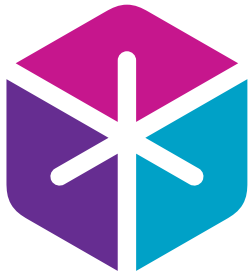
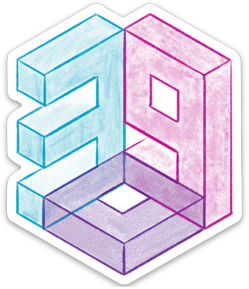




https://introml.mit.edu/

6.390 Intro to Machine Learning

Lecture 2: Regression and Regularization

Shen Shen

Sep 14, 2026

2:30pm, Room 10-250

Slides and Lecture Recording

6.390 topics in order:

- Intro to ML
- Regression and Regularization
- Gradient Descent
- Linear Classification
- Features, Neural Networks I
- Neural Networks II (Backprop)
- Convolutional Neural Networks
- Representation Learning
- Transformers
- Markov Decision Processes
- Reinforcement Learning
- Non-parametric Models

Recall week 1:

- Terminologies:

  training data | features | label | hypothesis | hypothesis class | loss function | parameters

  regression | training error | test data | classification | supervised learning | test error

- Ordinary least squares regression:
  - matrix-vector form objective $J(\theta) = \frac{1}{n}(X\theta - Y)^\top (X\theta - Y)$
  - closed-form solution $\theta^* = \left(X^\top X\right)^{-1} X^\top Y$

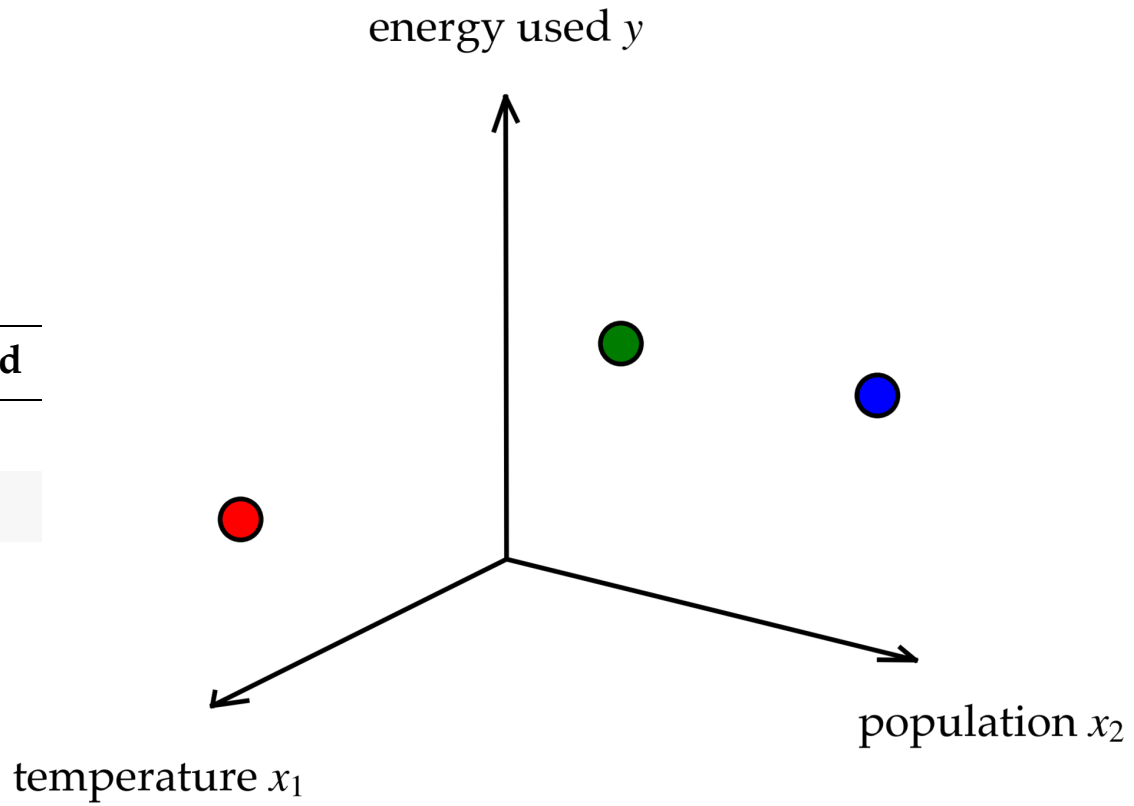
training data

City	Features		Label
	Temperature	Population	Energy Used
Chicago	1	0	2
New York	0	1	3
Boston	1	1	5

n (Chicago, New York, Boston)

d (Temperature, Population)

1 (Energy Used)



