Lecture 9

HW 2 due Friday.

Last time

- Accelerated gradient descent.

Today

- Lower bounds
- Review of smooth optimization
- Structured nonsmooth optimization

Lower bounds

Assumption: The given method produces iterates satisfying

Subspace spanned by

$$x_k \in x_0 + \text{span}\{\nabla f(x_0), \dots, \nabla f(x_{k-1})\}$$

Dimension dependent

Theorem For any $1 \le k \le \frac{1}{2}(d-1)$ and $L \ge 0$, there exists a function $f: \mathbb{R}^d \to \mathbb{R}$ with L-Lips grad such that for any algo. satisfying Assumption 1, we have

$$f(x_k) - \min f \ge \frac{3L \|x_0 - x^*\|^2}{32(k+1)^2}$$

$$\|x_k - x^*\|^2 \geq \frac{1}{2}\|x_0 - x^*\|^2.$$

Proof: Next, we will build "the worst function in the world."

Let

$$A_k = \left[\begin{array}{c|c} \overbrace{\begin{matrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & & & \\ & & & \ddots & & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{matrix}}^{k} & \overbrace{}^{d-k} \\ \hline & 0 \end{array}\right]$$

Let

$$f_k(x) = \frac{L}{4}\left[\frac{x^T A_k x}{2} - e_1^T x\right].$$

By the HW 1

$$\nabla f(x) = \frac{L}{4}\left[A_k x - e_1\right],$$

$$\nabla^2 f(x) = \frac{L}{4} A_k.$$

WLOG we take x_0, otherwise we could define $\tilde{f}_k(x) = f_k(x - x_0)$.

Intuition

If $x_0 = 0$, then x_i can only have the first ith components being nonzero. But we will see that the solution x^* has nonzeros in its first k entries.

Claim 1: Any algo satisfying

$$x_i \in \text{span}\{\nabla f_k(x_0), \dots, \nabla f_k(x_{i-1})\},$$

has $\text{span}\{\nabla f_k(x_0), \dots, \nabla f_k(x_i)\} \subseteq \mathbb{R}^{i+1} \times \{0\}^{d-i-1}$ for all $i \leq k$.

Proof Claim 1: We use induction

Base case: $i = 0 \Rightarrow \nabla f(x_0) = -\frac{L}{4} e_1$. ✓

Inductive case: Assume it holds for $i-1$

$$\Rightarrow \nabla f_k(x_{i-1}) = \frac{L}{4}[A x_{i-1} - e_1]$$

$$\in \frac{L}{4} A \cdot \text{span}\{\nabla f_k(x_\ell)\}_{\ell=0}^{i-1}$$

Since A_k is tridiagonal (check!)

$$\subseteq \frac{L}{4} A \cdot \mathbb{R}^i \times \{0\}^{d-i}$$

$$= \frac{L}{4} \mathbb{R}^{i+1} \times \{0\}^{d-i-1}.$$

$\square$

Claim 2: The function f_K is convex and have L-Lipschitz gradients.

Proof: By our characterizations these amounts to showing

$$0 \le \lambda_{min}(\nabla^2 f_K(x)) \le \lambda_{max}(\nabla f_K(x)) \le L$$

$\frac{L}{4}A_K$

clearly positive

$$\Rightarrow \quad s A_K s = \frac{L}{4}\left[(s_{(i)})^2 + \sum_{i=1}^{K-1}(s_{(i)} - s_{(i+1)})^2 + (s_{(K)})^2\right]$$

$$\le \frac{L}{4}\left[s_{(i)}^2 + 2\sum_{i=1}^{K-1}(s_{(i)}^2 + s_{(i+1)}^2) + s_{(k)}^2 \right]$$

$$\le \frac{L}{4}\sum_{i=1}^{K} 4 s_{(i)}^2$$

$$\le L \|s\|^2 \qquad \square$$

Claim 3: The vector $\bar{x}$ with entries

$$\bar{x}_{(i)} = \begin{cases} 1 - \frac{i}{K+1} & i \in \{1, \dots, K\}, \\ 0 & \text{otherwise}, \end{cases}$$

satisfies $\nabla f_k(\bar{x}) = 0$.

Proof: Follows by verifying $A_k \bar{x} = e_1$ $\square$
(check!)

