Lower Bounds for Non-smooth Convex Optimization

$$f^* = \min_{\substack{x \in \mathbb{R}^n \\ \|x\| \le R}} f(x)$$

Theorem Let $M > 0$ and $R > 0$ be fixed. Then, for <u>any</u> first-order algorithm running for $K \ge 1$ iterations $\exists\ f: \mathbb{R}^n \to \mathbb{R}$, $n \ge K$ such that

1. f is convex, it is Lipschitz with $M > 0$.
2. For the output x_K we have:

$$f(x_k) - f^* \ge \frac{MR}{2\sqrt{K}} .$$

Proof. (continue)

Let $\delta > 0$ - arbitrarily small.

Fixed parameters $\xi_1, \dots, \xi_K \in \{-1, 1\}$

and a permutation

$$t \mapsto \sigma(t) \in \{1, \dots, n\}$$

<u>Family of objectives:</u>

$$f_n(x) = M \cdot \max_{1 \le i \le k} \left[\xi_i \underbrace{\langle e_{\sigma(i)}, x \rangle}_{x^{(\sigma(i))}} - (i-1)\delta \right]$$

At iteration $k \geq 1$: the method returns x_k

We choose:

$$\sigma(k) = \underset{\substack{1 \leq i \leq n \\ i \notin \{\sigma(1), \ldots, \sigma(k-1)\}}}{\operatorname{argmax}} |x_k^{(i)}|$$

$$\xi_k = \operatorname{sign} x_k^{(\sigma(k))}.$$

Observe that, for $s < k$:

$$f_k(x) = \max\{ f_s(x), M \cdot \max_{s < i \leq k} [\xi_i \langle e_{\sigma(i)}, x \rangle - (i-1)\delta] \}$$

$$\boxed{f_k(x_s) = f_s(x_s)}$$

$$f_s(x_s) = f_{s+1}(x_s) = \ldots = f_k(x_s)$$

For any $\|x - x_s\| \leq \delta$:

$$f_k(x) \equiv f_s(x). \Rightarrow \partial f_k(x_s) = \partial f_s(x_s)$$

1. $f_k^* \leq -\frac{MR}{\sqrt{k}}$.

2. $f(x) \equiv f_k(x)$ For the output:

$$f(x_k) = f_k(x_k) \geq \underbrace{\xi_k \langle e_{\sigma(k)}, x_k \rangle}_{\geq 0} - (k-1)\delta$$

$$\geq -(k-1)\delta$$

Finally: $$f(x_n) - f^* \geq \frac{MR}{\sqrt{K}} - (K-1)\delta \geq \frac{MR}{2\sqrt{K}}$$

$$\text{when} \quad \delta \leq \frac{MR}{2(K-1)\sqrt{K}}. \qquad \square$$

Consider one step:

$$f_{n+1}(x) = \max\{ f_n(x), M[\xi_{n+1}\langle e_{\sigma(n+1)}, x\rangle - K\delta]\}$$

$$\underline{f_n(x_n)} \geq \underline{M\xi_{n+1}\langle e_{\sigma(n+1)}, x_n\rangle} + M\delta \qquad \star$$

$$\varphi(x) = f_n(x) - M\xi_{n+1}\langle e_{\sigma(n+1)}, x\rangle$$ - Lipschitz with const $M > 0$.

$$\varphi(x_n) \geq +M\delta$$

$$|\varphi(x_n) - \varphi(x)| \leq M\|x_n - x\|$$

$$\varphi(x) \geq \varphi(x_n) - M\|x_n - x\|$$
$$\geq M\delta - M\delta = 0.$$

$$\Rightarrow \qquad f_n(x) \geq M\xi_{n+1}\langle e_{\sigma(n+1)}, x\rangle$$

Complexity of Non-smooth Convex Optimization

Interior-Point Algorithms

moderate size

$n \le 10^4$

low-dimensional optimization

$n \le 10$-20

Large-Scale optimization

$n = 1$

$\downarrow$

Binary search

$n \to +\infty$

Cutting Plane Schemes

Center of gravity: $O\left(n \log \frac{MR}{\varepsilon}\right)$

Ellipsoid Method: $O\left(n^2 \log \frac{MR}{\varepsilon}\right)$

Subgradient Method

$O\left(\left[\frac{MR}{\varepsilon}\right]^2\right)$

$\Rightarrow$ Subgradient Method is better than Ellipsoid,

$$n \ge \frac{MR}{\varepsilon}.$$

Smooth Convex Optimization: $O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$ FGM

