EECS 245, Winter 2026

LEC 18 Gradient Descent

→ Read Ch. 8.1 - 8.4

Agenda

- Recap: Gradients and "The Big 3" rules
- Minimizing $R_{sq}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2$ using gradients
- Using gradients to minimize functions that can't be minimized by hand

Announcements

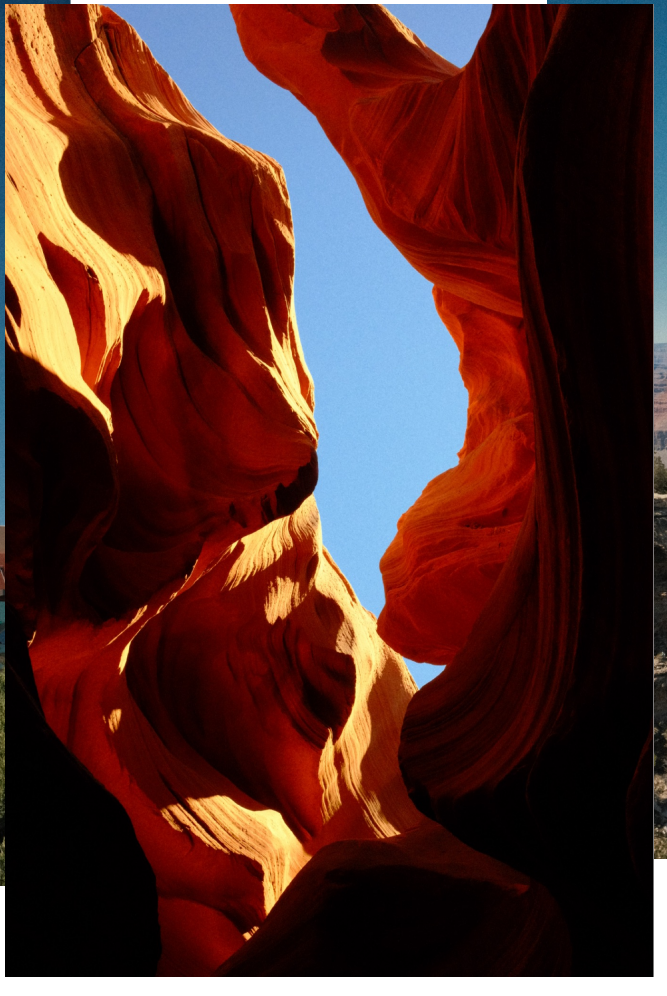
- HW 9A due on Friday
- HW 9B out tomorrow; due after Midterm 2
- MT 2 on Monday: details on next slide

