Course notes: Convex Analysis and Optimization

Dmitriy Drusvyatskiy

April 12, 2019

Contents

1 Review of Fundamentals 1

1.1 Inner products and linear maps 1

1.2 Norms 2

1.3 Eigenvalue and singular value decompositions of matrices . . 4

1.4 Point-set topology and differentiability 4

1.5 Fundamental theorems of calculus & accuracy in approximation 8

2 Smooth minimization 13

2.1 Optimality conditions: Smooth Unconstrained 13

2.2 Convexity, a first look 15

2.3 Rates of convergence 19

2.4 Two basic methods 20

2.4.1 Majorization view of gradient descent 21

2.4.2 Newton's method 25

2.5 Computational complexity for smooth convex minimization . 26

2.6 Conjugate Gradient Method 29

2.7 Optimal methods for smooth convex minimization 34

2.7.1 Fast gradient methods 34

2.7.2 Fast gradient methods through estimate sequences . . 40

2.7.3 Optimal quadratic averaging 46

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* $\|X\|_F = \sqrt{\mathrm{tr}\,(X^TX)}$.

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

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

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

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

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:

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

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 $\|v\|_{\mathcal{A}} = \sqrt{\langle \mathcal{A}v, v\rangle}$. Show that the dual norm with respect to the original inner product is $\|v\|^*_{\mathcal{A}} = \|v\|_{\mathcal{A}^{-1}} = \sqrt{\langle \mathcal{A}^{-1}v, v\rangle}$.

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 $\alpha, \beta \geq 0$ satisfying

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

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

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

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*

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

The reader should verify the inequality

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

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

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

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 $m \leq n$, 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. Using the property $\mathrm{tr}\,(AB) = \mathrm{tr}\,(BA)$, prove $\nabla F(X) = -X^{-1}$ and $\nabla^2 F(X)[V] = X^{-1}VX^{-1}$ for any $X \succ 0$. [**Hint:** To compute $\nabla F(X)$, justify
$$f(X+tV)-f(X)+t\langle X^{-1}, V\rangle = -\ln\det(I+X^{-1/2}VX^{-1/2})+\mathrm{tr}\,(X^{-1/2}VX^{-1/2}).$$
By rewriting the expression in terms of eigenvalues of $X^{-1/2}VX^{-1/2}$, deduce that the right-hand-side is $o(t)$. To compute the Hessian, observe
$$(X+V)^{-1} = X^{-1/2}\left(I + X^{-1/2}VX^{-1/2}\right)^{-1} X^{-1/2},$$
and then use the expansion
$$(I+A)^{-1} = I - A + A^2 - A^3 + \ldots = I - A + O(\|A\|^2_{op}),$$
whenever $\|A\|_{op} < 1$.]

3. Show
$$\langle \nabla^2 F(X)[V], V\rangle = \|X^{-\frac{1}{2}} V X^{-\frac{1}{2}}\|_F^2$$
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 $f\colon U \to \mathbf{R}$ 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*

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

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

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

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 $\int_0^1\gamma(t)\,dt=\Big(\int_0^1\gamma_1(t)\,dt,\ldots,\int_0^1\gamma_m(t)\,dt\Big)$, where γ_i are the coordinate functions of γ. The main observation is that whenever γ_i are integrable, the inequality

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

To see this, define $w=\int_0^1\gamma(t)\,dt$ and simply observe

$$\|w\|^2=\int_0^1\langle\gamma(t),w\rangle\,dt\le\|w\|\int_0^1\|\gamma(t)\|\,dt.$$

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:

$$\|F(y)-F(x)-\nabla F(x)(y-x)\|\le\frac{\beta}{2}\|y-x\|^2.$$

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 $x \in U$.

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.

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 convex neighborhood U of x such that $f(x) \leq f(y)$ for all $y \in U$. After all, gradients and Hessians provide only local information on the function.

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

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

Theorem 2.1. *(First-order necessary conditions) Suppose that x is a local minimizer of a function $f : U \to \mathbf{R}$. 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 2.1 shows that all local minimizers of f are critical points. In general, even finding local minimizers is too ambitious, and we will for the most part settle for critical points.

