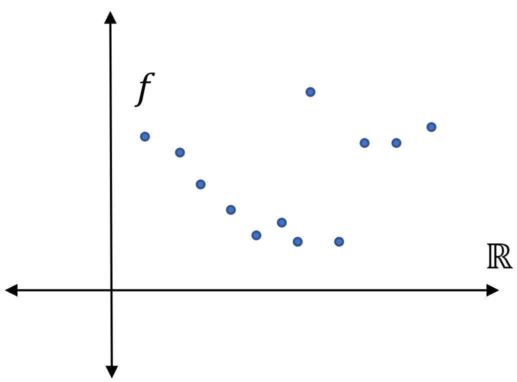
Introduction to Optimization Theory

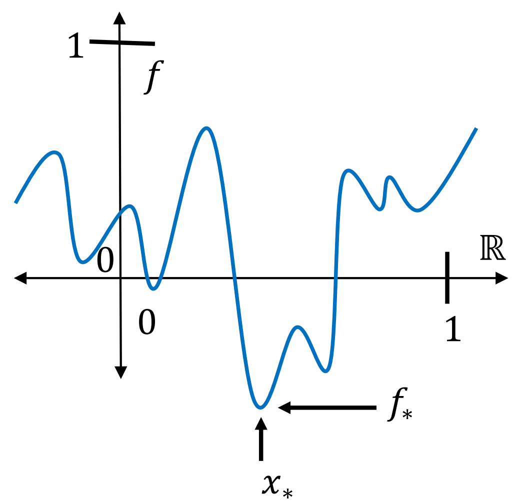
Lecture #10 - 10/15/20

MS&E 213 / CS 2690



Aaron Sidford

sidford@stanford.edu



Plan for Today

Recap

Extension #3

- Coordinate descent

Review

Tuesday

Next Unit : Convex Sets and Lipschitz Functions

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

Recap

Problem

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

Regularity	Oracle	Goal	Algorithm	Iterations
L-smooth μ-strongly convex	gradient	ϵ-optimal	gradient descent	$O\left(\sqrt{\frac{L}{\mu}} \log\left(\frac{f(x_0) - f_*}{\epsilon}\right)\right)$
L-smooth convex	gradient	ϵ-optimal	Accelerated gradient descent	$O\left(\sqrt{\frac{L\|x_0 - x_*\|_2^2}{\epsilon}}\right)$

Theorem

Suppose that $f(x_{k+1}) \leq \min_{x \in \mathbb{R}^n} f(x) + \frac{L}{2}\|x - x_k\|^2$ for all $k \geq 0$ and suppose that f is μ-strongly convex with respect to arbitrary norm $\|\cdot\|$ then

$$f(x_k) - f_* \leq \min\left\{\left(1 - \frac{\mu}{L + \mu}\right)^k [f(x_0) - f_*], \frac{LD^2}{k+3}\right\}$$

Recap

Problem

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

Theorem

Suppose that $f(x_{k+1}) \leq \min_{x \in \mathbb{R}^n} f(x) + \frac{L}{2}\|x - x_k\|^2$ for all $k \geq 0$ and suppose that f is μ-strongly convex with respect to arbitrary norm $\|\cdot\|$ then

$$f(x_k) - f_* \leq \min\left\{\left(1 - \frac{\mu}{L+\mu}\right)^k [f(x_0) - f_*], \frac{LD^2}{k+3}\right\}$$

Lemma

If $f: \mathbb{R}^n \to \mathbb{R}$ is defined for all $x \in \mathbb{R}^n$ by $f(x) = g(x) + \psi(x)$ for $g: \mathbb{R}^n \to \mathbb{R}$ that is L-smooth with respect to $\|\cdot\|$ and convex and

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}} g(x_k) + \nabla g(x_k)^\top (x - x_k) + \frac{L}{2}\|x - x_k\|^2 + \psi(x)$$

then $f(x_{k+1}) \leq \min_{x \in \mathbb{R}^n} f(x) + \frac{\mu}{2}\|x - x_k\|^2$.

