Chapter 9

Nonsmooth Optimization Methods

The gradient method for smooth functions f, described in Chapter 3, is intuitive in that is that it follows the direction of steepest descent at each iteration, which is a guaranteed direction of descent for f. Generalizing this method to nonsmooth functions f is not straighforward, as the "gradient" may not be uniquely defined, even for convex f, as we saw in Chapter 8. A natural idea would be to choose a vector from the subdifferential ∂f and use the negative of this vector as a search direction, but simple examples show that sucha direction may not given descent in f.

The simplest example is the absolute value function $f(x) = |x|$, where $x \in \mathbb{R}$. At the minimizing value $x = 0$, the subdifferential is $\partial|0| = [-1, 1]$, and any vector drawn from this interval (except for the very special choice $g = 0$) will step away from 0 and thus increase the function value. The situation only gets worse in higher dimensions. Consider the two-dimensional function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x_1, x_2) = |x_1| + 2|x_2|,$$

whose optimum is $(0, 0)$. At the point $(1, 0)$, the subdifferential here is the compact (but infinite) set

$$\partial f(1, 0) = \{(1, z) \,|\, |z| \leq 2\}$$

For the particular subgradient $g = (1, 2)$, the directional derivative in the negative of this direction is

$$f'((1, 0); (-1, -2)) = \sup_{g \in \partial f(1,0)} -g_1 - 2g_2 = -1 + 4 = 3\,,$$

showing that the function *increases* along this direction. These trivial examples, and another example illustrated in Figure 9.1 show that it is not obvious how to design a method that follows subgradients.

However, methods based on subgradients exist and are effective, and we describe three of them in this chapter. First, show how to compute the direction of steepest descent of a convex nonsmooth function, and will reveal that this direction is the negative of a subgradient (albeit a very special subgradient). Second, we show how using carefully selected stepsizes will allow us to follow *arbitrary* subgradients, even ones that increase the function, and still get provable convergence behavior over the long term. (Convergence of these methods is however quite slow, both in theory and practice.) Finally, we describe *proximal* methods, which exploit the structure of some interesting special cases of nonsmooth functions to obtain faster convergence.

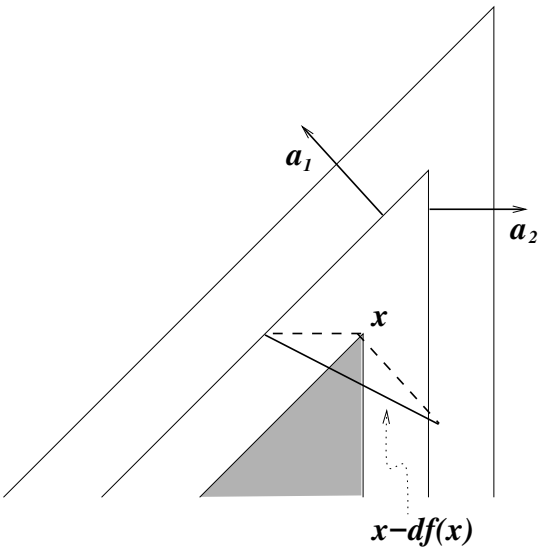

Figure 9.1: Subgradient of a function that is the max of two planes defined by vectors a_1 and a_2. Given a point x at which both planes achieve the maximum, the subgradient is $\partial f(x) = \{\lambda a_1 + (1-\lambda)a_2 \,|\, \lambda \in [0,1]\}$. The set of points $\{x - g \,|\, g \in \partial f(x)\}$ is a line segment (illustrated). The shaded region is the set of points with a smaller function value than $f(x)$. Note that some points of the form $x - \alpha g$ for $\alpha > 0$ and $g \in \partial f(x)$ have $f(x - \alpha g) < f(x)$. However there are other points with the same form for which $f(x - \alpha g) > f(x)$ for all $\alpha > 0$. That is, some but not all negative subgradients yield descent in f.

9.1 Subgradient Descent

When x is not a minimizer of f, the subdifferential $\partial f(x)$ always contains a vector g such that $-g$ is a descent direction for f. The vector $g_{\min}$ with *smallest norm* in $\partial f(x)$ has this property, and in fact $-g_{\min}$ is the direction of *steepest descent*. We define

$$g_{\min} := \arg \min_{z \in \partial f(x)} \|z\|_2 \,. \tag{9.1}$$

$g_{\min}$ exists because $\partial f(x)$ is nonempty and compact.

