IE 521 Convex Optimization

Lecture 21: Large-Scale Optimization

Bundle Methods

Niao He

1st May 2019

Outline

Recap Subgradient Method

Simple Constrained Convex Problems

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in X \end{aligned}$$

Subgradient Method

$$x_{t+1} = \Pi_X(x_t - \gamma_t g_t), \quad g_t \in \partial f(x_t)$$

Problem Class	Stepsize	Convergence
Convex Lipschitz	$O(\frac{1}{\sqrt{t}})$	$O(\frac{D_X M_f}{\sqrt{t}})$
Strongly Convex Lipschitz	$O(\frac{1}{\mu t})$	$O(\frac{M_f^2}{\mu t})$

+ Simple: requires only subgradients and projections
+ Sublinear rate in general
− No general stopping criteria
− Not fully exploit past information

Bundle Methods

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

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

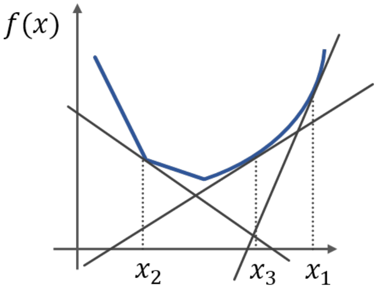


Figure: The bundle of affine underestimates

Polyhedral Model of the Objective

Definition. The piecewise linear function:

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

is called the t-th *polyhedral model* of convex function f.

Remark.

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. $f_1(x) \le \ldots \le f_t(x) \le f(x)$

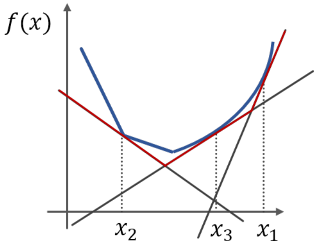


Figure: Polyhedral model

Kelley's Cutting Plane Method

(Kelley, 1960)

0. Initialize $x_1 \in X$
1. For $t \geq 1$, do

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

+ The algorithm converges so long as X is compact.
+ Auxiliary problem is easy to solve when X is polyhedron.
− Instability (solution to auxiliary problem is not unique)
− Suboptimal convergence rate: $O(1/\epsilon^3)$
− Poor performance in both theory and practice

How to stabilize Kelly's method?

Modified Kelley Method

Regularized approach: update x_{t+1} by $\hat{x}_{t+1}$

$$\hat{x}_{t+1} = \arg\min_{x \in X} \left\{ f_t(x) + \frac{\alpha_t}{2} \|x - x_t\|_2^2 \right\}$$

if objective is "sufficiently decreased".

Trust-region approach: update x_{t+1} by $\hat{x}_{t+1}$

$$\hat{x}_{t+1} = \arg\min_{x \in X} \{ f_t(x) : \|x - x_t\|_2 \leq \delta_t \}$$

if objective is "sufficiently decreased".

Drawbacks

- Unclear how to set parameters α_t, δ_t
- Unclear how to determine sufficient decrease
- Hard to analyze the convergence

Level-set Method

(Lemarchal, Nemirovski, Nesterov, 1995)

At each iteration, choose a level ℓ_t and update

$$x_{t+1} = \arg\min_{x \in X} \left\{ \frac{1}{2}\|x - x_t\|_2^2 : f_t(x) \le \ell_t \right\}$$

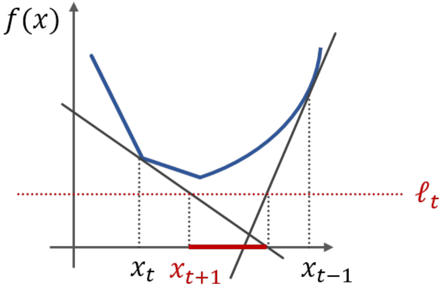


Figure: Level-set Method

Level-set 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 $\underline{f}_1 \le \ldots \le \underline{f}_t \le \ldots \le f^* \le \ldots \le \bar{f}_t \le \ldots \le \bar{f}_1$

- Define the level set

$$L_t = \left\{x : f_t(x) \le \ell_t := (1-\alpha)\underline{f}_t + \alpha \bar{f}_t\right\} \qquad \text{(level set)}$$

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

Level-set Method

(Lemarchal, Nemirovski, Nesterov, 1995)

0. Initialize $x_1 \in X$
1. For $t \geq 1$
   - Compute $\underline{f}_t$ and $\overline{f}_t$
   - Set $\ell_t = (1-\alpha)\underline{f}_t + \alpha \overline{f}_t$ and update

