COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Spring 2026.

Lecture 21

Logistics

- Quiz this week

- Problem Set 4 is posted and due next Friday 5/8 at 11:59pm. .
- After today you should be able to solve all the problems.
- Final exam Tuesday 5/12.

Summary

Last Class:

- Multivariable calculus review (directional derivative ← gradient $\langle \nabla f(x), v \rangle = D_v f(x)$)
- Introduction to gradient descent. Motivation as a greedy algorithm.
- Convex functions
- Lipschitz functions

This Class:

- Lipschitz functions
- Analysis of gradient descent for convex Lipschitz functions
- Extension to projected gradient descent for constrained optimization.

Gradient Descent Psuedocode

$\theta^* = \text{argmin } f(\theta)$

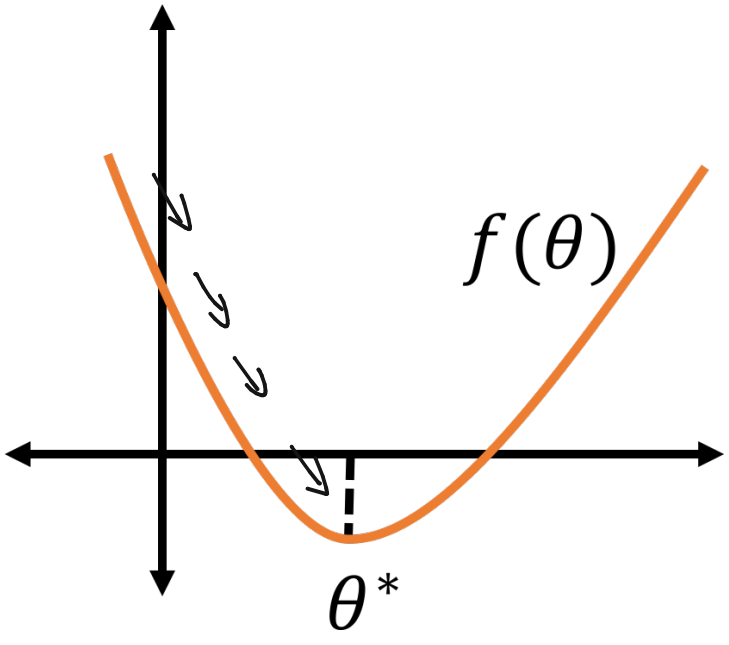
Gradient Descent

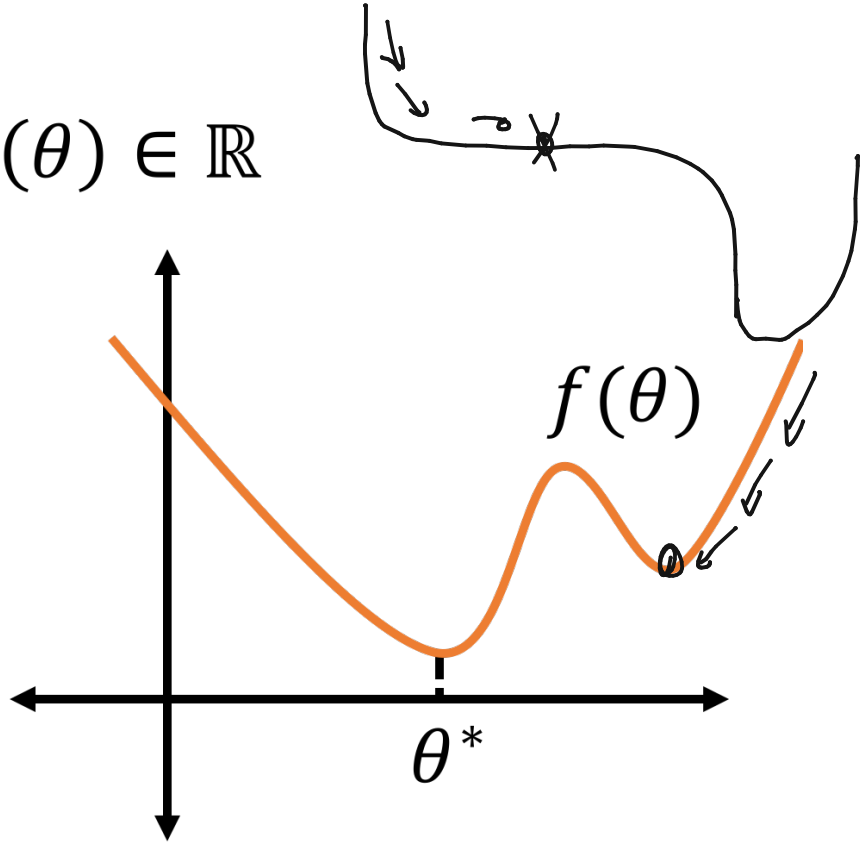
- Choose some initialization $\vec{\theta}^{(0)}$.
- For $i = 1, \ldots, t$
  - $\vec{\theta}^{(i)} = \vec{\theta}^{(i-1)} - \eta \nabla f(\vec{\theta}^{(i-1)})$
- Return $\vec{\theta}^{(t)}$, as an approximate minimizer of $f(\vec{\theta})$.

