University of California, Berkeley Lecturer: Benjamin Recht

CS227C/STAT260 Convex Optimization and Approximation: Optimization for Modern Data Analysis

Feburary 26, 2015 Notes: Benjamin Recht

Lecture 11: The Subgradient and Proximal Point Methods

1 Subgradient Descent

Though we have seen examples where certain elements of the subdifferential are *ascent* directions, there aways exists a descent direction in the subdifferential. Indeed, it is the element with smallest norm. Let g_{mn} be the minimum norm element of the subgradient:

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

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

Proposition 1 *Either $0 \in \partial f(x)$ or $-g_{\mathrm{mn}}$ is a descent direction.*

Proof Note that in order to be the minimum norm subgradient,

$$\langle g_{\mathrm{mn}}, \hat{g} - g_{\mathrm{mn}} \rangle \geq 0 \tag{1}$$

for all $\hat{g} \in \partial f(x)$. To see this, suppose otherwise. Then

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

for all $t \in [0,1]$, and

$$\frac{d}{dt}\|g_{\mathrm{mn}} + t(\hat{g} - g_{\mathrm{mn}})\|^2 \Big|_0 = 2\langle g_{\mathrm{mn}}, \hat{g} - g_{\mathrm{mn}} \rangle < 0$$

which contradicts that g_{mn} has minimum norm.

Using (1), we have

$$\langle \hat{g}, g_{\mathrm{mn}} \rangle \geq \|g_{\mathrm{mn}}\|_2^2$$

for all $\hat{g} \in \partial f(x)$.

Now we have

$$\begin{aligned} f'(x; -g_{\mathrm{mn}}) &= \sup_{g \in \partial f(x)} \langle -g_{\mathrm{mn}}, g \rangle \\ &= - \inf_{g \in \partial f(x)} \langle g_{\mathrm{mn}}, g \rangle \\ &\leq -\|g_{\mathrm{mn}}\|_2^2 \end{aligned}$$

proving that $-g_{\mathrm{mn}}$ is a descent direction whenever it is nonzero. ■

This suggests a natural algorithm for minimizing convex, nonsmooth functions. Compute the minimum norm element of the subdifferential and follow it down hill. There is only one problem with this approach: computing the minimum norm element might be prohibitively expensive. In the next section, we show that a very naive algorithm that simply follows arbitrary subgradients will actually converge.

2 The subgradient method

The subgradient method is rather simple: at each step k, we choose an 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 be taking ridiculous steps, it turns out that the average of the iterates

$$\bar{x}_k = \left(\sum_{j=1}^{k} \alpha_k \right)^{-1} \sum_{j=1}^{k} \alpha_k x_k$$

converges to a minimizer of f.

The proof of this method is *nearly identical* to the proof of the converge of the stochastic gradient method for convex functions with bounded stochastic gradients. We simply assume that for all x,

$$\|g\|_2 \leq G$$

for all $g \in \partial f(x)$. Note that this assumption implies that f must be Lipschitz with constant G (why?).

As we did with the stochastic gradient method. α_k be our stepsize sequence. Define

$$\lambda_k = \frac{1}{\sum_{j=0}^{k} \alpha_j} \,.$$

This is the sum of all the stepsizes up to iteration k. Also define

$$\bar{x}_k = \lambda_k^{-1} \sum_{j=0}^{k} \alpha_j x_j \,.$$

$\bar{x}_k$ is the average of the iterates weighted by the stepsize. We are going to analyze the deviation of $f(\bar{x}_k)$ from optimality.

Let $D_0 = \|x_0 - x_\star\|^2$. D_0 is the initial distance to an optimal solution.

