# Chapter 3

# 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 fundamental methods of this type and analyze their convergence and complexity properties. This chapter can be read as an introduction both to elementary 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 of this chapter are suited to the case in which $f$ and its gradient $\nabla f$ 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 $\nabla f$ cannot be evaluated exactly without a complete sweep through the data set, but unbiased estimates of $\nabla f$ can be obtained easily.
- Situations in which it is much less expensive to evaluate an individual component or a subvector of $\nabla f$ than the full gradient vector.
- Smooth but nonconvex $f$.

Extensions to the fundamental methods of his chapter to these more general situations 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 $\nabla f$, 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 $d$ with unit norm, the one that minimizes $d^T \nabla f(x)$ is $d = -\nabla f(x)/\|\nabla f(x)\|$. For this reason, we refer to $-\nabla f(x)$ as the *steepest descent* direction. Perhaps the simplest method for optimization of a smooth function makes use of this direction, defining its iterates by

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

for some steplength $\alpha_k > 0$. At each iteration, we are guaranteed that there is either some nonnegative step $\alpha$ that 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. (If $f$ is also convex, this point will be a global minimizer of $f$.) The algorithm defined by (3.2) is called the *gradient method* or the *steepest descent method*. In the next section, we will discuss the choice of steplengths $\alpha_k$, and analyze how many iterations are required to find points where the gradient nearly vanishes.

## 3.2 Steepest Descent

We focus first on the question of choosing the stepsize $\alpha_k$ for the steepest descent method (3.2). If $\alpha_k$ is too large, we risk taking a step that increases the function value. On the other hand, if $\alpha_k$ 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, which can be implemented when $f$ is $L$-smooth (see (2.7)) with a known value of the parameter $L$. By setting $\alpha_k$ to be a constant value $\alpha$, the formula (3.2) becomes

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

To estimate the amount of decrease in $f$ obtained at each iterate of this method, we use Lemma 2.2, which is a consequence of Taylor's theorem (Theorem 2.1). We obtain

$$f(x+\alpha d) \leq f(x) + \alpha \nabla f(x)^T d + \alpha^2 \frac{L}{2} \|d\|^2. \tag{3.4}$$

For $d = -\nabla f(x)$, the value of $\alpha$ that minimizes the expression on the right-hand side is $\alpha = 1/L$. By substituting this value into (3.4), and setting $x = x^k$, 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 $f$ has a global lower bound. That is, we assume that there is a value $\bar{f}$ that satisfies

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

(In the case that $f$ has a global minimizer $x^*$, $\bar{f}$ could be any value such that $\bar{f} \leq f(x^*)$.) By summing the inequalities (3.5) over $k = 0, 1, \ldots, T-1$, and canceling terms, 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

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

which implies that $\lim_{T\to\infty}\|\nabla f(x^T)\| = 0$. Moreover, we have

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

By taking square roots of both sides of this expression, we have

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

$$\min_{0\leq k\leq T-1}\|\nabla f(x^k)\| \leq \sqrt{\frac{2L[f(x^0) - \bar{f}]}{T}}\,. \tag{3.7}$$

Note that this convergence rate is *slow*, and tells us only that we will find a point $x^k$ that is nearly stationary. We need to assume stronger properties of $f$ 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 $x^*$. Define $f^* := f(x^*)$. 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

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

as required. □

### 3.2.3 Strongly Convex Case

Recall from (2.19) that the smooth function $f : \mathbb{R}^n \to \mathbb{R}$ is *strongly convex with modulus $m$* 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 positive 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 $\mu$ large enough. (Exercise: Prove this!)

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

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

Thus, if $\|\nabla f(x)\| < \delta$, we have

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

Thus, when the gradient is small, we are close to having found a point with minimal function value.

We can derive an estimate of the distance of $x$ to the optimal point $x^*$ in terms of the gradient by using (3.9) and the Cauchy-Schwarz inequality. We have

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

By rearranging terms, we have

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

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

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

where $f^* := f(x^*)$ as before. Subtracting $f^*$ from both sides of this inequality gives the recursion

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

Thus the sequence of function values converges *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 general 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 initial distance to optimality $\|x^0 - x^*\|$ rather than 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 $\epsilon$, whereas the sublinear rates depend on $1/\epsilon$ or $1/\epsilon^2$. When $\epsilon$ 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 $m/L$ is of the same order as $\epsilon$. 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 do not know $L$? Even when we do know it, is the steplength $\alpha_k \equiv 1/L$ good in practice? 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 Descent Methods: Convergence

In the previous section we considered the short-step gradient method that stepped along the negative gradient with a stepsize $1/L$ determined by the global curvature of the gradient. 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 $\alpha_k$ 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 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 $\alpha_k$ 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 in the previous section 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 a scheme that guarantees (3.19)? The following elementary theorem shows one basic property.

**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

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

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 $\nabla f$, 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 $\infty$. (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 $(1 - 2mC)$.

