EECS 245, Winter 2026

LEC 17

The Gradient Vector

→ Read: Ch. 7.2, 8.1, 8.2

Agenda

Ch. 8 is the last chapter in scope for MT 2

- Recap: multiple linear regression
- The gradient vector: a new way to minimize

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

→ Review: Partial derivatives

→ Gradients: "the big 3" rules"

→ Another way of coming up with the normal equation

Announcements

- HW 8 due Saturday → typo in 4c; redownload
  → TAs added remote OH on Saturday
- HW 7 scores + sol'ns up
- PrairieLearn solutions from last lab are up
- Suggested courses for next sem. on Ed

$h(\text{dept hour}_i, \text{day of month}_i)$

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

"Feature engineering" : creating new features

$h(x_i) = w_0 + \sin(w_1 x_i)$

not a linear model, since w is in sin

	departure_hour	day	day_of_month	minutes
0	10.816667	Mon	15	68.0
1	7.750000	Tue	16	94.0
2	8.450000	Mon	22	63.0
3	7.133333	Tue	23	100.0
4	9.150000	Tue	30	69.0

General case: n data points, d features

$$X_{n\times(d+1)} = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(d)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(d)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(d)} \end{bmatrix}$$

$\text{Aug}(\vec{x}_2)^T$

feature d, row 2

$$h(\vec{x}_i) = w_0 + w_1 \underline{x_i^{(1)}} + w_2 \underline{x_i^{(2)}} + \cdots + w_d \underline{x_i^{(d)}}$$

$$= \vec{w} \cdot \underbrace{\begin{bmatrix} 1 \\ x_i^{(1)} \\ x_i^{(2)} \\ \vdots \\ x_i^{(d)} \end{bmatrix}} = \vec{w} \cdot \text{Aug}(\vec{x}_i)$$

"augmented feature vector"

	departure_hour	day	day_of_month	minutes
0	10.816667	Mon	15	68.0
1	7.750000	Tue	16	94.0
2	8.450000	Mon	22	63.0
3	7.133333	Tue	23	100.0
4	9.150000	Tue	30	69.0

For row 3,
if we use 2 features:
1. dept hour
2. day of month,

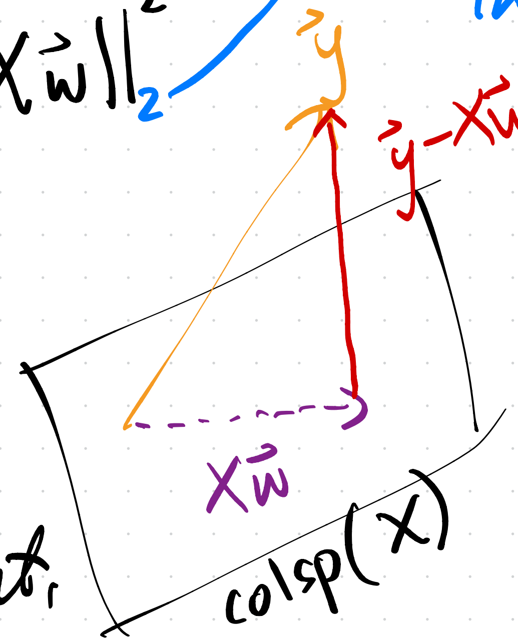
$$\text{Aug}(\vec{x}_3) = \begin{bmatrix} 1 \\ 8.45 \\ 22 \end{bmatrix}$$

How did we find $\vec{w}^*$, the optimal parameter vector?

