NOTES ON FIRST-ORDER METHODS FOR MINIMIZING NON-SMOOTH FUNCTIONS

1. Introduction. Consider a non-smooth optimization problem:

$$\underset{x\in\mathbf{R}^n}{\text{minimize}}\, f(x), \tag{1.1}$$

where $f : \mathbf{R}^n \to \mathbf{R}$ is not necessarily smooth. Before we delve into methods for minimizing such functions, we note that (1.1) is quite general. For example, the (generic) constrained optimization problem

$$\underset{x\in S}{\text{minimize}}\, f(x). \tag{1.2}$$

is equivalent to

$$\underset{x\in\mathbf{R}^n}{\text{minimize}}\, f(x) + I_S(x),$$

where I_S is the (convex) indicator function[1] of the feasible set S. By considering $f(x) + I_S(x)$ as a (non-smooth) objective function, we see that (1.2) is an instance of (1.1).

2. Subgradients. For now, we assume the function f is convex. Convexity is especially convenient because it suggests a natural generalization of gradients: *subgradients*.

DEFINITION 2.1. *The domain of a convex function $f : \mathbf{R}^n \to \mathbf{R}$ is*

$$\mathbf{dom}\, f := \{x \in \mathbf{R}^n : f(x) < \infty\}.$$

DEFINITION 2.2. *A point $g \in \mathbf{R}^n$ is a subgradient of a convex function $f : \mathbf{R}^n \to \mathbf{R}$ at a point $x \in \mathbf{dom}\, f$ if*

$$f(y) \geq f(x) + g^T(y - x) \text{ for any } y. \tag{2.1}$$

The set of all subgradients at x is called the subdifferential at $x : \partial f(x)$.

Subgradients are an especially convenient generalization of gradients to non-smooth, but convex functions. For example, they appear in the generalization of the usual zero-gradient optimality condition: a point $x^* \in \mathbf{R}^n$ is

[1]The convex or optimizer's indicator function is given by $I_S(x) = \begin{cases} 0 & x \in S \\ \infty & \text{otherwise} \end{cases}$.

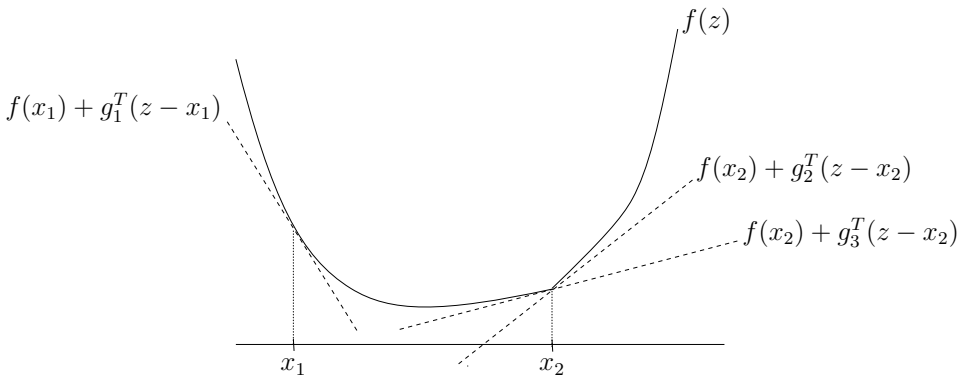

Fig 1: Boyd, Duchi and Vandenberghe (2015) At x_1, f is differentiable and $g_1 = \nabla f(x_1)$ is the unique subgradient at x_1. At x_2, f is not differentiable and has many subgradients.

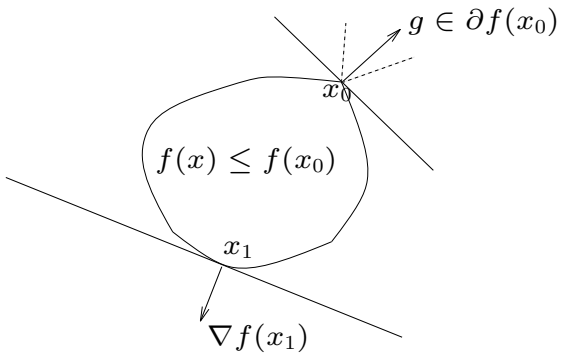

Fig 2: By rearranging (2.1), we obtain $g^T(y - x) \leq f(y) - f(x)$. For any y in the sublevel set $\{y \in \mathbf{R}^n \mid f(y) \leq f(x)\}$, we conclude $g^T(y - x) \leq 0$. At x_1, f is differentiable and $\nabla f(x_1)$ is the (outward) normal of the sublevel set at x_1. At x_0, f is not differentiable and any subgradient is a *supporting hyperplane* to the sublevel set at x_0.

an optimum of (1.1) if and only if $0 \in \partial f(x^*)$. Subgradients also (almost) always exist. There are pathological convex functions that are not subdifferentiable, but it is safe to assume the convex functions that we encounter are subdifferentiable.

