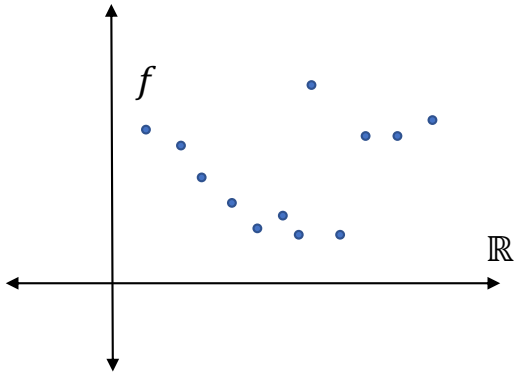
Introduction to Optimization Theory

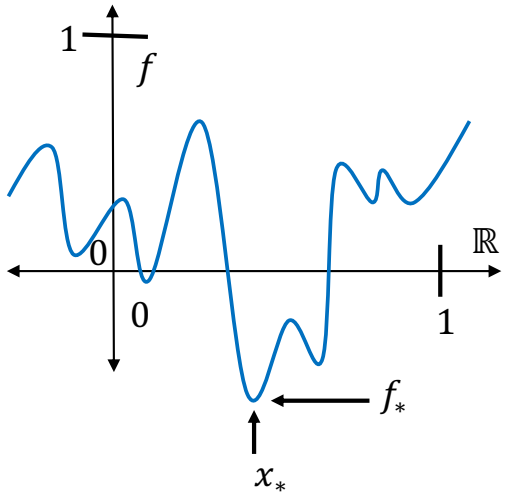
Lecture #9 - 10/13/20

MS&E 213 / CS 2690



Aaron Sidford

sidford@stanford.edu



Plan for Today

Recap

- Extension #1
- General norms

Extension #2

- Composite functions

Thursday

Review session

Recap

Problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

Regularity	Oracle	Goal	Algorithm	Iterations
$n = 1, f(x) \in [0,1], x_* \in [0,1]$	value	½-optimal	anything	∞
$n = 1, x_* \in [0,1]$, L-Lipschitz	value	ϵ-optimal	ϵ-net	$\Theta(L/\epsilon)$
$x_* \in [0,1]$, L-Lipschitz in $\|\cdot\|_\infty$	value	ϵ-optimal	ϵ-net	$\left(\Theta(L/\epsilon)\right)^n$
L-smooth and bounded	value, gradient	ϵ-optimal	ϵ-net	exponential
L-smooth	gradient	ϵ-critical	gradient descent	$O\left(\frac{L(f(x_0) - f_*)}{\epsilon^2}\right)$
L-smooth μ-strongly convex	gradient	ϵ-optimal	gradient descent	$O\left(\frac{L}{\mu}\log\left(\frac{f(x_0) - f_*}{\epsilon}\right)\right)$
L-smooth convex	gradient	ϵ-optimal	gradient descent	$O\left(\frac{L\|x_0 - x_*\|_2^2}{\epsilon}\right)$

Accelerated Gradient Descent: $O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{f(x_0)-f_*}{\epsilon}\right)\right)$ ***and*** $O\left(\sqrt{\frac{L\|x_0-x_*\|_2^2}{\epsilon}}\right)$

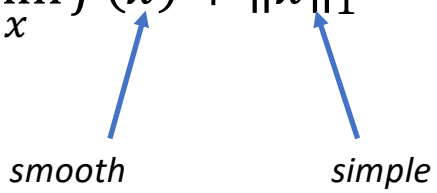
Extensions

Iterative Method Landscape

- So far – first order methods (gradient / value oracle) and $\|\cdot\|_2$
- Our machinery extends to many different settings and oracles
- **Goal**: see broader theory and understand extensions

Cases

- Different norms (e.g. $\|\cdot\|_\infty$)
- Constraints, e.g. $\min_{x \in S} f(x)$
- Composite functions, e.g. $\min_x f(x) + \|x\|_1$
- Coordinate descent



General Norms

$$\frac{\mu}{2}\|x-y\|^2 \le f(y) - [f(x) + \nabla f(x)^\top (y-x)] \le \frac{L}{2}\|x-y\|^2$$

Upper Bound Oracle!

- $x_{k+1} = \underset{x}{\operatorname{argmin}}\, f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{L}{2}\|x - x_k\|^2$
- $\Rightarrow f(x_{k+1}) \le f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_*^2$

Lemma

- $\frac{1}{2L}\|\nabla f(x)\|_*^2 \le f(x) - f_* \le \frac{L}{2}\|x - x_*\|^2$
- $\frac{1}{2\mu}\|\nabla f(x)\|_*^2 \ge f(x) - f_* \le \frac{\mu}{2}\|x - x_*\|^2$

Theorem: Gradient descent computes ϵ-optimal point with $O\left(\frac{L}{\mu}\log\left(\frac{[f(x_0)-f_*]}{\epsilon}\right)\right)$ gradient queries

Acceleration?

Depends on norm!

$\mu = 0$

Next extension!