The argument of Section 3.2.2 concerning rate of convergence for the (non-strongly) convex case cannot be generalized to the setting of (3.19), though similar results can be obtained by another technique under an additional assumption, as we show next.

**Theorem 3.6.** *Suppose that $f$ is convex and smooth, where $\nabla f$ 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*

$$f(x^T) - f^* \leq \frac{R_0^2}{CT} \quad T = 1, 2, \ldots . \tag{3.20}$$

*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 $\Delta_k$ becomes

$$\Delta_{k+1} \leq \Delta_k - \frac{C}{R_0^2}\Delta_k^2 = \Delta_k \left(1 - \frac{C}{R_0^2}\Delta_k\right). \tag{3.21}$$

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 (3.21) 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 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 $\bar{\epsilon}$, $\gamma_1$, $\gamma_2$:

$$0 < \bar{\epsilon} \leq \frac{-(d^k)^T \nabla f(x^k)}{\|\nabla f(x^k)\| \|d^k\|}, \tag{3.22a}$$

$$0 < \gamma_1 \leq \frac{\|d^k\|}{\|\nabla f(x^k)\|} \leq \gamma_2. \tag{3.22b}$$

Condition (3.22a) says that the angle between $-\nabla f(x^k)$ and $d^k$ is acute, and bounded away from $\pi/2$ for all $k$; while condition (3.22b) 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 negative gradient search direction $d^k = -\nabla f(x^k)$, the conditions (3.22) 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 $d = d^k$ in (3.4), we obtain

$$\begin{aligned} f(x^{k+1}) &= f(x^k + \alpha d^k) \\ &\leq f(x^k) + \alpha \nabla f(x^k)^T d^k + \alpha^2 \frac{L}{2} \|d^k\|^2 \\ &\leq f(x^k) - \alpha \bar{\epsilon} \|\nabla f(x^k)\| \|d^k\| + \alpha^2 \frac{L}{2} \|d^k\|^2 \\ &\leq f(x^k) - \alpha \left( \bar{\epsilon} - \alpha \frac{L}{2} \gamma_2 \right) \|\nabla f(x^k)\| \|d^k\|, \end{aligned} \tag{3.23}$$

where we used (3.22) for the last two inequalities. It is clear from this expression that for all values of $\alpha$ 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.

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 $\gamma_1$ and $\gamma_2$ are positive quantities as in (3.22). The condition (3.22b) holds, by definition of $S^k$, and condition (3.22a) 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 that the Hessian $\nabla^2 f(x)$ has eigenvalues uniformly bounded in the range $[1/\gamma_2, 1/\gamma_1]$ for all $x$.

- The Gauss-Southwell variant of coordinate descent chooses $d^k = -[\nabla f(x^k)]_{i_k} e_{i_k}$, where $i_k = \arg\max_{i=1,2,\ldots,n} |[\nabla f(x^k)]_i|$ and $e_{i_k}$ is the vector containing all zeros except for a 1 in position $i_k$. (We leave it as an exercise to show that the conditions (3.22) 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 (3.22) 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 condition (3.22a) holds in an expected sense. Since $E(\|d^k\|^2) = \frac{1}{n}\|\nabla f(x^k)\|_2^2$, the norms of $\|d^k\|$ and $\|\nabla f(x^k)\|$ are also similar to within a scale factor, so (3.22b) 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 $\xi_k$ 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 (3.22) 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 $\xi_k$.

## 3.5 Line-Search Methods: Choosing the Steplength

Assuming now that the search direction $d^k$ in (3.18) satisfies the properties (3.22), we turn to the choice of steplength $\alpha_k$, for which a well designed procedure is often used. We describe some methods that make use of the Lipschitz constant $L$ from (2.7), and other methods that do not assume knowledge of $L$, but still satisfy a sufficient decrease like (3.19).

**Constant Stepsize.** As we have seen in Section 3.2, constant stepsizes can yield useful convergence results. One drawback of the constant stepsize method is that 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 $\alpha$ as large as possible such that the algorithm does not 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, $\alpha_k$ 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, a special case of which was investigated already in Section 3.2, is to base the choice of $\alpha_k$ on knowledge of the global properties of the function $f$, particularly on the Lipschitz constant $L$ for the gradient (see (2.7)) or the modulus of convexity $m$ (see (2.18)). Given the expression (3.23) above, for example, and supposing we have estimates of all the quantities $\bar{\epsilon}$, $\gamma_2$, and $L$ that appear therein, we could choose $\alpha$ to maximize the coefficient of the last term. Setting $\alpha = \bar{\epsilon}/(L\gamma_2)$, we obtain from (3.23) and (3.22) that

$$f(x^{k+1}) \leq f(x^k) - \frac{\bar{\epsilon}^2}{2L\gamma_2}\|\nabla f(x^k)\|\|d^k\| \geq f(x^k) - \frac{\bar{\epsilon}^2\gamma_1}{2L\gamma_2}\|\nabla f(x^k)\|^2. \tag{3.24}$$