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

Theorem 2.2. *(Second-order conditions)*
Consider a C^2-smooth function $f\colon U \to \mathbf{R}$ 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*

$$\nabla f(x) = 0 \quad \textit{and} \quad \nabla^2 f(x) \succeq 0.$$

2. *(Sufficient conditions) If the relations*

$$\nabla f(x) = 0 \quad \textit{and} \quad \nabla^2 f(x) \succ 0$$

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 2.1 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 2.2) is merely to determine whether a putative point x is a local minimizer of a smooth function f. Such a viewpoint is far too limited.

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

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

2.2 Convexity, a first look

Finding a global minimizer of a general smooth function $f\colon \mathbf{E} \to \mathbf{R}$ 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 2.3 (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 2.4. 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 $\mathbf{E} \times \mathbf{R}$.

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

$$f\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leq \sum_{i=1}^{k} \lambda_i f(x_i),$$

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

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

Exercise 2.6. Consider an arbitrary set T and a family of convex functions $f_t \colon U \to (-\infty, +\infty]$ 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 2.7 (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 2.6), we deduce that f is convex, establishing (a).

Suppose now that f is C^2-smooth. Then for any fixed $x \in U$ and $h \in \mathbf{E}$, 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 2.8. Show that the functions f and F in Exercise 1.10 are convex.

Exercise 2.9. 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. $0 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{\beta}{2}\|x - y\|^2$
2. $f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y)$
3. $\frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle$

4. $0 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle \leq \beta \|x - y\|^2$

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

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

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

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

Corollary 2.10 (Minimizers of convex functions)**.** *For any C^1-smooth convex function $f\colon U \to \mathbf{R}$ 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 2.7. □

Exercise 2.11. 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 2.7 in a natural ways yields an important subclass of convex functions. These are the functions for which numerical methods have a chance of converging at least linearly.

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

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

Figure 2.1 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:

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

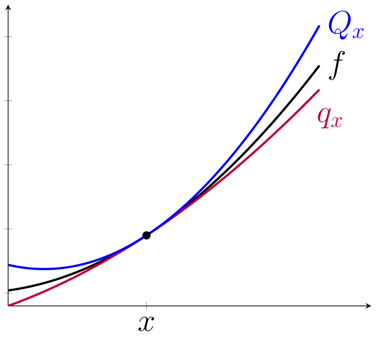

Figure 2.1: Illustration of a β-smooth and α-strongly convex function f, where $Q_x(y) := f(x) + \langle \nabla f(x), y - x\rangle + \frac{\beta}{2}\|y - x\|^2$ is an upper models based at x and $q_x(y) := f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y - x\|^2$ is a lower model based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

Exercise 2.13. 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 2.7 for strongly convex functions.

Theorem 2.14 (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 2.13, property (a) holds if and only if $f - \frac{\alpha}{2}\|\cdot\|^2$ is convex, which by Theorem 2.7, is equivalent to (b). Suppose now that f is C^2-smooth. Theorem 2.7 then shows that $f - \frac{\alpha}{2}\|\cdot\|^2$ is convex if and only if (c) holds. □

2.3 Rates of convergence

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

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

1. We will say that a_k converges *sublinearly* if there exist constants $c, q > 0$ satisfying
$$a_k \leq \frac{c}{k^q} \qquad \text{for all } k.$$
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
$$c \cdot (1-q)^k \leq \varepsilon \quad \iff \quad k \geq \frac{-1}{\ln(1-q)} \ln\left(\frac{c}{\varepsilon}\right).$$
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.

