

## Week 2

9/2/2025

- ↳ Quiz!
- ↳ How many read the notes? Important!
- ↳ Encourage practice, problem sets due every week
  - First 2 problems still 9/22
  - In general, due Mondays 11:59 pm
- ↳ Office hours change:  
Monday + Wednesday 12:30 to 2
- ↳ Post questions, come to office hours :
  - ↳ username on discord, so I can follow up!
  - ↳ How many on discord?

Last week was warm up for machine learning!

## Review

What about when  $M$  is not square?

$$\underline{A} \in \mathbb{R}^{d \times d}$$

$$\underline{A} = \sum_{i=1}^r \underline{v}_i \lambda_i \underline{w}_i^T$$

$\lambda_1, \dots, \lambda_r \in \mathbb{R}$  eigenvalues

$\underline{v}_1, \dots, \underline{v}_r \in \mathbb{R}^d$  right eigenvectors

$\underline{w}_1, \dots, \underline{w}_r \in \mathbb{R}^d$  left eigenvectors

Use: Page Rank where we powered up matrix multiplication!

$$\underline{A} \underline{v}_i = \lambda_i \underline{v}_i$$

$$\underline{w}_i^T \underline{A} = \lambda_i \underline{w}_i^T$$

$$\underline{v}_i^T \underline{v}_j = \mathbb{I}[i=j] = \underline{w}_i^T \underline{w}_j$$

$$\underline{w}_j^T \underline{A} \underline{v}_i = \underline{w}_j^T \lambda_i \underline{v}_i$$

$\Leftrightarrow$

$$\underline{w}_j^T \underline{v}_i = \mathbb{I}[i=j]$$

$$\lambda_j \underline{w}_j^T \underline{v}_i = \lambda_i \underline{w}_j^T \underline{v}_i$$

$$(\lambda_j - \lambda_i) \underline{w}_j^T \underline{v}_i = 0$$

## Supervised Learning

Problems like:

- predicting temperature
- identifying objects in an image
- generating next word

Labelled data

$$\underline{x} \in \mathbb{R}^d \quad y \in \mathbb{R}$$

n points...

$$(\underline{x}^{(1)}, y^{(1)}), \dots, (\underline{x}^{(n)}, y^{(n)})$$

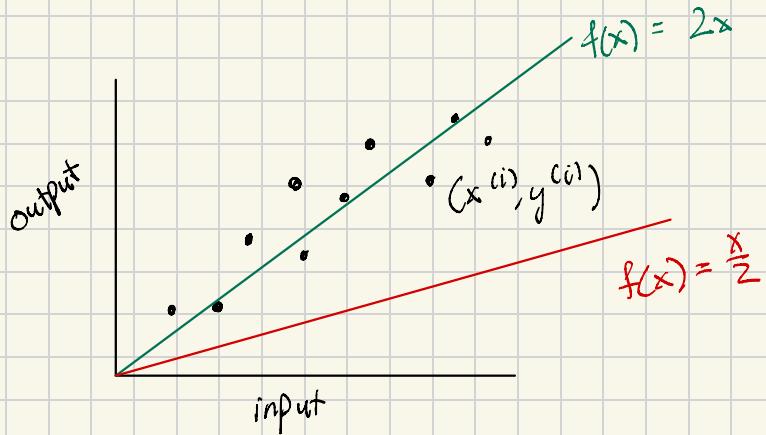
Goal: Find function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$   
so that  $f(\underline{x}^{(i)}) \approx y^{(i)} \quad \forall i$

Empirical risk minimization:

- ① Function class  $\mathcal{F}$  from which to select  $f$
- ② Loss to measure how well  $f$  fits data
- ③ Optimizer, method to select  $f$  with low loss