For now, we defer an account of the mathematical properties of subgradients and focus on the practical issue of how to evaluate them. Subgradients obey many of the familiar calculus rules for gradients:

1. *non-negative homogeneity:* $\partial[\alpha f](x) = \alpha \partial f(x)$ for any $\alpha \geq 0$.
2. *linearity:* $\partial \left[\sum_{i=1}^n f_i\right](x) = \sum_{i=1}^n \partial f(x)$.
3. *chain rule:* $\partial \left[f \circ \mathcal{A}\right](x) = A^T \partial f(Ax+b)$ for any affine mapping $\mathcal{A}(x) = Ax + b$.

Subgradients also obey additional calculus rules. An especially important one is the *pointwise maximum rule:* a subgradient of $f(x) := \max_{i \in [m]} f_i(x)$ at x is given by any g that is a subgradient (at x) of an f_i that attains the maximum. This follows from

$$f(y) \geq f_i(y) \geq f(x) + g^T(y-x) = f(x) + g^T(y-x).$$

This property extends to the supremum of infinitely many functions when the supremum is attained.

3. Subgradient descent. The natural generalization of gradient descent to nonsmooth functions is *subgradient descent.* It starts at an initial point x_0 and generates successive iterates by

$$x_{k+1} \leftarrow x_k - \alpha_k g_k \text{ for some } g_k \in \partial f(x_k).$$

The step length α_k is either kept fixed or diminishes: $\alpha_k \to 0$, $\sum_k \alpha_k = \infty$. When f is differentiable, then $\partial f(x)$ is a singleton consisting of $\nabla f(x)$, and subgradient descent is gradient descent.

To study the convergence of subgradient descent, we focus on functions whose domain is $\mathbf{R}^n$. We begin by deriving a bound on the distance to the optimal set:

$$\begin{aligned} \|x_{k+1} - x^*\|_2^2 &= \|x_k - \alpha_k z_k - x^*\|_2^2 \\ &= \|x_k - x^*\|_2^2 - 2\alpha_k g_k^T(x_k - x^*) + \alpha_k^2 \|g_k\|_2^2 \\ &\leq \|x_k - x^*\|_2^2 - 2\alpha_k(f(x_k) - f^*) + \alpha_k^2 \|g_k\|_2^2, \end{aligned} \tag{3.1}$$

where the third inequality follows by Definition 2.2. By induction on k,

$$\|x_k - x^*\|_2^2 \leq \|x_0 - x^*\|_2^2 - 2\sum_{l=0}^{k-1} \alpha_l (f(x_l) - f^*) + \sum_{l=0}^{k-1} \alpha_l^2 \|g_l\|_2^2.$$

We rearrange to obtain

$$\begin{aligned} 2\sum_{l=0}^{k-1}\alpha_l(f(x_l)-f^*) &\le \|x_0-x^*\|_2^2-\|x_k-x^*\|_2^2+\sum_{l=0}^{k-1}\alpha_l^2\|g_l\|_2^2 \\ &\le \|x_0-x^*\|_2^2+\sum_{0=1}^{k-1}\alpha_l^2\|g_l\|_2^2. \end{aligned} \tag{3.2}$$

Plugging step sizes into (3.2) gives corresponding convergence rates. Let $f_k^* := \min_{l<k} f(x_l)$. For a fixed step size α,

$$2\alpha k\,(f_k^*-f^*) \le \|x_0-x^*\|_2^2+\alpha^2 k\|g_l\|_2^2.$$

If the subgradients are bounded, i.e. $\|g_l\|_2 \le L$, we obtain

$$f_k^*-f^* \le \frac{\|x_0-x^*\|_2^2}{2\alpha k}+\frac{\alpha L^2}{2}.$$

Since the bound does not decay to zero as k grows, subgradient descent with a fixed step size may not converge!

On the other hand, for diminishing step sizes: $\alpha_k \to 0$,

$$2(f_k^*-f^*)\sum_{l=0}^{k-1}\alpha_l \le \|x_0-x^*\|_2^2+\|g_l\|_2^2\sum_{l=0}^{k-1}\alpha_l^2.$$

Again, if the subgradients are bounded, we obtain

$$f_k^*-f^* \le \frac{\|x_0-x^*\|_2^2+L^2\sum_{l=0}^{k-1}\alpha_l^2}{2\sum_{l=0}^{k-1}\alpha_l}.$$

When $\alpha_k \to 0$ and $\sum_k \alpha_k = \infty$, it's possible to show

$$\frac{\sum_l \alpha_l^2}{\sum_l \alpha_l} \to 0.$$

Reassuringly, subgradient descent with diminishing step sizes converges. It's natural to ask what are the optimal step sizes when certain problem parameters are known. As we shall see, the crucial parameter is the initial distance to the optimal set $\|x_0-x^*\|_2$.

