IE 598: Big Data Optimization **Fall 2016**

Lecture 14: Subgradient Method – October 11

Lecturer: Niao He *Scribers: Kaiqing Zhang*

Overview: From this lecture on, we are going to focus on nonsmooth convex optimization problems, where the objective functions are not necessarily continuously differentiable. We study the first algorithm, *subgradient method*, to address such problems. We investigate the convergence rates of the Subgradient Method under various problem settings.

14.1 Problem Setting: Nonsmooth Convex Optimization

In the previous lectures, we mainly discussed smooth convex optimization problems, where the objective function is continuously differentiable and the gradient is Lipschitz continuous. However, such differentiability may not hold for many applications, such as LASSO [Tib96], Support Vector Machine [SuyVan99], or Optimal Control [Vin00]. We now consider nonsmooth convex optimization problems

$$\min_{x\in X} f(x)$$

where

- $X \subset \mathbb{R}^n$ is convex and compact.
- $f : X \to \mathbb{R}$ is convex, possibly non-differentiable, but Lipschitz continuous on X.

Accordingly, we can define two important quantities of X and f as

- $\Omega = \Omega(X) = \max\limits_{x,y\in X} \frac{1}{2}\|x-y\|_2^2$ is the diameter of X.
- $M = M_{\|\cdot\|_2}(f) = \sup\limits_{x,y\in X} \frac{|f(x)-f(y)|}{\|x-y\|_2} < +\infty$ is the constant that characterizes the Lipschitz continuity.

Note that as discussed in Lecture 4, the convexity of the function f can actually lead to local Lipschitz continuity. We will simply make this an assumption in the in the subsequent text. In addition, due to the convexity, subgradient of f always exists, which motivates the Subgradient Method as follows.

14.2 Subgradient Method [N. Z. Shor, 1967]

The subgradient method, also called *subgradient descent*, was first proposed by Dr. Naum Zuselevich Shor in 1967. It works as follows

Algorithm 1 Subgradient Method

```
Pick $x_1 \in X$
while the iteration converges do
    Pick $g(x_t) \in \partial f(x_t)$
    $x_{t+1} = \Pi_X(x_t - \gamma_t g(x_t))$
```

where

- $g(x_t) \in \partial f(x_t)$ is a subgradient of f at x_t, which implies that $f(y) \geq f(x_t) + g(x_t)^T(y - x_t), \forall y \in X$.
- $\gamma_t > 0$ is the stepsize.
- $\Pi_X(x) = \arg\min_{y \in X} \|x - y\|_2^2$ is the projection operation.

Remark 14.1 *Unlike Gradient Descent (GD), Subgradient Descent (SD) method is not a descent method, i.e. moving along negative direction subgradient is not necessarily decreasing the objective function.*

To illustrate this, consider the function $f(x) = |x_1| + 2|x_2|$ with the level set contour as in Figure 15.1. For example, at $x = (1, 0)$, the the subdifferentiable set is $\partial f(x) = \{(1, 2a), a \in [-1, 1]\}$. If we choose $g = (1, 2)$ as shown, $-g$ is obviously not a descent direction.

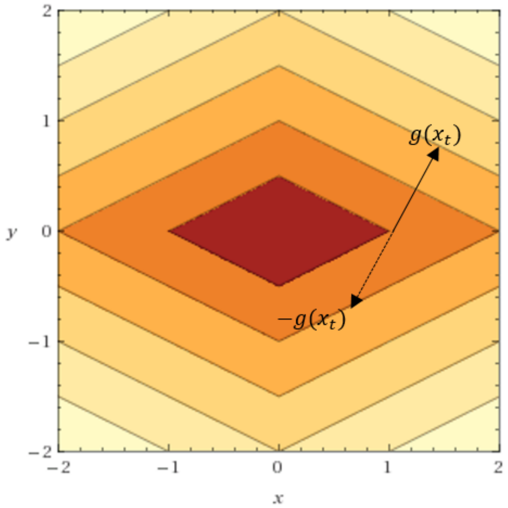

Figure 14.1: The contour of the function $f(x) = |x_1| + 2|x_2|$ and one of the subgradient direction at $(1, 0)$.

