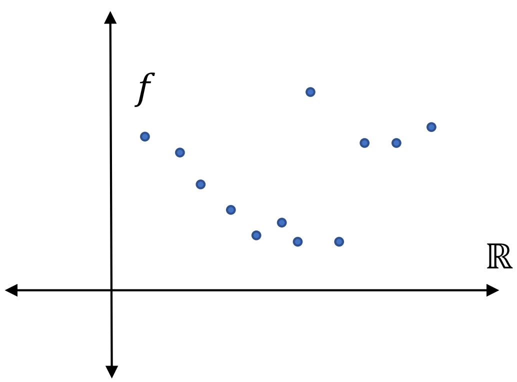
Introduction to Optimization Theory

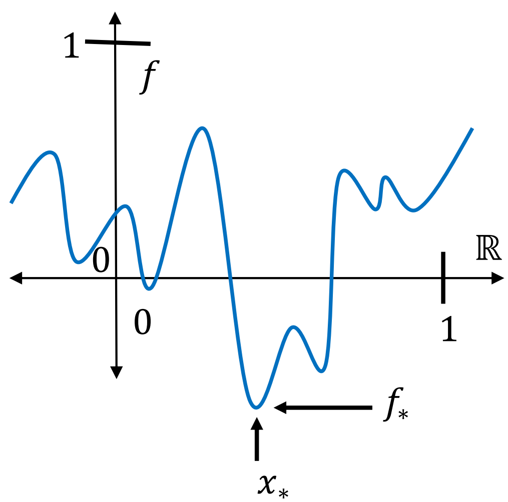
Lecture #4 - 9/24/20

MS&E 213 / CS 2690



Aaron Sidford

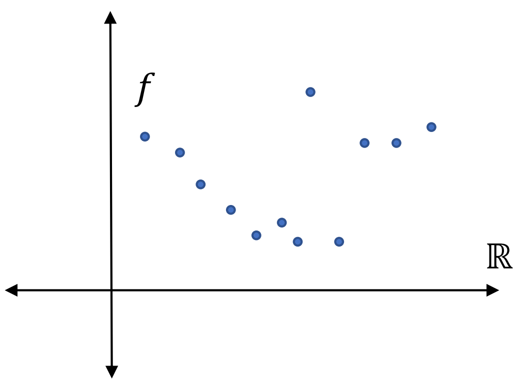
sidford@stanford.edu



Introduction to Optimization Theory

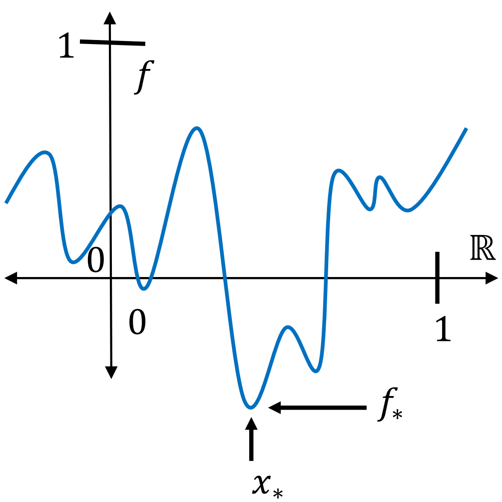
Lecture #4 - 9/24/20

MS&E 213 / CS 2690



Aaron Sidford

sidford@stanford.edu



High Level Lecture Plan

Brief Recap

Wrap up Chapters 1

Wrap up Chapters 2

Start Chapter 3

Recap

Goal
$\min_{x \in S} f(x)$ given by an oracle provably efficiently with few assumptions

- Objective function $f: \mathbb{R}^n \to \mathbb{R}$
- Constraint set $S \subseteq \mathbb{R}^n$

(Next many lectures, unconstrained $S = \mathbb{R}^n$)

query

e.g. $x \in \mathbb{R}^n$

↓

oracle

↓

output

e.g. $f(x) \in \mathbb{R}$ [value]
e.g. $\nabla f(x)$ [gradient]

Optimality Criteria

ϵ-(sub)optimal point / ϵ-additive function error:

- $x \in S$ s.t. $f(x) \leq f_* + \epsilon$ where $f_* = \min_{x \in S} f(x)$

ϵ-critical point:

- $x \in S$ s.t. $\|\nabla f(x)\|_2 \leq \epsilon$

Efficiency

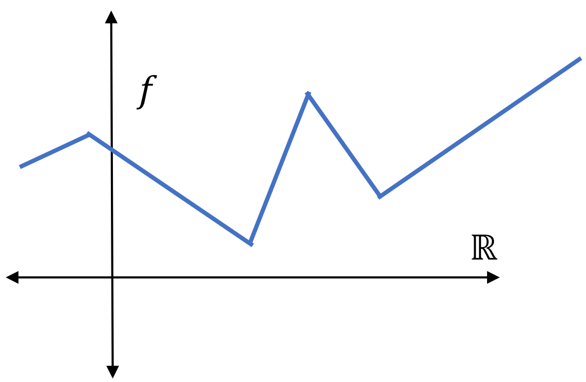
- Oracle complexity = #calls to oracle
- Runtime = # oracle calls × (average computational cost per oracle call)

Recap: structure so far

Today: complete the unit on each and introduce new assumption towards efficient ϵ-optimal point computation.

