EE 381V: Large Scale Optimization **Fall 2012**

Lecture 19 — November 6

Lecturer: Caramanis & Sanghavi *Scribe: Brian Bates & Ying Chen*

19.1 Topics Covered

- Recap of Key Concepts from the Previous Lecture
- Implications of Subgradients
- Convergence Analysis

In the last lecture, we introduced Legendre-Fenchel transform and the concept of subdifferential and subgradients. By doing this, we were approaching the basis of a new descent method for the optimization problem of a nonsmooth function. In this lecture, we will make further discussion on properties of subgradients, and then develop the subgradient method and complete the convergence analysis.

19.2 Recap of Previous Lecture

Definition 1. *$f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, where the right-hand-side is a set of extended real numbers. Then the Legendre-Fenchel transform is defined as*

$$f^\star(x^\star) = \sup_x \{\langle x^\star, x\rangle - f(x)\}. \tag{19.1}$$

As can be seen, if $h(x)$ is an affine function corresponding to a non-vertical supporting hyperplane, such that $h(x) \leq f(x)$ for $\forall x$, and $F^\star$ is the collection of all $h \subseteq \mathbb{R}^{n+1}$, then $F^\star$ is the epigraph of the convex function $f^\star$.

Theorem 19.1. *If f is closed and convex, $f^{\star\star} = f$.*

This is true since f is closed and convex if and only if epi$f$ is convex and closed.

According to the definition of Legendre-Fenchel transform, we know that

$$\langle x^\star, x\rangle \geq f^\star(x^\star) + f(x) \tag{19.2}$$

always holds. We can then define the concept of subdifferential and subgradients.

Definition 2. *For a convex function f, $x \in \partial f(x)$ if*

$$\langle x^\star, x\rangle = f^\star(x^\star) + f(x). \tag{19.3}$$

Recall that for a differentiable function f, f is convex if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \text{ for } \forall x, y. \tag{19.4}$$

Thus we can have another definition of subgradients.

Definition 3. *For a convex function f, the subdifferential of f at x is a set, $\partial f(x)$, of vectors, called subgradients, and $g \in \partial f(x)$ if*

$$f(y) \geq f(x) + \langle g, y - x \rangle \textit{ for } \forall y. \tag{19.5}$$

Hence $\partial f(x) = \{\nabla f(x)\}$ if f is a smooth (differentiable) function.

Lemma 19.2. *For an unconstrained optimization problem min : $f(x)$, $x^\star$ is optimal if and only if $0 \in \partial f(x^\star)$, which is $0 = \nabla f(x^\star)$ for a smooth function.*

Lemma 19.3. *For a constrained optimization problem min : $f(x)$, s.t. $x \in X$, $x^\star$ is optimal if and only if $0 \in \partial f(x^\star) + N_X(x^\star)$.*

Proof: Denote an indicator function

$$I_X(x) = \begin{cases} 0, & x \in X \\ +\infty, & \text{otherwise.} \end{cases} \tag{19.6}$$

Then the original constrained optimization problem can be rewritten as an unconstrained minimization problem *min* : $f(x) + I_X(x)$. It has been shown in the homework that

- $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$ and
- $\partial I_X(x) = N_X(x)$.

Substituting into lemma (19.2) we complete the proof. □

19.3 Implications of Subgradients

As we saw in the last lecture, we cannot simply substitute a subgradient in for a gradient in a descent algorithm. There are a few properties that play an important role in convergence.

19.3.1 Key properties subgradients do not have

- Subgradients may not define descent directions, which means $f(x - \epsilon \nabla f(x)) < f(x)$ may not hold. This was shown at the end of the last lecture.
- It may not be true that $\nabla f(x) \to 0$ as $x \to x^\star$. For example, the subgradient of $\|x\|_1$ is $+1$ or -1 arbitrarily close to the optimal solution.

19.3.2 Key properties of subgradients

Theorem 19.4. *Monotonicity.* $\langle u - v, x - y\rangle \geq 0$, *where* $u \in \partial f(x)$ *and* $v \in \partial f(y)$.

Corollary 19.5. $\langle g, x - x^\star\rangle \geq 0$ *for* $\forall g \in \partial f(x)$.

Proof: Take v in the previous theorem to be 0 and we will get the result. □

Now we are ready for one of the most important properties of subgradients, from which the subgradient descent method is constructed. The notation we will use in the theorem is as follows.

$$x_+ = x - h \cdot g, \tag{19.7}$$

where h is the step size and $g \in \partial f(x)$. We want to show although it may happen that $f(x_+) > f(x)$, it is always true that the distance to sublevel sets decreases for sufficiently small step size.