Proposition 9.1. *For a convex function f, and $x \in \operatorname{dom} f$ that is not a minimizer of f, the vector $-g_{\min}$ defined from (9.1) is a descent direction at x.*

Proof. Note first that we have

$$\langle g_{\min}, \hat{g} - g_{\min} \rangle \geq 0, \quad \text{for all } \hat{g} \in \partial f(x). \tag{9.2}$$

To see this, suppose otherwise. Then since

$$g_{\min} + t(\hat{g} - g_{\min}) \in \partial f(x)$$

for all $t \in [0, 1]$, we have

$$\frac{d}{dt} \|g_{\min} + t(\hat{g} - g_{\min})\|^2 \Big|_{t=0} = 2\langle g_{\min}, \hat{g} - g_{\min} \rangle < 0$$

which contradicts the fact that $g_{\min}$ minimizes $\|g\|^2$ over $g \in \partial f(x)$. From (9.2), we have $\langle \hat{g}, g_{\mathrm{mn}} \rangle \geq \|g_{\mathrm{mn}}\|_2^2$ for all $\hat{g} \in \partial f(x)$, so that

$$f'(x; -g_{\min}) = \sup_{g \in \partial f(x)} \langle -g_{\min}, g \rangle = - \inf_{g \in \partial f(x)} \langle g_{\min}, g \rangle \leq -\|g_{\min}\|_2^2,$$

proving that $-g_{\min}$ is a descent direction whenever it is nonzero. To see that $-g_{\min}$ is the *steepest* descent direction, we use a min-max argument. Note that

$$\inf_{\|v\| \leq 1} f'(x; v) = \inf_{\|v\| \leq 1} \sup_{g \in \partial f(x)} \langle v, g \rangle \leq \sup_{g \in \partial f(x)} \inf_{\|v\| \leq 1} \langle v, g \rangle = \sup_{g \in \partial f(x)} -\|g\| = -\|g_{\mathrm{mn}}\| \,.$$

The inequality in this expression follows from *weak duality*, which we cover in the next chapter. The reader can verify, however, that for any function $\varphi(x, z)$, we have

$$\inf_x \sup_z \varphi(x, z) \leq \sup_z \inf_x \varphi(x, z) \,.$$

completing the proof. □

Example 9.1. *Consider the function $f(x) = \|x\|_1$, whose minimizer is $x = 0$. At any nonzero x, the subdifferential $\partial \|x\|_1$ consists of vectors g such that*

$$g_i \in \begin{cases} \{+1\} & \text{if } x_i > 0 \\ \{-1\} & \text{if } x_i < 0 \\ [-1, 1] & \text{if } x_i = 0. \end{cases}$$

The minimum-norm subgradient is thus $g_{\min}$, where

$$(g_{\min})_i = \begin{cases} +1 & \text{if } x_i > 0 \\ -1 & \text{if } x_i < 0 \\ 0 & \text{if } x_i = 0. \end{cases}$$

Proposition 9.1 suggests a natural algorithm for minimizing convex, nonsmooth functions: Compute the minimum norm element of the subdifferential and search along the negative of this direction. The problem with this approach is that *computing the minimum norm element might be prohibitively expensive*. In the next section, we show that a naive algorithm that simply follows arbitrary subgradients can converge, under appropriate choices of steplengths.

9.2 The Subgradient Method

The subgradient method is rather simple: Starting from a point x_1, at each step k, we choose *any* element of the subdifferential $g_k \in \partial f(x_k)$ and set

$$x_{k+1} = x_k - \alpha_k g_k \,.$$

Though we have already pointed out that this method may take steps that increase f, the weighted average of all iterates encountered so far, defined by

$$\bar{x}_T = \lambda_T^{-1} \sum_{j=1}^{T} \alpha_k x_k, \quad \text{where } \lambda_T := \sum_{j=1}^{T} \alpha_j \tag{9.3}$$

is well behaved, and may even converge to a minimizer of f.

The analysis of this method is nearly identical to the proof of convergence of the stochastic gradient method for convex functions with bounded stochastic gradients. We assume that for all

$$\|g\|_2 \leq G, \quad \text{for all } g \in \partial f(x) \text{ and all } x.$$

Note that this assumption implies that f must be Lipschitz with constant G (why?). We also denote by $x_\star$ a minimizer of f, and define

$$D_0 := \|x_1 - x_\star\|, \tag{9.4}$$

which is the distance of the initial point to a minimizer of f.

To proceed with our analysis of the behavior of the weighted-average iterate $\bar{x}_T$, we expand the distance to an optimal solution of iterate x_{k+1}:

$$\begin{aligned} \|x_{k+1} - x_\star\|^2 &= \|x_k - \alpha_k g_k - x_\star\|^2 \\ &= \|x_k - x_\star\|^2 - 2\alpha_k \langle g_k, x_k - x_\star \rangle + \alpha_k^2 \|g_k\|^2 \\ &\leq \|x_k - x_\star\|^2 - 2\alpha_k \langle g_k, x_k - x_\star \rangle + \alpha_k^2 G^2 \,. \end{aligned} \tag{9.5}$$

Note that this exactly looks like our basic inequality for the subgradient method (5.26), except there are no expected values here. We can rearrange (9.5) to obtain

$$\alpha_k \langle g_k, x_k - x_\star \rangle \leq \frac{1}{2}\|x_k - x_\star\|^2 - \frac{1}{2}\|x_{k+1} - x_\star\|^2 + \frac{1}{2} G^2 \alpha_k^2. \tag{9.6}$$

