Chapter 3

Elementary Descent Methods

Methods that use information about gradients to obtain descent in the objective function at each iteration form the basis of all of the schemes studied in this book. We describe several methods of this type, along with analysis of their convergence and complexity properties. This chapter can be read as an introduction both to the gradient methods and to the fundamental tools of analysis that are used to understand optimization algorithms.

Throughout the chapter, we consider the unconstrained minimization of a smooth convex function:

$$\min_{x \in \mathbb{R}^n} f(x). \tag{3.1}$$

The algorithms we consider in this chapter are suited to the case in which f and its gradient ∇f can be evaluated—exactly, in principle—at arbitrary points x. Bearing in mind that this setup may not hold for many data analysis problems, we focus on those fundamental algorithms that can be extended to more general situations, for example:

- Objectives consisting of a smooth convex term plus a nonconvex regularization term;
- Minimization of smooth functions over simple constraint sets, such as bounds on the components of x;
- Functions for which f of ∇f cannot be evaluated exactly without a complete sweep through the data set, but unbiased estimates of ∇f can be obtained easily.
- Situations in which it is much cheaper to evaluate an individual component or a subvector of ∇f than the full gradient vector.
- Smooth but nonconvex f.

Extensions to the fundamental methods of his chapter, which allow us to handle these more general cases, will be considered in subsequent chapters.

3.1 Descent Directions

Most of the algorithms we will consider in this book generate a sequence of iterates $\{x^k\}$ for which the function values decrease at each iteration, that is, $f(x^{k+1}) < f(x^k)$ for each $k = 0, 1, 2, \ldots$.

Line-search methods proceed by identifying a direction d from each x such that f decreases as we move in the direction d. This notion can be formalized by the following definition:

Definition 3.1. *d is a descent direction for f at x if $f(x+td) < f(x)$ for all $t > 0$ sufficiently small.*

A simple, sufficient characterization of descent directions is given by the following proposition.

Proposition 3.2. *If f is continuously differentiable in a neighborhood of x, then any d such that $d^T \nabla f(x) < 0$ is a descent direction.*

Proof. We use Taylor's theorem — Theorem 2.1. By continuity of ∇f, we can identify $\bar{t} > 0$ such that $\nabla f(x+td)^T d < 0$ for all $t \in [0, \bar{t}]$. Thus from (2.3), we have for any $t \in (0, \bar{t}]$ that

$$f(x+td) = f(x) + t\nabla f(x+\gamma td)^T d, \quad \text{some } \gamma \in (0,1),$$

from which it follows that $f(x+td) < f(x)$, as claimed. □

Note that among all directions with unit norm,

$$\inf_{\|d\|=1} d^T \nabla f(x) = -\|\nabla f\|, \quad \text{achieved when } \; d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}.$$

For this reason, we refer to $-\nabla f(x)$ as the direction of *steepest descent*.

Since this direction always provides a descent direction, perhaps the simplest method for optimization of a smooth function has the iterations

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k), \quad k = 0, 1, 2, \ldots, \tag{3.2}$$

for some steplength $\alpha_k > 0$. At each iteration, we are guaranteed that there is either some nonnegative step α which decreases the function value unless $\nabla f(x_k) = 0$. But note that when $\nabla f(x) = 0$, we will have found a point which satisfies a necessary condition of local optimality. Moreover, if f is convex, we will have computed a global minimizer of f. This algorithm is called the *gradient method* or the *method of steepest descent*. In the next section, we will analyze how many iterations are required to find points where the gradient nearly vanishes.

3.2 Steepest Descent

We first focus on the question of choosing the stepsize α_k for the steepest descent method (3.3). If α_k is too large, we risk taking a step that increases the function value. On the other hand, if α_k is too small, we risk making too little progress and thus requiring too many iterations to find a solution.

The simplest stepsize protocol is the short-step variant of steepest descent. We assume here that f is L-smooth (see the definition in (2.7)). In this case, we can set α_k to be a constant value α, and simply iterate as follows:

$$x^{k+1} = x^k - \alpha \nabla f(x^k), \quad k = 0, 1, 2, \ldots . \tag{3.3}$$

To estimate the amount of decrease in f obtained at each iterate of this method, we use Taylor's theorem. By setting $p = \alpha d$ in (2.2), we obtain

$$\begin{aligned} f(x+\alpha d) &= f(x) + \alpha \nabla f(x)^T d + \alpha \int_0^1 [\nabla f(x+\gamma\alpha d) - \nabla f(x)] d \, d\gamma \\ &\leq f(x) + \alpha \nabla f(x)^T d + \alpha \int_0^1 \|\nabla f(x+\gamma\alpha d) - \nabla f(x)\| \|d\| \, d\gamma \\ &\leq f(x) + \alpha \nabla f(x)^T d + \alpha^2 \frac{L}{2} \|d\|^2, \end{aligned} \tag{3.4}$$

where we used (2.7) for the last line. For $x = x^k$ and $d = -\nabla f(x^k)$, the value of α that minimizes the expression on the right-hand side is $\alpha = 1/L$. By substituting these values, we obtain

$$f(x^{k+1}) = f(x^k - (1/L)\nabla f(x^k)) \leq f(x^k) - \frac{1}{2L} \|\nabla f(x^k)\|^2. \tag{3.5}$$

This expression is one of the foundational inequalities in the analysis of optimization methods. It quantifies the amount of decrease we can obtain from the function f to two critical quantities: the norm of the gradient $\nabla f(x^k)$ at the current iterate, and the Lipschitz constant L of the gradient. Depending on the other assumptions about f, we can derive a variety of different convergence rates from this basic inequality, as we now show.

3.2.1 General Case

From (3.5) alone, we can already say something about the rate of convergence of steepest descent, provided we assume that f is bounded below. That is, we assume that there is a value $\bar{f}$ such that

$$f(x) \geq \bar{f}, \quad \text{for all } x. \tag{3.6}$$

In the case that f has a minimizer x^*, we can define $\bar{f} = f(x^*)$.

Unwinding the inequalities (3.5), we find that

$$f(x^T) \leq f(x^0) - \frac{1}{2L} \sum_{k=0}^{T-1} \|\nabla f(x^k)\|^2$$

Since $\bar{f} \leq f(x^T)$, we have

$$\sum_{k=0}^{T-1} \|\nabla f(x^k)\|^2 \leq 2L[f(x^0) - \bar{f}] \, .$$

This implies that $\lim_{T\to\infty} \|\nabla f(x_T)\| = 0$. More concretely,

$$\min_{0\leq k\leq T-1} \|\nabla f(x^k)\| \leq \sqrt{\frac{2L[f(x^0) - f(x^T)]}{T}} \leq \sqrt{\frac{2L[f(x^0) - \bar{f}]}{T}} \, .$$

Thus, we have shown that after T steps of steepest descent, we can find a point x satisfying

$$\|\nabla f(x)\| \leq \sqrt{\frac{2L[f(x^0) - \bar{f}]}{T}}. \tag{3.7}$$

Note that this convergence rate is very slow, and only tells us that we will find a nearly stationary point. We need more structure about f to guarantee faster convergence and global optimality.