Theorem 19.6. *Let f be a convex function. If $f(x) > f(z)$, $g \in \partial f(x)$, and $x_+ = x - h \cdot g$, where h is small enough, then $\|x_+ - z\| < \|x - z\|$ holds for $z = x^\star$.*

Proof:

$$\begin{aligned}\|x - hg - z\|_2^2 &= \|x - z\|_2^2 - 2h\langle g, x - z\rangle + h^2\|g\|_2^2 &(19.8)\\ &\leq \|x - z\|_2^2 - 2h\left(f\left((x) - f(z)\right) + h^2\|g\|_2^2\right. &(19.9)\end{aligned}$$

The inequality is true because of the under-estimating property of subgradients (the second definition). With $f(x) > f(z)$ and h small enough, we can conclude

$$\|x - hg - z\|_2^2 < \|x - z\|_2^2.$$

□

This theorem is actually true for all z, but we will not show that in this lecture.

19.4 Convergence Analysis

Now we are ready to analyse the convergence of subgradient descent. To make things easier, assume f is G-Lipschitz. In other words, $\|g\| \leq G$, where $g \in \partial f(x)$ (Note that the same convergence can be obtained without this condition, but we will not show that here). From the subgradient descent algorithm

$$x_{k+1} = x_k - h_k g_k,$$

where $g_k \in \partial f(x_k)$. Therefore we have

$$\begin{aligned}
\|x_{k+1} - x^*\|_2^2 &= \|x_k - x^* - h_k g_k\|_2^2 && (19.10)\\
&= \|x_k - x^*\|_2^2 - 2h_k\langle g_k, x_k - x^*\rangle + h_k^2\|g_k\|_2^2 && (19.11)\\
&\le \|x_k - x^*\|_2^2 - 2h_k(f(x_k) - f^*) + h_k^2 G^2 && (19.12)\\
&\le \|x_0 - x^*\|_2^2 - 2\sum_{i=0}^{k} h_i(f(x_i) - f^*) + G^2\sum_{i=0}^{k} h_i^2 && (19.13)\\
&\le \|x_0 - x^*\|_2^2 - 2(f(x_{\text{best}}) - f^*)\sum_{i=0}^{k} h_i + G^2\sum_{i=0}^{k} h_i^2. && (19.14)
\end{aligned}$$

The last inequality follows if we let

$$x_{\text{best}} = \text{argmin}_{0\le i\le k} f(x_i). \tag{19.15}$$

Denoting R as the initial error we obtain

$$0 \le R^2 - 2(f(x_{\text{best}}) - f^*)\sum_{i=0}^{k} h_i + G^2\sum_{i=0}^{k} h_i^2. \tag{19.16}$$

Thus, after k steps, the optimality gap is bounded by

$$f(x_{\text{best}}) - f^* \le \frac{R^2 + G^2\sum_i h_i^2}{2\sum_i h_i}. \tag{19.17}$$

What should we use for the step lengths $\{h_i\}$? Some options are to use:

1. a fixed size $h_i = h$.
2. a decreasing non-summable sequence. In other words, choose $\{h_i\}$ such that $\sum_i h_i^2 < \infty$ and $\sum_i h_i = \infty$.
3. a sequence that minimizes the upperbound by $N + 1$.

(Exercise) Show that the problem of finding an h that minimizes the upperbound $F(h_0, ..., h_n) = \frac{R^2+G^2\sum_i h_i^2}{\sum_i h_i}$ is convex over $\mathbb{R}^{N+1} \ge 0$ and that the solution is $h_i = \frac{R/G}{\sqrt{N+1}}$.

Proof: The problem

$$\min_h \frac{R^2 + G^2\|h\|_2^2}{\|h\|_1} \tag{19.18}$$

can be split into

$$\min_l \min_{l=\|h\|_1} \frac{R^2 + G^2\|h\|_2^2}{l}, \tag{19.19}$$

which is convex. Additionally, since h is symmetric we have that the average,

$$h_i = h^* = \frac{R/G}{\sqrt{N+1}}, \tag{19.20}$$

is the optimal solution. □

This means that the error after N steps is $\epsilon \leq \frac{RG}{\sqrt{N+1}}$. In summary, after N steps, $\epsilon = O\left(\frac{1}{N}\right)$. Equivalently, if we want to bound the error below ϵ, we need at least $N = O\left(\frac{1}{\epsilon^2}\right)$ steps. Note that this is slow compared to a linear rate $O\left(\log(\frac{1}{2})\right)$ steps. This can, however, be useful in very large scale problems where we have a fair tolerance for error (say within 5%) and each iteration does not require much computation.