Step size η is chosen ahead of time or adapted during the algorithm (details to come).

When Does Gradient Descent Work?

$\theta \in \mathbb{R} \quad \nabla f(\theta) \in \mathbb{R}$



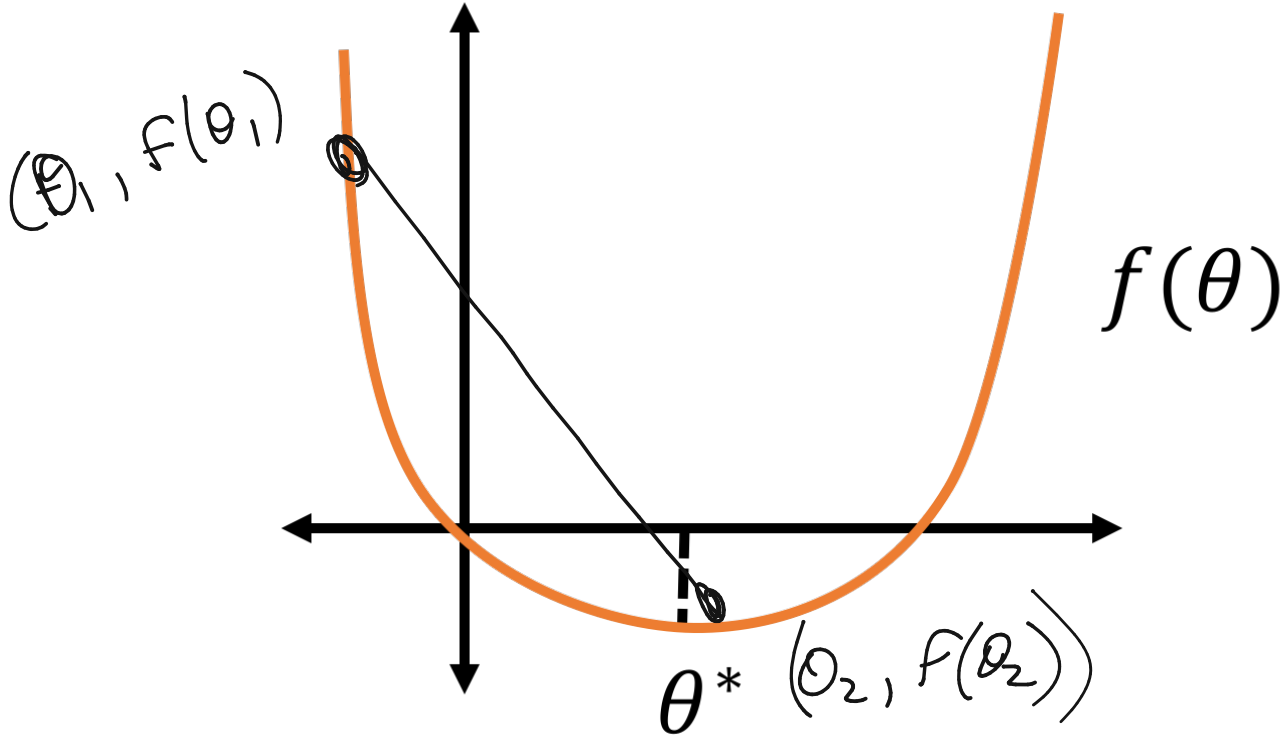


Gradient Descent Update: $\vec{\theta}_{i+1} = \vec{\theta}_i - \eta \nabla f(\vec{\theta}_i)$

