Introduction to Optimization Theory

Lecture #5 - 9/28/20
MS&E 213 / CS 2690

Aaron Sidford
sidford@stanford.edu
Plan for Today

Recap
- Gradient descent for smooth function
- Notions of convexity

Convexity
- Prove smoothness / convexity equivalences
- Example functions
- Implications of assumptions

Algorithm
- Gradient descent
- Algorithm analysis
### Recap

<table>
<thead>
<tr>
<th>Regularity</th>
<th>Oracle</th>
<th>Goal</th>
<th>Algorithm</th>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 1, f(x) \in [0,1], x_* \in [0,1]$</td>
<td>value</td>
<td>$\frac{1}{2}$-optimal</td>
<td>anything</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$n = 1, x_* \in [0,1], L$-Lipschitz</td>
<td>value</td>
<td>$\epsilon$-optimal</td>
<td>$\epsilon$-net</td>
<td>$\Theta(L/\epsilon)$</td>
</tr>
<tr>
<td>$x_* \in [0,1]^n, L$-Lipschitz in $| \cdot |_\infty$</td>
<td>value</td>
<td>$\epsilon$-optimal</td>
<td>$\epsilon$-net</td>
<td>$(\Theta(L/\epsilon))^n$</td>
</tr>
<tr>
<td>$L$-smooth and bounded</td>
<td>value, gradient</td>
<td>$\epsilon$-optimal</td>
<td>$\epsilon$-net</td>
<td>exponential</td>
</tr>
<tr>
<td>$L$-smooth</td>
<td>gradient</td>
<td>$\epsilon$-critical</td>
<td>gradient descent</td>
<td>$O \left( \frac{L(f(x_0) - f_*)}{\epsilon^2} \right)$</td>
</tr>
</tbody>
</table>

### Today

*What if want an $\epsilon$-optimal point with no dependence on dimension?*
Assumptions for Efficient $\varepsilon$-optimal Point

**Notion #1**: Hessian Lower Bound

- $f$ is twice differentiable and $z^T \nabla^2 f(x) z \geq \mu \|z\|^2$ for all $x, z$
- $\iff \lambda_{\min}(\nabla^2 f(x)) \geq \mu$

**Variational characterization of eigenvalues?**

**Notion #2**: Quadratic Lower Bounds

- $f$ is differentiable and $f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2} \|y-x\|^2 \overset{\text{def}}{=} L_y(x)$

**Notion #3**: $\mu$-strongly convex with respect to $\| \cdot \|$ (by default $\| \cdot \|_2$)

- $f(ty + (1-t)x) \leq t \cdot f(y) + (1-t) \cdot f(x) - \frac{\mu}{2} t(1-t) \|y-x\|^2$

For all $x, y$ and $t \in [0,1]$

Say $f$ is convex $\iff f$ is 0-strongly convex

**Theorem**

These three notions are equivalent for twice differentiable functions
Plan for Today

Recap
- Gradient descent for smooth function
- Notions of convexity

Convexity
- Smoothness / convexity equivalences
- Example functions
- Implications of assumptions

Algorithm
- Gradient descent
- Algorithm analysis
Equivalent Notions of Convexity

**Notion #1**: Hessian Lower Bound
- $f$ is twice differentiable and $z^T \nabla^2 f(x) z \geq \mu \|z\|_2^2$ for all $x, z$
- $\iff \lambda_{\min}(\nabla^2 f(x)) \geq \mu$

**Notion #2**: Quadratic Lower Bounds
- $f$ is differentiable and $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2 \overset{\text{def}}{=} L_y(x)$

**Notion #3**: $\mu$-strongly convex with respect to $\| \cdot \|$ (by default $\| \cdot \|_2$)
- $f(ty + (1 - t)x) \leq t \cdot f(y) + (1 - t) \cdot f(x) - \frac{\mu}{2} t(1 - t)\|y - x\|^2$
  For all $x, y$ and $t \in [0,1]$

Say $f$ is convex $\iff f$ is 0-strongly convex
Helpful Technical Lemma

For all $\alpha, \beta \in \mathbb{R} \cup \{\pm \infty\}$ and twice differentiable $f : \mathbb{R}^n \to \mathbb{R}$ and any norm $\| \cdot \|$ the following three conditions are equivalent:

- $\frac{\alpha}{2} \| x - y \|^2 \leq f(y) - [f(x) + \nabla f(x)^T (y - x)] \leq \frac{\beta}{2} \| x - y \|^2$ for all $x, y \in \mathbb{R}^n$

- $\frac{\alpha}{2} \| x - y \|^2 \leq (\nabla f(x) - \nabla f(y))^T (x - y) \leq \frac{\beta}{2} \| x - y \|^2$ for all $x, y \in \mathbb{R}^n$

- $\alpha \| z \|^2 \leq z^T \nabla^2 f(x) z \leq \beta \| z \|^2$ for all $x, z \in \mathbb{R}^n$

- Implies equivalence of upper and lower bounds implied by smoothness and Hessian eigenvalue bound.
- Implies some convexity equivalences

Proof is technical (but useful) and in smoothness notes.
Corollary

For all $\mu \geq 0$ and twice differentiable $f : \mathbb{R}^n \to \mathbb{R}$ and any norm $\| \cdot \|$ the following three conditions are equivalent:

• $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \| x - y \|^2$ for all $x, y \in \mathbb{R}^n$

• $(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{\mu}{2} \| x - y \|^2$ for all $x, y \in \mathbb{R}^n$

• $z^T \nabla^2 f(x) z \geq \mu \| z \|^2$ for all $x, z \in \mathbb{R}^n$

When $f$ is differentiable the first two conditions are equivalent to

$f(ty + (1 - t)x) \leq t \cdot f(y) + (1 - t) \cdot f(x) - \frac{\mu}{2} t(1 - t) \| y - x \|^2$

For all $x, y$ and $t \in [0,1]$. (See Note.)
What do $\mu$-strongly convex functions look like?

- $f(ty + (1 - t)x) \leq t \cdot f(y) + (1 - t) \cdot f(x) - \frac{\mu}{2} t(1 - t)\|x - y\|_2^2$
Example Convex Functions

\[ f(x) = \frac{1}{2} ||Ax - b||^2 \]

\[ f(x) = ||x|| \]

\[ f(x) = \exp(x) \]

\[ f(x) = x^p \text{ for even } p \]

\[ f(x) = -\log x \text{ for } x \geq 0 \]

\[ f(x) = x \log x \]

\[ f(x) = g(x) + h(x) \text{ for convex } g \text{ and } h \]

\[ f(x) = f(Ax) \text{ for convex } f \]

\[ f(x) = c \cdot f(x) \text{ for convex } f \text{ and } c \geq 0 \]

\[ \ldots \]
Goal: Minimize Smooth Convex Functions

**Theorem:** $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is $L$-smooth and $\mu$-strongly convex (with respect to $\| \cdot \|_2$) if and only if the following hold for all $x, y$

- $f(y) \leq U_x(y) \overset{\text{def}}{=} f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \| y - x \|_2^2$
- $f(y) \geq L_x(y) \overset{\text{def}}{=} f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \| y - x \|_2^2$

**Question:** is this assumption and a gradient oracle enough to obtain dimension independent efficient algorithms for $\epsilon$-optimal points?
Plan for Today

Recap
- Gradient descent for smooth function
- Notions of convexity

Convexity
- Smoothness / convexity equivalences
- Example functions
- Implications of assumptions

Algorithm
- Gradient descent
- Algorithm analysis
Goal: Minimize Smooth Convex Functions

Theorem: \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is \( L \)-smooth and \( \mu \)-strongly convex (with respect to \( \| \cdot \|_2 \)) if and only if the following hold for all \( x, y \)

\[
\begin{align*}
\bullet & \quad f(y) \leq U_x(y) \overset{\text{def}}{=} f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \| y - x \|_2^2 \\
\bullet & \quad f(y) \geq L_x(y) \overset{\text{def}}{=} f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \| y - x \|_2^2
\end{align*}
\]

Question: is this assumption and a gradient oracle enough to obtain dimension independent efficient algorithms for \( \epsilon \)-optimal points?
Algorithm?

Gradient Descent!

• For \( t = 0, \ldots, T - 1 \)
  • \( x_{t+1} = x_t - \frac{1}{L} \nabla f(x_t) \)
• Output \( x_T \)

• Goal: compute \( \epsilon \text{-optimal point} \)
• Assumption \( f: \mathbb{R}^n \to \mathbb{R} \) is \( L \)-smooth and \( \mu \)-strongly convex
• Given: \( x_0 \in \mathbb{R}^n \) and a gradient oracle

Upper Bound Analysis

• \( f(x_{t+1}) \leq U_{x_t}(x_{t+1}) \)
• \( U_{x_t}(x_{t+1}) = f(x_t) - \frac{1}{2L} \left\| \nabla f(x_t) \right\|^2 \)
• \( f(x_{t+1}) \leq f(x_t) - \frac{1}{2L} \left\| \nabla f(x_t) \right\|^2 \)

Question

How lower bound?
Algorithm?

Gradient Descent!
• For $t = 0, \ldots, T - 1$
  • $x_{t+1} = x_t - \frac{1}{L} \nabla f(x_t)$
• Output $x_T$

Upper Bound Analysis
• $f(x_{t+1}) \leq U_{x_t}(x_{t+1})$
• $U_{x_t}(x_{t+1}) = f(x_t) - \frac{1}{2L} \|\nabla f(x_t)\|^2$
• $f(x_{t+1}) \leq f(x_t) - \frac{1}{2L} \|\nabla f(x_t)\|^2$

Progress Measures
• $\|\nabla f(x)\|^2$ - norm of gradient
• $f(x) - f_*$ where $f_* = \inf_{x \in \mathbb{R}^n} f(x)$ - function error
• $\|x - x_*\|^2$ - for minimizer $x_*$ (i.e. $f(x_*) = f_*$)

Goal: compute $\epsilon$-optimal point
Assumption $f: \mathbb{R}^n \to \mathbb{R}$ is $L$-smooth and $\mu$-strongly convex
Given: $x_0 \in \mathbb{R}^n$ and a gradient oracle

Question
How lower bound?
Smoothness Implication

**Lemma** If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable then $\nabla f(x_*) = 0$. If $f$ is $L$-smooth then

$$\frac{1}{2L} \|\nabla f(x)\|_2^2 \leq f(x) - f(x_*) \leq \frac{L}{2} \|x - x_*\|_2^2$$

**Proof**

- Differentiability and $\nabla f(x) \neq 0 \Rightarrow f(x - \eta \nabla f(x)) < f(x)$ for small $\eta$
- $f_* \leq f(x - (1/L)\nabla f(x)) \leq f(x) - \frac{1}{2L} \|\nabla f(x)\|_2^2$
- $f(x) \leq f(x_*) + \nabla f(x_*)^T (x - x_*) + \frac{L}{2} \|x - x_*\|_2^2$

$$= f(x_*) + \frac{L}{2} \|x - x_*\|_2^2$$
Convexity Implication

**Lemma** If \( f : \mathbb{R}^n \to \mathbb{R} \) is differentiable and \( \mu \)-strongly convex

\[
\frac{1}{2\mu} \| \nabla f(x) \|_2^2 \geq f(x) - f(x^*) \geq \frac{\mu}{2} \| x - x^* \|_2^2
\]

**Proof**

- \( f(x) \geq f(x^*) + \nabla f(x^*)^T (x - x^*) + \frac{L}{2} \| x - x^* \|_2^2 \)
  \[
  = f(x^*) + \frac{L}{2} \| x - x^* \|_2^2
  \]

- \( f(x^*) \geq \min_y L_x(y) = \min_y f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \| y - x \|_2^2 \)
  \[
  = \min_y f(x) - \frac{1}{2\mu} \| \nabla f(x) \|_2^2 + \frac{\mu}{2} \| y - (x - (1/\mu) \nabla f(x)) \|_2^2
  \]
  \[
  = f(x) - \frac{1}{2\mu} \| \nabla f(x) \|_2^2
  \]
Strongly Convex Case

- Goal: compute $\epsilon$-optimal point
- Assumption $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is $L$-smooth and $\mu > 0$-strongly convex
- Given: $x_0 \in \mathbb{R}^n$ and a gradient oracle
- Algorithm: $x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$

- $\epsilon_k \overset{\text{def}}{=} f(x_k) - f^*$
- $f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \norm{\nabla f(x_k)}_2^2 \Rightarrow \epsilon_{k+1} \leq \epsilon_k - \frac{1}{2L} \norm{\nabla f(x_k)}_2^2$
- $\norm{\nabla f(x_k)}_2^2 \geq 2\mu[f(x_k) - f^*] = 2\mu \cdot \epsilon_k$
- $\Rightarrow \epsilon_{k+1} \leq \left(1 - \frac{\mu}{L}\right)\epsilon_k$
- $\Rightarrow \epsilon_k \leq \left(1 - \frac{\mu}{L}\right)^k \epsilon_0 \leq \exp\left(-\frac{k\mu}{L}\right)\epsilon_0$ [as $1 + x \leq \exp(x)$ for all $x$]
- $\Rightarrow k = \left\lfloor \frac{L}{\mu} \log\left(\frac{\epsilon_0}{\epsilon}\right) \right\rfloor$ then $\epsilon_k \leq \epsilon$

**Theorem**

Gradient descent computes $\epsilon$-critical point with $O\left(\frac{L}{\mu} \log\left(\frac{f(x_0) - f^*}{\epsilon}\right)\right)$ gradient queries.
Non-Strongly Convex Case ($\mu = 0$)

**Lemma** If $f$ is differentiable and convex then for all minimizers $x^*$

$$f(x) - f^* \leq \|\nabla f(x)\|_2 \cdot \|x - x^*\|_2^2$$

**Proof**

- $f(x^*) \geq f(x) + \nabla f(x)^T(x^* - x)$
  
  $$\geq f(x) - \|\nabla f(x)\|_2 \cdot \|x^* - x\|_2$$

$$\frac{1}{2\mu} \|\nabla f(x)\|_2^2 \geq f(x) - f(x^*) \geq \frac{\mu}{2} \|x - x^*\|_2^2$$
Convex Case

- Goal: compute $\epsilon$-optimal point
- Assumption $f: \mathbb{R}^n \to \mathbb{R}$ is $L$-smooth and convex
- Given: $x_0 \in \mathbb{R}^n$ and a gradient oracle
- Algorithm: $x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$

\[
\epsilon_k \overset{\text{def}}{=} f(x_k) - f_* \quad \text{and} \quad D \overset{\text{def}}{=} \max_{k \geq 0} \min_{x_*: f(x_*) = f_*} \|x_k - x_*\|_2
\]

- $f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|_2^2$ so $\epsilon_{k+1} \leq \epsilon_k - \frac{1}{2L} \|\nabla f(x_k)\|_2^2$

- $\epsilon_k \leq \|\nabla f(x_k)\|_2 \cdot D$ so $\epsilon_{k+1} \leq \epsilon_k - \frac{1}{2L} \left( \frac{\epsilon_k}{D} \right)^2$

- $\Rightarrow \frac{1}{\epsilon_k} \leq \frac{1}{\epsilon_{k+1}} - \frac{\epsilon_k}{2LD^2\epsilon_{k+1}} \leq \frac{1}{\epsilon_{k+1}} - \frac{1}{2LD^2}$

- $\Rightarrow \frac{1}{\epsilon_k} \geq \frac{1}{\epsilon_0} + \frac{k}{2LD^2}$

- $\epsilon_0 \leq \frac{L}{2} D^2 (f(x_k) - f_*) \leq \frac{L}{2} \|x_k - x_*\|_2^2$

- $\Rightarrow \epsilon_k \leq \frac{2LD^2}{k+4}$

**Theorem**

Gradient descent computes $\epsilon$-critical point with $O \left( \frac{LD^2}{\epsilon} \right)$ gradient queries.

**Note:** can improve to $O \left( \frac{L\|x_0 - x_*\|_2^2}{\epsilon} \right)$ for $\|:\|_2$
Plan for Today

Recap
- Gradient descent for smooth function
- Notions of convexity

Convexity
- Smoothness / convexity equivalences
- Example functions
- Implications of assumptions

Algorithm
- Gradient descent
- Algorithm analysis

Thursday
Geometry and optimality