Choices of Stepsizes Stepsize γ_t is an important parameter that need to be selected during the iterations, which will affect the convergence analysis as we will show later. Four most commonly used stepsizes include:

1. Constant stepsize: $\gamma_t \equiv \gamma > 0$.
2. Scaled stepsize: $\gamma_t = \frac{\gamma}{\|g(x_t)\|_2}$.
3. Non-summable but diminishing stepsize satisfying:

$$\sum_{t=1}^{\infty} \gamma_t = \infty, \qquad \lim_{t \to \infty} \gamma_t = 0$$

e.g., $\gamma_t \sim O(\frac{1}{\sqrt{t}})$.

4. Non-summable but square-summable stepsize satisfying:
$$\sum_{t=1}^{\infty} \gamma_t = \infty, \qquad \sum_{t=1}^{\infty} \gamma_t^2 < \infty,$$
e.g., $\gamma_t \sim O(\frac{1}{t})$.

5. Polyak stepsize: Assuming $f_* = f(x_*)$ is known, choose
$$\gamma_t = \frac{f(x_t) - f_*}{\|g(x_t)\|_2^2}.$$

14.3 Convergence Analysis

To be discussed later, it turns out that subgradient descent behaves substantially different from gradient descent. The choices of stepsize, rates of convergence, and criterion used to measure the convergence are different. As mentioned in Remark 14.1, subgradient descent is not a descent method. Hence we will need to introduce other quantities to measure the convergence, instead of the quantity $f(x_t) - f_*$ as used earlier.

14.3.1 Convex Case

For convex f, we have the following theorem.

Theorem 14.2 *Assume f is convex, then subgradient method satisfies*
$$\min_{1\le t\le T} f(x_t) - f_* \le \left(\sum_{t=1}^{T} \gamma_t\right)^{-1} \left(\frac{1}{2}\|x_1 - x_*\|_2^2 + \frac{1}{2}\sum_{t=1}^{T} \gamma_t^2 \|g(x_t)\|_2^2\right) \tag{14.1}$$
and
$$f(\hat{x}_T) - f_* \le \left(\sum_{t=1}^{T} \gamma_t\right)^{-1} \left(\frac{1}{2}\|x_1 - x_*\|_2^2 + \frac{1}{2}\sum_{t=1}^{T} \gamma_t^2 \|g(x_t)\|_2^2\right) \tag{14.2}$$
where $\hat{x}_T = \left(\sum_{t=1}^{T} \gamma_t\right)^{-1} \left(\sum_{t=1}^{T} \gamma_t x_t\right) \in X$.

Proof: The proof uses the similar technique as in the convergence for smooth problems. First
$$\begin{aligned}\|x_{t+1} - x_*\|_2^2 =& \|\Pi_X(x_t - \gamma_t g(x_t)) - \Pi_X(x_*)\|_2^2 \\ \le& \|x_t - \gamma_t g(x_t) - x_*\|_2^2 \\ =& \|x_t - x_*\|_2^2 - 2\gamma_t g(x_t)^T (x_t - x_*) + \gamma_t^2 \|g(x_t)\|_2^2\end{aligned}$$
the inequality comes from the non-expansiveness of the projection operation. Therefore we have
$$\gamma_t g(x_t)^T (x_t - x_*) \le \frac{1}{2}\left(\|x_t - x_*\|_2^2 - \|x_{t+1} - x_*\|_2^2 + \gamma_t^2 \|g(x_t)\|_2^2\right). \tag{14.3}$$

Due to the convexity of f, we have

$$\gamma_t g(x_t)^T(x_t - x_*) \geq \gamma_t \left(f(x_t) - f_*\right). \tag{14.4}$$

Combining (14.3) and (14.4) and adding both sides of the inequality from $t = 1$ to $t = T$, we obtain

$$\begin{aligned}\sum_{t=1}^{T} \gamma_t \left(f(x_t) - f_*\right) \leq & \frac{1}{2}\left(\|x_1 - x_*\|_2^2 - \|x_{T+1} - x_*\|_2^2 + \sum_{t=1}^{T} \gamma_t^2 \|g(x_t)\|_2^2\right) \\ \leq & \frac{1}{2}\left(\|x_1 - x_*\|_2^2 + \sum_{t=1}^{T} \gamma_t^2 \|g(x_t)\|_2^2\right). \end{aligned} \tag{14.5}$$