THEOREM 3.1. *When $f : \mathbf{R}^n \to \mathbf{R}$ is convex and its subgradients are bounded, i.e. $\|g\|_2 \le L$ for any $g \in \partial f(x)$ at any x, subgradient descent starting at x_0 such that $\|x_0-x^*\|_2 \le R$ with step sizes $\alpha_l \leftarrow \frac{R}{\sqrt{k}\|g_l\|_2}$ satisfy $f_k^*-f^* \le \frac{LR}{\sqrt{k}}$ after k iterations.*

PROOF. Let $\beta_l = \alpha_l \|g_l\|_2$. By (3.2), at the k-th iteration, we have

$$2(f_k^* - f^*) \sum_{l=0}^{k-1} \frac{\beta_l}{\|g_l\|_2} \leq \|x_0 - x^*\|_2^2 + \sum_{l=0}^{k-1} \beta_l^2 \leq R^2 + \sum_{l=0}^{k-1} \beta_l^2.$$

Since the subgradients are bounded,

$$\frac{2}{L}(f_k^* - f^*) \sum_{l=0}^{k-1} \beta_l \leq R^2 + \sum_{l=0}^{k-1} \beta_l^2.$$

We rearrange to obtain

$$f_k^* - f^* \leq \frac{R^2 + \sum_{l=1}^{k} \beta_l^2}{\frac{2}{L} \sum_{l=1}^{k} \beta_l}.$$

Since the bound is symmetric in $\{\beta_l\}$, the bound is minimized when all the β_l's are equal. For a given k, the bound is minimized at $\frac{R}{\sqrt{k}}$. The optimized bound is $f_k^* - f^* \leq \frac{LR}{\sqrt{k}}$. □

Thus, for a given number of iterations k, subgradient descent with a fixed step length attains $O(\frac{1}{\sqrt{k}})$-suboptimality after k iterations. In other words, subgradient descent is extremely slow: it obtain an ϵ-suboptimal point after at most $O\big(\frac{1}{\epsilon^2}\big)$ iterations. However, its simplicity and generality makes it well worth knowing.

3.1. *Polyak's step sizes.* Polyak (1987) proposes optimizing (3.1) in α_k to obtain the step size $\alpha_k \leftarrow \frac{f(x_k) - f^*}{\|g_k\|_2}$. Although Polyak's step size depends on the optimality gap, the method attains the $O\big(\frac{1}{\sqrt{k}}\big)$ convergence rate of subgradient descent with a fixed step length. Further, you may imagine f^* is rarely known in practice, but we shall see that's not the case.

THEOREM 3.2. *When $f : \mathbf{R}^n \to \mathbf{R}$ is convex and its subgradients are bounded, i.e. $\|g\|_2 \leq L$ for any $g \in \partial f(x)$ at any x, the iterates of subgradient descent with Polyak's step sizes starting at x_0 such that $\|x_0 - x^*\|_2 \leq R$ satisfy $f_k^* - f^* \leq \frac{LR}{\sqrt{k}}$ after k iterations.*

PROOF. The bound (3.1) is minimized w.r.t. α_k at $\frac{f(x_k) - f^*}{\|g_k\|_2}$. The optimized bound is

$$\|x_{k+1} - x^*\|_2^2 \leq \|x_k - x^*\|_2^2 - \frac{(f(x_k) - f^*)_2^2}{\|g_k\|_2^2}.$$

We rearrange to obtain

$$\frac{(f(x_k) - f^*)_2^2}{\|g_k\|_2^2} \leq \|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2.$$

Averaging the bounds for $l = 0, 1, \ldots, k-1$, we obtain

$$\frac{1}{k}\sum_{l=0}^{k-1} \frac{(f(x_l) - f^*)_2^2}{\|g_l\|_2^2} \leq \frac{1}{k}\left(\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2\right) \leq \frac{R^2}{k}. \tag{3.3}$$

Since the subgradients are bounded,

$$\begin{aligned} \frac{1}{k}\sum_{l=0}^{k-1} \frac{(f(x_l) - f^*)_2^2}{\|g_l\|_2^2} &\geq \frac{1}{L^2 k}\sum_{l=0}^{k-1} (f(x_l) - f^*)_2^2 \\ &\geq \frac{1}{L^2}\left(\frac{1}{k}\sum_{l=0}^{k-1} f(x_l) - f^*\right)^2 \\ &\geq \frac{1}{L^2}(f_k^* - f^*)^2. \end{aligned} \tag{3.4}$$

where the second inequality is by Jensen's inequality. We combine (3.3) and (3.4) to obtain

$$(f_k^* - f^*)^2 \leq \frac{L^2 R^2}{k}.$$

Taking the square root gives the stated result. □

To give a practical application of subgradient descent with Polyak's step sizes, consider the problem of finding a point in the (nonempty) intersection of convex sets $C_1, \ldots, C_m \subset \mathbf{R}^n$:

$$\underset{x}{\text{minimize}}\, f(x) := \max\{\text{dist}_{C_1}(x), \ldots, \text{dist}_{C_m}(x)\}, \tag{3.5}$$

where $\text{dist}_{C_j}(x) := \left\|x - P_{C_j}(x)\right\|_2$. By the calculus rule for pointwise maxima, a subgradient of f is given by

$$\nabla \text{dist}_{C_j}(x) = \frac{x - P_{C_j}(x)}{\text{dist}_{C_j}(x)},$$

where $j \in [m]$ is an index such that $\text{dist}_{C_j}(x) = f(x)$. In other words, the residual from the projection onto the farthest set is a subgradient.