Corollary: in the setting of this lemma can compute an ϵ-optimal point with $O\left(\min\left\{\left\lceil\frac{L}{\mu}\right\rceil \log\left(\frac{f(x_0) - f_*}{\epsilon}\right), \frac{LD^2}{\epsilon}\right\}\right)$ gradient queries to g.

Plan for Today

Recap ✓

Extension #3

- Coordinate descent

Review

Tuesday

Next Unit : Convex Sets and Lipschitz Functions

Coordinate Descent

*<u>**Question**</u>: stronger or weaker than gradient oracle?*

- *Weaker! Less information per query.*
- *Can implement 1-partial query from gradient query*
- *May need n-partial queries to implement 1 gradient query*

- Idea: Weaker Oracle
  - More iterations
  - Maybe lower cost per iteration
  - More fined grained analysis

*<u>**Question**</u>: why consider?*

- *One partial derivative evaluation could be $\frac{1}{n}$ of gradient cost!*

- <u>**Partial Derivative Oracle**</u>: for $f: \mathbb{R}^n \to \mathbb{R}$

<u>query</u>
$x \in \mathbb{R}^n, i \in [n]$ ⟶ **oracle** ⟶ <u>output</u>

$$\frac{\partial}{\partial x_i} f(x) = [\nabla f(x)]_i \in \mathbb{R}$$

Example Problem

Problem

- solve $Ax = b$ where $A \in \mathbb{R}^{n \times n}$ is symmetric, $A = A^\top$, and A is positive definite (PD), i.e. $z^\top A z > 0$ for all $z \neq 0$

Approach

- $\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} x^\top A x - b^\top x$

Oracle

- $\nabla f(x) = Ax - b$ *[considers every entry of A]*
- $\frac{\partial}{\partial x_i} f(x) = [\nabla f(x)]_i = [a_i^\top x - b_i]$ *for row a_i [considers one row of A]*

What Assumption?

1_i is indicator vector or i, i.e. $[1_i]_j = \begin{cases} 1 & i = j \\ 0 & otherwise \end{cases}$

Idea

- Coordinate smoothness!

Definition

- Differentiable $f: \mathbb{R}^n \to \mathbb{R}$ is L_i-smooth in coordinate i if and only if $g_x(h) = f(x + h\vec{1}_i)$ is L_i-smooth for all $x \in \mathbb{R}$

Picture

$$\left|\left[\nabla f\left(x + h\vec{1}_i\right)\right]_i - [\nabla f(x)]_i\right| \leq L \cdot |h|$$

How to use?

- f is L_i-smooth in coordinate i
- f μ-strongly convex
- Partial derivative oracle

Coordinate Descent Step

- $g(h) = f(x + h\vec{1}_i)$ is L_i-smooth
- $g'(h) = \nabla f\left(x + h\vec{1}_i\right)^\top \vec{1}_i = \left[\nabla f\left(x + h\vec{1}_i\right)\right]_i$
- $\Rightarrow g\left(h - \frac{1}{L_i} g'(h)\right) \leq g(h) - \frac{1}{2L_i}[g'(h)]^2$
- $\Rightarrow f\left(x - \frac{1}{L_i}[\nabla f(x)]_i \vec{1}_i\right) \leq f(x) - \frac{1}{2L_i}[\nabla f(x)]_i^2$

How to use?

- f is L_i-smooth in coordinate i
- f μ-strongly convex
- Partial derivative oracle

<u>Coordinate Descent Step</u>

- $\Rightarrow f\left(x - \frac{1}{L_i}[\nabla f(x)]_i \vec{1}_i\right) \leq f(x) - \frac{1}{2L_i}[\nabla f(x)]_i^2$

<u>How to use?</u>

- **<u>Idea</u>** find coordinate which maximizes $\frac{1}{2L_i}[\nabla f(x)]_i^2$
- **<u>Problem</u>**: takes $O(n)$ queries
- **<u>Idea</u>**: pick random coordinate!

