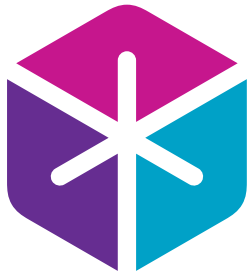
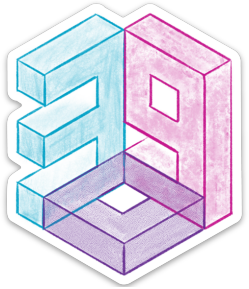




https://introml.mit.edu/

6.390 Intro to Machine Learning

Lecture 2: Regularization and Cross-validation

Shen Shen

Feb 9, 2026

3pm, Room 10-250

Slides and Lecture Recording

Recall

Let

$$X = \begin{bmatrix} — \; x^{(1)\top} \; — \\ \vdots \\ — \; x^{(n)\top} \; — \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & \cdots & x_d^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(n)} & \cdots & x_d^{(n)} \end{bmatrix} \qquad Y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}$$

$$\in \mathbb{R}^{n \times d} \qquad \in \mathbb{R}^{n \times 1} \qquad \in \mathbb{R}^{d \times 1}$$

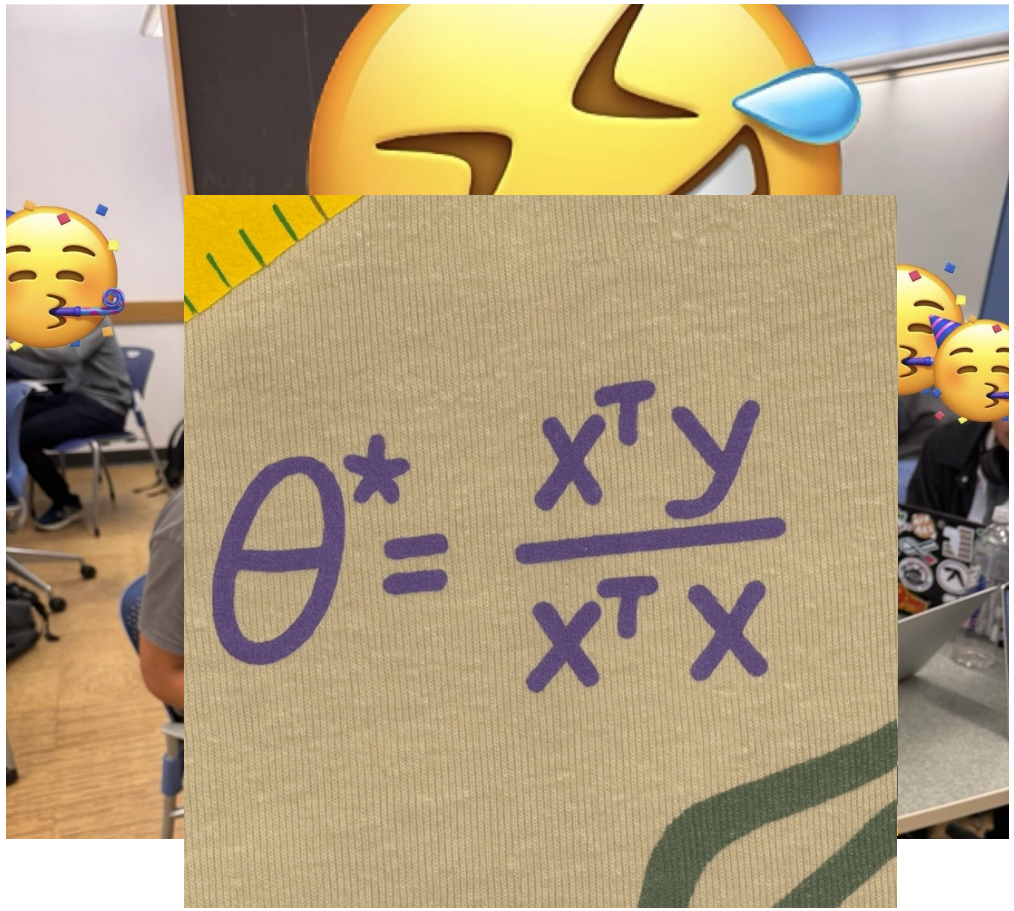
Then $J(\theta) = \frac{1}{n}(X\theta - Y)^\top (X\theta - Y) \quad \in \mathbb{R}^{1 \times 1}$