3.2.2 Convex Case

When f is also convex, we have the following stronger result for the steepest descent method.

Theorem 3.3. *Suppose that f is convex and L-smooth, and that (3.1) has a solution x^*. Then the steepest descent method with stepsize $\alpha_k \equiv 1/L$ generates a sequence $\{x^k\}_{k=0}^{\infty}$ that satisfies*

$$f(x^T) - f^* \leq \frac{L}{2T}\|x^0 - x^*\|^2, \quad T = 1, 2, \ldots . \tag{3.8}$$

Proof. By convexity of f, we have $f(x^*) \geq f(x^k) + \nabla f(x^k)^T(x^* - x^k)$, so by substituting into the key inequality (3.5), we obtain for $k = 0, 1, 2, \ldots$ that

$$\begin{aligned} f(x^{k+1}) &\leq f(x^*) + \nabla f(x^k)^T(x^k - x^*) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \\ &= f(x^*) + \frac{L}{2}\left(\|x^k - x^*\|^2 - \|x^k - x^* - \frac{1}{L}\nabla f(x^k)\|^2\right) \\ &= f(x^*) + \frac{L}{2}\left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right). \end{aligned}$$

By summing over $k = 0, 1, 2, \ldots, T-1$, we have

$$\begin{aligned} \sum_{k=0}^{T-1}(f(x^{k+1}) - f^*) &\leq \frac{L}{2}\sum_{k=0}^{T-1}\left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right) \\ &= \frac{L}{2}\left(\|x^0 - x^*\|^2 - \|x^T - x^*\|^2\right) \\ &\leq \frac{L}{2}\|x^0 - x^*\|^2. \end{aligned}$$

Since $\{f(x^k)\}$ is a nonincreasing sequence, we have

$$f(x^T) - f(x^*) \leq \frac{1}{T}\sum_{k=0}^{T-1}(f(x^{k+1}) - f^*) \leq \frac{L}{2T}\|x^0 - x^*\|^2,$$

as required. □

3.2.3 Strongly Convex Case

Recall from (2.19) that the smooth function $f : \mathbb{R}^d \to \mathbb{R}$ is *strongly convex* if there is a scalar $m > 0$ such that

$$f(z) \geq f(x) + \nabla f(x)^T(z - x) + \frac{m}{2}\|z - x\|^2 \tag{3.9}$$

Strong convexity asserts that f can be lower bounded by quadratic functions. These functions change from point to point, but only in the linear term. It also tells us that the curvature of the function is bounded away from zero. Note that if f is strongly convex *and* L-smooth, then f is bounded above and below by simple quadratics (see (2.9) and (2.19)). This "sandwiching" effect enables us to prove the linear convergence of the gradient method.

The simplest strongly convex function is the squared Euclidean norm $\|x\|^2$. Any convex function can be perturbed to form a strongly convex function by adding any small multiple of the squared Euclidean norm. In fact, if f is any L-smooth function, then

$$f_\mu(x) = f(x) + \mu\|x\|^2$$

is strongly convex for μ large enough. Verifying this fact is a interesting exercise.

As another canonical example, note that a quadratic function $f(x) = \frac{1}{2}x^TQx$ is strongly convex if and only if the smallest eigenvalue of Q is strictly positive. We saw in Theorem 2.8 that a strongly convex f has a unique minimizer, which we denote by x^*.

Strongly convex functions are in essence the "easiest" functions to optimize by first-order methods. First, the norm of the gradient provides useful information about how far away we are from optimality. Suppose we minimize both sides of the inequality (3.9) with respect to z. The minimizer on the left-hand side is clearly attained at $z = x^*$, while on the right-hand side it is attained at $x - \nabla f(x)/m$. By plugging these optimal values into (3.9), we obtain

$$\begin{aligned} f(x^*) &\geq f(x) - \nabla f(x)^T\left(\frac{1}{m}\nabla f(x)\right) + \frac{m}{2}\left\|\frac{1}{m}\nabla f(x)\right\|^2 \\ &= f(x) - \frac{1}{2m}\|\nabla f(x)\|^2. \end{aligned}$$

By rearrangement, we obtain

$$\|\nabla f(x)\|^2 \geq 2m[f(x) - f(x^*)]. \tag{3.10}$$

If $\|\nabla f(x)\| < \delta$ then

$$f(x) - f(x^*) \leq \frac{\|\nabla f(x)\|^2}{2m} \leq \frac{\delta^2}{2m}.$$

Thus, when the gradient is small, we are close to having found a point with minimal function value. We can even derive a stronger result about the distance of x to the optimal point x^*. Using (3.9) and Cauchy-Schwartz, we have

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T(x^* - x) + \frac{m}{2}\|x - x^*\|^2 \\ &\geq f(x) - \|\nabla f(x)\|\,\|x^* - x\| + \frac{m}{2}\|x - x^*\|^2 \end{aligned}$$

Rearranging terms proves that

$$\|x - x^*\| \leq \frac{2}{m}\|\nabla f(x)\|. \tag{3.11}$$

This says we can estimate the distance to the optimal value purely in terms of the norm of the gradient.

We summarize this discussion in the following

Lemma 3.4. *Let f be a strongly convex function with modulus m. Then we have*

$$f(x) - f(x^*) \leq \frac{\|\nabla f(x)\|^2}{2m} \tag{3.12}$$

$$\|x - x^*\| \leq \frac{2}{m}\|\nabla f(x)\|. \tag{3.13}$$

We can now proceed to analyze the convergence of gradient descent on strongly convex functions. By substituting (3.12) into our basic inequality (3.5), we obtain

$$f(x^{k+1}) = f\left(x^k - \frac{1}{L}\nabla f(x^k)\right) \leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \leq f(x^k) - \frac{m}{L}(f(x^k) - f^*).$$