Strongly convex Func: $O\left(\sqrt{\frac{L}{\mu}} \log \frac{LR^2}{\varepsilon}\right)$

Stochastic Optimization

Problem
$$\min_{x \in Q} f(x), \quad Q \subseteq \mathbb{R}^n \text{ - is a bounded convex set}$$

$$D = \operatorname{diam}(Q) = \max_{x,y \in Q} \|x-y\|$$

Denote by M - Lipschitz constant of f.

Stochastic Oracle

For any $x \in Q$, we can sample random variable ξ,

$g(x,\xi) \in \mathbb{R}^n$

stochastic substitute for the (sub)gradient.

1. Unbiased estimator of some subgradient

$$\mathbb{E}_{\xi}\, g(x,\xi) = f'(x) \in \partial f(x)$$

2. Bounded variance:

$$\mathbb{E}_{\xi} \|g(x,\xi) - f'(x)\|^2 \le \sigma^2, \quad \star$$

$\sigma > 0$ is a parameter of problem class.

$$\mathbb{E}\|g(x,\xi)\|^2 \le \sigma^2 + \|f'(x)\|^2 \le \sigma^2 + M^2.$$

Motivation

Subgradient Method with Normalized Stepsizes:

$$x_{k+1} = \pi_Q\left(x_k - \gamma \frac{f'(x_k)}{\|f'(x_k)\|}\right), \quad \gamma := \frac{D}{\sqrt{K+1}}$$

number of iterations

- We want to be flexible, K
- Stochastic Optimization?

Difficult to analyze

$$x_{k+1} = \pi_Q\left(x_k - \gamma \boxed{\frac{g_k}{\|g_k\|}}\right), \quad g_k = g(x_k, \xi_k)$$

small (depend on σ)

Algorithm: Stochastic Subgradient Method with Adaptive Stepsizes.

Init: $x_0 \in Q$, $S_0 = 0$.

For $k = 0 \dots K-1$ iterate:

1. Sample ξ_k, compute $g_k = g(x_k, \xi_k)$

2. Update: $S_{k+1} := S_k + \|g_k\|^2 = \|g_0\|^2 + \dots + \|g_k\|^2$

   Set $\beta_k = \frac{\sqrt{S_{k+1}}}{D}$.

3. Perform the Step:

$$x_{k+1} = \operatorname*{argmin}_{y \in Q} \left\{ \langle g_k, y - x_k \rangle + \frac{\beta_k}{2} \|y - x_k\|^2 \right\}$$

$$= \pi_Q\left(x_k - \frac{1}{\beta_k} g_k\right)$$

$$\beta_k = \frac{1}{D} \sqrt{\|g_0\|^2 + \dots + \|g_k\|^2}$$

Return: $\bar{x}_k = \frac{1}{K} \sum_{i=0}^{K-1} x_i$

Remark In Subgradient Method: $\approx \frac{\sqrt{K+1}}{D} \cdot \|f'(x_k)\|$

$$x_{k+1} = \pi_Q\left(x_k - \frac{1}{\alpha_k} f'(x_k)\right)$$

$$\alpha_k = \frac{\|f'(x_k)\| \cdot \sqrt{K+1}}{D}$$

Analysis

$$m_k(y) = \langle g_k, y - x_k \rangle + \frac{\beta_k}{2}\|y - x_k\|^2$$

$$m_k(y) \geq m_k(x_{k+1}) + \frac{\beta_k}{2}\|y - x_{k+1}\|^2 \quad \forall y.$$

$$m_k(x_{k+1}) = \min_{y \in Q} \langle g_k, y - x_k \rangle + \frac{\beta_k}{2}\|y - x_k\|^2$$
$$\geq \min_{h \in \mathbb{R}^n} \langle g_k, h \rangle + \frac{\beta_k}{2}\|h\|^2 = \underbrace{-\frac{\|g_k\|^2}{2\beta_k}}.$$

