10-725: Optimization **Fall 2012**

Lecture 7: September 18

Lecturer: Ryan Tibshirani *Scribes: David Bamman, William Chan and Yanchuan Sim*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

7.1 Convergence analysis of subgradient descent

7.1.1 Class Questions

If there is more than one subgradient, which do you choose? It's pretty rare (and difficult) to be able to compute all subgradients and in many applications, we are only able to compute one. If there is more than one subgradient, e.g., $|x|, x \in [-1, +1]$), an option is to take a random value between $[-1, +1]$. The theory we have, however, holds for *all* subgradients.

Backtracking for subgradients? The convergence of subgradients uses fixed stepsize rules. Step size choices are not usually adaptive. Ryan: "people don't use adaptive choices for subgradients as far as I know".

7.1.2 Quick Review

A subgradient of convex function $f : \mathrm{R}^n \to \mathrm{R}$ at x is any $g \in \mathrm{R}^n$ such that

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

Subgradients characterize optimality:

$$f(x^*) = \min_{x \in \mathrm{R}^n} f(x) \Leftrightarrow 0 \in \partial f(x*)$$

Note: this is true even if f is non-convex, but often we can't compute subgradients for non-convex (so moot point).

Step size choices: we either take fixed step sizes or we take constants such that:

$$\sum_{k=1}^{\infty} t_k^2 \leq \infty$$

$$\sum_{k=1}^{\infty} t_k = \infty$$

For fixed step sizes, we have

$$\lim_{k \to \infty} f(x_{best}^{(k)}) \leq f(x^*) + \frac{t}{2} G^2$$

Note: We are guaranteed to be $\frac{t}{2}G^2$ away from the optimal point if we run the method to infinity.

For diminishing step sizes, we have:

$$\lim_{k\to\infty} f(x_{best}^{(k)}) = f(x^*)$$

Note: We can guarantee that we reach the optimal point (after possibly an infinite amount of steps).

7.1.3 Convergence Rate

The question is, after k iterations, what is the error $f(x_{best}^{(k)}) - f(x^*)$?

Consider taking $t_i = \frac{R}{G\sqrt{k}}$, all $i = 1 \dots k$. The basic bound is:

$$\frac{R^2 + G^2 \sum_{i=1}^k t_i^2}{2\sum_{i=1}^k} = \frac{RG}{\sqrt{k}}$$

This means the subgradient method has convergence rate $O\left(\frac{1}{\sqrt{k}}\right)$; to get $f(x_{best}^{(k)}) - f(x^*) \leq \epsilon$, we need $O\left(\frac{1}{\sqrt{\epsilon^2}}\right)$ iterations.

This is actually the best we can do; e.g., we can't get do better than $O(\frac{1}{\sqrt{k}})$.

7.2 Subgradients and Alternating Projections

Using the problem of finding a point in the intersection of convex sets as an example, we derive the alternating projections algorithm using subgradients.

7.2.1 Intersection of sets

Problem: Given m closed convex sets $C_1, C_2, \dots, C_m$, we want to find $x^* \in \bigcap_i^m C_i$.

First, we define

$$f(x) = \max_{i=1,\dots,m} \text{dist}(x, C_i)$$

where

$$\text{dist}(x, C) = \min_{u\in C} \|x - u\|$$

is the closest we can get to x if we have to stay in the set C.

Also,

$$f(x^*) = 0 \iff x^* \in \bigcap_i^m C_i$$

Therefore, the optimization problem is to minimize

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

which, when equal to 0 is the point we are looking for.

Since C is closed and convex, there is a unique point $u^* = P_C(x)$. This unique point is the projection of x onto C, and it minimizes $\|x - u\|$ over $u \in C$. We can thus write

$$\mathrm{dist}(x, C) = \|x - P_C(x)\|$$

7.2.2 Finding subgradient of f_i

We want to calculate the subgradient of f because if we can do so, we can apply subgradient methods and obtain an algorithm to solve our problem.