As we shall see, subgradient descent on (3.5) with Polyak's step sizes is a form of *alternating projections.* Since the intersection is nonempty, we know

$f^* = 0$. At the k-th iteration, Polyak's step size is given by $\text{dist}_{C_j}(x)$. The subgradient descent step is

$$x_{k+1} \leftarrow x_k - \text{dist}_{C_j}(x)\frac{x_k - P_{C_j}(x_k)}{\text{dist}_{C_j}(x)} = P_{C_j}(x_k).$$

The iteration is very simple: at each iteration, project the current iterate onto the farthest set. When there are only two sets, the algorithm simply projects the iterate onto the other set. By Theorem 3.2, we know the algorithm has produced a point within distance $O\big(\frac{1}{\sqrt{k}}\big)$ of the intersection after at most k iterations.

Before moving on, we mentioned that subgradient descent is also "optimal" in the sense that Nesterov's 1983 method is optimal. That is, on convex functions with bounded subgradients, among methods that chooses iterates in

$$\mathcal{K}_k[f] = x_0 + \text{span}\{g_0, \ldots, g_{k-1}\} \text{ for } k = 1, 2, \ldots \tag{3.6}$$

To that subgradient descent is optimal, we exhibit a function and initial guess such that any method that chooses iterates in (3.6) has at best $O(\frac{1}{\sqrt{k}})$ convergence. We refer to Nesterov (2004), Theorem 3.2.1 for the details.

4. Smoothing. The basic approach of smoothing is to replace a non-smooth f with a smooth approximation f_μ (parametrized by μ) and minimize f_μ by a method for smooth optimization. Since first-order methods for minimizing smooth functions converge (much) faster than subgradient descent, we hope that minimizing f_μ gives an ϵ-suboptimal point in fewer iterations than subgradient descent on f.

For a concrete example, consider the robust regression problem

$$\underset{x \in \mathbf{R}^n}{\text{minimize}}\, f(x) := \sum_{i=1}^{n} \left|a_i^T x - b_i\right|. \tag{4.1}$$

We replace the absolute value function with the Huber function:

$$h_\mu(x) = \begin{cases} \frac{x^2}{2\mu} & |x| \le \mu \\ |x| - \frac{\mu}{2} & |x| \ge \mu \end{cases}.$$

The parameter μ trades off accuracy and smoothness. It is possible to show

$$h_\mu(x) \le |x| \le h_\mu(x) + \frac{\mu}{2}.$$

Thus smaller values of μ give a more accurate approximation to the absolute value function. On the other hand, it is also possible to show h_μ is strongly smooth with constant $\frac{1}{\mu}$. Since first-order method usually converge fast on more strongly smooth functions, larger values of μ give problems that can be solved more efficiently.

More precisely, we have

$$f(x) - f^* \leq f_\mu(x) - f^* + \frac{\mu n}{2} \leq f_\mu(x) - f_\mu^* + \frac{\mu n}{2} \text{ for any } x.$$

Thus, we must obtain an $\epsilon - \frac{\mu n}{2}$-suboptimal solution to the smoothed problem to ensure it is an ϵ-suboptimal solution to the original problem. If we solve the smoothed problem with Nesterov's 1983 method, we must take at most $O((\mu(\epsilon - \frac{\mu n}{2}))^{-\frac{1}{2}})$ iterations. By choosing $\mu = \frac{\epsilon}{n}$, we see that it takes at most $O(\frac{1}{\epsilon})$ iterations to obtain an ϵ-suboptimal solution to the original problem. By comparison, subgradient descent on the original problem takes at most $O(\frac{1}{\epsilon^2})$ iterations to obtain a comparably suboptimal point.

4.1. *Smoothing by the convex conjugate.* We have seen that smoothing allows us to minimize non-smooth function more efficiently than subgradient descent. The tricky part is forming a good smooth approximation. As we shall see, the conjugate representation of a convex function suggests a principled way to form a smooth approximation.

DEFINITION 4.1. *For any function $f : \mathbf{R}^n \to \mathbf{R}$, its convex conjugate is*

$$f^*(y) := \sup_{x \in \mathbf{dom}\, f} x^T y - f(x).$$

Since f^* is the supremum of affine functions, it is always convex (even if f is not convex). As the name conjugate suggests, the conjugate of the conjugate, or *biconjugate*, of a convex function is the function itself. Thus convex functions occur in pairs (f, f^*). Again, there are pathological examples that do not agree with their biconjugates, but it is safe to assume that the convex functions that we encounter agree with their biconjugates. The conjugate representation of a convex function f is

$$f(x) = [f^*](x) = \sup_{y \in \mathbf{dom}\, f^*} x^T y - f^*(y). \tag{4.2}$$

A consequence of the definition of the convex conjugate is *Fenchel's inequality.* It generalizes Young's inequality: $\frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|y\|_2^2 \geq x^T y$.