<u>f is L_1-Lipschitz w.r.t. $\|\cdot\|$</u>

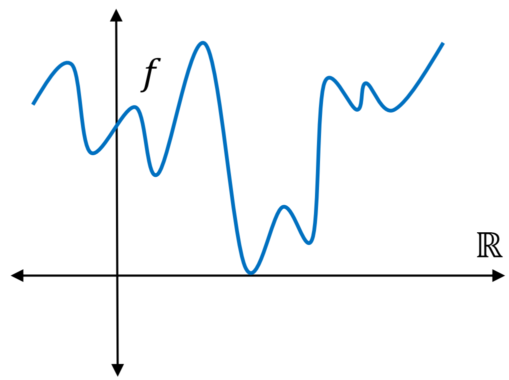
$$|f(x) - f(y)| \leq \|x - y\|$$
for all $x, y \in \mathbb{R}^n$



(bounded slope)
(bounded 1st derivatives)

<u>f is L_2-Lipschitz</u>

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L_2 \|x - y\|_2$$
for all $x, y \in \mathbb{R}^n$



(bounded curvature)
(bounded 2nd derivative)

High Level Lecture Plan

Brief Recap

Wrap up Chapters 1

Wrap up Chapters 2

Let’s be more specific :-)

Start Chapter 3

Detailed Lecture Plan

Lipschitz

- Recap: Lipschitz function minimization
- High dimensional upper / lower bounds
- Properties, characterizations

Smooth

- Recap: critical point computation of smooth functions
- Smooth function minimization lower bound
- General minimization strategy

Convex

- Introduce assumptions enabling efficient computation of ϵ-optimal points

Tuesday

Design and analyze algorithms.

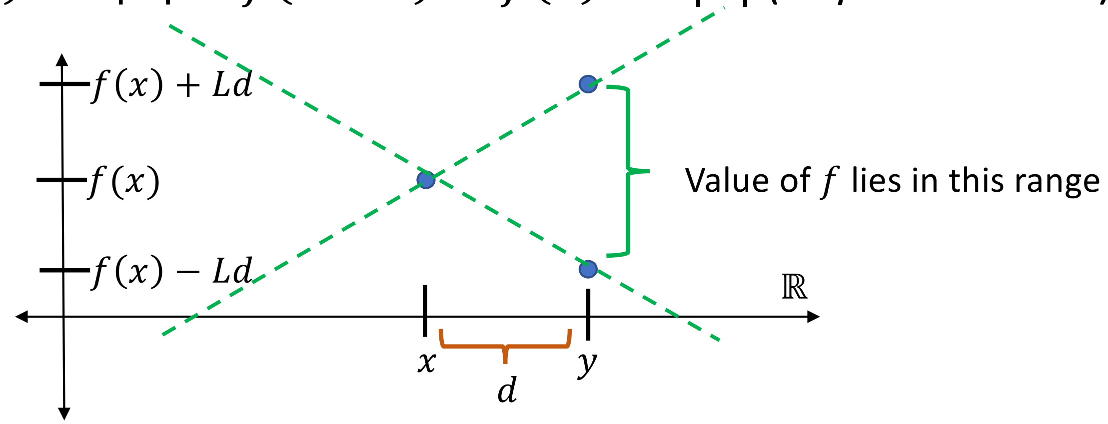
Recap: L-Lipschitz Function

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 $x, y \in \mathbb{R}^n$

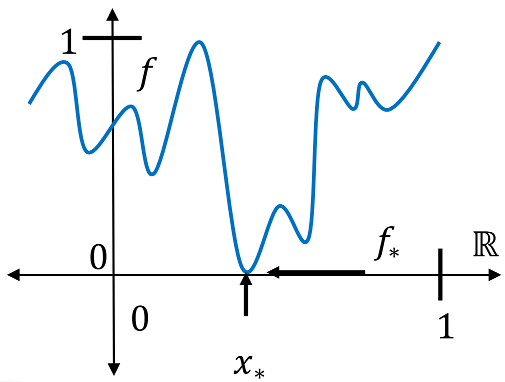
- If $n = 1$ and $\|\cdot\| = \|\cdot\|_p$ *(i.e. $\|x\| = \|x\|_p = (|x|^p)^{1/p} = |x|$)* then

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



Recap: 1d-Lipschitz Function Minimization

- $f: \mathbb{R} \to \mathbb{R}$ *(one dimensional)*
- Have evaluation oracle *(can compute $f(x)$ with 1 query)*
- $\exists x_* \in [0,1]$ such that $f(x) = f_* = \inf_{y \in \mathbb{R}} f(y)$
- $f(x) \in [0,1]$ for all $x \in \mathbb{R}$
- f is L-Lipschitz with respect to ℓ_∞
- **Goal**: compute ϵ-optimal point for $\epsilon \in (0,1)$