To proceed, we just expand the distance to an optimal solution:

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

Now we have

$$f\left(\bar{x}_T\right) - f(x_\star) \leq \lambda_T^{-1} \sum_{t=1}^{T} \alpha_t (f(x_t) - f(x_\star)) \tag{3a}$$

$$\leq \lambda_T^{-1} \sum_{t=1}^{T} \alpha_t \langle g_t, x_t - x_\star \rangle \tag{3b}$$

$$\leq \lambda_T^{-1} \sum_{t=1}^{T} \tfrac{1}{2} \left(\|x_t - x_\star\|^2 - \|x_{t+1} - x_\star\|^2 \right) + \tfrac{1}{2} \alpha_t^2 G^2 \tag{3c}$$

$$= \frac{\|x_0 - x_\star\|^2 - \|x_0 - x_T\|^2 + G^2 \sum_{t=0}^{T} \alpha_t^2}{2\lambda_T} \tag{3d}$$

$$\leq \frac{D_0^2 + G^2 \sum_{t=0}^{T} \alpha_t^2}{2 \sum_{t=1}^{T} \alpha_t} \tag{3e}$$

Here, (3a) follows because f is convex, (3b) is the definition of a subgradient, and (3c) uses (2).

Note that we also immediately have the bound

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

so our analysis works for both the average iterate and the *best* iterate.

2.1 Stepsizes

Let's dive into different possibilities for our stepsize.

Constant step size. First, we can just pick $\alpha_t = \alpha$. In this case, we get

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

The choice of $\alpha = \frac{\theta D_0}{G\sqrt{T}}$ results in the convergence rate

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

Constant step length. An alternative choice is to choose $\alpha_k = \frac{\alpha}{\|g_k\|}$. This ensures that the *length* of each step is a constant. In this case, we have the bound

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

With $\alpha = \frac{\theta D_0}{\sqrt{T}}$, we get the (worst-case) convergence rate

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

This is exactly the same as what we had before! But note that here we only had to estimate the distance to optimality, not the maximal norm of the gradient, to get a decent rate of convergence.

2.2 Diminishing Step Size

Diminishing stepsizes are attractive as one doesn't have to choose a final T, and we don't have to estimate any quantities about f.

Note that for any sequence α_k such that α_k tends to zero, but $\sum_k \alpha_k$ diverges, then

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

This is particularly easy to see if

$$\sum_k \alpha_k^2 = M < \infty$$

In this case,

$$f(\bar{x}_T) - f_\star \leq \frac{D_0^2 + G^2 \sum_{t=0}^T \alpha_t^2}{2\sum_{t=1}^T \alpha_t} \leq \frac{D_0^2 + G^2 M}{2\sum_{t=1}^T \alpha_t}$$

and the left-hand side clearly tends to zero.

To see that this works for general diminishing stepsizes, one only needs to prove that

$$\frac{\sum_{k=1}^\infty \alpha_k^2}{\sum_{k=1}^\infty \alpha_k} \to 0$$

whenever the sum of α_k diverges but α_k tends to zero. We close this section by deriving more quantitative bounds for an explicit stepsize choice. Suppose we set $\alpha_k = \frac{\theta}{\sqrt{k}}$. Then we have

$$f(\bar{x}_T) - f_\star \leq \frac{D_0^2 + G^2\theta^2 \sum_{t=0}^T t^{-1}}{2\theta \sum_{t=1}^T t^{-1/2}} \leq \frac{D_0^2 + G^2\theta^2 \log(T)}{2\theta\sqrt{T}}\,.$$

Note that this bound tends to zero at a rate of $\log(T)/\sqrt{T}$. This is slower than the $T^{-1/2}$ rate of a constant stepsize, but we are guaranteed asymptotic convergence to zero.

Note that any other choice of $\alpha_k \propto k^{-p}$ for $p \in (0,1)$ would have yielded a worse convergence rate.