Since $g_k \in \partial f(x_k)$, we have by the definition of subgradient that

$$f(x_k) - f(x_\star) \leq \langle g_k, x_k - x_\star \rangle. \tag{9.7}$$

By multiplying both sides of (9.7) by $\alpha_k > 0$, combining with (9.6), summing both sides from $k = 1$ to $k = T$, and using convexity of f, we obtain

$$\begin{aligned} f(\bar{x}_T) - f(x_\star) &\leq \lambda_T^{-1} \sum_{k=1}^{T} \alpha_k (f(x_k) - f(x_\star)) \\ &\leq \lambda_T^{-1} \frac{1}{2} \sum_{k=1}^{T} \left(\|x_k - x_\star\|^2 - \|x_{k+1} - x_\star\|^2 \right) + \frac{1}{2} \lambda_T^{-1} G^2 \sum_{k=1}^{T} \alpha_k^2 \\ &\leq \lambda_T^{-1} \frac{1}{2} \left(\|x_1 - x_\star\|^2 - \|x_{T+1} - x_\star\|^2 \right) \right) + \frac{1}{2} \lambda_T^{-1} G^2 \sum_{k=1}^{T} \alpha_k^2 \\ &\leq \frac{D_0^2 + G^2 \sum_{k=1}^{T} \alpha_k^2}{2 \sum_{k=1}^{T} \alpha_k}. \end{aligned} \tag{9.8}$$

We also immediately have the bound

$$\min_{t \leq T} f(x_t) - f(x_\star) \leq \lambda_T^{-1} \sum_{t=1}^{T} \alpha_t (f(x_t) - f(x_\star))$$

so our analysis works for both the weighted average of the first T iterates and the *best* of these iterates.

9.2.1 Stepsizes

Let us look at different possibilities for our stepsizes α_k, $k = 1, 2, \ldots$.

Constant step size. First, we can just pick $\alpha_k = \alpha$ for all k. In this case, we from (9.8) that

$$f(\bar{x}_T) - f(x_\star) \leq \frac{D_0^2 + TG^2\alpha^2}{2T\alpha}.$$

The choice $\alpha = \frac{\theta D_0}{G\sqrt{T}}$ for some parameter $\theta > 0$ yields

$$f(\bar{x}_T) - f(x_\star) \leq \tfrac{1}{2} \left(\theta + \theta^{-1} \right) \frac{D_0 G}{\sqrt{T}},$$

and the bound is minimized when we set $\theta = 1$.

Constant step length. An alternative is to choose $\alpha_k = \frac{\alpha}{\|g_k\|}$, so that the *length* of each step $\alpha_k g_k$ is constant. A slight modification of the analysis above yields the bound

$$f(\bar{x}_T) - f(x_\star) \leq \frac{D_0^2 + T\alpha^2}{2T\alpha/G}.$$

Setting $\alpha = \frac{\theta D_0}{\sqrt{T}}$, we obtain

$$f\left(\bar{x}_T\right) - f(x_\star) \leq \tfrac{1}{2}\left(\theta + \theta^{-1}\right) \frac{D_0 G}{\sqrt{T}}\,,$$

which matches our bound for the constant step size. Note that here our step size depends only D_0 (distance of x_1 to optimality) and not G (maximal subgradient norm).

An interesting feature of both choices so far is that the bound is not very sensitive to errors in the estimates of D_0 and G. Such errors can be captured in the parameter θ, and we see that the bound increases by only the modest factor $\frac{1}{2}(\theta + \theta^{-1})$ when θ moves away from its optimal value of 1.

9.2.2 Diminishing Step Size

The fixed stepsizes above required us to make a prior choice of T, the number of iterates to be taken. We now consider making choices of α_k that depend on k, and that decrease as k increases. Such choices do not require us to choose T in advance, and guarantee convergence to the optimal value of f as the number of iterates goes to ∞.

From (9.8), we see that for any sequence $\alpha_k > 0$ such that $\alpha_k \to 0$, but $\sum_{k=1}^{T} \alpha_k \uparrow \infty$ as $T \to \infty$, then

$$\lim_{T\to\infty} f\left(\bar{x}_T\right) = f(x_\star)\,.$$

This is particularly easy to see if $\sum_k \alpha_k^2 = M < \infty$, because we have from (9.8) that

$$f(\bar{x}_T) - f_\star \leq \frac{D_0^2 + G^2 \sum_{j=1}^{T} \alpha_j^2}{2\sum_{t=1}^{T} \alpha_t} \leq \frac{D_0^2 + G^2 M}{2\sum_{j=1}^{T} \alpha_j},$$

and the left-hand side clearly tends to zero as $T \to \infty$. To see that this approach works for general diminishing stepsizes, one only needs to prove that

$$\frac{\sum_{j=1}^{T} \alpha_j^2}{\sum_{j=1}^{T} \alpha_j} \to 0, \quad \text{as } T \to \infty,$$

whenever α_k tends to zero but $\sum_{k=1}^{T} \alpha_k$ diverges.