Convexity

Definition – Convex Function: A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathbb{R}^d$ and $\lambda \in [0, 1]$:

$$(1 - \lambda) \cdot f(\vec{\theta}_1) + \lambda \cdot f(\vec{\theta}_2) \geq f\left((1 - \lambda) \cdot \vec{\theta}_1 + \lambda \cdot \vec{\theta}_2\right)$$



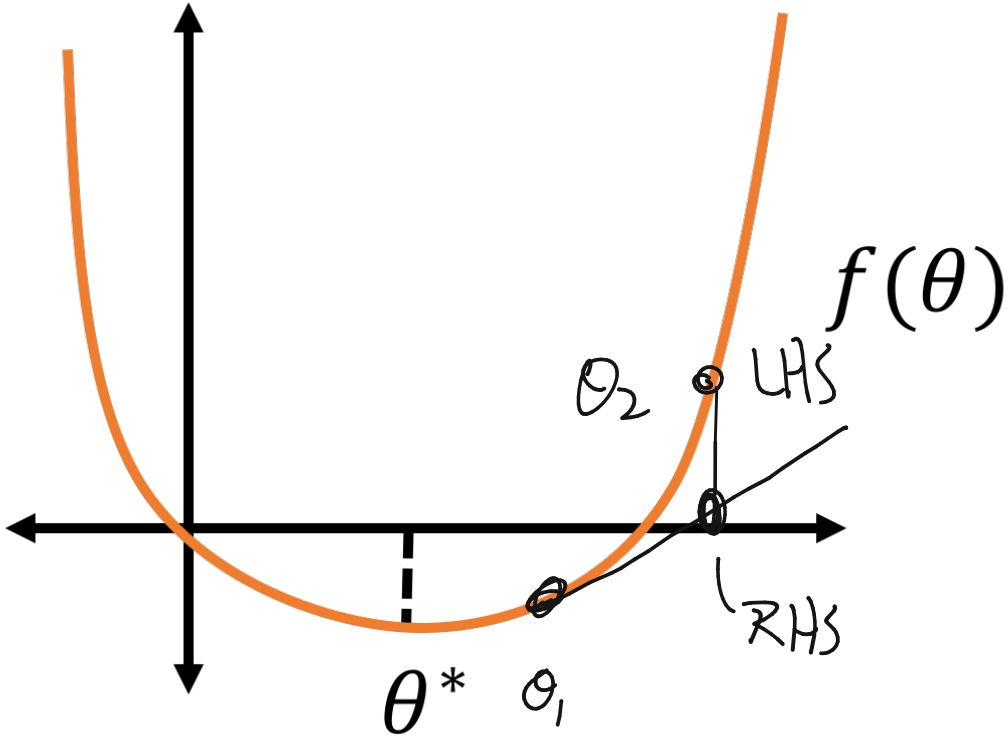
Convexity

Corollary – Convex Function: A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathbb{R}^d$ and $\lambda \in [0, 1]$:

$$f(\vec{\theta}_2) - f(\vec{\theta}_1) \geq \vec{\nabla} f(\vec{\theta}_1)^T \left(\vec{\theta}_2 - \vec{\theta}_1\right)$$