Subtracting f^* from both sides of this inequality gives us the recursion

$$f(x^{k+1}) - f^* \leq \left(1 - \frac{m}{L}\right)(f(x^k) - f^*). \tag{3.14}$$

Thus the function values converge *linearly* to the optimum. After T steps, we have

$$f(x^T) - f^* \leq \left(1 - \frac{m}{L}\right)^T (f(x^0) - f^*). \tag{3.15}$$

3.2.4 Comparison Between Rates

It is straightforward to convert these convergence expressions into complexities, using the techniques of Appendix A.2. We have from (3.7) that an iterate k will be found such that $\|\nabla f(x^k)\| \leq \epsilon$ for some $k \leq T$, where

$$T \geq \frac{2L(f(x^0) - f^*)}{\epsilon^2}.$$

For the weakly convex case, we have from (3.8) that $f(x^k) - f^* \leq \epsilon$ when

$$k \geq \frac{L\|x^0 - x^*\|^2}{2\epsilon}. \tag{3.16}$$

For the strongly convex case, we have from (3.15) that $f(x^k) - f^* \leq \epsilon$ for all k satisfying

$$k \geq \frac{L}{m}\log((f(x^0) - f^*)/\epsilon). \tag{3.17}$$

Note that in all three cases, we can get bounds in terms of the distance initial distance to optimality $\|x^0 - x^*\|$ rather than in terms of the initial optimality gap $f(x^0) - f^*$ by using the inequality

$$f(x^0) - f^* \leq \frac{L}{2}\|x^0 - x^*\|^2.$$

The linear rate (3.17) depends only logarithmically on ϵ, whereas the sublinear rates depend on $1/\epsilon$ or $1/\epsilon^2$. When ϵ is small (for example $\epsilon = 10^{-6}$), the linear rate would appear to be dramatically faster, and indeed this is usually the case. The only exception would be when m is extremely small, so that L/m is of the same order as ϵ. The problem is extremely ill conditioned in this case, and there is little difference between the linear rate (3.17) and the sublinear rate (3.16).

All of these bounds depend on knowledge of the curvature parameter L. What happens when we don't know L? Even when we do know it, is the steplength $\alpha_k \equiv 1/L$ good? We have reason to suspect not, since the inequality (3.5) on which it is based uses the conservative global upper bound L on curvature. (A sharper bound could be obtained in terms of the curvature in the neighborhood of the current iterate x^k.) In the remainder of this chapter, we expand our view to more general choices of search directions and stepsizes.

3.3 Line-Search Methods: Convergence

In the previous section we considered the short-step gradient method that followed the negative gradient with a stepsize determined by the global curvature of the gradient, $1/L$. In this section, we generalize the convergence results to more generic descent methods. Suppose each step has the form

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, 2, \ldots, \tag{3.18}$$

where d^k is a descent direction and α_k is a positive stepsize. What do we need to guarantee convergence to a stationary point at a particular rate? What do we need to guarantee convergence of the iterates themselves?

Recall that our analysis of steepest-descent algorithm with fixed stepsize $1/L$ in the previous section was based on the bound (3.5), which showed that the amount of decrease in f at iteration k is at least a multiple of $\|\nabla f(x^k)\|^2$. In the discussion below, we show that the same estimate of function decrease, except for a different constant, can be obtained for many line-search methods of the form (3.18), provided that d^k and α_k satisfy certain intuitive properties. Specifically, we show that the following inequality holds:

$$f(x^{k+1}) \leq f(x^k) - C\|\nabla f(x^k)\|^2, \quad \text{for some } C > 0. \tag{3.19}$$

The remainder of the analyses used properties about the function f itself that were independent of the algorithm: smoothness, convexity, and strong convexity. For a general descent method, we can provide similar analyses based on the property (3.19).

What can we say about the sequence of iterates $\{x^k\}$ generated by such a scheme? We state an elementary theorem.

Theorem 3.5. *Suppose that f is bounded below, with Lipschitz continuous gradient. Then all accumulation points $\bar{x}$ of the sequence $\{x^k\}$ generated by a scheme that satisfies* (3.19) *are stationary, that is, $\nabla f(\bar{x}) = 0$. If in addition f is convex, each such $\bar{x}$ is a solution of (3.1).*

Proof. Note first from (3.19) that

$$\|\nabla f(x^k)\|^2 \leq [f(x^k) - f(x^{k+1})]/C, \quad k = 0, 1, 2, \ldots, \tag{3.20}$$

and since $\{f(x^k)\}$ is a decreasing sequence that is bounded below, it follows that $\lim_{k\to\infty} f(x^k) - f(x^{k+1}) = 0$. If $\bar{x}$ is an accumulation point, there is a subsequence $\mathcal{S}$ such that $\lim_{k\in\mathcal{S},k\to\infty} x^k = \bar{x}$. By continuity of ∇f, we have $\nabla f(\bar{x}) = \lim_{k\in\mathcal{S},k\to\infty} \nabla f(x^k) = 0$, as required. If f is convex, each such $\bar{x}$ satisfies the first-order sufficient conditions to be a solution of (3.1). □

It is possible for the the sequence $\{x^k\}$ to be unbounded and have no accumulation points. For example, some descent methods applied to the scalar function $f(x) = e^{-x}$ will generate iterates that diverge to ∞. (This function is convex and bounded below but does not attain its minimum value.)

We can prove other results about *rates* of convergence of algorithms (3.18) satisfying (3.19), using almost identical proofs to those of Section 3.2. For example, for the case in which f is bounded below by some quantity $\bar{f}$, we can show using the techniques of Section 3.2.1 that

$$\min_{0\leq k\leq T-1} \|\nabla f(x^k)\| \leq \sqrt{\frac{f(x^0) - \bar{f}}{CT}}.$$

For the case in which f is strongly convex with modulus m (and unique solution x^*), we can combine (3.12) with (3.19) to deduce that

$$f(x^{k+1}) - f(x^*) \leq f(x^k) - f(x^*) - C\|\nabla f(x^k)\|^2 \leq (1 - 2mC)[f(x^k) - f(x^*)],$$

which indicates linear convergence with rate $(1 - 2mC)$.

Interestingly, the argument of Section 3.2.2 concerning rate of convergence for the (non-strongly) convex case cannot be generalized to this setting, though similiar results can be obtained by another technique under an additional assumption, as we now show.

Theorem 3.6. *Suppose that f is convex and smooth, where ∇f has Lipschitz constant L, and that (3.1) has a solution x^*. Assume moreover that the level set defined by x^0 is bounded in the sense that $R_0 < \infty$, where*

$$R_0 := \max \|\|x - x^*\| \mid f(x) \leq f(x^0)\}.$$

Then a descent method satisfying (3.19) generates a sequence $\{x^k\}_{k=0}^{\infty}$ that satisfies

$$f(x^T) - f^* \leq \frac{R_0^2}{CT} \quad T = 1, 2, \ldots . \tag{3.21}$$

Proof. Defining $\Delta_k := f(x^k) - f(x^*)$, we have that

$$\Delta_k = f(x^k) - f(x^*) \leq \nabla f(x^k)^T (x^k - x^*) \leq R_0 \|\nabla f(x^k)\|.$$

By substituting this bound into (3.19), we obtain

$$f(x^{k+1}) \leq f(x^k) - \frac{C}{R_0^2}\Delta_k^2,$$

which after subtracting $f(x^*)$ from both sides and using the definition of Δ_k becomes

$$\Delta_{k+1} \leq \Delta_k - \frac{C}{R_0^2}\Delta_k^2 = \Delta_k \left(1 - \frac{C}{R_0^2}\Delta_k\right). \tag{3.22}$$

By inverting both sides, we obtain

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} \frac{1}{1 - \frac{C}{R_0^2}\Delta_k}$$

Since $\Delta_{k+1} \geq 0$, we have from (3.22) that $\frac{C}{R_0^2}\Delta_k \in [0, 1]$, so using the fact that $\frac{1}{1-\epsilon} \geq 1 + \epsilon$ for all $\epsilon \in [0, 1]$, we obtain

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k}\left(1 + \frac{C}{R_0^2}\Delta_k\right) = \frac{1}{\Delta_k} + \frac{C}{R_0^2}.$$

By applying this formula recursively, we have for any $T \geq 1$ that

$$\frac{1}{\Delta_T} \geq \frac{1}{\Delta_0} + \frac{TC}{R_0^2} \geq \frac{TC}{R_0^2},$$

and we obtain the result by taking the inverse of both sides in this bound and using $\Delta_T = f(x^T) - f(x^*)$. □

3.4 Line Search Methods: Choosing the Direction

In this section, we turn to consider methods analysis of generic line-search descent methods, which take steps of the form (3.18), where $\alpha_k > 0$ and d^k is a search direction that satisfies the following properties, for some positive constants $\bar{\epsilon}$, γ_1, γ_2:

$$0 < \bar{\epsilon} \leq \frac{-(d^k)^T \nabla f(x^k)}{\|\nabla f(x^k)\| \|d^k\|}, \tag{3.23a}$$

$$0 < \gamma_1 \leq \frac{\|d^k\|}{\|\nabla f(x^k)\|} \leq \gamma_2. \tag{3.23b}$$

Condition (3.23a) says that the angle between $-\nabla f(x^k)$ and d^k is acute, and bounded away from $\pi/2$, while condition (3.23b) ensures that d^k and $\nabla f(x^k)$ are not too much different in length. (If x^k is a stationary point, we have $\nabla f(x^k) = 0$ so our algorithm will set $d^k = 0$ and terminate.)

For the "obvious" choice of search direction—the negative gradient $d^k = -\nabla f(x^k)$—the conditions (3.23) hold trivially, with $\bar{\epsilon} = \gamma_1 = \gamma_2 = 1$.

We can use Taylor's theorem to bound the change in f when we move along d^k from the current iteration x^k. By setting $x = x^k$ and $p = \alpha d^k$ in (2.2), we obtain

$$\begin{aligned}
f(x^{k+1}) &= f(x^k + \alpha d^k) \\
&= f(x^k) + \alpha \nabla f(x^k)^T d^k + \alpha \int_0^1 [\nabla f(x^k + \gamma \alpha d^k) - \nabla f(x^k)] d^k \, d\gamma \\
&\leq f(x^k) + \alpha \nabla f(x^k)^T d^k + \alpha \int_0^1 \|\nabla f(x^k + \gamma \alpha d^k) - \nabla f(x^k)\| \|d^k\| \, d\gamma \\
&\leq f(x^k) + \alpha \nabla f(x^k)^T d^k + \alpha^2 \int_0^1 \gamma L \|d^k\|^2 \, d\gamma \\
&\leq f(x^k) - \alpha \bar{\epsilon} \|\nabla f(x^k)\| \|d^k\| + \alpha^2 \frac{L}{2} \|d^k\|^2 \\
&\leq f(x^k) - \alpha \left(\bar{\epsilon} - \alpha \frac{L}{2} \gamma_2 \right) \|\nabla f(x^k)\| \|d^k\|,
\end{aligned} \tag{3.24}$$

where we used (2.7) for the second-last line and (3.23) throughout. It is clear from this expression that for all values of α sufficiently small—to be precise, for $\alpha \in (0, 2\bar{\epsilon}/(L\gamma_2))$—we have $f(x^{k+1}) < f(x^k)$, unless of course x^k is a stationary point.

In deriving the bound (3.24), we did not require convexity of f, only Lipschitz continuity of the gradient ∇f. The same is true for most of the analysis in this section. Convexity is used only in proving rates of convergence to a solution x^*, in Sections 3.3 and 3.2. (Even there, we could relax the convexity assumption to obtain results about convergence to stationary points.)

We mention a few possible choices of d^k apart from the negative gradient direction $-\nabla f(x^k)$.

- The transformed negative gradient direction $d^k = -S^k \nabla f(x^k)$, where S^k is a symmetric positive definite matrix with eigenvalues in the range $[\gamma_1, \gamma_2]$, where γ_1 and γ_2 are positive quantities as in (3.23). The second condition in (3.23) hold, by definition of S^k, and the first condition holds with $\bar{\epsilon} = \gamma_1/\gamma_2$, since

$$-(d^k)^T \nabla f(x^k) = \nabla f(x^k)^T S^k \nabla f(x^k) \geq \gamma_1 \|\nabla f(x^k)\|^2 \geq (\gamma_1/\gamma_2) \|\nabla f(x^k)\| \|d^k\|.$$

Newton's method, which chooses $S^k = \nabla^2 f(x^k)$, would satisfy this condition provided the true Hessian has eigenvalues uniformly bounded in the range $[1/\gamma_2, 1/\gamma_1]$ for all x^k.

- The Gauss-Southwell variant of coordinate descent chooses $d^k = -[\nabla f(x^k)]_{i_k}$, where $i_k = \arg\min_{i=1,2,\ldots,n} |[\nabla f(x^k)]_i|$. (We leave it as an exercise to show that the conditions (3.23) are satisfied for this choice of d^k.) There does not seem to be an obvious reason to use this search direction. Since it is defined in terms of the full gradient $\nabla f(x^k)$, why not use $d^k = -\nabla f(x^k)$ instead? The answer (as we discuss further in Chapter 6) is that for some important kinds of functions f, the gradient $\nabla f(x^k)$ can be updated efficiently to obtain $\nabla f(x^{k+1})$ provided that x^k and x^{k+1} differ in only a single coordinate. These cost savings make coordinate descent methods competitive with, and often faster than, full-gradient methods.

- Some algorithms make *randomized* choices of d^k in which the conditions (3.23) hold in the sense of expectation, rather than deterministically. In one variant of stochastic coordinate descent, we set $d^k = -[\nabla f(x^k)]_{i_k}$, for i_k chosen uniformly at random from $\{1, 2, \ldots, n\}$ at each k. Taking expectations over i_k, we have

$$\mathbb{E}_{i_k}\left((-d^k)^T \nabla f(x^k)\right) = \frac{1}{n}\sum_{i=1}^{n} [\nabla f(x^k)]_i^2 = \frac{1}{n}\|\nabla f(x^k)\|^2 \geq \frac{1}{n}\|\nabla f(x^k)\|\|d^k\|,$$

  where the last inequality follows from $\|d^k\| \leq \|\nabla f(x^k)\|$, so the first condition in (3.23) holds in an expected sense. We have that $E(\|d^k\|^2) = \frac{1}{n}\|\nabla f(x^k)\|_2^2$, so the norms of $\|d^k\|$ and $\|\nabla f(x^k)\|$ are also similar to within a scale factor, so the first part of (3.23) also holds in an expected sense. Rigorous analysis of these methods is presented in Chapter 6.