For the proof of (14.1), by definition, the left hand side of (14.5) can be lower bounded by

$$\sum_{t=1}^{T} \gamma_t \left(f(x_t) - f_*\right) \geq \left(\sum_{t=1}^{T} \gamma_t\right) \cdot \left(\min_{1 \leq t \leq T} f(x_t) - f_*\right)$$

For the proof of (14.2), first note that $\hat{x}_T$ is a convex combination of $\{x_1, \cdots, x_T\}$. Due to the convexity of f, we have

$$\sum_{t=1}^{T} \gamma_t f(x_t) \geq \left(\sum_{t=1}^{T} \gamma_t\right) \cdot f(\hat{x}_T)$$

and left hand side of (14.5) is thus lower bounded by

$$\sum_{t=1}^{T} \gamma_t \left(f(x_t) - f_*\right) \geq \left(\sum_{t=1}^{T} \gamma_t\right) \cdot \left(f(\hat{x}_T) - f_*\right).$$

Bounds (14.1) and (14.2) are hence proved. ■

Remark. Invoking the definition of Ω and M, we have $\frac{1}{2}\|x_1 - x_*\|_2^2 \leq \Omega$ and $\|g(x_t)\|_2 \leq M$. As a corollary,

$$\begin{aligned}\min_{1 \leq t \leq T} f(x_t) - f_* &\leq \left(\Omega + \frac{1}{2}\sum_{t=1}^{T} \gamma_t^2 M^2\right) \Big/ \left(\sum_{t=1}^{T} \gamma_t\right), \\ f\left(\frac{\sum_{t=1}^{T} \gamma_t x_t}{\sum_{t=1}^{T} \gamma_t}\right) - f_* &\leq \left(\Omega + \frac{1}{2}\sum_{t=1}^{T} \gamma_t^2 M^2\right) \Big/ \left(\sum_{t=1}^{T} \gamma_t\right). \end{aligned}$$

Remark 2. Slightly modifying the summation or averaging from T_0 to T instead of from 1 to T, we end up with a even more general results

$$\begin{aligned}\min_{T_0 \leq t \leq T} f(x_t) - f_* &\leq \left(\Omega + \frac{1}{2}\sum_{t=T_0}^{T} \gamma_t^2 M^2\right) \Big/ \left(\sum_{t=T_0}^{T} \gamma_t\right), \quad \forall 1 \leq T_0 \leq T \\ f\left(\frac{\sum_{t=T_0}^{T} \gamma_t x_t}{\sum_{t=T_0}^{T} \gamma_t}\right) - f_* &\leq \left(\Omega + \frac{1}{2}\sum_{t=T_0}^{T} \gamma_t^2 M^2\right) \Big/ \left(\sum_{t=T_0}^{T} \gamma_t\right), \quad \forall 1 \leq T_0 \leq T. \end{aligned}$$

Convergence with various stepsizes. It is interesting to see how the bounds in (14.1) and (14.1) would imply the convergence and even the convergence rate with different choices of stepsizes. By abuse of notation, we denote both $\min_{1\le t\le T} f(x_t) - f_*$ and $f(\hat{x}_T) - f_*$ as ϵ_T.

1. *Constant stepsize:* with $\gamma_t \equiv \gamma$,

$$\epsilon_T \le \frac{\Omega + (T/2)\gamma^2 M^2}{T\gamma} = \frac{\Omega}{T}\cdot\frac{1}{\gamma} + \frac{M^2}{2}\gamma \xrightarrow{T\to\infty} \frac{M^2}{2}\gamma.$$

   It is worth noticing that the error upper-bound does not diminish to zero as T grows to infinity, which shows one of the drawbacks of using arbitrary constant stepsizes. In addition, to optimize the upper bound, we can select the optimal stepsize γ_* to obtain:

$$\gamma_* = \frac{\sqrt{2\Omega}}{M\sqrt{T}} \Rightarrow \epsilon_T \le \frac{\sqrt{2\Omega}M}{\sqrt{T}}.$$

   It is shown that under this optimal choice $\epsilon_T \sim O(\frac{\sqrt{\Omega}M}{\sqrt{T}})$. However, this exhibits another drawback of constant stepsize that in practice T is not known in prior for evaluating the optimal γ_*.

