ORIE 6326: Convex Optimization

Lower bounds

Professor Udell
Operations Research and Information Engineering
Cornell

April 25, 2017
So?

how good are the convergence rates we proved? defining

$$\| x^{(0)} - x^* \|^2 \leq R,$$

we proved

- for $f$ $L$-Lipschitz,
  $$\bar{f}(k) - p^* \leq LR \sqrt{k}.$$

- for $f$ convex and $\beta$-smooth,
  $$f(x^{(k)}) - p^* \leq \frac{\beta R^2}{2k}.$$

- for $f$ convex, $\beta$-smooth, and $\alpha$-strongly convex,
  $$f(x^{(k)}) - p^* \leq \frac{R^2}{\exp\left(-\frac{k}{\kappa}\right)},$$

where $\kappa = \frac{\beta}{\alpha} \geq 1$ is condition number
for $\epsilon > 0$, we say $x \in \mathbb{R}^n$ is $\epsilon$-optimal for the problem

$$\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in X
\end{align*}$$

if $x \in X$ is feasible and $f(x) \leq p^* + \epsilon$.

sometimes $x$ is called an “$\epsilon$-optimal solution”
Black box model

oracles for a function $f : \mathbb{R}^n \to \mathbb{R}$

- a **0th order oracle** takes $x \in \mathbb{R}^n$ and outputs $f(x)$
- a **1st order oracle** takes $x \in \mathbb{R}^n$ and outputs $\tilde{\nabla} f(x)$
- (for twice-differentiable $f$) a **2nd order oracle** takes $x \in \mathbb{R}^n$ and outputs $\nabla^2 f(x)$
Black box model

oracles for a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$

- a **0th order oracle** takes $x \in \mathbb{R}^n$ and outputs $f(x)$
- a **1st order oracle** takes $x \in \mathbb{R}^n$ and outputs $\tilde{\nabla} f(x)$
- (for twice-differentiable $f$) a **2nd order oracle** takes $x \in \mathbb{R}^n$ and outputs $\nabla^2 f(x)$

oracles for a set $\mathcal{X}$

- a **separation oracle** takes $x \in \mathbb{R}^n$ as input and outputs either
  - $x \in \mathcal{X}$, or
  - the hyperplane with normal $g \in \mathbb{R}^n$ that separates $x$ from $\mathcal{X}$
Black box model

oracles for a function $f : \mathbb{R}^n \to \mathbb{R}$

- a 0th order oracle takes $x \in \mathbb{R}^n$ and outputs $f(x)$
- a 1st order oracle takes $x \in \mathbb{R}^n$ and outputs $\tilde{\nabla}f(x)$
- (for twice-differentiable $f$) a 2nd order oracle takes $x \in \mathbb{R}^n$ and outputs $\nabla^2 f(x)$

oracles for a set $\mathcal{X}$

- a separation oracle takes $x \in \mathbb{R}^n$ as input and outputs either
  - $x \in \mathcal{X}$, or
  - the hyperplane with normal $g \in \mathbb{R}^n$ that separates $x$ from $\mathcal{X}$

why invoke oracles?

- problem difficulty = minimum $\#$ oracle calls to find $\epsilon$-optimal point $x$
- independent of computational effort
Black box model for gradient method

key insight: for a gradient method,

$$x^{(k)} \in \text{span}\{x^0, \nabla f(x^0), \ldots, \nabla f(x^{k-1})\}$$
for any first-order method and for any $k < n$, there is a convex function so that if $\|x^0 - x^*\| \leq R$, then $f(x^{(k)}) - p^* \geq O(\cdot)$, where $\cdot =$

