algorithm. Let us now eliminate γ_k. To this end note from (2.33) $t_{k+1}\alpha = \beta t_{k+1}^2 - (1 - t_{k+1})\gamma_{k+1}$, and hence

$$\zeta_k := \tfrac{t_{k+1}\gamma_{k+1}(1-t_k)}{t_k(\beta t_{k+1}^2 + t_{k+1}\gamma_{k+1})} = \tfrac{\gamma_{k+1}(1-t_k)}{t_k(\beta t_{k+1}+\gamma_{k+1})} = \tfrac{t_k(1-t_k)}{t_{k+1}+t_k^2},$$

where the last equality uses $\gamma_{k+1} = \beta t_k^2$. Finally plugging in $\gamma_{k+1} = \beta t_k^2$ into (2.33) yields

$$t_{k+1}^2 = (1 - t_{k+1})t_k^2 + \tfrac{\alpha}{\beta} t_{k+1}.$$

Thus γ_k is eliminated from the scheme.

Algorithm 5: Simplified fast gradient method

```
Input: Starting point $x_0 \in \mathbf{E}$ and $t_0 \in (0, 1)$.
Set $k = 0$ and $y_0 = x_0$;
for $k = 0, \ldots, K$ do
    Set
        $x_{k+1} = y_k - \frac{1}{\beta}\nabla f(y_k).$
    Compute $t_{k+1} \in (0, 1)$ from the equation
        $t_{k+1}^2 = (1 - t_{k+1})t_k^2 + \tfrac{\alpha}{\beta} t_{k+1}$        (2.39)
    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 2.47. *Setting $t_0 = \frac{\alpha}{\beta}$ in Algorithm 5 yields iterates satisfying*

$$f(x_k) - f^* \le \beta \min\left\{ \left(1 - \sqrt{\frac{\alpha}{\beta}}\right)^k, \frac{4}{(k+2)^2} \right\} \cdot \|x_0 - x^*\|^2.$$

It is important to note that in the case $\alpha = 0$, Algorithm 5 is exactly Algorithm 2 with $a_k = t_k$. Indeed, equality (2.39) can be rewritten as

$$\frac{1 - t_{k+1}}{t_{k+1}^2} = \frac{1}{t_k^2},$$

which is exactly the equality in (2.25). 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 (2.39). Thus the interpolation coefficients in the definition of y_k are exactly the same.

2.8.3 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 $f\colon \mathbf{E} \to \mathbf{R}$ be a β-smooth and α-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 $x, y \in \mathbf{E}$, we let $\texttt{line\_search}\,(x, y)$ be the minimizer of f on the line between x and y. We assume throughout this section that $\texttt{line\_search}\,(x, y)$ 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 $x \in \mathbf{E}$, we define the following two points

$$x^+ := x - \tfrac{1}{\beta}\nabla f(x) \qquad \text{and} \qquad x^{++} := x - \tfrac{1}{\alpha}\nabla f(x).$$

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

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

yields

$$f(x) \geq \left(f(\bar{x}) - \frac{\|\nabla f(\bar{x})\|^2}{2\alpha} \right) + \frac{\alpha}{2}\left\|x - \bar{x}^{++}\right\|^2. \tag{2.40}$$

Suppose we have now available two quadratic lower-estimators:

$$f(x) \geq Q_A(x) := v_A + \frac{\alpha}{2}\left\|x - x_A\right\|^2,$$
$$f(x) \geq Q_B(x) := v_B + \frac{\alpha}{2}\left\|x - x_B\right\|^2.$$

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 λ yields the tightest lower-bound on the minimal value of f? To answer this question, observe the equality

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

where

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

and

$$v_\lambda = v_B + \left(v_A - v_B + \frac{\alpha}{2}\left\|x_A - x_B\right\|^2\right)\lambda - \left(\frac{\alpha}{2}\left\|x_A - x_B\right\|^2\right)\lambda^2. \tag{2.41}$$

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

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

With this choice of λ, we call the quadratic function $\overline{Q} = \bar{v} + \frac{\alpha}{2}\|\cdot - \bar{c}\|^2$ the *optimal averaging* of Q_A and Q_B. See Figure 2.6 for an illustration.

An algorithmic idea emerges. Given a current iterate x_k, form the quadratic lower-model $Q(\cdot)$ in (2.40) 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.

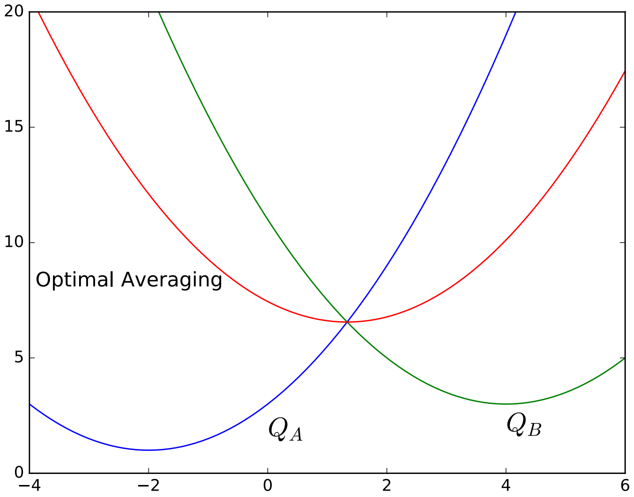

Figure 2.6: The optimal averaging of $Q_A(x) = 1 + 0.5(x + 2)^2$ and $Q_B(x) = 3 + 0.5(x - 4)^2$.

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.

Algorithm 6: Optimal Quadratic Averaging

```
**Input**: Starting point $x_0$ and strong convexity constant $\alpha > 0$.
**Output**: Final quadratic $Q_K(x) = v_K + \frac{\alpha}{2}\|x - c_K\|^2$ and $x_K^+$.
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
$c_0 = x_0^{++}$;
**for** $k = 1, \ldots, K$ **do**
    Set $x_k = \texttt{line\_search}\left(c_{k-1}, x_{k-1}^+\right)$;
    Set $Q(x) = \left(f(x_k) - \frac{\|\nabla f(x_k)\|^2}{2\alpha}\right) + \frac{\alpha}{2}\left\|x - x_k^{++}\right\|^2$ ;
    Let $Q_k(x) = v_k + \frac{\alpha}{2}\|x - c_k\|^2$ be the optimal averaging of $Q$ and
    $Q_{k-1}$ ;
**end**
```

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

Lemma 2.48. *Suppose that $\overline{Q} = \bar{v} + \frac{\alpha}{2}\| \cdot - \bar{c}\|^2$ is the optimal averaging of the quadratics $Q_A = v_A + \frac{\alpha}{2}\| \cdot - x_A\|^2$ and $Q_B = v_B + \frac{\alpha}{2}\| \cdot - x_B\|^2$. Then the quantity $\bar{v}$ is nondecreasing in both v_A and v_B. Moreover, whenever the* *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 (2.41) we deduce

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

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

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

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. □

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

Theorem 2.49 (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 minorants 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

$$Q_k(x) = v_k + \frac{\alpha}{2}\left\|x - c_k\right\|^2,$$

is the optimal averaging of $Q_A(x) = v_A + \frac{\alpha}{2}\left\|x - x_k^{++}\right\|^2$ and $Q_B(x) = v_B + \frac{\alpha}{2}\left\|x - c_{k-1}\right\|^2$. 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:

$$\begin{aligned} v_B \geq f(x_{k-1}^+) - \frac{\alpha}{2}r_{k-1} &\geq f(x_k) - \frac{\alpha}{2}r_{k-1} \\ &\geq f(x_k^+) + \frac{1}{2\beta}\left\|\nabla f(x_k)\right\|^2 - \frac{\alpha}{2}r_{k-1} \\ &= f(x_k^+) - \frac{\alpha}{2}\left(r_{k-1} - \frac{1}{\alpha^2\kappa}\left\|\nabla f(x_k)\right\|^2\right) := \hat{v}_B. \end{aligned}$$

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

$$\begin{aligned} v_k \geq v_A \geq \hat{v}_A &= f(x_k^+) - \frac{\alpha}{2}h^2\left(1 - \frac{1}{\kappa}\right) \\ &\geq f(x_k^+) - \frac{\alpha}{2}r_{k-1}\left(\frac{1 - \frac{1}{\kappa}}{2}\right) \\ &\geq f(x_k^+) - \frac{\alpha}{2}r_{k-1}\left(1 - \frac{1}{\sqrt{\kappa}}\right) \\ &= f(x_k^+) - \frac{\alpha}{2}r_k. \end{aligned}$$

Hence in this case, the proof is complete.

Next suppose $h^2 > \frac{r_{k-1}}{2}$ and let $v + \frac{\alpha}{2}\|\cdot - c\|^2$ be the optimal average of the two quadratics $\hat{v}_A + \frac{\alpha}{2}\|\cdot - x_k^{++}\|^2$ and $\hat{v}_B + \frac{\alpha}{2}\|\cdot - c_{k-1}\|^2$. By Lemma 2.48, the inequality $v_k \geq v$ holds. We claim that equality

$$v = \hat{v}_B + \frac{\alpha}{8}\frac{(d^2 + \frac{2}{\alpha}(\hat{v}_A - \hat{v}_B))^2}{d^2} \qquad \text{holds.} \tag{2.42}$$

This follows immediately from Lemma 2.48, once we show $\frac{1}{2} \geq \frac{|\hat{v}_A - \hat{v}_B|}{\alpha d^2}$. To this end, note first the equality $\frac{|\hat{v}_A - \hat{v}_B|}{\alpha d^2} = \frac{|r_{k-1} - h^2|}{2d^2}$. The choice $x_k = \texttt{line\_search}\left(c_{k-1}, x_{k-1}^+\right)$ ensures:

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

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

$$r_{k-1} - h^2 < \frac{r_{k-1}}{2} < h^2 \leq d^2. \tag{2.43}$$

Hence we can be sure that (2.42) 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 (2.43), 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 2.50. Fix t quadratics $Q_i(x) := v_i + \frac{\alpha}{2}\|x - c_i\|^2$, with $i \in \{1, \ldots, t\}$. Define the matrix $C = \begin{bmatrix} c_1 & c_2 & \ldots & c_t \end{bmatrix}$ and vector $v = \begin{bmatrix} v_1 & v_2 & \ldots & v_t \end{bmatrix}^T$.