Midterm 2 on Monday, March 30th, 7-9PM

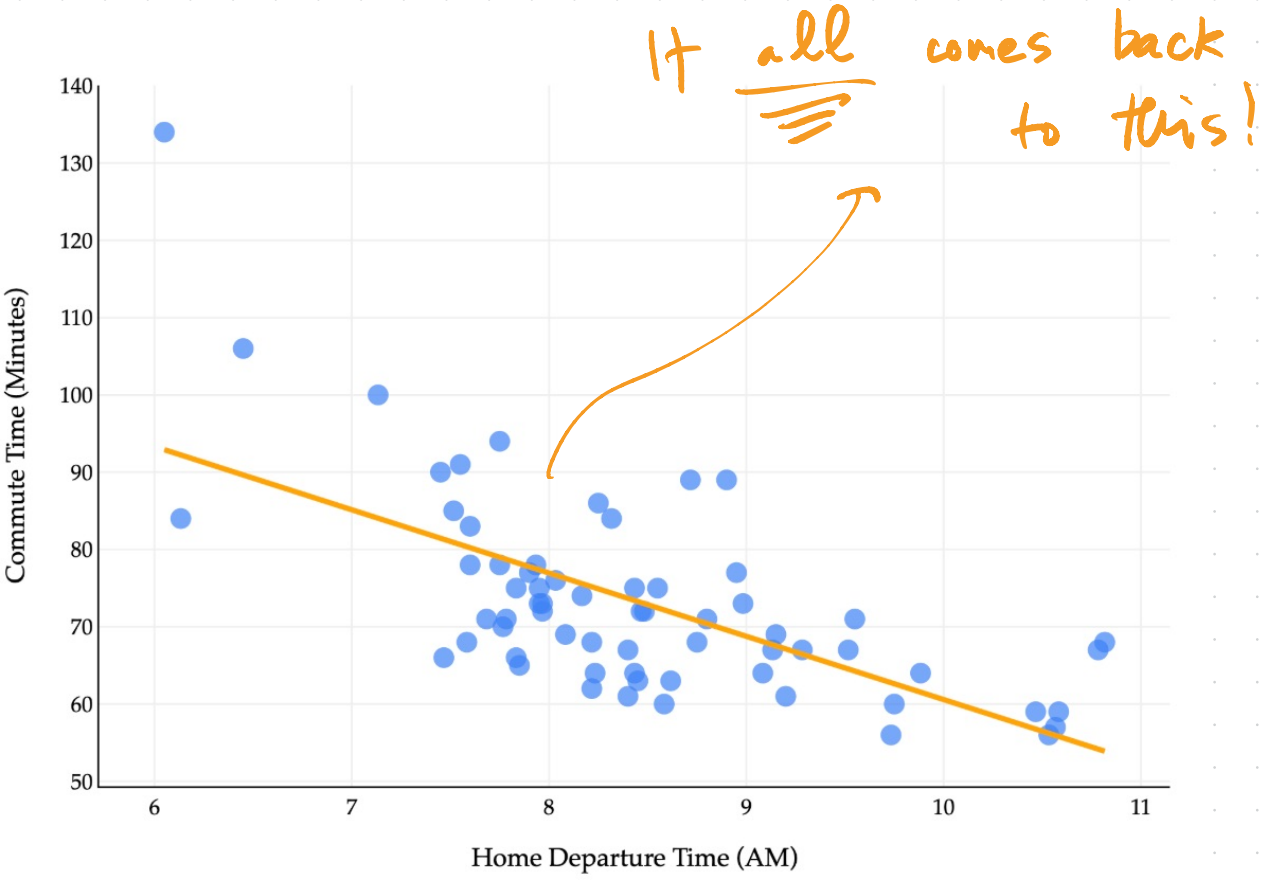
- Room: 1670 BBB (unless you have accomodations)
- Content: Lectures 11-19, Chapters 5-8, HW 6, 7, 8, 9A. Labs 7-10
- Similar format to MT1
- Can bring two double-sided 8.5 x 11" handwritten notes sheets
- Mock MT2: 2:30-5:30 PM, 1014 DOW

convexity will be on the final, but not MT2

Mock MT2 for 2 hours, office hours for last




me


It all comes back to this!
Commute Time (Minutes)
140
130
120
110
100
90
80
70
60
50
6
7
8
9
10
11
Home Departure Time (AM)

Recap: gradient

$f : \mathbb{R}^d \to \mathbb{R}$ "vector-to-scalar"

$$\nabla f(\vec{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}$$

gradient is a vector-to-vector function!

→ gradient points in direction of steepest ascent!