**Exact Line Search.** A second option is to perform a one-dimensional line search along direction $d^k$ to find the minimizing value of $\alpha$, that is,

$$\min_{\alpha>0} f(x^k + \alpha d^k). \tag{3.25}$$

This technique requires evaluation of $f(x^k + \alpha d^k)$ (and possibly also its derivative with respect to $\alpha$, namely $(d^k)^T\nabla f(x^k + \alpha d^k)$) economically, for arbitrary positive values of $\alpha$. 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 gradient along the search direction, 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 functions $f$, 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 (3.24), this bound 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. Many line-search methods were proposed 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 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 defined as follows:

$$f(x^k + \alpha d^k) \leq f(x^k) + c_1\alpha\nabla f(x^k)^T d^k, \tag{3.26a}$$

$$\nabla f(x^k + \alpha d^k)^T d^k \geq c_2\nabla f(x^k)^T d^k\,. \tag{3.26b}$$

Here, $c_1$ and $c_2$ are constants that satisfy $0 < c_1 < c_2 < 1$. The condition (3.26a) 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 (3.26b), which we call the "gradient condition," ensures that $\alpha_k$ 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 $\alpha_k$ 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

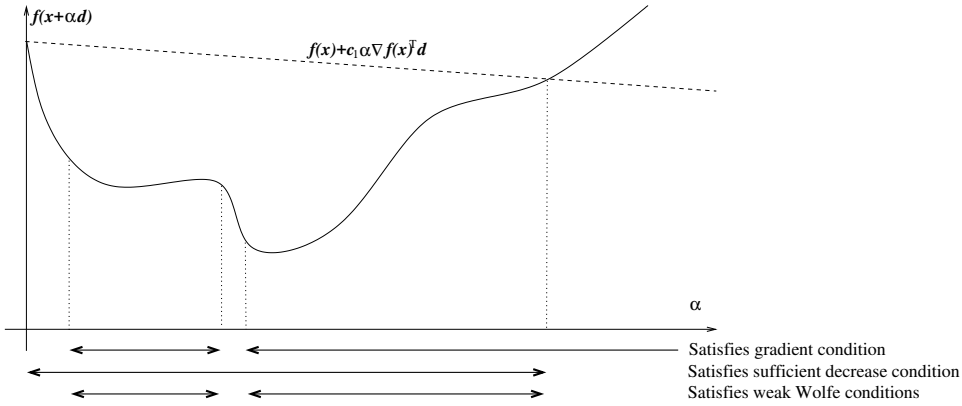


Figure 3.1: Weak Wolfe conditions are satisfied when both the gradient condition (3.26b) and the sufficient decrease condition (3.26a) hold.

(3.24)), we argue as follows. First, from condition (3.26b) and the Lipschitz property for $\nabla f$, 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

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

By substituting into (3.26a), and using the (3.22a), we obtain

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

Algorithm 3.1 (from [21]) describes an approach that combines extrapolation with bisection to find a steplength $\alpha$ satisfying the conditions (3.26). This method maintains a subinterval $[L, U]$ of the positive real line (initially $L = 0$ and $U = \infty$) that contains a point satisfying (3.26), along with a current guess $\alpha \in (L, U)$ of this point. If the sufficient decrease condition (3.26a) is violated by $\alpha$, then the current guess is too long, so the upper bound $U$ is assigned the value $\alpha$, and the new guess is taken to be the midpoint of the new interval $[L, U]$. If the sufficient decrease condition holds but the condition (3.26b) is violated, the current guess of $\alpha$ is too short. In this case, we move the lower bound up to $\alpha$, and take the next guess of $\alpha$ 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 $\alpha$ satisfying (3.26) can be found in Section A.3 in the Appendix.

**Backtracking Line Search.** Another popular approach to determining an appropriate value for $\alpha_k$ is known as "backtracking." It is widely used in situations where evaluation of $f$ is economical and practical, while evaluation of the gradient $\nabla f$ is more difficult. It is easy to implement (no

**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

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 $\alpha_k$ 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 (3.26a) is satisfied. Note that backtracking does not require a condition like (3.26b) to be checked. The purpose of such a condition is to ensure that $\alpha_k$ is not too short, but this is not a concern in backtracking, because we know that $\alpha_k$ is either the fixed value $\bar{\alpha}$, or is within a factor $\beta$ 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 (3.22) that

$$f(x^{k+1}) \leq f(x^k) + c_1 \bar{\alpha} \nabla f(x^k)^T d^k \leq f(x^k) - c_1 \bar{\alpha} \bar{\epsilon} \gamma_1 \|\nabla f(x^k)\|^2. \tag{3.27}$$

When backtracking is needed, we have from the fact that the test (3.26a) 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 (3.23), 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 (3.26a) with $\alpha = \alpha_k$ (note that this condition is satisfied for this value of $\alpha$)

and then using (3.22), we obtain

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + c_1\alpha_k \nabla f(x^k)^T d^k \\ &\leq f(x^k) - \frac{2}{L}\beta(1-c_1)c_1 \frac{(\nabla f(x^k)^T d^k)^2}{\|d^k\|^2} \\ &\leq f(x^k) - \frac{2}{L}\beta c_1(1-c_1)\bar{\epsilon}^2 \|\nabla f(x^k)\|^2. \end{aligned} \tag{3.28}$$

