Chapter 4

Gradient Methods Using Momentum

The steepest descent method described in Chapter 3 always steps in the negative gradient direction, which is orthogonal to the boundary of the level set for f at the current iterate. This direction can change sharply from one iteration to the next. For example, when the contours of f are narrow and elongated, the search directions at successive iterations may point almost in opposite directions and may be almost orthogonal to the direction in which the minimizer lies. The resulting small steps may produce only slow convergence toward the solution.

The steepest descent method is "greedy" in that it steps in the direction that is apparently most productive at the current iterate, making no explicit use of knowledge gained about the function f at earlier iterations. In this chapter, we examine methods that encode knowledge of the function in various ways, and exploit this knowledge in their choice of search directions and step lengths. One such class of techniques makes use of *momentum*, in which the search direction tends to be similar to that one used on the previous step, with a small tweak in the direction of a negative gradient evaluated at the current point or a nearby point. Each search direction is thus a combination of all gradients encountered so far during the search — a compact encoding of the history of the search. Momentum methods in common use include the heavy-ball method, the conjugate gradient method, and Nesterov's accelerated gradient methods.

4.1 Motivation from Differential Equations

One way to build intuition for momentum methods is to consider an optimization algorithm as a dynamical system. The continuous limit of an algorithm often traces out the solution path of a differential equation. For instance, the gradient method is akin to moving down a potential well, where the dynamics are driven by the gradient of f:

$$\frac{dx}{dt} = -\nabla f(x) \tag{4.1}$$

This differential equation has fixed points precisely when $\nabla f(x) = 0$, which are minimizers of a convex smooth function f.

There are, however, other differential equations whose fixed points occur precisely at the points for which $\nabla f(x) = 0$. Consider the second-order differential equation that governs a particle with mass moving in a potential defined by the gradient of f:

$$\mu \frac{d^2x}{dt^2} = -\nabla f(x) - \mu b \frac{dx}{dt}, \tag{4.2}$$

where $\mu \geq 0$ governs the *mass* of the particle and $b \geq 0$ governs the friction dissipated during the evolution of the system. As before, points x for which $\nabla f(x) = 0$ are fixed points of this ODE. In the limit as the mass $\mu \to 0$, and $b \to \infty$ with μb approaching a positive constant (a situation of "infinite friction"), we recover a scaled version of the system (4.1). For μ and b both positive, trajectories governed by (4.2) show evidence of momentum, continuing to move along similar directions with a slight turn toward the direction indicated by $-\nabla f(x)$.

A simple finite-difference approximation to (4.2) yields

$$\mu \frac{x(t+\Delta t) - 2x(t) + x(t-\Delta t)}{\Delta t^2} \approx -\nabla f(x(t)) - \mu b \frac{x(t+\Delta t) - x(t)}{\Delta t} \tag{4.3}$$

By rearranging terms and defining α and β appropriately (see the Exercises) we obtain

$$x(t+\Delta t) = x(t) - \alpha \nabla f(x(t)) + \beta(x(t) - x(t-\Delta t)). \tag{4.4}$$

By using this formula to generate a sequence $\{x^k\}$ of estimates of the vector x along the trajectory defined by (4.2), we obtain

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1})\,, \tag{4.5}$$

where $x^{-1} := x^0$. The algorithm defined by (4.5) is *Heavy-Ball Method* of Polyak. With a small modification, we obtain a related method known as becomes *Nesterov's optimal method* (see below). When f is a convex quadratic, approaches of the form (4.5) (possibly with adaptive choices of α and β that vary between iterations) are known as *Chebyshev iterative methods.*

Upon defining

$$p^k = x^{k+1} - x^k = -\alpha \nabla f(x^k) + \beta(x^k - x^{k-1}) = -\alpha \nabla f(x^k) + \beta p^{k-1},$$

(with $p^{-1} = 0$), we can rewrite the iteration (4.5) in terms of two sequences:

$$\begin{aligned} x^{k+1} &= x^k + p^k \\ p^k &= -\alpha \nabla f(x^k) + \beta p^{k-1}. \end{aligned}$$

Nesterov's optimal method (also known as *Nesterov's accelerated gradient method*) is defined by the formula

$$x^{k+1} = x^k - \alpha \nabla f(x^k + \beta(x^k - x^{k-1})) + \beta(x^k - x^{k-1})\,. \tag{4.6}$$

The only difference from (4.5) is that he gradient ∇f is evaluated at $x^k + \beta(x^k - x^{k-1})$ rather than at x^k. By introducing an intermediate sequence $\{y^k\}$, and allowing α and β to have possibly different values at each iteration, this method can be rewritten as follows:

$$y^k = x^k + \beta_k(x^k - x^{k-1}) \tag{4.7a}$$

$$x^{k+1} = y^k - \alpha_k \nabla f(y^k)\,, \tag{4.7b}$$

where we define $x^{-1} = x^0$ as before, so that $y^0 = x^0$.

4.2 Nesterov's Method: Convex Quadratics

In this section, we analyze the convergence behavior of Nesterov's optimal method (4.6) when applied to convex quadratic objectives f, and derive suitable values for its parameters α and β. We consider

$$f(x) = \frac{1}{2} x^T Q x - b^T x + c, \tag{4.8}$$

with positive definite Hessian Q and eigenvalues

$$0 < m = \lambda_n \le \lambda_{n-1} \le \cdots \le \lambda_2 \le \lambda_1 = L. \tag{4.9}$$

The condition number of Q is thus

$$\kappa := L/m. \tag{4.10}$$

Note that $x^* = Q^{-1}b$ is the minimizer of f, and that $\nabla f(x) = Qx - b = Q(x - x^*)$.