⇒ Minimizing mean squared error

$$R_{sq}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|_2^2$$

implicitly always there



⇒ the shortest error vector satisfies

$$\underbrace{X^TX\vec{w} = X^T\vec{y}}_{\text{the normal eq'n}}$$

if X's cols are linearly independent, unique $\vec{w}^*$

Ch 7.2, Activity 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)
```

assume X's cols are linearly independent

ws: optimal parameter vec

preds: vector of preds

squares: sum of squared errors

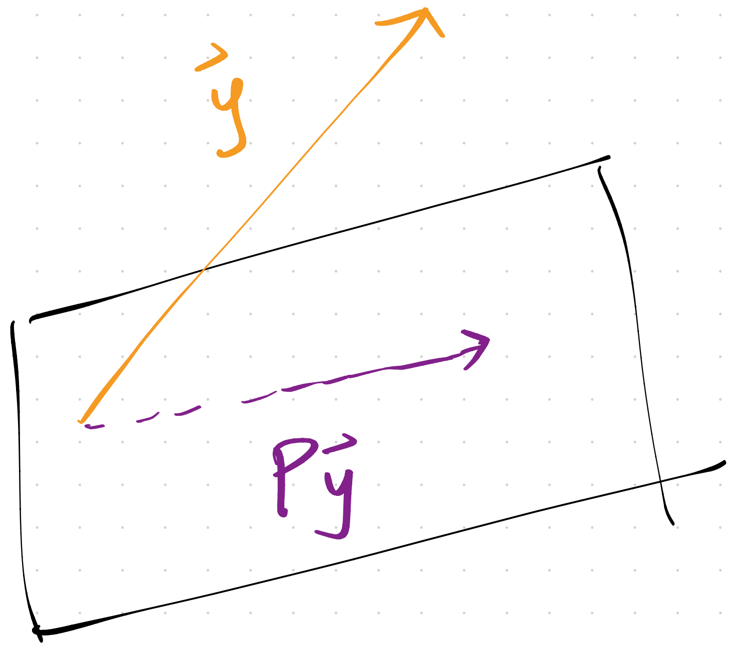
every column has at least 1 right answer!

	preds	ws	squares	np.sum(y - preds)
0				✓
$\lVert\vec{y} - X\vec{w}^*\rVert^2$			✓	
$X^TX\vec{w}^* - X^T\vec{y}$				✓ $\vec{0} \in \mathbb{R}^{d+1}$
$1^T(\vec{y} - X\vec{w}^*)$				✓ 0 scalar
$(X^TX)^{-1}X^T\vec{y}$		✓		
$X(X^TX)^{-1}X^T\vec{y}$	✓			

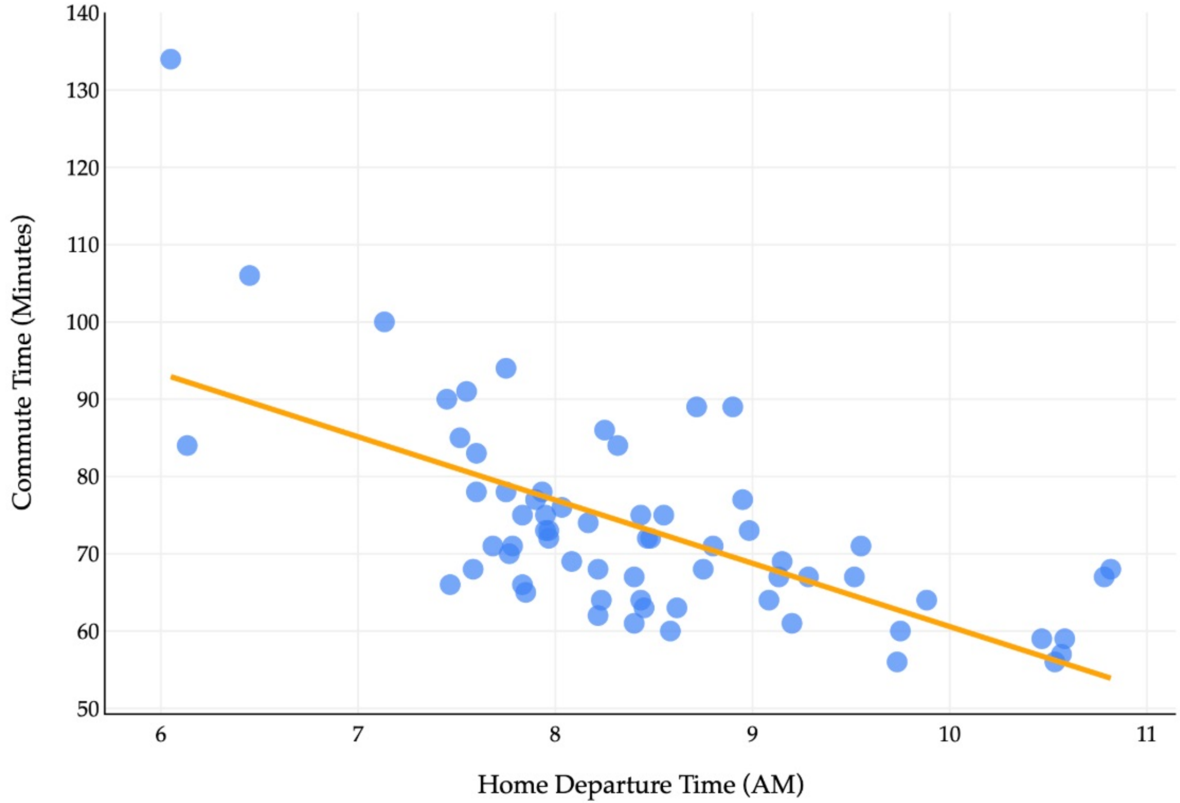
$\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}$

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

the sum of the error vector!



$\vec{y}$
$P\vec{y}$



Commute Time (Minutes)
140
130
120
110
100
90
80
70
60
50
6
7
8
9
10
11
Home Departure Time (AM)

so far, we've minimized

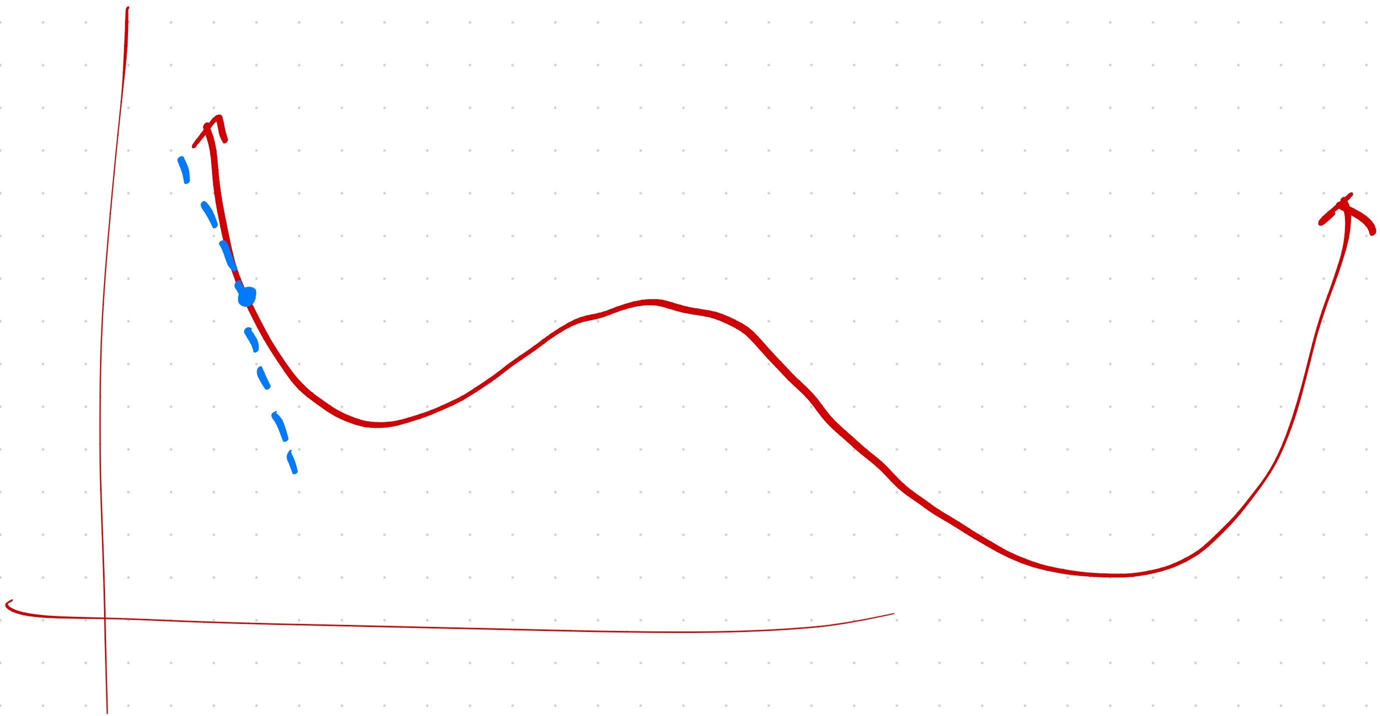
"vector-to-scalar"

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

using geometric arguments

$\Rightarrow$ what if we tried to use calculus?

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



Vector-to-scalar function example

$x^2 + y^2 - 3xy$

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

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

$$\frac{\partial f}{\partial x_1} = 2x_1 - 3x_2 \qquad \frac{\partial f}{\partial x_2} = 2x_2 - 3x_1$$