## 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 for a smooth, possibly nonconvex function $f$, which are

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \text{ positive semidefinite} \tag{3.29}$$

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

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq M\|x-y\|, \quad \text{for all } x, y \in \text{dom}\,(f). \tag{3.30}$$

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 $f$:

$$f(x+p) \leq f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x) p + \frac{1}{6} M\|p\|^3. \tag{3.31}$$

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 (3.31) 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 (3.29), that is,

$$\|\nabla f(x)\| \leq \epsilon_g, \quad \lambda_{\min}(\nabla^2 f(x)) \geq -\epsilon_H, \tag{3.32}$$

where $\epsilon_g$ and $\epsilon_H$ 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 (3.2) with $\alpha_k = 1/L$.

(ii) Otherwise, define $\lambda_k$ 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

$$x^{k+1} = x^k + \alpha_k p^k, \quad \text{where } \alpha_k = \frac{2|\lambda_k|}{M}. \tag{3.33}$$

If neither of these conditions hold, then $x^k$ satisfies the necessary conditions (3.32), so is an approximate second-order-necessary point.

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

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \leq f(x^k) - \frac{\epsilon_g^2}{2L}. \tag{3.34}$$

For a step of type (ii), we have from (3.31) that

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \alpha_k \nabla f(x^k)^T p^k + \frac{1}{2}\alpha_k^2 (p^k)^T \nabla^2 f(x^k) p^k + \frac{1}{6} M \alpha_k^3 \|p^k\|^3 \\ &\leq f(x^k) - \frac{1}{2}\left(\frac{2|\lambda_k|}{M}\right)^2 |\lambda_k| + \frac{1}{6} M \left(\frac{2|\lambda_k|}{M}\right)^3 \\ &= f(x^k) - \frac{2}{3}\frac{|\lambda_k|^3}{M^2} \\ &\leq f(x^k) - \frac{2}{3}\frac{\epsilon_H^3}{M^2}. \end{aligned} \tag{3.35}$$

By aggregating (3.34) and (3.35), we have that at each $x^k$ for which the condition (3.32) 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 (3.32) 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 to obtain this result.) 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 Mirror Descent

The steps of the steepest descent method (3.2) can also be obtained from the solution of simple quadratic problems:

$$x^{k+1} = \arg\min\ f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2\alpha_k}\|x - x^k\|^2. \tag{3.36}$$

Thus, we can think of the new iterate being obtained from a first-order Taylor-series model, with a quadratic penalty term, based on the Euclidean norm, that penalizes our move away from the current iterate. Moreover, as $\alpha_k$ decreases, the penalty becomes more severe, so the step is shorter. (This viewpoint is useful in later chapters, where we consider constrained and regularized problems.)

In this section, we consider a framework like (3.36) but with the final term replaced by a general class of distance measures called *Bregman divergences*, and denoted by $D_h(\cdot,\cdot)$. The steps have the form

$$x^{k+1} = \arg\min\ f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{\alpha_k} D_h(x, x^k). \tag{3.37}$$

The subscript "$h$" refers to a function that is smooth and strongly convex *in some norm*. That is, it satisfies (2.19) for some $m > 0$, but the norm in the final term $(m/2)\|y - x\|^2$ of this definition can be any norm, not necessarily the Euclidean norm that we use elsewhere in this book. This function $h$ is said to *generate* the Bregman divergence $D_h(\cdot,\cdot)$ by means of the following formula:

$$D_h(x, z) := h(x) - h(z) - \nabla h(z)^T(x - z), \tag{3.38}$$

which is the difference between $h(x)$ and the first-order Taylor-series approximation of $h$ at $z$, evaluated at $x$. See the illustration in Figure 3.2.

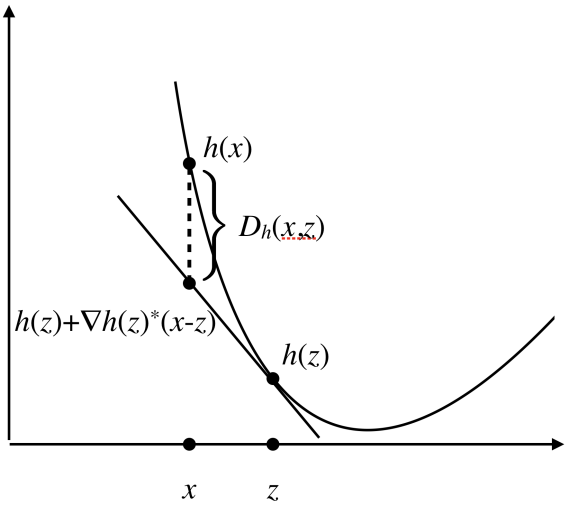


Figure 3.2: Illustration of how to compute a Bregman divergence $D_h(x, z)$.