$f = h + g$

$\nabla f = \nabla h + \nabla g$



Practice with Definitions

$h(\theta) = -g(\theta)$ $\quad f(\theta) = 0$

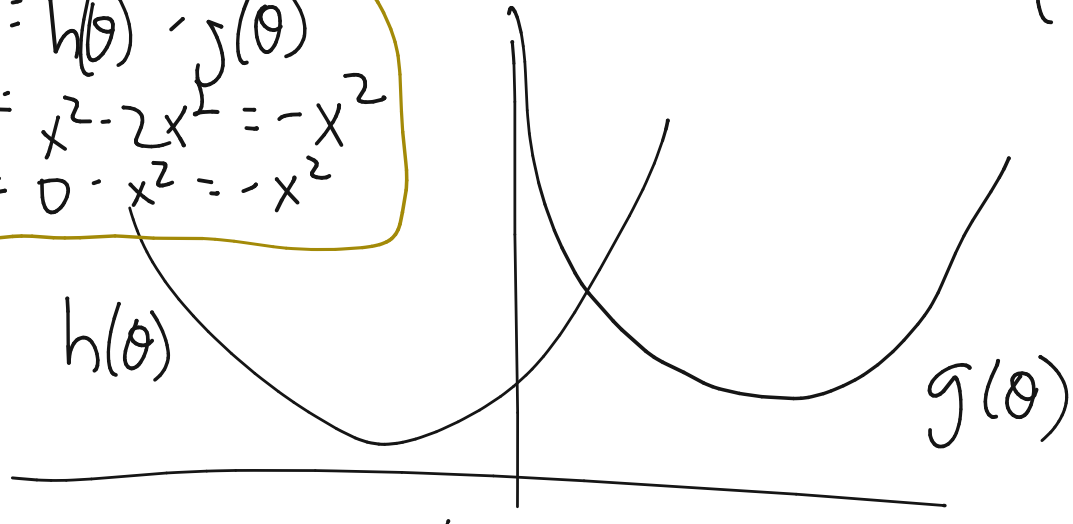
$h(\theta)$ is convex
$g(\theta) = -h(\theta)$ is convex

Let $f(\theta) = h(\theta) + g(\theta)$ where h and g are convex. Is f also convex?

yes always

$f(\theta) = h(\theta) \cdot g(\theta)$
$= x^2 \cdot -2x^2 = -x^2$
$= 0 \cdot x^2 = -x^2$

not always convex



$\lambda h(\theta_1) + (1-\lambda) h(\theta_2) \geq h(\lambda\theta_1 + (1-\lambda)\theta_2)$

$\lambda g(\theta_1) + (1-\lambda) g(\theta_2) \geq g(\lambda\theta_1 + (1-\lambda)\theta_2)$

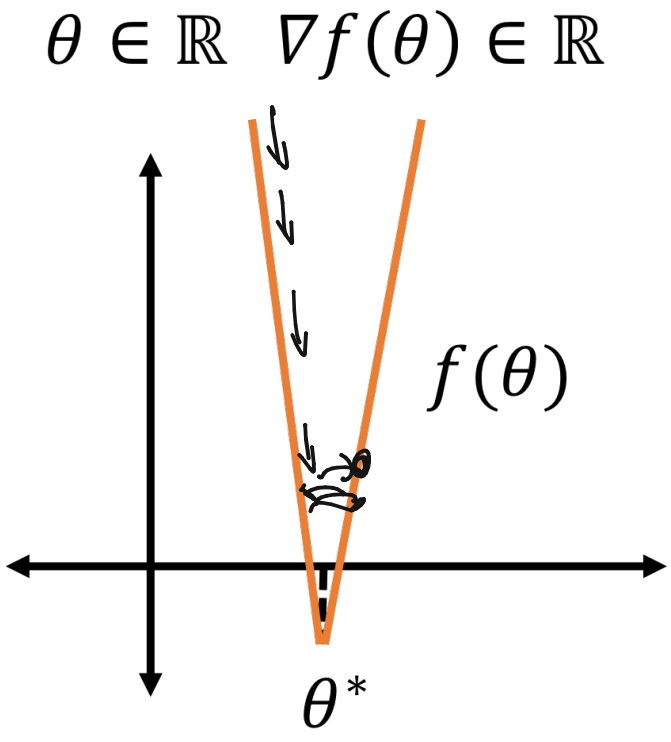
$\lambda(h(\theta_1) + g(\theta_1)) + (1-\lambda)(h(\theta_2) + g(\theta_2)) \geq h(\ldots) + g(\ldots)$

$\lambda f(\theta_1) + (1-\lambda) f(\theta_2) \geq f(\lambda\theta_1 + (1-\lambda)\theta_2)$

Practice with Definitions

Let $f(\theta) = h(\theta) + g(\theta)$ where h and g are convex. Is f also convex?