2. *Scaled stepsize:* with $\gamma_t = \frac{\gamma}{\|g(x_t)\|_2}$,

$$\epsilon_T \le \frac{\Omega + (1/2)\gamma^2 T}{\gamma\sum_{t=1}^{T} 1/\|g(x_t)\|_2} \le M\left(\frac{\Omega}{T}\cdot\frac{1}{\gamma} + \frac{1}{2}\gamma\right) \xrightarrow{T\to\infty} \frac{M}{2}\gamma.$$

   Similarly, we can select the optimal γ by minimizing the right hand side, i.e. $\gamma_* = \frac{\sqrt{2\Omega}}{\sqrt{T}}$:

$$\gamma_t = \frac{\sqrt{2\Omega}}{\sqrt{T}\|g(x_t)\|_2} \Rightarrow \epsilon_T \le \frac{\sqrt{2\Omega}M}{\sqrt{T}}.$$

   The same convergence rate is achieved while the same drawback about not knowing T in prior still exists in choosing γ_t.

3. *Non-summable but diminishing stepsize:*

$$\begin{aligned}\epsilon_T &\le \left(\Omega + \frac{1}{2}\sum_{t=1}^{T}\gamma_t^2 M^2\right) \Big/ \left(\sum_{t=1}^{T}\gamma_t\right)\\ &\le \left(\Omega + \frac{1}{2}\sum_{t=1}^{T_1}\gamma_t^2 M^2\right) \Big/ \left(\sum_{t=1}^{T}\gamma_t\right) + \left(\frac{M^2}{2}\sum_{t=T_1+1}^{T}\gamma_t^2\right) \Big/ \left(\sum_{t=T_1+1}^{T}\gamma_t\right)\end{aligned}$$

   where $1 \le T_1 \le T$. When $T \to \infty$, select large T_1 and the first term on the right hand side $\to 0$ since γ_t is non-summable. The second term also $\to 0$ because γ_t^2 always approaches zero faster than γ_t. Consequently, we know that

$$\epsilon_T \xrightarrow{T\to\infty} 0.$$

   An example choice of the stepsize is $\gamma_t = O\left(\frac{1}{t^q}\right)$ with $q \in (0, 1]$. As in the above cases, if we choose $\gamma_t = \frac{\sqrt{2\Omega}}{M\sqrt{t}}$, then

$$\gamma_t = \frac{\sqrt{2\Omega}}{M\sqrt{t}} \Rightarrow \epsilon_T \le O\left(\frac{\sqrt{\Omega}M\ln(T)}{\sqrt{T}}\right).$$

In fact, if we choose the averaging from $\frac{T}{2}$ instead of 1, we have

$$\min_{T/2\leq t\leq T} f(x_t) - f_* \leq O\left(\frac{M\cdot\sqrt{\Omega}}{\sqrt{T}}\right).$$

4. *Non-summable but square-summable stepsize:* It is obvious that

$$\epsilon_T \leq \left(\Omega + \frac{M^2}{2}\sum_{t=1}^{T}\gamma_t^2\right) \Bigg/ \left(\sum_{t=1}^{T}\gamma_t\right) \xrightarrow{T\to\infty} 0.$$

A typical choice of $\gamma_t = \frac{1}{t^{1+q}}, q>0$ also result in the rate of $O(\frac{1}{\sqrt{T}})$.