By specializing (4.6) to (4.8), and adding and subtracting x^* at several points in this expression, we obtain

$$x^{k+1} - x^* = (x^k - x^*) - \alpha Q(x^k + \beta(x^k - x^{k-1}) - x^*) + \beta \Big((x^k - x^*) - (x^{k-1} - x^*) \Big) . \tag{4.11}$$

By concatenating the error vector $x^k - x^*$ over two successive steps, we can express (4.11) in matrix form:

$$\begin{bmatrix} x^{k+1} - x^* \\ x^k - x^* \end{bmatrix} = \begin{bmatrix} (1+\beta)(I - \alpha Q) & -\beta(I - \alpha Q) \\ I & 0 \end{bmatrix} \begin{bmatrix} x^k - x^* \\ x^{k-1} - x^* \end{bmatrix} \tag{4.12}$$

By defining

$$w^k := \begin{bmatrix} x^{k+1} - x^* \\ x^k - x^* \end{bmatrix}, \quad T := \begin{bmatrix} (1+\beta)(I - \alpha Q) & -\beta(I - \alpha Q) \\ I & 0 \end{bmatrix} \tag{4.13}$$

we can write the iteration (4.12) as

$$w^k = T w^{k-1}, \quad k = 1, 2, \ldots . \tag{4.14}$$

For later reference, we define $x^{-1} := x^0$, so that

$$w^0 = \begin{bmatrix} x^0 - x^* \\ x^0 - x^* \end{bmatrix}. \tag{4.15}$$

Before stating a convergence result for Nesterov's method applied to (4.8), we recall the definition of the *spectral radius* of a matrix T, which is denoted by $\rho(T)$ and defined as follows:

$$\rho(T) := \max\{ |\lambda| \,|\, \lambda \text{ is an eigenvalue of } T \}. \tag{4.16}$$

For appropriate choices of α and β in (4.6), we have that $\rho(T) < 1$, which implies convergence of the sequence $\{w^k\}$ to zero. We develop this theory in the remainder of this section.

Theorem 4.1. *Consider Nesterov's optimal method* (4.6) *applied to the convex quadratic* (4.8) *with Hessian eigenvalues satisfying* (4.9). *If we set*

$$\alpha := \frac{1}{L}, \quad \beta := \frac{\sqrt{L} - \sqrt{m}}{\sqrt{L} + \sqrt{m}} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}, \tag{4.17}$$

then the matrix T defined in (4.13) *has the following eigenvalues*

$$\nu_{i,1} = \frac{1}{2}\left[(1+\beta)(1-\alpha\lambda_i) + i\sqrt{4\beta(1-\alpha\lambda_i) - (1+\beta)^2(1-\alpha\lambda_i)^2}\right], \tag{4.18a}$$

$$\nu_{i,2} = \frac{1}{2}\left[(1+\beta)(1-\alpha\lambda_i) - i\sqrt{4\beta(1-\alpha\lambda_i) - (1+\beta)^2(1-\alpha\lambda_i)^2}\right]. \tag{4.18b}$$

Moreover, $\rho(T) \le 1 - 1/\sqrt{\kappa}$.

Proof. We write the eigenvalue decomposition of Q as $Q = U\Lambda U^T$, where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. By defining the permutation matrix Π as follows:

$$\Pi_{ij} = \begin{cases} 1 & i \text{ odd}, \; j = (i+1)/2 \\ 1 & i \text{ even}, \; j = n + (i/2) \\ 0 & \text{otherwise.} \end{cases}$$

we have by applying a similarity transformation to the matrix T that

$$\begin{aligned} &\Pi \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}^T \begin{bmatrix} (1+\beta)(I-\alpha Q) & -\beta(I-\alpha Q) \\ I & 0 \end{bmatrix} \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix} \Pi^T \\ &= \Pi \begin{bmatrix} (1+\beta)(I-\alpha\Lambda) & -\beta(I-\alpha\Lambda) \\ I & 0 \end{bmatrix} \Pi^T \\ &= \begin{bmatrix} T_1 & 0 & \ldots & 0 \\ 0 & T_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & T_n \end{bmatrix}, \end{aligned}$$

where

$$T_i = \begin{bmatrix} (1+\beta)(1-\alpha\lambda_i) & -\beta(1-\alpha\lambda_i) \\ 1 & 0 \end{bmatrix}, \quad i = 1, 2, \ldots, n.$$

The eigenvalues of T are the eigenvalues of T_i, for $i = 1, 2, \ldots, n$, which are the roots of the following quadratic:

$$u^2 - (1+\beta)(1-\alpha\lambda_i)u + \beta(1-\alpha\lambda_i) = 0,$$

which are given by (4.18). Note first that for $i = 1$, we have from $\alpha = 1/L$ and $\lambda_1 = L$ that $\nu_{1,1} = \nu_{1,2} = 0$. Otherwise the roots (4.18) are distinct complex numbers when $1 - \alpha\lambda_i > 0$ and $(1+\beta)^2(1-\alpha\lambda_i) < 4\beta$. It can be shown that these inequalities hold when α and β are defined in (4.17) and $\lambda_i \in (m, L)$. Thus for $i = 2, 3, \ldots, n$, the magnitude of both $\nu_{i,1}$ and $\nu_{i,2}$ is

$$\begin{aligned} &\frac{1}{2}\sqrt{(1+\beta)^2(1-\alpha\lambda_i)^2 + 4\beta(1-\alpha\lambda_i) - (1+\beta)^2(1-\alpha\lambda_i)^2} \\ =&\frac{1}{2}\sqrt{4\beta(1-\alpha\lambda_i)} = \sqrt{\beta}\sqrt{1-(\lambda_i/L)} \end{aligned}$$