- Another important class of randomized schemes are the stochastic gradient methods discussed in Chapter 5. In place of an exact gradient $\nabla f(x^k)$, these method typically have access to a vector $g(x^k, \xi_k)$, where ξ_k is a random variable, such that $\mathbb{E}_{\xi_k} g(x^k, \xi_k) = \nabla f(x^k)$. That is, $g(x^k, \xi_k)$ is an unbiased (but often very noisy) estimate of the true gradient $\nabla f(x^k)$. Again, if we set $d^k = -g(x^k, \xi_k)$, the conditions (3.23) hold in an expected sense, though the bound $\mathbb{E}(\|d^k\|) \leq \gamma_2 \|\nabla f(x^k)\|$ requires additional conditions on the distribution of $g(x^k, \xi_k)$ as a function of ξ_k. Further analysis of stochastic gradient methods appears in Chapter 5.

3.5 Line Search Methods: Choosing the Steplength

Each iteration of a typical descent algorithm has two ingredients: a *search direction* d^k, which is typically related to the negative of the search direction, and a *step length* $\alpha_k > 0$, which is the scalar multiple applied to the search direction to obtain the step. The iteration therefore has the form

$$x^{k+1} = x^k + \alpha_k d^k. \tag{3.25}$$

We assume for this discussion that d^k satisfies the properties (3.23). We now turn to the issue of choosing α_k, which often requires a subroutine designed specifically for computing α_k at each iteration. We emphasize that even for the gradient method, when you don't know the parameter L, some method will be required to find a stepsize to guarantee a sufficient decrease like (3.19).

There are several alternative approaches, of varying theoretical and practical validity.

Constant Stepsize. As we have seen in Section 3.2, constant stepsizes can yield rapid convergence rates. The main drawback of the constant stepsize method is that one needs some prior information to properly choose the stepsize.

The first approach to choosing a constant stepsize (one commonly used in machine learning, where the step length is often known as the "learning rate") is trial and error. Extensive experience in applying gradient (or stochastic gradient) algorithms to a particular class of problems may reveal that a particular stepsize is reliable and reasonably efficient. Typically, a reasonable heuristic is to pick α as large as possible such that the algorithm doesn't diverge. In some sense, this approach is estimating the Lipschitz constant of the gradient of f by trial and error. Slightly enhanced variants are also possible, for example, α_k may be held constant for many successive iterations then decreased periodically. Since such schemes are highly application- and problem-dependent, we cannot say much more about them here.

A second approach is to base the choice of α_k on knowledge of the global properties of the function f, for example, on the Lipschitz constant L for the gradient (see (2.7)) or the modulus of convexity μ (see (2.18)). We call such variants "short-step" methods. Given the expression (3.24) above, for example, and supposing we have estimates of all the quantities γ_1, γ_2, and L that appear therein, we could choose α to maximize the coefficient of the last term. Setting $\alpha = \bar{\epsilon}/(L\gamma_2)$, we obtain from (3.24) and (3.23) that

$$f(x^{k+1}) \leq f(x^k) - \frac{\bar{\epsilon}^2}{2L\gamma_2}\|\nabla f(x^k)\|\|d^k\| \geq f(x^k) - \frac{\bar{\epsilon}^2\gamma_1}{2L\gamma_2}\|\nabla f(x^k)\|^2. \tag{3.26}$$

Thus, the amount of decrease in f at iteration k is at least a positive multiple of the squared gradient norm $\|\nabla f(x^k)\|^2$.

Exact Line Search. Once we have chosen a descent direction, we can minimizing the function restricted to this direction. That is, we can perform a one-dimensional line search along direction d^k to find an approximate solution of the following problem:

$$\min_{\alpha>0} f(x^k + \alpha d^k). \tag{3.27}$$

This technique requires evalaution of $f(x^k + \alpha d^k)$ (and possibly also its derivative with respect to α, namely $(d^k)^T\nabla f(x^k + \alpha d^k)$) economically, for arbitrary positive values of α. There are many cases where these line searches can be computed at low cost. For example, if f is a multivariate polynomial, the line search amounts to minimizing a univariate polynomial. Such a minimization can be performed by finding the roots of the polynomial, and then testing each root to find the minimum. In other settings, such as coordinate descent methods of Chapter 6, it is possible to evaluate $f(x^k + \alpha d^k)$ cheaply for certain f, provided that d^k is a coordinate direction. Convergence analysis for exact line search methods tracks that for the short-step methods above. Since the exact minimizer of $f(x^k + \alpha d^k)$ will achieve at least as much reduction in f as the choice $\alpha = \bar{\epsilon}/(L\gamma_2)$ used to derive the estimate (3.26), it is clear that (3.26) also holds for exact line searches.

Approximate Line Search. In full generality, exact line searches are expensive and unnecessary. Better empirical performance is achieved by approximate line search. There was a lot of research in the 1970s and 1980s on finding conditions that should be satisfied by *approximate* line searches so as to guarantee good convergence properties, and on identifying line-search procedures which find

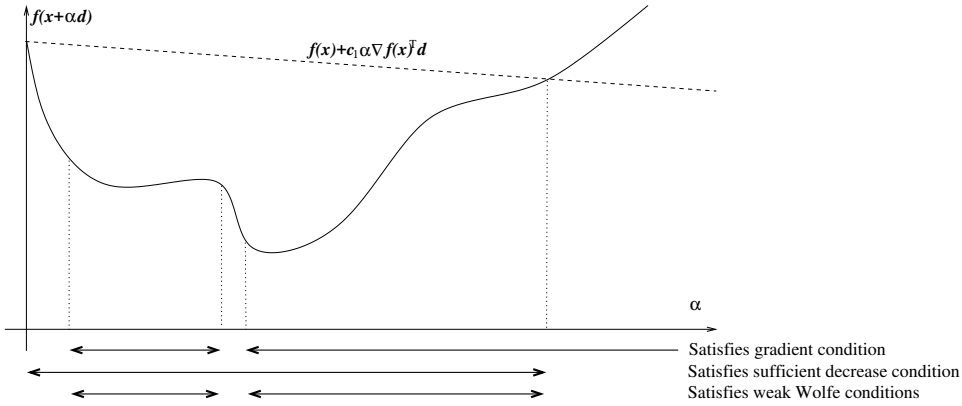

Figure 3.1: Weak Wolfe conditions are satisfied when both the gradient condition (3.28b) and the sufficient decrease condition (3.28a) hold.

such approximate solutions economically. (By "economically," we mean that an average of three or less evaluations of f are required.) One popular pair of conditions that the approximate minimizer $\alpha = \alpha_k$ is required to satisfy, called the *Weak Wolfe Conditions*, is as follows:

$$f(x^k + \alpha d^k) \leq f(x^k) + c_1 \alpha \nabla f(x^k)^T d^k, \tag{3.28a}$$

$$\nabla f(x^k + \alpha d^k)^T d^k \geq c_2 \nabla f(x^k)^T d^k \,. \tag{3.28b}$$

Here, c_1 and c_2 are constants that satisfy $0 < c_1 < c_2 < 1$. The condition (3.28a) is often known as the "sufficient decrease condition," because it ensures that the actual amount of decrease in f is at least a multiple c_1 of the amount suggested by the first-order Taylor expansion. The second condition (3.28b), which we call the "gradient condition," ensures that α_k is not too short; it ensures that we move far enough along d^k that the directional derivative of f along d^k is substantially less negative than its value at $\alpha = 0$, or is zero or positive. These conditions are illustrated in Figure 3.1.

