# Chapter 2

# Gradient Methods

The gradient method forms the foundation of all of the schemes studied in this book. We will provide several complementary perspectives on this algorithm that highlight the many different ways we can analyze and interpret optimization methods. This chapter can be read as both an introduction to the gradient method and to the fundamental tools for understanding optimization algorithms.

Throughout the chapter, we will be considered with the unconstrained minimization of a smooth convex function:

$$\min_{x \in \mathbb{R}^n} f(x). \tag{2.1}$$

The algorithms we consider in 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 cheaper 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, which allow us to handle these more general cases, will be considered in subsequent chapters.

## 2.1 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 2.1.** *$d$ is a descent direction for $f$ at $x$ if $f(x + td) < f(x)$ for all $t > 0$ sufficiently small.*

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

**Proposition 2.2.** *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 A.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 (A.6), 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.*

Note that when $\nabla f(x) = 0$, there are no directions $d$ that satisfy $\nabla f(x)^T d < 0$. This makes sense, since the condition $\nabla f(x) = 0$ is a necessary and sufficient condition for $x$ to be a minimizer of a convex function. This fact follows from the combination of Theorem 1.6 and Theorem 1.8.

In the next two sections, we consider the steepest descent method, in which the iterations are generated according to the following formula:

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k), \quad k = 0, 1, 2, \ldots, \tag{2.2}$$

for some steplength $\alpha_k > 0$.

## 2.2 Fixed point iteration

Our first view of the steepest descent method is as a fixed point iteration. Given that a solution of (2.1) for smooth convex $f$ is characterized by $\nabla f(x^*) = 0$, we can treat the minimization problem (2.1) by this equivalent system of nonlinear equations. A popular method for solving this system is *fixed point iteration.* In this approach, we identify some mapping $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ such that $x^* = \Phi(x^*)$ if and only if $\nabla f(x^*) = 0$. Fixed point iteration proceeds by iterating this mapping, setting

$$x^{k+1} = \Phi(x^k), \quad k = 0, 1, 2, \ldots . \tag{2.3}$$

A simple candidate is for our mapping

$$\Phi(x) = x - \alpha \nabla f(x), \quad \text{for some fixed } \alpha > 0. \tag{2.4}$$

This leads to a constant-step version of steepest descent. The $\alpha$ is a stepsize parameter that controls how much to follow the gradient at each iteration.

We assume typical conditions for a fixed-point iterative process, that is,

1. There exits an $x^*$ with $\nabla f(x^*) = 0$.
2. $\Phi(x) = x - \alpha \nabla f(x)$ is *contractive* for some $\alpha > 0$: i.e., there is a $\beta \in [0, 1)$ such that

$$\|\Phi(x) - \Phi(z)\| \le \beta \|x - z\| \quad \text{for all } x,\, z. \tag{2.5}$$

Then if we run the fixed-point method (2.4) starting at $x_0$, then we have the following recursive formula:

$$\begin{aligned} \|x^{k+1} - x^*\| &= \|x^k - \alpha \nabla f(x^k) - x^*\| \\ &= \|\Phi(x^k) - \Phi(x^*)\| \\ &\le \beta \|x^k - x^*\| \\ &\;\vdots \\ &\le \beta^{k+1} \|x_0 - x_\star\|. \end{aligned}$$

This derivation reveals that $x^k$ converges *linearly* to $x_\star$. That is, at every iteration, the distance to the optimal solution is decreased by at least a constant factor $\beta$. It is easy to show (see Section A.2) that the number of iterations $T$ required to obtain $\|x^k - x^*\| \le \epsilon$ is

$$T \ge \frac{\log(\|x^0 - x^*\|/\epsilon)}{|\log \beta|}.$$

What properties of $f$ are needed to guarantee that $\Phi$ satisfies the contraction property (2.5)? Let us consider the case of $f$ twice continuously differentiable, for which the result (A.7) holds, from Taylor's theorem. Suppose that the eigenvalues of $\nabla^f(x)$ are uniformly bounded in the range $[m, L]$ for all $x$. We then have

$$\begin{aligned} \|\Phi(x) - \Phi(z)\| &= \|(x - z) - \alpha(\nabla f(x) - \nabla f(z))\| \\ &= \left\| (x - z) - \alpha \int_0^1 \nabla^2 f(z + \gamma(x - z))(x - z)\, d\gamma \right\| \\ &= \left\| \left( I - \alpha \int_0^1 \nabla^2 f(z + \gamma(x - z)) \right) (x - z) \right\|. \end{aligned}$$