Thus for $\lambda_i \geq m$, we have

$$\begin{aligned}\sqrt{\beta}\sqrt{1-(\lambda_i/L)} \leq \sqrt{\beta}\sqrt{1-(m/L)} &= \left(\frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}} \cdot \frac{L-m}{L}\right)^{1/2} \\ &= \left(\frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}} \cdot \frac{(\sqrt{L}-\sqrt{m})(\sqrt{L}+\sqrt{m})}{L}\right)^{1/2} \\ &= \frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}} = 1-\sqrt{m/L}\,,\end{aligned}$$

with equality in the case of $\lambda_i = m$ (that is, $i = n$). We thus have

$$\rho(T) = \max_{i=1,2,\dots,n} \max(|\nu_{i,1}|, |\nu_{i,2}|) = 1 - 1/\sqrt{\kappa},$$

as required. □

We now examine the consequence of T having a spectral radius less than 1. A famous result in numerical linear algebra called *Gelfand's Formula* [8] states that

$$\rho(T) = \left(\lim_{k\to\infty} \|T^k\|\right)^{1/k}. \tag{4.19}$$

A consequence of this result is that for any $\epsilon > 0$, there is $C > 1$ such that

$$\|T^k\| \leq C(\rho(T)+\epsilon)^k. \tag{4.20}$$

Thus from (4.14), we have

$$\|w^k\| = \|T^k w^0\| \leq \|T^k\|\|w^0\| \leq (C\|w^0\|)(\rho(T)+\epsilon)^k,$$

which implies R-linear convergence provided that we choose $\epsilon \in (0, 1-\rho(T))$. Thus when $\rho(T) < 1$, we have from (4.20) that the sequence $\{w^k\}$ (hence also $\{x^k - x^*\}$) converges R-linearly to zero, with rate arbitrarily close to $\rho(T)$.

Let us compare the linear convergence of Nesterov's method against steepest descent, on convex quadratics. Recall from (3.18) that the steepest-descent method with constant step $\alpha = 1/L$ requires $O((L/m)\log\epsilon)$ iterations to obtain a reduction of factor ϵ in the function error $f(x^k) - f^*$. The rate defined by β in Theorem 4.1 suggests a complexity of $O(\sqrt{L/m}\log\epsilon)$ to obtain a reduction of factor ϵ in $\|w^k\|$ (a different quantity). For problems in which the condition number $\kappa = L/m$ is moderate to large, the heavy-ball method has a significant advantage. For example, if $\kappa = 1000$, the improved rate translates into a factor-of-30 reduction in number of iterations required, with similar workload per iteration (one gradient evaluation and a few vector operations).

A similar convergence result can be obtained by using Lyapunov functions. A function $V : \mathbb{R}^d \to \mathbb{R}$ is a Lyapunov function for an algorithm if

1. $V(w) > 0$ for all $w \neq w^*$, for some $w^* \in \mathbb{R}^d$;

2. $V(w^*) = 0$.

Lyapunov functions can be used to show convergence of an iterative process. For example, if we can show that $V(w^{k+1}) < \rho^2 V(w^k)$ for the sequence $\{w^k\}$ and some $\rho < 1$, we have demonstrated a kind of linear convergence of the sequence to its optimal point.

We construct a Lyapunov function for Nesterov's optimal method by defining a matrix P from the following theorem.

Theorem 4.2. *Let A be a square real matrix. Then for a given positive scalar ρ, we have that $\rho(A) < \rho$ if and only if there exists a $P \succ 0$ satisfying $A^T PA - \rho^2 P \prec 0$.*

Proof. If $\rho(A) < \rho$, then the matrix

$$P := \sum_{k=0}^{\infty} \rho^{-2k}(A^k)^T(A^k)$$

is well defined, positive definite (because the first term in the sum is a multiple of the identity), and satisfies $A^T PA - \rho^2 P = -\rho^2 I_d \prec 0$, proving the "only if" part of the result. For the converse, assume that the linear matrix inequality $A^T PA - \rho^2 P \prec 0$ has a solution $P \succ 0$, and let λ be an eigenvalue of A with corresponding eigenvector v. Then

$$0 > v^T A^T PAv - \rho^2 v^T Pv = (|\lambda|^2 - \rho^2)v^T Pv$$

But since $v^T Pv > 0$, we must have that $|\lambda| < \rho$. □

We apply this result to Nesterov's method by setting $A = T$ in (4.13). If there exists a $P \succ 0$ satisfying $T^T PT - \rho^2 P \prec 0$, we have

$$(w^k)^T Pw^k < \rho^2 (w^{k-1})^T Pw^{k-1}. \tag{4.21}$$

Iterating (4.21) down to $k = 0$, we see that

$$(w^k)^T Pw^k < \rho^{2k}(w^0)^T Pw^0,$$

where w^0 is defined in (4.15). We thus have

$$\lambda_{\min}(P)\|x^k - x^*\|^2 \le \lambda_{\min}(P)\|w^k\|^2 \le \rho^{2k}\|P\|\|w^0\|^2 = 2\rho^{2k}\|P\|\|x^0 - x^*\|^2,$$

so that

$$\|x^k - x^*\| \le \sqrt{2\mathrm{cond}(P)}\|x^0 - x^*\|\rho^k,$$

where cond(P) is the condition number of P. The function $V(w) := w^T Pw$ is a Lyapunov function for the algorithm, with optimum at $w^* = 0$. This function strictly decreases over all trajectories and thus certifies that the algorithm is *stable*, that is, it converges to nominal values.

For quadratic f, we are able to construct a quadratic Lyapunov function by doing an elementary eigenvalue analysis. But this proof does not generalize to the non-quadratic case. We show in the next section how to construct a Lyapunov function for Nesterov's optimal method that guarantees convergence for all strongly convex functions.