<table>
<thead>
<tr>
<th>$f$</th>
<th>$L$-Lipschitz</th>
<th>$\beta$-smooth</th>
</tr>
</thead>
<tbody>
<tr>
<td>not strongly convex</td>
<td>$\frac{RL}{\sqrt{k}}$</td>
<td>$\frac{\beta R^2}{k}$</td>
</tr>
<tr>
<td>$\alpha$-strongly convex</td>
<td>$\frac{L^2}{\alpha k}$</td>
<td>$\frac{R^2}{\exp(-\frac{k}{\sqrt{\kappa}})}$</td>
</tr>
</tbody>
</table>
Lower bound: nonsmooth

consider, for each $k = 1, \ldots, n$,

$$f(x) = \gamma \max_{1 \leq i \leq k} x_i + \frac{\alpha}{2} \|x\|^2$$

with domain $\|x\| \leq R$

- compute

$$\partial f(x) = \alpha x + \gamma \text{conv}\{e_i : i \in \arg\max_{1 \leq i \leq k} x_i\}$$

notice $f$ is $\alpha R + \gamma$-Lipschitz
Lower bound: nonsmooth

consider, for each $k = 1, \ldots, n,$

$$f(x) = \gamma \max_{1 \leq i \leq k} x_i + \frac{\alpha}{2} \|x\|^2$$

with domain $\|x\| \leq R$

- compute

$$\partial f(x) = \alpha x + \gamma \text{conv}\{e_i : i \in \arg\max_{1 \leq i \leq k} x_i\}$$

notice $f$ is $\alpha R + \gamma$-Lipschitz

- suppose 1st-order oracle returns

$$\tilde{\nabla} f(x) = \alpha x + \gamma e_j, \quad j = \inf(\arg\max_{1 \leq i \leq k} x_i)$$
Lower bound: nonsmooth

consider, for each \( k = 1, \ldots, n \),

\[
f(x) = \gamma \max_{1 \leq i \leq k} x_i + \frac{\alpha}{2} \|x\|^2
\]

with domain \( \|x\| \leq R \)

- compute

\[
\partial f(x) = \alpha x + \gamma \text{conv}\{e_i : i \in \arg\max_{1 \leq i \leq k} x_i\}
\]

notice \( f \) is \( \alpha R + \gamma \)-Lipschitz

- suppose 1st-order oracle returns

\[
\tilde{\nabla} f(x) = \alpha x + \gamma e_j, \quad j = \inf(\arg\max_{1 \leq i \leq k} x_i)
\]

- starting at \( x^{(0)} = 0 \), generates \( e_1 \), so \( x^{(1)} \in \text{span}\{e_1\} \)
Lower bound: nonsmooth

consider, for each $k = 1, \ldots, n$,

$$f(x) = \gamma \max_{1 \leq i \leq k} x_i + \frac{\alpha}{2} \|x\|^2$$

with domain $\|x\| \leq R$

- compute

$$\partial f(x) = \alpha x + \gamma \text{conv}\{e_i : i \in \arg \max_{1 \leq i \leq k} x_i\}$$

notice $f$ is $\alpha R + \gamma$-Lipschitz

- suppose 1st-order oracle returns

$$\tilde{\nabla} f(x) = \alpha x + \gamma e_j, \quad j = \inf(\arg \max_{1 \leq i \leq k} x_i)$$

- starting at $x^{(0)} = 0$, generates $e_1$, so $x^{(1)} \in \text{span}\{e_1\}$

- given $x^{(t-1)} \in \text{span}\{e_1, \ldots, e_{t-1}\}$, generates $x^{(t)} \in \text{span}\{e_1, \ldots, e_t\}$. 
Lower bound: nonsmooth

\[ f(x) = \gamma \max_{1 \leq i \leq k} x_i + \frac{\alpha}{2} \|x\|^2 \]

- if \( x(t) \in \text{span}\{e_1, \ldots, e_t\} \), how small can \( f(x(t)) \) be?
Lower bound: nonsmooth

\[ f(x) = \gamma \max_{1 \leq i \leq k} x_i + \frac{\alpha}{2} \|x\|^2 \]

- if \( x(t) \in \text{span}\{e_1, \ldots, e_t\} \), how small can \( f(x(t)) \) be? \( f(x(t)) \geq 0. \)
- find the solution: note that if

\[
y_i = \begin{cases} 
- \frac{\gamma}{\alpha k} & i = 1, \ldots, k \\
0 & \text{otherwise}
\end{cases}
\]

then \( 0 \in \partial f(y) \)
Lower bound: nonsmooth

\[ f(x) = \gamma \max_{1 \leq i \leq k} x_i + \frac{\alpha}{2} \|x\|^2 \]

- if \( x^{(t)} \in \text{span}\{e_1, \ldots, e_t\} \), how small can \( f(x^{(t)}) \) be? \( f(x^{(t)}) \geq 0 \).
- find the solution: note that if

\[ y_i = \begin{cases} -\frac{\gamma}{\alpha k} & i = 1, \ldots, k \\ 0 & \text{otherwise} \end{cases} \]

then \( 0 \in \partial f(y) \)

- so \( p^* = f(y) = -\frac{\gamma^2}{\alpha k} + \frac{\alpha}{2} \frac{\gamma^2}{\alpha^2 k} = -\frac{\gamma^2}{2\alpha k} \)
Lower bound: nonsmooth

\[ f(x) = \gamma \max_{1 \leq i \leq k} x_i + \frac{\alpha}{2} \|x\|^2 \]

- if \( x^{(t)} \in \text{span}\{e_1, \ldots, e_t\} \), how small can \( f(x^{(t)}) \) be? \( f(x^{(t)}) \geq 0 \).
- find the solution: note that if

\[
y_i = \begin{cases} 
-\frac{\gamma}{\alpha k} & i = 1, \ldots, k \\
0 & \text{otherwise}
\end{cases}
\]

then \( 0 \in \partial f(y) \)
- so \( p^* = f(y) = -\frac{\gamma^2}{\alpha k} + \frac{\alpha}{2} \frac{\gamma^2}{\alpha^2 k} = -\frac{\gamma^2}{2\alpha k} \)
- and so for any \( t \leq k \),

\[ f(x^{(t)}) - p^* \geq 0 - p^* = \frac{\gamma^2}{2\alpha k} \]
Lower bound: nonsmooth

for any $t \leq k$,

$$f(x^{(t)}) - p^* \geq \frac{\gamma^2}{2\alpha k}$$

to prove lower bound for $L$-Lipschitz functions:

- take $\gamma = L \frac{\sqrt{k}}{1+\sqrt{k}}$, $\alpha = \frac{L}{R} \frac{1}{1+\sqrt{k}}$
- check Lipschitz constant $\gamma + \alpha R = L$
- notice solution $y$ has $\|y\|^2 = R^2$ for this choice:

$$y_i = \begin{cases} 
- \gamma \alpha k = - \frac{R}{\sqrt{k}} & i = 1, \ldots, k \\
0 & \text{otherwise}
\end{cases}$$

- so for $t \leq k$,

$$f(x^{(t)}) - p^* \geq \frac{\gamma^2}{2\alpha k} = LR \frac{2}{2(1 + \sqrt{k})}$$
Lower bound: nonsmooth and strongly convex

for any $t \leq k$,

$$f(x^{(t)}) - p^* \geq \frac{\gamma^2}{2\alpha k}$$

to prove lower bound for $L$-Lipschitz $\alpha$-strongly convex functions:

- take $\gamma = \frac{L}{2}$, $R = \frac{L}{2\alpha}$
- notice solution $y$ has $\|y\|^2 = R^2/k \leq R^2$ for this choice:

$$y_i = \begin{cases} -\frac{\gamma}{\alpha k} = -\frac{L/2}{Lk/2R} = -\frac{R}{k} & i = 1, \ldots, k \\ 0 & \text{otherwise} \end{cases}$$

- so for $t \leq k$,

$$f(x^{(t)}) - p^* \geq \frac{\gamma^2}{2\alpha k} = \frac{L}{2\alpha k}$$
References

- Nesterov, Introductory Lectures on Convex Optimization
- Bubeck, Convex Optimization: Algorithms and Complexity (Section 3.5)
Lower bounds

for any first-order method and for any $k < n$, there is a convex function so that if $\|x^0 - x^*\| \leq R$, then $f(x^{(k)}) - p^* \geq \mathcal{O}(\cdot)$, where $\cdot =$