Randomized Step

$$f\left(x - \frac{1}{L_i}[\nabla f(x)]_i \vec{1}_i\right) \leq f(x) - \frac{1}{2L_i}[\nabla f(x)]_i^2$$

Lemma: If $f: \mathbb{R} \to \mathbb{R}$ is L_i-smooth in coordinate i for each $i \in [n]$ and and $j \in [n]$ is chosen at random with $\Pr[j = i] = p_i$ for all $i \in [n]$ and $y = x - \frac{1}{L_j}[\nabla f(x)]_j \vec{1}_j$ then $\mathbb{E} f(y) \leq f(x) - \sum_{i \in [n]} \frac{p_i}{2L_i}[\nabla f(x)]_i^2$.

Proof

- $\mathbb{E} f(y) = \sum_{i \in [n]} p_i f\left(x - \frac{1}{L_i}[\nabla f(x)]_i \vec{1}_i\right)$
- $\qquad \leq \sum_{i \in [n]} p_i \left[f(x) - \frac{1}{2L_i}[\nabla f(x)]_i^2\right]$

Randomized Coordinate Descent

Lemma: If $f : \mathbb{R} \to \mathbb{R}$ is L_i-smooth in coordinate i for each $i \in [n]$ and and $j \in [n]$ is chosen at random with $\Pr[j = i] = p_i$ for all $i \in [n]$ and $y = x - \frac{1}{L_j}[\nabla f(x)]_j \vec{1}_j$ then

$$\mathbb{E} f(y) \leq f(x) - \sum_{i \in [n]} \frac{p_i}{2L_i} [\nabla f(x)]_i^2 .$$

Theorem: Let $f : \mathbb{R}^n \to \mathbb{R}$ be μ-convex and L_i-smooth in coordinate i for all $i \in [n]$.
Let $x_0 \in \mathbb{R}^n$ and for all $k \geq 0$ repeat repeat

- $i_k \in [n]$ chosen independently at random with $\Pr[i_k = j] = \frac{L_j}{S}$ with $S = \sum_{j \in [n]} L_j$
- $x_{k+1} = x_k - \frac{1}{L_{i_k}} [\nabla f(x)]_{i_k} \vec{1}_{i_k}$

Then $\mathbb{E} f(x_k) - f_* \leq \min\left\{\left(1 - \frac{\mu}{S}\right)^k [f(x_0) - f_*], \frac{L\|x_0 - x_*\|^2}{k+4}\right\}$ and therefore $O\left(\min\left\{\frac{S}{\mu} \log\left(\frac{[f(x_0) - f_*]}{\epsilon}\right), \frac{S\|x_0 - x_*\|_2^2}{\epsilon}\right\}\right)$ partial derivatives suffices to compute an expected ϵ-optimal point.

Proof: $\mathbb{E} f(x_{k+1}) \leq f(x_k) - \frac{1}{2S} \|\nabla f(x_k)\|_2^2$

Improvable?

- $O\left(\min\left\{\frac{S}{\mu}\log f\left(\frac{[f(x_0)-f_*]}{\epsilon}\right), \frac{S\|x_0-x_*\|_2^2}{\epsilon}\right\}\right)$ partial derivatives suffice
- Theorem: $O\left(\min\left\{\sum_{i\in[n]}\sqrt{\frac{L_i}{\mu}}\log\left(\frac{[f(x_0)-f_*]}{\epsilon}\right), \sum_{i\in[n]}\|x_0 - x_*\|_2\sqrt{\frac{L_i}{\epsilon}}\right\}\right)$

Example: $\min_x f(x) = \frac{1}{2}x^\top A x$

- $g(h) = f(x + h\vec{1}_i)$ and $g''(h) = \vec{1}_i^\top \nabla^2 f\left(x + h\vec{1}_i\right)\vec{1}_i = A_{ii}$
- A_{ii}-Lipschitz for all $i \in [n]$ and $S = \sum_{i\in[n]} A_{ii} = \mathrm{tr}(A) = \sum_{i\in[n]} \lambda_i(A)$
- GD: $O\left(\frac{\lambda_n(A)}{\lambda_1(A)}\log\left(\frac{\epsilon_0}{\epsilon}\right)\right)$ evals
- RCD: $O\left(\frac{\sum_{i\in[n]}\lambda_i(A)}{\lambda_1(A)}\log\left(\frac{\epsilon_0}{\epsilon}\right)\right)$ evals

If each is n-factor easier then is faster by however much $\frac{1}{n}\lambda_i(A)$ is smaller than $\lambda_n(A)$!