We close this section by deriving more quantitative bounds for an explicit stepsize choice. Setting $\alpha_k = \frac{\theta}{\sqrt{k}}$, we have

$$f(\bar{x}_T) - f_\star \leq \frac{D_0^2 + G^2\theta^2 \sum_{j=1}^{T} j^{-1}}{2\theta \sum_{j=1}^{T} j^{-1/2}} \leq \frac{D_0^2 + G^2\theta^2(\log T + 1)}{2\theta\sqrt{T}}\,. \tag{9.9}$$

The upper bound in the numerator comes from the Riemann-sum bound

$$\sum_{j=1}^{T} j^{-1} \leq 1 + \int_{t=1}^{T} \frac{1}{t}\, dt \leq \log T + 1,$$

while the lower bound in the denominator comes from

$$\sum_{j=1}^{T} j^{-1/2} \geq \sum_{j=1}^{T} T^{-1/2} = T^{1/2}.$$

Note that this bound tends to zero at a rate of $\log(T)/\sqrt{T}$. This is very slightly slower than the $1/\sqrt{T}$ rate of a constant stepsize, but we are guaranteed asymptotic convergence to zero, and can continue to iterate well beyond a fixed number of iterations.

The alternative diminishing stepsize choice $\alpha_k \propto k^{-p}$ for $p \in (0, 1)$ yields a worse convergence bound than for $p = 1/2$.

More sophisticated schemes for choosing stepsizes involve a combination of fixed and diminishing sizes. The stepsize is fixed for a number of consecutive iterations (sometimes called an *epoch*), and then decreased to a smaller value, which again is fixed for a number of consecutive iterations.

9.3 Proximal-Gradient Algorithms for Regularized Optimization

While provably correct, the $1/\sqrt{T}$ rate of the subgradient method is considerably slower than the rates achievable for smooth functions. In this section, we explore how to exploit partial smoothness to accelerate the convergence rates for nonsmooth convex optimization. In particular, we describe an elementary but powerful approach for solving the regularized optimization problem

$$\min_{x \in \mathbb{R}^n} \phi(x) := f(x) + \tau\psi(x), \tag{9.10}$$

where f is a smooth convex function, ψ is a convex regularization function (known simply as the "regularizer"), and $\tau \geq 0$ is a regularization parameter. The technique we describe here is a natural extension of the steepest-descent approach, in that it reduces to the steepest-descent method analyzed in Theorem 3.3 applied to f when the regularization term is not present ($\tau = 0$). It is useful when the regularizer ψ has a simple structure that is easy to account for explicitly, as is true for many regularizers that arise in data analysis, such as the ℓ_1 function ($\psi(x) = \|x\|_1$) and the indicator function for a simple set Ω ($\psi(x) = I_\Omega(x)$), such as a box $\Omega = [l_1, u_1] \otimes [l_2, u_2] \otimes \ldots \otimes [l_n, u_n]$. Moreover, as we will see, the convergence rate will be dictated by the smooth part of the decomposition in (9.10) even thought he function ϕ is not smooth.

Each step of the algorithm is defined as follows:

$$x^{k+1} := \text{prox}_{\alpha_k \tau\psi}(x^k - \alpha_k \nabla f(x^k)), \tag{9.11}$$

for some steplength $\alpha_k > 0$, and the prox operator defined in (8.25). By substituting into this definition, we can verify that x^{k+1} is the solution of an approximation to the objective ψ of (9.10), namely:

$$x^{k+1} := \arg\min_z \nabla f(x^k)^T (z - x^k) + \frac{1}{2\alpha_k}\|z - x^k\|^2 + \tau\psi(z). \tag{9.12}$$

One way to verify this equivalence is to note that the objective in (9.12) can be written as

$$\frac{1}{\alpha_k}\left\{\frac{1}{2}\left\|z - (x^k - \alpha_k \nabla f(x^k))\right\|^2 + \alpha_k \tau\psi(x)\right\},$$

(modulo a term $\alpha_k\|\nabla f(x^k)\|^2$ that does not involve z). The subproblem objective in (9.12) consists of a linear term $\nabla f(x^k)^T(z-x^k)$ (the first-order term in a Taylor-series expansion), a proximality term $\frac{1}{2\alpha_k}\|z-x^k\|^2$ that becomes more strict as $\alpha_k \downarrow 0$, and the regularization term $\tau\psi(x)$ in unaltered form. When $\tau = 0$, we have $x^{k+1} = x^k - \alpha_k\nabla f(x^k)$, so the iteration (9.11) (or (9.12)) reduces to the usual steepest-descent approach discussed in Chapter 3 in this case. It is useful to continue thinking of α_k as playing the role of a line-search parameter, though here the line search is expressed implicitly through a proximal term.