LEMMA 4.2. *For any convex function $f : \mathbf{R}^n \to \mathbf{R}$ and any points x, y,*

$$f(x) + f^*(y) \geq x^T y.$$

Further, equality holds if and only if $y \in \partial f(x)$.

PROOF. To show the inequality, we plug x into the definition of f^* :

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} x^T y - f(x) \geq x^T y - f(x) \text{ for any } x \in \mathbf{dom}\, f.$$

Rearranging gives the inequality. To show equality holds when $y \in \partial f(x)$, we rearrange the first-order convexity condition to obtain

$$(x')^T y - f(x') \leq x^T y - f(x) \text{ for any } x'.$$

Thus $\sup_{x' \in \mathbf{dom}\, f} (x')^T y - f(x)$ is attained at x and

$$f^*(y) = x^T y - f(x).$$

To show $f(x) + f^*(y) = x^T y$ implies $y \in \partial f(x)$, we rearrange the equality to obtain $f^*(y) = x^T y - f(x)$. By the definition of $f^*(y)$,

$$x^T y - f(x) \geq (x')^T y - f(x') \text{ for any } x' \in \mathbf{dom}\, f.$$

Rearranging gives $f(x') \geq f(x) + y^T(x' - x)$, which implies $y \in \partial f(x)$. □

Before moving on, we describe some calculus rules for taking conjugates:

1. *scalar multiple:* $[\alpha f]^*(y) = \alpha f^*(\frac{y}{\alpha})$,
2. *sum:* $[f + g]^*(y) = \inf_z f(z) + g(y - z)$.

We are especially interested in the convex conjugates of strongly convex functions because they are continuously differentiable and strongly smooth. Both are desirable properties of a smooth approximation.

LEMMA 4.3. *For any strongly convex function $f : \mathbf{R}^n \to \mathbf{R}$ with strong convexity constant μ, its convex conjugate f^* is*

1. *continuously differentiable*
2. *strongly smooth with constant $\frac{1}{\mu}$.*

PROOF. By the pointwise maximum rule, a subgradient of f^* at y is

$$\arg\max_{x \in \mathbf{dom}\, f} x^T y - f(x).$$

Since f is strongly convex, the arg max is unique for any y. Thus the subgradient is unique, and we conclude f^* is differentiable.

To show f^* is strongly smooth, we show ∇f^* is Lipschitz continuous. By the pointwise maximum rule,

$$\nabla f^*(y_1) = \arg\max_{x \in \mathbf{dom}\, f} x^T y_1 - f(x) \text{ and}$$
$$\nabla f^*(y_2) = \arg\max_{x \in \mathbf{dom}\, f} x^T y_2 - f(x).$$

Thus

$$f^*(y_1) = \nabla f^*(y_1)^T y_1 - f(\nabla f^*(y_1)),$$
$$f^*(y_2) = \nabla f^*(y_2)^T y_2 - f(\nabla f^*(y_2)).$$

Rearranging, we see that Fenchel's inequality holds with equality:

$$f(\nabla f^*(y_1)) + f^*(y_1) = \nabla f^*(y_1)^T y_1,$$
$$f(\nabla f^*(y_1)) + f^*(y_2) = \nabla f^*(y_2)^T y_2.$$

By Lemma 4.2, we know that $y_1 \in \partial f(\nabla f^*(y_1))$, $y_2 \in \partial f(\nabla f^*(y_2))$. Since f is strongly convex, we have

$$\begin{aligned} f(\nabla f^*(y_2)) &\geq f(\nabla f^*(y_1)) + y_1^T(\nabla f^*(y_2) - \nabla f^*(y_1)) \\ &\quad + \frac{\mu}{2}\|\nabla f^*(y_1) - \nabla f^*(y_2)\|_2^2 \\ f(\nabla f^*(y_1)) &\geq f(\nabla f^*(y_2)) + y_2^T(\nabla f^*(y_1) - \nabla f^*(y_2)) \\ &\quad + \frac{\mu}{2}\|\nabla f^*(y_1) - \nabla f^*(y_2)\|_2^2. \end{aligned}$$

We add the inequalities and rearrange to obtain

$$(y_1 - y_2)^T(\nabla f^*(y_1) - \nabla f^*(y_2)) \geq \mu\|\nabla f^*(y_1) - \nabla f^*(y_2)\|_2^2.$$

We apply the Cauchy-Schwartz inequality to conclude ∇f^* is Lipschitz continuous with constant $\frac{1}{\mu}$. □