First, we consider $f_i(x)$ of C_i. It turns out that $f_i(x)$ is differentiable (but not discussed in class). For each i, if we take a point not in C_i, i.e $x \notin C_i$ and $\|x - P_C(x)\| \neq 0$, it turns out that

$$\frac{x - P_C(x)}{\|x - P_C(x)\|} \tag{7.1}$$

is a subgradient of $f_i(x)$. We obtain this by just taking the projected point and finding the gradient without the chain rule.

Now we are going to show that (7.1) is a subgradient of f_i at x. By the definition of the projected point $u = P_C(x)$,

$$(x - u)^T (y - u) \leq 0$$

for all $y \in C$. To find u, we are actually minimizing,

$$\min_{u \in C} \frac{1}{2} \|x - u\|^2$$

We can remove the constraints by using the indicator function,

$$\min \frac{1}{2} \|x - u\|^2 + I_C(x)$$

Differentiating that and setting it to 0, we get

$$\begin{aligned} & 0 \in (u - x) + \mathcal{N}_C(u) \\ \Rightarrow & x - u \in \mathcal{N}_C(u) \\ \Rightarrow & (x - u)^T u \geq (x - u)^T y \quad \forall y \in C \\ \Rightarrow & (x - u)^T (y - u) \leq 0 \end{aligned} \tag{7.2}$$

Equation (7.2) follows from the properties of the normal cone.

We can say that C is contained in the halfspace

$$C \subseteq H = \{y : (x - u)^T (y - u) \leq 0\}$$

and we claim that

$$\mathrm{dist}(y, C) \geq \frac{(x - u)^T (y - u)}{\|x - u\|} \quad \forall y$$

When $y \in H$, the RHS ≤ 0 by definition of the halfspace. When $y \notin H$,

$$\begin{aligned}
\text{dist}(y, H) &= \|y - u\| \sin \phi \\
&= \frac{\|x - u\| \|y - u\| \sin \phi}{\|x - u\|} \\
&= \frac{(x - u)^T (y - u)}{\|x - u\|} \\
&\leq \text{dist}(y, C)
\end{aligned}$$

because C is contained in H.

Now we have proved this inequality, we can rewrite it as

$$\begin{aligned}
\text{dist}(y, C) &\geq \frac{(x - u)^T (y - x + x - u)}{\|x - u\|} \\
&= \|x - u\| + \frac{(x - u)^T}{\|x - u\|}(y - x)
\end{aligned}$$

Hence, the term $\frac{(x-u)^T}{\|x-u\|}$, which is what we had in (7.1), is a subgradient of $\text{dist}(x, C)$.

7.2.3 Finding subgradients of f

Using a rule we learnt from earlier on in the course, if

$$f(x) = \max_{i=1,\ldots,m} f_i(x)$$

then,

$$\partial f(x) = \text{conv}\left(\bigcup_{j : f_j(x) = f(x)} \partial f_j(x) \right)$$

What this means is that the subgradient of $f(x)$ is equal to the convex hull of the union of all maximal $f_j(x)$'s, and take the respective subdifferentials.

If $f_i(x) = f(x) \neq 0$ (when it is 0, we are done), then

$$\frac{x - P_C(x)}{\|x - P_C(x)\|} \in \partial f(x)$$

This gives us a prescription for finding the subgradients.

7.2.4 Subgradient descent

We will use a particular stepsize, known as the Polyak stepsize, because this particular choice will give us a famous algorithm that is a special case of the subgradient method. For the purpose of illustration, the Polyak stepsize is

$$t_k = f(x^{(k-1)})$$

and the subgradient descent update rule is

$$\begin{aligned} x^{(k)} &= x^{(k-1)} - t_k \partial f(x^{(k-1)}) \\ &= x^{(k-1)} - f(x^{(k-1)}) \frac{x - P_{C_i}(x)}{\|x - P_{C_i}(x)\|} \text{ where } x^{(k-1)} \text{ is farthest from } C_i \\ &= x^{(k-1)} - x^{(k-1)} + P_{C_i}(x) \\ &= P_{C_i}(x) \end{aligned}$$

So the update rule is just to take $x^{(k-1)}$ and project it to the set it is farthest from.