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

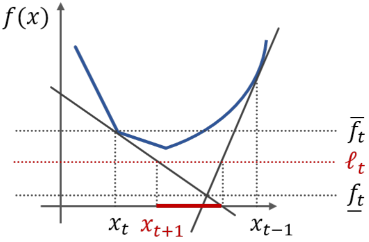


Figure: Level-set Method

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

Convergence of Level-set Method

Theorem. For $\alpha \in (0,1)$, whenever

$$T > \frac{1}{(1-\alpha)^2\alpha(2-\alpha)}\left(\frac{M_f D_X}{\epsilon}\right)^2,$$

we have

$$\min_{1\le t\le T} f(x_t) - f^* \le \bar{f}_T - \underline{f}_T \le \epsilon,$$

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

Corollary. Particularly, setting $\alpha^* = \frac{1}{2+\sqrt{2}}$, we have the efficiency estimate

$$T(\epsilon) \le \frac{4D_X^2 M_f^2}{\epsilon^2}.$$

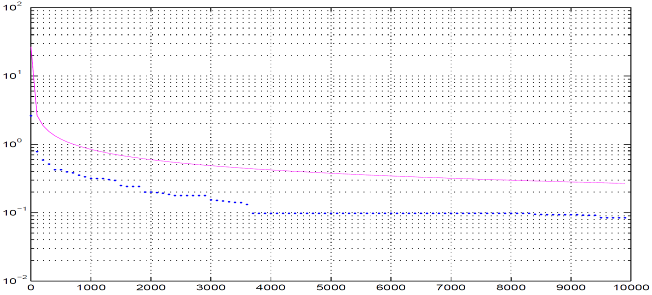
Remarks

- Same $O(\frac{1}{\epsilon^2})$ complexity as the subgradient method.
- Require computing projections onto polyhedrons.
- Require extra memory cost.
- Perform much better in practice and experimental evidence of polynomial-time complexity:

$$\mathcal{O}\left(\frac{\mathrm{Var}_X(f)}{\sqrt{t}}\right) \text{ vs. } \mathcal{O}\left(e^{-\frac{t}{n}}\mathrm{Var}_X(f)\right)$$

Illustration: ℓ_1-minimization

$$\min_{x \in \mathbb{R}^n : \|x\|_2 \leq 1} \quad \|Ax - b\|_1, \, n = 50$$



SD, accuracy vs. iteration count

(blue: errors; red: efficiency estimate $3\frac{\mathrm{Var}_{\|\cdot\|_2, X}(f)}{\sqrt{t}}$)

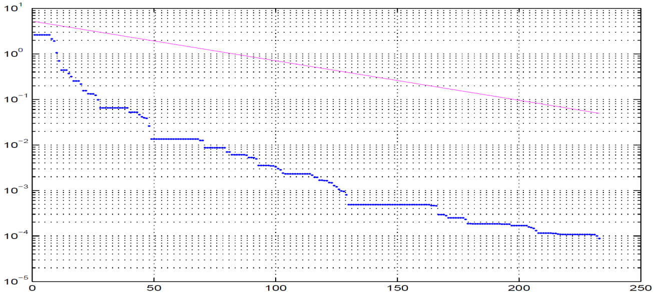
$\epsilon_{10000} = 0.084$

Figure: Subgradient Method

Figure from Chapter 5.3.2. in Ben-Tal & Nemirovski

Illustration: ℓ_1-minimization

$$\min_{x \in \mathbb{R}^n : \|x\|_2 \leq 1} \quad \|Ax - b\|_1, \, n = 50$$



BL, accuracy vs. iteration count
(blue: errors; red: efficiency estimate $e^{-\frac{t}{n}}\mathrm{Var}_X(f)$)
$\epsilon_{233} < 1.e-4$

Figure: Level-set Method

Figure from Chapter 5.3.2. in Ben-Tal & Nemirovski

Proof of Convergence

Define

$$\Delta_t := \bar{f}_t - \underline{f}_t$$

Immediate Observations:

1. $\Delta_1 \geq \Delta_2 \geq \cdots \geq \Delta_t \geq \cdots \geq 0$
2. $f_t(x_t) - f_t(x_{t+1}) \geq \bar{f}_t - \ell_t = (1-\alpha)\Delta_t$
3. $f_t(x)$ is also M_f-Lipschitz continuous
4. $\|x_t - x_{t+1}\|_2 \geq \frac{(1-\alpha)}{M_f}\Delta_t$

We want to find iteration T such that $\Delta_T \leq \epsilon$.
In other words, if $\Delta_T \geq \epsilon$, we want to show that

$$T \leq C(\alpha)\left(\frac{M_f D_X}{\epsilon}\right)^2$$

.