2.4 Two basic methods

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

2.4.1 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:

$$m_k(x) := f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{\beta}{2}\|x - x_k\|^2. \tag{2.1}$$

Clearly m_k agrees with f at x_k, while Corollary 1.14 shows that the inequality $m_k(x) \geq f(x)$ holds for all $x \in \mathbf{E}$, 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

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \langle \nabla f(x_k), \beta^{-1}\nabla f(x_k)\rangle + \frac{\beta}{2}\|\beta^{-1}\nabla f(x_k)\|^2 \\ &= f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|^2. \end{aligned} \tag{2.2}$$

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 2.15 (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 2.16 (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

$$\min_{i=1,\ldots k} \|\nabla f(x_i)\| \leq \frac{2\beta\|x_0 - x^*\|}{k}. \tag{2.3}$$

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

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

We conclude

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

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 2.17 (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

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

Thus we deduce

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

Rearranging yields

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

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, $1 - \frac{2}{Q+1} \le 1 - Q^{-1}$, 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 2.3, we can be sure that the inequality $\|x_k - x^*\|^2 \le \varepsilon$ holds after $k \ge \frac{Q+1}{2}\ln\left(\frac{\|x_0 - x^*\|^2}{\varepsilon}\right)$ iterations.

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

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

for some sequence $\alpha_k \ge 0$.

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

$$\frac{1}{2}\|x_{k+1} - x^*\|^2 \le \frac{1}{2}\|x_k - x^*\|^2 - \gamma_k(f(x_k) - f(x^*)) + \frac{\gamma_k^2}{2}\|\nabla f(x_k)\|^2. \tag{2.4}$$

2. Supposing that you know the minimal value f^* of f, show that the sequence $\gamma_k = \frac{f(x_k) - f^*}{\|\nabla f(x_k)\|^2}$ minimizes the right-hand-side of (2.4) in γ, thereby yielding the guarantee

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

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

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

If f is in addition α-strongly convex, derive the guarantee

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

2.4.2 Newton's method

In this section we consider Newton's method, an algorithm much different from gradient descent. Consider the problem of minimizing a C^2-smooth function $f\colon \mathbf{E} \to \mathbf{R}$. 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 2.19 (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*

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

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 2.19, 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.

2.5 Computational complexity for smooth convex minimization

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

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

Definition 2.20 (Linearly-expanding first-order method)**.** An algorithm is called a *linearly-expanding first-order method* if when applied to any 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 2.21 (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

$$f(x_k) - f^* \geq \frac{3\beta\|x_0 - x^*\|^2}{32(k+1)^2}, \tag{2.5}$$

$$\|x_k - x^*\|^2 \geq \tfrac{1}{8}\|x_0 - x^*\|^2, \tag{2.6}$$

where x^* is any minimizer of f.

For simplicity, we will only prove the bound on functional values (2.5). 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

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

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

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

and therefore

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

Exercise 2.22. Establish the following properties of f_k.

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

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

   with optimal value

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

2. Taking into account the standard inequalities,

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

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

3. Fix indices $1 < i < j < n$ and a point $x \in \mathbf{R}^i \times \{0\}^{n-i}$. 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 2.21 is now easy. Fix k and apply the linearly-expanding first order method to $f := f_{2k+1}$ staring at $x_0 = 0$. Let x^* be the minimizer of f and f^* the minimum of f. By Exercise 2.22 (part 3), the iterate x_k lies in $\mathbf{R}^k \times \{0\}^{n-k}$. 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 2.22, 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 2.21 do not depend on strong convexity constants. When the target function class consists of β-smooth strongly convex functions, the analogous complexity bounds become

$$f(x_k) - f^* \geq \left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{2k} \|x_0 - x^*\|^2, \tag{2.7}$$

$$\|x_k - x^*\|^2 \geq \frac{\alpha}{2}\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{2k} \|x_0 - x^*\|^2, \tag{2.8}$$

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

Let us now compare efficiency estimates of gradient descent with the lower-complexity bounds we have just discovered. Consider a 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 2.16, 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 (2.5) 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 (2.6) 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 2.16 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 (2.8), 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".