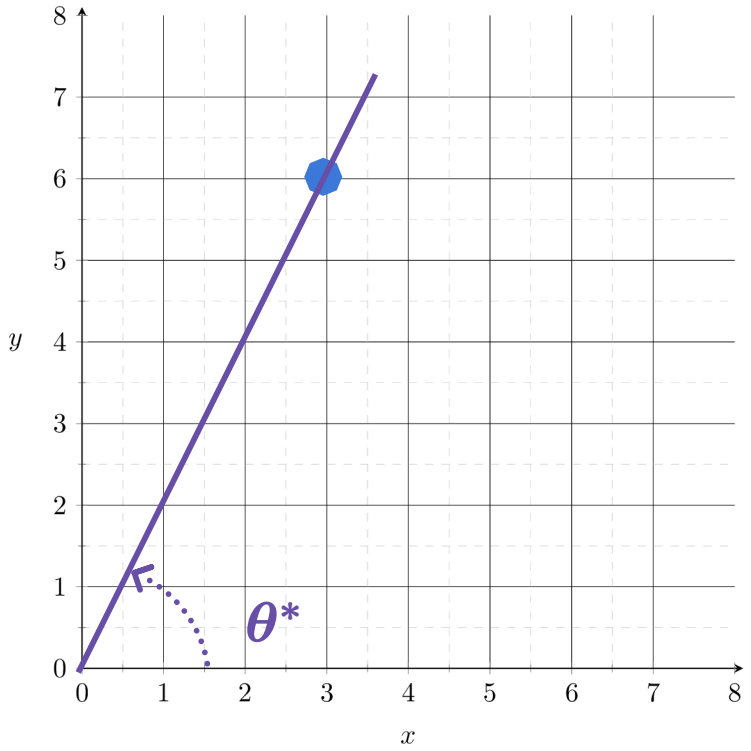
By matrix calculus and optimization

$$\theta^* = \left(X^\top X\right)^{-1} X^\top Y \qquad \in \mathbb{R}^{d \times 1}$$

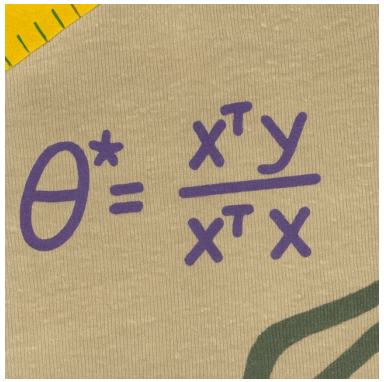
$$\theta^* = \left(X^\top X\right)^{-1} X^\top Y$$

