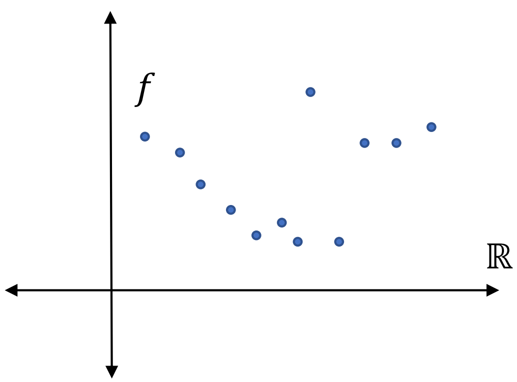
Introduction to Optimization Theory

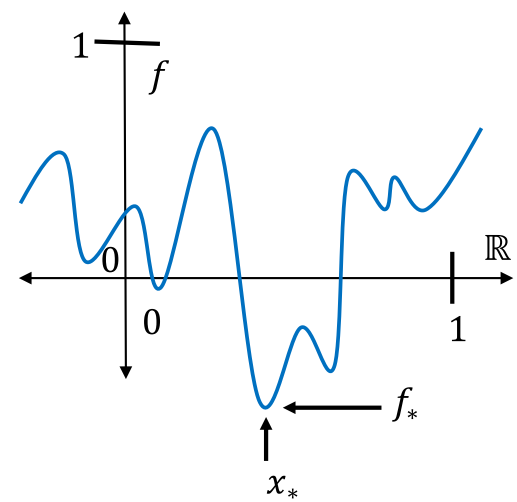
Lecture #5 - 9/28/20

MS&E 213 / CS 2690



Aaron Sidford

sidford@stanford.edu



Plan for Today

Recap

- Gradient descent for smooth function
- Notions of convexity

Convexity

- Prove smoothness / convexity equivalences
- Example functions
- Implications of assumptions

Algorithm

- Gradient descent
- Algorithm analysis

Recap

Problem

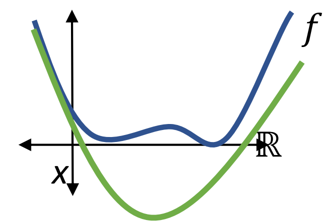
$$\min_{x \in \mathbb{R}^n} f(x)$$

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	$\Theta(L/\epsilon)$
$x_* \in [0,1]^n$, 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	$O\left(\frac{L(f(x_0) - f_*)}{\epsilon^2}\right)$

Today

What if want an ϵ-optimal point with no dependence on dimension?

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
- $\Leftrightarrow \lambda_{min}\left(\nabla^2 f(x)\right) \geq \mu$

Variational characterization of eigenvalues?

Notion #2: Quadratic Lower Bounds

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

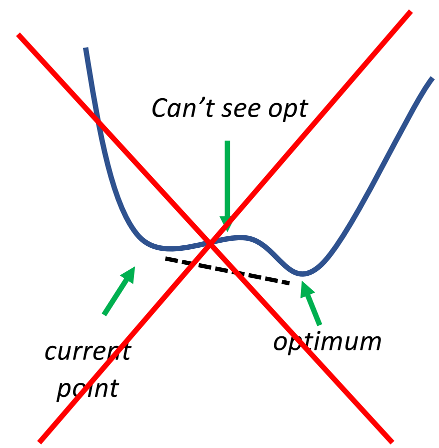
Notion #3: μ-strongly convex with respect to $\|\cdot\|$ (*by default* $\|\cdot\|_2$)

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

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



Plan for Today

Recap

- Gradient descent for smooth function
- Notions of convexity

Convexity

- Smoothness / convexity equivalences
- Example functions
- Implications of assumptions

Algorithm

- Gradient descent
- Algorithm analysis

Equivalent Notions of Convexity

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
- $\Leftrightarrow \lambda_{min}\left(\nabla^2 f(x)\right) \geq \mu$

Great for proving convexity

Notion #2: Quadratic Lower Bounds
- f is differentiable and $f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2} \|y - x\|^2 \stackrel{\text{def}}{=} L_y(x)$

