Proof. Suppose to the contrary, some $\|g\|_2 > M$, $g \in \partial f(x)$

Look at $y = x + g$.

$$f(y) \geq f(x) + g^T(y - x)$$
$$= f(x) + \|g\|_2^2$$
$$> f(x) + \|g\|_2 \cdot M$$

$\Rightarrow f(x+g) - f(x) > M\|g\|_2$. $\times$ $\Rightarrow$ Bound subgrad. $\square$

<u>Theorem</u> For any convex f that is M-Lipschitz,

then for α_k s.t.

$$\sum_{i=0}^{\infty} \alpha_i = \infty, \quad \sum_{i=0}^{\infty} \alpha_i^2 \leq C$$

we have $\lim_{k\to\infty} \min_{i \leq k} \{f(x_i) - f(x^*)\} \to 0$.

(some subsequence converges in obj gap.)

or some $f(x_k) = f(x^*)$

Proof. By our Lemma,

$$0 \leq \|x_{k+1} - x^*\|_2^2 \leq \|x_k - x^*\|_2^2 - 2\alpha_k(f(x_k) - f(x^*)) + \alpha_k^2\|g_k\|_2^2$$

(induction) $\leq \underline{\|x_0 - x^*\|^2} - 2\sum_{i=0}^{k}\alpha_i(f(x_i) - f(x^*)) + \sum_{i=0}^{k}\alpha_i^2\|g_i\|_2^2$

$$\Rightarrow \frac{\sum_{i=0}^{k}\alpha_i(f(x_i) - f(x^*))}{\sum_{i=0}^{k}\alpha_i} \leq \frac{\|x_0 - x^*\|^2 + \sum \alpha_i^2\|g_i\|_2^2}{2\sum_{i=0}^{k}\alpha_i}$$

$\min_{i \leq k}\{f(x_i) - f(x^*)\} \leq$ Weighted Average of gaps.

Want $\searrow 0$.

$$\min_{i \le k} \{ f(x_i) - f(x^*) \} \le \frac{\|x_0 - x^*\|^2 + M^2 \sum_{i=0}^{k} \alpha_i^2}{2 \sum_{i=0}^{k} \alpha_i}$$ (by 2nd Lemma)

$\to 0$ whenever $\sum^{\infty} \alpha_i = \infty$

$\sum^{\infty} \alpha_i^2 \le C$

$(\alpha_k = \frac{1}{k}, \quad C = \frac{\pi^2}{6})$

Finite-Time Guarantees.

Constant $\alpha_k = \alpha$.

$$\Rightarrow \min_{i \le k} \{ f(x_i) - f(x^*) \} \le \frac{\|x_0 - x^*\|^2 + M^2 \alpha^2 \cdot k}{2 \alpha k}$$

$$= \frac{\|x_0 - x^*\|^2}{2 \alpha k} + \frac{M^2 \alpha}{2}.$$

Want LHS $\le \varepsilon$, need $\frac{\|x_0 - x^*\|^2}{2 \alpha k} + \frac{M^2 \alpha}{2} = \varepsilon$

aim $\frac{\varepsilon}{2} + \frac{\varepsilon}{2}$

$\frac{M^2 \|x_0 - x^*\|^2}{2 \varepsilon \cdot k}$ Need $\boxed{\alpha = \frac{\varepsilon}{M^2}}$

Need $\boxed{k = \frac{M^2 \|x_0 - x^*\|^2}{\varepsilon^2}}.$

For comparison, GD: $f(x_k) - f(x_i^*) \le C/k$

$\Rightarrow k \ge C/\varepsilon.$

$\Leftrightarrow \varepsilon \le \frac{M \|x_0 - x^*\|}{\sqrt{k}}.$

Accel: $\Rightarrow k = \sqrt{C/\varepsilon}.$

Sadly this is as good as possible.

"Theorem" = there exists a bad function $f: \mathbb{R}^d \to \mathbb{R}$, and subgradient oracle $g(x) \in \partial f(x)$ s.t.

using $k+1 = d$, after k steps

$$f(x_k) - f(x^*) \geq \frac{M \, \|x_0 - x^*\|}{2(1+\sqrt{k+1})} \text{"}.$$

Proof. Nesterov's Convex OPT book.

$$f(x) = \max\{x_1, \dots, x_d\} + \frac{M}{2}\|x\|_2^2 .$$

$\Rightarrow$ The subgradient method is optimal (but still slow).

Extra Directions

Speedup under Strong Convexity: $O(1/k)$

Nonconvex Guarantees [Davis, 2018] $O\left(\frac{1}{k^{1/4}}\right)$

↑ adviser

Overview So Far

1. Necessary + Sufficient Optimality Cond

2. First-Order Methods

   Smooth OPT

   (linesearchs,
   GD,
   Accel)

   Equiv Characterizations for Smooth, Convex,
   Strongly Convex

   Proximal First-Order Methods

   (Mirrors smooth setting)

   Nonsmooth OPT more generally

   Subgradients

   Subgrad Method is optimal.

Midterm

   Stochastic OPT

3. Second-Order Methods

   Newton

   "Quasi-Netwon"