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

$$c_\lambda = C\lambda \qquad \text{and} \qquad v_\lambda = \left\langle \frac{\alpha}{2}\mathrm{diag}\,(C^T C) + v, \lambda \right\rangle - \frac{\alpha}{2}\|C\lambda\|^2.$$

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

$$\min_{\lambda \in \Delta_t} \frac{\alpha}{2} \|C\lambda\|^2 - \left\langle \frac{\alpha}{2} \mathrm{diag}\,(C^T C) + v, \lambda \right\rangle .$$

Chapter 3

Minimizing Sums of Smooth and Simple Functions

In this chapter, we study minimization of the sum of a 'simple' and a smooth function. This modeling mechanism lets us incorporate prior information about the decision variable, including structure (e.g. sparsity or smoothness), and the feasible region (e.g. non-negativity or box constraints). The simple function must be convex, but is allowed to be non-smooth, and in particular can take on infinite values.

First-order methods are easily modified to account for the simple term. The modifications preserve the rates of convergence from Chapter 2, and can be analyzed using analogous techniques to those already presented. We gain flexibility at essentially no computational cost. We start with a few motivating examples, and then provide the analysis.

Example 3.1 (Optimization with Simple Constraints)**.** Consider a smooth model $f(x)$ from Chapter 2, e.g. any learning problem arising from a general linear model. Suppose you are also given side information about the domain of the predictors x. For example:

- some components of x are non-negative
- some components of x have lower and upper bounds
- x must be in the level set of some convex function, e.g. $\|x\|_2 \leq \tau$.
- x must be in a certain affine subspace, e.g. $Ax = b$.

All of these constraints can be concisely written as $x \in C$, where C is a closed convex set. The modified optimization problem is then

$$\min f(x) + \delta\left(x \mid C\right),$$

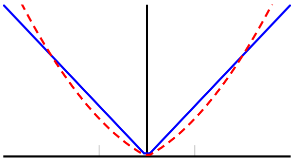

Figure 3.1: 1-norm (blue) and elastic net (red dashed) both have nonsmooth behavior at the origin.

where

$$\delta\left(x \mid C\right) := \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}.$$

is called the *convex indicator function* of C. We consider $\delta\left(\cdot \mid C\right)$ 'simple' when C admits an efficiently computable projection:

$$\mathrm{proj}_C(z) = \operatorname*{argmin}_{x \in C} \frac{1}{2}\|x - z\|^2.$$

Example 3.2 (Sparse regularization)**.** The notion of *sparsity* is fundamental to modern numerical analysis. Analogously to matrix sparsity, 'x is sparse' means either that most $x_i = 0$, or that the magnitudes of $|x_i|$ are quickly decaying. Modelers exploit sparsity in a range of settings.

1. **Compressive sensing.** Many signals are sparse in particular transform domains. For example, superpositions of periodic signals have a sparse Fourier representation. If a typical photograph is represented using *wavelets*, the magnitudes of the wavelet coefficients decay rapidly. Wavefields generated by earthquakes can be efficiently represented using *curvelets*. Applications such as image denoising and deblurring, seismic inverse problems, and image compression benefit from these ideas. The problems are captured by the formulation

$$\min_x \|b - AWx\|^2 + r(x),$$

   where A is a specially designed measurement matrix, typically with far fewer rows than columns, W is the transform where the signal of interest admits a sparse representation (e.g. Fourier, wavelets or curvelets), and $r(\cdot)$ is a non-smooth function that promotes sparsity of the input. Two common convex examples are $r(x) = \|x\|_1$, and $r(x) = \alpha\|x\|_1 + (1-\alpha)\|x\|^2$, known as the *elastic net*, see Figure 3.1. The curvature of the elastic net helps it find groups of correlated predictors in practice.

2. **Statistical learning problems.** We cannot expect that general learning problems will have sparse solutions x. However, for many

models, we want to discover the most important predictors. We can therefore consider the parametrized family of solutions

$$x(\lambda) \in \arg\min_x f(x) + \lambda r(x),$$

with $r(x)$ a nonsmooth regularizer. When λ is larger than $\|\nabla f(0)\|_\infty$, $x(\lambda) = 0$. As λ decreases, x_i 'activate'. The earliest activated entries can indicate the most important predictors. This kind of analysis is known as the Lasso, and is used in conjunction with all general linear models.

Example 3.3 (More non-smooth regularizers)**.** While the 1-norm penalty is ubiquitously used to promote sparsity, many other related regularizers are also used in a range of learning and inverse problems.

- The *OWL norm* $r(x) = \alpha\|x\|_1 + (1-\alpha)\|x\|_\infty$ can detect groups of correlated predictors even better than the elastic net.
- The *group lasso* penalty $r(x) = \sum_j \|x_j\|$ forces pre-specified groups of indices x_j to be jointly included or excluded.
- The *total variation* penalty $r(x) = \|Dx\|_1$, gives piecewise constant signals along directions determined by differential operator D.

Example 3.4 (Sparse covariance estimation)**.** Suppose we are given a symmetric positive definite sample covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$. Its inverse F is the (Fisher) information matrix, and the equality $F_{ij} = 0$ implies conditional independence of variables i and j. The *graphical Lasso* problem looks for sparse information by solving the problem

$$\min_{X\succeq 0} \left\{\log\det(X) + \operatorname{tr}(\Sigma X) + \lambda\|X\|_1\right\}.$$

Example 3.5 (Convex matrix completion)**.** Suppose we observe some entries a_{ij} of a large matrix $A \in \mathbf{R}^{m\times n}$, with ij ranging over some small index set $\mathcal{I}$, and wish to to recover A (i.e. fill in the missing entries). A classic approach is to penalize the *nuclear norm*, leading to the problem

$$\min_X \ \frac{1}{2}\sum_{ij\in\mathcal{I}} \|X_{ij} - a_{ij}\|^2 + \|X\|_*.$$

Compare this formulation to the smooth factorization approach.

Example 3.6 (Portfolio Estimation with Simplex Constraints)**.** Markowitz portfolio estimation is a foundational topic in computational finance. Given a set of N stocks, we consider their returns over T time steps, encoded by a matrix $F \in \mathbb{R}^{N\times T}$. From this information it is straightforward to compute a vector of mean returns $\mu \in \mathbb{R}^N$ and a covariance matrix $\Sigma \in \mathbb{R}^{N\times N}$, assuming the returns process is stationary. We want to choose investment weights for the N assets to minimize a measure of risk for a given return α. A common risk measure is the variance of the portfolio, $w^T \Sigma w$, so we have

$$\min_{w \in \Delta} w^T \Sigma w \quad \text{such that} \quad w^T \mu = \alpha.$$

The set $\Delta = \{w : w_i \in [0,1], 1^T w_i = 1\}$ is the unit simplex, which forces purchases must be non-negative (no shorting) and investment of all assets (one asset is typically a 'safe' option such as a bond or index fund).

3.1 Proximal Gradient Method

Consider the problem

$$\min_x f(x) = g(x) + h(x),$$

with g, h convex and g a β-smooth map. Analogously to steepest descent, we can design an iterative method by minimizing a simple upper bound obtained from g:

$$\begin{aligned} x^+ &= \operatorname*{argmin}_y g(x) + \langle \nabla g(x), y - x \rangle + \frac{\beta}{2}\|y - x\|^2 + h(y) \\ &= \operatorname*{argmin}_y \frac{\beta}{2}\|y - (x - \beta^{-1}\nabla g(x))\|^2 + h(y) \end{aligned}$$

Minimizing the sum of $h(y)$ and a small quadratic can be viewed as an atomic operation.

Definition 3.7 (Proximity Operator)**.** For a convex function $h(y) : \mathbb{R}^n \to \mathbb{R} \cup \infty$, define the *proximity* operator $\text{prox}_{\alpha h} : \mathbb{R}^n \to \mathbb{R}^n$ by

$$\operatorname*{prox}_{\gamma h}(z) = \operatorname*{argmin}_x \frac{1}{2\gamma}\|x - z\|^2 + h(x).$$

Note that the optimization problem defining $\text{prox}_{\gamma h}$ is strongly convex, so the solution is unique. The iteration for x^+ can therefore be written more compactly as

$$x^+ = \operatorname*{prox}_{\beta^{-1}h}(x - \beta^{-1}\nabla g(x)).$$

To analyze this algorithm, we introduce the proximal gradient map

$$G_t(x) := \frac{1}{t}\left(x - \text{prox}_{th}(x - t\nabla g(x))\right),$$

which behaves similarly to the gradient of a smooth function. For example, the proximal gradient iteration is written

$$x^+ = x - \beta^{-1} G_{\beta^{-1}}(x).$$

To understand the map G and its consequences, we first need to extend the notion of derivative to nonsmooth convex functions.

Definition 3.8 (Subgradient and Subdifferential)**.** Let $h : U \to \mathbb{R}$ be a convex function. A *subgradient* of h at x is a vector $v \in \mathbb{R}^n$ that satisfies

$$h(y) \geq h(x) + \langle v, y - x\rangle \quad \text{for all } y \in U.$$

The *subdifferential* of h at x is the set of all subgradients, and is denoted by $\partial h(x)$. Equivalently,

$$\partial h(x) := \{v \in \mathbb{R}^n : h(y) \geq h(x) + \langle v, y - x\rangle \text{ for all } y \in U.\}$$

When there is only one point in $\partial h(x)$, then h is differentiable at x and $\partial h(x) = \nabla f(x)$. If $0 \in \partial h(x)$, then immediately from the definition we have $h(y) \geq h(x)$ for all $y \in U$ so x must be a global minimizer.

Example 3.9 (Subdifferential of $\|\cdot\|_1$)**.** Suppose $h(x) = |x|$. Then

$$\partial h(x) = \begin{cases} \{1\} & \text{if } x > 0 \\ [-1, 1] & \text{if } x = 0 \\ \{-1\} & \text{if } x < 0 \end{cases}$$

Since $\|x\|_1 = \sum |x_i|$, the subdifferential of the 1-norm can be computed by applying the above formula to each coordinate.