Plan for Today

Recap

- Extension #1
- General norms

Extension #2

- Composite functions

Thursday

Review session

Composite Function Minimization

Problem $\min_{x \in \mathbb{R}^n} f(x)$ where $f(x) = g(x) + \psi(x)$

- g: $\mathbb{R}^n \to \mathbb{R}$ is L-smooth with respect to $\|\cdot\|$ and convex
- ψ: $\mathbb{R}^n \to \mathbb{R}$ is "given / simple" (TBD)
- f: $\mathbb{R} \to \mathbb{R}$ is μ-strongly convex with respect to $\|\cdot\|$

Examples

- **Constrained minimization**: $\min_{x \in S} f(x) \to \min_{x \in \mathbb{R}^n} f(x) + \psi(x)$ where $\psi(x) = 0$ if $x \in S$ and $\psi(x) = \infty$ otherwise
- **Regularization**
  - ℓ_1-regularization: $f(x) = g(x) + \lambda \|x\|_1$ (*encourage sparsity*)
  - ℓ_2-regularization: $f(x) = g(x) + \lambda \|x - x_0\|_2^2$ (*strong convexity*)
  - Many more!

Composite Function Minimization

Problem $\min_{x \in \mathbb{R}^n} f(x)$ where $f(x) = g(x) + \psi(x)$

- $g: \mathbb{R}^n \to \mathbb{R}$ is L-smooth with respect to $\| \cdot \|$ and convex
- $\psi: \mathbb{R}^n \to \mathbb{R}$ is "given / simple" (TBD)
- $f: \mathbb{R} \to \mathbb{R}$ is μ-strongly convex with respect to $\| \cdot \|$

Question

- How to optimize?
- Note: f may not be smooth! May not be differentiable!
  - e.g. $f(x) = g(x) + \lambda \|x\|_1$

Upper Bound Oracle

- $\min_{x \in \mathbb{R}^n} f(x)$ where $f(x) = g(x) + \psi(x)$
- $g: \mathbb{R}^n \to \mathbb{R}$ is L-smooth with respect to $\|\cdot\|$ and convex
- $\psi: \mathbb{R}^n \to \mathbb{R}$ is "given / simple" (TBD)
- $f: \mathbb{R} \to \mathbb{R}$ is μ-strongly convex with respect to $\|\cdot\|$

Algorithm

- $x_{k+1} = \operatorname{argmin}_y U_{x_k}(y) \stackrel{\text{def}}{=} g(x_k) + \nabla g(x_k)^\top (y - x_k) + \frac{L}{2}\|y - x_k\|^2 + \psi(y)$

Note

- Only need 1 gradient evaluation! (if ψ is known)
- ψ "given" $\Leftrightarrow$ can solve above problem

Analysis

- $\min_{x \in \mathbb{R}^n} f(x)$ where $f(x) = g(x) + \psi(x)$
- $g: \mathbb{R}^n \to \mathbb{R}$ is L-smooth with respect to $\|\cdot\|$ and convex
- $\psi: \mathbb{R}^n \to \mathbb{R}$ is "given / simple" (TBD)
- $f: \mathbb{R} \to \mathbb{R}$ is μ-strongly convex with respect to $\|\cdot\|$

Algorithm

- $x_{k+1} = \min_y U_{x_k}(y) \stackrel{\text{def}}{=} g(x_k) + \nabla g(x_k)^\top (y - x_k) + \frac{L}{2}\|y - x_k\|^2 + \psi(y)$

Lemma: $f(y) \le U_{x_k}(y) \le f(y) + \frac{L}{2}\|y - x_k\|^2$ for all $x_k, y \in \mathbb{R}^n$

Proof:

- Smoothness: $f(y) \le U_{x_k}(y)$
- Convexity: $g(y) \ge g(x_k) + \nabla g(x_k)^\top (y - x_k) \Rightarrow U_{x_k}(y) \le f(y) + \frac{L}{2}\|y - x_k\|^2$

Corollary: $f(x_{k+1}) \le \min_{y \in \mathbb{R}^n} f(y) + \frac{L}{2}\|y - x_k\|^2$

Is this property enough to obtain ϵ-optimal points?

Proximal Point Method: $x_{k+1} = \operatorname{argmin}_y f(y) + \frac{L}{2}\|y - x_k\|^2$

Is progress of proximal point method enough?

