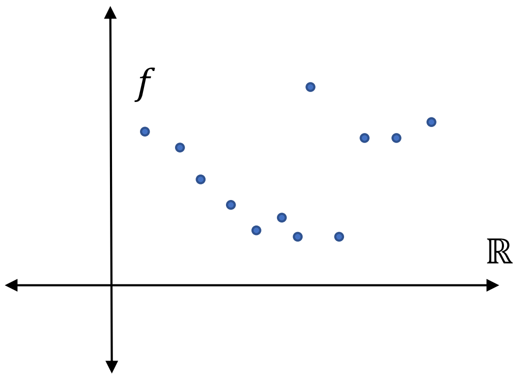
Introduction to Optimization Theory

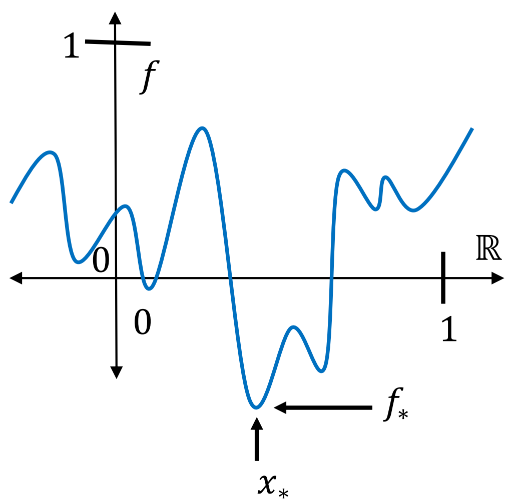
Lecture #2 - 9/17/20

MS&E 213 / CS 2690



Aaron Sidford

sidford@stanford.edu



Lecture Plan

Recap
- Oracles, minimization, efficiency, and iterative methods

Material
- Continuity, smoothness, and critical points

Tuesday
- Continuity, ϵ-nets, and lower

Recap

Goal

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

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

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

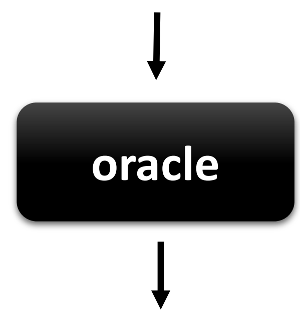
provably efficiently with few assumptions

Access to f?

- Through an "oracle"

query

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



output

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

Recap

$\min_{x \in S \subseteq \mathbb{R}^n} f(x)$ provably efficiently with few assumptions

Minimize? Progress Measure?

ϵ-(sub)optimal point or a point with **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. $\|\nabla f(x)\|_2 \leq \epsilon$ where $\|y\|_2 \stackrel{\text{def}}{=} \sqrt{\sum_{i \in [n]} y_i^2}$

Efficency?

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

Recap

Iterative Method Approach

- Start at initial point x_0
- For $t = 0, \dots, T - 1$
  - Query oracle
  - Take "local step" to obtain x_{t+1}
  - Repeat
- Output aggregation of the x_t

e.g.

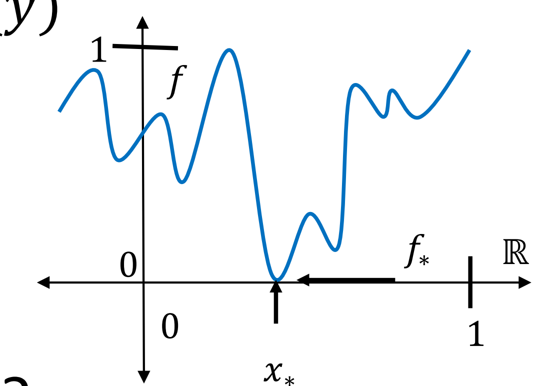
- **Last iterate**: x_{T-1}
- **Average iteration**: $\frac{1}{T} \sum_{k \in [T-1]} x_k$

Analysis

- Oracle complexity = # iterations
- Runtime = # iterations * cost per iteration (iteration complexity)