Theorem 19.7. *The subgradient method is "optimal" uniformly in the dimension.*

What do we mean by optimal? If we use an oracle (blackbox) to obtain $f(x)$ and $g \in \partial f(x)$ from x, then by optimal, we mean in the number of calls to the oracle. The oracle model allows us to obtain lower bounds.

Unless we know additional information about the structure of the problem, we cannot get a better convergence than subgradient descent. Consider the following cases:

1. If f smooth and strongly convex, with condition number K, gradient descent yields:

$$\left(\frac{K-1}{K+1}\right)^{2N} \|x_0 - x^*\|^2,$$

while the optimal convergence is $O\left(\left(\frac{\sqrt{K}-1}{\sqrt{K}+1}\right)^{2N}\right)$.

2. If f smooth and convex, gradient descent yields:

$$\frac{1}{N}\|x_0 - x_*\|^2,$$

while the optimal convergence is $O\left(\frac{1}{N^2}\right)$

3. If f convex subgradient is optimal:

$$\frac{1}{\sqrt{N}}\|x_0 - x_*\|^2,$$

which is the optimal.

If we had additional information about the structure of a problem, we can potentially do better than subgradient descent. For example, recall the problem where we wish to minimize

$$f(x) = \|Ax - b\|_2^2 + \lambda\|x\|_1.$$

For this problem we know that we can obtain better convergence than $O\left(\frac{1}{\sqrt{N}}\right)$.

19.4.1 Projected Subgradient

The projected subgradient algorithm gives us another look at constrained optimization:

$$\min f(x)$$

$$\text{subject to } x \in X,$$

where f and X are convex.

In projected subgradient method, our update step is

$$x_+ = \text{Proj}_X(x - hg) = \Pi_X(x - hg) \tag{19.21}$$

Lemma 19.8. *For $x \in X$, and x_0 potentially not in X,*

$$\begin{aligned} \|x - \Pi_X(x_0)\|_2^2 + \|\Pi_X(x_0) - x_0\|_2^2 &\leq \|x - x_0\|_2^2 & (19.22) \\ \Rightarrow \|\Pi_X(x_0) - x\|_2^2 &\leq \|x_0 - x\|_2^2. & (19.23) \end{aligned}$$

In other words, projecting can only get us closer.

Proof: The proof immediately follows from the fact that $\|\Pi_X(x_0) - x_0\|_2^2$ is non-negative. □

Thus,

$$\begin{aligned} \|x_{k+1} - x^*\|_2^2 &= \|\Pi_X(x_k - h_k g_k) - x^*\|_2^2 & (19.24) \\ &\leq \|x_k - h_k g_k - x^*\|_2^2 & (19.25) \end{aligned}$$

again converges $O\left(\frac{1}{\sqrt{N}}\right)$ or $O\left(\frac{1}{\epsilon^2}\right)$.

Theorem 19.9. *If f is convex with L-Lipschitz gradient and $x, y \in \mathbb{R}^n$, the following are all equivalent:*

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|x - y\|_2^2 \tag{19.26}$$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \tag{19.27}$$

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \tag{19.28}$$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L\|x - y\|_2^2 \tag{19.29}$$

$$\alpha f(x) + (1-\alpha) f(y) \geq f(\alpha x + (1-\alpha) y) + \frac{\alpha(1-\alpha)}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \tag{19.30}$$

$$\alpha f(x) + (1-\alpha) f(y) \leq f(\alpha x + (1-\alpha) y) + \alpha(1-\alpha)\frac{L}{2}\|x - y\|_2^2 \tag{19.31}$$

Proof:

(19.26) can be obtained directly from *Lemma* 4.1.

(19.27): Fix $x_0 \in \mathbb{R}^n$. Consider the function

$$g(y) = f(y) - \langle \nabla f(x_0), y \rangle. \tag{19.32}$$

Note that $g(y)$ is also convex with L-Lipschitz gradient and its optimal point is $y^\star = x_0$. Therefore, we can utilize (19.26) and claim that

$$g(y^\star) \leq g(y - \frac{1}{L}\nabla g(y)) \leq g(y) - \frac{1}{2L}\|\nabla g(y)\|^2. \tag{19.33}$$

We then get (19.27) since $\nabla g(y) = \nabla f(y) - \nabla f(x_0)$.