Therefore,

$$
\begin{aligned}
\min f_k &= f_k(\bar{x}) \\
&= \frac{L}{4}\left(\frac{1}{2}\bar{x}^T A_k \bar{x} - e_1^T \bar{x}\right) \\
&= \frac{L}{4}\left(\frac{1}{2} e_1^T \bar{x} - e_1^T \bar{x}\right) \\
&= -\frac{L}{8} e_1^T \bar{x} \\
&= -\frac{L}{8}\left(1 - \frac{1}{k+1}\right).
\end{aligned}
$$

$$
\|\bar{x}\|^2 = \sum_{i=1}^{k}\left(1 - \frac{i}{k+1}\right)^2 = \frac{1}{(k+1)} \sum_{i=1}^{k} (k-i+1)^2
$$

$$
= \frac{1}{(k+1)} \sum_{i=1}^{k} i^2 \overset{\text{sum of } k \text{ squares}}{=} \frac{1}{(k+1)^2} \frac{k(k+1)(2k+1)}{6}
$$

$$
\le \frac{2k+1}{6} \le \frac{k+1}{3}.
$$

Armed with these facts we can now prove the lower bound.

For any fixed k, set $d = 2k+1$ and

$$f(x) = f_{2k+1}(x).$$

Let x_k be the output of an algo satisfying Assumption 1. Then

$$f(x_k) = f_{2k+1}(x_k) \underset{\text{Claim 1}}{=} f_k(x_k) \geq \min f_k$$

Then,

$$\frac{f(x_k) - \min f}{\|x_0 - \bar{x}\|^2} \geq \frac{\min f_k - \min f_{2k+1}}{\|\bar{x}\|^2}$$

$\bar{x} \in \operatorname{argmin} f$

(☺) & (♡)

$$\geq \frac{\frac{L}{8}\left(\cancel{1} + \frac{1}{k+1} - \cancel{1} - \frac{1}{2k+2}\right)}{(2k+2)/3}$$

$$= \frac{3L}{8} \frac{(\overbrace{2k+2-k-1}^{k+1})}{(2k+2)^2(k+1)}$$

$$\geq \frac{3L}{32}\,\frac{1}{(k+1)^2}.$$

To prove the second part of the theorem, let's lower bound

$$\|x_k - \underset{\text{argmin } f_{2k+1}}{\bar{x}}\|^2 \overset{\text{Claim 1}}{\geq} \sum_{i=k+1}^{2k+1} (\bar{x}_{(i)})^2 = \sum_{i=k+1}^{2k+1}\left(1-\frac{i}{2k+2}\right)^2 \qquad \frac{1}{2k+2}\sum_{i=k+1}^{2k+1}(2k+2-i)^2$$

$$= \frac{1}{(2k+2)^2}\sum_{i=1}^{k+1} i^2$$

$$= \frac{1}{6\,(2\cancel{k+2})^2}\,(\cancel{k+1})(\cancel{k+2})(2k+2)$$

By (♡)

$$\geq \frac{1}{3\cdot 2}\,(2k+2)$$

$$\geq \frac{1}{2}\,\|x_0 - \bar{x}\|^2.$$

$\square$

Summary of guarantees for smooth optimazation.

So far we have proved the following table of results

Method	Generic rate (L-smooth)	Quadratic growth
Gradient Descent (for nonconvex f)	$\frac{1}{T}\sum_{k=0}^{T-1} \|\nabla f(x_k)\|^2 \le \Theta\left(\frac{1}{T}\right)$	$f(x_T) - f(x^*) \le \Theta\left(\left(1 - \frac{\mu^2}{4L^2}\right)^T\right)$ (Local rate for $\nabla f(x^*) \succ 0$)
Gradient Descent (for convex f)	$f(x_T) - \min f \le \Theta\left(\frac{1}{T}\right)$	$f(x_T) - \min f \le \Theta\left(\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2T}\right)$ (μ-strongly convex)
Accelerated Gradient (for convex f)	$f(y_T) - \min f \le \Theta\left(\frac{1}{T^2}\right)$ Optimal	$f(x_T) - \min f \le \Theta\left(\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^{2T}\right)$ (μ-strongly convex) HW2 P3 (Also optimal).

What's next? Structured nonsmooth optimization

1. Motivating problems
2. The proximal operator
3. Constraints and projections
4. Proximal gradient method
5. Acceleration
6. More proximal methods.