Theorem

$\left\lceil \frac{L}{2\epsilon} \right\rceil + 1$ queries suffice

Theorem

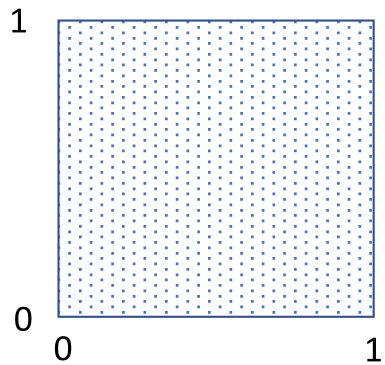
$\frac{L}{2\epsilon} - 2$ queries are needed

Recap: Higher Dimensions

- $f: \mathbb{R}^n \to \mathbb{R}$ *via evaluation oracle*
- $\exists x_* \in [0,1]^n$ *such that* $f(x) = f_*$
- $f(x) \in [0,1]$ *for all* $x \in \mathbb{R}^n$
- f *is L-Lipschitz w.r.t* $\|\cdot\|_\infty$
- *Goal: compute* ϵ*-optimal point*

Algorithm *(ϵ-net)*

- Pick $k \in \mathbb{Z}_{\geq 0}$
- Query $\left(\frac{i_1}{k}, \frac{i_2}{k}, \dots, \frac{i_k}{k}\right)^\top$ 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

Lower bound?

How to constructure difficult family of Lipschitz functions?

Lipschitz Function Facts

- $f: \mathbb{R}^n \to \mathbb{R}$ *is* L_f*-Lipschitz*
- $g: \mathbb{R}^n \to \mathbb{R}$ *is* L_g*-Lipschitz*
- $c \in \mathbb{R}$ *,* $a \in \mathbb{R}^n$
- $h: \mathbb{R}^n \to \mathbb{R}$ *defined for all* $x \in \mathbb{R}^n$ *by …*

- (**sum**) $h(x) \overset{\text{def}}{=} f(x) + g(x)$ is $L_f + L_g$-Lipschitz
- (**min**) $h(x) \overset{\text{def}}{=} \min\{f(x), g(x)\}$ is $\max\{L_f, L_g\}$-Lipschitz
- (**max**) $h(x) \overset{\text{def}}{=} \max\{f(x), g(x)\}$ is $\max\{L_f, L_g\}$-Lipschitz
- (**scaling**) $h(x) = c \cdot f(x)$ is $|c| \cdot L_f$-Lipschitz
- (**shifting**) $h(x) = f(x - a)$ is L_f-Lipschitz
- etc.

Lipschitz Functions Closed Under $\boldsymbol{min}$

If $f, g \in \mathbb{R}^n \to \mathbb{R}$ are L-Lipschitz and $h: \mathbb{R}^n \to \mathbb{R}$ with $h(x) \stackrel{\text{def}}{=} \min\{f(x), g(x)\}$ then h is L-Lipschitz.

- $h(y) \leq \min\{f(x) + L\|x - y\|, g(x) + L\|x - y\|\}$
- $\quad = \min\{f(x), g(x)\} + L\|x - y\|$
- $\quad = h(x) + L\|x - y\|$
- $\Rightarrow h(x) \leq h(y) + L\|x - y\|$
- $\Rightarrow |h(x) - h(y)| \leq L\|x - y\|$

A Non-trivial Lipschitz Function

Reverse Triangle Inequality

Any norm $\|\cdot\|: \mathbb{R}^n \to \mathbb{R}$ is 1-Lipschitz with respect to that norm.

$\big|\|x\| - \|y\|\big| \leq \|x - y\|$ for all $x, y \in \mathbb{R}^n$.

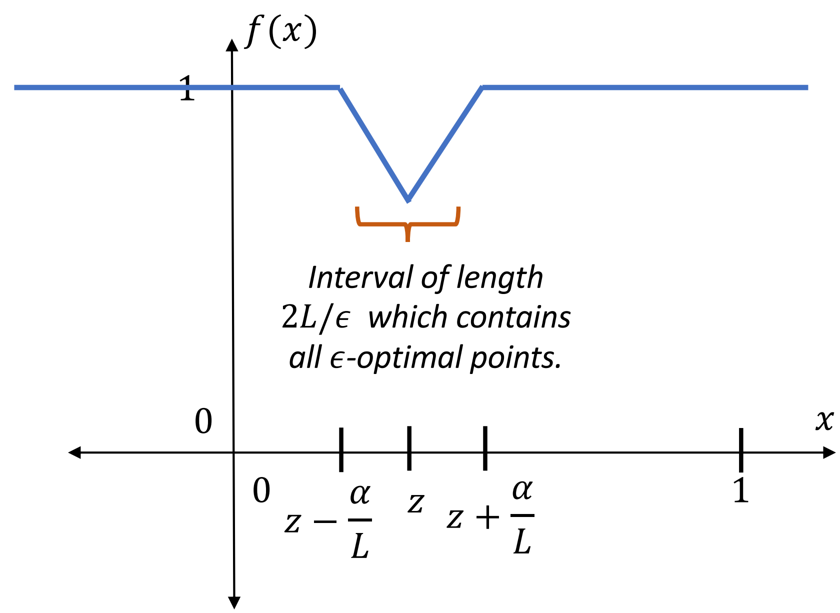
- $\|x\| = \|x - y + y\| \leq \|x - y\| + \|y\|$
- $\|x\| - \|y\| \leq \|x - y\|$
- $\|y\| - \|x\| \leq \|x - y\|$

