EECS 245 Fall 2025

Math for ML

Lecture 17: The Gradient Vector

→ Read : Ch. 4.1

Agenda

1. Brief recap: multiple linear regression } Ch. 3.2 (new examples)
2. The gradient vector: a new approach to minimizing $R_{sq}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2$

   → Review: derivatives / partial derivatives

   → Gradients; important gradient rules

   → The "alternative derivation" of the normal equations

} Ch. 4.1

Announcements: HW 8 due Friday, HW 6 grades out, HW 7 sol'ns out (read!), class suggestions on Ed

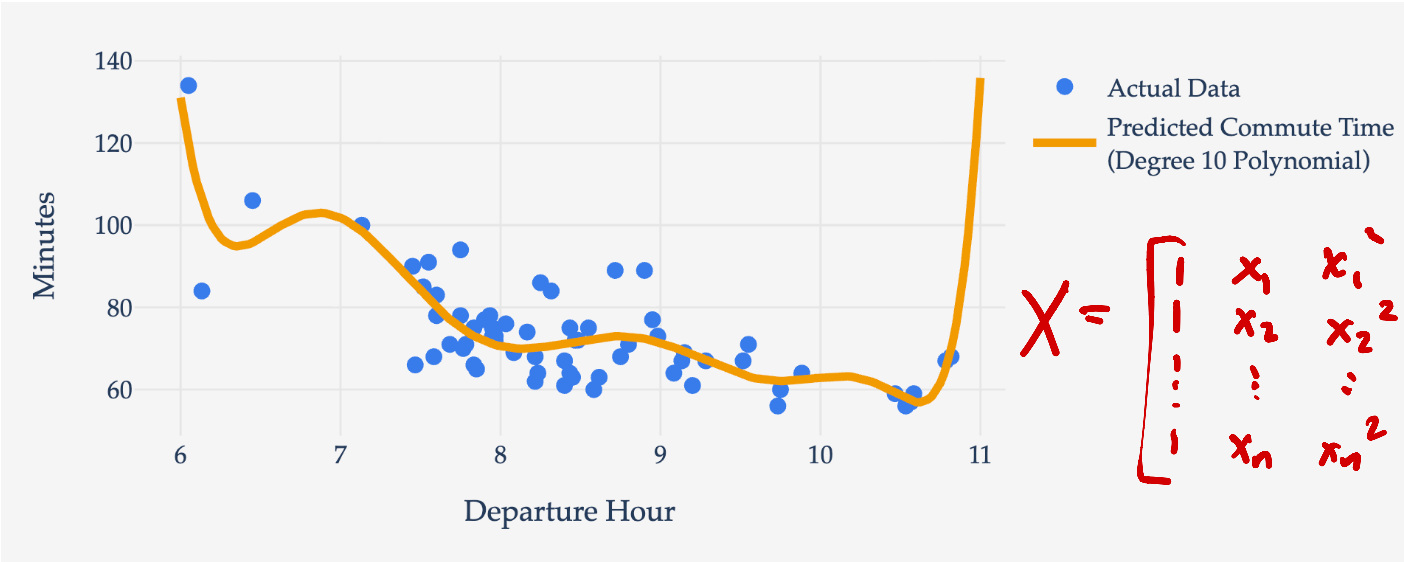
Midterm 2 is in 2 weeks....

LOCK IN!

- read old HW/lab solutions
- do all activities/examples in notes
- come to office hours
- actually come to lecture and read the notes

Big idea in Chapter 3.2 / Homework 8:

overfitting



$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{10} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{10} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{10} \end{bmatrix}$$

Why is adding more features not always a good thing?

overfitting: won't generalize to unseen data

$$\text{pred commute}_i = w_0 + w_1 \, \text{dept hour}_i + w_2 \, \text{day of month}_i$$

$$+ w_3 \log(\text{dept hour}_i)$$

$$+ w_4 \cos^{-1}\left(\frac{\text{day of month}_i^2}{\text{dept hour}_i}\right)$$

$$= \vec{w} \cdot \text{Aug}(\vec{x}_i)$$

"linear in the parameters"