<table>
<thead>
<tr>
<th>$f$</th>
<th>L-Lipschitz</th>
<th>$\beta$-smooth</th>
</tr>
</thead>
<tbody>
<tr>
<td>not strongly convex</td>
<td>$\frac{RL}{\sqrt{k}}$</td>
<td>$\frac{\beta R^2}{k}$</td>
</tr>
<tr>
<td>$\alpha$-strongly convex</td>
<td>$\frac{L^2}{\alpha k}$</td>
<td>$R^2 \exp\left( - \frac{k}{\sqrt{\kappa}} \right)$</td>
</tr>
</tbody>
</table>
Upper bounds

for any convex function with $p^* = \inf_{\|x^0 - x^*\| \leq R} f(x)$, we proved gradient descent can achieve $f(x^{(k)}) - p^* \geq O(\cdot)$, where $\cdot =$

<table>
<thead>
<tr>
<th>$f$</th>
<th>$L$-Lipschitz</th>
<th>$\beta$-smooth</th>
</tr>
</thead>
<tbody>
<tr>
<td>not strongly convex</td>
<td>$\frac{RL}{\sqrt{k}}$</td>
<td>$\frac{\beta R^2}{k}$</td>
</tr>
<tr>
<td>$\alpha$-strongly convex</td>
<td>$\frac{L^2}{\alpha k}$</td>
<td>$R^2 \exp\left(-\frac{k}{\kappa}\right)$</td>
</tr>
</tbody>
</table>
Spot the difference?

smooth, strongly convex:

- lower bound: \( R^2 \exp\left(-\frac{k}{\sqrt{\kappa}}\right) \)
- upper bound: \( R^2 \exp\left(-\frac{k}{\kappa}\right) \)
Spot the difference?

smooth, strongly convex:

- lower bound: $R^2 \exp\left(-\frac{k}{\sqrt{\kappa}}\right)$
- upper bound: $R^2 \exp\left(-\frac{k}{\kappa}\right)$

does this matter?
Spot the difference?

smooth, strongly convex:

- lower bound: $R^2 \exp\left(-\frac{k}{\sqrt{\kappa}}\right)$
- upper bound: $R^2 \exp\left(-\frac{k}{\kappa}\right)$

does this matter?

- yes
Spot the difference?

smooth, strongly convex:

- lower bound: $R^2 \exp\left(-\frac{k}{\sqrt{\kappa}}\right)$
- upper bound: $R^2 \exp\left(-\frac{k}{\kappa}\right)$

does this matter?

- yes

can we fix it?

- Nesterov, 1983: Acceleration!
Spot the difference?

smooth, strongly convex:

- lower bound: $R^2 \exp(-\frac{k}{\sqrt{\kappa}})$
- upper bound: $R^2 \exp(-\frac{k}{\kappa})$

does this matter?

- yes

can we fix it?

- Nesterov, 1983: Acceleration!
- Bubeck, Lee, Singh, 2015: Geometric Descent Method
Accelerated gradient descent

**Algorithm 1** Accelerated subgradient method.

given a starting point \( x^{(0)} \in \text{dom} f \),
parameters \( b^{(k)} \in [0, 1) \) and \( a^{(k)} \in [0, 2 + 2b^{(k)}) \),
and auxiliary point \( y^{(0)} = 0 \in \mathbb{R}^n \).

for \( k = 1, 2, \ldots \)

1. **Auxiliary update.** \( y^{(k+1)} = b^{(k)}y^{(k)} + \tilde{\nabla}f(x^{(k)}) \).
2. **Update.** \( x^{(k+1)} = x^{(k)} - a^{(k)}y^{(k+1)} \).

until stopping criterion is satisfied.

- if \( b^{(k)} = b = 0 \), reduces to gradient descent
- otherwise, go a bit farther in direction you went at last iterations
Accelerated gradient descent

Intuition and more (required reading): Gabriel Goh on momentum

http://distill.pub/2017/momentum/