A second assumption: Lipschitzness

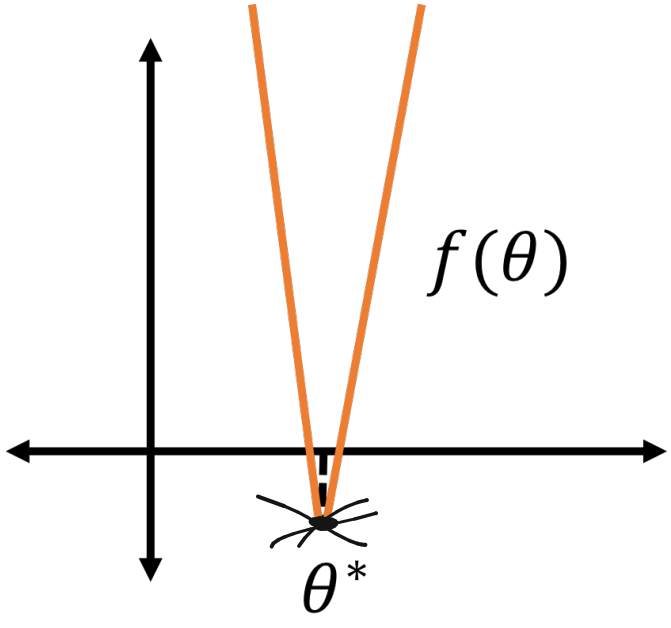


Gradient Descent Update:

$\vec{\theta}_{i+1} = \vec{\theta}_i - \eta \nabla f(\vec{\theta}_i)$

A second assumption: Lipschitzness

$\theta \in \mathbb{R} \quad \nabla f(\theta) \in \mathbb{R}$



$f(x) = |x| \qquad G = 1$

$f(x) = x^2 \qquad G = \infty$

$f'(x) = 2x$

Gradient Descent Update:

$\vec{\theta}_{i+1} = \vec{\theta}_i - \eta \nabla f(\vec{\theta}_i)$

Need to assume that the function is Lipschitz (size of gradient is bounded): There is some G s.t.:

$$\forall \vec{\theta}: \quad \|\vec{\nabla} f(\vec{\theta})\|_2 \leq G \Leftrightarrow \forall \vec{\theta}_1, \vec{\theta}_2: \quad |f(\vec{\theta}_1) - f(\vec{\theta}_2)| \leq G \cdot \|\vec{\theta}_1 - \vec{\theta}_2\|_2$$

Well-Behaved Functions

Definition – Convex Function: A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathbb{R}^d$ and $\lambda \in [0, 1]$:

$$(1 - \lambda) \cdot f(\vec{\theta}_1) + \lambda \cdot f(\vec{\theta}_2) \geq f\left((1 - \lambda) \cdot \vec{\theta}_1 + \lambda \cdot \vec{\theta}_2\right)$$

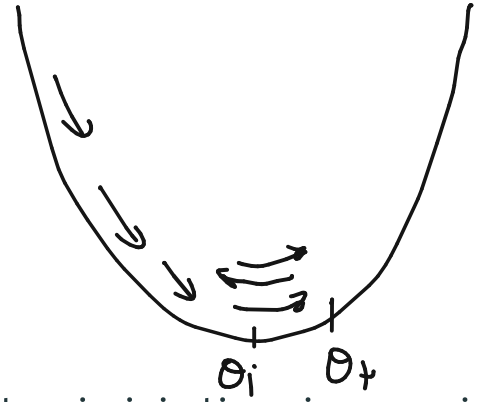
Corollary – Convex Function: A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathbb{R}^d$ and $\lambda \in [0, 1]$:

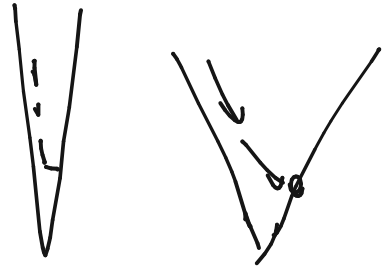
$$f(\vec{\theta}_2) - f(\vec{\theta}_1) \geq \vec{\nabla} f(\vec{\theta}_1)^T \left(\vec{\theta}_2 - \vec{\theta}_1\right)$$

Definition – Lipschitz Function: A function $f : \mathbb{R}^d \to \mathbb{R}$ is G-Lipschitz if $\|\vec{\nabla} f(\vec{\theta})\|_2 \leq G$ for all $\vec{\theta}$.

GD Analysis – Convex Functions

Assume that:





- f is convex.
- f is G-Lipschitz.
- $\|\vec{\theta}_1 - \vec{\theta}_*\|_2 \leq R$ where $\vec{\theta}_1$ is the initialization point.

