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 in almost opposite directions and may be almost orthogonal to the direction in which the minimizer lies. The method may thus take small steps that 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 several ways and exploit this knowledge in their choice of search directions and steplengths. One such class of techniques makes use of *momentum*, in which the search direction tends to be similar to the one used on the previous step but adds a small component from the negative gradient of f, 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 include the heavy-ball method, the conjugate gradient method, and Nesterov’s accelerated gradient methods.

The analysis of momentum methods tends to be laborious and not very intuitive. But these methods often achieve significant practical improvements over steepest descent, so it is worthwhile to gain some theoretical understanding. Several approaches to the analysis have been proposed. Here, we begin with strictly convex *quadratic* functions (Section 4.2) and present a convergence analysis of Nesterov’s accelerated gradient method that uses tools from linear algebra. We relate this analysis technique to the notion of *Lyapunov functions*, which we then use as a tool to analyze first strongly convex functions

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

(Section 4.3) and then general convex functions (Section 4.4). We make some remarks about the conjugate gradient method in Section 4.5 and then discuss lower bounds on global convergence rates in Section 4.6. (Lower bounds define a "speed limit" for methods of a certain class; methods that achieve these bounds are known as "optimal methods.")

One way to motivate momentum methods is to relate them to techniques for differential equations. We do this next.

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 (as the steplength goes to zero) 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, as follows:

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

This differential equation has fixed points precisely when $\nabla f(x) = 0$, which are minimizers of a convex smooth function f. Equation (4.1) is not the only differential equation 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:

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

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, the points x for which $\nabla f(x) = 0$ are fixed points of this ODE. In the limit as the mass $\mu \to 0$, we recover a scaled version of system (4.1). For positive values of μ, trajectories governed by (4.2) show evidence of momentum, gradually changing their orientations toward the direction indicated by $-\nabla f(x)$.

A simple finite-difference approximation to (4.2) yields

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

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

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*, described by Polyak (1964). With a small modification, we obtain a related method known as *Nesterov's optimal method*, discussed later. When applied to a convex quadratic function f, approaches of the form (4.5) (possibly with adaptive choices of α and β that vary between iterations) are known as *Chebyshev iterative methods*.

Nesterov's optimal method (also known as Nesterov's accelerated gradient method (Nesterov, 1983)) 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 the 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 $y^0 = x^0$. Note that we obtain y^k by taking a pure momentum step based on the last two x-iterates, while we obtain x^{k+1} by taking a pure gradient step from y^k. In this sense, the momentum step and the gradient step are teased apart, rather than being combined in a single step.

Note that each of these methods has a fixed point with $x^k = x^*$, where x^* is a minimizer of f. (For Nesterov's method, we also need $y^* = x^*$.) The rest of the chapter is devoted to finding conditions under which these accelerated algorithms converge to x^* at provable global rates. As we will see, with proper setting of parameters, these methods converge faster than the steepest-descent method.

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

4.2 Nesterov's Method: Convex Quadratics

We now 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

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

with positive definite Hessian Q and eigenvalues

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

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 $\nabla f(x) = Qx - b = Q(x - x^*)$.

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

$$\begin{aligned} &x^{k+1} - x^* \\ &= (x^k - x^*) - \alpha Q(x^k + \beta(x^k - x^{k-1}) - x^*) + \beta\left((x^k - x^*) - (x^{k-1} - x^*)\right). \end{aligned}$$

By concatenating the error vector $x^k - x^*$ over two successive steps, we can restate this expression in matrix form as follows:

$$\begin{bmatrix} x^{k+1} - x^* \\ x^k - x^* \end{bmatrix} = \begin{bmatrix} (1+\beta)(I - \alpha Q) & -\beta(I - \alpha Q) \\ I & 0 \end{bmatrix} \begin{bmatrix} x^k - x^* \\ x^{k-1} - x^* \end{bmatrix}. \tag{4.11}$$