Strongly Convex Case

***Challenge**! f may not be differentiable!*

Lemma: If $f: \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex with respect to $\|\cdot\|$ then

$$\min_y p(y) = f(y) + \frac{L}{2}\|y - x\|^2 \leq f(x) - \frac{\mu}{L+\mu}[f(x) - f_*]$$

Proof: $x_t = x + t(x_* - x)$

- $\min_y p(y) \leq \min_{t\in[0,1]} p(x_t) = \min_{t\in[0,1]} f(x_t) + \frac{L}{2}\cdot t^2\|x - x_*\|^2$
- $f(x_t) \leq t\cdot f(x_*) + (1-t)\cdot f(x) - \frac{\mu}{2}t(1-t)\|x - x_*\|^2$
- $\min_y p(y) \leq \min_{t\in[0,1]} f(x) - t[f(x) - f(x_*)] + \frac{t}{2}[Lt - \mu(1-t)]\cdot\|x - x_*\|^2$
- Picking $t = \frac{\mu}{L+\mu}$ yields result

Strongly Convex Case

- $\min_{x \in \mathbb{R}^n} f(x)$ where $f(x) = g(x) + \psi(x)$
- $g: \mathbb{R}^n \to \mathbb{R}$ is L-smooth with respect to $\|\cdot\|$ and convex
- $\psi: \mathbb{R}^n \to \mathbb{R}$ is "given / simple" (TBD)
- $f: \mathbb{R} \to \mathbb{R}$ is μ-strongly convex with respect to $\|\cdot\|$

Lemma: If $f: \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex with respect to $\|\cdot\|$ then

$$\min_y p(y) = f(y) + \frac{L}{2}\|y - x\|^2 \leq f(x) - \frac{\mu}{L+\mu}[f(x) - f_*]$$

Implications:

- $x_{k+1} = \operatorname{argmin}_x U_{x_k}(x) \Rightarrow f(x_{k+1}) \leq \min_y f(y) + \frac{L}{2}\|y - x_k\|^2$
- $\Rightarrow f(x_k) - f_* \leq \left(1 - \frac{\mu}{\mu+L}\right)^k [f(x_0) - f_*]$
- $\Rightarrow f(x_k) - f_* \leq \epsilon$ in $O\left(\frac{L}{\mu}\log\left(\frac{[f(x_0)-f_*]}{\epsilon}\right)\right)$ oracle queries

Question: did we need ψ convex?

***No**! But can't be too non-convex.*

Non-Strongly Convex Case

Lemma: If $f: \mathbb{R}^n \to \mathbb{R}$ is convex and x_* is a minimizer

$$\min_y p(y) = f(y) + \frac{L}{2}\|y - x\|^2 \leq f(x) - \frac{1}{2}[f(x) - f_*] \cdot \min\left\{\frac{f(x) - f_*}{L\|x - x_*\|^2}, 1\right\}$$

Proof: $x_t = x + t(x_* - x)$

- $\min_y p(y) \leq \min_{t\in[0,1]} p(x_t) = \min_{t\in[0,1]} f(x_t) + \frac{L}{2} \cdot t^2\|x - x_*\|^2$
- $f(x_t) \leq t \cdot f(x_*) + (1 - t) \cdot f(x)$
- $\min_y p(y) \leq \min_{t\in[0,1]} f(x) - t[f(x) - f(x_*)] + \frac{Lt^2}{2} \cdot \|x - x_*\|^2$
- Picking $t = \frac{f(x)-f_*}{L\|x-x_*\|^2}$ yields result if $t \leq 1$
- Otherwise $L\|x - x_*\|^2 \leq f(x) - f_*$ and picking $t = 1$ yields the result

$$\min_y f(y) + \frac{L}{2}\|y - x\|^2 \le f(x) - \frac{1}{2}[f(x) - f_*] \cdot \min\left\{\frac{f(x) - f_*}{L\|x - x_*\|^2}, 1\right\}$$

Non-Strongly Convex Case

Lemma: If f convex and x_k satisfy $f(x_{k+1}) \le \min_x f(x) + \frac{L}{2}\|x - x_k\|^2$ and $D = \max_{x_0, x_1, \ldots} \min_{\text{minimizer } x_*} \|x - x_*\|$ then $f(x_k) - f_* \le \frac{2LD^2}{k+3}$ $k \ge 1$

Proof: $\epsilon_k = f(x_k) - f_*$

- $\epsilon_{k+1} \le \epsilon_k - \frac{\epsilon_k}{2}\min\left\{\frac{\epsilon_k}{LD^2}, 1\right\}$
- $\epsilon_{k+1} \le \epsilon_k$
- $\epsilon_k \le \frac{L}{2} \cdot D^2$ for all $k \ge 1$
- $\min\left\{\frac{\epsilon_k}{LD^2}, 1\right\} \ge \min\left\{\frac{\epsilon_k}{LD^2}, \frac{2\epsilon_{k+1}}{LD^2}\right\} \ge \frac{\epsilon_{k+1}}{LD^2}$
- $\epsilon_{k+1} \le \epsilon_k - \frac{\epsilon_k \epsilon_{k+1}}{LD^2}$
- $\frac{1}{\epsilon_{k+1}} - \frac{1}{\epsilon_k} = \frac{\epsilon_k - \epsilon_{k+1}}{\epsilon_k \epsilon_{k+1}} \ge \frac{1}{2LD^2}$
- $\epsilon_k \le \frac{2LD^2}{k+3}$

Acceleration? *Yes! If norm is $\|\cdot\|_2$!*

Plan for Today

- Extension #1
- General norms

- Composite functions

Thursday

Review session