Since $h$ is convex, $D_h(x, z)$ is nonnegative and strongly convex in the first argument. It may not satisfy other familiar properties of squared norms, but it does satisfy a "three point property." This property holds for the squared Euclidean norm, where it is known as the "law of cosines:" For any three points $x$, $y$, $z$ in $\mathbb{R}^n$, we have

$$\begin{aligned}\|x - y\|^2 &= \|x - z\|^2 + \|z - y\|^2 - 2(x - z)^T(y - z)\\ &= \|x - z\|^2 + \|z - y\|^2 - 2\|x - z\|\|y - z\|\cos\gamma,\end{aligned}$$

where $\gamma$ is the angle made at $z$ by the vectors $(x - z)$ and $(y - z)$. When $\gamma$ is $\pi/2$, then $x$, $y$, and $z$ form a right-angled triangle, and this law reduces to the Pythagorean theorem.

Bregman divergences share the "three-point property:" We can show that

$$D_h(x,y) = D_h(x,z) - (x-z)^T(\nabla h(y) - \nabla h(z)) + D_h(z,y). \tag{3.39}$$

The proof is just algebra (see the Exercises). Remarkably, this property is all we need to "mirror" the analysis of our standard convergence proofs for steepest descent.

**Example 3.1** (Squared Euclidean Norm)**.** *For $h(x) = \frac{1}{2}\|x\|^2$, we have*

$$D_h(x,z) = \frac{1}{2}\|x\|^2 - \frac{1}{2}\|z\|^2 - z^T(x-z) = \frac{1}{2}\|x-z\|^2,$$

*so that* (3.36) *is a special case of* (3.37) *when the generating function is the squared Euclidean norm.*

**Example 3.2** (Negative Entropy)**.** *Consider the $n$-simplex of probability distributions, defined by $\Delta_n := \{p \in \mathbb{R}^n \,|\, p \geq 0,\ \sum_{i=1}^n p_i = 0\}$. Take $h(p) = \sum_{i=1}^n p_i \log p_i$ to be the* negative entropy *of the distribution $p$. This function is convex, and for any $p, q \in \Delta_n$, we have*

$$\begin{aligned} D_h(p,q) &= \sum_{i=1}^n p_i \log p_i - \sum_{i=1}^n q_i \log q_i - \sum_{i=1}^n (\log q_i - 1)(p_i - q_i) \\ &= \sum_{i=1}^n p_i \log p_i - \sum_{i=1}^n p_i \log q_i - \sum_{i=1}^n (p_i - q_i) \\ &= \sum_{i=1}^n p_i \log\left(\frac{p_i}{q_i}\right). \end{aligned}$$

*This measure is the* The Kullback-Liebler Divergence *or* KL Divergence *between $p$ and $q$. The function $h$ of this example is strongly convex with respect to the norm $\|\cdot\|_1$ on the interior of $\Delta_n$, with modulus* 1*. That is, we have*

$$h(p) \geq h(q) + \nabla h(q)^T(p-q) + \frac{1}{2}\|p-q\|_1^2, \quad \textit{for all } p, q \in \operatorname{int} \Delta_n.$$

*This bound is known as* Pinsker's Inequality*.*

We now consider the mirror descent algorithm, which defines its iterates by (3.37). Because from (3.38) we have that

$$\nabla_x D_h(x,z) = \nabla h(x) - \nabla h(z),$$

the optimality conditions for (3.37) are

$$\nabla f(x^k) + \frac{1}{\alpha}\nabla h(x^{k+1}) - \frac{1}{\alpha}\nabla h(x^k) = 0$$

We can thus write the next iterate $x^{k+1}$ explicitly as

$$x^{k+1} = (\nabla h)^{-1}\left\{\nabla h(x^k) - \alpha \nabla f(x^k)\right\},$$

where $(\nabla h)^{-1}$ is the inverse function of $h$. In fact, this inverse function is rarely computable, but for our special cases of Examples 3.1 and 3.2, it *can* be computed explicitly. For $h(x) = \frac{1}{2}\|x\|^2$, we have $(\nabla h)^{-1}(v) = v$. For $h(p) = \sum_{i=1}^n p_i \log p_i$, we can show that

$$(\nabla h)^{-1}(v)_i = \frac{e^{v_i}}{\sum_{j=1}^n e^{v_j}}, \quad i = 1, 2, \ldots, n.$$

Examples 3.1 and 3.2 cover almost the whole range of applications of mirror descent. There are not many other strongly convex functions out there whose gradient maps have simple inverses. But in principle, any such function $h$ would define its own Bregman divergence and hence its own mirror descent algorithm.

**Mirror Descent Analysis.** Because one of the key applications of mirror descent (Example 3.2) restricts the iterates to a subset of $\mathbb{R}^n$, we are more careful than usual here in setting up and analyzing the method over something less than the whole space $\mathbb{R}^n$.