By defining

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

we can write the iteration (4.11) as

$$w^k = Tw^{k-1}, \quad k = 1, 2, \ldots. \tag{4.13}$$

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

$$w^0 = \begin{bmatrix} x^0 - x^* \\ x^0 - x^* \end{bmatrix}. \tag{4.14}$$

Before stating a convergence result for Nesterov's method applied to (4.8), we recall that the *spectral radius* of a matrix T is defined as follows:

$$\rho(T) := \max\{|\lambda| : \ \lambda \ \text{ is an eigenvalue of } T\}. \tag{4.15}$$

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

For appropriate choices of α and β in (4.6), we have $\rho(T) < 1$, 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*

$$\alpha := \frac{1}{L}, \quad \beta := \frac{\sqrt{L} - \sqrt{m}}{\sqrt{L} + \sqrt{m}} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}, \tag{4.16}$$

the matrix T defined in (4.12) *has complex eigenvalues*

$$\nu_{i,1} = \frac{1}{2}\left[(1+\beta)(1-\alpha\lambda_i) + i\sqrt{4\beta(1-\alpha\lambda_i) - (1+\beta)^2(1-\alpha\lambda_i)^2}\right], \tag{4.17a}$$

$$\nu_{i,2} = \frac{1}{2}\left[(1+\beta)(1-\alpha\lambda_i) - i\sqrt{4\beta(1-\alpha\lambda_i) - (1+\beta)^2(1-\alpha\lambda_i)^2}\right]. \tag{4.17b}$$

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

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

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

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

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

where

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

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

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 (4.17). 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 (4.17) 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 (4.16) 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

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

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

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

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* (Gelfand, 1941) states that

$$\rho(T) = \lim_{k\to\infty} \left(\|T^k\|\right)^{1/k}. \tag{4.18}$$

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

$$\|T^k\| \leq C_\epsilon(\rho(T)+\epsilon)^k. \tag{4.19}$$

Thus, from (4.13), we have

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

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

which implies R-linear convergence, provided we choose $\epsilon \in (0, 1 - \rho(T))$. Thus, when $\rho(T) < 1$, we have from (4.19) 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 (3.17) 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 $\|w^k\|$ (which is obviously a different quantity from $f(x^k) - f^*$, but one that also shrinks to zero as $x^k \to x^*$). For problems in which the condition number $\kappa = L/m$ is moderate to large, Nesterov's method has a significant advantage. For example, if $\kappa = 1{,}000$, the improved rate translates into an approximate 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 the notion of *Lyapunov functions*. A Lyapunov function $V : \mathbb{R}^D \to \mathbb{R}$ has two essential properties:

1. $V(z) > 0$ for all $z \neq z^*$, for some $z^* \in \mathbb{R}^D$
2. $V(z^*) = 0$.

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

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 symmetric matrix $P \succ 0$ 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, is positive definite (because the first term in the sum is a multiple of the identity and all other terms are at least positive semidefinite), and satisfies $A^T PA - \rho^2 P = -\rho^2 I \prec 0$, proving the "only if" part of the result. For the converse, assume that the linear matrix inequality

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

$A^T PA - \rho^2 P \prec 0$ has a solution $P \succ 0$, and let $\lambda \in \mathbb{C}$ be an eigenvalue of A with corresponding eigenvector $v \in \mathbb{C}^D$. Then

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

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

We apply this result to Nesterov's method by setting $A = T$ in (4.12). If there exists a $P \succ 0$ satisfying $T^T PT - \rho^2 P \prec 0$, we have from (4.13) that

$$(w^k)^T Pw^k < \rho^2 (w^{k-1})^T Pw^{k-1}. \tag{4.20}$$

Iterating (4.20) 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 (4.14). 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

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

where $\text{cond}(P)$ is the condition number of P. In other words, the function $V(w) := w^T Pw$ is a Lyapunov function for the Nesterov algorithm, with optimum at $w^* = 0$. This function decreases strictly 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. This proof does not generalize to the nonquadratic case, however. 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 (2.18) (with $m > 0$) and the L-smooth property (2.7).