Lemma 4.3 suggests a principled approach to forming a smooth approximation of a convex function. Recall a convex function f, has a conjugate representation in terms of f^* :

$$f(x) = [f^*]^*(x) = \sup_{y \in \mathbf{dom}\, f^*} x^T y - f^*(y).$$

Consider the conjugate of the function $f^*(y) + \frac{\mu}{2}\|y\|_2^2$:

$$f_\mu = \left[f^* + \frac{\mu}{2}\|\cdot\|_2^2\right]^*(x) = \sup_y x^T y - f^*(y) - \frac{\mu}{2}\|y\|_2^2.$$

Since $f^* + \frac{\mu}{2}\|\cdot\|_2^2$ is strongly convex with constant μ, by Lemma 4.3, f_μ is continuously differentiable and strongly smooth with constant $\frac{1}{\mu}$. For example, the conjugate representation of the absolute value function is

$$|x| = \sup_y xy - I_{[-1,1]}(x),$$

where $I_{[-,1,1]}$ is the (convex) indicator function of the interval $[-1,1]$. It's smooth approximation by adding $\frac{\mu}{2}\|\cdot\|_2^2$ is the Huber function:

$$f_\mu = \sup_y xy - I_{[-1,1]}(y) - \frac{\mu}{2}\|y\|_2^2 = h_\mu(x)$$

More generally, we can smooth f by adding any strongly convex function d to f^* and taking the conjugate:

$$f_\mu = [f^* + \mu d]^*(x) = \sup_{y \in \mathbf{dom}\, f^*} x^T y - f^*(y) - \mu d(y) \tag{4.3}$$

To wrap up, we study the efficiency of minimizing a non-smooth function by smoothing. To ensure a good solution to the smoothed problem is also a good solution to the original problem, the smoothed function must be an accurate approximation of the original function. As we shall see, a class of functions for which smoothing by conjugation gives accurate approximations are functions whose conjugate has bounded domain.

LEMMA 4.4. *For a convex function $f : \mathbf{R}^n \to \mathbf{R}$ whose conjugate f^* has bounded domain, its smooth approximation f_μ given by (4.3) sandwiches f :*

$$f_\mu(x) \leq f(x) \leq f_\mu(x) + \mu B,$$

where $B := \sup_{y \in \mathbf{dom}\, f^*} d(y)$.

PROOF. The lower bounds follows from

$$\begin{aligned} f_\mu(x) &= \sup_{y \in \mathbf{dom}\, f^*} x^T y - f^*(y) - \mu d(y) \\ &\leq \sup_{y \in \mathbf{dom}\, f^*} x^T y - f^*(y) = f(x). \end{aligned}$$

The upper bound follows from

$$\begin{aligned} f_\mu(x) &= \sup_{y \in \mathbf{dom}\, f^*} x^T y - f^*(y) - \mu d(y) \\ &\geq \sup_{y \in \mathbf{dom}\, f^*} x^T y - f^*(y) - \mu B \\ &= f(x) - \mu B. \end{aligned}$$

Rearranging gives the upper bound. □

By Lemma 4.4,

$$f(x) - f^* \leq f_\mu(x) - f^* + \mu B \leq f_\mu(x) - f_\mu^* + \mu B \text{ for any } x.$$

Thus, to obtain an ϵ-suboptimal solution to the original problem, we must obtain an $\epsilon - \mu B$-suboptimal solution to the smoothed problem. By setting

$\mu = \frac{\epsilon}{2B}$, we see that it takes at most $O(\frac{1}{\epsilon})$ iterations for accelerated gradient descent to obtain an ϵ-suboptimal solution to the original problem. By comparison, subgradient descent (on the original problem) takes $O(\frac{1}{\epsilon^2})$ iterations to obtain a comparably suboptimal point. In practice, the efficiency of smoothing can be further improved by decreasing μ gradually.

The gain in efficiency is possibly offset by the cost of evaluating the gradient of the smoothed function. In general, $\nabla f_\mu(x)$ is the solution to

$$\arg\max_{y \in \mathbf{dom}\, f^*} x^T y - f^*(y) - \frac{\mu}{2}\|y\|_2^2 .$$

Thus evaluating ∇f_μ may be more costly than obtaining subgradients of f. However, when f_μ is given in closed form, smoothing is often a good idea.

4.2. *The proximal point method.* When the strongly convex function $\frac{1}{2}\|\cdot\|_2^2$ is added to the conjugate representation of a function f, the resulting smooth approximation is known as the *Moreau-Yosida envelope* of f :

$$f_\mu(x) = \left[f^* + \frac{\mu}{2}\|\cdot\|_2^2\right]^*(x) = \inf_{y \in \mathbf{dom}\, f} f(y) + \frac{1}{2\mu}\|x-y\|_2^2 . \tag{4.4}$$

The second equality follows from the fact that the conjugate of a sum is the *infimal convolution* of the conjugates:

$$[f+g]^*(x) = \inf_{y \in \mathbf{dom}\, f^*} f^*(y) + g^*(x-y).$$

Thus, minimizing f_μ is an *exact* reformulation of minimizing f : the minimum of f_μ is also the minimum of f. Thus it takes an accelerated gradient method at most $O(\frac{1}{\sqrt{\epsilon}})$ iterations to obtain an $\epsilon-$suboptimal solution

By the pointwise maximum rule, the gradient of the Moreau envelope is

$$\begin{aligned} \nabla f_\mu(x) &= \arg\max_{y \in \mathbf{dom}\, f^*} x^T y - f^*(y) - \frac{\mu}{2}\|y\|_2^2 \\ &= \arg\min_{y \in \mathbf{dom}\, f^*} f^*(y) + \frac{\mu}{2}\Big\|\frac{x}{\mu} - y\Big\|_2^2. \end{aligned} \tag{4.5}$$

This is our first encounter with the *proximal mapping*. It is the basic step in a class of optimization methods called *proximal methods*.