Plan for Today

Recap

Extension #3

- Coordinate descent

Review

Tuesday

Next Unit : Convex Sets and Lipschitz Functions

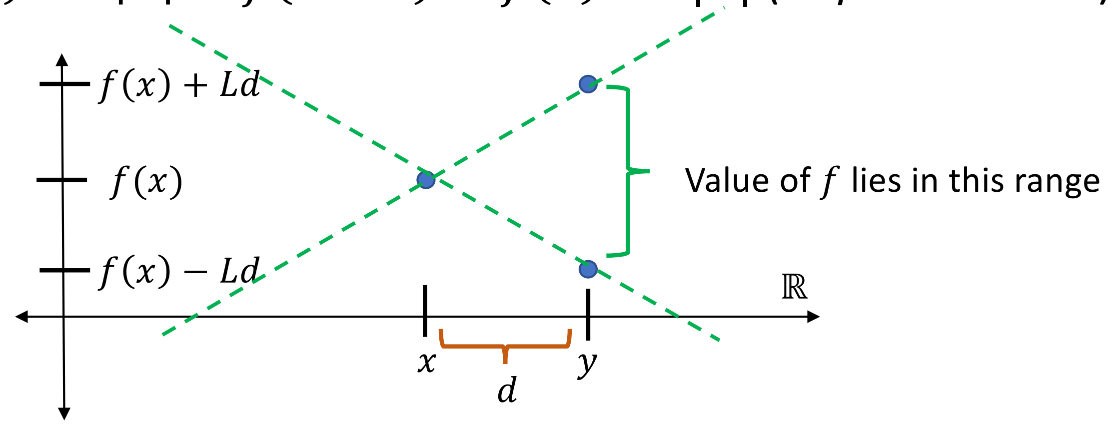
L-Lipschitz Functions

f is L-Lipschitz w.r.t. $\|\cdot\|$ if $|f(x) - f(y)| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^n$

- $\Leftrightarrow -L\|x - y\| \leq f(y) - f(x) \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^n$
- $\Leftrightarrow f(x) - L\|x - y\| \leq f(y) \leq f(x) + L\|x - y\|$ for all $x, y \in \mathbb{R}^n$

- If $n = 1$ and $\|\cdot\| = \|\cdot\|_p$ *(i.e. $\|x\| = \|x\|_p = (|x|^p)^{1/p} = |x|$)* then

$$\Leftrightarrow f(x) - L|d| \leq f(x + d) \leq f(x) + L|d| \text{ (slope at most } L\text{)}$$



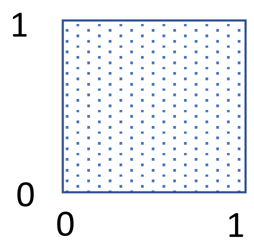
Optimal up to constants!
($(cL/\epsilon)^n$ queries are needed)

$f: \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to norm $\|\cdot\|$ if for all $x, y \in \mathbb{R}^n$ it is the case that $|f(x) - f(y)| \leq L\|x - y\|$.

Theorem: If $f: \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to $\|\cdot\|_\infty$ can compute ϵ-optimal point with $\left\lceil \frac{L}{\epsilon} \right\rceil^n$ queries to a value oracle.

Algorithm *(ϵ-net)*

- Pick $k \in \mathbb{Z}_{\geq 0}$
- Query $\left(\frac{i_1}{k}, \frac{i_2}{k}, \ldots, \frac{i_k}{k}\right)^\top$ for all possible $i_j \in [k]$
- Return point of minimum value



Analysis

- $\forall i \in [n], \exists j \in [k]$ s.t. $\left| x_*(i) - \frac{j}{k} \right| \leq \frac{1}{k}$
- $\exists q$ queried s.t. $\|x_* - q\|_\infty \leq \frac{1}{k}$
- $f(q) \leq f(x_*) + \frac{L}{k}$
- Point output is $\frac{L}{k}$-optimal
- k^n queries are made
- $\left\lceil \frac{L}{\epsilon} \right\rceil^n$-queries suffice

