IE 521: Convex Optimization **Spring 2017, UIUC**

Lecture 25: Subgradient Method and Bundle Methods – April 24

Instructor: Niao He *Scribe: Shuanglong Wang*

Courtesy warning: These notes do not necessarily cover everything discussed in the class. Please email TA (swang157@illinois.edu) if you find any typos or mistakes.

In this lecture, we cover the following topics

- Subgradient Method
- Bundle Method
  - Kelley cutting plane method
  - Level method

Reference: Nesterov.2004. Chapter 3.2.3, 3.3

25.1 Subgradient Method

Recall that subgradient method works as follows

$$x_{t+1} = \Pi_X(x_t - \gamma_t g_t), \quad t = 1, 2, ...$$

where $g_t \in \partial f(x_t)$, $\gamma_t > 0$ and $\Pi_X(x) = \arg\min_{y \in X} \parallel y - x \parallel_2$ is the projection operator.

Note that the projection on X is easy to compute when X is simple, e.g. X is a ball, box, simplex, polyhedron, etc.

Lemma 25.1 (Projection) *$\forall x \; in \mathbf{R}^n$, $z \in X$,*

$$\parallel x - z \parallel_2^2 \geq \parallel x - \Pi_X(x) \parallel_2^2 + \parallel z - \Pi_X(x) \parallel_2^2$$

Proof: When $x \in X$, the inequality immediately hold true. Let $x \notin X$. By definition.

$$\Pi_X(x) = \arg\min_{x \in X} \parallel z - x \parallel_2^2$$

By optimality condition, this implies $2(\Pi_X(x) - x)^T(z - \Pi_X(x)) \geq 0, \forall z \in X$. Hence,

$$\begin{aligned} \parallel x - z \parallel_2^2 &= \parallel x - \Pi_X(x) + \Pi_X(x) - z \parallel_2^2 \\ &\geq \parallel x - \Pi_X(x) \parallel_2^2 + \parallel \Pi_X(x) - z \parallel_2^2 \end{aligned}$$

■

Lemma 25.2 (Key relation) *For the subgradient method, we have*

$$\| x_{t+1} - x^* \|_2^2 \quad \geq \quad \| x_t - x^* \|_2^2 - 2\gamma_t(f(x_t) - f^*) + \gamma_t^2 \| g_t \|_2^2 \qquad (\star)$$

Proof:

$$\begin{aligned} \| x_{t+1} - x^* \|_2^2 &= \| \Pi_X(x_t - \gamma_t g_t) - x^* \|_2^2 \\ &\leq \| x_t - \gamma_t g_t - x^* \|_2^2 \\ &= \| x_t - x^* \|_2^2 - 2\gamma_t g_t^T(x_t - x^*) + \gamma_t^2 \| g_t \|^2 \end{aligned}$$

Due to convexity of f, we have $f^* \geq f(x_t) + g_t^T(x^* - x_t)$, i.e.

$$g_t^T(x_t - x^*) > f(x_t) - f^*$$

Combing there two inequalities leads to the desired result. ■

Remark: Note that when f^* is known, we can choose the 'optimal' γ_t by minimizing the right hand side of $(\star)$: $\gamma_t^* = \frac{f(x_t) - f^*}{\|g_t\|_2^2}$, which is the Polyak's stepsize.

In fact, knowing f^* is not a problem sometimes. For instance, when the goal is to solve the convex feasibility problem: Find $x^* \in X$, s.t. $g_i(x) \leq 0$, $i = 1, ..., m$. We can formulate this as

$$\min_{x \in X} \max_{1 \leq i \leq m} g_i(x) \quad \text{or} \quad \min_{x \in X} \sum_{i=1}^{m} \max(g_i(x), 0)$$

The optimal value f^* is known to be 0 in this case.

If f^* is not known, one can replace f^* by its online estimate.

Theorem 25.3 (Convergence) *Suppose $f(x)$ is convex and Lipschitz continuous on X:*

$$|f(x) - f(y)| \leq M_f \| x - y \|_2, \quad \forall x, y \in X$$

where $M_f < +\infty$. Then the subgradient method satisfies:

$$f(\hat{x}_T) - f^* \leq \frac{\| x_1 - x^* \|_2^2 + \sum_{t=1}^{T} \gamma_t^2 M_f^2}{2 \sum_{t=1}^{T} \gamma_t}$$

where $\hat{x}_T = (\sum_{t=1}^{T} \gamma_t)^{-1} (\sum_{t=1}^{T} \gamma_t x_t)$

Proof: The Lipschitz continuity implies that $\| g_t \|_2 \leq M_f, \forall t$. Summing up the key relation $(\star)$ from $t = 1$ to $t = T$, we obtain

$$2 \sum_{t=1}^{T} \gamma_t (f(x_t) - f^*) \leq \| x_1 - x^* \|_2^2 - \| x_{T+1} - x^* \|_2^2 + \sum_{t=1}^{T} \gamma_t^2 M_f^2$$