Jane street shirt

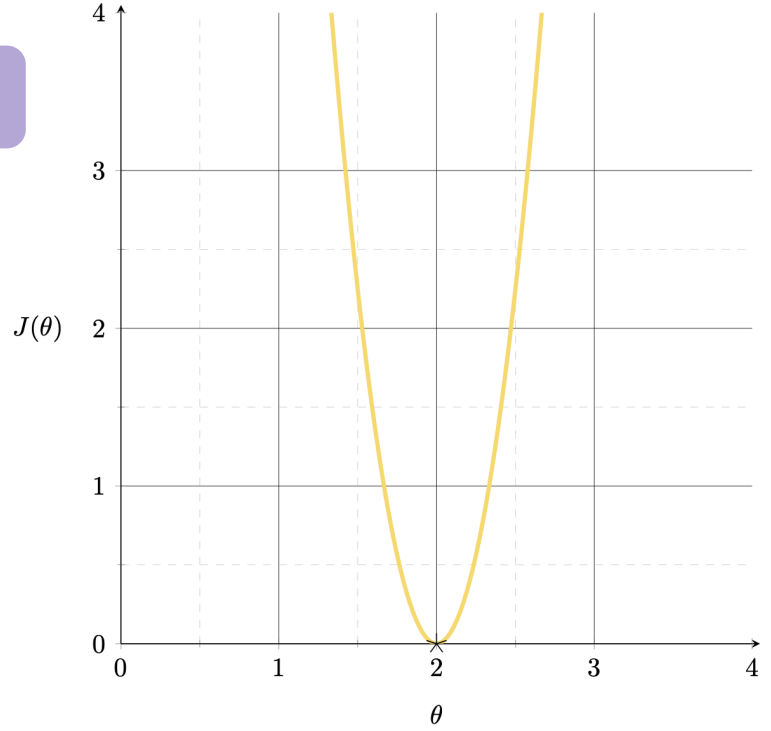




$$\theta^* = (X^\top X)^{-1} X^\top Y$$

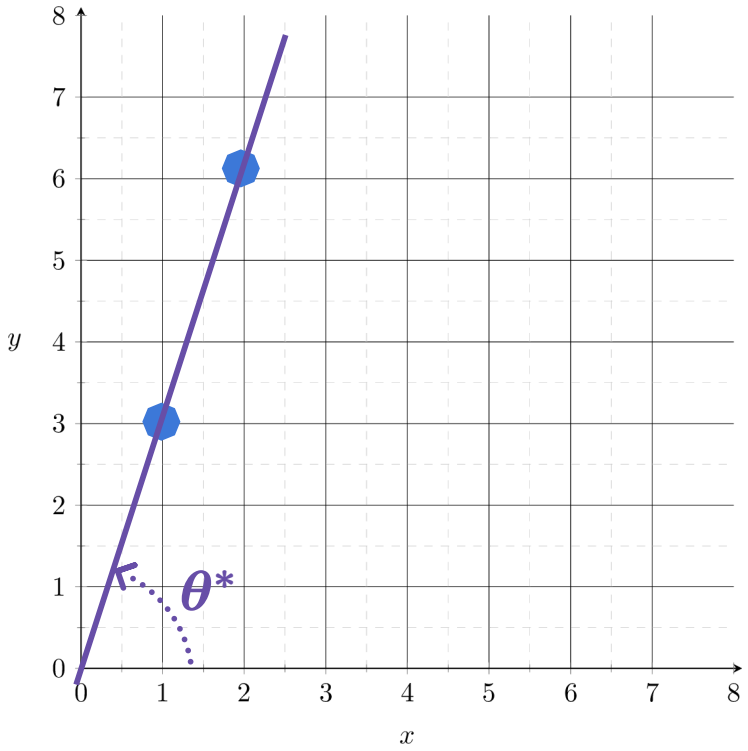


$$J(\theta) = (3\theta - 6)^2$$

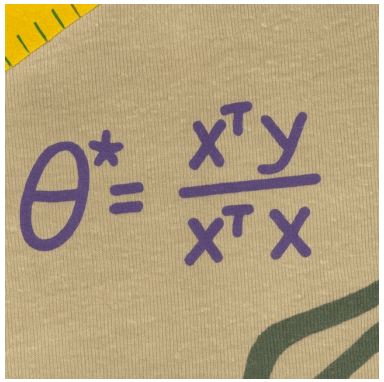


$$X = [3] \qquad Y = [6]$$

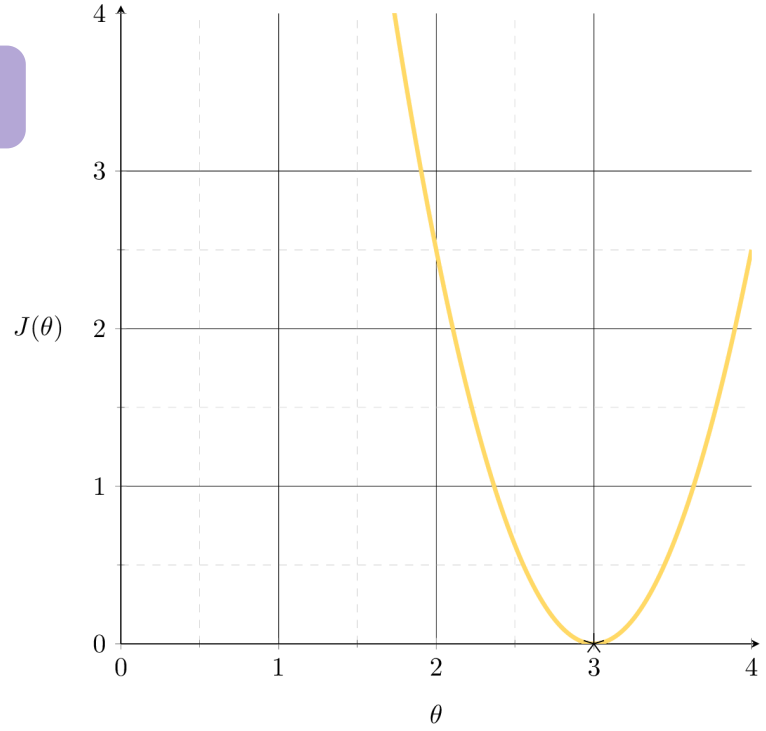
$$\theta^* = (X^\top X)^{-1}(X^\top Y) \quad = \frac{X^\top Y}{X^\top X} = \frac{6}{3} = 2$$



$$\theta^* = (X^\top X)^{-1} X^\top Y$$



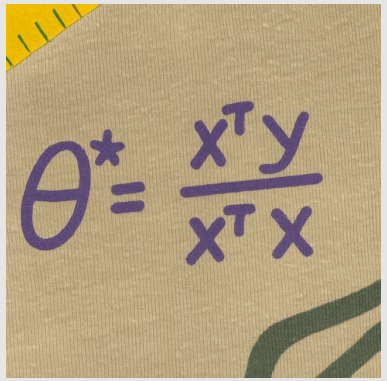
$$J(\theta) = \frac{1}{2}\left[(\theta - 3)^2 + (2\theta - 6)^2\right]$$



$$X = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \qquad Y = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$$

$$\theta^* = (X^\top X)^{-1}(X^\top Y) \quad = \frac{X^\top Y}{X^\top X} = \frac{15}{5} = 3$$

$$\theta^* = \left(X^\top X\right)^{-1} X^\top Y$$



$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \quad Y = \begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix}$$

$$X^\top X = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \quad X^\top Y = \begin{bmatrix} 7 \\ 8 \end{bmatrix}$$

$$\theta^* = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 7 \\ 8 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

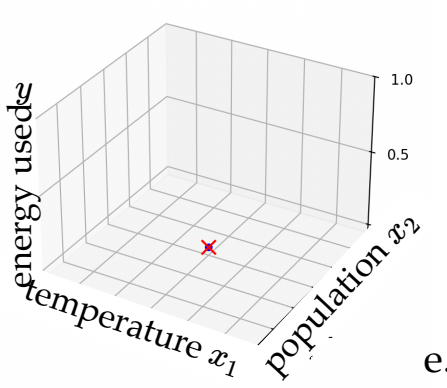
Outline

- The "*trouble*" with the closed-form solution
  - visually, practically, mathematically
- Regularization and ridge regression
- Cross-validation

https://shenshen.mit.edu/demos/ridge/n-less-d-interactive.html?embed

https://shenshen.mit.edu/demos/ridge/collinear-interactive.html?embed

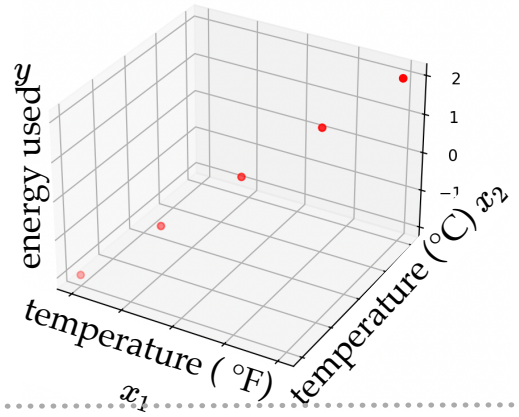
data



(a) $n < d$

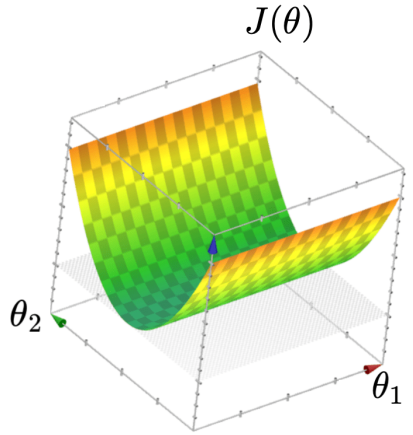
e.g. genomics, NLP

(b) Linearly-dependent features:



e.g. temp °F/°C,
age/birth_year, ...