Smooth Functions

f is L-smooth if $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^n$

Implication

- $x, y \in \mathbb{R}^n$, $x_t = x + t(y - x)$ for $t \in [0,1]$
- $f(y) - [f(x) + \nabla f(x)^\top (y - x)] = \int_0^1 (\nabla f(x_\alpha) - \nabla f(x))^\top (y - x) d\alpha$
- $|f(y) - [f(x) + \nabla f(x)^\top (y - x)]| \leq \int_0^1 \left|(\nabla f(x_\alpha) - \nabla f(x))^\top (y - x)\right| d\alpha$
- $\left|(\nabla f(x_\alpha) - \nabla f(x))^\top (y - x)\right| \leq L\|x_\alpha - x\|_2 \|y - x\|_2 = L\alpha\|y - x\|_2^2$
- $|f(y) - [f(x) + \nabla f(x)^\top (y - x)]| \leq \frac{L}{2}\|y - x\|_2^2$

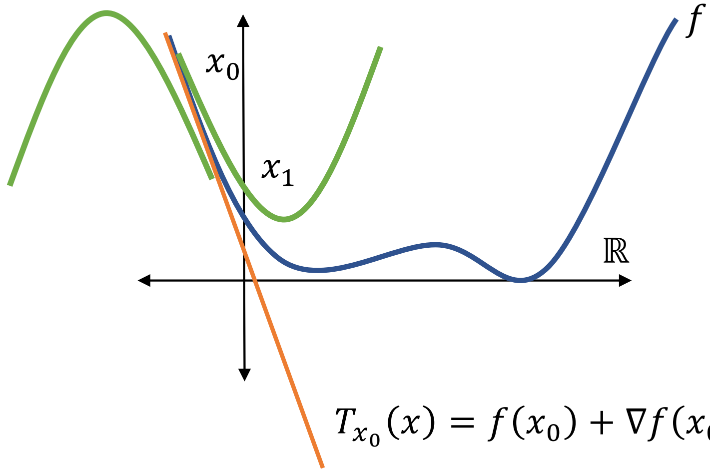
Picture?

Corollary: If L-smooth and $x, y \in \mathbb{R}^n$:

$$|f(y) - [f(x) + \nabla f(x)^\top (y - x)]| \leq \frac{L}{2} \|x - y\|_2^2$$

$$L_{x_0}(x) = f(x_0) + \nabla f(x_0)^\top (x - x_0) - \frac{L}{2} \|x - x_0\|_2^2$$

$$U_{x_0}(x) = f(x_0) + \nabla f(x_0)^\top (x - x_0) + \frac{L}{2} \|x - x_0\|_2^2$$



$$T_{x_0}(x) = f(x_0) + \nabla f(x_0)^\top (x - x_0)$$

Corollary implies that $L_{x_0}(x) \leq f(x) \leq U_{x_0}(x)$ *for all x!*

Gradient descent!

$$x_{k+1} = \operatorname*{argmin}_x U_{x_k}(x) = x_k - \frac{1}{L} \nabla f(x_k) \text{ !!!}$$

Critical Points

Corollary: If L-smooth and $x, y \in \mathbb{R}^n$:

$$|f(y) - [f(x) + \nabla f(x)^\top (y - x)]| \leq \frac{L}{2} \|x - y\|_2^2$$

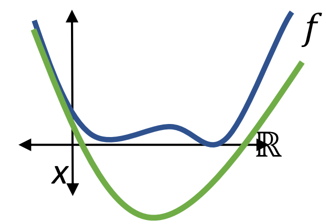
Theorem: $\leq 2L[f(x_0) - f_*]/\epsilon^2$ queries suffices to compute ϵ-critical ($\|\nabla f(x)\|_2 \leq \epsilon$) point of L-smooth function.