2.6 Conjugate Gradient Method

Before describing optimal first-order methods for general smooth convex minimization, it is instructive to look for inspiration at the primordial subclass of smooth optimization problems. We will consider minimizing strongly convex quadratics. For this class, the conjugate gradient method – well-known in numerical analysis literature – achieves rates that match the worst-case bound (2.7) 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 2.23. 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 2.24. 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

$$r_{k+1} = r_k - t_k A v_k. \tag{2.9}$$

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

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

$$\langle r_{k+1}, v_i \rangle = 0 \qquad \textit{holds for all } i = 1, \ldots, k \tag{2.10}$$

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 (2.10) holds with k replaced by $k-1$. Taking into account the recursion (2.9) and Exercise 2.24, 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 (2.10). Now clearly x_{k+1} lies in $x_1 + \operatorname{span}\{v_1, \ldots v_k\}$. On the other hand, equation (2.10) 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 2.26. *The conjugate direction method finds* x^* *after at most* n *iterations.*

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

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

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 2.25 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 2.27. *Consider the iterates x_k generated by the conjugate gradient method. Supposing $x_k \neq x^*$, we have*

$$\langle r_k, r_i\rangle = 0 \qquad \textit{for all} \quad i = 0, 1, \ldots, k-1, \tag{2.11}$$

$$\langle v_k, v_i\rangle_A = 0 \qquad \textit{for all} \quad i = 0, 1, \ldots, k-1, \tag{2.12}$$

and

$$\operatorname{span}\{r_0, r_1, \ldots, r_k\} = \operatorname{span}\{v_0, v_1, \ldots, v_k\} = \mathcal{K}_k(r_0). \tag{2.13}$$

Proof. We have already proved equation (2.12), as this was the motivation for the conjugate gradient method. Equation (2.11) follows by observing the inclusion $r_i \in \operatorname{span}\{v_i, v_{i-1}\}$ and appealing to Theorem 2.25. We prove the final claim (2.13) 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

$$\operatorname{span}\{r_0, r_1, \ldots, r_{k+1}\} \subseteq \operatorname{span}\{v_0, v_1, \ldots, v_{k+1}\} \tag{2.14}$$

holds since r_i lie in $\operatorname{span}\{v_i, v_{i-1}\}$. Taking into account the induction assumption, we deduce $v_{k+1} \in \operatorname{span}\{r_{k+1}, v_k\} \subseteq \operatorname{span}\{r_0, r_1, \ldots, r_{k+1}\}$. Hence equality holds in (2.14).

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

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

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

$$f(x_{k+1}) - f^* \leq \frac{1}{2}\|x_0 - x^*\|_A^2 \cdot \max_{i=1,\ldots,n} q(\lambda_i)^2 \qquad (2.15)$$

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

Proof. Let $\gamma_1, \ldots, \gamma_m$ be the distinct eigenvalues of A and define the degree m polynomial $q(\lambda) := \frac{(-1)^m}{\gamma_1 \cdots \gamma_m}(\lambda - \gamma_1) \cdots (\lambda - \gamma_m)$. Observe $q(0) = 1$. Moreover, clearly equality $0 = q(\gamma_i)$ holds for all indices i. Inequality (2.15) 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 (2.15) 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 2.2 for an illustration.

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

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

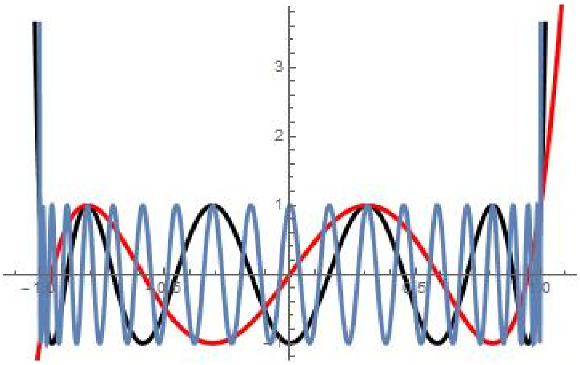

Figure 2.2: T_5, T_{10}, T_{40} are shown in red, black, and violet, respectively, on the interval $[-1, 1]$.

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 (2.15), 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 (2.7).

2.7 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 (2.5) and (2.7).