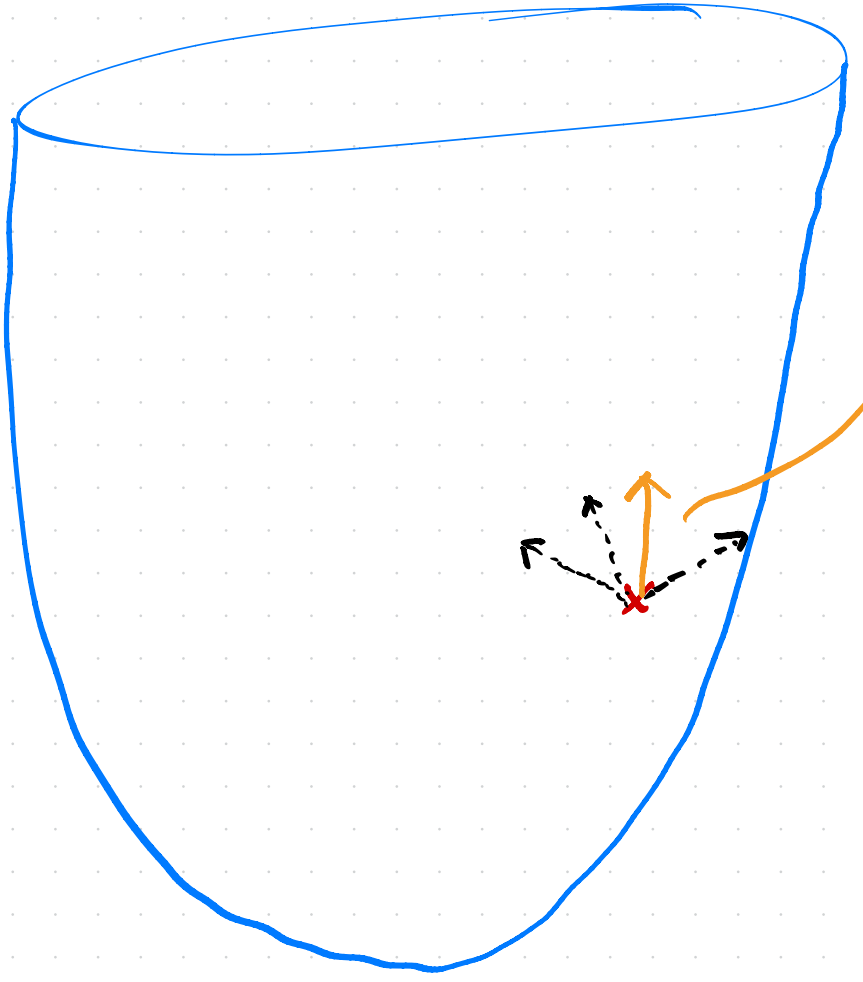
⇒ extension of derivative

$$f(\vec{x}) = \|\vec{x}\|^2 = x_1^2 + x_2^2 + \cdots + x_d^2$$

$$\frac{\partial f}{\partial x_i} = 2x_i$$

$$\nabla f(\vec{x}) = \begin{bmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_d \end{bmatrix} = 2\vec{x}$$

$\mathbb{R}^2 \to \mathbb{R}$



gradient: direction of steepest ascent

The Big 3 rules

- $f(\vec{x}) = \vec{a} \cdot \vec{x} = \vec{a}^T \vec{x}$

  $\vec{a} \in \mathbb{R}^d$ "constant"

  $\nabla f(\vec{x}) = \vec{a}$

- $f(\vec{x}) = \|\vec{x}\|^2$

  $\nabla f(\vec{x}) = 2\vec{x}$

- $f(\vec{x}) = \vec{x}^T A \vec{x}$ "quadratic form"

  $\nabla f(\vec{x}) = (A + A^T)\vec{x}$

Example: $f(\vec{x}) = \|\vec{x}\|^p = \left(\sqrt{x_1^2 + x_2^2 + \cdots + x_d^2}\right)^p$

Gradient?

1. Expanding def'n and using partials
2. Chain rule

$$f(\vec{x}) = \|\vec{x}\|_2^p = \left(\sqrt{x_1^2 + x_2^2 + \cdots + x_d^2}\right)^p$$

$$= \left(\sum_{i=1}^{d} x_i^2\right)^{p/2}$$

$$\frac{\partial f}{\partial x_i} = \frac{p}{2}\left(\sum_{i=1}^{d} x_i^2\right)^{\frac{p}{2}-1} \underbrace{\frac{\partial}{\partial x_i}\left(\sum_{i=1}^{d} x_i^2\right)}_{= 2x_i}$$