$\theta_* = \arg\min f(\theta)$

Gradient Descent

- Choose some initialization $\vec{\theta}_1$ and set $\eta = \frac{R}{G\sqrt{t}}$.
- For $i = 1, \ldots, t-1$
  - $\vec{\theta}_{i+1} = \vec{\theta}_i - \eta \vec{\nabla} f(\vec{\theta}_i)$
- Return $\hat{\theta} = \arg\min_{\vec{\theta}_1, \ldots, \vec{\theta}_t} f(\vec{\theta}_i)$.

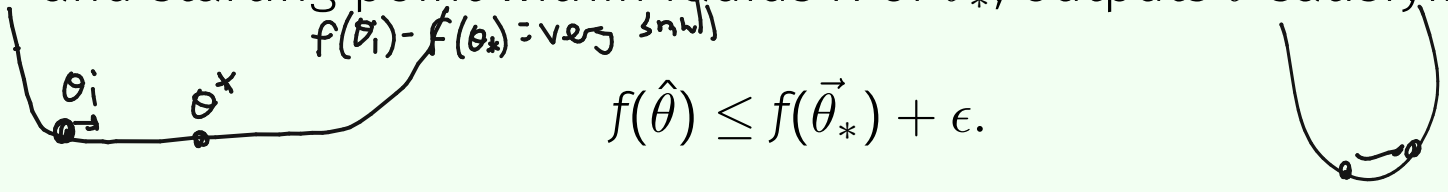
GD Analysis Proof

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

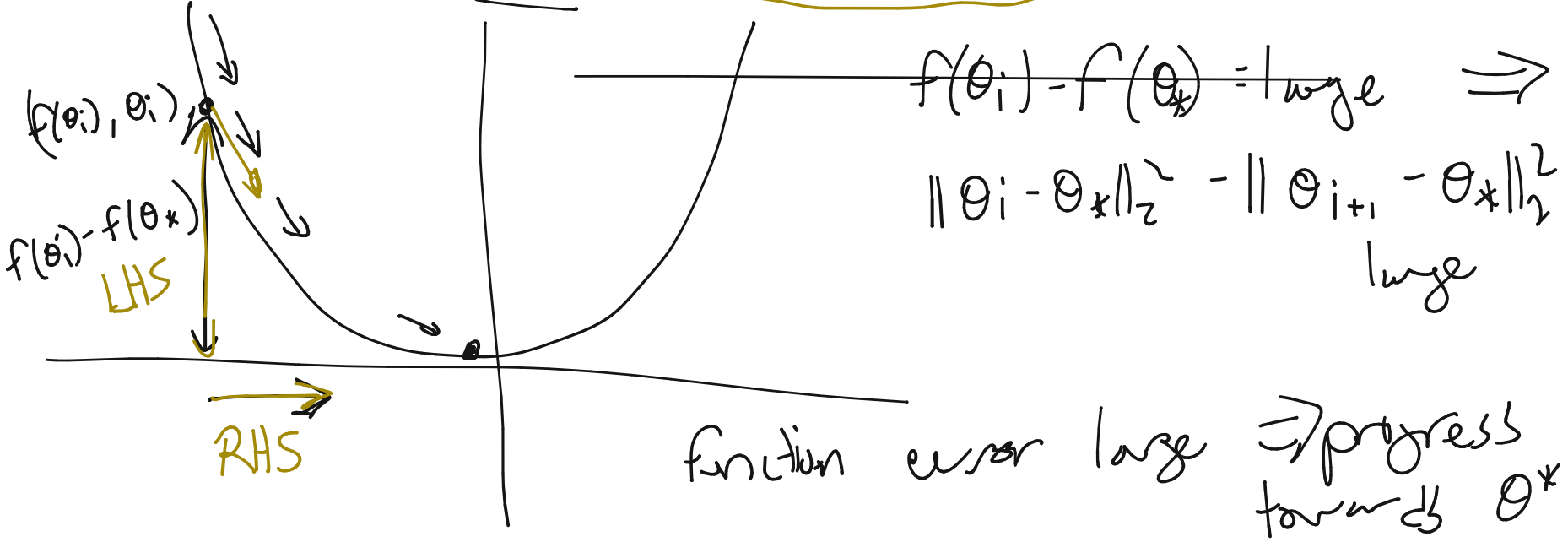
GD Analysis Proof

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:



$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

Step 1: For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$. Visually:



GD Analysis Proof

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

$\|a+b\|_2^2 = \|a\|_2^2 + \|b\|_2^2 + 2a^Tb$

Step 1: For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \theta_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$. Formally:

$$\|\theta_{i+1} - \theta_*\|_2^2 = \|\theta_i - \eta\nabla f(\theta_i) - \theta_*\|_2^2$$

$$= \|\underbrace{\theta_i - \theta_*}_{a} - \underbrace{\eta\nabla f(\theta_i)}_{b}\|_2^2 = \|\theta_i - \theta_*\|_2^2 + \|\eta\nabla f(\theta_i)\|_2^2 - 2\eta\nabla f(\theta_i)^T(\theta_i - \theta_*)$$

$\eta^2 G^2$

$$2\eta\nabla f(\theta_i)^T(\theta_i - \theta_*) \leq \|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2 + \eta^2 G^2$$

$$\nabla f(\theta_i)^T(\theta_i - \theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

GD Analysis Proof

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

Step 1: For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$.

Step 1.1: $\vec{\nabla} f(\vec{\theta}_i)^T (\vec{\theta}_i - \vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$

$\forall \theta_1, \theta_2 \quad f(\theta_2) - f(\theta_1) \geq \nabla f(\theta_1)^T (\theta_2 - \theta_1)$

$\theta_1 = \theta_i \quad \theta_2 = \theta_*$

$f(\theta_*) - f(\theta_i) \geq \nabla f(\theta_i)^T (\theta_* - \theta_i)$

$f(\theta_i) - f(\theta_*) \leq \nabla f(\theta_i)^T (\theta_i - \theta_*)$

GD Analysis Proof

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

Step 1: For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$.

Step 1.1: $\vec{\nabla} f(\vec{\theta}_i)^T (\vec{\theta}_i - \vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$ $\Longrightarrow$ Step 1 by convexity.

GD Analysis Proof

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

Step 1: For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$

$$\frac{1}{t}\sum_{i=1}^{t} f(\theta_i) - f(\theta_*) = \frac{1}{t}\sum_{i=1}^{t} \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

"telescoping sum"

$$\|\theta_1 - \theta_*\|_2^2 - \|\theta_2 - \theta_*\|_2^2 + \|\theta_2 - \theta_*\|_2^2 - \|\theta_3 - \theta_*\|_2^2 \ldots$$

$$\|\theta_1 - \theta_*\|_2^2 - \|\theta_{t+1} - \theta_*\|_2^2$$

"average error"

upper bounds

minimum error

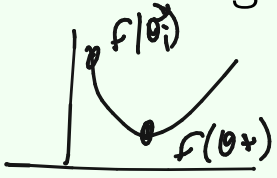
$$f(\bar{\theta}) - f(\theta^*) \leq \frac{1}{t} \cdot \frac{R^2}{2\eta} + \frac{\eta G^2}{2}$$

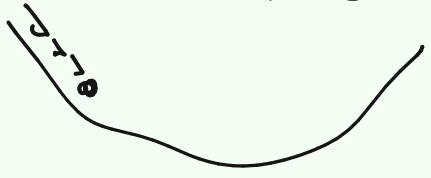
2η

GD Analysis Proof

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$





Step 1: For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \Longrightarrow$

undershooting error overshooting error

Step 2: $\frac{1}{t}\sum_{i=1}^{t} f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{R^2}{2\eta \cdot t} + \frac{\eta G^2}{2}$. $< \epsilon$

$t = \frac{R^2G^2}{\epsilon^2}$

$$\min_i f(\theta_i) - f(\theta^*) \leq \frac{R^2}{2\eta t} + \frac{\eta G^2}{2}$$

$$\leq \frac{R^2}{\frac{2Rt}{G\sqrt{t}}} + \frac{RG^2}{2G\sqrt{t}} = \frac{RG}{2\sqrt{t}} + \frac{RG}{2\sqrt{t}} \leq \frac{RG}{\sqrt{t}} \leq \epsilon$$

GD Analysis Proof

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

Step 2: $\frac{1}{t}\sum_{i=1}^{t} f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{R^2}{2\eta \cdot t} + \frac{\eta G^2}{2}$.

$\min_{\theta \in S} f(\theta)$