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

It follows from the analysis of Section 3.2 that $V(x) := f(x) - f^*$ is actually a Lyapunov function for the steepest-descent method (see (3.14)). For Nesterov's method, we need to define a specially adapted Lyapunov function. First, for any variable v, we define $\tilde{v}^k := v^k - v^*$ for any sequence $\{v^k\}$ that converges to v^*. Next, we define the Lyapunov function as follows:

$$V_k = f(x^k) - f^* + \frac{L}{2}\|\tilde{x}^k - \rho^2\tilde{x}^{k-1}\|^2. \tag{4.21}$$

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

$$V_{k+1} \leq \rho^2 V_k \quad \text{for some } \rho < 1, \tag{4.22}$$

provided that α_k and β_k are chosen as in (4.16); that is,

$$\alpha_k \equiv \frac{1}{L}, \quad \beta_k \equiv \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}. \tag{4.23}$$

To do so, we only make use of the standard chain of inequalities for strongly convex functions with Lipschitz gradients that we used extensively in Chapter 3 for studying the gradient methods. Namely, we use inequalities (2.9) and (2.19), restated here for convenience:

$$\begin{aligned} &f(z) + \nabla f(z)^T(w-z) + \frac{m}{2}\|w-z\|^2 \\ &\quad\leq f(w) \\ &\quad\leq f(z) + \nabla f(z)^T(w-z) + \frac{L}{2}\|w-z\|_2^2, \quad \text{for all } w \text{ and } z. \end{aligned} \tag{4.24}$$

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:

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

Here, formula (4.25a) follows from the right-hand inequality in (4.24), with $w = x^{k+1}$ and $z = y^k$, and (4.25b) is obtained by adding and subtracting

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

the same term several times. We now invoke the left-hand inequality in (4.24) twice. By setting $w = y^k$ and $z = x^k$ 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 $w = x^*$ and $z = y^k$ 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

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

where

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

The bound (4.26b) suffices to prove (4.21), 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* (4.23)*, and setting $\rho^2 = (1 - 1/\sqrt{\kappa})$, we have for R_k defined in* (4.27) *that*

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

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

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 (4.23). 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 that one could derive a faster bound (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 $V_{k+1} \leq \rho^2 V_k$. 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* (4.23), *and setting $\rho^2 = (1 - 1/\sqrt{\kappa})$, we have*

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

Proof We have from $V_{k+1} \leq \rho^2 V_k$ and the definition of V_k in (4.22) 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 (4.22) 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 the heavy-ball method applied to quadratics: $1 - 1/\sqrt{\kappa}$ for Nesterov and approximately $1 - 2/\sqrt{\kappa}$ for heavy-ball. (We prove the latter rate in the Exercises, using a similar technique to the one in Section 4.2.) 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.

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

4.4 Convergence for Weakly Convex Functions

We can prove convergence of Nesterov's optimal method (4.7) for weakly convex functions by modifying the analysis of Section 4.3. We need to allow β_k to vary with k (and, hence, ρ_k also) 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 with ρ_k in the addition/subtraction steps. By setting $m = 0$ in (4.26a), we obtain

$$\begin{aligned}
V_{k+1} \le \rho_k^2 &\left[f(x^k) - f^* + \frac{L}{2}\|\tilde{x}^k - \rho_{k-1}^2 \tilde{x}^{k-1}\|^2 \right] \\
&+ L(\tilde{u}^k)^T(\tilde{y}^k - \rho_k^2 \tilde{x}^k) - \frac{L}{2}\|\tilde{u}^k\|^2 \\
&+ \frac{L}{2}\|\tilde{x}^{k+1} - \rho_k^2 \tilde{x}^k\|^2 - \frac{\rho_k^2 L}{2}\|\tilde{x}^k - \rho_{k-1}^2 \tilde{x}^{k-1}\|^2 \\
= \rho_k^2 &\left[f(x^k) - f^* + \frac{L}{2}\|\tilde{x}^k - \rho_{k-1}^2 \tilde{x}^{k-1}\|^2 \right] \\
&+ \frac{L}{2}\|\tilde{y}^k - \rho_k^2 \tilde{x}^k\|^2 - \frac{\rho_k^2 L}{2}\|\tilde{x}^k - \rho_{k-1}^2 \tilde{x}^{k-1}\|^2 \qquad (4.29a) \\
= \rho_k^2 &V_k + W_k, \qquad (4.29b)
\end{aligned}$$