- $x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$
- $f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^\top (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|_2^2 \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|_2^2$
- $\sum_{i \in [k]} f(x_i) \leq \sum_{i \in [k]} \left[f(x_{i-1}) - \frac{1}{2L} \|\nabla f(x_{i-1})\|_2^2 \right]$
- $f(x_k) - f(x_0) \leq -\frac{1}{2L} \sum_{i \in [k]} \|\nabla f(x_{i-1})\|_2^2$
- $\frac{1}{k} \sum_{i \in [k]} \|\nabla f(x_{i-1})\|_2^2 \leq \frac{2L[f(x_0) - f(x_k)]}{k} \leq \frac{2L[f(x_0) - f_*]}{k}$
- $\Rightarrow \exists i \in [0, k-1]$ s.t. $\|\nabla f(x_i)\|_2^2 \leq \frac{2L[f(x_0) - f_*]}{k}$
- $\Rightarrow \epsilon$-critical point found when $k \geq 2L[f(x_0) - f_*]/\epsilon^2$!

Assumptions for Efficient ϵ-optimal Point

Notion #1: Hessian Lower Bound

- f is twice differentiable and $z^\top \nabla^2 f(x) z \geq \mu \|z\|_2^2$ for all x, z
- $\Leftrightarrow \lambda_{min}\left(\nabla^2 f(x)\right) \geq \mu$



Notion #2: Quadratic Lower Bounds

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

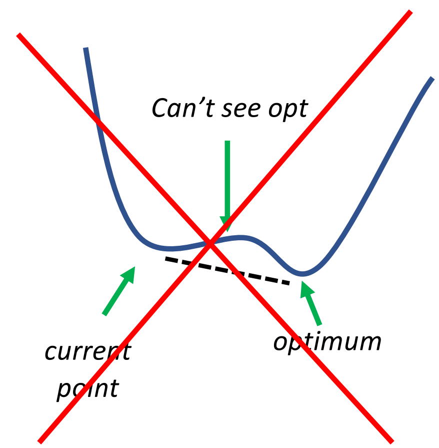
Notion #3: μ-strongly convex with respect to $\|\cdot\|$ (*by default* $\|\cdot\|_2$)

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

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

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

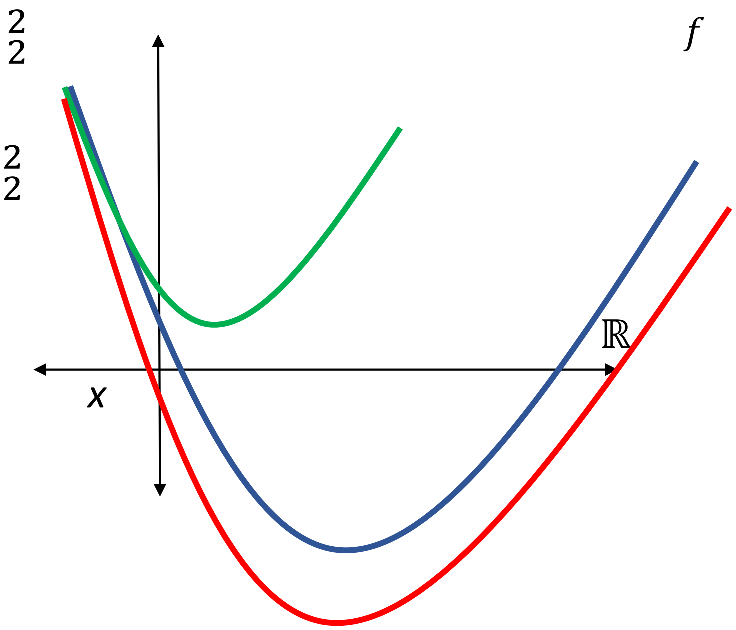
Theorem
These three notions are equivalent for twice differentiable functions



Minimizing Smooth Convex Functions

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

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



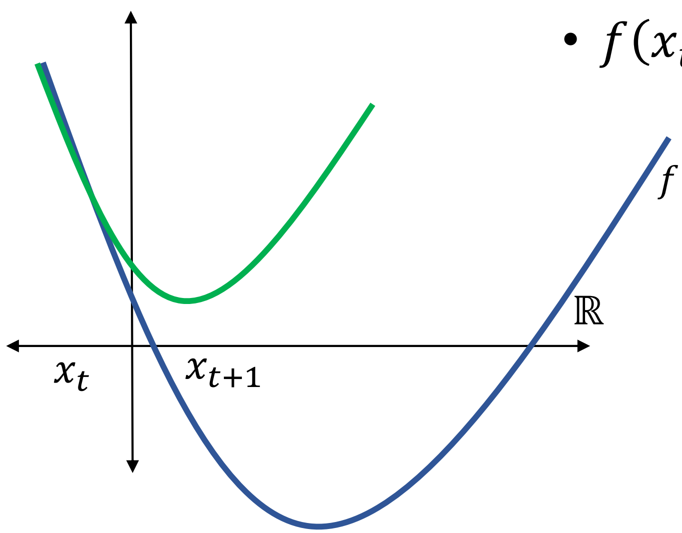
***Question**: is this assumption and a gradient oracle enough to obtain dimension independent efficient algorithms for ϵ-optimal points?*