MSE



closed-form formula $\theta^* = \left(X^\top X\right)^{-1} X^\top Y$ undefined

∞ many optimal θ^*

optimal solution

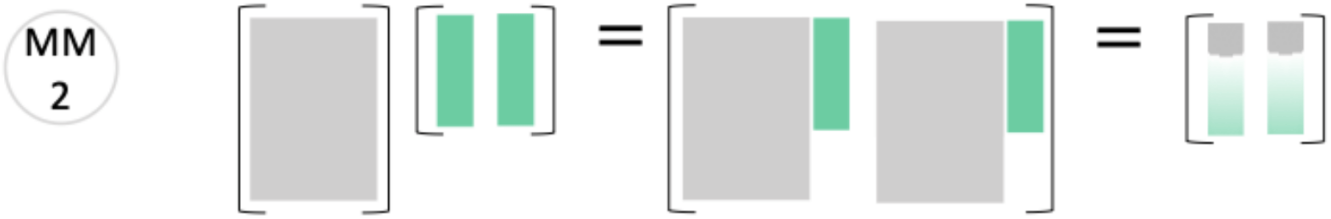
Not enough info to pin down a unique solution

mathematically,

$(X^\top X)$ is singular

$\Updownarrow$

X is not full column rank



Ax and Ay are linear combinations of columns of A.

$$\begin{bmatrix}1 & 2\\ 3 & 4\\ 5 & 6\end{bmatrix}\begin{bmatrix}x_1 & y_1\\ x_2 & y_2\end{bmatrix} = A[\boldsymbol{x} \quad \boldsymbol{y}] = [A\boldsymbol{x} \quad A\boldsymbol{y}]$$

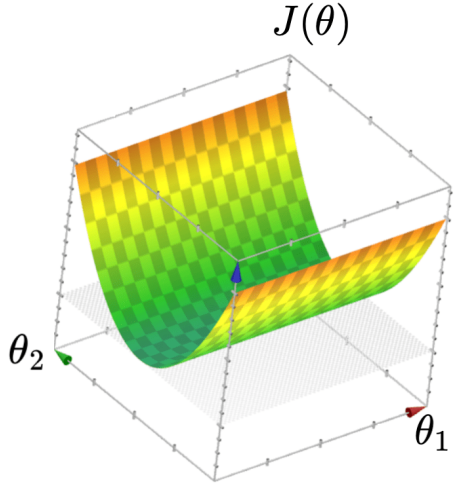
demo (a) $n < d$: 1 sample, 2 features

$$X^\top X = \begin{bmatrix}2\\ 3\end{bmatrix}\begin{bmatrix}2 & 3\end{bmatrix} = \begin{bmatrix}4 & 6\\ 6 & 9\end{bmatrix}$$

demo (b) Collinear: $x_2 = 1.5 \cdot x_1$

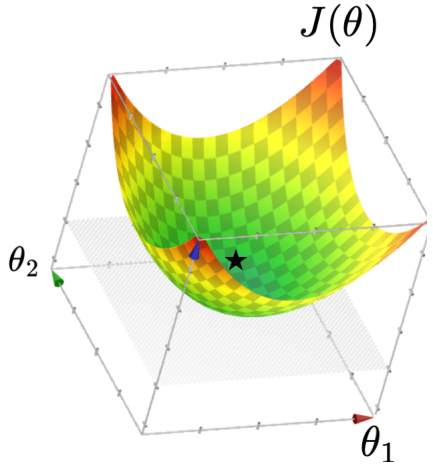
$$X^\top X = \begin{bmatrix}2 & 4 & 6\\ 3 & 6 & 9\end{bmatrix}\begin{bmatrix}2 & 3\\ 4 & 6\\ 6 & 9\end{bmatrix} = \begin{bmatrix}56 & 84\\ 84 & 126\end{bmatrix}$$

https://github.com/kenjihiranabe/The-Art-of-Linear-Algebra
https://www.3blue1brown.com/topics/linear-algebra



When X is not full column rank

- $J(\theta)$ has a "flat" bottom
- This 👉 formula is not well-defined
- Infinitely many optimal hyperplanes



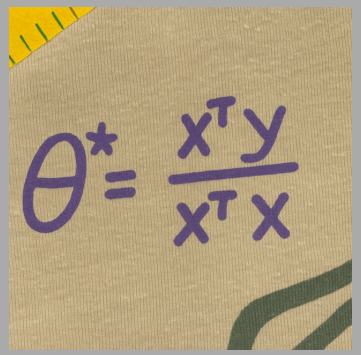
Typically, X is full column rank

- $J(\theta)$ "curves up" everywhere
- $\theta^* = \left(X^\top X\right)^{-1} X^\top Y$
- unique optimal hyperplane