Recap: an impossible setting

- $f: \mathbb{R} \to \mathbb{R}$ *(one dimensional)*
- Have evaluation oracle *(can compute $f(x)$ with 1 query)*
- Promised $\exists x_* \in [0,1]$ such that $f(x) = f_* = \inf_{y \in \mathbb{R}} f(y)$
- Promised $f(x) \in [0,1]$ for all $x \in \mathbb{R}$
- Goal: compute 1/2-optimal point
  - i.e. compute x with $f(x) \le f(x_*) + 1/2$

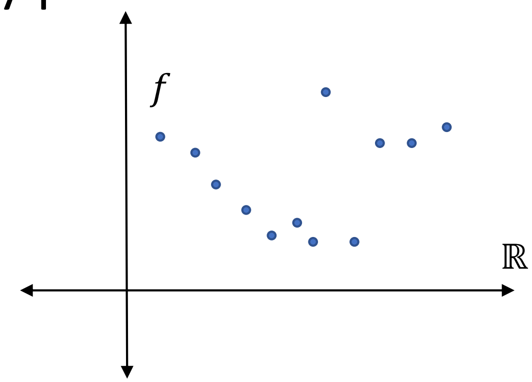


- **Question**: what oracle complexity achievable?
- **Answer**: ∞ is optimal

We will discuss lower bound a little more formally next week.

Recap

Problem: oracle gives only pointwise information, no local information.



Solution:

- This is a class on *continuous* optimization
- **Today**: assume more structure and analyze a working method

Lecture Plan

Recap ✓

- Oracles, minimization, efficiency, and iterative methods

Material

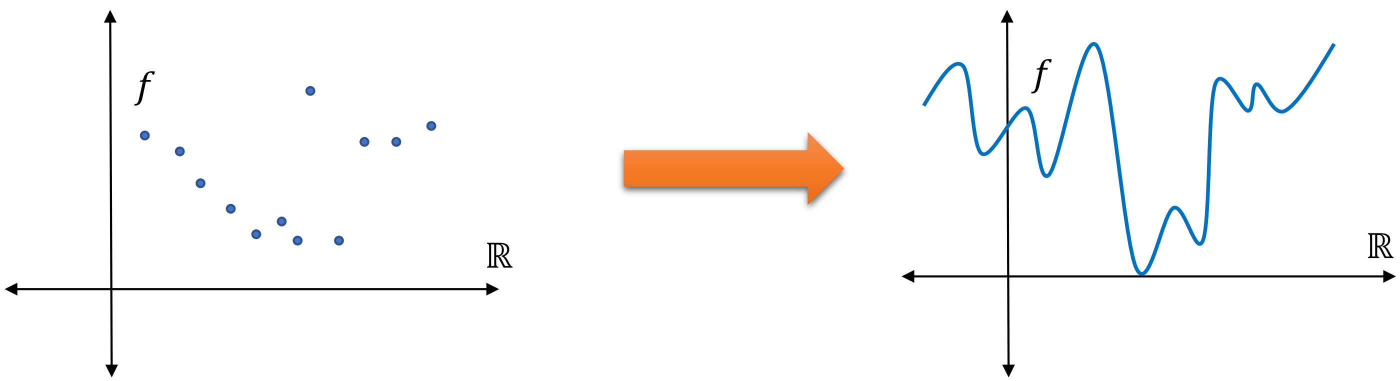
- Continuity, smoothness, and critical points

Tuesday

- Continuity, ϵ-nets, and lower

Continuous Function Minimization

Problem: oracle gives only pointwise information, no local information.



Idea:

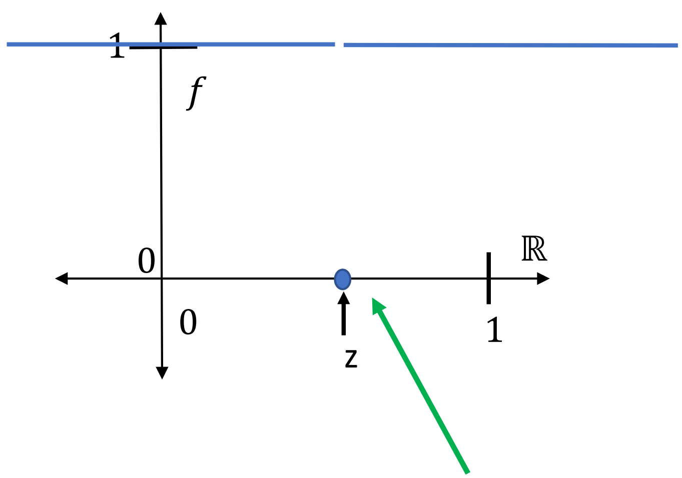
- Make assumptions so oracle give some local / global information
- Frequent assumption: continuity

Question: is continuity enough?