2.7.1 Fast gradient methods

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

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

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

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 2.16). 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

$$\begin{aligned}\frac{1}{a_k^2}(f(x_{k+1})-f^*) \leq \frac{1-a_k}{a_k^2}&(f(x_k)-f^*)\\ &+\frac{\beta}{2}\Big(\|x^*-[x_k-a_k^{-1}(x_k-y_k)]\|^2 \\ &\qquad -\|x^*-[x_k-a_k^{-1}(x_k-x_{k+1})]\|^2\Big).\end{aligned} \tag{2.16}$$

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

$$z_{k+1} := x_k - a_k^{-1}(x_k - x_{k+1}). \tag{2.17}$$

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 (2.17) 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 (2.16) becomes

$$\frac{1}{a_k^2}(f(x_{k+1})-f^*) + \frac{\beta}{2}\|x^*-z_{k+1}\|^2 \leq \frac{1-a_k}{a_k^2}(f(x_k)-f^*) + \frac{\beta}{2}\|x^*-z_k\|^2. \tag{2.18}$$

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

$$\frac{1-a_k}{a_k^2} \leq \frac{1}{a_{k-1}^2}. \tag{2.19}$$

Then the right-hand-side of (2.18) is upper-bounded by the same term as the left-hand-side with k replaced by $k-1$. Iterating the recurrence (2.18) yields

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

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

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

Looking back at (2.19), the choices $a_k = \frac{2}{k+2}$ are valid, and will yield the efficiency estimate

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

Thus the scheme is indeed optimal for minimizing β-smooth convex functions, since this estimate matches the lower complexity bound (2.6). A slightly faster rate will occur when choosing $a_k \in (0,1]$ to satisfy (2.19) with equality, meaning

$$a_{k+1} = \frac{\sqrt{a_k^4 + 4a_k^2} - a_k^2}{2}. \tag{2.20}$$

Exercise 2.30. Suppose $a_0 = 1$ and a_k is given by (2.20) for each index $k \geq 1$. Using induction, establish the bound $a_k \leq \frac{2}{k+2}$, for each $k \geq 0$.

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

Algorithm 2: Fast gradient method for smooth convex minimization

```
Input: Starting point $x_0 \in \mathbf{E}$.
Set $k = 0$ and $a_0 = a_{-1} = 1$;
for $k = 0, \ldots, K$ do
    Set
        $$y_k = x_k + a_k(a_{k-1}^{-1} - 1)(x_k - x_{k-1})$$
        $$x_{k+1} = y_k - \frac{1}{\beta}\nabla f(y_k) \qquad (2.21)$$
    Choose $a_{k+1} \in (0,1)$ satisfying
        $$\frac{1 - a_{k+1}}{a_{k+1}^2} \leq \frac{1}{a_k^2}. \qquad (2.22)$$
    $k \leftarrow k + 1$.
end
```

Theorem 2.31 (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*

$$f(x_k) - f^* \leq \frac{2\beta\|x^* - x_0\|^2}{(k+1)^2}. \tag{2.23}$$

Let us next analyze the rate at which Algorithm 2 forces the gradient to tend to zero. One can try to apply the same reasoning as in the proof of Theorem 2.16. One immediately runs into a difficulty, however, namely there is no clear relationship between the values $f(y_k)$ and $f(x_k)$. This difficulty can be overcome by introducing an extra gradient step in the scheme. A simpler approach is to take slightly shorter gradient steps in (2.21).

Theorem 2.32 (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* (2.21) *by $x_{k+1} = y_k - \frac{1}{2\beta}\nabla f(y_k)$. Then the iterates generated by the algorithm satisfy*

$$f(x_k) - f^* \leq \frac{4\beta\|x^* - x_0\|^2}{(k+1)^2}, \tag{2.24}$$

$$\min_{i=1,\ldots,k} \|\nabla f(y_i)\| \leq \frac{8\sqrt{3}\cdot\beta\|x^* - x_0\|}{\sqrt{k(k+1)(2k+1)}}. \tag{2.25}$$

Proof. The proof is a slight modification of the argument outlined above of Theorem 2.31. 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

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

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 (2.24). 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 (2.25). □

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 (2.22). The argument behind the resulting scheme relies on a different algebraic technique called *estimate sequences.* This technique is more intricate and more general than the arguments we outlined for sublinear rates of convergence. We will explain this technique in Section 2.7.2.