formula isn't wrong, data is trouble-making 🥺

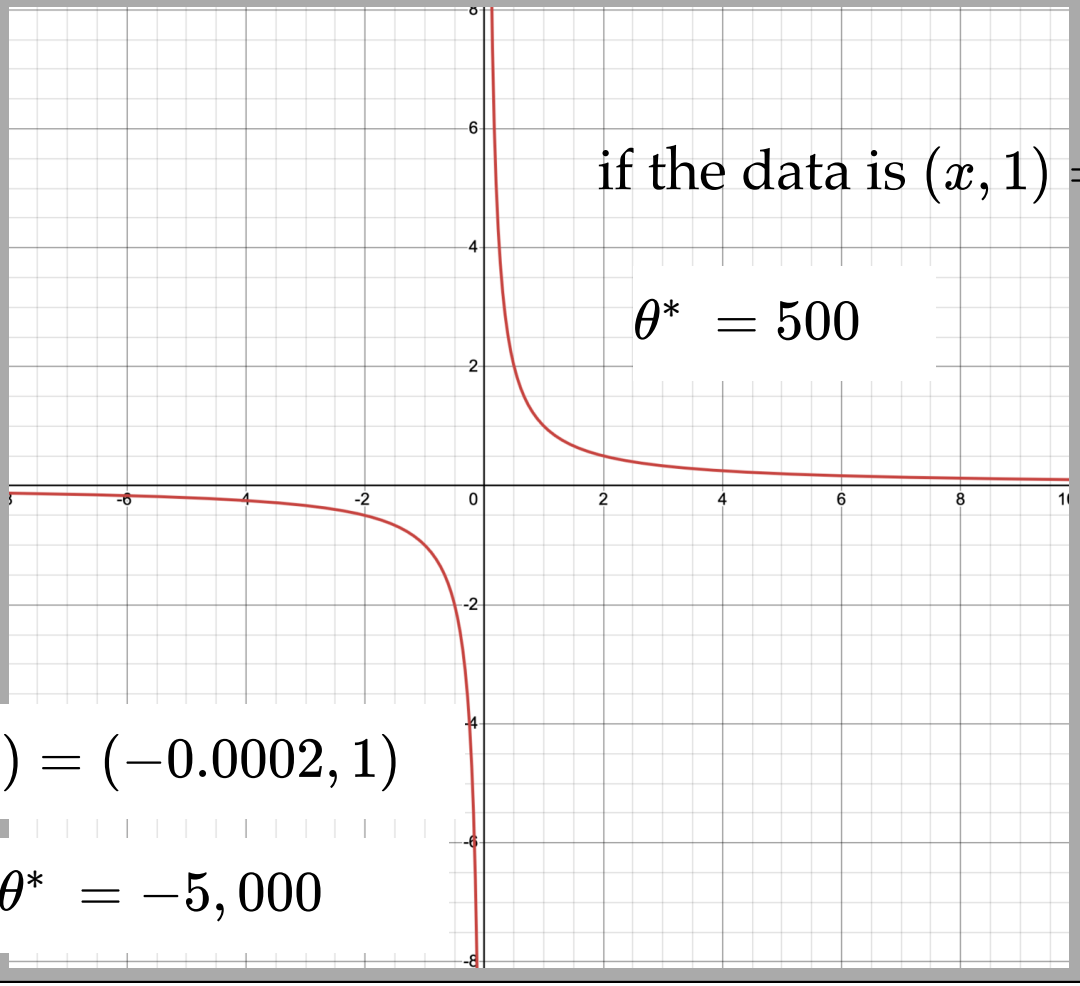
$X^\top X$ is "more" invertible 🥰

minimum eigenvalue of$(X^\top X)$ increasing



assume $n = 1$ and $y = 1$

then $\theta^* = \frac{1}{x}$



if the data is $(x, 1) = (0.002, 1)$

$\theta^* = 500$

if the data is $(x, y) = (-0.0002, 1)$

$\theta^* = -5,000$

https://shenshen.mit.edu/demos/ridge/sensitive-ols.html

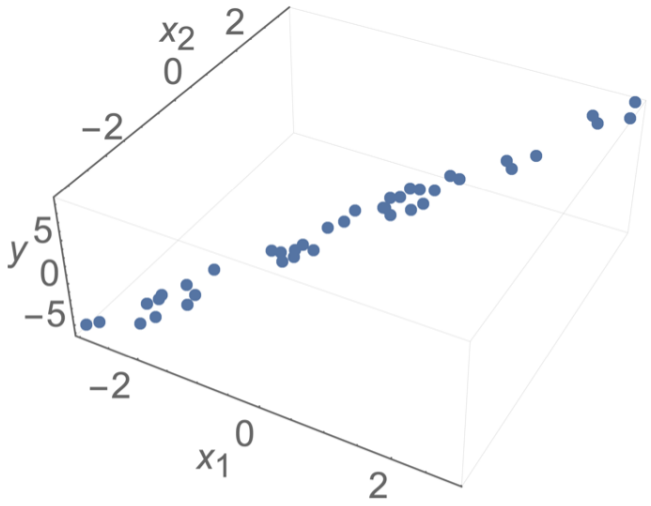
when $X^\top X$ is *almost* singular

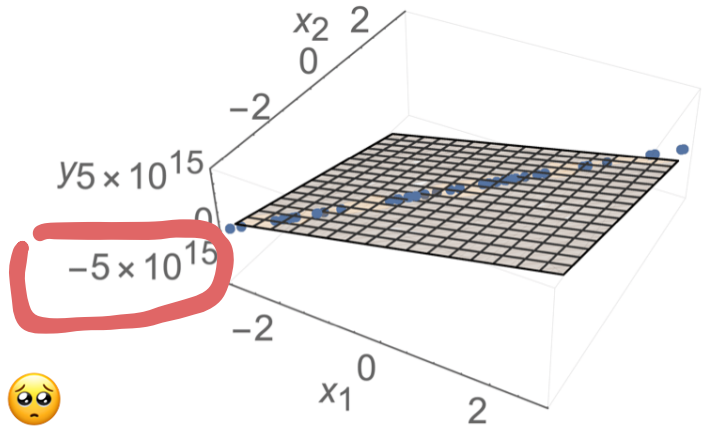
technically, $\theta^* = \left(X^\top X\right)^{-1} X^\top Y$ exists and gives the unique optimal hyperplane