4.3 Convergence for Strongly Convex Functions

We have shown that methods that use momentum are faster on convex quadratic functions than steepest-descent methods, and the proof techniques build some intuition for the case of general strongly convex functions. But they do not generalize directly. In this section, we propose a different Lyapunov function that allows us to prove convergence of Nesterov's method for the case of strongly convex smooth functions, satisfying (2.6) (with $m > 0$) and (2.16).

It follows from the analysis of Section 3.3 that $f(x) - f^*$ is a Lyapunov function for the steepest descent method (see (3.15)). For Nesterov's method, we need a specially adapted Lyapunov function. First, for any variable v, we define $\tilde{v} := v - v^*$, where v^* is the fixed point for that variable. (Thus $\tilde{x}^k = x^k - x^*$, $\tilde{y}^k = y^k - x^*$, and so on.) Next, we define the Lyapunov function as follows:

$$V_k = f(x^k) - f^* + \frac{L}{2}\|\tilde{x}^k - \rho^2 \tilde{x}^{k-1}\|^2 . \tag{4.22}$$

(We have omitted the dependence of V_k on x^k and x^{k-1} for clarity.) We will show that

$$V_{k+1} \le \rho^2 V_k \quad \text{for some } \rho < 1, \tag{4.23}$$

provided that α_k and β_k are chosen as in (4.17), that is,

$$\alpha_k \equiv \frac{1}{L}, \quad \beta_k \equiv \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}. \tag{4.24}$$

A key device in our proof is Lemma 2.11.

For compactness of notation, we define $u^k := \frac{1}{L}\nabla f(y^k)$. (Since $u^* = 0$, we have $\tilde{u}^k = u^k$.) The decrease in the Lyapunov function at iteration k is developed as follows:

$$\begin{aligned} V_{k+1} &= f(x^{k+1}) - f^* + \frac{L}{2}\|\tilde{x}^{k+1} - \rho^2 \tilde{x}^k\|^2 \\ &\le f(y^k) - f^* - \frac{L}{2}\|\tilde{u}^k\|^2 + \frac{L}{2}\|\tilde{x}^{k+1} - \rho^2\tilde{x}^k\|^2 && \text{(4.25a)} \\ &= \rho^2(f(y^k) - f^* + L(\tilde{u}^k)^T(\tilde{x}^k - \tilde{y}^k)) - \rho^2 L(\tilde{u}^k)^T(\tilde{x}^k - \tilde{y}^k) && \text{(4.25b)} \\ &\quad + (1-\rho^2)(f(y^k) - f^* - L(\tilde{u}^k)^T\tilde{y}^k) + (1-\rho^2)L(\tilde{u}^k)^T\tilde{y}^k \\ &\quad - \frac{L}{2}\|\tilde{u}^k\|^2 + \frac{L}{2}\|\tilde{x}^{k+1} - \rho^2\tilde{x}^k\|^2. \end{aligned}$$

Here, formula (4.25a) follows from the right-hand inequality in Lemma 2.11, with $y = x^{k+1}$ and $x = y^k$, while (4.25b) is obtained by adding and subtracting the same term, several times. We now invoke the left-hand inequality in Lemma 2.11 twice. By setting $x = y^k$ and $y = x^k$, and using $\tilde{u}^k = u^k = \frac{1}{L}\nabla f(y^k)$, we obtain

$$\begin{aligned} f(y^k) &\le f(x^k) - \nabla f(y^k)^T(x^k - y^k) - \frac{m}{2}\|x^k - y^k\|^2 \\ &= f(x^k) - L(\tilde{u}^k)^T(\tilde{x}^k - \tilde{y}^k) - \frac{m}{2}\|\tilde{x}^k - \tilde{y}^k\|^2 \end{aligned}$$

By setting $x = y^k$ and $y = x^*$ in this same bound, we obtain

$$\begin{aligned} f(x^*) &\ge f(y^k) + \nabla f(y^k)^T(x^* - y^k) + \frac{m}{2}\|y^k - x^*\|^2 \\ &= f(y^k) - L(\tilde{u}^k)^T\tilde{y}^k + \frac{m}{2}\|\tilde{y}^k\|^2. \end{aligned}$$

By substituting these bounds into (4.25b), we obtain

$$\begin{aligned}
V_{k+1} &\le \rho^2(f(x^k) - f^* - \frac{m}{2}\|\tilde{x}^k - \tilde{y}^k\|^2) - \frac{m(1-\rho^2)}{2}\|\tilde{y}^k\|^2 \\
&\quad - \rho^2 L(\tilde{u}^k)^T(\tilde{x}^k - \tilde{y}^k) + (1-\rho^2)L(\tilde{u}^k)^T\tilde{y}^k \\
&\quad - \frac{L}{2}\|\tilde{u}^k\|^2 + \frac{L}{2}\|\tilde{x}^{k+1} - \rho^2\tilde{x}^k\|^2 \\
&= \rho^2(f(x^k) - f^* + \frac{L}{2}\|\tilde{x}^k - \rho^2\tilde{x}^{k-1}\|^2) \qquad (4.26a) \\
&\quad - \frac{m\rho^2}{2}\|\tilde{x}^k - \tilde{y}^k\|^2 - \frac{m(1-\rho^2)}{2}\|\tilde{y}^k\|^2 + L(\tilde{u}^k)^T(\tilde{y}^k - \rho^2\tilde{x}^k) - \frac{L}{2}\|\tilde{u}^k\|^2 \\
&\quad + \frac{L}{2}\|\tilde{x}^{k+1} - \rho^2\tilde{x}^k\|^2 - \frac{\rho^2 L}{2}\|\tilde{x}^k - \rho^2\tilde{x}^{k-1}\|^2 \qquad (4.26b) \\
&= \rho^2 V_k + R_k, \qquad (4.26c)
\end{aligned}$$

where

$$\begin{aligned}
R_k := &-\frac{m\rho^2}{2}\|\tilde{x}^k - \tilde{y}^k\|^2 - \frac{m(1-\rho^2)}{2}\|\tilde{y}^k\|^2 + L(\tilde{u}^k)^T(\tilde{y}^k - \rho^2\tilde{x}^k) - \frac{L}{2}\|\tilde{u}^k\|^2 \\
&+ \frac{L}{2}\|\tilde{x}^{k+1} - \rho^2\tilde{x}^k\|^2 - \frac{\rho^2 L}{2}\|\tilde{x}^k - \rho^2\tilde{x}^{k-1}\|^2. \qquad (4.27)
\end{aligned}$$