It can be shown that there exist values of α_k that satisfy both weak Wolfe conditions simultaneously. To show that these conditions imply a reduction in f that is related to $\|\nabla f(x^k)\|^2$ (as in (3.26)), we argue as follows. First, from condition (3.28b) and the Lipschitz property for ∇f, we have

$$-(1 - c_2)\nabla f(x^k)^T d^k \leq [\nabla f(x^k + \alpha_k d^k) - \nabla f(x^k)]^T d^k \leq L\alpha_k \|d^k\|^2,$$

and thus

$$\alpha_k \geq -\frac{(1 - c_2)}{L} \frac{\nabla f(x^k)^T d^k}{\|d^k\|^2}.$$

Substituting into (3.28a), and using the first condition in (3.23), then yields

$$\begin{aligned} f(x^{k+1}) = f(x^k + \alpha_k d^k) &\leq f(x^k) + c_1 \alpha_k \nabla f(x^k)^T d^k \\ &\leq f(x^k) - \frac{c_1(1 - c_2)}{L} \frac{(\nabla f(x^k)^T d^k)^2}{\|d^k\|^2} \\ &\leq f(x^k) - \frac{c_1(1 - c_2)}{L} \bar{\epsilon}^2 \|\nabla f(x^k)\|^2. \end{aligned}$$

Algorithm 3.1 (from [12]) describes an approach that combines extrapolation with bisection to find a steplength α satisfying the conditions (3.42). This method maintains a subinterval $[L, U]$ of the positive real line (intially $L = 0$ and $U = \infty$) that contains a point satisfying (3.42), along with a current guess $\alpha \in (L, U)$ of this point. If the sufficient decrease condition (3.28a) is violated by α, then the current guess is too long, so the upper bound U is assigned the value α, and the new guess is taken to be the midpoint of the new interval $[L, U]$. If the sufficient decrease condition holds by the condition (3.28b) is violated, the current guess of α is too short. In this case, we move the lower bound up to α, and take the next guess of α to be either the midpoint of $[L, U]$ (if U is finite), or double the previous guess (if U is still infinite).

A rigorous proof that Algorithm 3.1 terminates with a value of α satisfying (3.42) can be found in Section A.3 in the Appendix.

Algorithm 3.1 Extrapolation-Bisection Line Search (EBLS)

Given $0 < c_1 < c_2 < 1$, set $L \leftarrow 0$, $U \leftarrow +\infty$, $\alpha \leftarrow 1$;
repeat
  if $f(x + \alpha d) > f(x) + c_1 \alpha \nabla f(x)^T d$ **then**
    Set $U \leftarrow \alpha$ and $\alpha \leftarrow (U + L)/2$;
  else if $\nabla f(x + \alpha d)^T d < c_2 \nabla f(x)^T d$ **then**
    Set $L \leftarrow \alpha$;
    if $U = +\infty$ **then**
      Set $\alpha \leftarrow 2L$;
    else
      Set $\alpha = (L + U)/2$;
    end if
  else
    Stop (Success!);
  end if
until Forever

Backtracking Line Search. Another popular approach to determining an appropriate value for α_k is known as "backtracking." It is widely used in situations where evaluation of f is economical and practical, while evaluation of the gradient ∇f is more difficult. It is easy to implement (no estimate of the Lipschitz constant L is required, for example) and still results in reasonably fast convergence.

In its simplest variant, we first try a value $\bar{\alpha} > 0$ as the initial guess of the steplength, and choose a constant $\beta \in (0, 1)$. The step length α_k is set to the first value in the sequence $\bar{\alpha}, \beta\bar{\alpha}, \beta^2\bar{\alpha}, \beta^3\bar{\alpha}, \ldots$ for which a sufficient decrease condition (3.28a) is satisfied. Note that backtracking does not require a condition like (3.28b) to be checked. The purpose of such a condition is to ensure that α_k is not too short, but this is not a concern in backtracking, because we know that α_k is either the fixed value $\bar{\alpha}$, or is within a factor β of a step length that is too long.

Under the assumptions above, we can again show that the decrease in f at iteration k is a positive multiple of $\|\nabla f(x^k)\|^2$. When no backtracking is necessary, that is, $\alpha_k = \bar{\alpha}$, we have from (3.23) that

$$f(x^{k+1}) \leq f(x^k) + c_1 \bar{\alpha} \nabla f(x^k)^T d^k \leq f(x^k) - c_1 \bar{\alpha} \bar{\epsilon} \gamma_1 \|\nabla f(x^k)\|^2. \tag{3.29}$$

When backtracking is needed, we have from the fact that the test (3.28a) is *not* satisfied for the previously tried value $\alpha = \beta^{-1}\alpha_k$ that

$$f(x^k + \beta^{-1}\alpha_k d^k) > f(x^k) + c_1\beta^{-1}\alpha_k \nabla f(x^k)^T d^k.$$

By a Taylor series argument like the one in (3.24), we have

$$f(x^k + \beta^{-1}\alpha_k d^k) \leq f(x^k) + \beta^{-1}\alpha_k \nabla f(x^k)^T d^k + \frac{L}{2}(\beta^{-1}\alpha_k)^2 \|d^k\|^2.$$

From the last two inequalities and some elementary manipulation, we obtain that

$$\alpha_k \geq -\frac{2}{L}\beta(1-c_1)\frac{\nabla f(x^k)^T d^k}{\|d^k\|^2}.$$

By substituting into (3.28a) with $\alpha = \alpha_k$ (note that this condition is satisfied for this value of α) and then using (3.23), we obtain

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + c_1\alpha_k \nabla f(x^k)^T d^k \\ &\leq f(x^k) - \frac{2}{L}\beta(1-c_1)c_1\frac{(\nabla f(x^k)^T d^k)^2}{\|d^k\|^2} \\ &\leq f(x^k) - \frac{2}{L}\beta c_1(1-c_1)\bar{\epsilon}^2\|\nabla f(x^k)\|^2. \end{aligned} \tag{3.30}$$