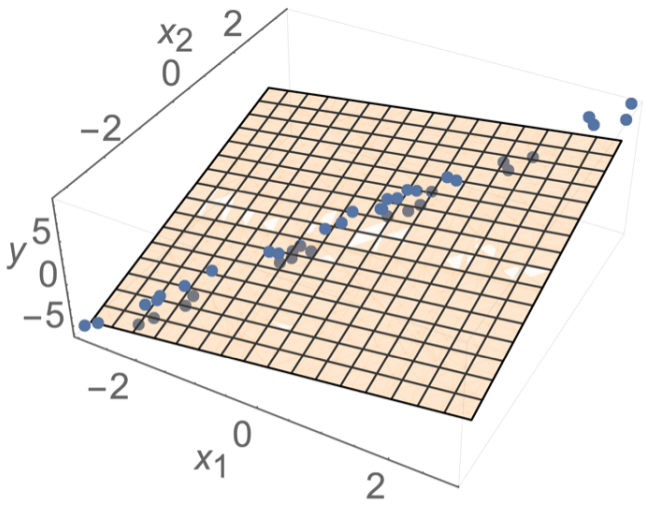
practically, θ^* tends to have huge magnitude

θ^* tends to be very sensitive to the small changes in the data

lots of other θs fit the training data almost equally well







Outline

- The "*trouble*" with the closed-form solution
- Regularization and ridge regression
  - hyperparameters
- Cross-validation

Ridge Regression: Objective

- Many θ give similar loss, but some have huge magnitude... unstable!
  - small change in $x \Rightarrow$ wildly different prediction
- Idea: penalize large θ in our objective, a.k.a. (explicit) *regularization*
- Ridge objective:

$$J_{\text{ridge}}(\theta) = \underbrace{\frac{1}{n}(X\theta - Y)^\top (X\theta - Y)}_{\text{MSE (on training data)}} + \underbrace{\lambda \|\theta\|^2}_{\text{penalty (on parameter magnitude)}}$$

$\lambda > 0$ controls how heavily we penalize magnitude relative to MSE

https://shenshen.mit.edu/demos/ridge/ridge-lambda.html?embed

Ridge Regression: Solution

$$J_{\text{ridge}}(\theta) = \frac{1}{n}(X\theta - Y)^\top (X\theta - Y) + \lambda \|\theta\|^2$$

$$\theta^*_{\text{ridge}} = \left(X^\top X + n\lambda I\right)^{-1} X^\top Y$$

for $\lambda > 0$: $X^\top X + n\lambda I$ is always invertible, so θ^*_{ridge} always exists and is unique

How does λ affect the learned θ?

- $\lambda = 0$?
- $\lambda = 1000$?
- $\lambda = -100$?