5. *Polyak stepsize:* The motivation of choosing this stepsize comes from the fact that

$$\begin{aligned}\|x_{t+1}-x_*\|_2^2 \leq & \|x_t - x_*\|_2^2 - 2\gamma_t g(x_t)^T(x_t - x_*) + \gamma_t^2\|g(x_t)\|_2^2 \\ \leq & \|x_t - x_*\|_2^2 - 2\gamma_t\left(f(x_t) - f_*\right) + \gamma_t^2\|g(x_t)\|_2^2\end{aligned}$$

as we showed in the proof of **Theorem** 14.2. The Polyak step $\gamma_t = \frac{f(x_t)-f_*}{\|g(x_t)\|_2^2}$ is exactly the minimizer of the right hand side. In fact, the stepsize yields

$$\|x_{t+1}-x_*\|_2^2 \leq \|x_t - x_*\|_2^2 - \frac{\left(f(x_t)-f_*\right)^2}{\|g(x_t)\|_2^2}, \tag{14.6}$$

which guarantees $\|x_t - x_*\|_2^2$ decreases each step. Applying (14.6) recursively, we obtain

$$\sum_{t=1}^{T}\left(f(x_t)-f_*\right)^2 \leq 2\Omega\cdot M < \infty.$$

Therefore we have $\epsilon_T \to 0$ as $T\to\infty$ and $\epsilon_T \leq O\left(\frac{1}{\sqrt{T}}\right)$ (otherwise ϵ_T will not be square-summable).

14.3.2 Strongly Convex Case

For strongly convex function f, we obtain the following theorem.

Theorem 14.3 *Assume f is μ-strongly convex, then subgradient method with stepsize $\gamma_t = \frac{1}{\mu t}$ satisfies*

$$\min_{1\leq t\leq T} f(x_t) - f_* \leq \frac{M^2(\ln(T)+1)}{2\mu T} \tag{14.7}$$

and

$$f\left(\hat{x}_T\right) - f_* \leq \frac{M^2(\ln(T)+1)}{2\mu T}, \qquad \textit{with } \hat{x}_T := \frac{1}{T}\sum_{t=1}^{T}x_t \tag{14.8}$$

where M is as defined in Section 14.1.

Proof: First recall that μ-strongly convex implies that

$$f(y) \geq f(x) + g(x)^T(y-x) + \frac{\mu}{2}\|x-y\|_2^2, \quad \forall x,y\in X.$$

Similarly as the proof for **Theorem** 14.2, the left hand side of (14.3) can be lower bounded by

$$\gamma_t g(x_t)^T(x_t - x_*) \geq \gamma_t \left(f(x_t) - f_* + \frac{\mu}{2}\|x_t - x_*\|_2^2 \right).$$

Combining (14.3) and plug in $\gamma_t = \frac{1}{\mu t}$ we have

$$f(x_t) - f_* \leq \left(\frac{\mu}{2}(t-1)\|x_t - x_*\|_2^2 - \frac{\mu}{2}t\|x_{t+1} - x_*\|_2^2 + \frac{1}{2\mu t}\|g(x_t)\|_2^2 \right).$$

By recursively adding both sides from $t = 1$ to $t = T$, we obtain

$$\sum_{t=1}^{T} f(x_t) - f_* \leq \sum_{t=1}^{T} \frac{1}{2\mu t}\|g(x_t)\|_2^2 \leq \frac{M^2}{2\mu}\sum_{t=1}^{T}\frac{1}{t} \leq \frac{M^2}{2\mu}(\ln(T) + 1).$$

In addition, we have $\sum\limits_{t=1}^{T} f(x_t) - f_* \geq T \cdot \epsilon_T$ for either $\min\limits_{1\leq t\leq T} f(x_t) - f_*$ or $f(\hat{x}_T) - f_*$ with $\hat{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$ as shown in the previous proof, which leads to the bounds (14.7) and (14.8). ■

With another choice of stepsize and averaging strategy, we can get rid of the log factor in the bound. The following theorem can be obtained [Bub14].

Theorem 14.4 *Assume f is μ-strongly convex, then subgradient method with stepsize $\gamma_t = \frac{2}{\mu(t+1)}$ satisfies*

$$\min_{1\leq t\leq T} f(x_t) - f_* \leq \frac{2M^2}{\mu(T+1)} \tag{14.9}$$

and

$$f(\hat{x}_T) - f_* \leq \frac{2M^2}{\mu(T+1)}, \qquad \text{with } \hat{x}_T = \sum_{t=1}^{T} \frac{2t}{T(T+1)} x_t. \tag{14.10}$$

, M is as defined in Section 15.1.