- $f: \mathbb{R} \to \mathbb{R}$ *(one dimensional)*
- Have evaluation oracle *(can compute $f(x)$ with 1 query)*
- Promised $\exists x_* \in [0,1]$ such that $f(x) = f_* = \inf_{y \in \mathbb{R}} f(y)$
- Promised $f(x) \in [0,1]$ for all $x \in \mathbb{R}$
- Promised f is continuous: $\lim_{y \to x} f(y) = f(x)$
- Goal: compute 1/2-optimal point
  - i.e. compute x with $f(x) \leq f(x_*) + 1/2$

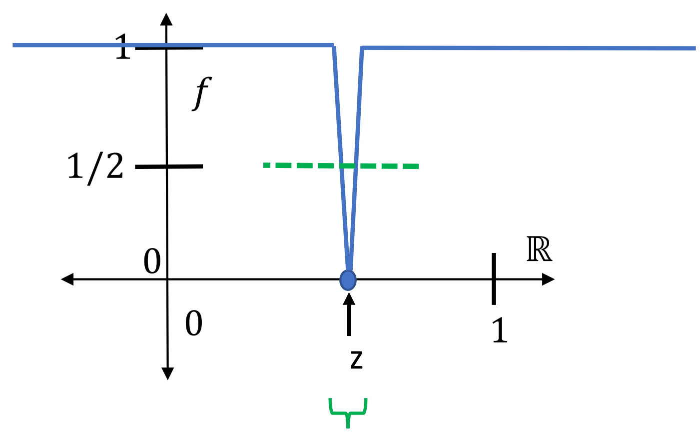
- **Question**: what oracle complexity achievable?
- **Answer**: ∞ is optimal

Proof Sketch by Picture



Only ½-optimal point

Problem: can make slope arbitrarily large



Can make arbitrarily small

Idea: assume bounded slope / quantify continuity

Quantifying Continuity

Are many different assumptions that could be made. Will discuss one family of assumptions for now.

Lipschitz Continuous:

- $f: \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to a norm $\|\cdot\|$ if and only if $|f(x) - f(y)| \leq L \cdot \|x - y\|$

Recall: $\|\cdot\|: \mathbb{R}^n \to \mathbb{R}$ is a norm if and only if $\forall \alpha \in \mathbb{R}$ and $x, y \in \mathbb{R}^n$

- $\|\alpha x\| = |\alpha| \cdot \|x\|$ (*absolute homogeneity*)
- $\|x + y\| \leq \|x\| + \|y\|$ (*triangle inequality*)
- $\|x\| = 0 \Leftrightarrow x = 0$ (*called a "semi-norm" if doesn't necessarily hold*)

Examples:

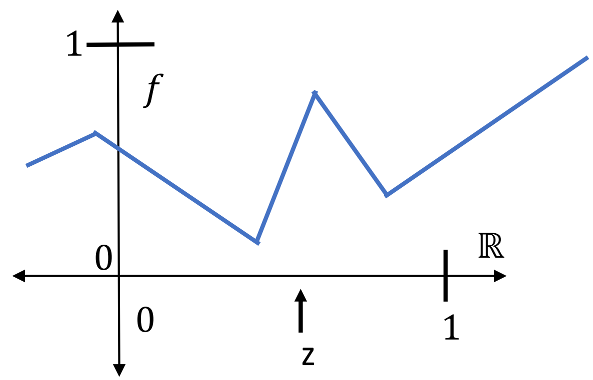
- $\|x\|_2 \stackrel{\text{def}}{=} \sqrt{\sum_i x_i^2}$ (Euclidean or ℓ_2 norm) ← Default if unspecified
- $\|x\|_p \stackrel{\text{def}}{=} (\sum_i |x_i|^p)^{1/p}$ (p-norm or ℓ_p-norm for $p \geq 1$)
- $\|x\|_\infty \stackrel{\text{def}}{=} \max_i |x_i|$

We will see many more in class.