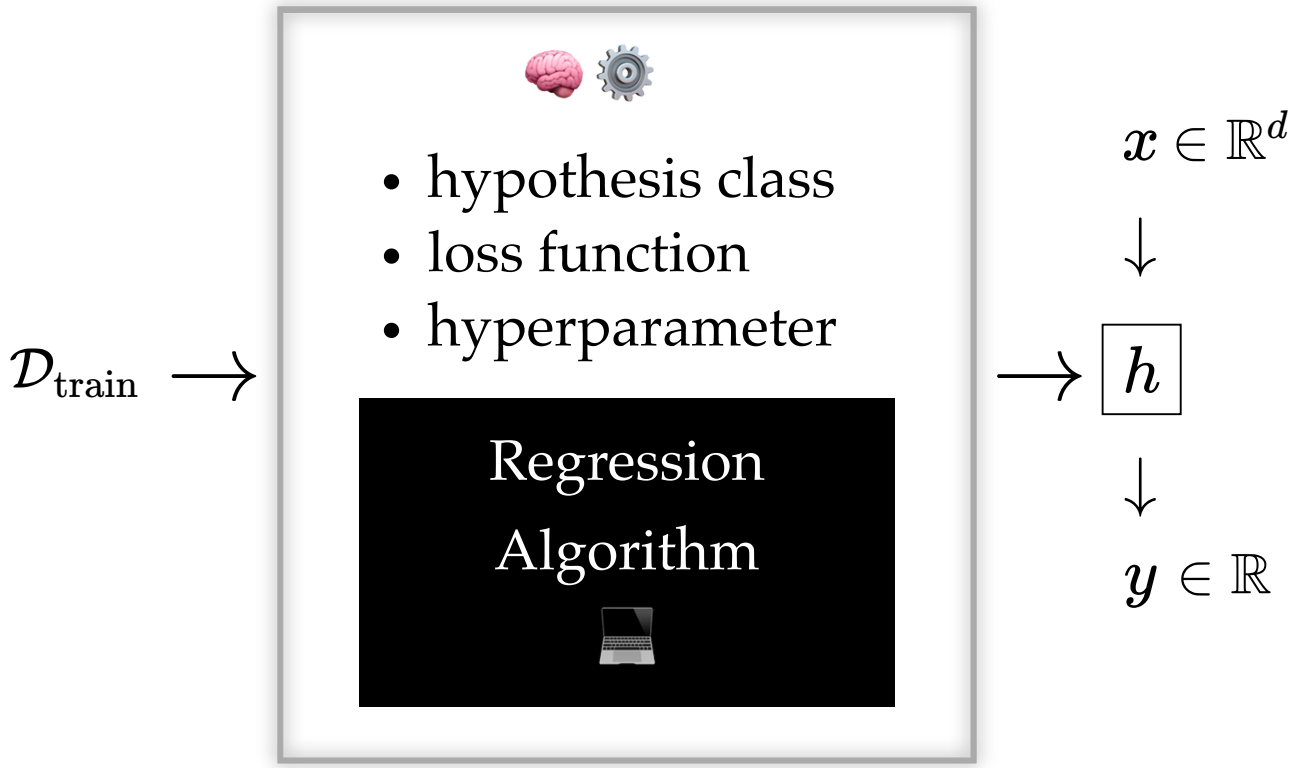
No penalty — reduces to OLS

Huge penalty — forces $\theta \approx 0$

Rewards large θ — counter-productive!

this is why we require $\lambda > 0$

λ is a hyperparameter



- affects learning outcome, and not learned by the algorithm
- we already saw a hyperparameter (in lab 1, how many random regressor tried)

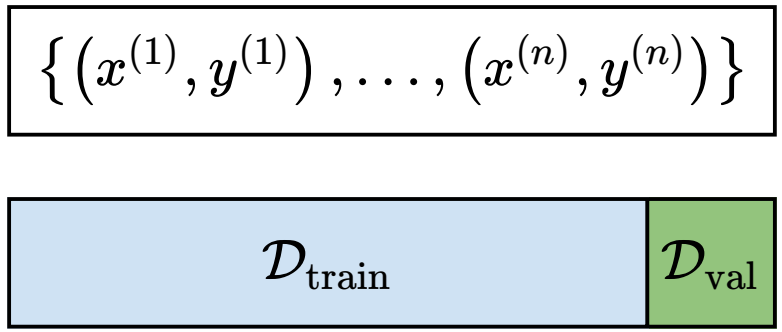
Outline

- The "*trouble*" with the closed-form solution
- Regularization and ridge regression
- Cross-validation

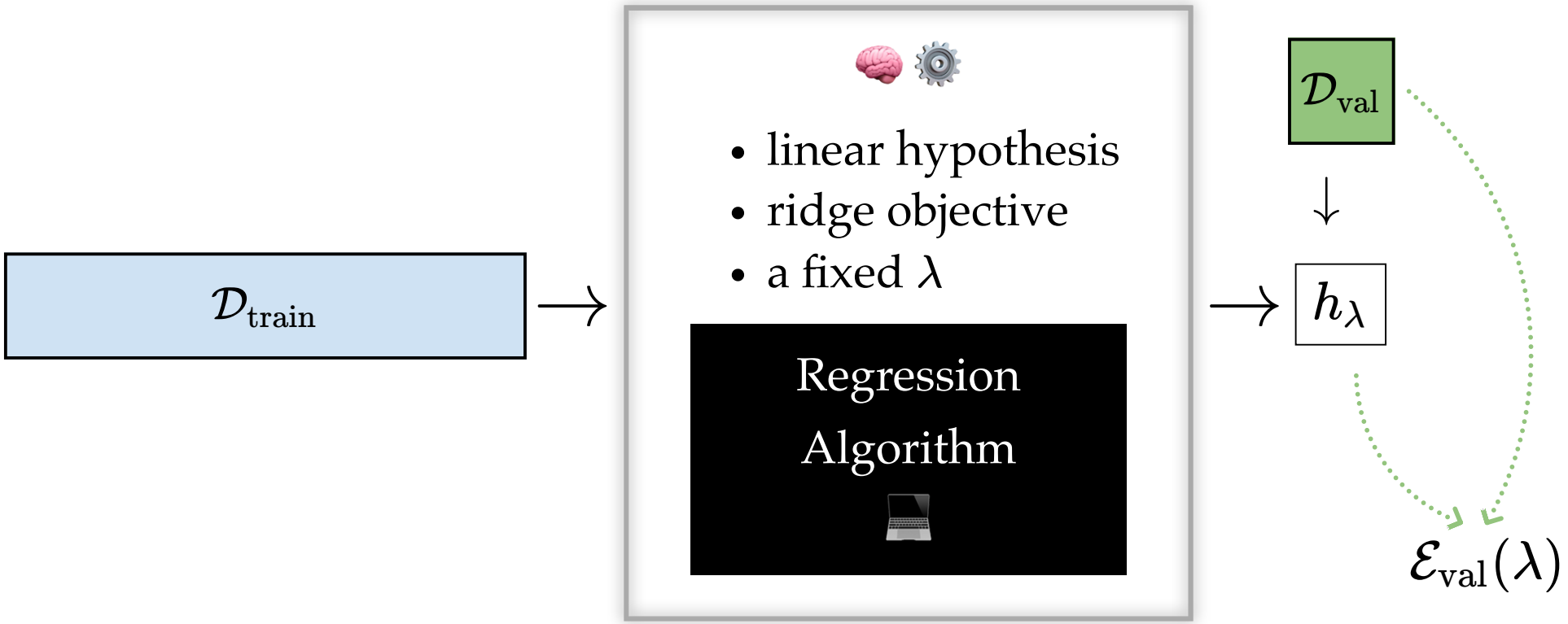
We need to choose hyperparameters (like λ)