By convexity of f: $(\sum_{t=1}^{T} \gamma_t) f(\hat{x}_T) \leq \sum_{t=1}^{T} \gamma_t f(x_t)$ This further leads to

$$(\sum_{t=1}^{T} \gamma_t)[f(\hat{x}_T) - f^*] \leq \frac{1}{2} \| x_1 - x^* \|^2 + \frac{M^2}{2} \sum_{t=1}^{T} \gamma_t^2$$

and concludes the proof. ■

Convergence under various stepsize Assume $D_X = \max_{x,y} \| x - y \|_2$ is the diameter of the set X. It is interesting to see how the bounds in the above theorem 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 $\gamma_t \equiv \gamma$,

$$\epsilon_T \le \frac{D_X^2 + T\gamma^2 M_f^2}{2T\gamma} = \frac{D_X^2}{2T}\cdot\frac{1}{\gamma} + \frac{M_f^2}{2}\gamma \xrightarrow{T\to\infty} \frac{M_f^2}{2}\gamma.$$

   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:

$$\gamma_* = \frac{D_X}{M_f\sqrt{T}} \Rightarrow \epsilon_T \le \frac{D_X M_f}{\sqrt{T}}.$$

   It is shown that under this optimal choice $\epsilon_T \sim O(\frac{D_X M_f}{\sqrt{T}})$. 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 $\gamma_t = \frac{\gamma}{\|g(x_t)\|_2}$,

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

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

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

   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:*

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

   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

$$\gamma_t = \frac{D_X}{M_f\sqrt{t}} \Rightarrow \epsilon_T \leq O\left(\frac{D_X M_f \ln(T)}{\sqrt{T}}\right).$$

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

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

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 stepsize yields

$$\|x_{t+1} - x_*\|_2^2 \leq \|x_t - x_*\|_2^2 - \frac{(f(x_t) - f_*)^2}{\|g(x_t)\|_2^2}, \tag{25.1}$$

which guarantees $\|x_t - x_*\|_2^2$ decreases each step. Applying (25.1) recursively, we obtain

$$\sum_{t=1}^{T}(f(x_t) - f_*)^2 \leq D_X^2 \cdot M_f < \infty.$$

Therefore we have $\epsilon_T \to 0$ as $T \to \infty$ and $\epsilon_T \leq O\left(\frac{1}{\sqrt{T}}\right)$.

Corollary 25.4 *When T is known, setting $\gamma_t \equiv \frac{D_X}{M_f\sqrt{T}}$, in particular, we have*

$$f(\hat{x}_T) - f^* \leq \frac{D_X M_f}{\sqrt{T}}$$

Remark: Subgradient method converges sublinearly. For an accuracy $\epsilon > 0$, need $O(\frac{D_X^2 M_f^2}{\epsilon^2})$ number of iterations.

25.2 Bundle Method

When running the subgradient method, we obtain a bundle of affine underestimate of $f(x)$:

$$f(x_t) + g_t^T(x - x_t), \quad t = 1, 2, ...$$

Definition 25.5 *The piecewise linear function:*

$$f_t(x) = \max_{1 \le i \le t} \left\{ f(x_i) + g_i^T (x - x_i) \right\}$$

where $g_i \in \partial f(x_i)$ is called the t-th model of convex function f.

Note

1. $f_t(x) \le f(x), \forall x \in X$
2. $f_t(x_i) = f(x_i), \forall 1 \le i \le t$
3. $f_1(x) \le f_2(x) \le ... \le f_t(x) \le ... \le f(x)$

25.2.1 Kelley method (Kelley, 1960)

The Kelley method works as follows:

$$x_{t+1} = \arg\min_{x \in X} f_t(x)$$

Obviously, the above algorithm converges so long as X is compact. The auxiliary problem is not so disturbing (reduces to LP) when X is polyhedron. However, the issue is that x_t is not unique and Kelley method can be very unstable. Indeed, the worst-case complexity of Kelley method is at least $O(\frac{1}{\epsilon^{(n-1)/2}})$.

Remedy: To prevent the instability issue, a possible remedy is update x_{t+1} by

$$x_{t+1} = \arg\min_{x \in X} \left\{ f_t(x) + \frac{\alpha_t}{2} \parallel x - x_t \parallel_2^2 \right\}$$

with properly selected $\alpha_t > 0$.

25.2.2 Level Method (Lemarchal, Nemirovski, Nesterov, 1995)

Denote

$$\underline{f}_t = \min_{x \in X} f_t(x) \qquad \text{(minimal value of the model)}$$

$$\bar{f}_t = \min_{1 \le i \le t} f(x_i) \qquad \text{(record value of the model)}$$

we have $\underline{f}_1 \le \underline{f}_2 \le ... \le f^* \le ... \le \bar{f}_2 \le \bar{f}_1$

Denote the level set

$$L_t = \left\{ x : f_t(x) \le l_t := (1 - \alpha)\underline{f}_t + \alpha \bar{f}_t \right\}$$

Note that L_t is nonempty, convex and closed, and doesn't contain the search points $\{x_1, ..., x_t\}$

The Level method works as follows

$$x_{t+1} = \Pi_{L_t}(x_t) = \arg\min_{x \in X} \left\{ \| x - x_t \|_2^2 \colon f_t(x) \leq l_t \right\}$$

Note when $\alpha = 0$, reduces to Kelley method. $\alpha = 1$, there will be no progress.

The auxiliary problem reduces to a quadratic program when X is polyhedron.

Theorem 25.6 *When* $t > \frac{1}{(1-\alpha)^2\alpha(2-\alpha)} \left(\frac{M_f D_X}{\epsilon}\right)^2$, *we have*

$$\bar{f}_t - \underline{f}_t \leq \epsilon$$

where M_f *is the Lipschitz constant and* D_X *is the diameter of set* X.

Remark: Level method achieves same complexity as the subgradient method (which indeed is unimprovable), but can perform much better than subgradient method in practice.