3.6 Convergence to Approximate Second-Order Necessary Points

The line-search methods that we described so far in this chapter asymptotically satisfy first-order optimality conditions with certain complexity guarantees. We now describe an elementary method that is designed to find points that satisfy the second-order necessary conditions, which are

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \text{ positive semidefinite} \tag{3.31}$$

(see Theorem 2.4). Our method makes a further smoothness assumption on f. In addition to Lipschitz continuity of the gradient ∇f, we assume Lipschitz continuity of the Hessian $\nabla^2 f$. That is, we assume that there is a constant M such that

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq M\|x - y\|, \quad \text{for all } x, y \in \text{dom}\,(f). \tag{3.32}$$

By extending Taylor's theorem (Theorem 2.1) to a third-order term, and using the definition of M, we obtain the following cubic upper bound on f:

$$f(x+p) \leq f(x) + \nabla f(x)^T p + \frac{1}{2}p^T \nabla^2 f(x) p + \frac{1}{6}M\|p\|^3. \tag{3.33}$$

As in Section 3.2, we make an additional assumption that f is bounded below by $\bar{f}$.

We describe an elementary algorithm that makes use of the expansion (3.33) as well as the steepest-descent theory of Subsection 3.2. Our algorithm aims to identify a point that *approximately* satisfies the second-order necessary conditions (3.31), that is,

$$\|\nabla f(x)\| \leq \epsilon_g, \quad \lambda_{\min}(\nabla^2 f(x)) \geq -\epsilon_H, \tag{3.34}$$

where ϵ_g and ϵ_H are two small constants.

Our algorithm takes steps of two types: a steepest-descent step, as in Section 3.2, or a step in a negative curvature direction for $\nabla^2 f$. Iteration k proceeds as follows:

(i) If $\|\nabla f(x^k)\| > \epsilon_g$, take the steepest descent step (3.3).

(ii) Otherwise, define λ_k to be the minimum eigenvalue of $\nabla^2 f(x^K)$, that is, $\lambda^k := \lambda_{\min}(\nabla^2 f(x^k))$. If $\lambda_k < -\epsilon_H$, choose p^k to be the eigenvector corresponding to the most negative eigenvalue of $\nabla^2 f(x^k)$. Choose the size and sign of p^k such that $\|p^k\| = 1$ and $(p^k)^T \nabla f(x^k) \leq 0$, and set

$$x^{k+1} = x^k + \alpha_k p^k, \quad \text{where } \alpha_k = \frac{2|\lambda_k|}{M}. \tag{3.35}$$

- If neither of these conditions hold, then x^k satisfies the necessary conditions (3.34), so is an approximate second-order-necessary point.

For the steepest-descent step (i), we have from (3.5) that

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \leq f(x^k) - \frac{\epsilon_g^2}{2L}. \tag{3.36}$$

For a step of type (ii), we have from (3.33) that

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \alpha_k \nabla f(x^k)^T p^k + \frac{1}{2}\alpha_k^2 (p^k)^T \nabla^2 f(x^k) p^k + \frac{1}{6} M \alpha_k^3 \|p^k\|^3 \\ &\leq f(x^k) - \frac{1}{2}\left(\frac{2|\lambda_k|}{M}\right)^2 |\lambda_k| + \frac{1}{6} M \left(\frac{2|\lambda_k|}{M}\right)^3 \\ &= f(x^k) - \frac{2}{3}\frac{|\lambda_k|^3}{M^2} \qquad (3.37) \\ &= f(x^k) - \frac{2}{3}\frac{\epsilon_H^3}{M^2}. \qquad (3.38) \end{aligned}$$

By aggregating (3.36) and (3.38), we have that at each x^k for which the condition (3.34) does *not* hold, we attain a decrease in the objective of at least

$$\min\left(\frac{\epsilon_g^2}{2L}, \frac{2}{3}\frac{\epsilon_H^3}{M^2}\right).$$

Using the lower bound $\bar{f}$ on the objective f, we see that the number of iterations K required to meet he condition (3.34) must satisfy the condition

$$K \min\left(\frac{\epsilon_g^2}{2L}, \frac{2}{3}\frac{\epsilon_H^3}{M^2}\right) \leq f(x^0) - \bar{f},$$

from which we conclude that

$$K \leq \max\left(2L\epsilon_g^{-2}, \frac{3}{2}M^2\epsilon_H^{-3}\right)\left(f(x^0) - \bar{f}\right).$$

Note that the maximum number of iterates required to identify a point for which just the approximate stationarity condition $\|\nabla f(x^k)\| \leq \epsilon_g$ holds is at most $2L\epsilon_g^{-2}(f(x^0) - \bar{f})$. (We can just omit the second-order part of the algorithm.) Note too that it is easy to devise *approximate* versions of this algorithm with similar complexity. For example, the negative curvature direction p^k in step (ii) above can be replaced by an approximation to the direction of most negative curvature, obtained by the Lanczos iteration with random initialization.

3.7 The KL and PL Properties

Some functions that are convex but not strongly convex have a property that allows convergence results to be proved with rates similar to those for strongly convex functions. The Polyak-Lojasiewicz (PL) condition [46, 25] holds when there exists $m > 0$ such that (3.10) holds, that is,

$$\|\nabla f(x)\|^2 \geq 2m[f(x) - f(x^*)], \tag{3.39}$$

where x^* is any minimizer of f. This condition can be combined with a bound of the form (3.19) on the per-iterate decrease, we obtain linear convergence rates. An example of a function satisfying PL but not strong convexity is the quadratic function $f(x) = \frac{1}{2}x^T Ax$, where $A \succeq 0$ but A is singular. Then $f^* = 0$ and the condition (3.39) holds where m is the smallest *nonzero* eigenvalue of A.

The PL condition is a special case of the Kurdyka-Lojasiewicz (KL) condition [35, 26], which again requires $\|\nabla f(x)\|$ to grow at a rate that depends on $f(x) - f(x^*)$ as x moves away from the solution set. The nature of this growth rate and of the algorithm for generating $\{x^k\}$ allows local convergence of $\{f(x^k)\}$ to $f(x^*)$ at various rates to be proved.

Notes and References

The proof for weakly convex is from Vandenberghe notes.

The proof of Theorem 3.6 is from [38, Theorem 2.1.14].

The weak Wolfe line search of Algorithm 3.1 is from [12].

Exercises

1. **Linear Rates.** Let $\{x_k\}$ be a sequence satisfying $x_{k+1} \leq (1-\beta)x_k$ for $0 < \beta < 1$, and $x_0 \leq C$. Prove that $x_k \leq \epsilon$ for all
$$k \geq \beta^{-1} \log\left(\tfrac{C}{\epsilon}\right) .$$

2. Verify that if f is twice continuously differentiable with the Hessian satisfying $mI \preceq \nabla^2 f(x)$ for all $x \in \mathrm{dom}\,(f)$, then the strong convexity condition (2.18) is satisfied.