Formula (4.26a) follows from adding and subtracting identical terms, together with some rearrangement.

The bound (4.26c) suffices to prove (4.22) provided we can show that R_k is negative. We state the result formally as follows.

Proposition 4.3. *For Nesterov's optimal method* (4.7) *applied to a strongly convex function, with α_k and β_k defined in* (4.24)*, and setting $\rho^2 = (1 - 1/\sqrt{\kappa})$, we have*

$$R_k = -\frac{1}{2}L\rho^2\left(\frac{1}{\kappa} + \frac{1}{\sqrt{\kappa}}\right)\|\tilde{x}^k - \tilde{y}^k\|^2.$$

This result is proved purely by algebraic manipulation, using the specification of Nesterov's optimal method along with the definitions of the various quantities and the steplength settings (4.24). We leave it as an Exercise. Note that any choice of ρ and β_k that make this quantity negative would suffice. It is possible one could derive a faster bounds (that is, a lower value of ρ) by making other choices of the parameters that lead to a nonpositive value of R_k.

Proposition 4.3 asserts that R_k is a negative square for appropriately chosen parameters. Hence we can conclude that $V_{k+1} \le \rho^2 V_k$. We summarize the convergence result in the following theorem.

Theorem 4.4. *For Nesterov's optimal method* (4.7) *applied to a strongly convex function, with α_k and β_k defined in* (4.24)*, and setting $\rho^2 = (1 - 1/\sqrt{\kappa})$, we have*

$$f(x^k) - f^* \le \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \left\{f(x_0) - f^* + \frac{m}{2}\|x_0 - x^*\|^2\right\}.$$

Proof. We have from $V_{k+1} \leq \rho^2 V_k$ and the definition of V_k in (4.23) that

$$f(x^k) - f^* \leq V_k \leq \rho^{2k} V_0 = \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k V_0.$$

Recalling that $x^{-1} := x^0$, we have from (4.23) that

$$\begin{aligned} V_0 &= f(x^0) - f^* + \frac{L}{2}\|(1-\rho^2)\tilde{x}^0\|^2 \\ &= f(x^0) - f^* + \frac{L}{2}\left(\frac{1}{\sqrt{\kappa}}\right)^2 \|x^0 - x^*\|^2 \\ &= f(x_0) - f^* + \frac{m}{2}\|x_0 - x^*\|^2, \end{aligned}$$

giving the result. □

We note that the provable convergence rate is slightly worse for Nesterov's method than for heavy-ball applied to quadratics: $1-1/\sqrt{\kappa}$ for Nesterov and approximately $1-2/\sqrt{\kappa}$ for heavy-ball. This worst-case bound suggests that Nesterov's method may require about twice as many iterates to reach a given tolerance threshold ϵ. This discrepancy is rarely observed in practice. Moreover, Nesterov's method can be adapted to a wider class of functions, as we show now.

4.4 Convergence for Weakly Convex Functions

We can prove convergence of Nesterov's optimal method for weakly convex functions by modifying the analysis of Section 4.3 for the strongly convex case. The basic idea is to use varying values of β_k and hence of the decrease parameter ρ_k, while maintaining a constant value for the α parameter: $\alpha_k \equiv 1/L$.

We start by redefining V_k to use a variable value of ρ, as follows:

$$V_k = f(x^k) - f^* + \frac{L}{2}\|\tilde{x}^k - \rho_{k-1}^2 \tilde{x}^{k-1}\|^2. \tag{4.28}$$

We can now proceed with the derivation of the previous section, substituting this modified definition of V_k into (4.25) and (4.26) and replacing ρ by ρ_k in the addition/subtraction steps. By setting $m = 0$ in (4.26b), we obtain

$$\begin{aligned} V_{k+1} &\leq \rho_k^2 (f(x^k) - f^* + \frac{L}{2}\|\tilde{x}^k - \rho_{k-1}^2 \tilde{x}^{k-1}\|^2) && \text{(4.29a)} \\ &\qquad + L(\tilde{u}^k)^T(\tilde{y}^k - \rho_k^2 \tilde{x}^k) - \frac{L}{2}\|\tilde{u}^k\|^2 + \frac{L}{2}\|\tilde{x}^{k+1} - \rho_k^2 \tilde{x}^k\|^2 - \frac{\rho_k^2 L}{2}\|\tilde{x}^k - \rho_{k-1}^2 \tilde{x}^{k-1}\|^2 \\ &= \rho_k^2 (f(x^k) - f^* + \frac{L}{2}\|\tilde{x}^k - \rho_{k-1}^2 \tilde{x}^{k-1}\|^2) + \frac{L}{2}\|\tilde{y}^k - \rho_k^2 \tilde{x}^k\|^2 - \frac{\rho_k^2 L}{2}\|\tilde{x}^k - \rho_{k-1}^2 \tilde{x}^{k-1}\|^2 && \text{(4.29b)} \\ &= \rho_k^2 V_k + R_k^{(\text{weak})}, && \text{(4.29c)} \end{aligned}$$

where

$$R_k^{(\text{weak})} := \frac{L}{2}\|\tilde{y}^k - \rho_k^2 \tilde{x}^k\|^2 - \frac{\rho_k^2 L}{2}\|\tilde{x}^k - \rho_{k-1}^2 \tilde{x}^{k-1}\|^2. \tag{4.30}$$

Formula (4.29b) follows by using the identity $\tilde{x}^{k+1} = x^{k+1} - x^* = y^k - u^k - x^* = \tilde{y}^k - \tilde{u}^k$, from (4.7b), and completing the square.