Great for designing algorithms

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 $t \in [0,1]$

Great for visualizing

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

Helpful Technical Lemma

- *Implies equivalence of upper and lower bounds implied by smoothness and Hessian eigenvalue bound.*
- *Implies some convexity equivalences*

For all $\alpha, \beta \in \mathbb{R} \cup \{\pm\infty\}$ and twice differentiable $f: \mathbb{R}^n \rightarrow \mathbb{R}$ and any norm $\|\cdot\|$ the following three conditions are equivalent:

- $\frac{\alpha}{2}\|x-y\|^2 \leq f(y) - [f(x) + \nabla f(x)^\top (y-x)] \leq \frac{\beta}{2}\|x-y\|^2$ for all $x, y \in \mathbb{R}^n$
- $\frac{\alpha}{2}\|x-y\|^2 \leq \left(\nabla f(x) - \nabla f(y)\right)^\top (x-y) \leq \frac{\beta}{2}\|x-y\|^2$ for all $x, y \in \mathbb{R}^n$
- $\alpha\|z\|^2 \leq z^\top \nabla^2 f(x) z \leq \beta\|z\|^2$ for al $x, z \in \mathbb{R}^n$

Proof is technical (but useful) and in smoothness notes.

Corollary

For all $\mu \geq 0$ and twice differentiable $f\colon \mathbb{R}^n \to \mathbb{R}$ and any norm $\|\cdot\|$ the following three conditions are equivalent:

- $f(y) \geq f(x) + \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|x-y\|^2$ for all $x, y \in \mathbb{R}^n$
- $\left(\nabla f(x) - \nabla f(y)\right)^\top (x-y) \geq \frac{\mu}{2}\|x-y\|^2$ for all $x, y \in \mathbb{R}^n$
- $z^\top \nabla^2 f(x) z \geq \mu \|z\|^2$ for al $x, z \in \mathbb{R}^n$

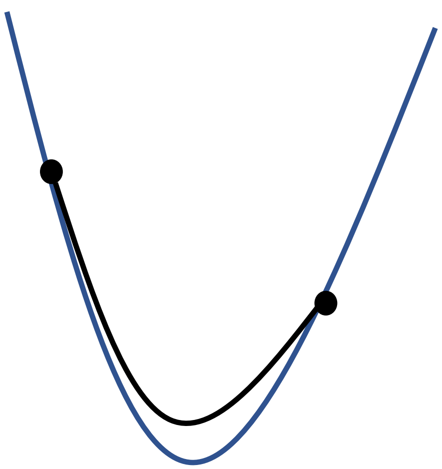
When f is differentiable the first two conditions are equivalent to

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

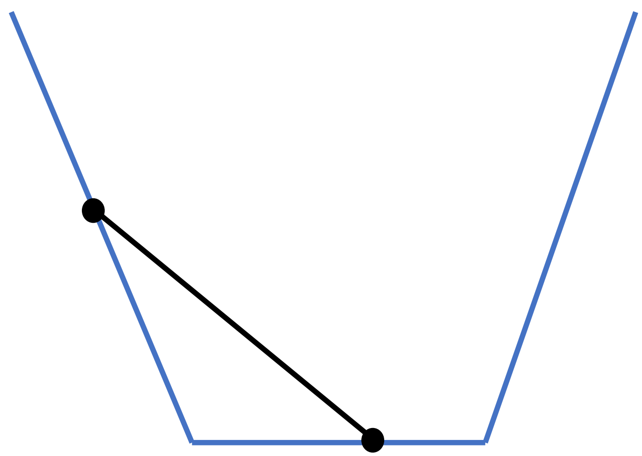
For all x, y and $t \in [0,1]$. (See Note.)

What do μ-strongly convex functions look like?