There is, however, a different approach to get a fast linearly convergent method simply by periodically restarting Algorithm 2. Let $f\colon \mathbf{E} \to \mathbf{R}$ 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 2.31 along with strong convexity yields the guarantee

$$f(x_k^i) - f^* \leq \frac{2\beta\|x^* - x_0^i\|^2}{(k+1)^2} \leq \frac{4\beta}{\alpha(k+1)^2}(f(x_0^i) - f^*). \tag{2.26}$$

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).

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** $i = 0, \ldots, K$ **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**
```

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

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

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 (2.7) for smooth strongly convex minimization.

2.7.2 Fast gradient methods through estimate sequences

In this section, we describe an algebraic technique for designing fast gradient method for minimizing a 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 2.33 (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

$$\phi_k(x) \leq (1 - \lambda_k)f(x) + \lambda_k \phi_0(x) \tag{2.27}$$

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

$$f(x_k) - f^* \leq \lambda_k(\phi_0^* - f^*). \tag{2.28}$$

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 2.34. 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 (2.27). (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 2.35. 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 2.34 preserve the same form:

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

where

$$\begin{aligned}
\gamma_{k+1} &= (1-t_k)\gamma_k + t_k\alpha, \\
v_{k+1} &= \frac{1}{\gamma_{k+1}}\left[(1-t_k)\gamma_k v_k + t_k\alpha y_k - t_k\nabla f(y_k)\right], \\
\phi_{k+1}^* &= (1-t_k)\phi_k^* + t_k f(y_k) - \frac{t_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|^2 \\
&\quad + \frac{t_k(1-t_k)\gamma_k}{\gamma_{k+1}}\Big(\frac{\alpha}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k\rangle\Big). \qquad (2.29)
\end{aligned}$$

Now having available an estimate sequence constructed above, let's try to find the sequence $\{x_k\}$ satisfying $f(x_k) \leq \phi_k^*$. Suppose we already have available a point x_k satisfying this condition; let us see how to choose x_{k+1}. Lowerbounding the term $\|y_k - v_k\|$ in (2.29) by zero, we deduce

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

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

$$\phi_{k+1}^* \geq \Big(f(y_k) - \frac{t_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|^2\Big) + (1-t_k)\langle \nabla f(y_k), \frac{t_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k\rangle.$$

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

$$t_k^2\beta = \gamma_{k+1} = (1-t_k)\gamma_k + t_k\alpha,$$

and set

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

We then deduce

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

Finally let us ensure

$$\frac{t_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k = 0,$$

by setting

$$y_k = \frac{t_k \gamma_k v_k + \gamma_{k+1} x_k}{\gamma_k + t_k \alpha}.$$

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

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 $k = 0, \ldots, K$ do
    Compute $t_k \in (0,1)$ from equation
        $\beta t_k^2 = (1 - t_k)\gamma_k + t_k \alpha.$    (2.30)
    Set
        $\gamma_{k+1} = (1 - t_k)\gamma_k + t_k \alpha$    (2.31)
        $y_k = \frac{t_k \gamma_k v_k + \gamma_{k+1} x_k}{\gamma_k + t_k \alpha}$    (2.32)
        $x_{k+1} = y_k - \frac{1}{\beta} \nabla f(y_k)$    (2.33)
        $v_{k+1} = \frac{(1 - t_k)\gamma_k v_k + t_k \alpha y_k - t_k \nabla f(y_k)}{\gamma_{k+1}}$    (2.34)
    Set $k \leftarrow k + 1$.
end
```

Appealing to (2.28) and exercise 2.34, we see that the point x_k generated by Algorithm 4 satisfy

$$f(x_k) - f^* \leq \lambda_k \left[f(x_0) - f^* + \frac{\gamma_0}{2} \|x_0 - x^*\|^2 \right], \tag{2.35}$$