Can we minimize Lipschitz functions?

We will discuss Tuesday

- $f: \mathbb{R} \to \mathbb{R}$
- Have evaluation oracle
- Promised $\exists x_* \in [0,1]$ such that $f(x) = f_* = \inf_{y \in \mathbb{R}} f(y)$
- Promised $f(x) \in [0,1]$ for all $x \in \mathbb{R}$
- Promised f is L-Lipschitz (with respect to ℓ_2)
- **Goal**: compute 1/2-optimal point
  - i.e. compute x with $f(x) \leq f(x_*) + 1/2$



L-Lipschitz implies that slope of lines is at most L.

Interesting, general, useful, recently popularized topic.

Today: Smoothness & Critical Points

Recall: $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}^n$ if exists $g \in \mathbb{R}^n$ such that

(Note: choice of norm does not affect definition)

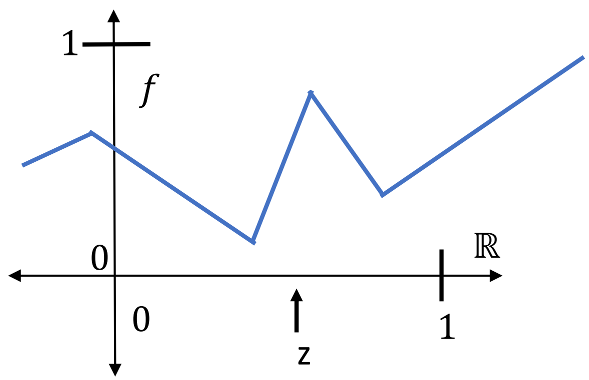
$$\lim_{h \to 0} \frac{|f(x+h) - [f(x) + \mathrm{g}^\top h|}{\|h\|_2} = 0$$

Further, when this holds, $g = \nabla f(x)$, i.e. $g_i = \frac{\partial}{\partial x_i} f(x)$.

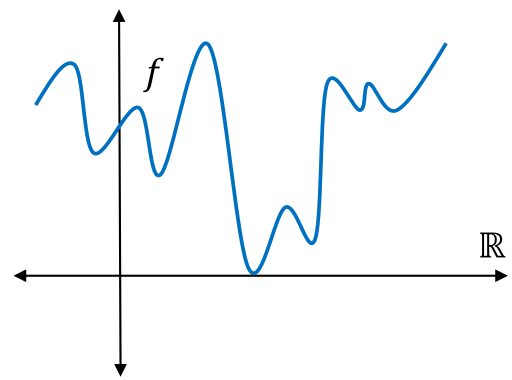
Smoothness: f is *L-smooth* (with respect to ℓ_2) if differentiable and for all $x, y \in \mathbb{R}^n$ we have $\|\nabla f(x) - \nabla f(y)\|_2 \le L \cdot \|x - y\|_2$.

Picture

We will characterize more later



Lipschitz
(bounded slope)
(bounded 1st derivatives)

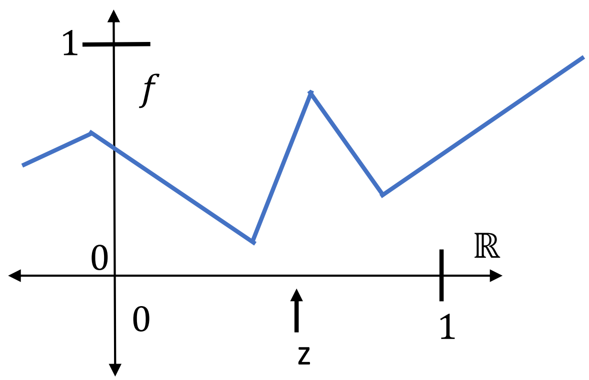


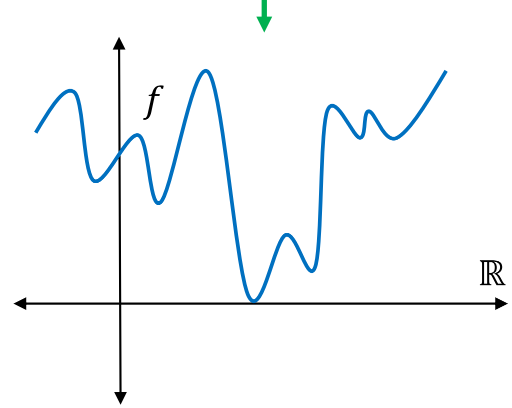
Smooth
(bounded curvature)
(bounded 2nd derivative)

Problem

Discuss more next week.