$$= \frac{p}{2}\left(\sum_{i=1}^{d} x_i^2\right)^{\frac{p}{2}-1} \cdot 2x_i = p\, x_i \left(\sum_{i=1}^{d} x_i^2\right)^{\frac{1}{2}(p-2)}$$

$$= p\, x_i \|\vec{x}\|^{p-2}$$

$$\frac{\partial f}{\partial x_i} = p x_i \|\vec{x}\|^{p-2}$$

$$f(\vec{x}) = \|\vec{x}\|^p$$

$$\nabla f(\vec{x}) = \begin{bmatrix} p x_1 \|\vec{x}\|^{p-2} \\ p x_2 \|\vec{x}\|^{p-2} \\ \vdots \\ p x_d \|\vec{x}\|^{p-2} \end{bmatrix} = p \|\vec{x}\|^{p-2} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

$$= p \|\vec{x}\|^{p-2} \vec{x}$$

e.g. $p = 1$, $\nabla f(\vec{x}) = 1 \|\vec{x}\|^{-1} \vec{x} = \frac{\vec{x}}{\|\vec{x}\|}$

② Chain rule

pretend:

$$\frac{d}{dx} h(g(x)) = h'(g(x))\, g'(x)$$

$$f(\vec{x}) = \|\vec{x}\|^p$$

$$g(\vec{x}) = \|\vec{x}\| \quad : \mathbb{R}^d \to \mathbb{R}$$

$$h(y) = y^p \quad : \mathbb{R} \to \mathbb{R}$$

$$f(\vec{x}) = h(g(\vec{x}))$$

$$\nabla f(\vec{x}) = \underbrace{h'(g(\vec{x}))}_{\text{scalar}} \begin{bmatrix} \frac{\partial g}{\partial x_1} \\ \frac{\partial g}{\partial x_2} \\ \vdots \end{bmatrix}_{\nabla g(\vec{x})} = \left(\frac{dh}{dx}(g(\vec{x}))\right) \nabla g(\vec{x})$$

⇒ back to MSE

$$R_{sq}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2$$

parameter vector $\mathbb{R}^{d+1}$

design matrix $n \times (d+1)$

observation vector $\mathbb{R}^n$

$R_{sq} : \mathbb{R}^{d+1} \to \mathbb{R}$

d features: d+1 parameters

⇒ gradient?

⇒ can't just use the chain rule directly

⇒ $\frac{2}{n}(\vec{y} - X\vec{w})$ is a vector in $\mathbb{R}^n$, not $\mathbb{R}^{d+1}$

⇒ need to do something else

$$R_{sq}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2 = \frac{1}{n}(\vec{y} - X\vec{w})^T(\vec{y} - X\vec{w})$$

$$= \frac{1}{n}\left(\vec{y}^T\vec{y} - \vec{y}^T X\vec{w} - (X\vec{w})^T\vec{y} + (X\vec{w})^T(X\vec{w})\right)$$

$$= \frac{1}{n}\left(\vec{y}^T\vec{y} - 2\vec{w}^T(X^T\vec{y}) + \vec{w}^T X^T X\vec{w}\right)$$

(the two underlined terms are the same!)

$\nabla(\vec{x}^T A\vec{x}) = (A + A^T)\vec{x}$

$$\nabla R_{sq}(\vec{w}) = \frac{1}{n}\left(0 - 2X^T\vec{y} + 2X^T X\vec{w}\right)$$

$X^T X + (X^T X)^T = 2X^T X$

$$\Rightarrow \quad \nabla R_{sq}(\vec{w}) = \frac{-2}{n}\left(X^T\vec{y} - X^TX\vec{w}\right)$$

→ here, we can set $\nabla R_{sq}(\vec{w}) = \vec{0}$ and solve!

$$-\frac{2}{n}\left(X^T\vec{y} - X^TX\vec{w}\right) = \vec{0}$$

$$X^T\vec{y} - X^TX\vec{w} = \vec{0}$$

$$X^TX\vec{w} = X^T\vec{y}$$

the same normal equation from before!!!!