Example 3.10 (Subdifferential of an indicator function)**.** Suppose $h(x) = \delta\left(x \mid C\right)$ where C is a closed convex set. If $x \in C$, $v \in \partial h(x)$ is characterized by

$$\delta\left(y \mid C\right) \geq \langle v, y - x\rangle + \delta\left(x \mid C\right) = \langle v, y - x\rangle.$$

This inequality always holds for $y \notin C$; if $y \in C$, it gives $0 \geq \langle v, y - x\rangle$. Therefore

$$\partial h(x) = \{v : 0 \geq \langle v, y - x\rangle \quad \text{for all } y \in C\},$$

which is called the *normal cone* to C at x.

Coming back to the map G, we show it is analogous to the gradient map in the smooth case.

Remark 3.11. $G_t(x) - \nabla g(x) \in \partial h(x^+)$, where $x^+ = \text{prox}_{th}(x - t\nabla g(x))$

Proof. Observe

$$x^+ = \underset{u}{\text{argmin}} \ \{g(x) + \langle \nabla g(x), u - x\rangle + \frac{1}{2t}\|u - x\|^2 + h(u)\}$$

Then differentiating the RHS of the above expression with respect to u at $u = x^+$ gives

$$0 \in \nabla g(x) + \underbrace{\frac{1}{t}(x^+ - x)}_{G_t(x)} + \partial h(x^+)$$

That is,

$$G_t(x) - \nabla g(x) \in \partial h(x^+).$$

□

It immediately follows that $G_t(x) = 0$ if and only if x minimizes $g + h$.

Theorem 3.12. *Suppose that g is β-smooth and α-convex, where α can be 0, and define $x^+ := \mathrm{prox}_{th}(x - t\nabla g(x))$, and assume that h is convex. Then we have*

$$f(y) \geq f(x^+) + \langle G_t(x), y - x\rangle + t\left(1 - \frac{\beta t}{2}\right)\|G_t(x)\|^2 + \frac{\alpha}{2}\|y - x\|^2. \quad (3.1)$$

Proof.

$$\begin{aligned}
f(x^+) &= g(x - tG_t(x)) + h(x^+) \\
&\leq g(x) - t\left\langle \nabla g(x), G_t(x)\right\rangle + \frac{\beta t^2}{2}\|G_t(x)\|^2 + h(x^+) \\
&\leq g(y) + \langle x - y, \nabla g(x)\rangle - \frac{\alpha}{2}\|y - x\|^2 - t\left\langle \nabla g(x), G_t(x)\right\rangle + \frac{\beta t^2}{2}\|G_t(x)\|^2 h(x^+) \\
&= g(y) + \left\langle x^+ - y, \nabla g(x)\right\rangle - \frac{\alpha}{2}\|y - x\|^2 + \frac{\beta t^2}{2}\|G_t(x)\|^2 + h(x^+) \\
&\leq f(y) + \left\langle x^+ - y, \nabla g(x)\right\rangle - \frac{\alpha}{2}\|y - x\|^2 + \frac{\beta t^2}{2}\|G_t(x)\|^2 + \langle G_t(x) - \nabla g(x), x^+ - y\rangle \\
&\leq f(y) + \left\langle x^+ - y, G_t(x)\right\rangle - \frac{\alpha}{2}\|y - x\|^2 + \frac{\beta t^2}{2}\|G_t(x)\|^2 \\
&= f(y) - \langle y - x, G_t(x)\rangle - \frac{\alpha}{2}\|y - x\|^2 - \left\langle x - x^+, G_t(x)\right\rangle + \frac{\beta t^2}{2}\|G_t(x)\|^2 \\
&= f(y) - \langle y - x, G_t(x)\rangle - \frac{\alpha}{2}\|y - x\|^2 - \left(t - \frac{\beta t^2}{2}\right)\|G_t(x)\|^2.
\end{aligned}$$

□

Remarks:

1. If $\alpha = 0$, taking $t = \frac{1}{\beta}$ and $y = x$, we have

$$f(x^+) \leq f(x) - \frac{1}{2\beta}\|G_t(x)\|^2.$$

2. Letting $y = x^*$ and $t = \frac{1}{\beta}$, we have

$$0 \geq f(x^+) - f(x^*) + \langle G_t(x), x^* - x\rangle + \frac{1}{2t}\|G_t(x)\|^2 + \frac{\alpha}{2}\|x^* - x\|^2$$

and in particular

$$\langle G_t(x), x - x^*\rangle \geq \frac{1}{2t}\|G_t(x)\|^2 + \frac{\alpha}{2}\|x^* - x\|^2$$

Rate for convex problems: The proximal gradient method with $\frac{1}{\beta}$ step satisfies

$$f(x_k) - f(x^*) \leq \frac{\beta}{2k}\|x_1 - x^*\|^2.$$

The key inequality is

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

Exercise 3.13. Derive the above inequality.

We immediately have Theorem 2.25 for the prox-gradient method.

Theorem 3.14 (Prox-gradient descent and convexity)**.** *Suppose that $h(x) = f(x) + g(x)$, with $f\colon \mathbf{E} \to \mathbf{R}$ is convex and β-smooth, and g convex. Then the iterates generated by the prox-gradient descent method satisfy*

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

Rate for strongly convex problems: If in addition f is α-convex, then

$$f(x_{k+1}) - f(x^*) \leq \frac{\beta}{2}\left(1 - \frac{\alpha}{\beta}\right)^k \|x_1 - x^*\|^2$$

and

$$\|x_{k+1} - x^*\|^2 \leq \left(1 - \frac{\alpha}{\beta}\right)^k \|x_1 - x^*\|^2.$$

Proof :

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 - 2t\,\langle G_t(x_k), x_k - x^*\rangle + t^2\|G_t(x_k)\|^2 \\
&\leq \|x_k - x^*\|^2 - 2t\left(\frac{t}{2}\|G_t(x_k)\|^2 + \frac{\alpha}{2}\|x_k - x^*\|^2\right) + t^2\|G_t(x_k)\|^2 \\
&= \|x_k - x^*\|^2 - \frac{\alpha}{\beta}\|x_k - x^*\|^2
\end{aligned}$$

Again, we can immediately state a theorem analogous to Theorem 3.15.

Theorem 3.15 (Prox-gradient descent and strong convexity)**.**
Suppose that $h(x) = f(x) + g(x)$, with $f\colon \mathbf{E} \to \mathbf{R}$ is α-strongly convex and β-smooth, and g convex. Then the iterates generated by the proximal 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.

In other words, adding a ‘prox-friendly’ convex function g preserves the rates of first order methods for f alone. If g is strongly convex but f is not, proximal gradient still has a linear rate (see the exercises).

Exercise 3.16. Show that if g is α_2-convex, then (3.1) can be strengthened to

$$f(y) \geq f(x^+)+(1+t\alpha_2)\langle G_t(x), y-x\rangle+t\left(1-\frac{\beta t+\alpha_2 t}{2}\right)\|G_t(x)\|^2+\frac{\alpha+\alpha_2}{2}\|y-x\|^2.$$

Exercise 3.17. Show that if g is α_2-convex, then with step $t = \frac{1}{\beta+\alpha_2}$ in Remark 2 we have

$$0 \geq f(x^+) - f(x^*) + \langle G_t(x), x^* - x\rangle + \frac{1}{2t}\|G_t(x)\|^2 + \frac{\alpha+\alpha_2}{2}\|x^* - x\|^2$$

and in particular

$$\langle G_t(x), x - x^*\rangle \geq \frac{1}{2t}\|G_t(x)\|^2 + \frac{\alpha+\alpha_2}{2}\|x^* - x\|^2$$

Exercise 3.18. State and prove the convergence rate under the additional assumption that g is α_2 convex.

Chapter 4

Convexity

Algorithms for minimizing smooth convex functions rely heavily on basic results of mathematical analysis, summarized in Section 1.5. Much in the same way, algorithms for nonsmooth convex optimization are based on a mathematical field, called *convex analysis.* This chapter is devoted to developing the main results of this subject.

4.1 Basic convex geometry

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

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

A set $Q \subseteq \mathbf{R}^n$ is *convex* if for any two points $x, y \in Q$, the line segment $[x, y]$ is also contained in Q. Recall also the definition of the *unit simplex*:

$$\Delta_n = \left\{ x \in \mathbf{R}^n : \sum_{i=1}^{n} x_i = 1, x \geq 0 \right\}.$$

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

Exercise 4.1. Show that a set $Q \subset \mathbf{E}$ is convex if and only if any convex combination of points $x_1, \ldots, x_t \in Q$ lies in Q for any integer $t \geq 1$.

Convexity is very stable property, being preserved under a variety of operations.

Exercise 4.2. Prove the following statements.

1. **(Pointwise sum)** For any two convex sets $Q_1, Q_2 \subset \mathbf{E}$, 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 $Q \subset \mathbf{E}$ and $L \in \mathbf{Y}$ and linear maps $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$ and $\mathcal{H}\colon \mathbf{Y} \to \mathbf{E}$, the image $\mathcal{A}Q$ and the preimage $\mathcal{H}^{-1}L$ are convex sets.

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

Exercise 4.3. For any set $Q \subset \mathbf{E}$, prove the equality:
$$\mathrm{conv}(Q) = \left\{ \sum_{i=1}^{k} \lambda_i x_i \ : \ k \in \mathbb{N}_+, \ x_1, \ldots, x_k \in Q, \ \lambda \in \Delta_k \right\}. \tag{4.1}$$

The following theorem shows that in the description (4.1), it is sufficient to take $k \le n + 1$.