## Univariate Linear Regression

$$x^{(1)}, \dots, x^{(n)} \in \mathbb{R}$$



① Function class:

$$f(x) = \omega x \quad \text{for } \omega \in \mathbb{R}$$

## ② Loss

Attempt #1:  $f(x) - y$

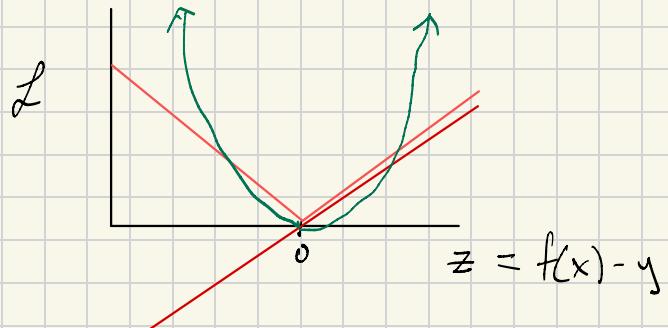
$\because y > f(x) \Rightarrow$  very negative

Attempt #2:  $|f(x) - y|$

$\because$  not differentiable at 0

Attempt #3:  $(f(x) - y)^2$

$\therefore$  far gets penalized more



Mean Squared Error (MSE) Loss

$$\mathcal{L}(\omega) = \frac{1}{n} \sum_{i=1}^n [f_\omega(x^{(i)}) - y^{(i)}]^2$$

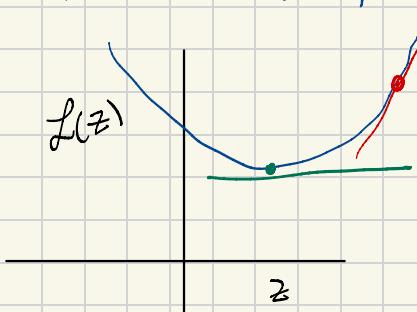
### ③ Exact Optimization

Key insight: Squared loss is convex

↳ differentiable

↳ single minima

Q: How do we find minima?



no improvement in any direction!

$$\begin{aligned}\frac{\partial}{\partial \omega} [\mathcal{L}(\omega)] &= \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \omega} [w x^{(i)} - y^{(i)}]^2 \\ &= \frac{1}{n} \sum_{i=1}^n 2[w x^{(i)} - y^{(i)}] \frac{\partial}{\partial \omega} (w x^{(i)} - y^{(i)}) \\ &\stackrel{\text{set}}{=} \frac{1}{n} \sum_{i=1}^n 2[w x^{(i)} - y^{(i)}] x^{(i)} \\ &= 0\end{aligned}$$

$$\begin{aligned}\stackrel{?}{=} \frac{1}{n} \sum_{i=1}^n w^* (x^{(i)})^2 &= \frac{2}{n} \sum_{i=1}^n y^{(i)} x^{(i)} \\ w^* &= \frac{\sum_{i=1}^n y^{(i)} x^{(i)}}{\sum_{i=1}^n [x^{(i)}]^2}\end{aligned}$$

In general,  $\underline{x}^{(i)} \in \mathbb{R}^d$

Again, but with some linear algebra

9/4/2025

Reminders

- ↳ Notes?
- ↳ Pset 2 due Monday  
(self grade following Friday)
- ↳ Quiz
  - ↳ grades released
  - ↳ let's chat ☺

Review

$$(\underline{x}^{(1)}, \underline{y}^{(1)}), \dots, (\underline{x}^{(n)}, \underline{y}^{(n)})$$

Goal :  $f$  so that  $f(\underline{x}^{(i)}) \approx \underline{y}^{(i)}$

(1) Function Class

(2) Loss

(3) Optimizer

Tuesday

dim?    d=1

(1) Linear

(2) MSE

(3) Exact

Thursday

d ≥ 1

Linear

MSE

Exact

# Multivariate Linear Regression

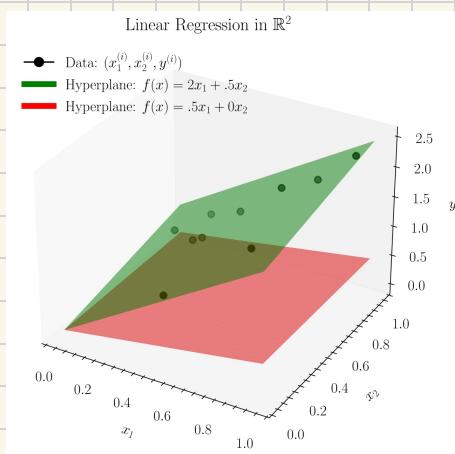
$(\underline{x}^{(1)}, y^{(1)}), \dots, (\underline{x}^{(n)}, y^{(n)})$      $\underline{x}^{(i)} \in \mathbb{R}^d$