Recap

- $f : \mathbb{R} \to \mathbb{R}$ *via evaluation oracle*
- $\exists x_* \in [0,1]^n$ *such that* $f(x) = f_*$
- $f(x) \in [0,1]$ *for all* $x \in \mathbb{R}$
- f *is L-Lipschitz w.r.t* $\|\cdot\|_\infty$
- *Goal: compute* ϵ*-optimal point*

***Lower bound** strategy find valid functions with disjoint* ϵ*-optimal points.*

- $f_{z,\alpha}(x) = \min\{1, 1 - \alpha + L|x - z|\}$



<u>Claims</u>

- x' is ϵ-optimal for $f_{z,\alpha}$ for $\alpha > \epsilon$ if and only if $|x' - z| \leq \epsilon/L$
- $f_{z,\alpha}$ is L-Lipschitz w.r.t $\|\cdot\|_\infty$

<u>Lower bound idea</u>

- If oracle outputs 1 and not enough queries, consistent with two $f_{z,\alpha}$

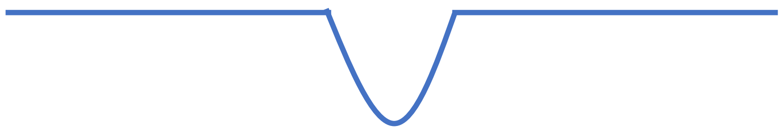
<u>Idea</u> generalize

Generalizing

- $f: \mathbb{R} \to \mathbb{R}$ *via evaluation oracle*
- $\exists x_*$ *with* $\|x_*\| \leq 1$ *such that* $f(x) = f_*$
- ~~$f(x) \in [0,1]$ *for all* $x \in \mathbb{R}$~~
- f *is* L*-Lipschitz w.r.t* $\|\cdot\|$
- *Goal: compute* ϵ*-optimal point*

Lower bound *strategy find valid functions with disjoint* ϵ*-optimal points.*

- $f_{z,\alpha}(x) = \min\{1, 1 - \alpha + L\|x - z\|\}$



Claims

- x' is ϵ-optimal for $f_{z,\alpha}$ for $\alpha > \epsilon$ iff $\|x' - z\| \leq \epsilon/L$
- $f_{z,\alpha}$ is L-Lipschitz w.r.t $\|\cdot\|$

Proof ϵ-opt for $\alpha > \epsilon$

- $f_{z,\alpha}(z) = 1 - \alpha = f_{z,\alpha}^*$
- $\|x' - z\| \leq \frac{\epsilon}{L} \Rightarrow f_{z,\alpha}(x') \leq 1 - \alpha + \epsilon$
- $\|x' - z\| > \frac{\epsilon}{L} \Rightarrow f_{z,\alpha}(x') > 1 - \alpha + \epsilon$

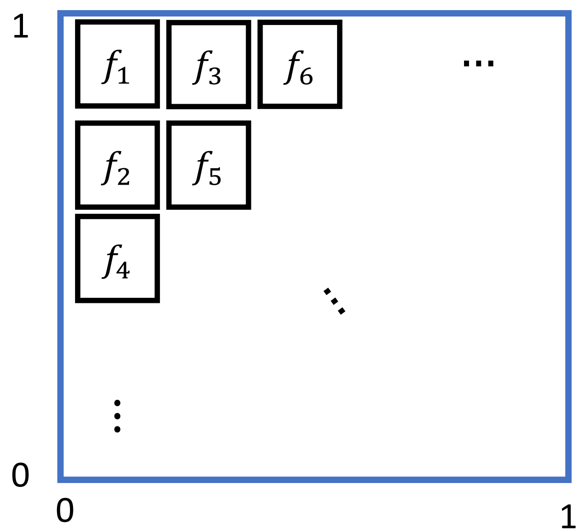
Proof L-Lipschitz

- $\|x - z\|$ is 1-Lipschitz
- $L\|x - z\|$ is L-Lipschitz
- $1 - \alpha$ is 0-Lipschitz
- $f_{z,\alpha}$ is L-Lipschitz

Assuming small enough L/ϵ

Higher Dimension Lower Bound

- *$f: \mathbb{R}^n \to \mathbb{R}$ via evaluation oracle*
- *$\exists x_* \in [0,1]^n$ such that $f(x) = f_*$*
- *$f(x) \in [0,1]$ for all $x \in \mathbb{R}^n$*
- *f is L-Lipschitz w.r.t $\|\cdot\|_\infty$*
- *Goal: compute ϵ-optimal point*