Hence, we get:

$$\underbrace{\frac{\|g_k\|^2}{2\beta_k}}_{\text{"Error"}} + \frac{\beta_k}{2}\|y - x_k\|^2 \geq \underbrace{\frac{\beta_k}{2}\|y - x_{k+1}\|^2}_{} + \underbrace{\langle g_k, x_k - y \rangle}_{\text{"Progress"}}$$

x^*

Note: $\mathbb{E}_{\xi_k} \langle g_k, x_k - x^* \rangle = \langle f'(x_k), x_k - x^* \rangle \geq f(x_k) - f^*$

AdaGrad-Norm Stepsizes

AdaGrad Method $\quad \beta_k \mapsto D_k$ - diagonal matrix

Lemma Let $\beta_{k+1} \geq \beta_k$. Then:

$$\boxed{\frac{\|g_k\|^2}{2\beta_k}} + \frac{\beta_k}{2}\|x^* - x_k\|^2 + \boxed{\frac{\beta_{k+1} - \beta_k}{2} D^2} \geq \underbrace{\frac{\beta_{k+1}}{2}\|x^* - x_{k+1}\|^2} + \langle g_k, x_k - x^* \rangle$$

error terms

$$\frac{\beta_{k+1}}{2}\|x^* - x_{k+1}\|^2 = \frac{\beta_k}{2}\|x^* - x_{k+1}\|^2 + \frac{\beta_{k+1} - \beta_k}{2}\underbrace{\|x^* - x_{k+1}\|^2}_{\leq D^2}.$$

$$(\beta_{k+1}-\beta_k)D^2 = \left(\sqrt{S_{k+2}}-\sqrt{S_{k+1}}\right)D =$$

$$= \frac{S_{k+2}-S_{k+1}}{\sqrt{S_{k+2}}+\sqrt{S_{k+1}}}\cdot D = \frac{\|g_{k+1}\|^2\cdot D}{\sqrt{S_{k+2}}+\sqrt{S_{k+1}}} \geq \frac{D\|g_{k+1}\|^2}{2\sqrt{S_{k+2}}} =$$

$$= \frac{\|g_{k+1}\|^2}{2\beta_{k+1}}$$

Telescoping :

$$\sum_{i=0}^{k-1}\langle g_i, x_i-x^*\rangle \leq \frac{\beta_0}{2}\overbrace{\|x^*-x_0\|^2}^{\leq D^2} + \frac{3}{2}D^2\cdot[\beta_K-\beta_0]$$

$$\leq 2D^2\beta_K.$$

1. $$\mathbb{E}\left[\sum_{i=0}^{k-1}\langle g_i, x_i-x^*\rangle\right] = \mathbb{E}\left[\sum_{i=0}^{k-1}\langle f'(x_i), x_i-x^*\rangle\right] \geq$$

$$\geq \mathbb{E}\left[\sum_{i=0}^{k-1} f(x_i)-f^*\right] \underset{\text{Jensen's ineq.}}{\geq} K\,\mathbb{E}\left[f(\bar{x}_k)-f^*\right]$$

2. $$\mathbb{E}\beta_k = \mathbb{E}\frac{\sqrt{\sum_{i=0}^{k-1}\|g_i\|^2}}{D} \leq \frac{\sqrt{\sum_{i=0}^{k-1}\mathbb{E}\|g_i\|^2}}{D} \leq \frac{\sqrt{K}\sqrt{\sigma^2+M^2}}{D}.$$

$$\leq \frac{\sqrt{K}(\sigma+M)}{D}$$

Finally :

$$\mathbb{E}\left[f(\bar{x}_k)-f^*\right] \leq \frac{D(\sigma+M)}{\sqrt{K}}.$$