## ① Function Class

Let weights  $\underline{w} \in \mathbb{R}^d$

$$f(\underline{x}) = \langle \underline{x}, \underline{w} \rangle = \underline{x}^\top \underline{w} = \sum_{k=1}^d w_k x_k$$

coefficient  
for each  
feature



$$f(\underline{x}) = [2 \quad .5] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$f(\underline{x}) = [.5 \quad 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

## ② MSE

$$\mathcal{L}(\underline{w}) = \frac{1}{n} \sum_{i=1}^n (\langle \underline{x}^{(i)}, \underline{w} \rangle - y^{(i)})^2$$

$$\underline{X} \in \mathbb{R}^{n \times d}$$

$$\begin{bmatrix} \underline{x}^{(1)\top} \\ \underline{x}^{(2)\top} \\ \vdots \\ \underline{x}^{(n)\top} \end{bmatrix}$$

$$\underline{y} \in \mathbb{R}^n$$

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

$$\frac{1}{n} \|\underline{X}\underline{w} - \underline{y}\|_2^2 = \mathcal{L}(\underline{w})$$

$$\frac{1}{n} \left\| \begin{bmatrix} \langle \underline{x}^{(1)}, \underline{w} \rangle \\ \langle \underline{x}^{(2)}, \underline{w} \rangle \\ \vdots \\ \langle \underline{x}^{(n)}, \underline{w} \rangle \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix} \right\|_2^2$$

(3)

## Optimizer

Key insight:

$\underline{w}^*$  when no improvement  
in any direction

$$\nabla_{\underline{w}} \mathcal{L}(\underline{w}) = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial w_1} \\ \frac{\partial \mathcal{L}}{\partial w_2} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial w_d} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\underline{e}_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i^{\text{th position}}$$

$$\frac{\partial \mathcal{L}(\underline{w})}{\partial w_i} = \lim_{\Delta \rightarrow 0} \frac{\mathcal{L}(\underline{w} + \Delta \underline{e}_i) - \mathcal{L}(\underline{w})}{\Delta}$$

$$= \lim_{\Delta \rightarrow 0} \frac{\|\underline{X}(\underline{w} + \Delta \underline{e}_i) - \underline{y}\|_2^2 - \|\underline{X}\underline{w} - \underline{y}\|_2^2}{\Delta} = (*)$$

$$\begin{aligned} \|\underline{a} + \underline{b}\|_2^2 &= \sum_{i=1}^d (a_i + b_i)^2 = (\underline{a} + \underline{b})^T (\underline{a} + \underline{b}) \\ &= \underline{a}^T \underline{a} + \underline{a}^T \underline{b} + \underline{b}^T \underline{a} + \underline{b}^T \underline{b} \\ &= \|\underline{a}\|_2^2 + 2 \langle \underline{a}, \underline{b} \rangle + \|\underline{b}\|_2^2 \end{aligned}$$

$$\underline{a} = \underline{X}\underline{w} - \underline{y} \quad \underline{b} = \Delta \underline{X}\underline{e}_i$$

$$(*) = \lim_{\Delta \rightarrow 0} \frac{\|\underline{X}\underline{w} - \underline{y}\|_2^2 + 2 \langle \underline{X}\underline{w} - \underline{y}, \Delta \underline{X}\underline{e}_i \rangle + \|\Delta \underline{X}\underline{e}_i\|_2^2 - \|\underline{X}\underline{w} - \underline{y}\|_2^2}{\Delta}$$

$$= \lim_{\Delta \rightarrow 0} \frac{2 \Delta \langle \underline{X}\underline{w} - \underline{y}, \underline{X}\underline{e}_i \rangle + \Delta^2 \|\underline{X}\underline{e}_i\|_2^2}{\Delta}$$