The eigenvalues of the matrix $I - \alpha \int_0^1 \nabla^2 f(z + \gamma(x - z))$ are in the range $[1 - \alpha L, 1 - \alpha m]$, so we will obtain the contraction property (2.5) provided that

$$-\beta \le 1 - \alpha L \le 1 - \alpha m \le \beta. \tag{2.6}$$

The second of these conditions requires that $m > 0$, that is, that $f$ is uniformly strongly convex. We conclude via the fixed-point analysis that the steepest-descent approach can converge at a linear rate. provided that $\nabla^2 f$ is uniformly positive definite.

From (2.6), we can also identify appropriate settings of $\alpha$ and $\beta$. Given $\beta < 1$, can we identify a range of $\alpha$ for which (2.6) holds? The first condition requires $1 - \alpha L \ge -\beta$, while the second requires $1 - \alpha m \le \beta$. Both conditions are satisfied when

$$\alpha \in \left[ \frac{1 - \beta}{m}, \frac{1 + \beta}{L} \right].$$

Note that this interval exists only when the lower bound is less than the upper bound. The value of $\beta$ for which the upper and lower bounds are equal is in fact the optimal value of $\beta$ — the smallest value for which it is possible to find $\alpha$ that satisfies (2.6). Elementary calculation shows that this value is $\beta = (L-m)/(L+m)$, and the corresponding value of $\alpha$ is $\alpha = 2/(L+m)$.

## 2.3 Steepest Descent

We return to the steepest descent method (2.7), and focus on the question of choosing the stepsize $\alpha_k$. 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. We assume here that $f$ is convex and smooth, and that its gradients satisfy the Lipschitz condition (1.20) with Lipschitz constant $L$. In this case, we can set $\alpha_k$ to be a constant, and simply iterate

$$x^{k+1} = x^k - (1/L)\nabla f(x^k), \quad k = 0, 1, 2, \dots . \tag{2.7}$$

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 (1.14), we obtain

$$\begin{aligned} f(x+\alpha d) &= f(x) + \alpha \nabla f(x)^T d + \alpha \int_0^1 [\nabla f(x+\gamma\alpha d) - \nabla f(x)] d \, d\gamma \\ &\le f(x) + \alpha \nabla f(x)^T d + \alpha \int_0^1 \|\nabla f(x+\gamma\alpha d) - \nabla f(x)\| \|d\| \, d\gamma \\ &\le f(x) + \alpha \nabla f(x)^T d + \alpha^2 \frac{L}{2}\|d\|^2, \end{aligned} \tag{2.8}$$

where we used (1.20) for the last line. For $x = x^k$ and $d = -\nabla f(x^k)$, the value of $\alpha$ that minimizes the expression on the right-hand side is $\alpha = 1/L$. By substituting these values, we obtain

$$f(x^{k+1}) = f(x^k - (1/L)\nabla f(x^k)) \le f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2. \tag{2.9}$$

This expression is one of the foundational inequalities in the analysis of optimization methods. Depending on the variety of assumptions about $f$, we can derive a variety of different convergence rates from this basic inequality.

### 2.3.1 General Case

From (2.9) alone, we can already prove convergence of the steepest descent method with no additional assumptions.

We can rearrange (2.9) and sum over the iterates of the first $T$ steps of the algorithm to find that

$$\begin{aligned} \sum_{k=0}^{T-1} ||\nabla f(x^k)||^2 &\le 2L \sum_{k=0}^{T-1} f(x^k) - f(x^{k+1}) \\ &= 2L[f(x^0) - f(x^n)] \\ &\le 2L[f(x^0) - f(x_\star)] \, . \end{aligned}$$

The second line follows because the sum telescopes.

