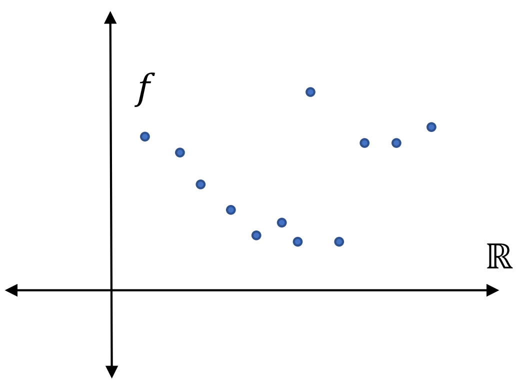
Introduction to Optimization Theory

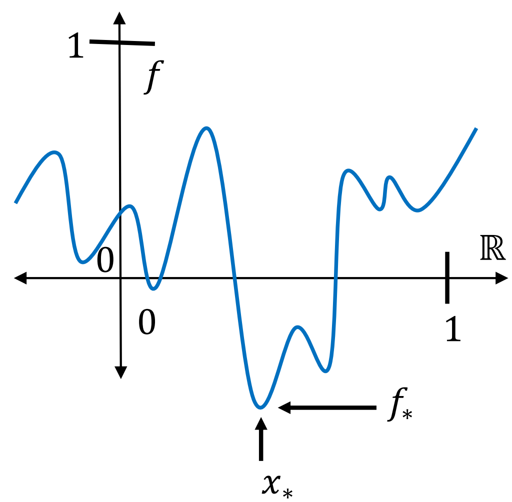
Lecture #8 - 10/8/20

MS&E 213 / CS 2690



Aaron Sidford

sidford@stanford.edu



Plan for Today

Recap

- Iterative methods

Extension #1

- General norms

Tuesday

- Composite functions
- More Extensions

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*** $O\left(\sqrt{\frac{L\|x_0-x_*\|_2^2}{\epsilon}}\right)$

Plan for Today

Recap

- Iterative methods

Extension #1

- General norms

Tuesday

- Composite functions
- More Extensions

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

Casess

- 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

smooth *simple*

Extension #1 – Arbitrary Norms

- **Definition**: $\|\cdot\|: \mathbb{R}^n \to \mathbb{R}$ is a norm if and only if for all $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$:

$$\|\alpha x\| = |\alpha| \cdot \|x\| \text{ , } \|x + y\| \leq \|x\| + \|y\| \text{ , and } \|x\| = 0 \Leftrightarrow x = 0$$

- **Definition**: For norm $\|\cdot\|$ its *dual norm* $\|\cdot\|_*$ is defined for all $x \in \mathbb{R}^n$ as $\|x\|_* = \max_{\|z\| \leq 1} z^\top x$.

- **Lemma**: $\|\cdot\|_*$ is a norm if $\|\cdot\|$ is a norm.

- **Examples**:
  - $\|x\|_1$ and $\|x\|_\infty$
  - $\|x\|_p$ and $\|x\|_q$ for $\frac{1}{p} + \frac{1}{q} = 1$ when $1 \leq p, q$ (e.g. $\|x\|_2$ and $\|x\|_2$)
  - $\|x\|_A = \sqrt{x^\top A x}$ for positive definite A and $\|x\|_{A^{-1}}$

- **Lemma ("Cauchy Schwarz")**: $|x^\top y| \leq \|x\| \cdot \|y\|_*$ for all x and y

- **Lemma:** $\min_y x^\top y + \frac{\alpha}{2} \|y\|^2 = -\frac{1}{2\alpha} \|x\|_*^2$

Analogous to

$$\min_y f(x) + \nabla f(y)^\top (x - y) + \frac{L}{2} \|x - y\|_2^2 = f(y) - \frac{1}{2L} \|\nabla f(y)\|_2^2$$

Example Proof

$\|x\|_* = \max_{\|z\| \le 1} z^\top x$

Same as $\|z\| = 1$ since can always increase argument without decreasing objective.

Lemma: $\min_y x^\top y + \frac{\alpha}{2}\|y\|^2 = -\frac{1}{2\alpha}\|x\|_*^2$

Proof:

- LHS $= -\max_y -x^\top y - \frac{\alpha}{2}\|y\|^2$
- $= -\max_{\beta \in \mathbb{R}, z \in \mathbb{R}^n : \|z\| = 1} -x^\top(\beta \cdot z) - \frac{\alpha}{2}\|\beta \cdot z\|^2$
- $= -\max_{\beta, \|z\| = 1} \beta \cdot (-x)^\top z - \frac{\alpha\beta^2}{2}$
- $= -\max_\beta \beta \cdot \|-x\|_* - \frac{\alpha\beta^2}{2}$

Maximizing $\beta = \frac{\|x\|_*}{\alpha}$

Arbitrary Norms

- **Definition**: $f: \mathbb{R}^n \to \mathbb{R}$ is *L-smooth with respect to* $\|\cdot\|$ if and only if

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \text{ for all } x, y \in \mathbb{R}^n$$

- **Definition**: $f: \mathbb{R}^n \to \mathbb{R}$ is μ-strongly with respect to $\|\cdot\|$ if and only if

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

Why?

$$O\left(\frac{L\|x - x_*\|_2^2}{k}\right) \quad versus \quad O\left(\frac{L\|x - x_*\|_\infty^2}{k}\right)$$

Can mean a $O(n)$ step improvement as $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$

Algorithms?