Let $\mathcal{X} \subseteq \mathcal{D} \subseteq \mathbb{R}^n$ be convex sets, and suppose that $h : \mathcal{X} \to \mathbb{R}$ is continuously differentiable. Let $\|\cdot\|$ be some arbitrary norm (not necessarily Euclidean) and assume that $h$ is strongly convex with modulus $m$ with respect to this norm, that is,

$$h(x) \geq h(z) + \langle \nabla h(z), x - z\rangle + \frac{m}{2}\|x - z\|^2, \quad \text{for all } x, z \in \mathcal{X}.$$

Also recall that a function $f$ is $L$-Lipschitz with respect to $\|\cdot\|$ if and only if $\|g\|_* \leq L$ for all $g \in \partial f(x)$, where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$.

Consider the mirror descent algorithm (3.37), modified slightly to confine its iterates to the set $\mathcal{X}$:

$$x^{k+1} = \arg\min_{x \in \mathcal{X}} f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{\alpha_k} D_h(x, x^k), \quad k = 0, 1, 2, \ldots .$$

The optimality conditions for this subproblem are

$$\left[\nabla f(x^k) + \frac{1}{\alpha_k}\nabla h(x^{k+1}) - \frac{1}{\alpha_k}\nabla h(x^k)\right]^T (x - x^{k+1}) \geq 0, \quad \text{for all } x \in \mathcal{X}. \tag{3.40}$$

Here (and also for later algorithms), we analyze the behavior of a *weighted average* of the iterates, rather than the iterates $x^k$ themselves. We define

$$\lambda_k = \sum_{j=0}^k \alpha_j, \quad \bar{x}^k = \lambda_k^{-1} \sum_{j=0}^k \alpha_j x^j . \tag{3.41}$$

We have the following result, whose proof is from Beck and Teboulle [5].

**Theorem 3.7.** *Let $\|\cdot\|$ be an arbitrary norm on $\mathcal{X}$, and suppose that $h$ is a $m$-strongly-convex function with respect to $\|\cdot\|$ on $\mathcal{X}$. Suppose that $f$ is convex and $L$-Lipschitz with respect to $\|\cdot\|$, and that a solution $x^*$ to the problem $\min_{x\in\mathcal{X}} f(x)$ exists, with objective $f^* = f(x^*)$. Then for any integer $T \geq 1$, we have*

$$f(\bar{x}^T) - f^* \leq \frac{D_h(x^*, x^0) + \frac{L^2}{2m}\sum_{t=0}^T \alpha_t^2}{\sum_{t=0}^T \alpha_t},$$

*where $\bar{x}^T$ is defined by* (3.41).

*Proof.* By adding and subtracting terms, we have

$$\begin{aligned}\alpha_k \nabla f(x^k)^T(x^k - x^*) = (&-\alpha_k \nabla f(x^k) - \nabla h(x^{k+1}) + \nabla h(x^k))^T(x^* - x^{k+1}) \\ &+ (\nabla h(x^{k+1}) - \nabla h(x^k))^T(x^* - x^{k+1}) + (\alpha_k \nabla f(x^k))^T(x^k - x^{k+1}).\end{aligned}$$

The first term on the right-hand side is nonpositive, because of the optimality conditions (3.40). The second term can be rewritten using the three-point property (3.39) as follows:

$$(\nabla h(x^{k+1}) - \nabla h(x^k))^T(x^* - x^{k+1}) = -D_h(x^*, x^{k+1}) - D_h(x^{k+1}, x^k) + D_h(x^*, x^k).$$

The final term can be bounded as

$$\alpha_k \nabla f(x^k)^T(x^k - x^{k+1}) \leq \alpha_k \|\nabla f(x^k)\|_\star \|x^k - x^{k+1}\| \leq \frac{\alpha_k^2}{2m}\|\nabla f(x^k)\|_\star^2 + \frac{m}{2}\|x^k - x^{k+1}\|^2,$$

where we used the bound $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ for any scalars $a$ and $b$. Finally, note that since $h$ is strongly convex with parameter $m$, we have

$$-D_h(x^{k+1}, x^k) + \frac{m}{2}\|x^k - x^{k+1}\|^2 = -h(x^{k+1}) + h(x^k) + \nabla h(x^k)^T(x^{k+1} - x^k) + \frac{m}{2}\|x^k - x^{k+1}\|^2 \leq 0\,.$$

By assembling all these inequalities and substituting into the original expression, we obtain

$$\alpha_k \nabla f(x^k)^T(x^k - x^*) \leq -D_h(x^*, x^{k+1}) + D_h(x^*, x^k) + \frac{\alpha_k^2}{2m}\|\nabla f(x^k)\|_\star^2\,. \tag{3.42}$$