Assemble in matrix-vector form:

features

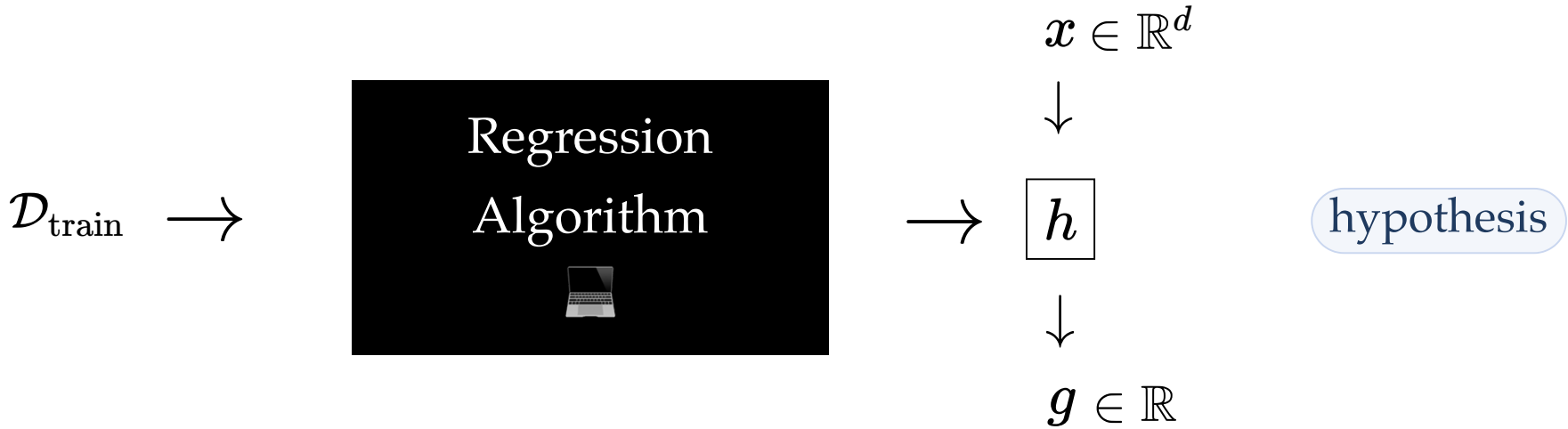
$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \in \mathbb{R}^{3\times 2}$$

label

$$Y = \begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix} \in \mathbb{R}^{3\times 1}$$

[contrived data]