We now choose ρ_k to force $R_k^{(\text{weak})} = 0$ for $k \ge 1$. From the definition (4.30), this will be true provided that

$$\tilde{y}^k - \rho_k^2 \tilde{x}^k = \rho_k \tilde{x}^k - \rho_k \rho_{k-1}^2 \tilde{x}^{k-1}. \tag{4.31}$$

By substituting $\tilde{y}^k = (1+\beta_k)\tilde{x}^k - \beta_k \tilde{x}^{k-1}$ (from (4.7b)), and setting the coefficients of $\tilde{x}^k$ and $\tilde{x}^{k-1}$ to zero, we find that the following conditions ensure (4.31):

$$1 + \beta_k - \rho_k^2 = \rho_k, \quad \beta_k = \rho_k \rho_{k-1}^2. \tag{4.32}$$

From an arbitrary choice of ρ_0 (about which more below), we can use these formulae to define subsequent values of β_k and ρ_k, for $k = 1, 2, \ldots$. By substituting for β_k, we obtain the following relationship between two successive values of ρ:

$$1 + \rho_k(\rho_{k-1}^2 - 1) - \rho_k^2 = 0, \tag{4.33}$$

which yields

$$\rho_k^2 = \frac{(1-\rho_k^2)^2}{(1-\rho_{k-1}^2)^2}, \quad k = 1, 2, \ldots. \tag{4.34}$$

Using the fact that $V_k \le \rho_{k-1}^2 V_{k-1}$ for $k = 1, 2, \ldots$ (from (4.29c) and $R_k^{(\text{weak})} = 0$ for $k = 1, 2, \ldots$, we obtain

$$V_k \le \rho_{k-1}^2 \rho_{k-2}^2 \ldots \rho_1^2 V_1 = \left\{ \prod_{j=1}^{k-1} \rho_j^2 \right\} V_1 = \frac{(1-\rho_{k-1}^2)^2}{(1-\rho_0^2)^2} V_1. \tag{4.35}$$

For a bound on V_1, we make the choices $\rho_0 = 0$ and $\rho_{-1} = 0$, use (4.29c) and (4.30), and recall that $y^0 = x^0$ to obtain

$$V_1 \le R_0^{(\text{weak})} = \frac{L}{2}\|\tilde{y}^0\|^2 = \frac{L}{2}\|x^0 - x^*\|^2,$$

which by substitution into (4.35) (setting $\rho_0 = 0$ again) yields

$$V_k \le (1-\rho_{k-1}^2)^2 \frac{L}{2}\|x^0 - x^*\|^2. \tag{4.36}$$

We now use an elementary inductive argument to show that

$$1 - \rho_k^2 \le \frac{2}{k+2}. \tag{4.37}$$

Note first that the choice $\rho_0 = 0$ ensures that (4.37) is satisfied for $k = 0$. Supposing that it is satisfied for some k, we want to show that $1 - \rho_{k+1}^2 \le 2/(k+3)$. Suppose for contradiction that this claim is *not* true. We then have

$$1 - \rho_{k+1}^2 > \frac{2}{k+3}, \quad \text{so that } \rho_{k+1}^2 < \frac{k+1}{k+3},$$

and thus

$$\frac{(1-\rho_{k+1}^2)^2}{\rho_{k+1}^2} > \left(\frac{2}{k+3}\right)^2 \frac{k+3}{k+1} = \frac{4}{(k+1)(k+3)}.$$

Since $(k+1)(k+3) < (k+2)^2$ for all k, this bound together with (4.37) contradicts (4.34). We conclude that (4.37) continues to hold when k is replaced by $k+1$, so by induction (4.37) holds for $k = 0, 1, 2, \ldots$.

By substituting (4.37) into (4.36), and using the definition (4.28), we obtain

$$f(x^k) - f^* \le V_k \le \frac{2L}{(k+1)^2} \|x^0 - x^*\|^2. \tag{4.38}$$

This sublinear rate is faster than the rate we proved for the steepest-descent method (see Theorem 3.3) by a square-root factor.

We summarize Nesterov's optimal method for the weakly convex case in Algorithm 4.1. Note that we have defined ρ_k and β_k to satisfy the formulas (4.32) and (4.33), for $k = 1, 2, \ldots$, and set $\alpha_k \equiv 1/L$ in (4.7b).