$$f_{z,\alpha}(x) = \min\{1, 1 - \alpha + L\|x - z\|_\infty\}$$

Proof Sketch

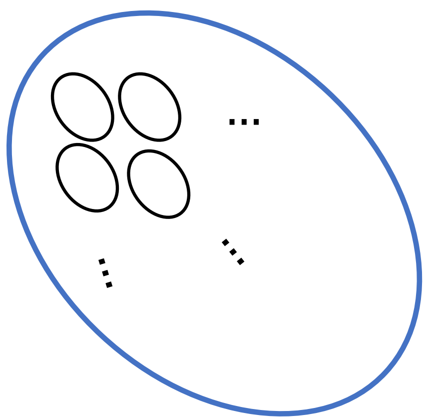
- The f are disjoint, cover $[0,1]^n$ and each have sides of length $2\alpha/L$ for $\alpha > \epsilon$.
- There are $\left[\frac{L}{2\alpha}\right]^n$ different f
- Resisting oracle: output 1. If don't query a point in two different f then algorithms is incorrect on some input.
- $\Rightarrow \left[\frac{L}{2\epsilon}\right]^n - 2$ queries are needed!
- *Recall: $\left[\frac{L}{\epsilon}\right]^n$ upper bound*

General Bounds

- $f: \mathbb{R} \to \mathbb{R}$ *via evaluation oracle*
- $\exists x_*$ *with* $\|x_*\| \leq 1$ *such that* $f(x) = f_*$
- f *is* L*-Lipschitz w.r.t* $\|\cdot\|$
- *Goal: compute* ϵ*-optimal point*

Upper Bound Strategy (ϵ-net)

- Query points such that for all y with $\|y\| \leq 1$ some query point x is within distance L/ϵ (i.e. $\|x - y\| \leq L/\epsilon$)



Lower Bound Strategy

- (1) Show that no matter what points are queried there are two points distance $> L/\epsilon$ from all queried points and each other.
- (2) Find set of points all at distance $> L/\epsilon$ from each other

We'll discuss Lipschitz functions more later in the course.

Detailed Lecture Plan

Lipschitz

- Recap: Lipschitz function minimization
- High dimensional upper / lower bounds
- Properties, characterizations

Smooth

- Recap: critical point computation of smooth functions
- Smooth function minimization lower bound
- General minimization strategy

Convex

- Introduce assumptions enabling efficient computation of ϵ-optimal points

Tuesday

Design and analyze algorithms.

Use O (resp. Ω) to hide additive and multiplicative constants in upper (resp. lower) bounds

Computing ϵ-Optimal Points

Does smoothness help?

f is L_1-Lipschitz w.r.t. $\|\cdot\|$

$$|f(x) - f(y)| \leq \|x - y\|$$

for all $x, y \in \mathbb{R}^n$

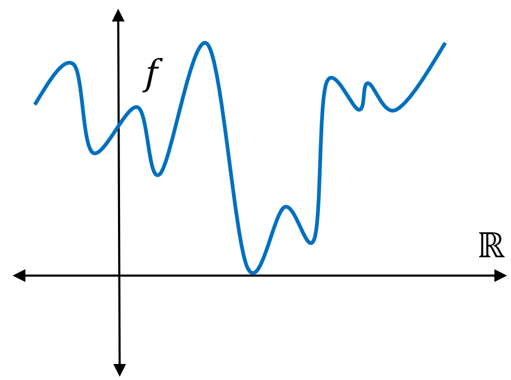
$\sim \left(\frac{cL}{\epsilon}\right)^n = \left(\Omega(L/\epsilon)\right)^n$ queries needed when bounded (where c depends on norm and dimension)

(bounded slope)
(bounded 1st derivatives)

f is L_2-Lipschitz

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L_2 \|x - y\|_2$$

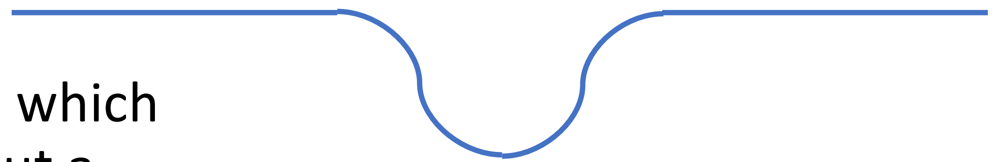
for all $x, y \in \mathbb{R}^n$



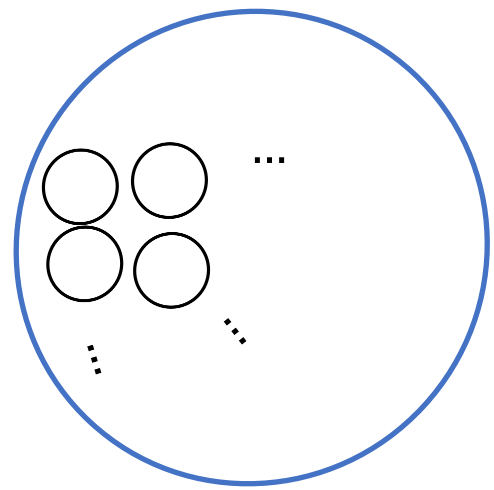
(bounded curvature)
(bounded 2nd derivative)

Problem

If $\|x_*\|_2 \leq 1$ *and* f *is* L*-smooth then* $O(nL/\epsilon)^{O(n)}$ *queries suffice and* $\Omega(L/\epsilon)^{O(n)}$ *queries are needed.*