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

$$\begin{aligned} \min_{0\le k\le T-1} ||\nabla f(x^k)|| &\le \sqrt{\frac{2L[f(x^0)-f(x^T)]}{T}} \\ &\le \sqrt{\frac{2L[f(x^0)-f(x_\star)]}{T}}. \end{aligned}$$

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

$$\|\nabla f(x)\| \le \sqrt{\frac{2L[f(x^0)-f(x_\star)]}{T}}. \tag{2.10}$$

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.

### 2.3.2 Convex Case

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

**Theorem 2.3.** *Suppose that $f$ is convex and smooth, where $\nabla f$ has Lipschitz constant $L$, and that* (2.1) *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*

$$f(x^T) - f_\star \le \frac{f(x^0)-f_\star}{T}. \tag{2.11}$$

*Proof.* By convexity of $f$, we have $f(x^*) \ge f(x^k) + \nabla f(x^k)^T(x^* - x^k)$, so by substituting into (2.9), we obtain for $k = 0, 1, 2, \ldots$ that

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

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

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

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

$$f(x^T) - f(x^*) \le \frac{1}{T}\sum_{k=0}^{T-1}(f(x^{k+1}) - f_\star) \le \frac{f(x^0)-f_\star}{T},$$

as required. □

### 2.3.3 Strongly Convex Case

A function $f : \mathbb{R}^d \to \mathbb{R}$ is *strongly convex* if there is a scalar $m > 0$ such that

$$f(z) \geq f(x) + \nabla f(x)^T (z - x) + \tfrac{m}{2}\|z - x\|^2 \tag{2.12}$$

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 a function is strongly convex *and* has $L$-Lipschitz gradients, then $f$ is bounded above and below by simple quadratics. This sandwiching will enable 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 differentiable function with $L$-Lipschitz gradients, then

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

is strongly convex for $\mu$ large enough. Verifying this fact is a fun exercise.

As another canonical example, note that a quadratic function $f(x) = \frac{1}{2}x^T Q x$ is strongly convex if and only if the smallest eigenvalue of $Q$ is strictly positive.

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. Note that if we minimize the right hand side of our strong convexity condition with respect to $x$, we find that the minimizer is $x - \frac{1}{m}\nabla f(x)$. Plugging that into (2.12), we find

$$\begin{aligned} f(z) &\geq \min_x f(x) + \nabla f(x)^T (z - x) + \frac{m}{2}||z - x||^2 \\ &\geq f(x) - \nabla f(x)^T \frac{1}{m}\nabla f(x) + \frac{m}{2}\left\|\frac{1}{m}\nabla f(x)\right\|^2 \\ &\geq f(x) - \frac{1}{2m}||\nabla f(x)||^2 \end{aligned} \tag{2.13}$$

Now if $||\nabla f(x)|| < \delta$ then

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

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 $x_\star$. Using (2.12) and Cauchy-Schwartz, we have

$$\begin{aligned} f(x_{opt}) &\geq f(x) + \nabla f(x)^T (x_\star - x) + \frac{m}{2}||x - x_\star||^2 \\ &\geq f(x) - ||\nabla f(x)||\; ||x_\star - x|| + \frac{m}{2}||x - x_\star||^2 \end{aligned}$$

Rearranging terms proves that

$$||x - x_\star|| \leq \frac{2}{m}||\nabla f(x)||\,. \tag{2.14}$$

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

An immediate consequence of (2.14) is the following

**Corollary 2.4.** *If $f$ is strongly convex then $f$ has a unique optimal solution.*

Essentially, strongly convex functions are nice, wide bowls, and we just need to roll downhill to the bottom.

We close this discussion of strong convexity by proving that for differentiable functions, strong convexity is equivalent to the Hessian having positive eigenvalues. We encountered this fact when we were discussing contraction mappings in the previous lecture.

