Lecture 9
HW 2 due Friday.
Scribe.

Last time
- Accelerated gradient descent.
- Lower bounds

Today
- Proof lower bound
- Review of smooth optimization
- Structured nonsmooth optimization

Lower bounds continued

Assumption: The given method produces iterates satisfying

\[ x_k \in x_0 + \text{span}\{ \nabla f(x_0), \ldots, \nabla f(x_{k-1}) \} \]

Subspace spanned by Dimension dependent

Theorem: For any \( 1 \leq k \leq \frac{1}{2}(d-1) \) and \( x_0 \), there exists a function \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) with \( L \)-Lips grad such that for any algo. satisfying Assumption 1, we have

\[
 f(x_k) - \min f \geq \frac{3L \| x_0 - x^* \|^2}{32 (k+1)^2}
\]
\[ \| x_k - x^* \|^2 \geq \frac{1}{2} \| x_0 - x^* \|^2. \]

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

Let

\[ A_k = \begin{pmatrix} 2 & -1 & \cdots & -1 \\ -1 & 2 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & 2 \end{pmatrix}. \]

Let

\[ f_k(x) = \frac{1}{4} \left[ x^T A_k x - e_1^T x \right]. \]

By the HW 1

\[ \nabla f(x) = \frac{1}{4} \left[ A_k x - e_1 \right], \]

\[ \nabla^2 f(x) = \frac{1}{4} A_k. \]

WLOG we take \( x_0 \), otherwise we could define \( f_k(x) = f_k(x - x_0) \).
Intuition

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

Claim 1: Any algo satisfying

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

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

Proof Claim 1: We use induction

Base case: $i = 0 \Rightarrow \nabla f(x_0) = -\frac{1}{4} e_1$.

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

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

$$\in \frac{L}{4} A \cdot \text{span}\{ \nabla f_k(x_2) \}_{j=0}^{i-1}$$

Since $A_k$ is tridiagonal (check!)

$$\leq \frac{L}{4} A \cdot \mathbb{R}^{i} \times \log^{d-i} y$$

$$= \frac{L}{4} \mathbb{R}^{i+1} \times \log^{d-i-1} y.$$
Claim 2: The function $f_k$ is convex and have $L$-Lipschitz gradients.

Proof: By our characterizations these amounts to showing

$$0 \leq \lambda_{\min} (\nabla^2 f_k(x)) \leq \lambda_{\max} (\nabla^2 f_k(x)) \leq L$$

Clearly positive

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

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

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

$$\leq L \|s\|^2$$

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

$$\vec{x}_{(i)} = \begin{cases} 1 - \frac{i}{K+1} & \text{if } i \in \{1, \ldots, K\}, \\ 0 & \text{otherwise}, \end{cases}$$
satisfies \( \forall f_k(x) = 0 \).

**Proof:** Follows by verifying \( A_k \bar{x} = e_1 \)

(check!)

Therefore,

\[
\min f_k = f_k(\bar{x}) \\
= \frac{L}{4} \left( \frac{1}{2} \bar{x}^T A_k \bar{x} - e_1^T \bar{x} \right) \\
= \frac{L}{4} \left( \frac{1}{2} e_1^T \bar{x} - e_1^T \bar{x} \right) \\
= -\frac{L}{8} e_1^T \bar{x} \\
= -\frac{L}{8} \left( 1 - \frac{1}{k+1} \right).
\]

\[
\| \bar{x} \|_2^2 = \sum_{i=1}^{K} \left( 1 - \frac{i}{k+1} \right)^2 = \frac{1}{(k+1)} \sum_{i=1}^{K} (k-i+1)^2 \sum \text{ of } k \text{ squares} \\
= \frac{1}{(k+1)} \sum_{i=1}^{K} i^2 = \frac{1}{(k+1)^2} \frac{k \cdot (k+1) \cdot (2k+1)}{6} \\
\leq \frac{2 \cdot k+1}{6} \leq \frac{k+1}{3}.
\]
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

\[
f(x_k) = f_{2k+1}(x_k) = f_k(x_k) \geq \min f_k
\]

Claim 1

Then,

\[
f(x_k) - \min f \geq \frac{\min f_k - \min f_{2k+1}}{\|x - x\|^2} \times \arg \min_{x \in X} f
\]

\[
\geq \frac{1}{8} \left( \left( 1 + \frac{1}{2k+1} \right) - \frac{1}{2k+1} \right)
\]

\[
= \frac{3}{8} \left( \frac{2k+2 - k - 1}{(2k+2)^2(k+1)} \right)
\]
To prove the second part of the theorem, let's lower bound

\[ \| x_k - \bar{x} \|^2 \geq \sum_{i=k+1}^{2k+1} (\bar{x}_i)^2 = \sum_{i=k+1}^{2k+1} \left( 1 - \frac{i}{2k+2} \right)^2 \]

\[ \begin{align*}
\text{Claim 1} & \quad \text{argmin}_{f_{2k+1}} \\
\geq & \quad \frac{1}{2} \sum_{i=1}^{k+1} i^2 \\
= & \quad \frac{1}{6} (k+1)(k+2)(2k+2) \\
\geq & \quad \frac{1}{2} (2k+2) \\
\geq & \quad \frac{1}{2} \| x_0 - x \|^2.
\end{align*} \]

By (10)

Summary of guarantees for smooth optimization.

So far we have proved the following table of results.
<table>
<thead>
<tr>
<th>Method</th>
<th>Generic rate (L-smooth)</th>
<th>Quadratic growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient Descent (for nonconvex $f$)</td>
<td>$\frac{1}{T} \sum_{k=0}^{T-1} | \nabla f(x_k) |^2 \leq \Theta(\frac{1}{T})$</td>
<td>$f(x_T) - f(x^<em>) \leq \Theta(\frac{1}{\sqrt{T}})$ (Local rate for $\nabla f(x^</em>) &gt; 0$)</td>
</tr>
<tr>
<td>Gradient Descent (for convex $f$)</td>
<td>$f(x_T) - \min f \leq \Theta(\frac{1}{T})$</td>
<td>$f(x_T) - \min f \leq \Theta(\frac{2}{\sqrt{\mu T}})$ ((\mu)-strongly convex)</td>
</tr>
<tr>
<td>Accelerated Gradient (for convex $f$)</td>
<td>$f(y_T) - \min f \leq \Theta(\frac{1}{T^2})$</td>
<td>$f(x_T) - \min f \leq \Theta(\frac{1}{\sqrt{\mu T}})$ ((\mu)-strongly convex)</td>
</tr>
</tbody>
</table>

Optimal

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