- There are smooth functions which are constant except for all but a small region which contains all ϵ-optimal points
- Can show number of queries need still scale exponentially with dimension. (*though better*)



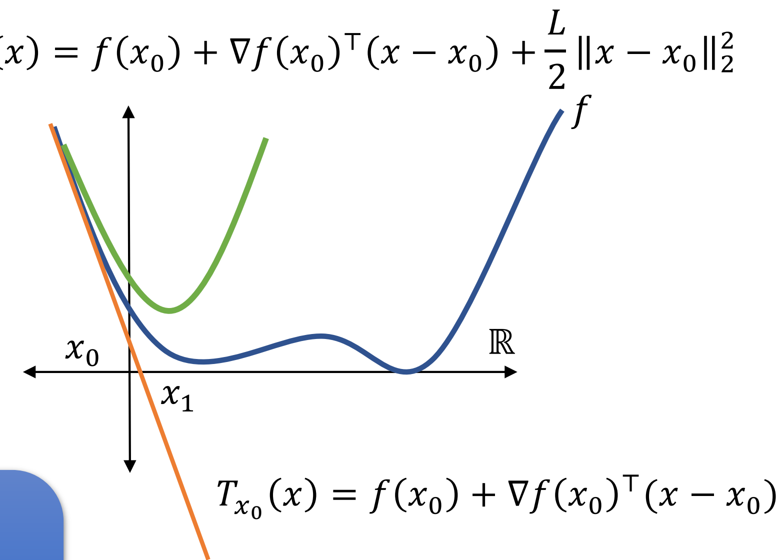
Previous Solution

Compute critical points instead.

Recap: Gradient Descent Method for Critical Points

Algorithm / Method (for L-smooth f)

- Initial point: $x_0 \in \mathbb{R}^n$
- For $k = 0,1,2, \ldots$
  - $x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$
  - If $\|\nabla f(x_k)\|_2 \leq \epsilon$ then output x_k

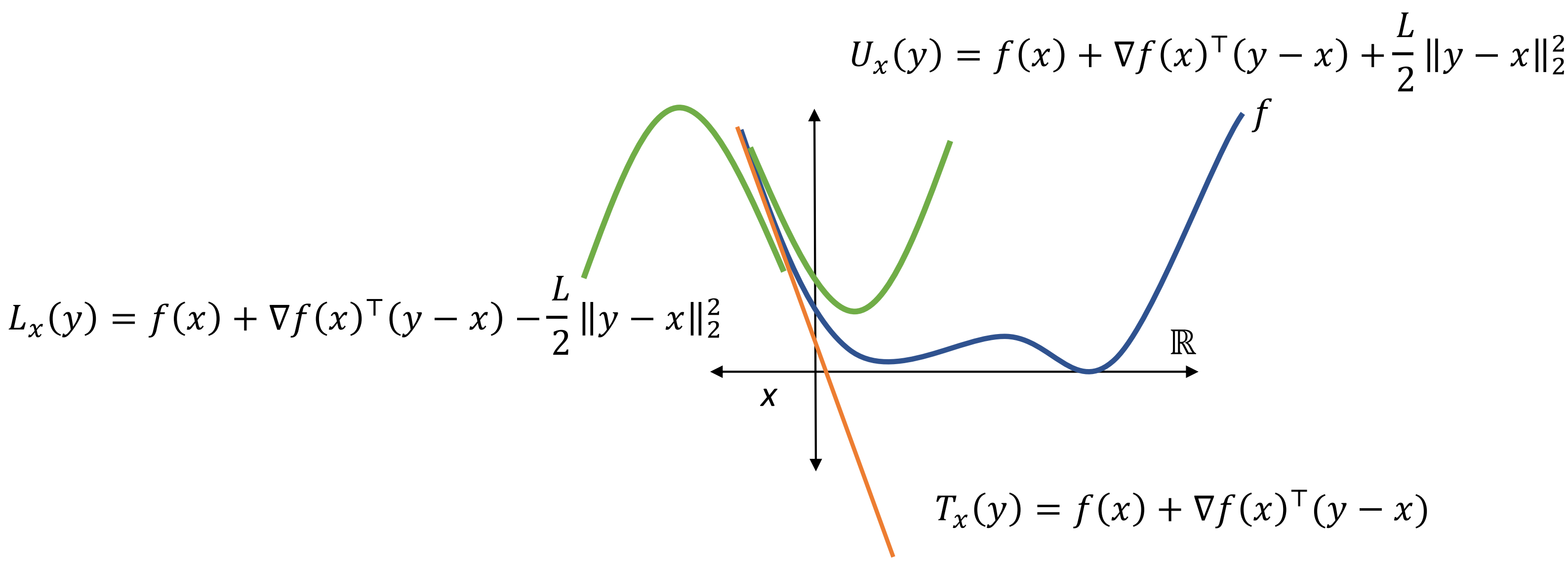


Theorem

ϵ-critical point in $\leq 2L[f(x_0) - f_*]/\epsilon^2$ steps / queries for $f_* = \inf_{x \in \mathbb{R}^n} f(x)$