where

$$W_k := \frac{L}{2}\|\tilde{y}^k - \rho_k^2 \tilde{x}^k\|^2 - \frac{\rho_k^2 L}{2}\|\tilde{x}^k - \rho_{k-1}^2 \tilde{x}^{k-1}\|^2. \tag{4.30}$$

Formula (4.29a) 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 by completing the square.

We now choose ρ_k to force $W_k = 0$ for $k \ge 1$. From the definition (4.30), this will be true, provided

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

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):

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

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

From an arbitrary choice of ρ_0 (more information about this is given in what follows), we can use these formulas 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 (4.29b) and $W_k = 0$ 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 (4.29b) and (4.30), and recall that $y^0 = x^0$ to obtain

$$V_1 \le W_0 = \frac{L}{2}\|\tilde{y}^0\|^2 = \frac{L}{2}\|x^0 - x^*\|^2,$$

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 $k = 0, 1, 2, \ldots$.

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

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

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

This sublinear rate is faster than the rate proved for the steepest-descent method in Theorem 3.3 in that $1/k$ convergence has become $1/k^2$ convergence.

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 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: Weakly Convex f

```
Given $x^0$ and constant $L$ satisfying (2.7), 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 $x^{k+1} := y^k - (1/L)\nabla f(y^k)$;
    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 Methods

A problem with the version of Nesterov's method described before is that we need to know the parameters L and m to compute the appropriate steplengths. (There are versions of these methods for which this prior knowledge is not required, and adaptive estimates of L are made (see Nesterov, 2015; Beck and Teboulle, 2009). The conjugate gradient method, developed in the early 1950s for systems of equations involving symmetric positive definite matrices (equivalently, minimizing strongly convex quadratic functions) does not require knowledge of these parameters. The conjugate gradient method, which is also a momentum method, can be extended and adapted to solve smooth (even nonconvex) optimization problems, as shown first by Fletcher and Reeves (1964).

Focusing for the moment on the case of strongly convex quadratic f, consider first the heavy-ball formula (4.5) in which α and β are allowed to vary across iterations, as follows:

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k(x^k - x^{k-1}). \tag{4.39}$$

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

We now introduce a vector p^k that captures the search direction, such that $x^{k+1} = x^k + \alpha_k p^k$ for all k. With some manipulation, we see that

$$\begin{aligned} p^k &= -\nabla f(x^k) + \frac{\beta_k}{\alpha_k}(x^k - x^{k-1}) = -\nabla f(x^k) + \frac{\beta_k \alpha_{k-1}}{\alpha_k} p^{k-1} \\ &= -\nabla f(x^k) + \gamma_{k-1} p^{k-1}, \end{aligned}$$

where we introduced a new scalar γ_{k-1} to replace $\beta_k \alpha_{k-1}/\alpha_k$. (Initially, we set $p^0 = -\nabla f(x^0)$.) The conjugate gradient method also keeps track of the residual $r^k = \nabla f(x^k) = Qx^k - b$, where we used the notation (4.8). Note that r^k can be updated to r^{k+1} as follows:

$$r^{k+1} = Qx^{k+1} - b = Qx^k - b + \alpha_k Qp^k = r^k + \alpha_k Qp^k.$$

Thus, the conjugate gradient method for strongly convex quadratic functions can be defined by the following three update formulas:

$$x^{k+1} \leftarrow x^k + \alpha_k p^k, \tag{4.40a}$$

$$r^{k+1} \leftarrow r^k + \alpha_k Qp^k, \tag{4.40b}$$

$$p^{k+1} \leftarrow -r^{k+1} + \gamma_k p^k, \tag{4.40c}$$

together with the formulas defining the scalars γ_k and α_k. We choose α_k by performing an exact minimization of $f(x^k + \alpha p^k)$ for α – which, for the convex quadratic (4.8), leads to the explicit formula

$$\alpha_k = \frac{(p^k)^T r^k}{(p^k)^T Qp^k}. \tag{4.41}$$

We choose γ_k to ensure that the two directions p^k and p^{k+1} satisfy *conjugacy* with respect to Q – that is, $(p^k)^T Qp^{k+1} = 0$. By substituting from (4.40c), we obtain

$$\gamma_k = \frac{(r^{k+1})^T Qp^k}{(p^k)^T Qp_k} = \frac{(r^{k+1})^T r^{k+1}}{(r^k)^T r^k}. \tag{4.42}$$

(The equality of the last two formulas is not obvious, and we leave it as an Exercise.) Formulas (4.40), (4.41), and (4.42), along with the initial iterate x^0 and search direction $p^0 = -(Qx^0 - b)$, give a complete description of the basic conjugate gradient method for the strongly convex quadratic function (4.8).

One remarkable property of the conjugate gradient method is that we do not just have conjugacy of two successive search directions p^k and p^{k+1}, ensured by formula (4.42), but, in fact, p^{k+1} is conjugate to *all* preceding search directions $p^k, p^{k-1}, \ldots, p^0$! It follows that these directions form a linearly independent set, and we can show in addition that x^{k+1} is the minimizer of

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