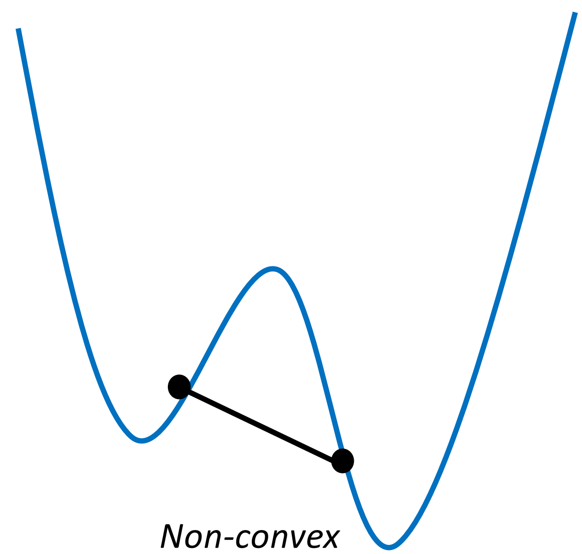
- $f(ty + (1-t)x) \leq t \cdot f(y) + (1-t) \cdot f(x) - \frac{\mu}{2} t(1-t) \|x - y\|_2^2$



Strongly convex



Convex



Non-convex

Example Convex Functions

$f(ty + (1-t)x) \leq t \cdot f(y) + (1-t) \cdot f(x)$

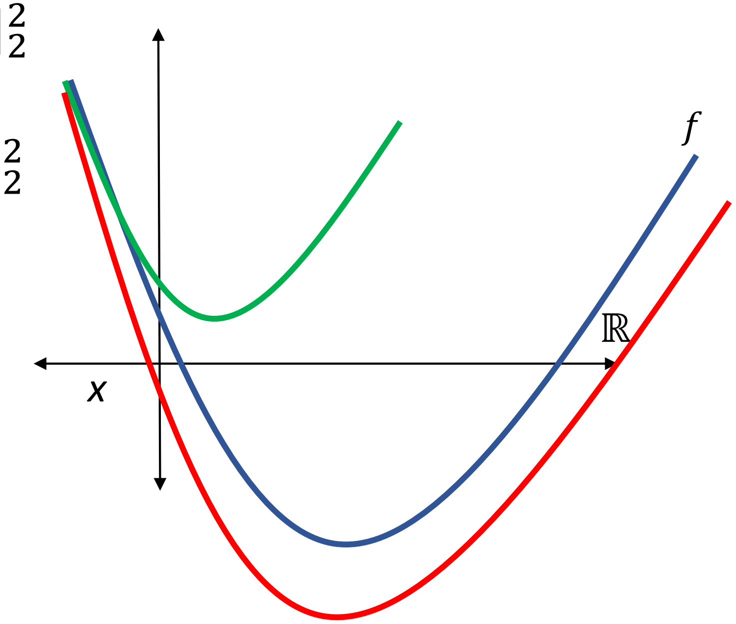
- $f(x) = \frac{1}{2}\|Ax - b\|_2^2$
- $f(x) = \|x\|$
- $f(x) = \exp(x)$
- $f(x) = x^p$ for even p
- $f(x) = -\log x$ for $x \geq 0$
- $f(x) = x \log x$
- …

- $f(x) = g(x) + h(x)$ for convex g and h
- $f(x) = f(\boldsymbol{A}x)$ for convex f
- $f(x) = c \cdot f(x)$ for convex f and $c \geq 0$
- …

Goal: Minimize 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?*

Plan for Today

Recap

- Gradient descent for smooth function
- Notions of convexity

Convexity

- Smoothness / convexity equivalences
- Example functions
- Implications of assumptions