Algorithm?

- *Goal: compute ϵ-optimal point*
- *Assumption $f: \mathbb{R}^n \to \mathbb{R}$ is L-smooth and μ-strongly convex*
- *Given: $x_0 \in \mathbb{R}^n$ and a gradient oracle*
- $f(y) \leq \boldsymbol{U_x(y)} \stackrel{\text{def}}{=} f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2$
- $f(y) \geq \boldsymbol{L_x(y)} \stackrel{\text{def}}{=} f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2$

Gradient Descent!

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



Upper Bound Analysis

- $f(x_{t+1}) \leq \boldsymbol{U_{x_t}(x_{t+1})}$
- $\boldsymbol{U_{x_t}(x_{t+1}) = f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|_2^2}$
- $f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|_2^2$

Question

How lower bound?

Algorithm?

- *Goal: compute ϵ-optimal point*
- *Assumption $f: \mathbb{R}^n \to \mathbb{R}$ is L-smooth and μ-strongly convex*
- *Given: $x_0 \in \mathbb{R}^n$ and a gradient oracle*
- $f(y) \leq \boldsymbol{U_x(y)} \stackrel{\text{def}}{=} f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2$
- $f(y) \geq \boldsymbol{L_x(y)} \stackrel{\text{def}}{=} f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2$

Gradient Descent!

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

Upper Bound Analysis

- $f(x_{t+1}) \leq \boldsymbol{U_{x_t}(x_{t+1})}$
- $\boldsymbol{U_{x_t}(x_{t+1}) = f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|_2^2}$
- $f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|_2^2$

Lower Bound Analysis ($\mu > 0$)

- $f_* \geq \boldsymbol{L_{x_t}(x_*)}$
- $f_* \geq \min_u L_{x_t}(u) = f(x_t) - \frac{1}{2\mu}\|\nabla f(x_t)\|_2^2$
- $\frac{1}{2\mu}\|\nabla f(x_t)\|_2^2 \geq f(x_t) - f_*$

Lower Bound Analysis ($\mu = 0$)

- $f_* \geq \boldsymbol{L_{x_t}(x_*) = f(x_t)} + \nabla f(x_t)^\top (x_* - x_t)$
- $f_* \geq f(x_t) - \|\nabla f(x_t)\|_2 \cdot \|x_* - x_t\|_2$

Strongly Convex Case

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

- $\epsilon_k \stackrel{\text{def}}{=} f(x_k) - f_*$
- $f(x_{k+1}) \le f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2 \Rightarrow \epsilon_{k+1} \le \epsilon_k - \frac{1}{2L}\|\nabla f(x_k)\|_2^2$
- $\|\nabla f(x_k)\|_2^2 \ge 2\mu[f(x_k) - f_*] = 2\mu \cdot \epsilon_k$
- $\Rightarrow \epsilon_{k+1} \le \left(1 - \frac{\mu}{L}\right)\epsilon_k$
- $\Rightarrow \epsilon_k \le \left(1 - \frac{\mu}{L}\right)^k \epsilon_0 \le \exp\left(-\frac{k\mu}{L}\right)\epsilon_0$ [as $1 + x \le \exp(x)$ for all x]
- $\Rightarrow k = \left\lceil \frac{L}{\mu} \log\left(\frac{\epsilon_0}{\epsilon}\right) \right\rceil$ then $\epsilon_k \le \epsilon$

Theorem

Gradient descent computes ϵ-critical point with $O\left(\frac{L}{\mu}\log\left(\frac{f(x_0) - f_*}{\epsilon}\right)\right)$ gradient queries.