f in the affine set defined by $x^0 + \text{span}\,\{p^0, p^1, \dots, p^k\}$. Thus, the conjugate gradient method is guaranteed to terminate at an exact minimizer of a strongly convex quadratic f in at most n iterations.

Many extensions of the conjugate gradient approach to nonquadratic and nonconvex functions have been proposed. These typically involve choosing α_k with a (possibly inexact) line search along the direction p^k and defining γ_k in a way that mimics (4.42) (and usually reduces to this formula when f is convex quadratic and α_k is exact). The many variants of nonlinear CG are discussed in Nocedal and Wright (2006, chapter 5). There are some convergence results for these methods, but they are generally not as strong and those proved for the accelerated gradient methods that are the main focus of this chapter. Because these methods often perform well, we expect them to become topics of further investigation, so stronger results can be expected in future. (In contrast, the convergence theory for the conjugate gradient method applied to the convex quadratic case is extraordinarily rich, as also discussed in Nocedal and Wright (2006, chapter 5).)

4.6 Lower Bounds on Convergence Rates

The term "optimal" is used in connection with Nesterov's method because the convergence rate achieved by the method is the best possible (up to a constant), among algorithms that make use of gradient information at the iterates x^k and functions with Lipschitz continuous gradients. This claim can be proved by means of a carefully designed function for which *no* method that makes use of all gradients 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 by Nesterov (2004) is a convex quadratic $f(x) = \frac{1}{2}x^T Ax - e_1^T x$, 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 $i = 1, 2, \dots, n$. It is easy to show that $\|A\|_2 \le 4$, so that this function is L-smooth with $L = 4$.

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

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, $j = 0, 1, \ldots, k$, 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

$$\|x^k - x^*\|^2 \geq \sum_{j=k+1}^{n} (x_j^*)^2 = \sum_{j=k+1}^{n} \left(1 - \frac{j}{n+1}\right)^2. \tag{4.43}$$

A little arithmetic (see Exercises) shows that

$$\|x^k - x^*\|^2 \geq \frac{1}{8}\|x^0 - x^*\|^2, \quad k = 1, 2, \ldots, \frac{n}{2} - 1. \tag{4.44}$$

It can be shown further that

$$f(x^k) - f^* \geq \frac{3}{8(k+1)^2}\|x^0 - x^*\|^2, \quad k = 1, 2, \ldots, \frac{n}{2} - 1. \tag{4.45}$$

This lower bound on $f(x^k) - x^*$ is within a constant factor of the upper bound (4.38) when we recall that $L = 4$ for this function.

The restriction $k < n/2$ in the preceding argument is not fully satisfying. A more compelling example would show that the lower bound (4.45) holds for all k.

Notes and References

A description of Chebyshev iterative methods for convex quadratics can be found in Golub and Van Loan (1996, chapter 10).

The use of ODE methodology to study continuous-time limits of momentum methods dates to the paper of Su et al. (2014). Many other papers that pursue this line of work have appeared in subsequent years; the following references give some idea of the scope of this work: Wibisono et al. (2016); Attouch et al. (2018); Maddison et al. (2018); Shi et al. (2018).

The heavy-ball method was described by Polyak (1964). Nesterov's method was described originally in Nesterov (1983). Convergence proofs based on bounding functions were given in the text (Nesterov, 2004). Our description of Lyapunov functions follows that of Lessard et al. (2016). The FISTA algorithm (Beck and Teboulle, 2009) extends a similar approach to problems in which the objective is a smooth convex function added to a simple (possibly nonsmooth)

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

convex function. (We consider functions with this structure further in Section 9.3.)

A momentum method whose analysis can be performed with geometric tools is described by Bubeck et al. (2015), and an approach based on “optimal quadratic averaging” is presented in Drusvyatskiy et al. (2018).

The conjugate gradient method was proposed by Hestenes and Steifel; their first comprehensive description is in Hestenes and Steifel (1952). There are many later treatments by other authors (for example, Golub and Van Loan, 1996). This method has become a workhorse in scientific computing for solving large systems of linear equations with symmetric positive definite matrices. Its extension to nonlinear function minimization was first proposed by Fletcher and Reeves (1964), and many variants followed. More information can be found in Nocedal and Wright (2006, chapter 5) and its extensive list of references.

Exercises

1. Define α and β in terms of b, μ, and Δt such that (4.4) corresponds to (4.3). Repeat the question for the case in which the term dx/dt is approximated by central differences:

$$\frac{x(t+\Delta t) - x(t-\Delta t)}{2\Delta t}.$$

2. Minimize a quadratic objective $f(x) = \frac{1}{2}x^T Ax$ 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); % different x0 for each trial
```