Algorithm 4.1 Nesterov's Optimal Algorithm: General Convex f

```
Given $x^0$ and constant $L$ satisfying (2.16), set $x^{-1} = x^0$, $\beta_0 = 0$, and $\rho_0 = 0$;
for $k = 0, 1, 2, \ldots$ do
    Set $y^k := x^k + \beta_k(x^k - x^{k-1})$;
    Set $x^{k+1} := y^k - (1/L)\nabla f(x^k)$;
    Define $\rho_{k+1}$ to be the root in $[0, 1]$ of the following quadratic: $1 + \rho_{k+1}(\rho_k^2 - 1) - \rho_{k+1}^2 = 0$;
    Set $\beta_{k+1} = \rho_{k+1}\rho_k^2$;
end for
```

4.5 Conjugate Gradient Method

The problem with Nesterov's method as presented is that one is required to know the convexity parameters L and m to compute the appropriate step-sizes. The conjugate gradient method is a simple modification of Nesterov's method that does not require knowledge of these parameters. Consider the momentum iterations

$$\begin{aligned} x^{k+1} &= x^k - \alpha_k y^k \\ y^k &= -\nabla f(x^k) + \beta_{k-1} y^{k-1} \end{aligned}$$

This is equivalent to our previous iteration after a change of variables. To choose α_k, we can pick the stepsize to minimize f along the direction y^k (i.e. $\min_{\alpha > 0} f(x^k + \alpha y^k)$). For quadratics, if we take derivatives, this gives us

$$\alpha_k = \frac{(y^k)^T r^k}{(y^k)^T Q y^k} \qquad \text{where } r^k = Qx^k - p$$

Definition 4.5. *We say that two non-zero vectors u and v are conjugate (with respect to Q) if*

$$u^T Q v = 0\,.$$

We now pick β_k so that $\langle y^k, Qy^{k-1}\rangle = 0$ (i.e to make y^k and y^{k-1} conjugate).

$$\begin{aligned}\langle y^k, Qy^{k-1}\rangle &= \langle -r^{k-1} + \beta y^{k-1}, Qy^{k-1}\rangle \\ &= \langle -r^{k-1}, Qy^{k-1}\rangle + \beta_k\langle y^{k-1}, Qy^{k-1}\rangle\end{aligned}$$

$$\beta_k = \frac{\langle r^{k-1}, Qy^{k-1}\rangle}{\langle y^{k-1}, Qy^{k-1}\rangle}$$

Conjugacy guarantees that the directions y^k are orthogonal with respect to the inner product

$$\langle u, v\rangle_Q := u^T Qv\,.$$

Consequently, walking along conjugate directions using exact line search yields convergence to x^* in n steps.

Unfortunately, the conjugate gradient method doesn't admit particularly rigorous analysis when f is not quadratic.

4.6 Lower Bounds on Rates

The term "optimal" in Nesterov's optimal method is used because the convergence rate achieved by the method is the best possible (possibly up to a constant), among algorithms that make use of gradient information at the iterates x^k. This claim can be proved by means of a carefully designed function, for which *no* method that makes use of all gradient observed up to and including iteration k (namely, $\nabla f(x^i)$, $i = 0, 1, 2, \dots, k$) can produce a sequence $\{x^k\}$ that achieves a rate better than (4.38). The function proposed in [14] is a convex quadratic $f(x) = (1/2)x^T Ax - e_1^T x$, where

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & \dots & 0 \\ 0 & -1 & 2 & -1 & 0 & \dots & 0 \\ & & \ddots & \ddots & \ddots & & \\ 0 & \dots & & -1 & 2 & -1 & \\ 0 & \dots & & & 0 & -1 & 2 \end{bmatrix}, \quad e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The solution x^* satisfies $Ax^* = e_1$; its components are $x_i^* = 1 - i/(n+1)$, for $i = 1, 2, \dots, n$. If we use $x^0 = 0$ as the starting point, and construct the iterate x^{k+1} as

$$x^{k+1} = x^k + \sum_{j=0}^{k} \gamma_j \nabla f(x^j),$$

for some coefficients γ_j, $j = 0, 1, \dots, k$, an elementary inductive argument shows that each iterate x^k can have nonzero entries only in its first k components. It follows that for any such algorithm, we have

$$\|x^k - x^*\|^2 \geq \sum_{j=k+1}^{n} (x_j^*)^2 = \sum_{j=k+1}^{n} \left(1 - \frac{j}{n+1}\right)^2. \tag{4.39}$$

A little arithmetic (see Exercises) shows that

$$\|x^k - x^*\|^2 \geq \frac{1}{8}\|x^0 - x^*\|^2, \quad k = 1, 2, \ldots, \frac{n}{2} - 1, \tag{4.40}$$

It can be shown further that

$$f(x^k) - f^* \geq \frac{3L}{32(k+1)^2}\|x^0 - x^*\|^2, \quad k = 1, 2, \ldots, \frac{n}{2} - 1, \tag{4.41}$$

where $L = \|A\|_2$. This lower bound on $f(x^k) - x^*$ is within a constant factor of the upper bound (4.38).

The restriction $k \leq n/2$ in the argument above is not fully satisfying. A more compelling example would show that the lower bound (4.41) holds for all k.

Notes and References

The discussion of Lyapunov functions follows Lessard et al. Our analysis of Nesterov's optimal method in the weakly convex case (Section 4.4) is similar to that of [24].[1]

Exercises

1. Define α and β in terms of b, μ, and Δt such that (4.4) corresponds to (4.3).

2. Minimize a quadratic objective $f(x) = (1/2)x^TAx$ with some first-order methods, generating the problems using the following Matlab code fragment (or its equivalent in another language) to generate a Hessian with eigenvalues in the range $[m, L]$.

```
mu=0.01; L=1; kappa=L/mu;
n=100;
A = randn(n,n); [Q,R]=qr(A);
D=rand(n,1); D=10.^D; Dmin=min(D); Dmax=max(D);
D=(D-Dmin)/(Dmax-Dmin);
D = mu + D*(L-mu);
A = Q'*diag(D)*Q;
epsilon=1.e-6;
kmax=1000;
x0 = randn(n,1); % use a different x0 for each of the 10 trials
```

   Run the code in each case until $f(x_k) \leq \epsilon$ for tolerance $\epsilon = 10^{-6}$. Implement the following methods.

   - Steepest descent with $\alpha_k \equiv 2/(m + L)$.
   - Steepest descent with $\alpha_k \equiv 1/L$.
   - Steepest descent with exact line search.