Many local queries are required to approximately find approximate minimizer. (Discuss more Tuesday)

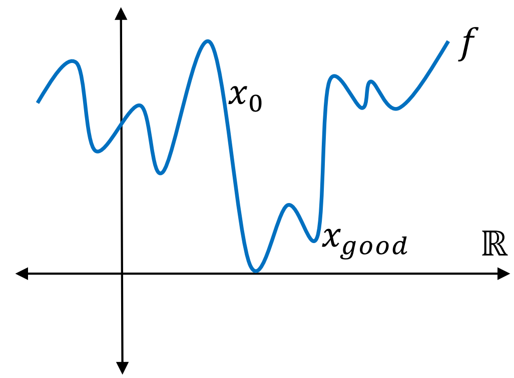




Today: seek critical points

Smooth
(bounded curvature)
(bounded 2^{nd} derivative)

Today's Setting



- **Problem**: Assume $f: \mathbb{R}^n \to \mathbb{R}$ is L-smooth and given some $x_0 \in \mathbb{R}^n$ such that *function error*, $f(x_0) - f_*$, is bounded.

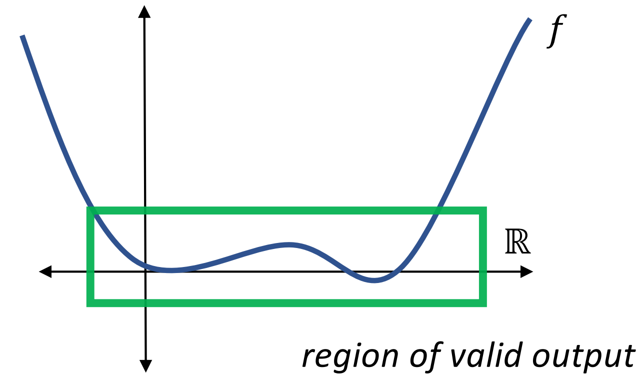
- **Oracle**: gradient oracle

<u>query</u> $x \in \mathbb{R}^n$ → **oracle** → <u>output</u> $\nabla f(x) \in \mathbb{R}^n$

- **Question**: how many queries needed to compute ϵ-critical point, i.e. $\|\nabla f(x)\|_2 \leq \epsilon$?

Why?

Critical Point Computation



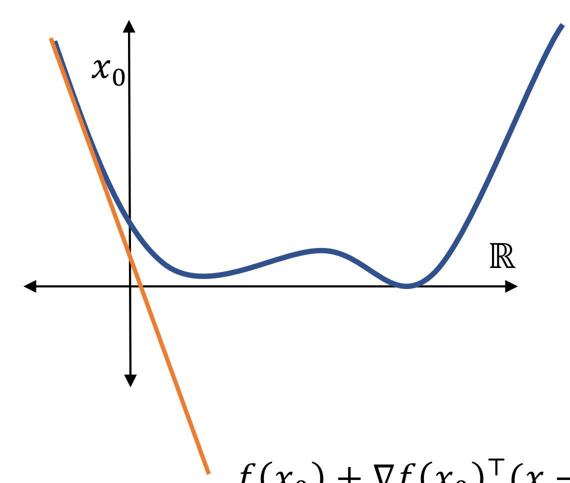
Idea

- Locally

$$f(x+h) \approx f(x) + \nabla f(x)^\top h$$

- So if $h = -\eta \nabla f(x)$ for small η (*or more broadly* $\nabla f(x)^\top h < 0$ for small enough h) function value decreases!