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

e.g. $\vec{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$

$$\nabla f\left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 1 \\ -4 \end{bmatrix}$$

Definition: Gradient Vector

Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a vector-to-scalar function. The **gradient vector** of f, denoted $\nabla f(\vec{x})$, is the vector in $\mathbb{R}^d$ of partial derivatives of f:

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

$\nabla f(\vec{x})$ itself is a vector-to-vector function; it takes in a vector $\vec{x} \in \mathbb{R}^d$ and outputs a new vector in $\mathbb{R}^d$, describing the rates of change of f along each dimension. The gradient, when evaluated at a point $\vec{x}_0$ describes the **direction of steepest ascent** of f at $\vec{x}_0$, i.e. the direction in which f is increasing most quickly.

8.2 : The "Big 3 Rules"

Consider the function

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

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

$$= a_1 x_1 + a_2 x_2 + \cdots + a_d x_d$$

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

$$\nabla f(\vec{x}) = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{bmatrix} = \vec{a}$$

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

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

$\Rightarrow$ Ponder : $\boxed{f(\vec{x}) = \|\vec{x}\| \Rightarrow \nabla f(\vec{x}) = \frac{\vec{x}}{\|\vec{x}\|}}$

the third "Big 3" rule

$$f(\vec{x}) = \underset{1\times n}{\vec{x}^T} \underset{n\times n}{A} \underset{n\times 1}{\vec{x}}$$

"quadratic form"

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

proof linked in the notes,
but just know it