Idea: consider upper bound!!

Lemma: If $f: \mathbb{R}^n \to \mathbb{R}$ is L-smooth with respect to $\|\cdot\|$ then for all $x, y \in \mathbb{R}^n$
$|f(y) - [f(x) + \nabla f(x)^\top (y - x)]| \leq \frac{L}{2} \|x - y\|^2$

Proof: $x_t = x + t(y - x)$

- $A = f(y) - [f(x) + \nabla f(x)^\top (y - x)] = \int_0^1 \left(\nabla f(x_t) - \nabla f(x)\right)^\top (y - x)\, dt$
- $|A| \leq \int_0^1 \left|\left(\nabla f(x_t) - \nabla f(x)\right)^\top (y - x)\right| dt$
- $\leq \int_0^1 \|\nabla f(x_t) - \nabla f(x)\|_* \|y - x\| dt$
- $\|\nabla f(x_t) - \nabla f(x)\|_* \leq L\|x_t - x\| = Lt\|y - x\|$

Equivalence?

Lemma: If $f: \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable with

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

then f is L-smooth, i.e. $\|\nabla f(x) - \nabla f(y)\|_* \le L\|x-y\|$.

Proof:

- Let $g(z) = f(z) - [f(x) + \nabla f(x)^\top (z-x)]$
- g is convex and $\nabla g(x) = 0$
- $0 = g(x) = \min_z g(z)$
- $g(z) \le f(y) + \nabla f(y)^\top (z-y) + \frac{L}{2}\|z-y\|^2] - [f(x) + \nabla f(x)^\top (z-x)]$
- $f(y) \le f(x) + \nabla f(x)^\top (y-x) + \frac{L}{2}\|y-x\|^2$
- $0 \le \min_z \left(\nabla f(y) - \nabla f(x)\right)^\top (z-y) + \frac{L}{2}\|z-y\|^2 + \frac{L}{2}\|y-x\|^2$
- $\quad = -\frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2 + \frac{L}{2}\|y-x\|^2$

$$\min_y x^\top y + \frac{\alpha}{2}\|y\|^2 = -\frac{1}{2\alpha}\|y\|_*^2$$

More Equivalences

Lemma: $f: \mathbb{R}^n \to \mathbb{R}$ is L-smooth and μ-strongly convex with respect to $\|\cdot\|$ if and only if for all $x, y \in \mathbb{R}^n$

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

Lemma: twice differentiable $f: \mathbb{R}^n \to \mathbb{R}$ is L-smooth and μ-strongly convex with respect to $\|\cdot\|$ if and only if for all $x, z \in \mathbb{R}^n$

$$\mu\|z\|^2 \leq z^\top \nabla^2 f(x) z \leq L\|z\|^2$$

Algorithm?

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

$$\min_y x^\top y + \frac{\alpha}{2}\|y\|^2 = -\frac{1}{2\alpha}\|y\|_*^2$$

Upper Bound Oracle!

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

Example: $\|\cdot\| = \|\cdot\|_\infty$

- $\operatorname{argmin}_{\beta, \|z\|_\infty = 1} f(x_k) + \nabla f(x_k)^\top (\beta \cdot z) + \frac{L}{2}\|\beta \cdot z\|_\infty^2$
- $z = -\operatorname{sgn}(\nabla f(x_k))$ where $\operatorname{sgn}(x)_i = \begin{cases} 1 & x_i > 0 \\ -1 & x_i < 0 \\ 0 & otherwise \end{cases}$
- $\operatorname{argmin}_\beta f(x_k) - \beta\|\nabla f(x_k)\|_1 + \frac{L\beta^2}{2} = f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_1^2$
- $x_{k+1} = x_k - \frac{\|\nabla f(x_k)\|_1}{L} \cdot \operatorname{sgn}(\nabla f(x_k))$

Analysis

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

Upper Bound Oracle!

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

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

Lemma

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

Acceleration?

$\mu = 0$

Depends on norm!

Next extension!

Plan for Today

Recap

- Iterative methods

Extension #1

- General norms

Tuesday

- Composite functions
- More Extensions

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 $|| \cdot ||$

Examples

- **Constrained minimization**: $\min_{x \in S} f(x) \to \min_{x \in S} f(x) + \psi(x)$ 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. $f(x) = g(x) + \lambda\|x\|_1$