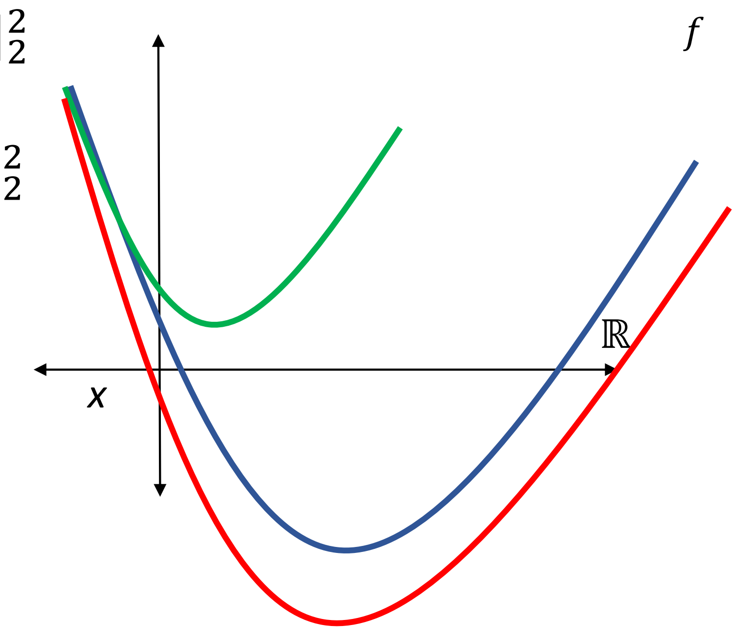
Algorithm

- Gradient descent
- Algorithm analysis

Goal: Minimize 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

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



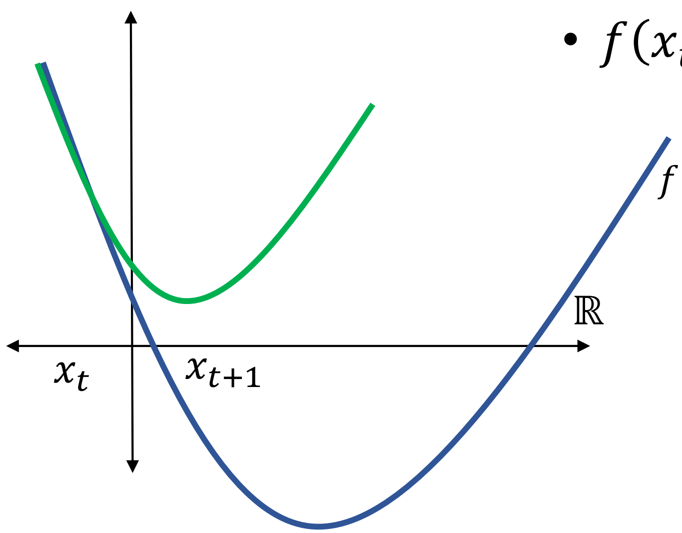
Question*: is this assumption and a gradient oracle enough to obtain dimension independent efficient algorithms for ϵ-optimal points?*

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*

Gradient Descent!

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



Upper Bound Analysis

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

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*

Gradient Descent!

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

Upper Bound Analysis

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

Progres Meassures

- $\|\nabla f(x)\|_2^2$ - norm of gradient
- $f(x) - f_*$ where $f_* = \inf_{x \in \mathbb{R}^n} f(x)$ - function error
- $\|x - x_*\|_2^2$ - for minimizer x_* (*i.e.* $f(x_*) = f_*$)

Question

How lower bound?

Smoothness Implication

- $\|\nabla f(x)\|_2^2$ - *norm of gradient*
- $f(x) - f_*$ *where* $f_* = \inf_{x\in\mathbb{R}^n} f(x)$ - *function error*
- $\|x - x_*\|_2^2$ - *for minimizer* x_* *(i.e.* $f(x_*) = f_*$*)*

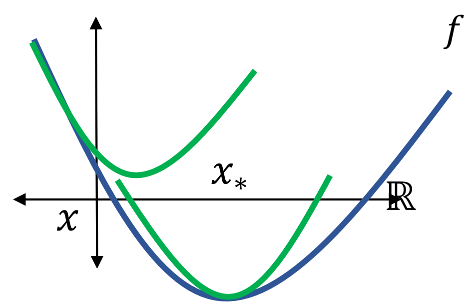
Lemma If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable then $\nabla f(x_*) = 0$. If f is L-smooth then

$$\frac{1}{2L}\|\nabla f(x)\|_2^2 \leq f(x) - f(x_*) \leq \frac{L}{2}\|x - x_*\|_2^2$$