Recap

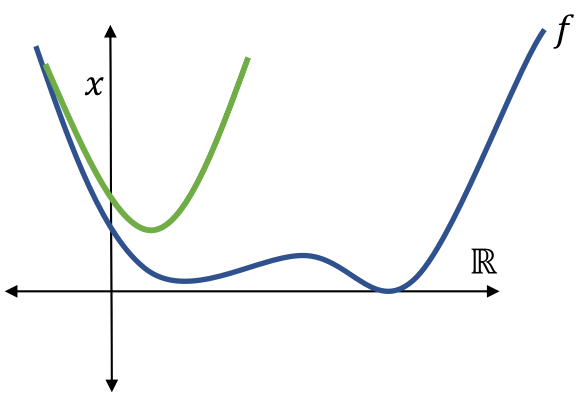
- f is L-smooth $\Leftrightarrow \|\nabla f(x) - \nabla f(y)\|_2 \leq L \cdot \|x - y\|_2$ for all $x, y \in \mathbb{R}^n$
- $\Rightarrow |f(y) - [f(x) + \nabla f(x)^\top (y - x)]| \leq L \cdot \|y - x\|_2^2$



Deriving Gradient Descent

- f is L-smooth $\Leftrightarrow \|\nabla f(x) - \nabla f(y)\|_2 \leq L \cdot \|x - y\|_2$ for all $x, y \in \mathbb{R}^n$
- $\Rightarrow |f(y) - [f(x) + \nabla f(x)^\top (y - x)]| \leq L \cdot \|y - x\|_2^2$
- $\Rightarrow f(y) \leq U_x(y) \stackrel{\text{def}}{=} f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|_2^2$
- Note: $U_x(x) = f(x)$ so set $x_{k+1} = \min_x U_{x_k}(x)$
- $\nabla U_{x_k}(x) = \nabla f(x_k) + L(x - x_k)$
- $x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$

***Note**: only need upper bound!*

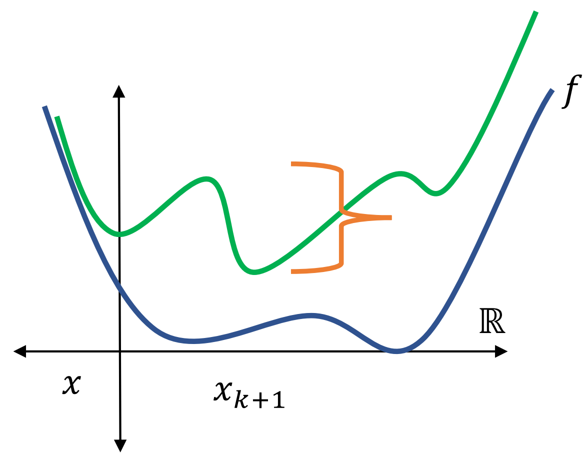


A General Framework

- **Upper Bound Oracle**

query $x \in \mathbb{R}^n$ → **oracle** → output $U_x : \mathbb{R}^n \to \mathbb{R}$
s.t. $U_x(x) = f(x)$
$U_x(y) \geq f(y)$ for all y

- **Algorithm**
  - $x_{k+1} = \min_x U_{x_k}(x)$



Analysis

- $f(x_{k+1}) - f(x_k) \leq U_{x_k}(x_{k+1}) - f(x_k)$
- $= \min_x U_{x_k}(x) - U_{x_k}(x_k) = \Delta_k$
- If $k \geq [f(x_0) - f_*]/\epsilon$ then some $\Delta_k \geq -\epsilon$

For sooth functions $\Delta_k = -\frac{1}{2L}\|\nabla f(x_k)\|_2^2$.

Make at least as much progress as make from minimizing upper bound.

How obtain upper bound?

Lemma: For $x_\alpha \stackrel{\text{def}}{=} x + \alpha(y - x)$

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \int_0^1 \left(\nabla f(x_\alpha) - \nabla f(x)\right)^\top (y - x) d\alpha$$

Definition: f is twice differentiable at x if $\boldsymbol{H} \in \mathbb{R}^{n \times n}$ satisfies

$$\lim_{h \to 0} \frac{\|\nabla f(x + h) - [\nabla f(x) + \boldsymbol{H}h]\|}{\|h\|} = 0 \quad .$$

Implies $\boldsymbol{H} = \nabla^2 f(x)$ is the Hessian of f at x with $[\nabla^2 f(x)]_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(x)$.

Lemma: If f is twice differentiable then for $x_\alpha = x + \alpha(y - x)$

$$\nabla f(y) - \nabla f(x) = \int_0^1 \nabla^2 f(x_\alpha)(y - x) d\alpha \quad .$$

Why Useful?