Run the code in each case until $f(x_k) \le \epsilon$ 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
- 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 what happens to the codes and algorithms in the previous question when we reset m to 0 (making f weakly convex). Comment in 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. Consider the function

$$f(x) = \begin{cases} 25x^2 & x < 1 \\ x^2 + 48x - 24 & 1 \le x \le 2 \\ 25x^2 - 48x + 72 & x > 2 \end{cases}.$$

(a) Prove f is strongly convex with parameter 2 and has L-Lipschitz gradients with $L = 50$.

(b) What is the global minimizer of f? Justify your answer.

(c) Run the gradient method with steplength $1/50$, Nesterov's method with steplength $1/50$ and $\beta = 2/3$, and the heavy-ball method with $\alpha = 1/18$ and $\beta = 4/9$, starting from $x_0 = 3$ in each case. Plot the function value versus the iteration counter for each method. For each method, also plot the worst-case upper bounds on the function value as derived for the case in which f is a strongly convex quadratic with $m = 2$ and $L = 50$. Explain how the actual performance relates to the worst-case upper bound for quadratic functions.

5. Prove using Gelfand's formula (4.18) that (4.19) is true for any $\epsilon > 0$, for some $C_\epsilon > 1$.

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press

6. 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 Section 4.2 to a large extent, proceeding 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 $\rho(T) = \sqrt{\beta} \approx 1 - \sqrt{\kappa}$.

7. 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 (4.23).

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

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

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

   (The bound (4.44) follows by setting $k = \frac{n}{2} - 1$ in this expression and noting that it is decreasing in k.)

10. Show that the two formulas in (4.42) for the parameter γ_k in the conjugate gradient method are, in fact, equal by making use of the formulas (4.40) and (4.41).

https://doi.org/10.1017/9781009004282.005 Published online by Cambridge University Press