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 2.36 (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 2.37. *Setting $\gamma_0 = \beta$ in Algorithm 4 yields iterates satisfying*

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

Proof. This follows immediately from inequality (2.35), Theorem 2.36, 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 (2.32) and plugging in this description into (2.34) 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 (2.31) and (2.30), while the last uses (2.33). Plugging in the analogous expression of v_{k+1} into (2.32) 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 (2.30) $t_{k+1}\alpha = \beta t_{k+1}^2 - (1 - t_{k+1})\gamma_{k+1}$, 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 (2.30) 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
        $$t_{k+1}^2 = (1 - t_{k+1})t_k^2 + \tfrac{\alpha}{\beta} t_{k+1} \qquad (2.36)$$
    Set
        $$y_{k+1} = x_{k+1} + \frac{t_k(1 - t_k)}{t_k^2 + t_{k+1}}(x_{k+1} - x_k).$$
**end**
```

Thus we have established the following.

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

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

It is important to note that in the case $\alpha = 0$, Algorithm 5 is exactly Algorithm 2 with $a_k = t_k$. Indeed, equality (2.36) can be rewritten as

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

which is exactly the equality in (2.22). 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 (2.36). Thus the interpolation coefficients in the definition of y_k are exactly the same.

2.7.3 Optimal quadratic averaging

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

Henceforth, let $f\colon \mathbf{E} \to \mathbf{R}$ 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 `line_search` (x, y) be the minimizer of f on the line between x and y. We assume throughout this section that `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

$$f(x) \geq \left(f(\bar{x}) - \frac{\|\nabla f(\bar{x})\|^2}{2\alpha} \right) + \frac{\alpha}{2} \left\| x - \bar{x}^{++} \right\|^2. \tag{2.37}$$

Suppose we have now available two quadratic lower-estimators:

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

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

$$v_\lambda = v_B + \left(v_A - v_B + \frac{\alpha}{2} \left\| x_A - x_B \right\|^2 \right) \lambda - \left(\frac{\alpha}{2} \left\| x_A - x_B \right\|^2 \right) \lambda^2. \tag{2.38}$$

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 2.3 for an illustration.

An algorithmic idea emerges. Given a current iterate x_k, form the quadratic lower-model $Q(\cdot)$ in (2.37) 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.

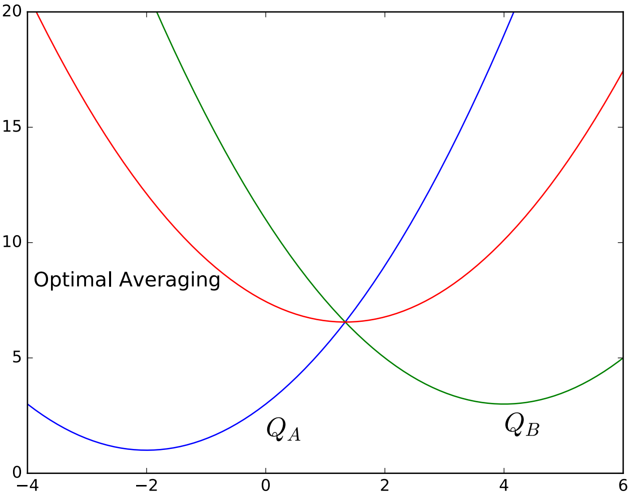

Figure 2.3: The optimal averaging of $Q_A(x) = 1 + 0.5(x+2)^2$ and $Q_B(x) = 3 + 0.5(x-4)^2$.

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

```
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 2.39. *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* $|v_A - v_B| \le \frac{\alpha}{2}\|x_A - x_B\|^2$ *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

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

If $\hat{\lambda}$ lies in $[0,1]$, equality $\bar{\lambda} = \hat{\lambda}$ holds, and then from (2.38) 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 $\frac{|v_A - v_B|}{\alpha\|x_A - x_B\|^2} \le \frac{1}{2}$. 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 2.40 (Convergence of optimal quadratic averaging). *In Algorithm 6, for every index $k \ge 0$, the inequalities $v_k \le f^* \le f(x_k^+)$ hold and we have*

$$f(x_k^+) - v_k \le \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k (f(x_0^+) - v_0).$$