Theorem 4.4 (Carathéodory)**.** *Consider a set $Q \subset \mathbf{E}$. Then for any point $x \in \mathrm{conv}(Q)$, there exist points $x_1, \ldots, x_{n+1} \in Q$ along with weights $\lambda \in \Delta_{n+1}$ satisfying $x = \sum_{i=1}^{n+1} \lambda_i x_i$.*

Proof. Since x belongs to $\mathrm{conv}(Q)$, 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 $k \le n + 1$ holds, then there is nothing to prove. Hence suppose $k \ge n + 2$. Then the vectors
$$x_2 - x_1, \ldots, x_k - x_1$$
are linearly dependent. That is there exists numbers μ_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 α 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
$$\sum_{i=1}^{k} (\lambda_i - \alpha \mu_i) = 1.$$

We will now choose α 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 $i^* \in \operatorname{argmin}_i\{\lambda_i/\mu_i : \mu_i > 0\}$. 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. The result follows. □

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 $\operatorname{aff} Q$, is the intersection of all affine sets containing Q. Clearly, $\operatorname{aff} Q$ is itself an affine set. The *relative interior* of Q, denoted $\operatorname{ri} Q$, is the interior of Q relative to $\operatorname{aff} Q$, that is

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

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

Theorem 4.5 (Relative interior is nonempty). *For any nonempty convex set $Q \subset \mathbf{E}$, the relative interior $\operatorname{ri} Q$ is nonempty.*

Proof. Without loss of generality, we may translate Q to contain the origin. Let d be the dimension of the linear subspace $\operatorname{aff} Q$. Observe that Q must contain some d linearly independent vectors $x_1, \ldots, x_d$, since otherwise $\operatorname{aff} Q$ would have a smaller dimension than d. Consider the linear map $A\colon \mathbf{R}^d \to \operatorname{aff} Q$, 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

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

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

The following is a useful topological property of convex sets.

Theorem 4.6 (Accessibility). *Consider a convex set Q and two points $x \in \operatorname{ri} Q$ and $y \in \operatorname{cl} Q$. Then the line segment $[x, y)$ is contained in $\operatorname{ri} Q$.*

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

Corollary 4.7. *For any nonempty convex set Q in $\mathbf{E}$, we have* $\mathrm{cl}\,(\mathrm{ri}\,Q) = \mathrm{cl}\,Q$.

Proof. The inclusion $\mathrm{ri}\,Q \subseteq Q$ immediately implies $\mathrm{cl}\,(\mathrm{ri}\,Q) \subseteq \mathrm{cl}\,Q$. Conversely, fix a point $y \in \mathrm{cl}\,Q$. Since $\mathrm{ri}\,Q$ is nonempty by Theorem 4.5, we may also choose a point $x \in \mathrm{ri}\,Q$. Theorem 4.6 then immediately implies $y \in \mathrm{cl}\,[x, y) \subseteq \mathrm{cl}\,(\mathrm{ri}\,Q)$. Since $y \in \mathrm{cl}\,Q$ is arbitrary, we have established the equality $\mathrm{cl}\,(\mathrm{ri}\,Q) = \mathrm{cl}\,Q$.

□

4.1.1 Separation theorem

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* of Q.

We begin with the following basic definitions. Along with any set $Q \subset \mathbf{E}$ we define the *distance function*

$$\mathrm{dist}_Q(y) := \inf_{x \in Q} \|x - y\|$$

and the *projection*

$$\mathrm{proj}_Q(y) := \{x \in Q : \mathrm{dist}_Q(y) = \|x - y\|\}.$$

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

Exercise 4.8. Show that for any nonempty set $Q \subseteq \mathbf{E}$, the function $\mathrm{dist}_Q \colon \mathbf{E} \to \mathbf{R}$ is 1-Lipschitz.

If Q is closed, then the nearest-point set $\mathrm{proj}_Q(y)$ is nonempty for any $y \in \mathbf{E}$. 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 $\mathrm{proj}_Q(y)$. When Q is convex, the set $\mathrm{proj}_Q(y)$ is not only nonempty, but is also a singleton.

Theorem 4.9 (Properties of the projection)**.** *For any nonempty, closed, convex set $Q \subset \mathbf{E}$, the set $\mathrm{proj}_Q(y)$ is a singleton, and the unique vector $z \in \mathrm{proj}_Q(y)$ is characterized by the property*

$$\langle y - z, x - z \rangle \leq 0 \qquad \text{for all } x \in Q. \tag{4.2}$$

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

The claim that any point z satisfying (4.2) lies in $\mathrm{proj}_Q(y)$ is an easy exercise. We therefore prove the converse. Since Q is closed, the set $\mathrm{proj}_Q(y)$ is nonempty. Fix a point $z \in \mathrm{proj}_Q(y)$ 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 $z \in \mathrm{proj}_Q(y)$. We conclude that (4.2) holds. To see that $\mathrm{proj}_Q(y)$ is a singleton, consider another point $z' \in \mathrm{proj}_Q(y)$. Then clearly z' lies in the intersection $(\mathrm{cl}\, B_r(y)) \cap (\mathbf{E} \setminus H)$. The definition of H on the other hand, implies that this intersection is the singleton $\{z\}$. □

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

Theorem 4.10 (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 $b \in \mathbf{R}$ satisfying*

$$\langle a, x\rangle \leq b < \langle a, y\rangle \quad \textit{for any } x \in Q.$$

Proof. Define the nonzero vector $a := y - \mathrm{proj}_Q(y)$. Then for any $x \in Q$, the condition (4.2) yields the inequalitites

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

as claimed. □

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

Exercise 4.11. Given a nonempty set $Q \subset \mathbf{E}$, define

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

Prove the equality

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

for any nonempty set $Q \subset \mathbf{E}$.

4.1.2 Cones and polarity

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

Exercise 4.12. Show that a set $K \subset \mathbf{E}$ 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 4.13. Prove for any convex cone $K \subset \mathbf{E}$ the equality $\operatorname{aff}(K) = K - K$.

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 $K \subset \mathbf{E}$, the *polar cone* is the set

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

Thus K° 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, $(\mathbf{R}^n_+)^\circ = \mathbf{R}^n_-$ and $(\mathbf{S}^n_+)^\circ = \mathbf{S}^n_-$.

Exercise 4.14 (Double-polar theorem)**.** For any cone K, prove the equality $(K^\circ)^\circ = \operatorname{cl}\operatorname{conv} K$. (Hint: use separation (Theorem 4.10))

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

Theorem 4.15 (Polarity calculus)**.** *For any linear mapping $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$ and a cone $K \subset \mathbf{Y}$, the chain rule holds*

$$(\mathcal{A}K)^\circ = (\mathcal{A}^*)^{-1}K^\circ.$$

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

$$(K_1 + K_2)^\circ = K_1^\circ \cap K_2^\circ$$

Proof. Observe the equivalence

$$\begin{aligned} y \in (\mathcal{A}K)^\circ &\iff \langle \mathcal{A}x, y\rangle \leq 0 \text{ for all } x \in K \\ &\iff \langle x, \mathcal{A}^*y\rangle \leq 0 \text{ for all } x \in K \\ &\iff y \in (\mathcal{A}^*)^{-1}K^\circ. \end{aligned}$$

This establishes the first equality. The sum rule follows by applying the chain rule to the expression $\mathcal{A}(K_1 \times K_2)$ with the mapping $\mathcal{A}(x, y) := x + y$. □

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 the applying the polarity operation for cones. Consider a set $Q \subset \mathbf{E}$ and let K be the cone generated by $Q \times \{1\} \subset \mathbf{E} \times \mathbf{R}$. That is

$$K = \{(\lambda x, \lambda) \in \mathbf{E} \times \mathbf{R} : x \in \mathbf{E}, \lambda \geq 0\}.$$

It is then natural to define the *polar set* as

$$Q^\circ := \{x \in \mathbf{E} : (x, -1) \in K^\circ\}.$$

Unraveling the definitions, the following algebraic description of the polar appears.

Exercise 4.16. Show for any set $Q \subset \mathbf{E}$, the equality

$$Q^\circ = \{v \in \mathbf{E} : \langle v, x\rangle \leq 1 \text{ for all } x \in Q\}.$$

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 4.14

Exercise 4.17 (Double polar)**.** For any set $Q \subset \mathbf{E}$ containing the origin, we have

$$(Q^\circ)^\circ = \operatorname{cl}\operatorname{conv} Q.$$

4.1.3 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 $Q \subset \mathbf{E}$ 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 $\mathbf{R}_+(x_i - \bar{x})$ 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 convex cone. Whenever Q is convex, this definition simplifies drastically.

Exercise 4.18. Show for any convex set $Q \subset \mathbf{E}$ and a point $\bar{x} \in Q$ the equality:

$$T_Q(\bar{x}) = \operatorname{cl}\ \mathbf{R}_+(Q - \bar{x}) := \operatorname{cl}\{\lambda(x - \bar{x}) : \lambda \geq 0, x \in Q\}.$$

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 full contained in the half-space $\{x \in \mathbf{E} : \langle v, x - \bar{x} \rangle \leq 0\}$ up to a first-order error. More precisely, the *normal cone* to a set $Q \subset \mathbf{E}$ at a point $\bar{x} \in Q$ is defined by

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

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

Again, when Q is convex, the definition simplifies.

Exercise 4.19. Show for any convex set $Q \subset \mathbf{E}$ and a point $\bar{x} \in Q$ the equality,

$$N_Q(\bar{x}) = \{v \in \mathbf{E} : \langle v, x - \bar{x} \rangle \leq 0 \quad \text{for all } x \in Q\}$$

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 4.20. 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 $\mathrm{argmax}_{x \in Q} \langle v, x \rangle$.

4. equality $\mathrm{proj}_Q(\bar{x} + \lambda v) = \bar{x}$ holds for all $\lambda \geq 0$,

5. equality $\mathrm{proj}_Q(\bar{x} + \lambda v) = \bar{x}$ holds for some $\lambda > 0$.

Exercise 4.21. Show for any convex cone K and a point $x \in K$, the equality

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

Exercise 4.22. Show for any convex set Q and a point $x \in K$, the equivalence

$$x \in \operatorname{int} Q \quad \Longleftrightarrow \quad N_Q(x) = \{0\}.$$

4.2 Convex functions: basic operations and continuity

We next move on to convex analysis – the study of convex functions. We will consider functions f mapping $\mathbf{E}$ to the extended-real-line $\overline{\mathbf{R}} := \mathbf{R} \cup \{\pm\infty\}$. 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 $f \colon \mathbf{E} \to \overline{\mathbf{R}}$ is called *proper* if it never takes the value $-\infty$ and is not identically equal to $+\infty$.