$$= 2 \langle \underline{X}\underline{w} - \underline{y}, \underline{X}\underline{e}_i \rangle$$

$$\underline{X} \underline{e}_i = \underline{x}_i$$

$$\begin{bmatrix} -\underline{x}^{(1)\top} \\ \vdots \\ -\underline{x}^{(n)\top} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\nabla_{\underline{w}} \mathcal{L}(\underline{w}) = \frac{1}{2} \begin{bmatrix} \underline{x}_1^\top (\underline{x}\underline{w} - \underline{y}) \\ \vdots \\ \underline{x}_d^\top (\underline{x}\underline{w} - \underline{y}) \end{bmatrix} = 2 \underline{X}^\top (\underline{x}\underline{w} - \underline{y})$$

$$\nabla_{\underline{w}} \mathcal{L}(\underline{w}^*) = 0 = 2 \underline{X}^\top (\underline{x}\underline{w}^* - \underline{y})$$

$$\underline{X}^\top \underline{X} \underline{w}^* = \underline{X}^\top \underline{y}$$

$$\underline{w}^* = (\underline{X}^\top \underline{X})^+ \underline{X}^\top \underline{y}$$

Singular Value Decomposition

$$\underline{X} = \sum_{i=1}^d \sigma_i \underline{u}_i \underline{v}_i^\top \quad \begin{matrix} d & \leftarrow \text{assume features indep} \\ n \times 1 & 1 \times d \end{matrix}$$

$$\underline{X}^+ = \sum_{i=1}^d \frac{1}{\sigma_i} \underline{v}_i \underline{u}_i^\top \quad \begin{matrix} d \times 1 & 1 \times n \end{matrix}$$

$$\underline{u}_i^\top \underline{u}_j = \mathbb{I}[i=j] = \underline{v}_i^\top \underline{v}_j$$

$$\underline{X}^\top \underline{X} = \sum_{i=1}^d \underline{v}_i \sigma_i^2 \underline{u}_i^\top$$

$$(\underline{X}^\top \underline{X})^+ \underline{X}^\top \underline{y} = \sum_{i=1}^d \underline{v}_i \frac{1}{\sigma_i^2} \underline{v}_i^\top \sum_{j=1}^d \sigma_j \underline{u}_j \underline{u}_j^\top \underline{y}$$

$$= \sum_{i=1}^d \underline{v}_i \frac{1}{\sigma_i^2} \underline{u}_i^\top \underline{y}$$

Q: Why squared loss?

Before:

↳ differentiable

↳ penalize far, more

Now:

Q2: Why use linear model?

$$y = \langle x, w^* \rangle + \eta$$

where  $\eta \sim N(0, \sigma^2)$

$\uparrow$  mean 0       $\nwarrow$  unknown variance

~ Central Limit Theorem ~

Equivalently,

$$y \sim N(\langle x, w^* \rangle, \sigma^2)$$

ERM says find model that most likely generated data...

$\Pr(y \text{ drawn from model } w)$

$$= \Pr(y \sim N(\langle x, w \rangle, \sigma^2))$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - \langle x, w \rangle)^2}{2\sigma^2}\right)$$

$$\arg \max_w \Pr(y^{(1)} \sim N(\langle x^{(1)}, w \rangle, \sigma^2), \dots, y^{(n)} \sim N(\langle x^{(n)}, w \rangle, \sigma^2))$$

$$= \arg \max_w \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \langle x_i, w \rangle)^2}{2\sigma^2}\right)$$

$$\arg \max_w \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \langle x, w \rangle)^2}{2\sigma^2}\right)$$

$$= \arg \max_w \log(\cdot)$$

$$= \arg \max_w \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{i=1}^n \log\left(\exp\left(-\frac{(y - \langle x, w \rangle)^2}{2\sigma^2}\right)\right)$$

$$= \arg \max_w \sum_{i=1}^n -(y - \langle x, w \rangle)^2$$

$$= \arg \min_w \sum_{i=1}^n (y - \langle x, w \rangle)^2$$

M S E :