3 The Proximal Point Method

Let P be an extended-real-valued convex function on $\mathbb{R}^n$. Define the operator

$$\mathrm{prox}_P(x) = \arg\min_y \tfrac{1}{2}\|x-y\|_2^2 + P(y) \tag{4}$$

Since the optimized function is strongly convex, it must have a unique optimal solution. Therefore, we can conclude that $\mathrm{prox}_P(x)$ is a well-defined mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$. By the first order optimality conditions, we conclude that $\mathrm{prox}_P(x)$ is the unique point satisfying

$$x - \mathrm{prox}_P(x) \in \partial P(\mathrm{prox}_P(x))\,. \tag{5}$$

The definition of prox_P also reveals that it is well-defined for all $x \in \mathbb{R}^n$, and maps onto the set $\mathrm{dom}(P) := \{z \in \mathbb{R}^n \,:\, P(z) < \infty\}$. The mapping prox_P is called the *proximity operator* or *proximal point mapping* associated with P.

Let's look at some examples.

1. If $\mathbb{I}_C$ is an indicator function for a convex set C

$$\mathbb{I}_C(x) = \begin{cases} 0 & x \in C \\ \infty & \text{otherwise} \end{cases} \tag{6}$$

then $\text{prox}_{\mathbb{I}_C}$ is the Euclidean projection onto C. That is, $\text{prox}_{\mathbb{I}_C}(x)$ is the closest point in the set C to x in Euclidean distance.

2. For $\mathbb{I}_{\mathbb{R}_+}$, this proximity mapping takes on the trivial form:

$$\mathbb{I}_{\mathbb{R}_+}(x)_i = \max(x_i, 0) \tag{7}$$

3. For $P(x) = \frac{\mu}{2}\|x\|_2^2$, $\text{prox}_P(x) = \frac{1}{1+\mu}x$. That is, $\text{prox}_P(x)$ is equal to a multiple of x, shrunk towards the origin.

4. For $P(x) = \mu\|x\|_1$,

$$\text{prox}_P(x)_i = \begin{cases} x_i + \mu & x_i < -\mu \\ 0 & -\mu \le x_i \le \mu \\ x_i - \mu & x_i > \mu \end{cases} \tag{8}$$

This function is called the *shrinkage* operator and has many applications in signal processing. To see that this is the correct form, one needs only to analyze the optimality conditions of the one dimensional problem

$$\text{minimize } \tfrac{1}{2}(x-y)^2 + \mu|y| \tag{9}$$

3.1 The Proximal Point Algorithm

Proximity operators have many algorithmic applications. As a warm up, consider the following simple iteration: pick $x_0 \in \mathbb{R}^n$ and $\alpha > 0$ and define the iteration $x_{k+1} = \text{prox}_{\alpha P}(x_k)$. This simple iteration can be shown to converge to a minimizer of the function P. To prove this, we need the following two lemmas. The first is a simple consequence of the convexity of P.

Lemma 2 *Let P be convex on X. Let x, $y \in X$, and let $g_y \in \partial P(y)$ and $g_x \in \partial P(y)$. Then $\langle g_x - g_y, x - y\rangle \ge 0$.*

Proof By the definition of the subdifferential, we have

$$\begin{aligned} P(x) - P(y) &\ge \langle g_y, x - y\rangle \\ P(y) - P(x) &\ge \langle g_x, y - x\rangle \end{aligned} \tag{10}$$

Adding these two equations gives $-\langle g_x - g_y, x - y\rangle \le 0$. ■

The second lemma uses this key inequality to establish several facts about the proximity operator. This lemma is proven in [**?**].

Lemma 3 *Let $Q_\alpha(x) := x - \text{prox}_{\alpha P}(x)$. Then we have*

1. $\alpha^{-1}Q_\alpha(x) \in \partial P(\text{prox}_{\alpha P}(x))$

2. $\langle \operatorname{prox}_{\alpha P}(x) - \operatorname{prox}_{\alpha P}(z), Q_\alpha(x) - Q_\alpha(z)\rangle \geq 0$

3. $\| \operatorname{prox}_{\alpha P}(x) - \operatorname{prox}_{\alpha P}(z)\|^2 + \|Q_\alpha(x) - Q_\alpha(z)\|^2 \leq \|x - z\|^2$

4. $\|x - z\| = \| \operatorname{prox}_{\alpha P}(x) - \operatorname{prox}_{\alpha P}(z)\|$ *if and only if* $x - z = \operatorname{prox}_{\alpha P}(x) - \operatorname{prox} \alpha P(z)$

Proof The first assertion follows from the definitions. The second assertion follows from (i), and Lemma 2. The third assertion follows from (ii) after expanding the identity

$$\|x - z\|^2 = \|[\operatorname{prox}_{\alpha P}(x) - \operatorname{prox}_{\alpha P}(z)] + [Q_\alpha(x) - Q_\alpha(z)]\|^2 .$$

(iv) follows immediately from (iii). ■

By Lemma 3 (3), we have

$$\| \operatorname{prox}_{\alpha P}(x) - \operatorname{prox}_{\alpha P}(z)\|^2 \leq \|x - z\|^2 \tag{11}$$

and we say that the proximity operator is *nonexpansive*. This is the essential property needed to prove the convergence of the proximal point method. That the proximity operator is nonexpansive also plays a role in the projected gradient algorithm, analyzed below. As a particular consequence, we have the following

Corollary 4 *Let C be a closed convex set. Let $\Pi_C(x)$ denote the closest point to x in the set C in the Euclidean distance. Then for all $x, y \in \mathbb{R}^n$, $\|\Pi_C(x) - \Pi_C(y)\| \leq \|x - y\|$.*

Proof Since $\Pi_C = \operatorname{prox}_{\mathbb{I}_C}$, this follows immediately from (11). ■

Using the nonexpansive property of the proximity operator, we can now verify the convergence of the proximal point method. Since $\operatorname{prox}_{\alpha P}$ is non-expansive, $\{z_k\}$ lies in a compact set and must have a limit point $\bar{z}$. Also for any $z_\star$ with $0 \in \partial P(z_\star)$,

$$\|z_{k+1} - z_\star\| = \| \operatorname{prox}_{\alpha P}(z_k) - \operatorname{prox}_{\alpha P}(z_\star)\| \leq \|z_k - z_\star\| \tag{12}$$

which means that the sequence $\|z_k - z_\star\|$ is monotonically non-increasing. Therefore

$$\lim_{k\to\infty} \|z_k - z_\star\| = \|\bar{z} - z_\star\| . \tag{13}$$

where $\bar{z}$ is any limit point of z_k. By continuity we have $\operatorname{prox}_{\alpha P}(\bar{z})$ is also a limit point of z_k. Therefore, we must have

$$\| \operatorname{prox}_{\alpha P}(\bar{z}) - \operatorname{prox}_{\alpha P}(z_\star)\| = \| \operatorname{prox}_{\alpha P}(\bar{z}) - z_\star\| = \|\bar{z} - z_\star\| \tag{14}$$

But this means that $\operatorname{prox}_{\alpha P}(\bar{z}) - \operatorname{prox}_{\alpha P}(z_\star) = \bar{z} - z_\star$, and in turn that $\operatorname{prox}_{\alpha P}(\bar{z}) = \bar{z}$ and $0 \in \partial P(\bar{z})$. Now using $\bar{z}$ for $z_\star$ in (13) shows that

$$\lim_{k\to\infty} \|z_k - \bar{z}\| = 0 \tag{15}$$

In other words, the sequence z_k converges to $\bar{z}$.

4 The proximal gradient algorithm

The projected gradient algorithm combines a proximal step with a gradient step. This lets us solve a variety of constrained optimization problems with simple constraints, and it lets us solve some non-smooth problems at linear rates.

We will aim to analyze a function h which admits a decomposition

$$h(x) = f(x) + P(x) \tag{16}$$

where f is smooth and P is a convex extended real valued function. Let us assume that ∇f is Lipschitz so that $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.