The key idea behind the proximal gradient algorithm is summed up by the following proposition that shows that every fixed point of (9.11) is a minimizer of ϕ:

Proposition 9.2. *Let f be differentiable and convex and let ψ be convex. $x_\star$ is an optimal solution of (9.10) if and only if $x_\star = \operatorname{prox}_{\alpha\tau\psi}(x_\star - \alpha\nabla f(x_\star))$ for all $\alpha > 0$.*

Proof. $x_\star$ is an optimal solution if and only if $-\nabla f(x_\star) \in \partial\tau\psi(x_\star)$. This is equivalent to

$$(x_\star - \alpha\nabla f(x_\star)) - x_\star \in \alpha\partial\tau\psi(x_\star)\,,$$

which is equivalent to $x_\star = \operatorname{prox}_{\alpha\tau\psi}(x_\star - \alpha\nabla f(x_\star))$. □

With regards to convergence, the linear convergence of the proximal gradient method when f is strongly convex problems can be derived in a nearly similar way to that of the projected gradient method. Indeed, we only need to invoke the nonexpansive property of the proximity operator (See Proposition 8.20) and then follow the argument in Section 7.3.2 to conclude

Proposition 9.3. *Let f be have L-Lipschitz gradients and strong convexity parameter m and let ψ be convex. Let $x_\star$ be the unique minimizer of $\phi = f + \tau\psi$. Then the iterates of the proximal gradient method with stepsize $\frac{2}{m+L}$ satisfy*

$$\|x_k - x_\star\| \leq \left(\frac{\kappa-1}{\kappa+1}\right)^k \|x_0 - x_\star\|\,. \tag{9.13}$$

For weakly convex functions, the proof is more delicate, and we now turn to a proof that will show that the proximal gradient method converges at a rate of $1/T$, just as in the case of smooth convex functions.

9.3.1 Convergence Rate for Weakly Convex f

We will demonstrate convergence of the method (9.11) at a sublinear rate, for functions f whose gradients satisfy a Lipschitz continuity property with Lipschitz constant L (see (2.7)), and for the constant steplength choice $\alpha_k = 1/L$. We follow closely the approach in the lecture on "Proximal Gradient Methods" of [32].

The proof makes use of a "gradient map" defined by

$$G_\alpha(x) := \frac{1}{\alpha}\left(x - \operatorname{prox}_{\alpha\tau\psi}(x - \alpha\nabla f(x))\right). \tag{9.14}$$

By comparing with (9.11), we see that this map defines the step taken at iteration k:

$$x^{k+1} = x^k - \alpha_k G_{\alpha_k}(x^k) \quad \Leftrightarrow \quad G_{\alpha_k} = \frac{1}{\alpha_k}(x^k - x^{k+1}). \tag{9.15}$$

The following technical lemma reveals some useful properties of $G_\alpha(x)$.

Lemma 9.4. *Suppose that in problem* (9.10)*, ψ is a closed convex function and that f is is convex with Lipschitz continuous gradient on $\mathbb{R}^n$, with Lipschitz constant L. Then for the definition* (9.14) *with $\alpha > 0$, the following claims are true.*

(a) $G_\alpha(x) \in \nabla f(x) + \tau \partial\psi(x - \alpha G_\alpha(x))$.

(b) For any z, and any $\alpha \in (0, 1/L]$, we have that

$$\phi(x - \alpha G_\alpha(x)) \leq \phi(z) + G_\alpha(x)^T(x - z) - \frac{\alpha}{2}\|G_\alpha(x)\|^2.$$

Proof. For part (a), we use the optimality property (8.26) of the prox operator, and make the following substitutions: $x - \alpha\nabla f(x)$ for "x", α for "λ", and $\tau\psi$ for "h" to obtain

$$0 \in \alpha\tau\partial\psi(\text{prox}_{\alpha\tau\psi}(x - \alpha\nabla f(x))) + (\text{prox}_{\alpha\tau\psi}(x - \alpha\nabla f(x)) - (x - \alpha\nabla f(x)).$$

We use definition (9.14) to make the substitution $\text{prox}_{\alpha\tau\psi}(x - \alpha\nabla f(x)) = x - \alpha G_\alpha(x)$, to obtain

$$0 \in \alpha\tau\partial\psi(x - \alpha G_\alpha(x)) - \alpha(G_\alpha(x) - \nabla f(x)),$$

and the result follows when we divide by α.