- Can't use training error (would always pick $\lambda = 0$)
- Can't use test error (we don't have test data)

- Hold-out some data



- Use $\mathcal{D}_{\text{val}}$ to evaluate how good a hyperparameter λ is



for each $\lambda \in \{0.1, 1, 10\}$:

train on $\mathcal{D}_{\text{train}}$ with λ

compute $\mathcal{E}_{\text{val}}(\lambda)$ on $\mathcal{D}_{\text{val}}$

return $\arg\min_\lambda \mathcal{E}_{\text{val}}(\lambda)$

in this example, compare $\mathcal{E}_{\text{val}}(.1)$, $\mathcal{E}_{\text{val}}(1)$, and $\mathcal{E}_{\text{val}}(10)$, return the λ corresponding to smallest validation error

Cross-validation

for each $\lambda \in \{0.1, 1, 10\}$:

  for $i = 1, \ldots, 5$:

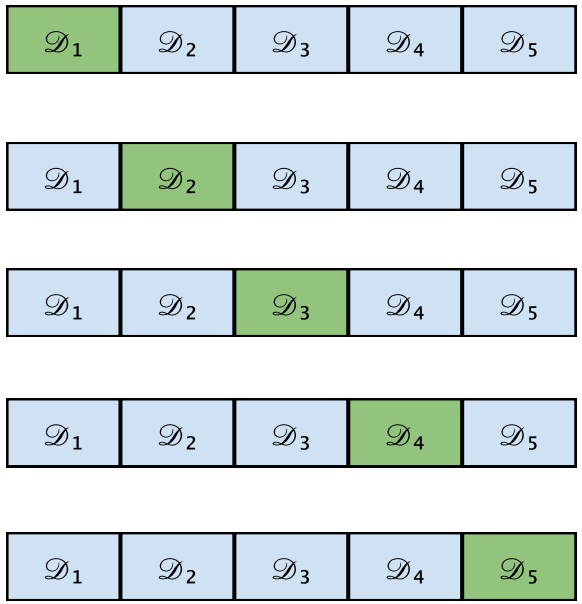
    train h_i on $\mathcal{D} \setminus \mathcal{D}_i$ with λ

    $\mathcal{E}_i$ = error on $\mathcal{D}_i$

  $\mathcal{E}_{\text{val}}(\lambda) = (\mathcal{E}_1 + \cdots + \mathcal{E}_5)/5$

return $\lambda^* = \arg\min_\lambda \mathcal{E}_{\text{val}}(\lambda)$

outer loop of $\lambda \in \{0.1, 1, 10\}$:



$\mathcal{E}_{\text{val}}(\lambda) = (\mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 + \mathcal{E}_4 + \mathcal{E}_5)/5$

How many hypotheses trained in this example to pick λ^*?

$$\theta^*_{\text{final}} = (X^\top X + n\lambda^* I)^{-1} X^\top Y$$

finally train using the chosen λ^*, and data from all of $\mathcal{D}$

Summary

- When $X^\top X$ is singular or ill-conditioned, OLS is undefined or overfits.
- Regularization combats overfitting by penalizing large θ.
- Ridge regression adds $\lambda \|\theta\|^2$ to the objective — still has a closed-form solution.
- λ is a hyperparameter that trades off fit vs. regularization.
- Validation and cross-validation provide principled ways to choose λ.

Cross-validation

for each $\lambda \in \{0.1, 1, 10\}$:

  for $i = 1, \ldots, 5$:

    train h_i on $\mathcal{D} \setminus \mathcal{D}_i$ with λ

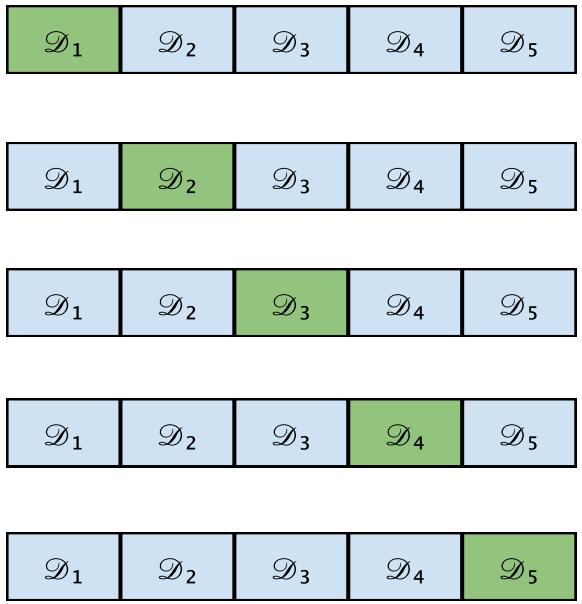
    $\mathcal{E}_i$ = error on $\mathcal{D}_i$

  $\mathcal{E}_{\text{val}}(\lambda) = (\mathcal{E}_1 + \cdots + \mathcal{E}_5)/5$

return $\lambda^* = \arg\min_\lambda \ \mathcal{E}_{\text{val}}(\lambda)$

$$\theta^*_{\text{final}} = (X^\top X + n\lambda^* I)^{-1} X^\top Y$$

outer loop of $\lambda \in \{0.1, 1, 10\}$:



$$\mathcal{E}_{\text{val}}(\lambda) = (\mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 + \mathcal{E}_4 + \mathcal{E}_5)/5$$

How many hypotheses trained in this example to pick λ^*?

finally train using data from all of $\mathcal{D}$