→ can use normal eq'ns to find $\vec{w}^*$

could put these features in a design matrix

$$\text{pred commute}_i = w_0 + \sin(w_1 \text{ dept hour}_i)$$

w_1 in the sin

$$\neq \vec{w} \cdot \text{Aug}(\vec{x}_i)$$

not linear in the parameters

$\Rightarrow$ can't use normal eq'ns to find $\vec{w}^*$

Ch 3.2, Act. 2

Suppose we use the code below to build a multiple linear regression model to predict the width of a fish, given its height and weight.

```
model = LinearRegression()
model.fit(X, y)

# Used in the answer choices below.
ws = np.append(model.intercept_, model.coef_)
preds = model.predict(X)
squares = X.shape[0] * mean_squared_error(y, preds)
```

n

$\vec{p} = X\vec{w}^*$

$\vec{w}^*: (X^TX)\vec{w}^* = X^T\vec{y}$

$X^T\vec{e} = \vec{0}$

$\Rightarrow \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \cdot \vec{e} = 0$

review!

$\Rightarrow \sum e_i = 0$

	preds	ws	squares	np.sum(y - preds)
0				✓
$\Vert\vec{y} - X\vec{w}^*\Vert^2$			✓	
$X^TX\vec{w}^* - X^T\vec{y}$				✓
$\vec{1}^T(\vec{y} - X\vec{w}^*)$				✓
$(X^TX)^{-1}X^T\vec{y}$		✓		
$X(X^TX)^{-1}X^T\vec{y}$	✓			

(normal eq'ns)

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

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

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

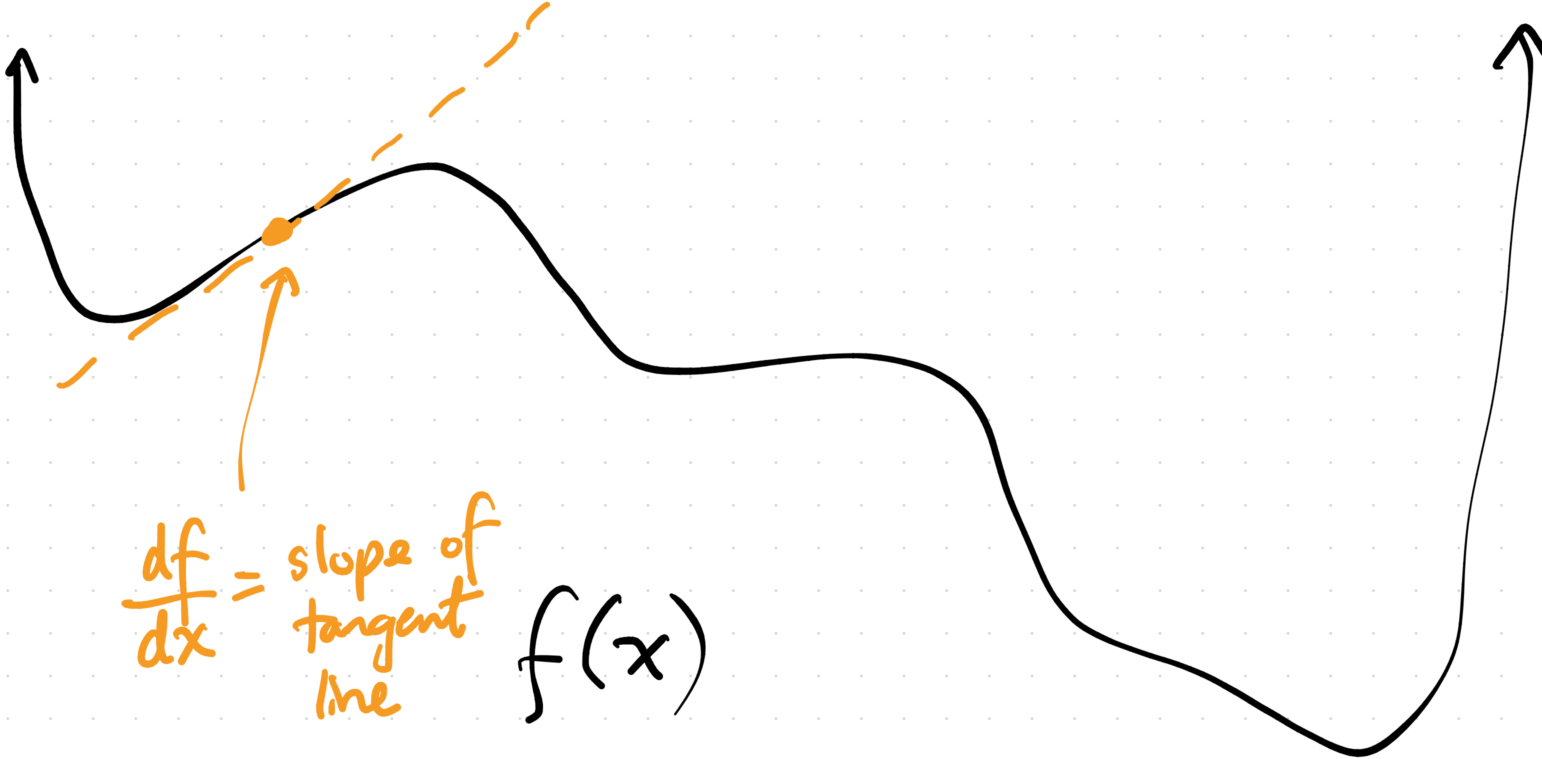
"vector-to-scalar" function $\quad f : \mathbb{R}^2 \to \mathbb{R}$