For (b), we start with the following consequence of Lipschitz continuity of ∇f, from Lemma **??**:

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2.$$

By setting $y = x - \alpha G_\alpha(x)$, for any $\alpha \in (0, 1/L]$, we have

$$\begin{aligned} f(x - \alpha G_\alpha(x)) &\leq f(x) - \alpha G_\alpha(x)^T\nabla f(x) + \frac{L\alpha^2}{2}\|G_\alpha(x)\|^2 \\ &\leq f(x) - \alpha G_\alpha(x)^T\nabla f(x) + \frac{\alpha}{2}\|G_\alpha(x)\|^2. \end{aligned} \tag{9.16}$$

(The second inequality uses $\alpha \in (0, 1/L]$.) We also have by convexity of f and ψ that for any z and any $v \in \partial\psi(x - \alpha G_\alpha(x)$ the following are true:

$$f(z) \geq f(x) + \nabla f(x)^T(z - x), \quad \psi(z) \geq \psi(x - \alpha G_\alpha(x)) + v^T(z - (x - \alpha G_\alpha(x))). \tag{9.17}$$

We have from part (a) that $v = (G_\alpha(x) - \nabla f(x))/\tau \in \partial\psi(x - \alpha G_\alpha(x))$, so by making this choice of v in (9.17) and also using (9.16) we have for any $\alpha \in (0, 1/L]$ that

$$\begin{aligned} \phi(x - \alpha G_\alpha(x)) &= f(x - \alpha G_\alpha(x)) + \tau\psi(x - \alpha G_\alpha(x)) \\ &\leq f(x) - \alpha G_\alpha(x)^T\nabla f(x) + \frac{\alpha}{2}\|G_\alpha(x)\|^2 + \tau\psi(x - \alpha G_\alpha(x)) \quad \text{(from (9.16))} \\ &\leq f(z) + \nabla f(x)^T(x - z) - \alpha G_\alpha(x)^T\nabla f(x) + \frac{\alpha}{2}\|G_\alpha(x)\|^2 \\ &\quad + \tau\psi(z) + (G_\alpha(x) - \nabla f(x))^T(x - \alpha G_\alpha(x) - z) \qquad \text{(from (9.17))} \\ &= f(z) + \tau\psi(z) + G_\alpha(x)^T(x - z) - \frac{\alpha}{2}\|G_\alpha(x)\|^2, \end{aligned}$$

where the last equality follows from cancellation of several terms in the previous line. Thus (b) is proved. □

Theorem 9.5. *Suppose that in problem* (9.10)*, ψ is a closed convex function and that f is is convex with Lipschitz continuous gradient on $\mathbb{R}^n$, with Lipschitz constant L. Suppose that* (9.10) *attains a minimizer x^* (not necessarily unique) with optimal objective value ϕ^*. Then if $\alpha_k = 1/L$ for all k in* (9.11)*, we have*

$$\phi(x^k) - \phi^* \leq \frac{L\|x^0 - x^*\|^2}{2k}, \quad k = 1, 2, \ldots .$$

Proof. Since $\alpha_k = 1/L$ satisfies the conditions of Lemma 9.4, we can use part (b) of this result to show that the sequence $\{\phi(x^k)\}$ is decreasing and that the distance to the optimum x^* also decreases at each iteration. Setting $x = z = x^k$ and $\alpha = \alpha_k$ in Lemma 9.4, and recalling (9.15), we have

$$\phi(x^{k+1}) = \phi(x^k - \alpha_k G_{\alpha_k}(x^k)) \leq \phi(x^k) - \frac{\alpha_k}{2}\|G_{\alpha_k}(x^k)\|^2,$$

justifying the first claim. For the second claim, we have by setting $x = x^k$, $\alpha = \alpha_k$, and $z = x^*$ in Lemma 9.4 that

$$\begin{aligned} 0 \leq \phi(x^{k+1}) - \phi^* &= \phi(x^k - \alpha_k G_{\alpha_k}(x^k)) - \phi^* \\ &\leq G_{\alpha_k}^T(x^k - x^*) - \frac{\alpha_k}{2}\|G_{\alpha_k}(x^k)\|^2 \\ &= \frac{1}{2\alpha_k}\left(\|x^k - x^*\|^2 - \|x^k - x^* - \alpha_k G_{\alpha_k}(x^k)\|^2\right) \\ &= \frac{1}{2\alpha_k}\left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right), \end{aligned} \tag{9.18}$$

from which $\|x^{k+1} - x^*\| \leq \|x^k - x^*\|$ follows.

By setting $\alpha_k = 1/L$ in (9.18), and summing over $k = 0, 1, 2, \ldots, K-1$, we obtain from a telescoping sum on the right-hand side that

$$\sum_{k=0}^{K-1}(\phi(x^{k+1}) - \phi^*) \leq \frac{L}{2}\left(\|x^0 - x^*\|^2 - \|x^K - x^*\|^2\right) \leq \frac{L}{2}\|x^0 - x^*\|^2.$$