Convex Case

- *Goal: compute ϵ-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: : $x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$*

- $\epsilon_k \stackrel{\text{def}}{=} f(x_k) - f_*$ *and* $D \stackrel{\text{def}}{=} \max_{k \geq 0} \min_{x_*: f(x_*) = f_*} \|x_k - x_*\|_2$
- $f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|_2^2$ 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 $\epsilon_{k+1} \leq \epsilon_k - \frac{1}{2L} \left(\frac{\epsilon_k}{D}\right)^2$
- $\Rightarrow \frac{1}{\epsilon_k} \leq \frac{1}{\epsilon_{k+1}} - \frac{\epsilon_k}{2LD^2 \epsilon_{k+1}} \leq \frac{1}{\epsilon_{k+1}} - \frac{1}{2LD^2}$
- $\Rightarrow \frac{1}{\epsilon_k} \geq \frac{1}{\epsilon_0} + \frac{k}{2LD^2}$
- $\epsilon_0 \leq \frac{L}{2} D^2$ $\left(f(x_k) - f_* \leq \frac{L}{2} \|x_k - x_*\|_2^2\right)$

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

Theorem

Gradient descent computes ϵ-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 $\|\cdot\|_2$

Recap

Problem

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

Regularity	Oracle	Goal	Algorithm	Iterations
L-smooth μ-strongly convex	gradient	ϵ-optimal	gradient descent	$O\left(\sqrt{\frac{L}{\mu}} \log\left(\frac{f(x_0) - f_*}{\epsilon}\right)\right)$
L-smooth convex	gradient	ϵ-optimal	Accelerated gradient descent	$O\left(\sqrt{\frac{L\|x_0 - x_*\|_2^2}{\epsilon}}\right)$

Theorem

Suppose that $f(x_{k+1}) \leq \min_{x \in \mathbb{R}^n} f(x) + \frac{L}{2}\|x - x_k\|^2$ for all $k \geq 0$ and suppose that f is μ-strongly convex with respect to arbitrary norm $\|\cdot\|$ then

$$f(x_{k+1}) - f_* \leq \min\left\{\left(1 - \frac{\mu}{L + \mu}\right)^k [f(x_0) - f_*], \frac{LD^2}{k+3}\right\}$$

Recap

Problem

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

Theorem

Suppose that $f(x_{k+1}) \leq \min_{x \in \mathbb{R}^n} f(x) + \frac{L}{2}\|x - x_k\|^2$ for all $k \geq 0$ and suppose that f is μ-strongly convex with respect to arbitrary norm $\|\cdot\|$ then

$$f(x_k) - f_* \leq \min\left\{\left(1 - \frac{\mu}{L+\mu}\right)^k [f(x_0) - f_*], \frac{LD^2}{k+3}\right\}$$

Lemma

If $f: \mathbb{R}^n \to \mathbb{R}$ is defined for all $x \in \mathbb{R}^n$ by $f(x) = g(x) + \psi(x)$ for $g: \mathbb{R}^n \to \mathbb{R}$ that is L-smooth with respect to $\|\cdot\|$ and convex and

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}} g(x_k) + \nabla g(x_k)^\top (x - x_k) + \frac{L}{2}\|x - x_k\|^2 + \psi(x)$$

then $f(x_{k+1}) \leq \min_{x \in \mathbb{R}^n} f(x) + \frac{\mu}{2}\|x - x_k\|^2$.

Corollary: in the setting of this lemma can compute an ϵ-optimal point with $O\left(\min\left\{\left\lceil\frac{L}{\mu}\right\rceil \log\left(\frac{f(x_0)-f_*}{\epsilon}\right), \frac{LD^2}{\epsilon}\right\}\right)$ gradient queries to g.

Plan for Today

- *What if are non-smooth?*
- *What if willing to obtain dimension dependent rates?*
- *Structure of convex sets?*
- *Online learning?*
- *Stochastic gradient descent?*
- *Newton's method?*
- *See you Tuesday!*

Recap

Extension #3

- Coordinate descent

Review

Tuesday

Next Unit : Convex Sets and Lipschitz Functions