- *Find it often easier to bound Hessian than difference*
- *(Similarly find it easier to bound gradient to prove Lipschitz)*

$\Leftrightarrow \lambda_{\max}\left(\nabla^2 f(x)\right) \leq L$ for all $i \in [n]$

Lemma: if $z^\top \nabla f(x) z \leq L\|z\|_2^2$ for all x, z

- $f(y) = f(x) + \nabla f(x)^\top (y - x) + \int_0^1 \int_0^t (y - x)^\top \nabla^2 f(x_\alpha)(y - x) d\alpha\, dt$

$$\leq f(x) + \nabla f(x)^\top (y - x) + \int_0^1 \int_0^t L\|y - x\|_2^2 d\alpha\, dt$$

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

Weaker assumption than smoothness

- Claim: twice differentiable f is L-smooth $\Leftrightarrow |z^\top \nabla^2 f(x) z| \leq L\|z\|_2^2$ for all x,

$\Leftrightarrow \left|\lambda_i\left(\nabla^2 f(x)\right)\right| \leq L$ for all $i \in [n]$

Detailed Lecture Plan

Lipschitz

- Recap: Lipschitz function minimization
- High dimensional upper / lower bounds
- Properties, characterizations

Smooth

- Recap: critical point computation of smooth functions
- Smooth function minimization lower bound
- General minimization strategy

Convex

- Introduce assumptions enabling efficient computation of ϵ-optimal points

Tuesday

Design and analyze algorithms.

Why Useful?

- Find it often easier to bound Hessian than difference
- (Similarly find it easier to bound gradient to prove Lipschitz)

Lemma: if $z^\top \nabla f(x) z \leq L\|z\|_2^2$ for all x, z $\Leftrightarrow \lambda_{\max}\left(\nabla^2 f(x)\right) \leq L$ for all $i \in [n]$

- $f(y) = f(x) + \nabla f(x)^\top (y-x) + \int_0^1 \int_0^t (y-x)^\top \nabla^2 f(x_\alpha)(y-x) d\alpha\, dt$

$$\leq f(x) + \nabla f(x)^\top (y-x) + \int_0^1 \int_0^t L\|y-x\|_2^2 d\alpha\, dt$$

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

<u>Weaker assumption than smoothness</u>

- Claim: twice differentiable f is L-smooth $\Leftrightarrow |z^\top \nabla^2 f(x) z| \leq L\|z\|_2^2$ for all x,

What if want more than critical points? (i.e. ϵ-optimal points)

$\Leftrightarrow \left|\lambda_i\left(\nabla^2 f(x)\right)\right| \leq L$ for all $i \in [n]$

Assumptions for obtaining ϵ-optimal point

- So far, smoothness just ensures progress relative to norm of gradient
- Problem: this progress might not be large relative to suboptimality
- We are only using upper bounds on function now
- Can we close gap by assuming lower bounds?

Notion #1

- 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

- f is differentiable and $f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2} \|y - x\|_2^2 \stackrel{\text{def}}{=} L_y(x)$

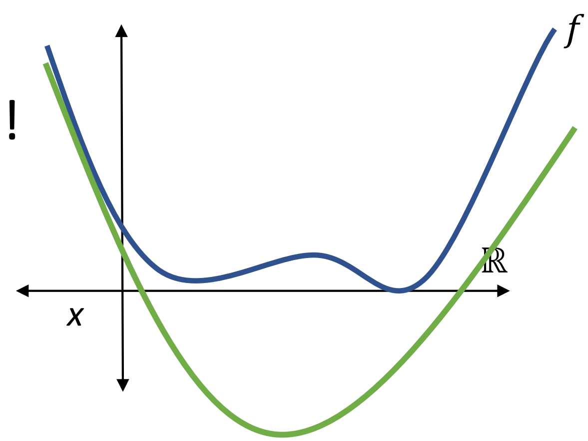
Assumptions for obtaining ϵ-optimal point

Notion #2

- f differentiable and $f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2} \|y - x\|_2^2$

Note

- $\nabla f(x) = 0 \Rightarrow f(x) = f_*$ under notion #2!
- Have solution if gradient descent step doesn't help



Another assumption

Next Week
discuss more and design and analyze algorithms.

- Motivation: what stops gradient descent from converging to minimum?

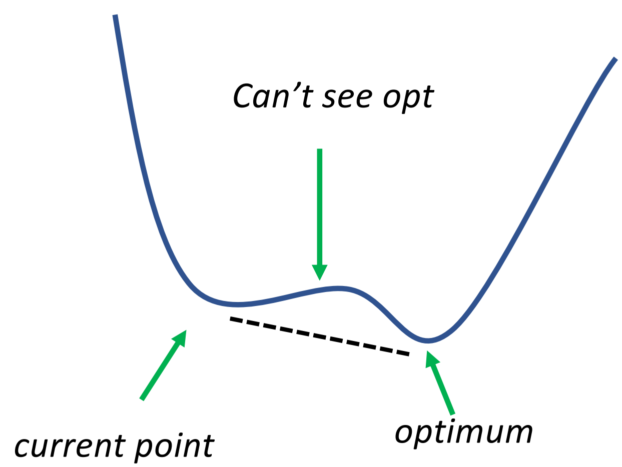
Notion #3: μ-strong convexity

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

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

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

Theorem: if twice differentiable all notions are equivalent!!!



Detailed Lecture Plan

Lipschitz

- Recap: Lipschitz function minimization
- High dimensional upper / lower bounds
- Properties, characterizations

Smooth

- Recap: critical point computation of smooth functions
- Smooth function minimization lower bound
- General minimization strategy

Convex

- Introduce assumptions enabling efficient computation of ϵ-optimal points

Tuesday

Design and analyze algorithms.