By monotonicity of $\{\phi(x^k)\}$, we have

$$K(\phi(x^K) - \phi^*) \leq \sum_{k=0}^{K-1}(\phi(x^{k+1}) - \phi^*).$$

The result follows immediately by combining these last two expressions. □

9.4 Proximal Coordinate Descent for Structured Nonsmooth Functions

Coordinate descent methods and proximal gradient methods can be combined n a fairly straightforward way to componentwise regularized objectives of the form

$$\min_{x \in \mathbb{R}^n} h(x) := f(x) + \lambda \sum_{i=1}^{n} \Omega_i(x_i), \tag{9.19}$$

where f is convex as before and each regularization term $\Omega_i : \mathbb{R} \to \mathbb{R}$ is convex but possibly nonsmooth. Mirroring the proximal gradient method, in place of the step (6.2) along coordinate i_k, we obtain the next iteration by solving the following scalar subproblem:

$$\chi^k := \arg\min_{\chi} (\chi - x^k_{i_k})^T \nabla_{i_k} f(x^k) + \frac{1}{2\alpha_k} |\chi - x^k_{i_k}|^2 + \lambda \Omega_{i_k}(\chi). \tag{9.20}$$

Which we recognize as

$$x^{k+1}_i = \operatorname{prox}_{\alpha\lambda\Omega_{i_k}}(x^k_i - \alpha_k \nabla_{i_k} f(x^k)) \tag{9.21}$$

In this section we prove a result for the randomized CD method, which applies the step (9.20), (9.21) to a component i_k selected randomly and uniformly from $\{1, 2, \ldots, n\}$ at each iteration. We prove the result for the case of strongly convex f, using a simplified version of the analysis from [27]. It makes use of the following assumption.

Assumption 2. *The function f in (9.19) is uniformly Lipschitz continuously differentiable and strongly convex with modulus $\mu > 0$ (see (2.17)). The functions Ω_i, $i = 1, 2, \ldots, n$ are convex.*

Under this assumption, coercivity implies that h attains its minimum value h^* at a unique point x^*.

Our result uses the coordinate Lipschitz constant $L_{\max}$ for f, as defined in (6.5). Note that the modulus of convexity μ for f is also the modulus of convexity for h. By elementary results for convex functions, we have that

$$h(\alpha x + (1-\alpha)y) \leq \alpha h(x) + (1-\alpha)h(y) - \frac{1}{2}\mu\alpha(1-\alpha)\|x - y\|^2. \tag{9.22}$$

Theorem 9.6. *Suppose that Assumption 2 holds. Suppose that the indices i_k in (9.20) are chosen independently for each k with uniform probability from $\{1, 2, \ldots, n\}$, and that $\alpha_k \equiv 1/L_{\max}$. Then for all $k \geq 0$, we have*

$$E\left(h(x^k)\right) - h^* \leq \left(1 - \frac{\mu}{nL_{\max}}\right)^k (h(x^0) - h^*). \tag{9.23}$$

Proof. Define the function

$$H(x^k, z) := f(x^k) + \nabla f(x^k)^T (z - x^k) + \frac{1}{2} L_{\max} \|z - x^k\|^2 + \lambda \Omega(z),$$

and note that this function is separable in the components of z, and attains its minimum over z at the vector z^k whose i_k component is defined in (9.20). Note by strong convexity (2.17), we have that

$$\begin{aligned} H(x^k, z) &\leq f(z) - \frac{1}{2}\mu\|z - x^k\|^2 + \frac{1}{2} L_{\max}\|z - x^k\|^2 + \lambda\Omega(z) \\ &= h(z) + \frac{1}{2}(L_{\max} - \mu)\|z - x^k\|^2. \end{aligned} \tag{9.24}$$

We have by minimizing both sides over z in this expression that

$$\begin{aligned}
H(x^k, z^k) &= \min_z \; H(x^k, z) \\
&\leq \min_z \; h(z) + \frac{1}{2}(L_{\max} - \mu)\|z - x^k\|^2 \\
&\leq \min_{\alpha \in [0,1]} \; h(\alpha x^* + (1-\alpha)x^k) + \frac{1}{2}(L_{\max} - \mu)\alpha^2 \|x^k - x^*\|^2 \\
&\leq \min_{\alpha \in [0,1]} \; \alpha h^* + (1-\alpha)h(x^k) + \frac{1}{2}\left[(L_{\max} - \mu)\alpha^2 - \mu\alpha(1-\alpha)\right]\|x^k - x^*\|^2 \\
&\leq \frac{\mu}{L_{\max}} h^* + \left(1 - \frac{\mu}{L_{\max}}\right) h(x^k), \qquad (9.25)
\end{aligned}$$

where we used (9.24) for the first inequality, (9.22) for the third inequality, and the particular value $\alpha = \mu/L_{\max}$ for the fourth inequality (for which value the coefficient of $\|x^k - x^*\|^2$ vanishes). Taking the expected value of $h(x^{k+1})$ over the index i_k, we have

$$\begin{aligned}
E_{i_k} h(x^{k+1}) &= \frac{1}{n}\sum_{i=1}^n \left[f(x^k + (z_i^k - x_i^k)e_i) + \lambda\Omega_i(z_i^k) + \lambda \sum_{j \neq i} \Omega_j(x_j^k) \right] \\
&\leq \frac{1}{n}\sum_{i=1}^n \left\{ f(x^k) + [\nabla f(x^k)]_i (z_i^k - x_i^k) + \frac{1}{2}L_{\max}(z_i^k - x_i^k)^2 \right. \\
&\qquad \left. + \lambda\Omega_i(z_i^k) + \lambda \sum_{j \neq i} \Omega_j(x_j^k) \right\} \\
&= \frac{n-1}{n} h(x^k) + \frac{1}{n}\left[f(x^k) + \nabla f(x^k)^T (z^k - x^k) \right. \\
&\qquad \left. + \frac{1}{2}L_{\max}\|z^k - x^k\|^2 + \lambda\Omega(z^k) \right] \\
&= \frac{n-1}{n} h(x^k) + \frac{1}{n} H(x^k, z^k).
\end{aligned}$$