Proof: Similarly as the proof for **Theorem** 14.3, we first have

$$\gamma_t g(x_t)^T(x_t - x_*) \geq \gamma_t \left(f(x_t) - f_* + \frac{\mu}{2}\|x_t - x_*\|_2^2 \right).$$

Substitute $\gamma_t = \frac{2}{\mu(t+1)}$ in 14.3, we have

$$f(x_t) - f_* \leq \left(\frac{\mu}{4}(t-1)\|x_t - x_*\|_2^2 - \frac{\mu}{4}(t+1)\|x_{t+1} - x_*\|_2^2 + \frac{1}{\mu(t+1)}\|g(x_t)\|_2^2 \right).$$

Multiplying both sides by t leads to

$$\begin{aligned} t\left(f(x_t) - f_*\right) &\leq \left(\frac{\mu}{4}t(t-1)\|x_t - x_*\|_2^2 - \frac{\mu}{4}t(t+1)\|x_{t+1} - x_*\|_2^2 + \frac{t}{\mu(t+1)}\|g(x_t)\|_2^2 \right) \\ &\leq \left(\frac{\mu}{4}t(t-1)\|x_t - x_*\|_2^2 - \frac{\mu}{4}t(t+1)\|x_{t+1} - x_*\|_2^2 + \frac{1}{\mu}\|g(x_t)\|_2^2 \right). \end{aligned}$$

Recursively adding both sides from $t = 1$ to $t = T$, we obtain

$$\sum_{t=1}^{T} t\left(f(x_t) - f_*\right) \leq \frac{T}{\mu}\|g(x_t)\|_2^2 \leq \frac{T}{\mu}M^2. \tag{14.11}$$

Moreover, by definition and convexity we have

$$\sum_{t=1}^{T} tf(x_t) \geq \left(\frac{T(T+1)}{2}\right) \cdot f\left(\sum_{t=1}^{T} \frac{2t}{T(T+1)} x_t\right),$$

and

$$\sum_{t=1}^{T} tf(x_t) \geq \left(\frac{T(T+1)}{2}\right) \cdot \min_{1 \leq t \leq T} f(x_t).$$

Combining with (14.11) and dividing both sides by $\frac{T(T+1)}{2}$, we conclude the proof.

■

14.4 Summary

It is worth comparing the convergence rate of subgradient method for nonsmooth convex optimization problems with the optimal rate for smooth problems. Actually, it is shown in Table 14.1 that in both convex and strongly convex cases, subgradient method achieves slower convergence than the accelerated gradient descent method. Particularly, subgradient method can only achieve sublinear convergence even under the strongly convex case, instead of linear rate in smooth case. By constructing worst case functions, we will show that this is actually the optimal rate one can get for nonsmooth convex optimization next lecture.

Table 14.1: Comparison of convergence rates for nonsmooth and smooth convex optimization problems.

	Convex	**Strongly Convex**
Subgradient method	$O\left(\frac{\sqrt{\Omega}M}{\sqrt{t}}\right)$	$O\left(\frac{M^2}{\mu t}\right)$
Accelerated gradient descent	$O\left(\frac{L \cdot D^2}{t^2}\right)$	$O\left(\left(\frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}\right)^{2t}\right)$

References

[Tib96] R. Tibshirani, "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[SuyVan99] J. Suykens, and J. Vandewalle, "Least squares support vector machine classifiers", *Neural processing letters*, 9(3), 293-300, 1999.

[Vin00] R. Vinter, "Optimal control. Systems & control: foundations & applications", *Birkauser, Boston*, 2000.

[Nes13] Y. Nesterov, "Introductory lectures on convex optimization: A basic course", *Springer Science & Business Media*, vol. 87, 2013.

[Bub14] S. Bubeck, "Convex optimization: Algorithms and complexity", arXiv preprint arXiv:1405.4980, 2014.