This is also known as the alternating projections algorithm. By using the subgradient method, we can now use what we know about subgradients to say things about the *alternating projections* algorithm (such as convergence rate and guarantees, etc).

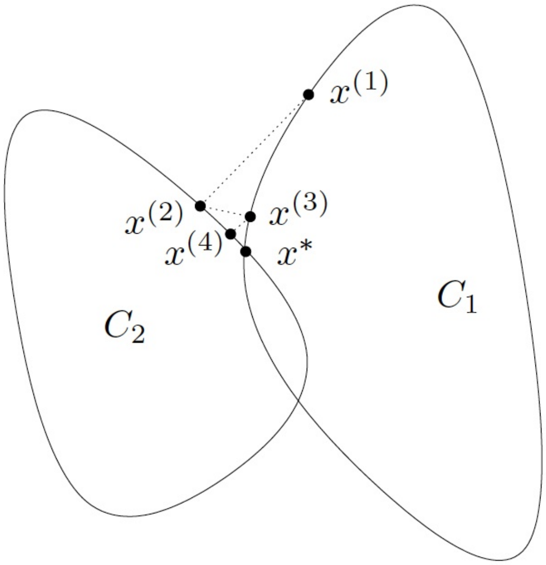

Figure 7.1: Alternating Projection from Boyd Notes

7.3 Nesterov's Theorem

Theorem 7.1 *Nesterov's Theorem: For any $k \leq n-1$ and starting point $x^{(0)}$, there is a function in the problem class such that any nonsmooth first-order method satisfies*

$$f(x^{(k)}) - f(x^*) \geq \frac{RG}{2(1+\sqrt{k+1})}$$

Proof: Let $k = n-1$ and $x^{(0)} = 0$.

$$f(x) = \max_{i=1\ldots n} x_i + \frac{1}{2}||x||^2$$

The optimal x^* here $= (-1/n, \ldots, -1/n)$, with the optimal function value $f(x^*) = -\frac{1}{2n}$. If $R = \frac{1}{\sqrt{n}}$, then f is Lipschitz with $G = 1 + \frac{1}{\sqrt{n}}$.

Claim: At any iteration i from 1 to n, all of the elements of x from x_{i+1} to x_n are 0. To show this, let us assume we have some oracle that gives us $g = e_j + x$, where j is the smallest index of x's maximum value, x_j. e_j is the basis vector:

$$\begin{aligned} e_1 &= (1, 0, 0, 0) \\ e_2 &= (0, 1, 0, 0) \\ \dots \end{aligned}$$

At iteration 1, $g = e_1 + x$ (since x has only one nonzero value, located at element 1); From this we can see that $\text{span}\{g^{(0)}, g^{(1)}\} \subseteq \text{span}\{e_1, e_2\}$. In general

$$\begin{aligned} \text{span}\{g^{(0)}, g^{(1)}\} &\subseteq \text{span}\{e_1, e_2\} \\ \text{span}\{g^{(0)}, g^{(1)}, g^{(2)}\} &\subseteq \text{span}\{e_1, e_2, e_3\} \\ \text{span}\{g^{(0)}, g^{(1)}, g^{(2)}, g^{(3)}\} &\subseteq \text{span}\{e_1, e_2, e_3, e_4\} \end{aligned}$$

Therefore, $f(x^{n-1}) \geq 0$. Since we know that the optimal function value is $-\frac{1}{2n}$

$$f(x^{n-1}) - f(x^*) \geq \frac{1}{2n} = \frac{RG}{2(1 + \sqrt{n})}$$

■

References

[Nesterov04] Y. Nesterov (2004), *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers.