Lecture 13

HW3 due Thursday
Midterm posted on Friday Morning
Scribe for Today?

Last time

- Guarantees for strongly convex
- Accelerated Forward Backward Method.
- More proximal methods
- Alternating Projections

Today

- Black-box convex optimization
- Things that break
- Analysis

Black-box convex optimization

What happens when we cannot solve for the prox?

Now we only assume that we are given a problem

$$\min_{x \in \mathbb{R}^d} f(x) \leftarrow \text{convex } f: \mathbb{R}^d \to \mathbb{R}$$

and that we can query for any x $f(x)$ and $g(x) \in \partial f(x)$.

We already saw a problem like this

in HW3:

$$\min_w \sum_i \max\{0, 1 - y_i x_i^T w\} + \frac{\lambda}{2}\|w\|^2$$

where computing a subgradient was easy, but solving the prox was hard.

A natural idea is to generalize GD

$$x_{k+1} \leftarrow x_k - \alpha_k g(x_k).$$

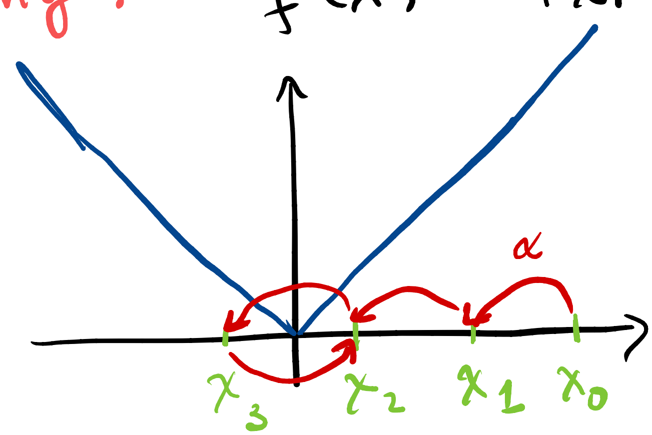
Things that break

Smooth optimization land was rather nice. In nonsmooth optimization we cannot have:

Guarantees with constant stepsize

Why? $f(x) = |x|$ $x_0 = 2.5\alpha$

Fixed step size

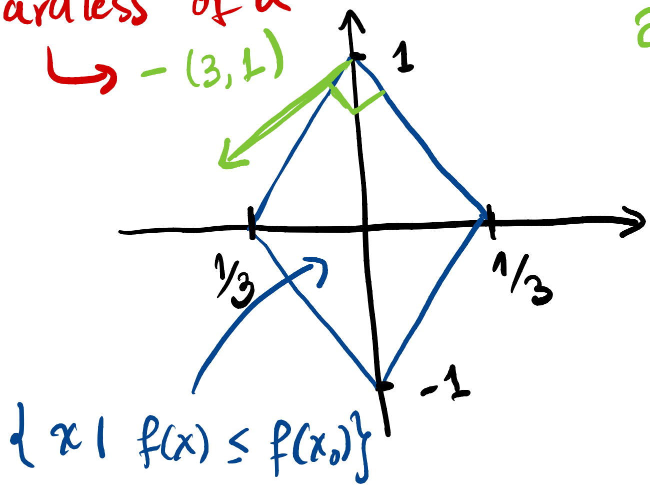


No guarantee of descent

Why? $f(x_1, x_2) = 3|x_1| + |x_2|$

with $x_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

No descent regardless of α



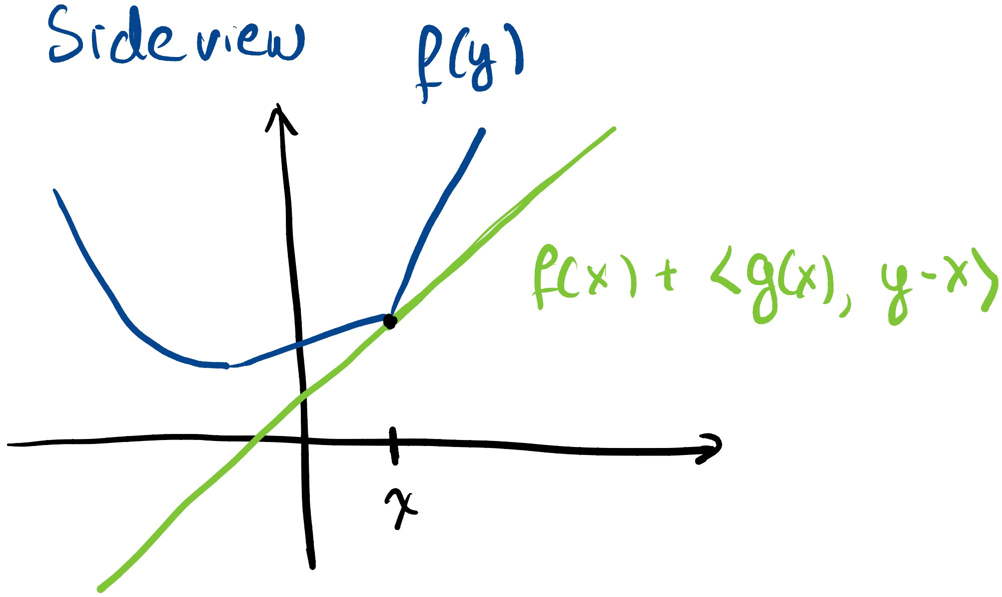
$$\partial f(0,1) = 3\partial(|x_1|)(0,1) + \partial(|x_2|)(0,1)$$

$$= \begin{bmatrix} 3[-1,1] \\ 1 \end{bmatrix}$$

$$\Rightarrow (3,1) \in \partial f(0,1)$$

Two perspectives on subgradients

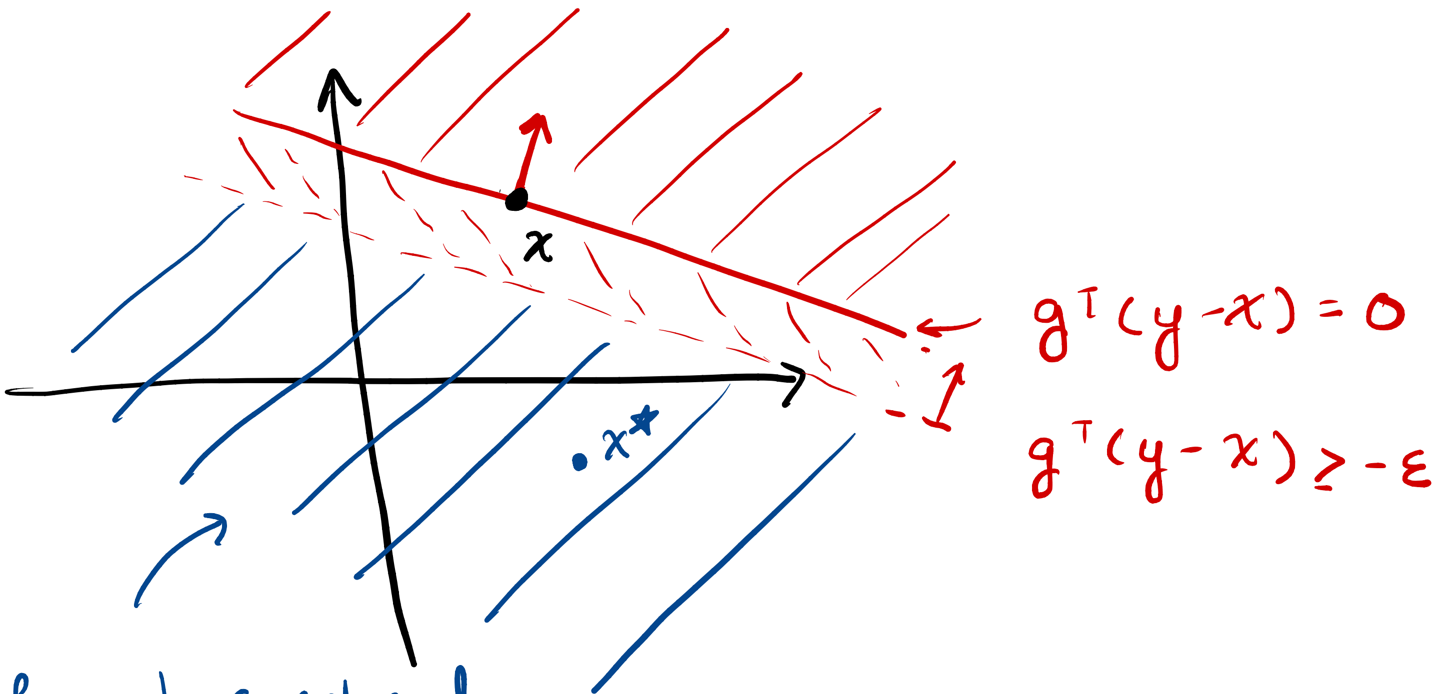
Sideview



We can also use this perspective to derive

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \left\{ f(x_k) + \langle g(x_k), x - x_k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}$$

Contour / overhead



If not ε-optimal at x, then optimal is here.

If $f(x) - \min f > \varepsilon \Rightarrow f(x) - \varepsilon > \min f$

if x' is such $g^T(y-x) \geq -\varepsilon \Rightarrow$

$f(x') \geq f(x) - \varepsilon > \min f$.

Lemma Assume that $f: \mathbb{R}^d \to \mathbb{R}$ is convex achieving a minimum at x^*. Then the iterates of subgradient descent satisfy

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

Proof: By definition

$$\|x_{k+1} - x^*\|^2 = \|x_k - \alpha_k g_k - x^*\|^2$$

$$= \|x_k - x^*\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle + \alpha_k^2 \|g_k\|^2$$

Atlas

$$\leq \|x_k - x^*\|^2 - 2\alpha_k (f(x_k) - f(x^*)) + \alpha_k^2 \|g_k\|^2.$$

$\square$

Intuition

We will get closer to the solution if

$$-2\alpha_k (f(x_k) - f(x^*)) + \alpha_k^2 \|g_k\|^2 < 0.$$

We can achive that if $\|g_k\|$ is bounded.

Lemma. If f is M-Lipschitz, then for all $x \in \mathbb{R}^d$, $g \in \partial f(x)$,

$$\|g\|_2 \leq M.$$

Proof: Seeking contradiction assume $\|g\|_2 > M$ for some $g \in \partial f(x)$. Then, if we take $y = x + g$

$$f(y) \geq f(x) + g^T(y - x)$$
$$\geq f(x) + \|g\|^2$$

$$\geq f(x) + \|g\| M.$$

Thus, $f(y) - f(x) \geq M\|g\| = M\|y - x\|.$

$\Downarrow$ $\square$

Exercise: Prove that the opposite implication in the previous Lemma also holds.

Theorem: Assume that $f: \mathbb{R}^d \to \mathbb{R}$ is an M-Lipschitz function, and suppose $x^* \in \operatorname{argmin} f(x)$. Then, the iterates of subgradient descent satisfy

$$\min_{k \leq T} \{f(x_k) - \min f\} \leq \frac{\|x_0 - x^*\|^2 + L^2 \sum_{k=0}^{T} \alpha_k^2}{2 \sum_{k=0}^{T} \alpha_k}.$$

In particular, if $\sum^{\infty} \alpha_k^2 < \infty$ and $\sum^{\infty} \alpha_k = \infty$, then

$$\lim_{T \to \infty} \min_{k \leq T} \{f(x_k) - \min f\} = 0.$$

Proof: For any k we have

$$2\alpha_k (f(x_k) - f(x^*)) \overset{\text{First Lemma}}{\leq} \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha_k^2 \|g_k\|^2$$

second lemma

$$\leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + L^2 \alpha_k^2.$$

Summing up for $k \leq T$

$$2\sum \alpha_k \left(f(x_k) - f(x^*)\right) \leq \|x_0 - x^*\|^2 + L^2 \sum \alpha_k^2$$

Lower bounding by $\min_{k \leq T} \left(f(x_k) - f(x^*)\right)$, yields

$$\min_{k \leq T} f(x_k) - f(x^*) \leq \frac{\|x_0 - x^*\|^2 + L^2 \sum^{T} \alpha_k^2}{2 \sum^{T} \alpha_k}.$$

Taking limits on both sides gives

$$\lim_{T \to \infty} \min_{k \leq T} f(x_k) - f(x^*) \leq \frac{\|x_0 - x^*\|^2 + L^2 \sum^{\infty} \alpha_k^2}{2 \sum^{\infty} \alpha_k},$$

when $\sum^{\infty} \alpha_k = \infty$ and $\sum^{\infty} \alpha_k^2 < \infty$, the right hand side goes to zero $\square$

<u>Corollary</u>: If we set $\alpha_k = \alpha$, then

$$\min_{k \leq T} \{ f(x_k) - \min f \} \leq \frac{\|x_0 - x^*\|^2}{2\alpha T} + \frac{M^2 \alpha}{2}.$$

If we set $\alpha = \varepsilon / M^2$ and $T \geq \frac{M^2 \|x_0 - x^*\|^2}{\varepsilon^2}$, then

$$\min \{ f(x_k) - \min f \} \leq \varepsilon .$$

Proof: First inequality follows trivially from the Theorem. Then

$$\frac{\|x_0 - x^*\|^2}{2\alpha T} + \frac{M^2 \alpha}{2} \overset{\alpha}{=} \frac{\|x_0 - x^*\|}{2 \varepsilon T} M^2 + \frac{\varepsilon}{2}$$

$$\overset{T}{\leq} \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon . \quad \square$$

Thus we need $T = \Omega(\frac{1}{\varepsilon^2})$ for an ε-min.
With GD we needed $T = \Omega(\frac{1}{\varepsilon})$
and with AGD we needed $T = (\frac{1}{\sqrt{\varepsilon}})$.

Theorem There exists a convex M-Lipschitz function $f: \mathbb{R}^d \to \mathbb{R}$ and a subgradient oracle $g(x) \in \partial f(x)$ s.t. any algorithm s.t.

$$x_{k+1} \in x_0 + \operatorname{span} \{ g(x_0), \dots, g(x_k) \}$$

satisfies that for $k < d$

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

You can find the proof in Nesterov's Book (Theorem 3.2.1)

Extensions

There are results for

- Strongly convex functions $O(\frac{1}{\varepsilon})$
- Weakly convex functions $O(\frac{1}{\varepsilon^4})$.