**Proposition 2.5.** *If $f$ is strongly convex and two-times differentiable, then $\nabla^2 f(x) \succeq mI$*

*Proof.* Using the fact that

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

and

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

we can add these two inequalities together and get

$$\langle \nabla f(x) - \nabla f(y), x-y \rangle \geq m||x-y||^2 .$$

Setting $x = u + \alpha v$ and $y = u$, for $u$ and $v$ in $\mathbb{R}^d$ yields

$$\langle \nabla f(u+\alpha d) - \nabla f(v), \alpha v \rangle \geq m\alpha^2 ||v||^2$$

Dividing through by $\alpha^2$ and taking the limit as $\alpha$ goes to zero proves

$$d^T \nabla^2 f(x) d \geq m||x-y||^2$$

as desired. □

Combining these facts about strongly convex functions with our basic inequality (2.9), we have

$$\begin{aligned} f(x^{k+1}) = f(x^k - (1/L)\nabla f(x^k)) & \qquad \leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \\ \leq f(x^k) - \frac{m}{L}(f(x^k) - f_\star) & \end{aligned}$$

Subtracting $f_\star$ from both sides of this inequality gives us the recursion

$$f(x^{k+1}) - f_\star \leq \left(1 - \frac{m}{L}\right)(f(x^k) - f_\star) . \tag{2.15}$$

This asserts that the function values converge *linearly* to the optimum. After $T$ steps, we have

$$f(x^T) - f_\star \leq \left(1 - \frac{m}{L}\right)^T (f(x^0) - f_\star) . \tag{2.16}$$

### 2.3.4 Comparison Between Rates

It is straightforward to convert these convergence expressions into complexities, using the techniques of Appendix A.2. We have from (2.10) that an iterate $k$ will be found such that $\|\nabla f(x^k)\| \le \epsilon$ for some $k \le T$, where

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

For the weakly convex case, we have from (2.11) that $f(x^k) - f^* \le \epsilon$ when

$$T \le \frac{f(x^0) - f^*}{\epsilon}. \tag{2.17}$$

For the strongly convex case, we have from (2.16) that $f(x^k) - f^* \le \epsilon$ for all $k$ satisfying

$$k \ge \frac{L}{m} \log((f(x^0) - f^*)/\epsilon). \tag{2.18}$$

This is the same rate of convergence we derived for the iterates using the contraction mapping analysis in Section 2.2. 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

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

The linear rate (2.18) 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 $L/m$ 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 (2.18) and the sublinear rate (2.17).

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 (2.9) 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.

## 2.4 Descent Methods

The simplest method for optimization of a smooth function when the gradient $\nabla f$ can be calculated efficiently is the steepest-descent method, in which we take a short step in the negative gradient direction $-\nabla f(x)$. This approach has intuitive appeal, and it can be seen immediately from Taylor's theorem (by setting $p = -\alpha \nabla f(x)$ in (1.14) for $\alpha$ small and positive) that it will produce decrease in $f$ at non-stationary points, provided that $\alpha$ is sufficiently small.

In this chapter, we consider methods in which each step has the form

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, 2, \dots, \tag{2.19}$$

where $\alpha_k > 0$ and $d^k$ is a search direction that satisfies the following properties:

$$-(d^k)^T \nabla f(x^k) \ge \bar{\epsilon} \|\nabla f(x^k)\| \|d^k\|, \quad \gamma_1 \|\nabla f(x^k)\| \le \|d^k\| \le \gamma_2 \|\nabla f(x^k)\|, \tag{2.20}$$

for some positive constants $\bar{\epsilon}$, $\gamma_1$, $\gamma_2$. The first condition says that the angle between $-\nabla f(x^k)$ and $d^k$ is acute, and bounded away from $\pi/2$, while the second condition 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 $\nabla f(x^k) = 0$ and thus $d^k = 0$ according to the second condition above—the method will not step away from a stationary point.