We now proceed with a telescoping sum argument. We first use convexity of $f$ then (3.42) to obtain

$$\begin{aligned}f(\bar{x}^T) - f^* &\leq \lambda_T^{-1} \sum_{k=0}^{T} \alpha_k (f(x^k) - f(x^*)) \\ &\leq \lambda_T^{-1} \sum_{k=0}^{T} \alpha_k \nabla f(x^k)^T(x^k - x^*) \\ &\leq \lambda_T^{-1} \sum_{k=0}^{T} \left\{ D_h(x^*, x^k) - D_h(x^*, x^{k+1}) + \frac{\alpha_k^2}{2m}\|\nabla f(x^k)\|_\star^2 \right\} \\ &\leq \frac{D_h(x^*, x^0) + \frac{1}{2m}\sum_{k=0}^{T} \alpha_k^2 \|\nabla f(x^k)\|_\star^2}{\lambda_T},\end{aligned}$$

where we used $D_h(x^*, x^{T+1}) \geq 0$ in the final inequality. Since $\|\nabla f(x^k)\|_\star \leq L$ by assumption, the proof is complete. □

We can use this result to make various choices of steplengths $\alpha_k$. Suppose that we have a bound $R$ on $D_h(x^*, x^0)$ (this may be easy to obtain if the set $\mathcal{X}$ is compact, for example), and knowledge of the constants $L$ associated with $f$ and $m$ associated with $h$. Then choosing the number of iterations $T$ in advance, the "optimal" choice of constant stepsize will be the value $\alpha$ that minimizes

$$\frac{R + \frac{L^2}{2m}\sum_{k=0}^{T} \alpha^2}{\sum_{k=0}^{T} \alpha} = \frac{R + \frac{L^2(T+1)}{2m}\alpha^2}{(T+1)\alpha}.$$

A short calculation shows that the minimizing value is

$$\alpha = \frac{\sqrt{2mR}}{L}\frac{1}{\sqrt{T+1}}, \tag{3.43}$$

which yields the following estimate:

$$f(\bar{x}^T) - f^* \leq \frac{L\sqrt{2R}}{\sqrt{m}}\frac{1}{\sqrt{T+1}}. \tag{3.44}$$

Note that this rate of $1/\sqrt{T}$ is asymptotically slower than the $1/T$ rate achieved for convex functions in Section 3.2.2. However, we note two points. First, mirror descent is not particularly sensitive to variations in the stepsize. For example, if the choice of fixed steplength in (3.43) is scaled by a constant $\theta > 0$ (because of mis-estimation of the constants $L$ and $R$, for example), the effect on the convergence expression (3.44) is modest; the right-hand side increases, but only by a factor related to $\theta$ and $\theta^{-1}$. The use of averaging results in slower convergence but greater robustness to choice of steplength. (If the steplength in the regular steepest descent method of Section 3.2.2 is chosen to be too long, the method may not converge at all.)

The second point is that the constants $L$, $R$, and $m$ may be smaller for a certain choice of Bregman divergence and norm than for the usual Euclidean norm. Returning to Example 3.2, where $\mathcal{X}$ is the unit simplex $\Delta_n$, we have by choosing $x^0$ to be the midpoint of the simplex $(1/n)\mathbf{1}$ that

$$R = \sup_{p\in\Delta_n} D_h(p, \tfrac{1}{n}\mathbf{1}) \leq \sup_{p\in\Delta_n} \sum_{i=1}^{n} (p_i \log p_i - p_i \log 1/n) \leq \log n.$$

We noted already that in Example 3.2, the function $h$ is strongly convex with respect to norm $\|\cdot\|_1$ with modulus $m = 1$. Moreover, using the dual norm $\|\cdot\|_\infty$, the constant $L$ bounds the supremum of $\|\nabla f(x)\|_\infty$ over $\mathcal{X}$, rather than $\|\nabla f(x)\|_2$, which may be larger by a factor of $n$. The advantage of this setup can be observed in practice. Mirror descent with the KL divergence is often considerably faster for optimizing a function over the simplex than the mirror descent variant based on the Euclidean norm, particularly when the gradients $\nabla f(x)$ are dense vectors.

## 3.8 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 [77, 51] holds when there exists $m > 0$ such that (3.10) holds, that is,

$$\|\nabla f(x)\|^2 \geq 2m[f(x) - f(x^*)], \tag{3.45}$$

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 to obtain linear convergence rates of the form (3.15). 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 (3.45) holds where $m$ is the smallest *nonzero* eigenvalue of $A$. (See Section A.7 in the Appendix for a proof of this claim.)

The PL condition is a special case of the Kurdyka-Lojasiewicz (KL) condition [64, 53], 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 of Theorem 3.3 is from the notes of L. Vandenberghe, while Theorem 3.6 is from [70, Theorem 2.1.14].

Additional information about line-search algorithms can be found in [75, Chapter 3].

## Exercises

1. Verify that if $f$ is twice continuously differentiable with the Hessian satisfying $mI \preceq \nabla^2 f(x)$ for all $x \in \mathrm{dom}\,(f)$, for some $m > 0$, then the strong convexity condition (2.18) is satisfied.