Proof. Since in each iteration, the algorithm only averages quadratic minorants of f, the inequalities $v_k \le f^* \le f(x_k^+)$ 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 $v_k \ge f(x_k^+) - \frac{\alpha}{2} r_k$ holds for all $k \ge 0$. The base case $k = 0$ is immediate, and so assume we have

$$v_{k-1} \ge f(x_{k-1}^+) - \frac{\alpha}{2} r_{k-1}$$

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:

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

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 2.39, the inequality $v_k \geq v$ holds. We claim that equality

$$v = \hat{v}_B + \frac{\alpha}{8}\frac{(d^2 + \frac{2}{\alpha}(\hat{v}_A - \hat{v}_B))^2}{d^2} \qquad \text{holds.} \tag{2.39}$$

This follows immediately from Lemma 2.39, once we show $\frac{1}{2} \geq \frac{|\hat{v}_A - \hat{v}_B|}{\alpha d^2}$. To this end, note first the equality $\frac{|\hat{v}_A - \hat{v}_B|}{\alpha d^2} = \frac{|r_{k-1} - h^2|}{2d^2}$. 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

$$r_{k-1} - h^2 < \frac{r_{k-1}}{2} < h^2 \leq d^2. \tag{2.40}$$

Hence we can be sure that (2.39) 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

$$r_{k-1} - \frac{1}{\kappa}h^2 - \frac{(d^2 + r_{k-1} - h^2)^2}{4d^2} \le \left(1 - \frac{1}{\sqrt{\kappa}}\right) r_{k-1}.$$

After rearranging, our task simplifies to showing

$$\frac{r_{k-1}}{\sqrt{\kappa}} \le \frac{h^2}{\kappa} + \frac{(d^2 + r_{k-1} - h^2)^2}{4d^2}.$$

Taking derivatives and using inequality (2.40), 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

$$\frac{h^2}{\kappa} + \frac{(d^2 + r_{k-1} - h^2)^2}{4d^2} \ge \frac{h^2}{\kappa} + \frac{r_{k-1}^2}{4h^2}.$$

Minimizing the right-hand-side over all h satisfying $h^2 \ge \frac{r_{k-1}}{2}$ yields the inequality

$$\frac{h^2}{\kappa} + \frac{r_{k-1}^2}{4h^2} \ge \frac{r_{k-1}}{\sqrt{\kappa}}.$$

The proof is complete. □

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

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

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

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

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

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

References.

The convergence guarantees of gradient descent in Section 2.4.1 are classical and can be found for example in the textbooks [2, 6]. Lower complexity bounds originate in [5]; our discussion in Subsection 2.5 follows the text [6]. Our discussion of the conjugate gradient algorithm in Section 2.6 follows the monograph [10]. The fast gradient method in Section 2.7.1 and the restart strategy originate with Nesterov [8]; the proof of convergence we present is from [1]. The technique of estimate sequences in Section 2.7.2 was introduced in [9]. The quadratic averaging algorithm originates in [4], and is formally equivalent to the earlier geometric descent algorithm [3]. Estimates on the gradient norms (2.3), (2.25) follow from [7].

Bibliography

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[2] Amir Beck. *First-order methods in optimization*, volume 25 of *MOS-SIAM Series on Optimization.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2017.

[3] Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

[4] Dmitriy Drusvyatskiy, Maryam Fazel, and Scott Roy. An optimal first order method based on optimal quadratic averaging. *SIAM J. Optim.*, 28(1):251–271, 2018.

[5] A.S. Nemirovsky and D.B. Yudin. *Problem complexity and method efficiency in optimization.* A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[6] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization.* Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

[7] Y. Nesterov. How to make the gradients small. *OPTIMA, MPS Newsletter*, (88):10–11, 2012.

[8] Yu. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

[9] Yu. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1, Ser. A):127–152, 2005.

[10] J. Nocedal and S.J. Wright. *Numerical optimization.* Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.