if X's cols lin ind: $\vec{w}^* = (X^TX)^{-1}X^T\vec{y}$

what if we can calculate

$$\nabla f(\vec{x})$$

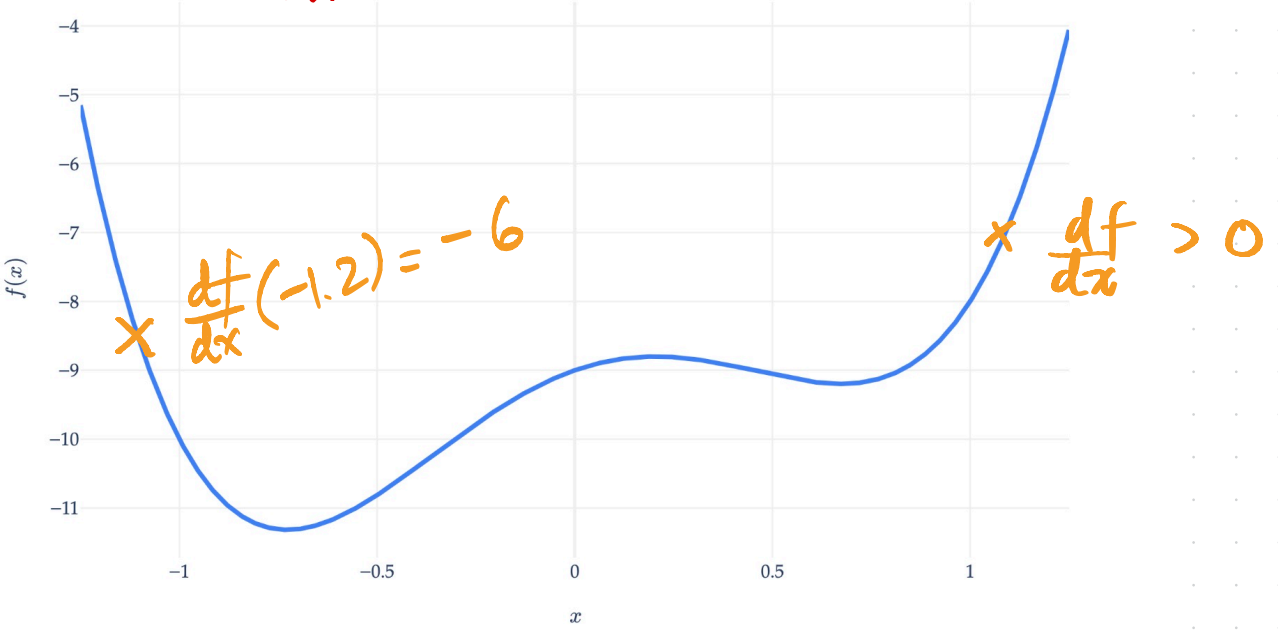
$$\nabla R(\vec{w})$$

but can't, on paper, solve for where

$$\nabla R(\vec{w}) = \vec{0} \; ?$$

$$f(x) = 5x^4 - x^3 - 5x^2 + 2x - 9$$

$$\frac{df}{dx}(x) = 20x^3 - 3x^2 - 10x + 2$$



See 8.3 for visuals!

Definition: Gradient Descent

Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a **differentiable** vector-to-scalar function, meaning that all of its partial derivatives are defined everywhere.

To find $\vec{x}^*$, the minimizer of f:

1. Choose a positive number, α. This number is called the **learning rate**, or **step size**.
2. Choose an **initial guess** for the minimizer, $\vec{x}^{(0)}$.
3. Then, repeatedly update the guess using the **update rule**:

$$\vec{x}^{(t+1)} = \vec{x}^{(t)} - \alpha \nabla f(\vec{x}^{(t)})$$

4. Terminate once the algorithm converges, which happens when the norm of the gradient, $||\nabla f(\vec{x}^{(t)})||$, is below some small **tolerance** level, e.g. 0.001 (since this must mean we're very close to a minimum).