For the "obvious" choice of search direction—the negative gradient $d^k = -\nabla f(x^k)$—the conditions (2.20) 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 (1.14), we obtain

$$\begin{aligned}
f(x^{k+1}) &= f(x^k + \alpha d^k) \\
&= f(x^k) + \alpha \nabla f(x^k)^T d^k + \alpha \int_0^1 [\nabla f(x^k + \gamma \alpha d^k) - \nabla f(x^k)] d^k \, d\gamma \\
&\le f(x^k) + \alpha \nabla f(x^k)^T d^k + \alpha \int_0^1 \|\nabla f(x^k + \alpha d^k) - \nabla f(x^k)\| \|d^k\| \, d\gamma \\
&\le f(x^k) - \alpha \bar{\epsilon} \|\nabla f(x^k)\| \|d^k\| + \alpha^2 \frac{L}{2} \|d^k\|^2 \\
&\le f(x^k) - \alpha \left( \bar{\epsilon} - \alpha \frac{L}{2} \gamma_2 \right) \|\nabla f(x^k)\| \|d^k\|, \qquad (2.21)
\end{aligned}$$

where we used (1.20) for the second-last line and (2.20) throughout. 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.

In deriving the bound (2.21), we did not require convexity of $f$, only Lipschitz continuity of the gradient $\nabla f$. The same is true for most of the analysis in this section. Convexity used only in proving rates of convergence to a solution $x^*$, in Sections 2.6, **??**, and **??**. (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 $\gamma_1$ and $\gamma_2$ are positive quantities as in (2.20). The second condition in (2.20) hold, by definition of $S^k$, and the first condition holds with $\bar{\epsilon} = \gamma_1/\gamma_2$, since

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

  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 $x^k$. But our focus here is on methods that use only first-derivative information, so we delay further consideration of these methods until Chapters 12 and **??**.

- 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 (2.20) 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 5) 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 (2.20) 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 $\{1, 2, \dots, n\}$ 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 (2.20) 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 (2.20) also holds in an expected sense. Rigorous analysis of these methods is presented in Chapter 5.

- Another important class of randomized schemes are the stochastic gradient methods discussed in Chapter 4. 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 (2.20) 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 $\xi_k$. Further analysis of stochastic gradient methods appears in Chapter 4.

## 2.5 Line Searches

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

$$x^{k+1} = x^k + \alpha_k d^k. \tag{2.22}$$

We assume for this discussion that $d^k$ satisfies the properties (2.20). We discuss here the issue of choosing $\alpha_k$. There are several alternative approaches, of varying theoretical and practical validity.

**Constant Stepsize** As we have seen in Sections 2.2 and **??**, 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—a common one 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 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, $\alpha_k$ 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 is to base the choice of $\alpha_k$ on knowledge of the global properties of the function $f$, for example, on the Lipschitz constant $L$ for the gradient (see (1.20)) or the modulus of convexity $\mu$ (see (1.6)). We call such variants "short-step" methods. Given the expression (2.21) above, for example, and supposing we have estimates of all the quantities $\gamma_1$, $\gamma_2$, 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 (2.21) and (2.20) that

$$f(x^{k+1}) \leq f(x^k) - \frac{\bar{\epsilon}^2}{2L\gamma_2}\|\nabla f(x^k)\|\|d^k\| \geq f(x^k) - \frac{\bar{\epsilon}^2\gamma_1}{2L\gamma_2}\|\nabla f(x^k)\|^2. \tag{2.23}$$

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:

$$\min_{\alpha>0} f(x^k + \alpha d^k). \tag{2.24}$$

This technique requires the ability to evaluate $f(x^k + \alpha d^k)$ (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 very low cost. For example, if $f$ is a multivariate polynomial, the line search problem 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 5, 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 (2.23), it is clear that (2.23) also holds for exact line searches.

**Approximate line search** In full generality, exact line searches are often very expensive, and better empirical performance is achieved by approximate line search. Indeed, it is usually not worth the effort to be *too* precise in identifying the exact minimizing $\alpha$, so 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 such approximate solutions economically. (By "economically," we mean that an average of three or less evaluations of $f$ are required.[1]) One popular pair of conditions that the approximate minimizer $\alpha = \alpha_k$ is required to satisfy, called the *Weak Wolfe Conditions*, is as follows:

$$f(x^k + \alpha d^k) \leq f(x^k) + c_1 \alpha \nabla f(x^k)^T d^k, \tag{2.25a}$$

$$\nabla f(x^k + \alpha d^k)^T d^k \geq c_2 \nabla f(x^k)^T d^k\,. \tag{2.25b}$$

[1] Yes, it's correct grammar to use "less" rather than "fewer" here, because this comparator refers to the "average," which is not necessarily a positive integer.