$$f(\vec{x}) = x_1^2 + 2x_1x_2 - \cos(x_2^3)$$

$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

before:

$f(x, y) = x^2 + 2xy - \cos(y^3)$



"scalar to scalar" functions

$$f: \mathbb{R} \to \mathbb{R}$$

Gradient

multivariate equivalent of derivative

\nabla

suppose $f : \mathbb{R}^d \to \mathbb{R}$

"vector-to-scalar"

the "gradient" of f is a vector in $\mathbb{R}^d$ containing all partial derivatives of f

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

technically $\nabla f(\vec{x})$ is a vector to vector function!

direction of "steepest ascent"

vector-to-scalar

ex.

$$f(\vec{x}) = x_1^2 + x_2^2 - 3x_1x_2$$

$$\nabla f(\vec{x}) = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} \\ \dfrac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 - 3x_2 \\ 2x_2 - 3x_1 \end{bmatrix}$$

at $\vec{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $\nabla f\left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 2(2) - 3(1) \\ 2(1) - 3(2) \end{bmatrix} = \begin{bmatrix} 1 \\ -4 \end{bmatrix}$

Example:

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

$$= a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

$\vec{a}$ fixed vector in $\mathbb{R}^n$,

$\vec{x} \in \mathbb{R}^n$

$$\nabla f(\vec{x}) = \begin{bmatrix} \partial f/\partial x_1 \\ \partial f/\partial x_2 \\ \vdots \\ \partial f/x_n \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \vec{a}$$

big rule 1

$$\frac{\partial f}{\partial x_1} = a_1$$

$$\frac{\partial f}{\partial x_2} = a_2$$

$$\vdots$$

Example:

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

$\vec{x} \in \mathbb{R}^n$

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

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

big rule 2

read 4.1

followup: what is $\nabla(\|\vec{x}\|)$? ans: $\dfrac{\vec{x}}{\|\vec{x}\|}$

Quadratic form : Given an $n \times n$ matrix A, a quadratic form is the function

$\vec{x} \in \mathbb{R}^n$

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

rule :

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

big rule 3

e.g. $A = \begin{bmatrix} 2 & 4 \\ 6 & 1 \end{bmatrix}$

vector to scalar

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

$$= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 6 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2x_1 + 4x_2 \\ 6x_1 + x_2 \end{bmatrix}$$

$$= x_1(2x_1 + 4x_2) + x_2(6x_1 + x_2)$$

$$= 2x_1^2 + (4+6)x_1x_2 + x_2^2 = 2x_1^2 + 10x_1x_2 + \cancel{1}x_2^2$$

usually, pick symmetric matrix for quadratic form,

i.e. $A = A^T$

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

if A symm: $= 2A\vec{x}$