Definition 4.5. *For any convex function f, its proximal mapping is*

$$\mathrm{prox}_f(x) := \arg\min_{y \in \mathbf{dom}\, f} f(x) + \frac{1}{2}\|x-y\|_2^2 \tag{4.6}$$

Proximal mappings generalize projections onto convex sets. In particular, the proximal mapping of the indicator function of a convex set is the projection onto the set. The proximal mapping of a function and its conjugate also induce *Moreau's decomposition* of any point in $\mathbf{R}^n$. Moreau's decomposition generalizes the decomposition of a point into its projection onto a subspace and the orthocomplement.

LEMMA 4.6. *For any point $x \in \mathbf{R}^n$ and convex function f,*

$$x = \operatorname{prox}_f(x) + \operatorname{prox}_{f^*}(x).$$

PROOF. The optimality condition of (4.6) says

$$x - \operatorname{prox}_f(x) \in \partial f(\operatorname{prox}_f(x)).$$

By Lemma 4.2, Fenchel's inequality holds with equality, and, by Lemma 4.2 again, we have

$$\operatorname{prox}_f(x) \in \partial f^*(x - \operatorname{prox}_f(x))$$

or, equivalently,

$$x - (x - \operatorname{prox}_f(x)) \in \partial f^*(x - \operatorname{prox}_f(x)).$$

We recognize $x - \operatorname{prox}_f(x)$ satisfies the optimality condition of the optimization problem defining $\operatorname{prox}_{f^*}(x)$ to deduce $\operatorname{prox}_{f^*}(x) = x - \operatorname{prox}_f(x)$. □

COROLLARY 4.7. *For any point $x \in \mathbf{R}^n$ and convex function f,*

$$x = \operatorname{prox}_{\alpha f}(x) + \frac{1}{\alpha} \operatorname{prox}_{\alpha f^*}(\alpha x),$$

PROOF. We apply the Moreau decomposition to μf

$$x = \operatorname{prox}_{\alpha f}(x) + \operatorname{prox}_{[\alpha f]^*}(x) = \operatorname{prox}_{\alpha f}(x) + \frac{1}{\alpha} \operatorname{prox}_{\alpha f^*}(\alpha x),$$

where the second equality follows from the fact that $[\alpha f]^*(y) = \alpha f^*(\frac{x}{\alpha})$. □

The Moreau decomposition gives us a very elegant way to express gradient descent on f_μ. By Corollary 4.7, a gradient step is

$$x_k - \mu \nabla f_\mu(x_k) = x_k - \mu \operatorname{prox}_{\mu^{-1} f^*}\left(\tfrac{x}{\mu}\right) = \operatorname{prox}_{\mu f}(x_k), \tag{4.7}$$

The parameter μ is a step size. We recognize (4.7) as the proximal point method by Rockafellar (1976). It is the grandfather of all *proximal methods*,

many of which are widely used today. We refer to Parikh and Boyd (2013) for an account of proximal methods.

Before exploring the lineage of the proximal point method, we study its convergence rate. By the argument in Section 4.1, we know smoothing takes at most $O(\frac{1}{\epsilon^2})$ iterations to obtain an ϵ-suboptimal point. As it turns out, since minimizing f_μ is an exact reformulation of minimizing f, it only takes $O(\frac{1}{\epsilon})$ iterations to obtain an ϵ-suboptimal point.

THEOREM 4.8. *When $f : \mathbf{R}^n \to \mathbf{R}$ is convex, the proximal point method satisfies*

$$f(x_k) - f^* \leq \frac{1}{2k\mu}\|x_0 - x^*\|_2^2.$$

PROOF. First, by the optimality of x_{k+1} for the optimization probem defining $\text{prox}_{\mu f}(x_k)$, we know

$$f(x_{k+1}) + \frac{1}{2\mu}\|x_{k+1} - x_k\|_2^2 \leq f(x_k).$$

Rearranging, we deduce the proximal point method is a descent method. Again, by the optimality of x_{k+1}, we know $\frac{1}{\mu}(x_k - x_{k+1}) \in \partial f(x_{k+1})$. Since f is convex,