Here, $c_1$ and $c_2$ are constants that satisfy $0 < c_1 < c_2 < 1$. The condition (2.25a) 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 (2.25b) ensures that $\alpha_k$ is not too short; it ensures that we move far enough along $d^k$ that the magnitude of the directional derivative of $f$ along $d^k$ is substantially reduced over its value at $\alpha = 0$. It can be shown that there exist values of $\alpha_k$ that satisfy both conditions simultaneously. To show that these conditions imply a reduction in $f$ that is related to $\|\nabla f(x^k)\|^2$ (as in (2.23)), we argue as follows. First, from condition (2.25b) and the Lipschitz property for $\nabla f$, we have

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

and thus

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

Substituting into (2.25a), and using the first condition in (2.20), then yields

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

**Backtracking Line Search** Another popular approach to determining an appropriate value for $\alpha_k$ is known as "backtracking." It is widely used when evaluation of $f$ is practical, because it is very easy to implement (no estimate of the Lipschitz constant $L$ is required, for example), requires no additional gradient evaluations, 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 $\bar{\alpha}, \beta\bar{\alpha}, \beta^2\bar{\alpha}, \beta^3\bar{\alpha}, \dots$ for which a sufficient decrease condition (2.25a) is satisfied. Note that backtracking does not require a condition like (2.25b) 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.

## 2.6 Convergence of Iterates

As is clear from the previous section, most schemes of descent type produce iterates that satisfy a bound of the form

$$f(x^{k+1}) \le f(x^k) - C\|\nabla f(x^k)\|^2, \quad \text{for some } C > 0. \tag{2.26}$$

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

**Theorem 2.6.** *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* (2.26) *are stationary, that is, $\nabla f(\bar{x}) = 0$. If in addition $f$ is convex, each such $\bar{x}$ is a solution of (2.1).*

*Proof.* Note first from (2.26) that

$$\|\nabla f(x^k)\|^2 \leq [f(x^k) - f(x^{k+1})]/C, \quad k = 0, 1, 2, \ldots, \tag{2.27}$$

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 (2.1). □

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 bounded below but does not attain its minimum value.) It is, however, an immediate corollary of Theorem 2.6 that if $\{x^k\}$ is bounded and if there is a unique solution $x^*$, then $\lim_{k\to\infty} x^k = x^*$.

## Notes and References

Other, slightly more complicated line-search procedures are discussed in [15, Chapter 3]. There is a discussion of step-size methods that perform a robust, efficient one-dimensional search to identify a value of $\alpha_k$ that satisfy the Wolfe conditions.

Stole the proof for weakly convex from Vandenberghe notes.

Weak Wolfe line search is adapted from Burke and Overton and Lewis.

## Exercises

1. Verify that if $f$ is twice continuously differentiable with the Hessian satisfying $mI \preceq \nabla^2 f(x)$ for all $x \in \text{dom}(f)$, then the strong convexity condition (1.6) is satisfied.

2. Show, as a corollary of Theorem 2.6 that if the sequence $\{x^k\}$ described in this theorem is bounded and if $f$ is strictly convex, we have $\lim_{k\to\infty} x^k = x^*$.

3. How much of the analysis of Sections 2.3, 2.4, 2.5, and 2.6 applies to smooth *nonconvex* functions? Specifically, state an analog of Theorem 2.6 that is true when the assumption of convexity of $f$ is dropped.

4. How is the analysis of Section 2.3 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 $\alpha$.

5. Find positive values of $\bar{\epsilon}$, $\gamma_1$, and $\gamma_2$ such that the Gauss-Southwell choice $d^k = -[\nabla f(x^k)]_{i_k}$, where $i_k = \arg\min_{i=1,2,\ldots,n} |[\nabla f(x^k)]_i|$ satisfies conditions (2.20).