By subtracting h^* from both sides of this expression, and using (9.25) to substitute for $H(x^k, z^k)$, we obtain

$$E_{i_k} h(x^{k+1}) - h^* \leq \left(1 - \frac{\mu}{nL_{\max}}\right)(h(x^k) - h^*).$$

By taking expectations of both sides of this expression with respect to the random indices $i_0, i_1, i_2, \ldots, i_{k-1}$, we obtain

$$E(h(x^{k+1})) - h^* \leq \left(1 - \frac{\mu}{nL_{\max}}\right)(E(h(x^k)) - h^*).$$

The result follows from a recursive application of this formula. □

A result similar to (6.8) can be proved for the case in which f is convex but not strongly convex, but there are a few technical complications (see [27]).

9.5 Proximal Point Method

The proximal-point method (due to Rockafellar [30]) is a fundamental method for solving the problem

$$\min_{x\in\mathbb{R}^n} \psi(x), \tag{9.26}$$

where ψ is a convex function. The iterates are obtained from

$$x^{k+1} := \arg\min_z \psi(z) + \frac{1}{2\alpha_k}\|z - x^k\|^2 = \text{prox}_{\alpha_k\psi}(x^k), \tag{9.27}$$

where $\alpha_k > 0$ is a steplength parameter. Note that smoothness of ψ is not required. The problem (9.26) is a special case of (9.10) in which we set $f = 0$ and $\tau = 1$. We can thus state convergence results are corollaries of the results in Section 9.3.

The subproblem to be solved in (9.27) for the proximal-point method contains the original objective ψ, thus would appear to be as difficult to solve as the original problem. The quadratic regularization term in (9.27) plays an important stabilizing role. In important special cases (such as the augmented Lagrangian methods described in Chapter 11), its presence actually makes the solution of problem (9.27) much easier than the solution of the original problem (9.26).

Because there is no smooth part f in (9.26) (when we compare the objectives in (9.10) and (9.26)), there are no restrictions on the steplengths α_k. In a constant-steplength variant of (9.27), we can fix $\alpha_k \equiv \alpha$ for any $\alpha > 0$, and set $L = 1/\alpha$ in Theorem 9.5 to obtain the following convergence result.

Theorem 9.7. *Suppose that ψ is a closed convex function in* (9.26) *and that* (9.26) *attains a minimizer x^* (not necessarily unique) with optimal objective value ψ^*. Then if $\alpha_k = \alpha > 0$ for all k in* (9.27)*, we have*

$$\psi(x^k) - \psi^* \le \frac{\|x^0 - x^*\|^2}{2\alpha k}, \quad k = 1, 2, \ldots .$$

We observe again a sublinear $1/k$ rate of convergence, with a constant term depending inversely on α. The dependence on α makes intuitive sense. If α is chosen to be large, the quadratic regularization in (9.27) is mild, and the constant factor $\|x^0 - x^*\|^2/(2\alpha)$ in the convergence expression is small. (In the extreme case, as $\alpha \to \infty$, the effect of regularization vanishes, and the approach (9.27) almost converges in one step. This is not surprising, as (9.27) is close to the original problem (9.26) in this case.) When α is smaller, the quadratic regularization is more significant, the constant in the convergence experession is corresponding larger, so overall convergence is slower, when measured in terms of iterations. However, in the latter case, each subproblem may be easier to solve, as we may be able to use the approximate solution of one subproblem as a "warm start" for the following subproblem. Overall, the optimal choice of parameter α will depend very much on the structure of ψ.

Exercises

1. Let $\{\alpha_k\}_{k=1,2,\ldots}$ be a sequence of positive number such that $\alpha_k \downarrow 0$ but $\sum_{k=1}^T \alpha_k \uparrow \infty$ as $T \to \infty$. Show that
$$\frac{\sum_{j=1}^T \alpha_j^2}{\sum_{j=1}^T \alpha_j} \to 0, \quad \text{as } T \to \infty.$$

2. Consider the subgradient method with diminishing stepsize of the form $\alpha_k = \theta/k^p$ for some fixed value of p in the range $(0, 1)$. Using the techniques of Section 9.2, find a bound on $f(\bar{x}_T) - f * x_\star)$ that generalizes the bound (9.9) for the particular choice $p = 1/2$. Verify that $p = 1/2$ yields the tightest bound for $p \in (0, 1)$.