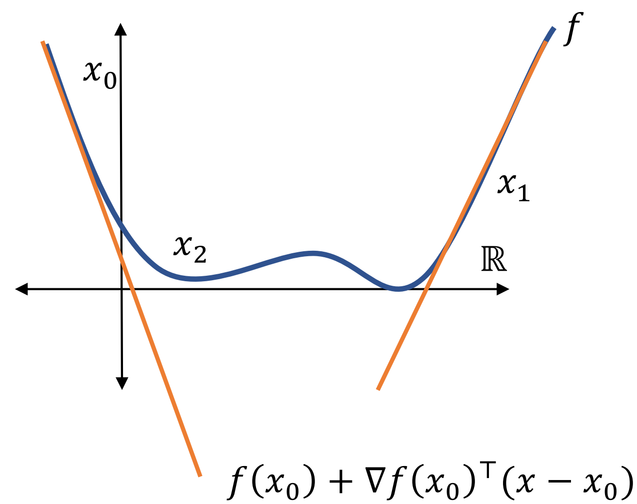
- **Hope**: smoothness makes progress substantial whenever non-critical.



Gradient Descent Method for Critical Points

Algorithm / Method

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



Step Size

- $\eta_k =$ "step size"
- Many step size schemes
- Often in this class, fixed step size, $\eta_k = \eta$

Convergence Analysis

Theorem: Gradient descent on L-smooth function with $\eta =$ (TBD) computes an ϵ-critical point with $\leq 2L[f(x_0) - f_*]/\epsilon^2$ queries

Lemma: If f is L-smooth and $y = x - \eta \nabla f(x)$ then

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

Plan: (1) Prove theorem using lemma (2) prove lemma (3) build some more intuition

Goal

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

Tool

Lemma: If f is L-smooth and $y = x - \eta\nabla f(x)$ then $\left|f(y) - [f(x) - \eta\|\nabla f(x)\|_2^2]\right| \leq \frac{\eta^2 L}{2}\|\nabla f(x)\|_2^2$

- $x_{k+1} = x_k - \eta\nabla f(x_k)$ so apply lemma with $x = x_k$ and $y = x_{k+1}$
- $f(x_{k+1}) \leq f(x_k) - \eta\|\nabla f(x_k)\|_2^2 + \frac{\eta^2 L}{2}\|\nabla f(x_k)\|_2^2$
- "best" η?
  - $g(\eta) = -\eta + \frac{\eta^2 L}{2}$ has $g'(\eta) = -1 + \eta L$. minimizer $\eta = \frac{1}{L}$
- $\Rightarrow f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2$

Function value decreases by amount depending on norm of gradient!!

Since function value can only decrease by $f(x_0) - f_*$ must find a small gradient!!

Goal

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

Tool

Lemma: If f is L-smooth and $y = x - \eta \nabla f(x)$ then

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

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

Optimal?

Open for decades (and when I first taught this class)

[CDHS18]
Yes !!!

(in worst-case up to constants, if depend on nothing else)

Proof Strategy

Goal

Lemma: If f is L-smooth and $y = x - \eta \nabla f(x)$ then

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

- **Goal**: analyze f change for "gradient descent step" $y = x - \eta \nabla f(x)$
- **Broader Goal**: analyze change in f between two points

- **How?** (*common proof strategy this course*)
  - Integrate: Taylor expansion
  - Bound: Cauchy Schwarz inequality

(A little multivariable-calculus recap, slower today.
Faster / see notes / ask questions later classes.)

Goal

Lemma: If $f: \mathbb{R}^n \to \mathbb{R}$ is differentiable, $x, y \in \mathbb{R}^n$, $x_t = x + t(y - x)$ for $t \in [0,1]$:

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

- Let $g(t) = f(x_t)$ for all $t \in [0,1]$
- $f(y) - f(x) = g(1) - g(0) = \int_0^1 g'(\alpha) d\alpha$ (fundamental theorem calculus)
- Since f is differentiable

$$0 = \lim_{h \to 0} \frac{|f(x_\alpha + h) - [f(x_\alpha) + \nabla f(x_\alpha)^\top h]|}{\|h\|_2}$$

- Let $h = t(y - x)$ for $t \to 0$ so $f(x_\alpha + h) = g(\alpha + t)$

$$0 = \lim_{t \to 0} \frac{|g(\alpha + t) - [g(\alpha) + t \cdot \nabla f(x_\alpha)^\top (y - x)]|}{\|t(y - x)\|_2}$$

- $\Rightarrow g'(\alpha) = \nabla f(x_\alpha)^\top (y - x)$
- $\Rightarrow f(y) - f(x) = \int_0^1 \nabla f(x_\alpha)^\top (y - x) d\alpha$ ☺