[1] **SJW:** is this true?

- Heavy-ball method, with $\alpha = 4/(\sqrt{L}+\sqrt{m})^2$ and $\beta = (\sqrt{L}-\sqrt{m})/(\sqrt{L}+\sqrt{m})$.
- Nesterov's optimal method, with $\alpha = 1/L$ and $\beta = (\sqrt{L}-\sqrt{m})/(\sqrt{L}+\sqrt{m})$.

(a) Tabulate the average number of iterations required, over 10 random starts.

(b) Draw a plot of the convergence behavior on a typical run, plotting iteration number against $\log_{10}(f(x_k) - f(x^*))$. (Use the same figure, with different colors for each algorithm.)

(c) Discuss your results, noting in particular whether the worst-case convergence analysis is reflected in the practical results.

3. Discuss happens to the codes and algorithms in the previous question when we reset m to 0 (making f weakly convex). Comment on particular on what happens when you use the uniform steplength $\alpha_k \equiv 2/(L+m)$ in steepest descent. Are these observations consistent with the convergence theory of Chapter 3?

4. Prove using Gelfand's formula (4.19) that (4.20) is true, for some $C > 1$.

5. Show that the heavy-ball method (4.5) converges at a linear rate on the convex quadratic (4.8) with eigenvalues (4.9), if we set

$$\alpha := \frac{4}{(\sqrt{L}+\sqrt{m})^2}, \quad \beta := \frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}}.$$

You can follow the proof technique of Section4.2 to a large extent, proeeding in the following steps.

(a) Write the algorithm as a linear recursion $w^{k+1} = Tw^k$ for appropriate choice of matrix T and state variables w^k.

(b) Use a transformation to express T as a block-diagonal matrix, with 2×2 blocks T_i on the diagonals, where each T_i depends on a single eigenvalue λ_i of Q.

(c) Find the eigenvalues $\bar{\lambda}_{i,1}$, $\bar{\lambda}_{i,2}$ of each T_i as a function of λ_i, α, and β.

(d) Show that for the given values of α and β, these eigenvalues are all complex.

(e) Show that in fact $|\bar{\lambda}_{i,1}| = |\bar{\lambda}_{i,2}| = \sqrt{\beta}$ for all $i = 1, 2, \ldots, n$, so that $\rho(T) = \sqrt{\beta} \approx 1 - \kappa^{-1/2}$.

6. Prove Proposition 4.3 by using (4.7); the definitions $\kappa = L/m$, $\tilde{u}^k = u^k = (1/L)\nabla f(y^k)$, and $\rho^2 = (1 - 1/\sqrt{\kappa})$; and (4.24).

7. Show that if $\rho_{k-1} \in [0,1]$, the quadratic equation (4.33) has a root ρ_k in $[0,1]$.

8. For the quadratic function of Section 4.6, prove the following bounds:

$$\|x^0 - x^*\|_2^2 \le n/3, \quad \|x^k - x^*\|^2 \ge \frac{(n-k)^3}{3(n+1)^2} \ge \frac{(n-k)^3}{n(n+1)^2}\|x^0 - x^*\|^2.$$

(The bound (4.40) follows by setting $k = \frac{n}{2} - 1$ in this expression and noting that it is decreasing in k.)

Chapter 5

Stochastic Gradient Methods

The stochastic gradient (SG) method is one of the most popular algorithms in modern data analysis and machine learning. It has a long history. Variants of the method have been invented and reinvented several times by different communities, under such names as "least mean squares," "back propagation," "online learning," and the "randomized Kaczmarz method." Most people attribute the SG approach to the 1951 work of Robbins and Monro [21], who were interested in devising efficient algorithms for computing random means and roots of scalar functions for which only noisy values are available. In this chapter, we explore some of the properties and implementation details of the SG method.

As before, our target problem is to minimize the multivariate convex function $f : \mathbb{R}^d \to \mathbb{R}$. The SG method differs from methods of Chapters 3 and 4 in the kind of information that is available about f. In place of an exact value of $\nabla f(x)$, we assume that we can compute or acquire a random function $g(x, \xi)$, which depends on a random variable ξ as well as x, such that

$$\nabla f(x) = \mathbb{E}_\xi [g(x, \xi)]. \tag{5.1}$$

(We assume that ξ belongs to some space Ξ with distribution P, and $\mathbb{E}_\xi$ denotes the expectation taken over $\xi \in \Xi$ according to distribution P.) SG proceeds by substituting $g(x, \xi)$ for the true gradient ∇f in the steepest-descent update formula, so each iteration is as follows:

$$x^{k+1} = x^k - \alpha_k g(x^k, \xi^k), \tag{5.2}$$

where the random variable ξ^k is chosen according to the distribution P (independently of the choices at other iterations) and $\alpha_k > 0$ is the steplength. The method steps in a direction that *in expectation* equals the steepest descent direction. Although $g(x^k, \xi^k)$ may differ substantially from $\nabla f(x^k)$ — it may contain a lot of "noise" — it also contains enough "signal" to yield provable progress over the long term.

The choice of steplength α_k is critical to the theoretical and practical behavior of the SG approach. We cannot expect to match the performance of steepest descent, in which we move along the true negative gradient direction $-\nabla f(x^k)$ rather than its noisy approximation $-g(x^k, \xi^k)$. In the steepest descent method, the constant steplength $\alpha_k \equiv 1/L$ (where L is the Lipschitz constant for ∇f) yields convergence; see Chapter 3. We can show that this constant-steplength choice will not yield the same convergence properties in the stochastic gradient context, by considering what happens if we initialize the method at the minimizer of f, that is, $x^0 = x^*$. Since $\nabla f(x^*) = 0$,