$$\begin{aligned} f(x_k) - f^* &\leq \frac{1}{\mu}(x_{k+1} - x_k)^T(x^* - x_{k+1}) \\ &\leq \frac{1}{\mu}\left((x^* - x_k)^T(x^* - x_{k+1}) - \|x_{k+1} - x^*\|_2^2\right) \\ &\leq \frac{1}{2\mu}\left(\|x_k - x^*\|_2^2 + \|x^* - x_{k+1}\|_2^2\right) - \frac{1}{\mu}\|x_{k+1} - x^*\|_2^2 \\ &= \frac{1}{2\mu}\left(\|x^* - x_k\|_2^2 - \|x_{k+1} - x^*\|\right), \end{aligned}$$

where the third inequality follows by Young's inequality. Summing over iterations, we have

$$\begin{aligned} \frac{1}{k}\sum_{l=1}^{k} f(x_k) - f^* &\leq \frac{1}{2k\mu}\left(\|x_0 - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2\right) \\ &\leq \frac{1}{2k\mu}\|x_0 - x^*\|_2^2 \end{aligned}$$

We recall $\{f(x_k)\}$ is decreasing to obtain the stated result. □

The interpretation of the proximal point method as gradient descent on the Moreau envelope suggests it is possible to accelerate the proximal point method. The resulting algorithm was proposed by Güler (1992). It takes $O(\frac{1}{\sqrt{\epsilon}})$ iterations to achieve an ϵ-suboptimal point.

Although it inspired many method that are widely used, the proximal point method is more of a "conceptual" method: useful for reasoning about methods, but not widely used in practice. As we shall see, the *augmented Lagrangian method* for linearly constrained optimization is an instance of the proximal point method.

Consider the linearly constrained problem

$$\begin{array}{ll} \underset{x\in\mathbf{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & Ax = b : y \end{array} \tag{4.8}$$

The Lagrangian is

$$L(x, y) := f(x) + y^T(Ax - b),$$

and the dual problem is

$$\underset{y\in\mathbf{R}^m}{\text{maximize}} -f^*(-A^Ty) - b^Ty,$$

which is equivalent to

$$\underset{y\in\mathbf{R}^m}{\text{minimize}}\, f^*(-A^Ty) + b^Ty.$$

The *augmented Lagrangian* with penalty paramter $\rho > 0$ is

$$L(x, y, \rho) := f(x) + y^T(Ax - b) + \frac{\rho}{2}\|Ax - b\|_2^2. \tag{4.9}$$

The augmented Lagrangian method solves (4.8) by

$$x_{k+1} \leftarrow \arg\min_x L(x, y_k, \rho), \tag{4.10}$$

$$y_{k+1} \leftarrow y_k - \rho(Ax_{k+1} - b). \tag{4.11}$$

We defer the details of the augmented Lagrangian method and focus on its interpretation as an instance of the proximal point method. It turns out that (4.11) is equivalent to a proximal point step on the dual problem.

LEMMA 4.9. *The proximal point of* $g(y) := f^*(-A^Ty) + b^Ty$ *at* y *is*

$$y_+ = y + \rho(Ax_\rho^*(y) - b), \tag{4.12}$$

where $x_\rho^*(y) := \arg\min_x L(x, y, \rho)$.

PROOF. By Definition 4.5, the proximal mapping of $g(y)$ is given by

$$\arg\min_z f^*(-A^T z) + b^T z + \frac{1}{2\rho}\|y - z\|_2^2. \tag{4.13}$$

We know y_+ satisfies the optimality conditions of (4.13):

$$b + \frac{1}{\rho}(y_+ - y) \in A\partial f^*(-A^T y_+). \tag{4.14}$$

To show (4.12), we check that $y + \rho A(x_\rho^*(y) - b)$ satisfies (4.14). The point $x_\rho^*(y)$ satisfies the optimality conditions of (4.9):

$$-A^T(y - \rho(Ax_\rho^*(y) - b)) \in \partial f(x_\rho^*(y)).$$

By Lemma 4.2, Fenchel's inequality holds with equality and

$$x_\rho^*(y) \in \partial f^*(-A^T(y - \rho(Ax_\rho^*(y) - b))).$$

Multipying by A,

$$Ax_\rho^*(y) \in A\partial f^*(-A^T(y - \rho(Ax_\rho^*(y) - b))).$$

We plug in $Ax_\rho^*(y) = \frac{1}{\rho}(y_+ - y + \rho b)$ to draw the desired conclusion. □

REFERENCES

BOYD, S., DUCHI, J. and VANDENBERGHE, L. (2015). Subgradients. EE 364b Lecture Notes.

GÜLER, O. (1992). New proximal point algorithms for convex minimization. *SIAM Journal on Optimization* **2** 649–664.

NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization* **87**. Springer Science & Business Media.

PARIKH, N. and BOYD, S. (2013). Proximal algorithms. *Foundations and Trends in Optimization* **1** 123–231.

POLYAK, B. T. (1987). *Introduction to Optimization*. Optimization Software New York.

ROCKAFELLAR, R. T. (1976). Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization* **14** 877–898.

YUEKAI SUN
STANFORD, CALIFORNIA
APRIL 29, 2015
E-MAIL: yuekai@stanford.edu