Let us define a projected gradient scheme to solve this problem. Let $\alpha_0, \ldots, \alpha_T, \ldots$, be a sequence of positive step sizes. Choose $x_0 \in X$, and iterate

$$x_{k+1} = \text{prox}_{\alpha_k P}(x_k - \alpha_k \nabla f(x_k)) \,. \tag{17}$$

The algorithm alternates between taking gradient steps and then taking proximal point steps.

The key idea behind this algorithm is summed up by the following proposition

Proposition 5 *Let f be differentiable and convex and let P be convex. $x_\star$ is an optimal solution of*

$$\text{minimize}_x \; f(x) + P(x) \tag{18}$$

if and only if $x_\star = \text{prox}_{\alpha P}(x_\star - \alpha \nabla f(x_\star))$ for all $\alpha > 0$.

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

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

which is equivalent to $x_\star = \text{prox}_{\alpha P}(x_\star - \alpha \nabla f(x_\star))$. ■

For non-convex f, we see that a fixed point of the projected gradient iteration is a stationary point of h. We first analyze the convergence of this projected gradient method for arbitrary smooth f, and then focus on strongly convex f.

To summarize, we have proven that proximal steps have all of the properties of projections. Hence, our analyses of projected gradient immediately apply to the proximal point method. Moreover, now that we have the tools of subgradient calculus, we can prove a more precise convergence rate for weakly convex functions.

4.1 New convergence proof for weakly convex functions.

The proof is borrowed from Lieven Vandenberghe's course notes[1]. Define the function:

$$Q_\alpha(x) = \tfrac{1}{\alpha}(x - \text{prox}_{\alpha P}(x))$$

Note that we have $Q_\alpha(x) \in \partial P(\text{prox}_{\alpha P})$.

[1] `http://www.seas.ucla.edu/~vandenbe/236C/lectures/ppm.pdf`

By the first order convexity condition

$$\begin{aligned} P(\text{prox}_{\alpha P}(x)) &\leq P(z) + Q_\alpha(x)^T(\text{prox}_{\alpha P}(x) - z) \\ &= P(z) + Q_\alpha(x)^T(x - z) - \alpha\|Q_\alpha(x)\|^2 \end{aligned} \tag{19}$$

Using (19) with $z = x$, we have

$$P(\text{prox}_{\alpha P}(x)) \leq P(x) - \alpha\|Q_\alpha(x)\|^2$$

implying that the proximity operator reduces the value of the function P. That is, the proximal point method is a descent method and the function values $P(x_k)$ form a decreasing sequence.

Using (19) with $z = x_\star$ proves that

$$\begin{aligned} P(\text{prox}_{\alpha P}(x)) - P_\star &\leq Q_\alpha^T(x - x_\star) - \alpha\|Q_\alpha(x)\|^2 \\ &= \frac{1}{2\alpha}\left(\|x - x_\star\|^2 - \|\,\text{prox}_{\alpha P}(x) - x_\star\|^2\right) - \frac{\alpha}{2}\|Q_\alpha(x)\|^2 \end{aligned}$$

In terms of the iterates of the algorithm, this means

$$P(x_{k+1}) - P_\star \leq \frac{1}{2\alpha}\left(\|x_k - x_\star\|^2 - \|x_{k+1} - x_\star\|^2\right) - \frac{1}{2\alpha}\|x_k - x_{k+1}\|^2 \tag{20}$$

Summing (20) over all iterates yields

$$\sum_{i=1}^{k} P(x_i) - P_\star + \frac{1}{2\alpha}\sum_{i=1}^{k}\|x_i - x_{i+1}\|^2 \leq \frac{1}{2\alpha}\|x_0 - x_\star\|^2$$

Since the function values form a non-increasing sequence this means

$$P(x_k) - P_\star \leq \frac{1}{2\alpha k}\|x_0 - x_\star\|^2 .$$

Note that in the previous argument, we were only able to show the iterates themselves converge. This proves that the function values converge at a rate that is asymptotically faster than the subgradient method. Also note that picking $\alpha = \infty$ would be ideal. But this is obvious! This is equivalent to minimizing P directly.