Proof

- Differentiability and $\nabla f(x) \neq 0 \Rightarrow f\big(x - \eta\nabla f(x)\big) < f(x)$ for small η
- $f_* \leq f(x - (1/L)\nabla f(x)) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2$
- $f(x) \leq f(x_*) + \nabla f(x_*)^\top(x - x_*) + \frac{L}{2}\|x - x_*\|_2^2$

$$= f(x_*) + \frac{L}{2}\|x - x_*\|_2^2$$

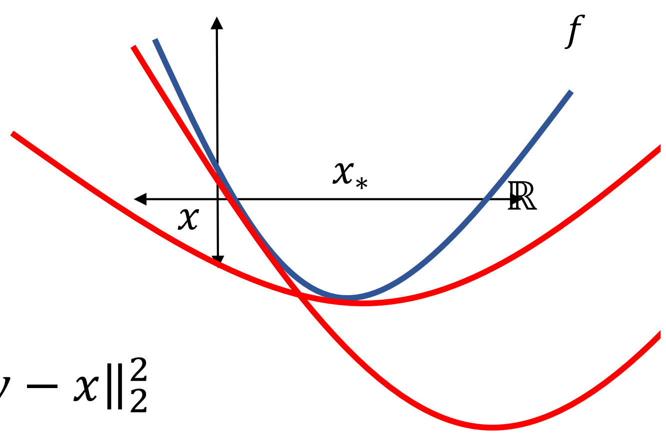


Convexity Implication

- $\|\nabla f(x)\|_2^2$ *- norm of gradient*
- $f(x) - f_*$ *where* $f_* = \inf_{x \in \mathbb{R}^n} f(x)$ *- function error*
- $\|x - x_*\|_2^2$ *- for minimizer* x_* *(i.e.* $f(x_*) = f_*$*)*

<u>**Lemma**</u> If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and μ-strongly convex

$$\frac{1}{2\boldsymbol{\mu}}\|\nabla f(x)\|_2^2 \geq f(x) - f(x_*) \geq \frac{\boldsymbol{\mu}}{2}\|x - x_*\|_2^2$$



<u>**Proof**</u>

- $$f(x) \geq f(x_*) + \nabla f(x_*)^\top (x - x_*) + \frac{L}{2}\|x - x_*\|_2^2$$
  $$= f(x_*) + \frac{L}{2}\|x - x_*\|_2^2$$
- $$f(x_*) \geq \min_y L_x(y) = \min_y f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2$$
  $$= \min_y f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2 + \frac{\mu}{2}\left\|y - \left(x - (1/\mu)\nabla f(x)\right)\right\|_2^2$$
  $$= f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2$$

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.

$$\frac{1}{2\mu}\|\nabla f(x)\|_2^2 \geq f(x) - f(x_*) \geq \frac{\mu}{2}\|x - x_*\|_2^2$$

Non-strongly Convex Case ($\mu = 0$)

Lemma If f is differentiable and convex then for all minimizers x_*

$$f(x) - f_* \leq \|\nabla f(x)\|_2 \cdot \|x - x_*\|_2^2$$

Proof

- $f(x_*) \geq f(x) + \nabla f(x)^\top (x_* - x)$
  $\geq f(x) - \|\nabla f(x)\|_2 \cdot \|x_* - x\|_2$

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 $\epsilon_{k+1} \leq \epsilon_k - \frac{1}{2L} \left(\frac{\epsilon_k}{D}\right)^2$
- $\Rightarrow \frac{1}{\epsilon_k} \leq \frac{1}{\epsilon_{k+1}} - \frac{\epsilon_k}{2LD^2 \epsilon_{k+1}} \leq \frac{1}{\epsilon_{k+1}} - \frac{1}{2LD^2}$
- $\Rightarrow \frac{1}{\epsilon_k} \geq \frac{1}{\epsilon_0} + \frac{k}{2LD^2}$
- $\epsilon_0 \leq \frac{L}{2} D^2$ $(f(x_k) - f_* \leq \frac{L}{2} \|x_k - x_*\|_2^2)$

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

Optimal?

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 $\|\cdot\|_2$

Plan for Today

Recap

- Gradient descent for smooth function
- Notions of convexity

Convexity

- Smoothness / convexity equivalences
- Example functions
- Implications of assumptions

Algorithm

- Gradient descent
- Algorithm analysis

Thursday

Geometry and optimality