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

3. 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 $\alpha$.

4. Find positive values of $\bar{\epsilon}$, $\gamma_1$, and $\gamma_2$ such that the Gauss-Southwell choice $d^k = -[\nabla f(x^k)]_{i_k} e_{i_k}$, where $i_k = \arg\min_{i=1,2,\ldots,n} |[\nabla f(x^k)]_i|$ and $e_{i_k}$ is the vector containing all zeros except for a 1 in position $i_k$, satisfies conditions (3.22).

5. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a strongly convex function with modulus $m$, an $L$-Lipschitz gradient, and (unique) minimizer $x^*$ with function value $f^* = f(x^*)$. Use the co-coercivity property (2.21) and the fact that $\nabla f(x^*) = 0$ to prove that the $k$th iterate of the gradient method applied to $f$ with stepsize $\frac{2}{m+L}$ satisfies

$$\|x^k - x^*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x^0 - x^*\|,$$

   where $\kappa = L/m$.

6. Let $f$ be a convex function with $L$-Lipschitz gradients. Assume that we know that the minimizer lies in a ball of radius $R$ about zero. In this exercise, we show that minimizing a nearby strongly convex function will yield an approximate minimizer of $f$ with good complexity. Consider running the gradient method on the strongly convex function

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

   where $0 < \epsilon \ll L$, initialized at some $x^0$ with $\|x^0\| \leq R$. Let $x_\epsilon^*$ denote the (unique) minimizer of $f_\epsilon$.

   (a) Prove that $f(z) - f(x^*) \leq f_\epsilon(z) - f_\epsilon(x_\epsilon^*) + \frac{\epsilon}{2}$, for any $z$ with $\|z\| \leq R$.

   (b) Prove that for an appropriately chosen stepsize, the gradient method applied to $f_\epsilon$ will find a solution such that

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

in at most approximately

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

Find a precise estimate of this rate, and write the constant stepsize that yields this convergence rate.

7. Let $A$ be an $N \times d$ matrix with $N < d$ and $\text{rank}(A) = N$, and consider the least-squares optimization problem

$$\min_x \; f(x) := \frac{1}{N}\|Ax - b\|^2 . \tag{3.46}$$

   (a) Assume there exists a $z$ such that $Az = b$. Characterize the solution space of the system $Ax = b$.

   (b) Write down the Lipschitz constant for the gradient of the function (3.46) in terms of $A$.

   (c) If you run the steepest descent method on (3.46) starting at $x^0 = 0$, with appropriate choice of steplength, how many iterations are required to find a solution with $\frac{1}{n}\|Ax - b\|^2 \leq \epsilon$?

   (d) Consider the *regularized* problem

$$\min \; f_\mu(x) := \frac{1}{n}\|Ax - b\|^2 + \mu\|x\|^2 . \tag{3.47}$$

   for some $\mu > 0$. Express the minimizer $x_\mu$ of of (3.47) in closed form.

   (e) If you run the gradient method on (3.47) starting at $x^0 = 0$, how many iterations are required to find a solution with $f_\mu(x) - f_\mu(x_\mu) \leq \epsilon$?

   (f) Suppose $\hat{x}$ satisfies $f_\mu(\hat{x}) - f_\mu(x_\mu) \leq \epsilon$. Find a tight upper bound on $f(\hat{x})$.

   (g) From Section 3.8, for $f$ defined in (3.46), find the value of $m$ that satisfies (3.45), in terms of the minimum eigenvalue of $A^T A$ (and possibly other quantities).

   (h) Referring to Section 3.2.3, define an appropriate choice of steplength for the steepest descent method applied to (3.46), and write down the linear convergence expression for the resulting method.

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

$$f(x^k + \alpha d^k) \leq f(x^k) + c_1 \alpha \nabla f(x^k)^T d^k, \tag{3.48a}$$

$$|\nabla f(x^k + \alpha d^k)^T d^k| \leq c_2 |\nabla f(x^k)^T d^k| , \tag{3.48b}$$

   where $c_1$ and $c_2$ are constants that satisfy $0 < c_1 < c_2 < 1$. (The difference with the weak Wolfe conditions (3.26) 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 (3.48b) 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 $L$, $\alpha$, and $U$ are required in these two cases.)

9. Consider the following function $f : \mathbb{R}^n \to \mathbb{R}$:

$$f(x) = \frac{1}{4} \sum_{l=1}^{d-1} \cos(x_l - x_{l+1}) + \sum_{l=1}^{n} l x_l^2 \, .$$

   (a) Compute a constant stepsize for which the gradient method is guaranteed to converge.

   (b) Characterize the stationary points $x$ (the points for which $\nabla f(x) = 0$). For each such 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 stationary point the algorithm converges. Explain your reasoning.

10. Prove the three-point property (3.39) for Bregman divergences.

11. Suppose the choice of constant steplength $\alpha$ (3.43) in the mirror descent algorithm is scaled by some positive constant $\theta$. Show how this modified choice changes the bound (3.44).