(19.28): Let $g(z) = f(z) - \langle \nabla f(x), z \rangle$ and $h(z) = f(z) - \langle \nabla f(y), z \rangle$. We can check that g and h are both L-Lipschitz gradient. Taking the derivative of z on both g and h, we find that the optimal solution of $g(z)$ is $z_g^\star = x$ and the optimal solution of $h(z)$ is $z_h^\star = y$. Hence, we have $g(x) \leq g(y) - \frac{1}{2L}\|\nabla g(y)\|_2^2$ and $h(y) \leq h(x) - \frac{1}{2L}\|\nabla h(x)\|_2^2$. By combining these two inequalities, we obtain the original statement.

(19.29): Using Cauchy-Schwarz inequality, we can get

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \|\nabla f(x) - \nabla f(y)\| \|x - y\| \leq L\|x - y\|_2^2. \tag{19.34}$$

(19.30): The statement can be obtained from (19.27), by applying exactly the same trick as in the proof of (19.31) below.

(19.31): Let $z = \alpha x + (1-\alpha)y$. Using (19.26), we can get

$$f(x) \leq f(z) + \langle \nabla f(z), x - z \rangle + \frac{L}{2}\|z - x\|^2. \tag{19.35}$$

$$f(y) \leq f(z) + \langle \nabla f(z), y - z \rangle + \frac{L}{2}\|z - y\|^2. \tag{19.36}$$

Multiplying (19.35) by α, (19.36) by $1-\alpha$ and adding, we can get

$$\begin{aligned}
\alpha f(x) + (1-\alpha)f(y) &\leq f(z) + \frac{\alpha L}{2}\|z - x\|^2 + \frac{(1-\alpha)L}{2}\|z - y\|^2 && (19.37)\\
&= f(z) + \frac{\alpha(1-\alpha)^2 L}{2}\|y - x\|^2 + \frac{(1-\alpha)\alpha^2 L}{2}\|y - x\|^2 && (19.38)\\
&= f(z) + \alpha(1-\alpha)\frac{L}{2}\|x - y\|_2^2. && (19.39)
\end{aligned}$$

□

Using Theorem 19.9, we can analyse the convergence of the projected subgradient method. Beginning with the update step we have,

$$\begin{aligned}
x_{k+1} &= x_k - h_k \nabla f(x_k) && (19.40)\\
\|x_{k+1} - x^*\|_2^2 &= \|x_k - x^* - h_k \nabla f(x_k)\|_2^2 && (19.41)\\
&= \|x_k - x^*\|_2^2 - 2h_k\langle \nabla f(x_k), x_k - x^* \rangle + h_k^2\|\nabla f(x_k)\|_2^2. && (19.42)
\end{aligned}$$

From (19.28),

$$\|x_{k+1} - x^*\|_2^2 \leq \|x_k - x^*\|_2^2 - h\left(\frac{2}{L} - h\right)\|\nabla f(x_k)\|_2^2. \tag{19.43}$$

From (19.26),

$$\begin{aligned}
\Leftrightarrow f(x)_{k+1} &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 && (19.44)\\
&= f(x_k) - h\left(1 - \frac{L}{2}\right) h\|\nabla f(x_k)\|_2^2 && (19.45)\\
\Leftrightarrow f(x_k) - f^* &\leq \langle \nabla f(x_k), x_k - x^*\rangle && (19.46)\\
&\leq \|\nabla f(x)\| \cdot \|x_k - x^*\| && (19.47)\\
&\leq \|\nabla f(x)\| \cdot \|x_0 - x^*\| && (19.48)\\
\Leftrightarrow f(x_{k+1}) - f^* &= f(x_k) - f^* + f(x_{k+1}) - f(x_k) && (19.49)\\
&\leq (f(x_k) - f^*) - h\left(1 - \frac{L}{2}\right) h\|\nabla f(x_k)\|_2^2 && (19.50)\\
&\leq f(x_k) - f^* - \frac{h(1 - L/2)h}{\|x_0 - x^*\|_2^2}(f(x_k) - f^*)^2. && (19.51)
\end{aligned}$$

If $h_k = \frac{1}{L}$, this implies that,

$$\begin{aligned}
f(X_k) - f^* \leq \frac{2L(f(x_0) - f^*)\|x_0 - x^*\|_2^2}{2L\|x_0 - x^*\|_2^2 + K(f(x_0) - f^*)} && (19.52)\\
\leq \frac{2L\|x_0 - x^*\|_2^2}{K + 4}. && (19.53)
\end{aligned}$$

Thus, we see that the convergence rate of the projected subgradient method is

$$\sim O\left(\frac{1}{K}\right). \tag{19.54}$$

19.5 Next Class

In the next class we will explore Stochastic Subgradient Methods.