Given a function $f \colon \mathbf{E} \to \overline{\mathbf{R}}$, the *domain* of f and the *epigraph* of f are

$$\begin{aligned} \operatorname{dom} f &:= \{x \in \mathbf{E} : f(x) < +\infty\}, \\ \operatorname{epi} f &:= \{(x, r) \in \mathbf{E} \times \mathbb{R} : f(x) \leq r\}, \end{aligned}$$

respectively. Thus $\operatorname{dom} f$ consists of all point x a which f is finite or evaluates to $-\infty$. The epigraph $\operatorname{epi} f$ 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 $f \colon \mathbf{E} \to \overline{\mathbf{R}}$ is *convex* if $\operatorname{epi} f$ is a convex set in $\mathbf{E} \times \mathbf{R}$. 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 $x, y \in \mathbf{E}$ and $\lambda \in (0, 1)$.

Exercise 4.23 (Jensen's Inequality)**.** Show that a proper functions $f \colon \mathbf{E} \to \overline{\mathbf{R}}$ is convex if and only if we have $f(\sum_{i=1}^{k} \lambda_i x_i) \leq \sum_{i=1}^{k} \lambda_i f(x_i)$ for any integer $k \in \mathbb{N}$, points $x_1, \ldots, x_k \in \mathbf{E}$, and weights $\lambda \in \Delta_k$.

Exercise 4.24. Let $f \colon \mathbf{E} \to \overline{\mathbf{R}}$ be a convex function. Show that if there exists a point $x \in \operatorname{ri}(\operatorname{dom} f)$ with $f(x)$ finite, then f must be proper.

We will call a function $f : \mathbf{E} \to \overline{\mathbf{R}}$ *closed* or *lower-semi-continuous* if $\operatorname{epi} f$ is a closed set.

Exercise 4.25. Show that $f \colon \mathbf{E} \to \overline{\mathbf{R}}$ is closed if and only if the inequality

$$\liminf_{y \to x} f(y) \geq f(x) \qquad \text{holds for any } x \in \mathbf{E}.$$

Consider a function $f \colon \mathbf{E} \to \overline{\mathbf{R}}$. It is easy to check that the set $\operatorname{cl}(\operatorname{epi} f))$ is itself an epigraph of some closed function. We call this function the *closed envelope* of f and denote it by $\operatorname{cl} f$. Similarly, $\operatorname{conv}(\operatorname{epi} f))$ is itself an epigraph of some closed function. We call this function the *convex envelope* of f and denote it by $\operatorname{co} f$. Combining the two operations, yields the close *closed convex envelope* of f, which we denote by $\overline{\operatorname{co}} f = \operatorname{cl}(\operatorname{co}(f))$. 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 $\mathbf{E}$, we say that g is a *minorant* of f if it satisfies $g(y) \leq f(y)$ for all $y \in \mathbf{E}$.

Exercise 4.26. Given a proper function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$, show the equalities

$$\begin{aligned}(\overline{\operatorname{co}} f)(x) &= \sup\{g(x) : g\colon \mathbf{E} \to \overline{\mathbf{R}} \text{ is a closed convex minorant of } f\}\\ &= \sup\{g(x) : g\colon \mathbf{E} \to \overline{\mathbf{R}} \text{ is an affine minorant of } f\}.\end{aligned}$$

Just like is it often easier to work with convex cones than with convex sets, it is often easier to work function whose epigraphs are convex cones. We say that $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is *sublinear* if its epigraph, $\operatorname{epi} f$, is a convex cone.

Exercise 4.27. Let $g\colon \mathbf{E} \to \overline{\mathbf{R}}$ be a proper function.

1. Show that g is sublinear if and only if $f(\lambda x + \mu y) \leq \lambda f(x) + \mu f(y)$ for all $x, y \in \mathbf{E}$ and $\lambda, \mu \geq 0$.
2. Show that if g is sublinear, then $\operatorname{cl} g$ is the support function of the set

$$Q = \{x : \langle x, y\rangle \leq g(y) \quad \forall y \in \mathbf{E}\}.$$

There are a number of convex functions that naturally arise from convex sets. Given a set $Q \subseteq \mathbf{E}$, define its *indicator function*

$$\delta_Q(x) = \begin{cases} 0, & x \in Q \\ +\infty, & x \notin Q \end{cases},$$

its *support function*

$$\delta_Q^\star(v) = \max_{x \in Q} \langle v, x\rangle,$$

and its *gauge function*

$$\gamma_Q(x) = \inf\{\lambda \geq 0\colon x \in \lambda Q\}.$$

Notice that support functions and gauges are sublinear. Conversely, by Exercise 4.27 closed sublinear functions are support functions. The notation $\delta_Q^\star(v)$ may seem strange at first, since it is not clear what the support function $\delta_Q^\star(v)$ has to do with the indication function $\delta_Q(x)$. The notation will make sense shortly, in light of Fenchel conjugacy (Section 4.3).

Exercise 4.28. Show that if Q is convex, then δ_Q, $\delta_Q^\star$, dist_Q and γ_Q are all convex.

Exercise 4.29. Show that for any closed, convex set Q containing the origin, we have $\gamma_Q(x) = \delta_{Q^\circ}^\star(x)$.

Convexity is preserved under a variety of operations.

1. **(Monotone convex composition)** If $f\colon \mathbf{E} \to \mathbf{R}$ is convex, and $\varphi\colon \mathbf{R} \to \mathbf{R}$ is convex and nondecreasing, then the composition $\varphi \circ f$ is convex.

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

3. (**Affine composition**) More generally, if $A\colon \mathbf{E} \to \mathbf{Y}$ is a linear map and $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ is a proper convex functions, 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 $\operatorname{epi} f = \bigcap_{i\in I} \operatorname{epi} f_i$.

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

6. **(Infimal Convolution)** The *infimal convolution* of two functions $f, g\colon \mathbf{E} \to \overline{\mathbf{R}}$ is the function
$$(f\Box g)(x) = \inf_y \{f(x-y) + g(y)\} \tag{4.3}$$
Equivalently, we may write
$$(f\Box g)(x) = \inf\{r : (x, r) \in \operatorname{epi} f + \operatorname{epi} g\}.$$
Hence infimal convolution is an example of a lower envelope with $Q := \operatorname{epi} f + \operatorname{epi} g$. We deduce that if f and g are convex, then so is the convolution $f\Box g$.

7. **(Infimal Projection)** Consider a convex function $g\colon \mathbf{E} \times \mathbf{Y} \to \overline{\mathbf{R}}$. The function
$$f(x) := \inf_y \ g(x, y)$$
is called the *infimal projection* of g. To see that this function is convex, write
$$\begin{aligned} f(x) &= \inf_{y,r}\{r : g(x, y) \le r\} \\ &= \inf\{r : \exists y \text{ with } (x, y, r) \in \operatorname{epi} g\} \\ &= \inf\{r : (x, r) \in \pi_{1,3}(\operatorname{epi} g)\}. \end{aligned} \tag{4.4}$$
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 $Q := \pi_{1,3}(\operatorname{epi} g)$, More concretely, we may write
$$\operatorname{epi} f = \pi_{1,3}(\operatorname{epi} g).$$

We end this section with a remarkable property: convex functions are always locally Lipschitz continuous on the relative interior of their domains.

Theorem 4.30. *Let f be a proper convex function and Q a compact subset of* $\mathrm{ri}\,(\mathrm{dom}\, f)$*. Then f is Lipschitz continuous on Q.*

Proof. Without loss of generality, by restricting to the affine hull, $\mathrm{aff}\,(\mathrm{dom}\, f)$, we can assume that $\mathrm{dom}\, f$ has nonempty interior. Choose $\epsilon > 0$ satisfying $\mathrm{cl}\,(Q + \epsilon B) \subset \mathrm{int}\,(\mathrm{dom}\, f)$, 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 $\bar{x} \in \mathrm{int}\,(\mathrm{dom}\, f)$. The points $(x_i, f(x_i))$ all lie in the boundary of $\mathrm{epi}\, f$

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 $(\bar{v}, \bar{\alpha}) \in N_{\mathrm{epi}\,(f)}(\bar{x}, f(\bar{x}))$, guaranteed to exist by Exercise 4.22. 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 $\bar{v} \in N_{\mathrm{dom}\, f}(\bar{x})$. By Exercise 4.22, this is a contradiction since $\bar{x}$ lies in the interior of $\mathrm{dom}\, f$.

Suppose now we are in the second case, $f(x_i) \to +\infty$. Choose nonzero vectors $(v_i, \alpha_i) \in N_{\mathrm{epi}\, f}(x_i, f(x_i))$. 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 $x \in \mathrm{dom}\, f$. Note if v_i is zero, then $f(x_i)$ 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, α_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 $\bar{v} \in N_{\mathrm{dom}\, f}(\bar{x})$, but this is impossible since $\bar{x}$ is in the interior of $\mathrm{dom}\, f$, yielding a contradiction. Thus, we have proved that f is bounded on $Q + \epsilon B$.

Let α_1 and α_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

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

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 $f\colon \mathbf{R}^2 \to \overline{\mathbf{R}}$ given by

15 down vote accepted A simpler solution for aligning fractions is to let TeX decide what space to add:

$$f(x,y) = \begin{cases} \frac{y^2}{x} & \text{if } x > 0 \\ 0 & \text{if } (x,y) = (0,0) \\ +\infty & \text{otherwise} \end{cases} . \tag{4.5}$$

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

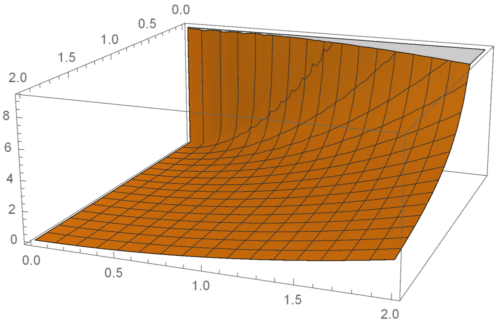

Figure 4.1: Plot of the function $f(x,y)$ in equation (4.5).

4.3 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 4.31. For a function $f : \mathbf{E} \to \overline{\mathbf{R}}$, define the *Fenchel conjugate* function $f^\star\colon \mathbf{E} \to \overline{\mathbf{R}}$ by

$$f^\star(y) = \sup_{x \in \mathbf{E}} \{\langle y, x\rangle - f(x)\}$$

This operation arises naturally from epigraphical geometry. Indeed, from the very definition of the Fenchel conjugate, observe that the epigraph, epi $f^\star$, consists of all pairs (y,r) satisfying $f(x) \geq \langle y,x\rangle - r$ for all points x. Thus epi $f^\star$ 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

$$\begin{aligned} f^\star(y) &= \sup_{x\in\mathbf{E}}\{\langle (y,-1),(x,f(x))\rangle\} \\ &= \sup_{(x,r)\in \operatorname{epi} f}\{\langle (y,-1),(x,r)\rangle\} \\ &= \delta^\star_{\operatorname{epi} f}(y,-1). \end{aligned}$$

Thus the conjugate $f^\star(y)$ is exactly the support function of 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 δ_Q is exactly the support function of Q, thereby explaining the notation $\delta^\star_Q$ 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 $f^\star(y) = \frac{1}{2}\langle A^{-1}y,y\rangle$.

Let us next see what happens when the conjugacy operation is applied twice $f^{\star\star} := (f^\star)^\star$. Let us look first at the simplest example of an affine function.

Exercise 4.32. Show that for any affine function $f(x) = \langle a,x\rangle + b$, we have $f^*(y) = -b + \delta_{\{a\}}(y)$. Deduce the equality $f^{\star\star} = f$.

We will also use the following elementary observation.

Exercise 4.33. For any function $g : \mathbf{E}\times\mathbf{Y} \to \overline{\mathbf{R}}$, 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 4.34 (Biconjugacy)**.** *For any proper convex function $f : \mathbf{E} \to \overline{\mathbf{R}}$, equality $f^{\star\star} = \overline{\text{co}} f$ holds.*

Proof. We begin by successively deducing:

$$\begin{aligned}
(f^\star)^\star(x) &= \sup_y\{\langle x,y\rangle - f^\star(y)\}\\
&= \sup_y\{\langle x,y\rangle - \sup_z\{\langle z,y\rangle - f(z)\}\}\\
&= \sup_y \inf_z\{\langle y, x-z\rangle + f(z)\}\\
&\leq \inf_z \sup_y\{\langle y, x-z\rangle + f(z)\}\\
&= \inf_z \begin{cases} +\infty & x\neq z\\ f(z) & x=z\end{cases}\\
&= f(x).
\end{aligned}$$

The inequality in the fourth line is immediate from Exercise 4.33.

Thus we have established $f^{\star\star}\leq f$. Notice that $f^{\star\star}$ is by definition closed and convex. Hence we deduce from Exercise 4.26, the inequality $f^{\star\star}\leq \overline{\text{co}} f$. 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 $g^\star \geq f^\star$. Taking into account Exercise 4.32 then yields $g=(g^\star)^\star \leq (f^\star)^\star \leq f$. Taking the supremum over all affine minorants g of f yields $\overline{\text{co}} f \leq f^{\star\star}$, 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 $\delta_K^\star = \delta_{K^\circ}$. Consequently, Theorem 4.34 shows

$$\delta_{\text{cl}\,\text{conv}\,K} = \overline{\text{co}}(\delta_K) = (\delta_K)^{\star\star} = \delta^\star_{K^\circ} = \delta_{K^{\circ\circ}}.$$

Hence we deduce $K^{\circ\circ} = \text{cl}\,\text{conv}\,K$. This is exactly the conclusion of Exercise 4.14.

Exercise 4.35. Show the following.

1. If $f,g : \mathbf{E}^n \to \overline{\mathbf{R}}$ are closed proper convex functions, then equalities hold:

$$(f\Box g)^\star = f^\star + g^\star \qquad \text{and} \qquad (f+g)^\star = \text{cl}\,(f^\star \Box g^\star).$$

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

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

4.4 Differential properties

We next turn to differential properties of convex functions.

Definition 4.36 (Subgradients and the Subdifferential)**.** Consider a convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and a point $x \in \mathbf{E}$, with $f(x)$ finite. Then $v \in \mathbf{E}$ is called a subgradient of f at x if the inequality

$$f(y) \geq f(x) + \langle v, y - x\rangle \qquad \text{holds for all } y \in \mathbf{E}.$$

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 $y \mapsto f(x) + \langle v, x - y\rangle$ globally minorizes f. The connection of subdifferentials to epigraphical geometry becomes clear by noting

$$v \in \partial f(x) \qquad \Longleftrightarrow \qquad (v, -1) \in N_{\operatorname{epi} f}(x, f(x)).$$

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

Exercise 4.37. Show that if $f\colon \mathbf{E} \to \mathbf{R}$ is convex and differentiable at a point x, then equality $\partial f(x) = \{\nabla f(x)\}$ holds.

Exercise 4.38 (Existence of subgradients)**.** Consider a proper convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$. Use Theorem 4.30 to show that for any point $x \in \operatorname{ri}\operatorname{dom} f$, 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 4.39. *Consider a convex function $f : \mathbf{E} \to \overline{\mathbf{R}}$ 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 $3 \Rightarrow 1 \Rightarrow 2$ 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 λ 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.

Theorem 4.40 (Fenchel-Young Inequality)**.** *Consider a convex function $f : \mathbf{E} \to \overline{\mathbf{R}}$. Then for any points $x, y \in \mathbf{E}$, the inequality*

$$f(x) + f^{\star}(y) \geq \langle x, y \rangle \qquad \textit{holds,}$$

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

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

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 4.41. *Suppose $f : \mathbf{E} \to \overline{\mathbf{R}}$ is proper, closed, and convex. Then*

$$y \in \partial f(x) \quad \Longleftrightarrow \quad x \in \partial f^{\star}(y).$$

Proof. From Theorem 4.40, we deduce $y \in \partial f(x)$ if and only if

$$\langle x, y \rangle = f(x) + f^{\star}(y).$$

On the other hand by Theorem 4.34, we have $f(x)+f^{\star}(y) = (f^{\star})^{\star}(x)+f^{*}(y)$. Applying Theorem 4.40 again we deduce $y \in \partial f(x)$ if and only if $x \in \partial f^{*}(y)$. □

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.

Definition 4.42 (Directional derivative)**.** Let $f : \mathbf{E} \to \overline{\mathbf{R}}$ 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 4.43. *Suppose $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ 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

$$\begin{aligned}\frac{f(x+\hat{\lambda}y)-f(x)}{\hat{\lambda}} &= \frac{f\left(\left(\frac{\lambda-\hat{\lambda}}{\lambda}\right)x + \frac{\hat{\lambda}}{\lambda}(x+\lambda y)\right) - f(x)}{\hat{\lambda}} \\ &\leq \frac{\frac{\lambda-\hat{\lambda}}{\lambda}f(x) + \frac{\hat{\lambda}}{\lambda}f(x+\lambda y) - f(x)}{\hat{\lambda}} \\ &= \frac{f(x+\lambda y)-f(x)}{\lambda}\end{aligned}$$

The result follows. □

The function $f'(x,\cdot) : \mathbf{R}^n \to \overline{\mathbf{R}}$ 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 4.44. *Consider a proper convex function $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ and fix a point $\bar{x} \in \operatorname{dom} f$. Then $\operatorname{cl} f'(x,\cdot)$ is precisely the support function of the subdifferential $\partial f(x)$.*

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

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

It follows immediately that

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

Conversely suppose $v \notin \partial f(x)$. Hence, there exists a vector z satisfying $f(x+z) < f(x) + \langle v, z\rangle$. Theorem 4.43 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

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

Appealing to Exercise 4.27, the result follows. □

Exercise 4.45. Let $f\colon \mathbf{E} \to \overline{\mathbf{R}}$ be a proper convex function and let Q be any open convex subset of $\operatorname{dom} f$. Prove the identity

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

4.5 Fenchel duality

The idea of duality has appeared throughout the previous sections on convexity, culminating in the definition of the Fenchel conjugate. In this section, we will consider a general class of structured optimization problems

$$(P) \qquad \inf_{x \in \mathbf{E}} \; h(\mathcal{A}x) + g(x),$$

where $h\colon \mathbf{Y} \to \overline{\mathbf{R}}$ and $g\colon \mathbf{E} \to \overline{\mathbf{R}}$ are proper, closed convex functions and $\mathcal{A}\colon \mathbf{E} \to \mathbf{Y}$ is a linear map. This problem is called the primal. Let us define a new convex optimization problem called the *dual*

$$(D) \qquad \sup_{y \in \mathbf{Y}} \; -h^\star(y) - g^\star(-\mathcal{A}^* y),$$

The main theorem of this section shows that under mild conditions, the optimal values of (P) and (D) are equal and are attained; the latter means that the inf and sup are really min and max. Throughout let val(P) and val(D) denote the optimal values of the primal and the dual problems, respectively.

The dual problem (D) arises naturally from a lower-bounding viewpoint. Let us try to find simple lower bounds for val(P). From the Fenchel-Young inequality, we have

$$h^\star(\mathcal{A}x) + h^\star(y) \geq \langle \mathcal{A}x, y\rangle \qquad \text{for all } y \in \mathbf{Y}.$$

Therefore any $y \in \operatorname{dom} h^\star$ yields the lower bound

$$\begin{aligned} \operatorname{val}(P) &\geq \min_{x \in \mathbf{E}} \{-h^\star(y) + \langle \mathcal{A}x, y\rangle + g(x)\} \\ &= -h^\star(y) - \sup_{x \in \mathbf{E}} \{\langle -\mathcal{A}^* y, x\rangle - g(x)\} \\ &= -h^\star(y) - g^\star(-\mathcal{A}^* y). \end{aligned}$$

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

$$\operatorname{val}(P) \geq \operatorname{val}(D).$$

The goal of this section is to show that under mild conditions, equality holds.

The analysis proceeds through a perturbation argument, by embedding our target primal problem in a larger family of optimization problems:

$$p(y) := \min_{x \in \mathbf{E}} \; h(\mathcal{A}x + y) + g(x).$$

The problem $p(0)$ is the primal (P). The variation of the value function $p(\cdot)$ will provide the means to analyze the relationship between (P) and (D).

Let us take a step-back and consider an arbitrary convex function $F\colon \mathbf{E}\times \mathbf{Y} \to \overline{\mathbf{R}}$ and the two families of optimization problems:

$$p(y) := \inf_x \ F(x,y) \qquad \text{and} \qquad q(x) := \sup_y -F^\star(x,y), \tag{4.6}$$

Exercise 4.35 (part 3) yields the equality $p^\star(y) = F^\star(0,y)$ and therefore

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

We will think of $p(0)$ and $q(0)$ as a primal-dual pair. Equality therefore can be understood through biconjugacy.

Theorem 4.46 (Strong duality)**.** *Suppose $F\colon \mathbf{E}\times\mathbf{Y} \to \overline{\mathbf{R}}$ is proper, closed, and convex.*

(a) *The inequality $p(0) \geq q(0)$ always holds.*

(b) *If $p(0)$ is finite, then*

$$\partial p(0) = \operatorname*{argmax}_y \ - F^\star(0,y).$$

Similarly if $q(0)$ is finite, then

$$\partial(-q)(0) = \operatorname*{argmin}_x \ F(x,0).$$

(c) *If $0 \in \mathrm{ri}\,(\mathrm{dom}\, p)$, then equality $p(0) = q(0)$ holds and the supremum $q(0)$ is attained, if finite. Similarly, if $0 \in \mathrm{ri}\,(\mathrm{dom}\,(-q))$, then equality $p(0) = q(0)$ holds and the infimum $p(0)$ is attained, if finite.*

Proof. Part (a) is immediate from the inequality $p(0) \geq p^{\star\star}(0)$. We can next suppose $p(0)$ is finite, since otherwise all the claimed statements are trivially true. By definition, a vector ϕ satisfies $\phi \in \partial p(0)$ if and only if

$$p(0) \leq p(y) - \langle \phi, y\rangle = \inf_x \left\{ F(x,y) - \left\langle \begin{pmatrix} 0 \\ \phi \end{pmatrix}, \begin{pmatrix} x \\ y \end{pmatrix} \right\rangle \right\} \qquad \forall y.$$

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

$$\partial p(0) = \operatorname*{argmax}_y -F^\star(0,y),$$

as claimed. Moreover, the condition $0 \in \mathrm{ri}\,(\mathrm{dom}\, p)$ along with the assumption that $p(0)$ is finite guarantees that p is proper (Exercise 4.24). Hence by Theorem 4.38, the subdifferential $\partial p(0)$ is nonempty, and therefore equality $p(0) = q(0)$ holds and the dual is attained. The symmetric argument for the dual is completely analogous. □

Let us now interpret this theorem for the primal problem (P). Set

$$F(x,y) := h(\mathcal{A}x + y) + g(x)$$

Let us compute the conjugate

$$F^\star(x,y) = \sup_{z,w}\{\langle (z,w),(x,y)\rangle - h(\mathcal{A}z + w) - g(z)\}$$

Making the substitution $v := \mathcal{A}z + w$, we get

$$\begin{aligned} F^\star(x,y) &= \sup_{z,v}\{\langle z,x\rangle + \langle v - \mathcal{A}z, y\rangle - h(v) - g(z)\} \\ &= \sup_{z}\{\langle z, x - \mathcal{A}^* y\rangle - g(z)\} + \sup_{v}\{\langle v,y\rangle - h(v)\} \\ &= g^\star(x - \mathcal{A}^\star y) + h^\star(y). \end{aligned}$$

Thus the Fenchel dual problem (D) is exactly the problem $q(0) = \sup_y F^\star(0,y)$.

4.6 Monotonicity

Definition 4.47. A *set-valued mapping* $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ maps $x \in \mathbb{R}^n$ to a subset $T(x) \subseteq \mathbb{R}^n$. Given $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, define *domain* to be

$$\text{dom } T = \{x \in \mathbb{R}^n : T(x) \neq \emptyset\}$$

and the *graph* of T to be

$$\text{gph } T = \{(x,y) \in \mathbb{R}^n \times \mathbb{R}^n : y \in T(x)\}.$$

A mapping $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is *monotone* if $\langle y_1 - y_2, x_1 - x_2\rangle \geq 0$ for any $x_1, x_2 \in \mathbb{R}^n, y_1 \in T(x_1), y_2 \in T(x_2)$.

Example 4.48. Given a convex function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, the mapping $\partial f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a set value mapping.

Proposition 4.49. *If f is convex then ∂f is monotone.*

Proof. Suppose $y_1 \in \partial f(x_1)$, $y_2 \in \partial f(x_2)$, then $f(x_2) \geq f(x_1) + \langle y_1, x_2 - x_1\rangle$ and $f(x_1) \geq f(x_2) + \langle y_2, x_1 - x_2\rangle$. By adding the equations we know $\langle y_1 - y_2, x_1 - x_2\rangle \geq 0$. □ □

Theorem 4.50. *If $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is closed and convex, then ∂f is maximal monotone.*

Proof. See Rockafellar. □

□

Chapter 5

Nonsmooth Convex Optimization

5.1 Subgradient methods

To solve our problem, we introduce the **subgradient algorithm**, which has two steps per iteration:

1. Get $v_k \in \partial f(x_k)$
2. Update $x_{k+1} = \text{proj}_Q(x_k - \alpha_k v_k)$.

Later we will see how to choose the step lengths α_k appropriately.

Question: What will happen if we're at a minimum?
Answer: We can move away!! Although 0 belongs to the subdifferential at the minimum, we might select a nonzero element in the subdifferential for the search direction, and move away from the minimum. As a result, the method will not be monotone. In the smooth case, the search directions v_k tend to 0, so the step lengths α_k don't need to. In the general case, we must ensure the step lengths tend to 0 so the algorithm doesn't bounce around indefinitely.

Analysis of Subgradient Method
All guarantees will be on function value (compare with strong convexity later.) The elements v_k define affine minorants for f, which give lower bounds on the optimal value $f(\bar{x})$. The idea is to close the gap. Define

$$\hat{l}_k = f(x_k) + \langle v_k \rangle \bar{x} - x_k,$$

and note $\hat{l}_k \leq f(\bar{x})$. We also have the following lower bounds on the optimal value, which use information up to the current iterate k:

$$l_k = \sum_{i=1}^{k} \frac{\alpha_i}{A_k} \hat{l}_i,$$

where $A_k = \sum_{i=1}^{k} \alpha_i$. (Note, neither $\hat{l}_k$ or l_k are computable and thus cannot be used as stopping criteria.) We show

$$\min\nolimits_{i=1,\ldots,k} f(x_i) - l_k \to 0,$$

under an appropriate choice of α_k. Observe

$$\begin{aligned} 0 \leq \|x_{k+1} - \bar{x}\|^2 &= \left\|\operatorname{proj}_Q(x_k - \alpha_k v_k) - \bar{x}\right\|^2 \\ &\leq \|x_k - \alpha_k v_k - \bar{x}\|^2 \\ &= \|x_k - \bar{x}\|^2 + 2\alpha_k \langle \bar{x} - x_k \rangle v_k + \alpha_k^2 \|v_k\|^2 \\ &\leq \|x_1 - \bar{x}\|^2 + \sum_{i=1}^{k} 2\alpha_i \langle v_i \rangle \bar{x} - x_i + L^2 \sum_{i=1}^{k} \alpha_i^2. \end{aligned}$$

Thus

$$\begin{aligned} 0 &\leq \min_{i=1,\ldots,k} f(x_i) - l_k \\ &\leq \sum_{i=1}^{k} \frac{\alpha_i}{A_k} f(x_i) - l_k \\ &= \sum_{i=1}^{k} \frac{\alpha_i}{A_k} \langle v_i \rangle x_i - \bar{x} \\ &\leq \frac{\|x_1 - \bar{x}\|^2 + L^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i}. \end{aligned}$$

We get convergence if the steps α_i are square-summable, but not summable, e.g., $\alpha_i = \frac{1}{i}$. If we run the algorithm for a fixed number of steps k, we can optimize over α_i in the previous bound to get the best choice of step lengths. More on this in the next section.

Last time: Consider the problem $\min\{f(x) \mid x \in Q\}$, where f and Q are closed and convex, and f is L-Lipschitz. Additionally, $\operatorname{proj}_Q(x)$ is computable. The subgradient method, for each k, obtains an arbitrary $v_k \in \partial f(x_k)$ and defines $x_{k+1} := \mathrm{proj}_Q(x_k - \alpha_k v_k)$. Then, we proved

$$\min_{i=1\ldots k} f(x_i) - f^* \leq \frac{\|x_i - x^*\|^2 + L^2 \sum\limits_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i}.$$

Conclusion: If $\sum\limits_{i=1}^{\infty} \alpha_i^2 < \infty$, but $\sum\limits_{i=1}^{\infty} \alpha_i = \infty$, then $\min\limits_{i=1\ldots k} f(x_i) \to f^*$.

Suppose we only run the algorithm up to iteration N. Minimizing the right hand side over $\alpha_1, \ldots, \alpha_N$ yields

$$\alpha_1 = \alpha_2 = \cdots = \alpha_N = \frac{\|x_1 - x^*\|}{L\sqrt{N}}.$$

So

$$f_{\mathrm{best}}^{(N)} - f^* \leq \frac{\|x_1 - x^*\|L}{\sqrt{N}}.$$

If we want $f_{\mathrm{best}}^{N} - f^* \leq \varepsilon$, we can be done in $N = \dfrac{\|x_1 - x^*\|^2 L^2}{\varepsilon^2}$ steps. What if we know f^*? Can the algorithm at least empirically be improved? Recall the inequality we used

$$\begin{aligned} 0 \leq \|x_{k+1} - x^*\| &\leq \|x_k - x^*\|^2 + 2\alpha_k \langle v_k, x^* - x_k\rangle + \alpha_k^2 \|v_k\|^2 \\ &\leq \|x_k - x^*\|^2 + 2\alpha_k (f^* - f(x_k)) + \alpha_k^2 \|v_k\|^2. \end{aligned}$$

If f^* is known, then the right hand side can be minimized in α_k yielding $\alpha_k = \dfrac{f(x_k) - f^*}{\|v_k\|^2}$.

Then, plugging these values back into the bound we get

$$\begin{aligned} 0 &\leq \|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \frac{(f^* - f(x_k))^2}{\|v_k\|^2} \\ &\leq \|x_1 - x^*\|^2 - \frac{1}{L^2} \sum_{i=1}^{k} (f^* - f(x_i))^2. \end{aligned}$$

We conclude as before

$$\min_{i=1\ldots k} f(x_i) - f^* \leq \frac{\|x^* - x_1\|L}{\sqrt{k}}.$$

Next we analyze if the convergence rate of the subgradient method is "optimal" among a large class of algorithms.

Problem Class:

- Convex function, f.
- Starting point x_0 with $\|x_0 - x^*\| \leq R$, for some R.
- Lipschitz constant L of f on $B(X^*, R)$.
- Oracle defining f: given x, returns $f(x)$ and some $v \in \partial f(x)$.

Algorithm Class:

- $x_k \in x_0 + \text{span}\{v_1, v_2, \ldots, v_{k-1}\}$

Test Problem: Minimize

$$f(x) = \min_{i=1\ldots k} x_i + \frac{1}{2}\|x\|^2,$$

where $x_0 = 0$.

Solution: Let $x^* = -\frac{1}{k}(1, 1, \ldots, 1, 0, 0, \ldots, 0)$. Then, $\partial f(x^*) = \text{conv}\{e_i\}_{i=1\ldots k} + x^*$. So, $0 = \sum_{i=1}^{k} \frac{1}{k} e_i + x^* \in \partial f(x^*)$, verifying that x^* is a minimizer (a unique one in fact). Here $R = \|x_0 - x^*\| = \dfrac{1}{\sqrt{k}}$ and $L = 1 + \dfrac{1}{\sqrt{k}}$.

Oracle: (Resistance Oracle) Given x, the oracle returns $f(x)$ and the subgradient $e_{\hat{\jmath}} + x$, where $\hat{\jmath} = \min\{j | x_j = \max_{i=1\ldots k} x_i\}$. Then for $i = 0, 1, \ldots, n-k$, the entries $(x_k)_{k+i} = 0$. So,

$$f_{\text{best}}^{k-1} - f^* \geq -f^* = \frac{1}{2k} = \frac{RL}{2\sqrt{1+k}}, \text{ for } k < n.$$

This rate has the same order of growth in R,L, and k as the rate for the subgradient method. In this sense, the subgradient method is optimal within the problem class above.

Last time, we saw the estimate for subgradient method was

$$f_k^{\text{best}} - f^* \leq \frac{\|x_1 - x^*\| L}{\sqrt{k}}$$

for the problem $\min\{f(x) : x \in Q\}$ and this was optimal in $(\|x - x^*\|, L, k)$. Be aware of the dependence on k and L. In this lecture, we see under what additional conditions on our function f can we achieve better convergence estimates.

Definition 5.1 ($\boldsymbol{\alpha}$-Strongly Convex)**.** A function $f : \mathbf{R}^n \to \bar{\mathbf{R}}$ is **$\boldsymbol{\alpha}$-strongly convex** ($\alpha \geq 0$) if

$$f(y) \geq f(x) + \langle v, y - x \rangle + \frac{\alpha}{2}\|y - x\|_2^2, \qquad \text{for all } x, y \text{ and } v \in \partial f(x)\ .$$

Observe that this is equivalent to saying that your function f is bounded below by a convex quadratic. If $\alpha = 0$, this is the same as f being convex.

Remark 5.2. The following are easy to verify

(1) A function f is α-strong convex if and only if $x \mapsto f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex.

(2) If f is C^2-smooth, then f is α-strongly convex if and only if $\lambda_{\min}(\nabla^2 f(x)) \geq \alpha$, or equivalently $\nabla^2 f(x) \succeq \alpha I$.

(3) If f is α-strongly convex and g is β-strongly convex convex, then $f + g$ is $(\alpha + \beta)$-strongly convex.

Question: Can we improve the convergence of the subgradient method if we assume f is α-convex ($\alpha > 0$)?
Answer: Yes!

Let's begin by analyzing the projected subgradient method for α-strongly convex functions. Recall, the algorithm is given by

$$v_k \in \partial f(x_k)$$

$$x_{k+1} = \text{proj}_Q(x_k + t_k v_k)$$

Proposition 5.3. *Under the projected subgradient method with* $f : \mathbf{R}^n \to \mathbf{R}$ *α-strongly convex and Lipschitz with constant L, we have*

$$f_k^{best} - f^* \leq \frac{2L^2}{\alpha(k+1)}$$

Proof. By using that the projection is Lipschitz with constant 1 and f is Lipschitz with constant L, plugging in the definition of x_{k+1} (see last lecture), we get

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &\leq \|x_k - x^*\|^2 + 2t_k\langle v_k, x^* - x_k\rangle + t_k^2 L^2 \\ &\leq \|x_k - x^*\|_2^2 + 2t_k\big(f(x^*) - f(x_k) - \tfrac{\alpha}{2}\|x^* - x_k\|_2^2\big) + t_k^2 L^2.\end{aligned}$$

Rewriting the expression yields

$$\begin{aligned}2t_k\big(f(x_k) - f(x^*)\big) &\leq \big(1 - \alpha t_k\big)\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 + t_k^2 L^2 \\ \Rightarrow \quad f(x_k) - f(x^*) &\leq \left(\frac{1 - \alpha t_k}{2t_k}\right)\|x_k - x^*\|_2^2 - \frac{1}{2t_k}\|x_{k+1} - x^*\|_2^2 + \frac{t_k}{2}L^2.\end{aligned}$$

Set $t_k := \frac{2}{\alpha(k+1)}$. (If you work really hard, you will get that this is the t_k is you want!) Plugging this value in and multiplying both sides by k, we get

$$k\big(f(x_k) - f(x^*)\big) \leq \frac{\alpha k(k-1)}{4}\|x_k - x^*\|_2^2 - \frac{\alpha k(k+1)}{4}\|x_{k+1} - x^*\|_2^2 + \frac{k}{\alpha(k+1)}L^2.$$

Summing up,

$$\sum_{i=1}^{k} i\big(f(x_i) - f(x^*)\big) \le \sum_{i=1}^{k} \frac{\alpha\, i(i-1)}{4} \|x_i - x^*\|_2^2 - \frac{\alpha\, i(i+1)}{4} \|x_{i+1} - x^*\|_2^2 + \sum_{i=1}^{k} \frac{i}{\alpha(i+1)} L^2.$$

The summand $\sum_{i=1}^{k} \frac{\alpha\, i(i-1)}{4} \|x_i - x^*\|_2^2 - \frac{\alpha\, i(i+1)}{4} \|x_{i+1} - x^*\|_2^2$ telescopes. Moreover we can bound the second summand by using $\frac{i}{i+1} \le 1$. Hence,

$$\sum_{i=1}^{k} i\big(f(x_i) - f(x^*)\big) \le \frac{kL^2}{\alpha}.$$

Therefore, we deduce that

$$f_k^{\text{best}} - f(x^*) \le \frac{2L^2}{\alpha(k+1)}.$$

□

Remark 5.4. This bound is optimal. You can see this by adjusting the function we had used in the previous lecture.

Key inequality

$$f\left(\frac{1}{k}\sum_{i=1}^{k} x_i\right) - f(x^*) \le \frac{1}{k}\left(\sum_{i=1}^{k} f(x_i) - f(x^*)\right) \le \frac{1}{k}\sum_{i=1}^{k} \langle \nabla f(x_i), x_i - x^* \rangle.$$

Dual averaging. Problem

$$\min_{x \in Q} \; f(x).$$

Define the linearization $l(y; x) = f(x) + \langle \nabla f(x), y - x \rangle$
Method

$$x_{t+1} = \operatorname*{argmin}_{x} \; \frac{1}{k}\sum_{i=1}^{k} l(x, x_i) + \tfrac{1}{2t_k}\|x - x_0\|^2.$$

Set

$$\phi_k(x) := \sum_{i=1}^{k} \langle \nabla f(x_i), x \rangle + \tfrac{k}{2t_k}\|x - x_0\|^2$$

and note

$$\phi_k(x) = \phi_{k-1}(x) + \langle \nabla f(x_k), x \rangle + \left(\tfrac{k}{2t_k} - \tfrac{k-1}{2t_{k-1}}\right) \|x - x_0\|^2.$$

Then

$$\phi_k(x_{k+1}) - \phi_k(x) \le -\tfrac{k}{2t_k}\|x_{k+1} - x\|^2.$$

Hence

$$\begin{aligned}\phi_k(x_{k+1}) - \phi_k(x) &= \phi_{k-1}(x_{k+1}) - \phi_{k-1}(x) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle \\ &+ \left(\tfrac{k}{2t_k} - \tfrac{k-1}{2t_{k-1}}\right)(\|x_{k+1} - x_0\|^2 - \|x - x_0\|^2) \\ &\geq \tfrac{k-1}{2t_{k-1}}\|x_{k+1} - x_k\|^2 + \langle \nabla f(x_k), x_{k+1} - x \rangle \\ &+ \left(\tfrac{k}{2t_k} - \tfrac{k-1}{2t_{k-1}}\right)(\|x_{k+1} - x_0\|^2 - \|x - x_0\|^2)\end{aligned}$$

So

$$-\left(\tfrac{k-1}{2t_{k-1}} + \tfrac{k}{2t_k}\right)\|x_{k+1} - x_k\|^2 + \langle \nabla f(x_k), x_k - x_{k+1} \rangle \geq \left(\tfrac{k}{2t_k} - \tfrac{k-1}{2t_{k-1}}\right)(\|x_{k+1} - x_0\|^2 - \|x_k - x_0\|^2)$$

Expanding the left-hand-side, we deduce

$$\frac{1}{k}\sum_{i=1}^{k}\langle \nabla f(x_i), x_{k+1} - x \rangle \leq \frac{1}{2t_k}(\|x - x_0\|^2 - \|x_{k+1} - x_0\|^2 - \|x_{k+1} - x\|^2).$$

Observe now the equality

$$\phi_k(x_{k+1}) - \phi_k(x) = \phi_{k-1}(x_{k+1}) - \phi_{k-1}(x)$$