3. Show, as a corollary of Theorem 3.5 that if the sequence $\{x^k\}$ described in this theorem is bounded and if f is strictly convex, we have $\lim_{k\to\infty} x^k = x^*$.

4. How much of the analysis of Sections 3.2, 3.4, 3.5, and 3.3 applies to smooth *nonconvex* functions? Specifically, state an analog of Theorem 3.5 that is true when the assumption of convexity of f is dropped.

5. How is the analysis of Section 3.2 affected if we take an even shorter constant steplength than $1/L$, that is, $\alpha \in (0, 1/L)$? Show that we can still attain a "$1/k$" sublinear convergence rate for $\{f(x^k)\}$, but that the rate involves a constant that depends on the choice of α.

6. Find positive values of $\bar{\epsilon}$, γ_1, and γ_2 such that the Gauss-Southwell choice $d^k = -[\nabla f(x^k)]_{i_k}$, where $i_k = \arg\min_{i=1,2,\ldots,n} |[\nabla f(x^k)]_i|$ satisfies conditions (3.23).

7. **Co-coercivity of the gradient map.** Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is an m-strongly convex function with L-Lipschitz gradients.

   (a) Show that $q(x) := f(x) - \frac{m}{2}\|x\|^2$ is convex with $L - m$ Lipschitz gradients.

   (b) Use part (a) to prove that

   $$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{mL}{m+L}\|x - y\|^2 + \frac{1}{m+L}\|\nabla f(x) - \nabla f(y)\|^2$$

   for all x and y.

   (c) Use part (b) and the fact that $\nabla f(x_\star) = 0$ to show that the kth iterate of the gradient method applied to f with stepsize $\frac{2}{m+L}$ satisfies

   $$\|x_k - x_\star\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x_0 - x_\star\|,$$

   where $\kappa = L/m$.

8. **Weakly convex optimization.** Let f be a convex function with L-Lipschitz gradients. Assume that we know the true optimal solution lies in a ball of radius R about zero. In this exercise, we will show that minimizing a nearby strongly convex function will quickly produce a solution that is an approximate minimizer of f. Consider running the gradient method on the function

   $$f_\epsilon(x) = f(x) + \frac{\epsilon}{2R^2}\|x\|^2$$

   initialized at some x_0 with $\|x_0\| \leq R$.

   (a) Let $x_\star^{(\epsilon)}$ denote an optimal solution of f_ϵ. Is $x_\star^{(\epsilon)}$ unique?

   (b) Prove that $f(z) - f(x_\star) \leq f_\epsilon(z) - f_\epsilon(x_\star^{(\epsilon)}) + \frac{\epsilon}{2}$.

   (c) Prove that for an appropriately chosen stepsize, the gradient method applied to f_ϵ will find a solution such that

   $$f_\epsilon(z) - f_\epsilon(x_\star^{(\epsilon)}) \leq \frac{\epsilon}{2}$$

   in at most

   $$\frac{R^2 L}{\epsilon} \log\left(\frac{8R^2}{\epsilon}\right)$$

   iterations. Find a constant stepsize that yields such a convergence rate.

9. **Regularized Least-Squares.** Let A be an $n \times d$ matrix with $n < d$ and $\operatorname{rank}(A) = n$. In this problem, we will study the least-squares optimization problem

   $$\text{minimize } \frac{1}{n}\|Ax - b\|^2. \tag{3.40}$$

   (a) Assume there exists a z such that $Az = b$. How many solutions of the equation $Ax = b$ are there?

(b) If you run the gradient method on (3.40) starting at $x_0 = 0$, how many iterations are required to find a solution with $\frac{1}{n}\|Ax - b\|^2 \leq \epsilon$?

(c) Consider the *regularized* problem

$$\text{minimize } \ell_\mu(x) := \frac{1}{n}\|Ax - b\|^2 + \mu\|x\|^2 \,. \tag{3.41}$$

where μ is some positive scalar. Let $x^{(\mu)}$ denote the minimizer of (3.41). Compute a closed form formula of $x^{(\mu)}$.

(d) If you run the gradient method on (3.41) starting at $x_0 = 0$, how many iterations are required to find a solution with $\ell_\mu(x) - \ell_\mu(x^{(\mu)}) \leq \epsilon$?

(e) Suppose $\hat{x}$ satisfies $\ell_\mu(\hat{x}) - \ell_\mu(x^{(\mu)}) \leq \epsilon$. Come up with as tight an upper bound as you can on the quantity $\frac{1}{n}\|A\hat{x} - b\|^2$.

10. Modify the Extrapolation-Bisection Line Search (Algorithm 3.1) so that it terminates at a point satisfying *strong* Wolfe conditions, which are

$$f(x^k + \alpha d^k) \leq f(x^k) + c_1\alpha\nabla f(x^k)^T d^k, \tag{3.42a}$$

$$|\nabla f(x^k + \alpha d^k)^T d^k| \leq c_2|\nabla f(x^k)^T d^k| \,, \tag{3.42b}$$

where c_1 and c_2 are constants that satisfy $0 < c_1 < c_2 < 1$. (The difference with the weak Wolfe conditions (3.42) is that the directional derivative $\nabla f(x^k+\alpha d^k)^T d^k$ is not only bounded below by $c_2|\nabla f(x^k)^T d^k|$ but also bounded *above* by this same quantity. That is, it cannot be too positive. (Hint: You should test separately for the two ways in which (3.42b) is violated, that is, $\nabla f(x^k+\alpha d^k)^T d^k < -c_2|\nabla f(x^k)^T d^k|$ and $\nabla f(x^k+\alpha d^k)^T d^k > c_2|\nabla f(x^k)^T d^k|$. Different adjustments of L, α, and U are required in these two cases.)

11. Prove the claim of Section 3.7, that the function $f(x) = \frac{1}{2}x^T Ax$, where $A \succeq 0$ but A is singular, satisfies the condition (3.39) with m being the smallest nonzero eigenvalue of A. (Hint: Use the eigenvalue decomposition $A = \sum_{i=1}^r \lambda_i u_i u_i^T$, where $r < n$ is the rank of A, λ_i are the eigenvalues, and $\{u_1, u_2, \ldots, u_r\}$ is the orthonormal set of eigenvectors.)

12. Consider the function $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \frac{1}{4}\sum_{k=1}^{d-1} \cos(x_k - x_{k+1}) + \sum_{k=1}^{d} kx^2 \,.$$

(a) Compute a constant stepsize with which the gradient method converges.

(b) Compute the set of points where $\nabla f(x) = 0$. For each point, determine if it is a local minima, local maxima, or a global minimum.

(c) Consider the gradient method with the constant stepsize you computed in part (a) and the initial point $x_0 = [1, 1, 1, \ldots, 1]^T$. Determine to which point in $\{x \ : \ \nabla f(x) = 0\}$ the algorithm converges. Explain your reasoning.