Lemma: If $f: \mathbb{R}^n \to \mathbb{R}$ is differentiable, $x, y \in \mathbb{R}^n$, $x_t = x + t(y - x)$ for $t \in [0,1]$:

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

- **New Goal**: Upper bound $\left| \int_0^1 \left(\nabla f(x_\alpha) - \nabla f(x)\right)^\top (y - x) d\alpha \right|$
- **Lemma**: (Cauchy-Schwarz Inequality) $\forall x, y \in \mathbb{R}^n$, $|x^\top y| \leq \|x\|_2 \|y\|_2$
- **Proof**:
  - $\|x\|_2^2 \cdot \|y\|_2^2 - |x^\top y|^2 = \left(\sum_i x_i^2\right)\left(\sum_j y_j^2\right) - \left(\sum_i x_i y_i\right)\left(\sum_j x_j y_j\right)$
  - $= \sum_i \sum_j \left(x_i^2 y_j^2 - x_i y_i x_j y_j\right)$
  - $= \sum_{i<j} \left(x_i^2 y_j^2 + x_j^2 y_i^2 - 2 x_i y_i x_j y_j\right) = \sum_{i<j} \left(x_i y_j - x_j y_i\right)^2 \geq 0$ ☺

***Strategy**: to show $a \leq b$ for $a, b \geq 0$, suffices to show $b^2 \geq a^2$ or equivalently $b^2 - a^2 \geq 0$.*

Lemma: If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, $x, y \in \mathbb{R}^n$, $x_t = x + t(y - x)$ for $t \in [0,1]$:

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

- **New Goal**: Upper bound $\left|\int_0^1 \left(\nabla f(x_\alpha) - \nabla f(x)\right)^\top (y - x) d\alpha\right|$
- **Lemma**: (Cauchy-Schwarz Inequality) $\forall x, y \in \mathbb{R}^n$, $|x^\top y| \leq \|x\|_2 \|y\|_2$
- **Corollary:** if f is L-smooth
  - $\left|\int_0^1 \left(\nabla f(x_\alpha) - \nabla f(x)\right)^\top (y - x) d\alpha\right|$
  - $\leq \int_0^1 \left|\left(\nabla f(x_\alpha) - \nabla f(x)\right)^\top (y - x)\right| d\alpha$
  - $\leq \int_0^1 \|\nabla f(x_\alpha) - \nabla f(x)\|_2 \|y - x\|_2 d\alpha$
  - $\leq \int_0^1 L \|x_\alpha - x\|_2 \|y - x\|_2 d\alpha$
  - $= \int_0^1 L\alpha \|y - x\|_2^2 d\alpha = \frac{L}{2} \|y - x\|_2^2$

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

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

☺

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

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

Simply apply with $y = x - \eta \nabla f(x)$

Lemma: If f is L-smooth and $y = x - \eta \nabla f(x)$ then

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

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

☺

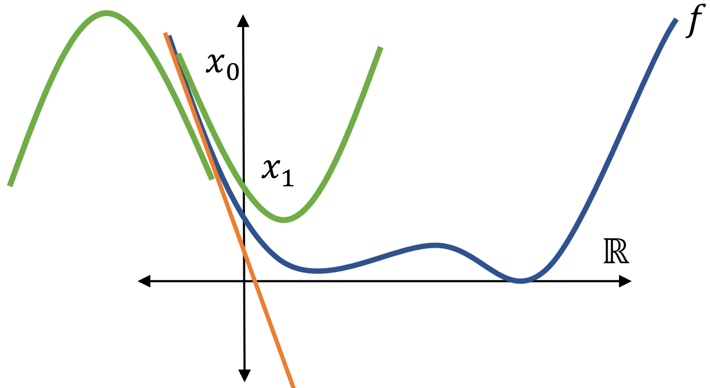
Picture?

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? $x_{k+1} = \min_x U_{x_k}(x)$!!!

Will build on this idea later in the course

Lecture Plan

Recap ✓

- Oracles, minimization, efficiency, and iterative methods

Material ✓

- Continuity, smoothness, and critical points

Tuesday

- Continuity, ϵ-nets, and lower