Proof of Convergence (cont'd)

Decompose the iterations into blocks such that

$$\underbrace{\Delta_T = \Delta_{t_1} \leq \cdots}_{J_1=\{t:\,\Delta_t\leq\frac{\Delta_{t_1}}{1-\alpha}\}} \leq \underbrace{\Delta_{t_2} \leq \cdots \leq}_{J_2=\{t:\,\Delta_t\leq\frac{\Delta_{t_2}}{1-\alpha}\}} \cdots \leq \underbrace{\Delta_{t_m} \leq \cdots \leq \Delta_{t_{m+1}} = \Delta_1}_{J_m=\{t:\,\Delta_t\leq\frac{\Delta_{t_m}}{1-\alpha}\}}$$

$$\Delta_{t_i} \leq \Delta_t \leq \frac{\Delta_{t_i}}{1-\alpha}, \forall t \in J_i, \forall i = 1, \dots, m$$

Claim: let $u_i = \mathrm{argmin}_{x\in X}\{f_{t_i}(x)\}$, then

$$u_i \in L_t := \{x : f_t(x) \leq \ell_t\}, \forall t \in J_i.$$

This is because

$$f_t(u_i) \leq f_{t_i}(u_i) = \underline{f}_{t_i} = \bar{f}_{t_i} - \Delta_{t_i} \leq \bar{f}_t - (1-\alpha)\Delta_t = \ell_t.$$

Proof of Convergence (cont'd)

- For $t \in J_i$, recall that $x_{t+1} = \Pi_{L_t}(x_t)$, we have

$$\begin{aligned}\|x_{t+1} - u_i\|_2^2 &\leq \|x_t - u_i\|_2^2 - \|x_{t+1} - x_t\|_2^2 \\ &\leq \|x_t - u_i\|_2^2 - \frac{(1-\alpha)^2}{M_f^2}\Delta_t^2 \\ &\leq \|x_t - u_i\|_2^2 - \frac{(1-\alpha)^2}{M_f^2}\Delta_{t_i}^2\end{aligned}$$

- Telescoping the sum over $t \in J_i$, we have

$$|J_i| \cdot \frac{(1-\alpha)^2}{M_f^2}\Delta_{t_i}^2 \leq D_X^2.$$

- It follows that

$$T = \sum_{i=1}^{m} |J_i| \leq \frac{D_X^2 M_f^2}{(1-\alpha)^2} \cdot \sum_{i=1}^{m} \frac{1}{\Delta_{t_i}^2}$$

Proof of Convergence (cont'd)

- On the other hand, of $\Delta_T = \Delta_{t_1} \geq \epsilon$, by construction of the blocks, we have

$$\Delta_{t_i} \geq \frac{1}{1-\alpha}\Delta_{t_{i-1}} \geq \cdots \geq \frac{1}{(1-\alpha)^{i-1}}\Delta_{t_1} \geq \frac{\epsilon}{(1-\alpha)^{i-1}}$$

- This implies that

$$\frac{1}{\Delta_{t_i}^2} \leq \frac{(1-\alpha)^{2i-2}}{\epsilon^2}$$

$$\sum_{i=1}^{m} \frac{1}{\Delta_{t_i}^2} \leq \sum_{i=1}^{m}(1-\alpha)^{2i-2}\frac{1}{\epsilon^2} = \frac{1}{\alpha(2-\alpha)\epsilon^2}$$

- Hence, we have

$$T = \sum_{i=1}^{m} |J_i| \leq \frac{D_X^2 M_f^2}{(1-\alpha)^2\alpha(2-\alpha)\epsilon^2}.$$

- This concludes the proof.

Further Improvements

Restricted Memory Schemes

- Simple approach: shrink bundle size to $O(n)$ whenever $\Delta_t \to \Delta_t/2$
- Other approaches: Truncated Proximal Bundle-level method (TPBL), Non-Euclidean Restricted Memory Level Method (NERML)

Mirror Descent Schemes

$$x_{t+1} = \operatorname*{argmin}_{x \in X} \{\langle \gamma_t g_t, x \rangle + D_\omega(x, x_t)\}$$

where $D_\omega(x, y) = \omega(x) - \omega(y) - \nabla\omega(y)^T(x - y)$.

$$\mathcal{O}\left(\frac{\mathrm{Var}_{X,\|\cdot\|_2}(f)}{\sqrt{t}}\right) \Longrightarrow \mathcal{O}\left(\frac{\mathrm{Var}_{X,\|\cdot\|}(f)}{\sqrt{t}}\right)$$

References

- Ben-Tal & Nemirovski, *Modern Convex Optimization*, Chapter 5.3.2