learning algorithm

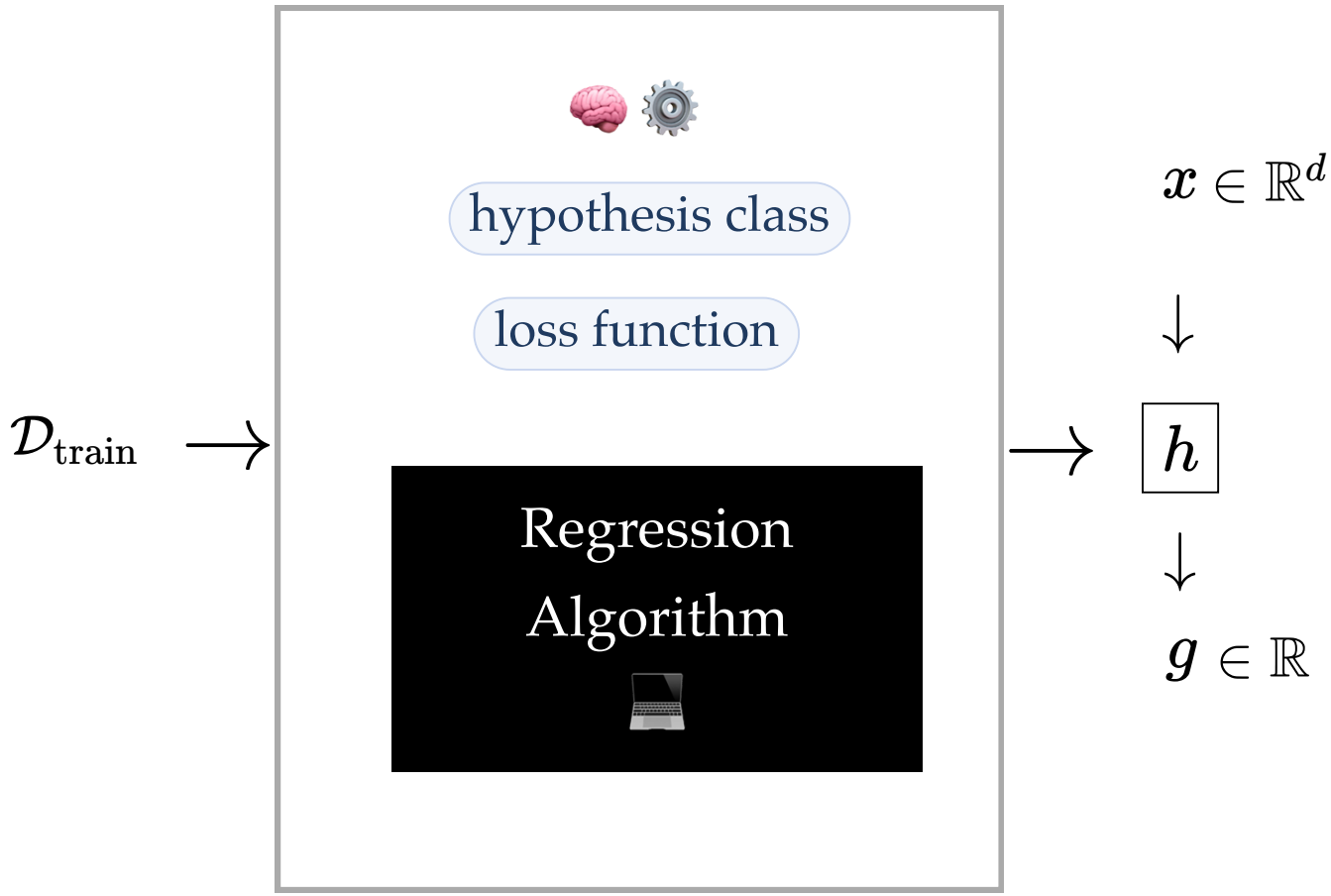


hypothesis

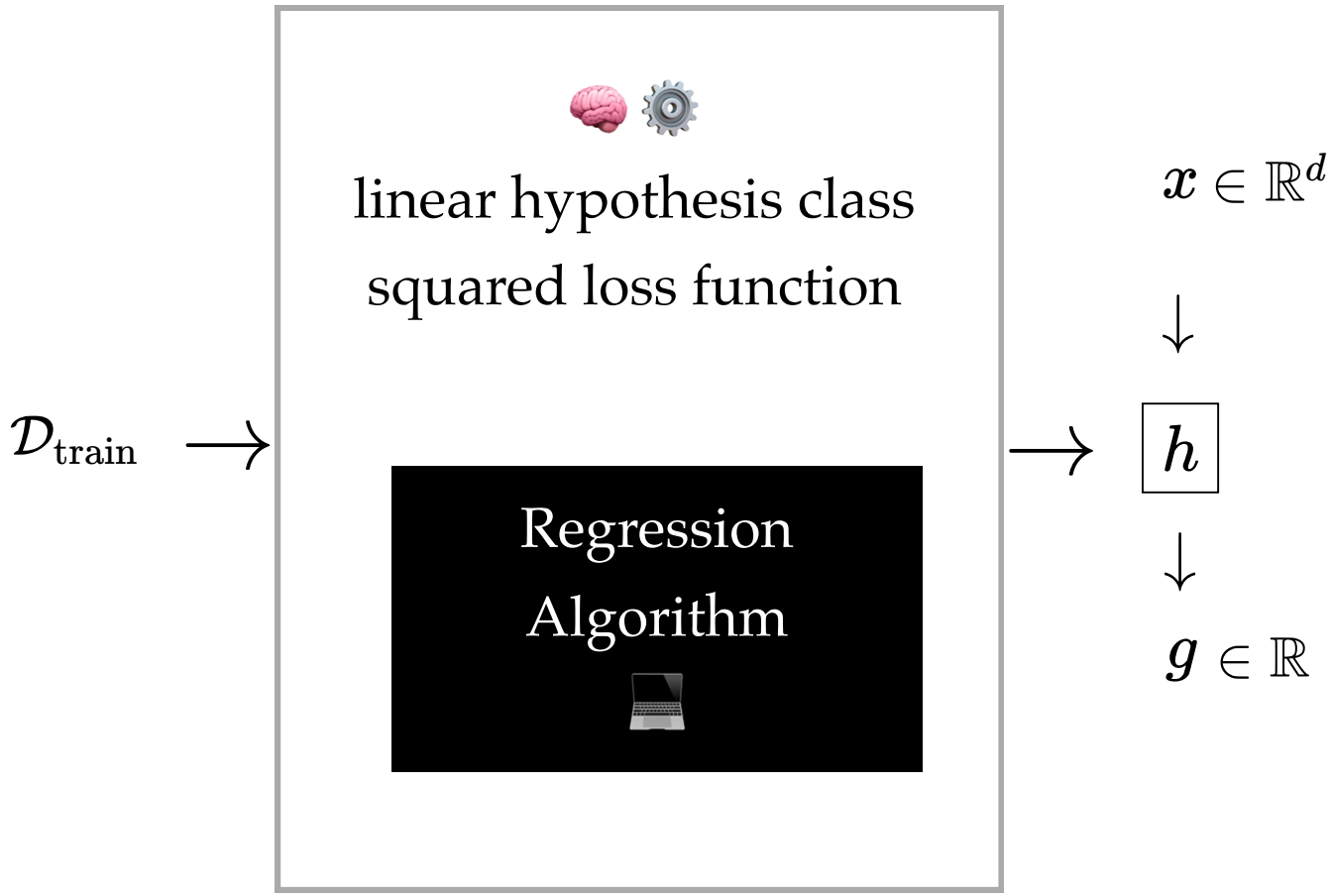
What do we want from the algorithm?

A good way to label *new* features, i.e. a *good* hypothesis.

The "good"ness depends on human choices, even before the algorithm does anything



Ordinary least squares regression



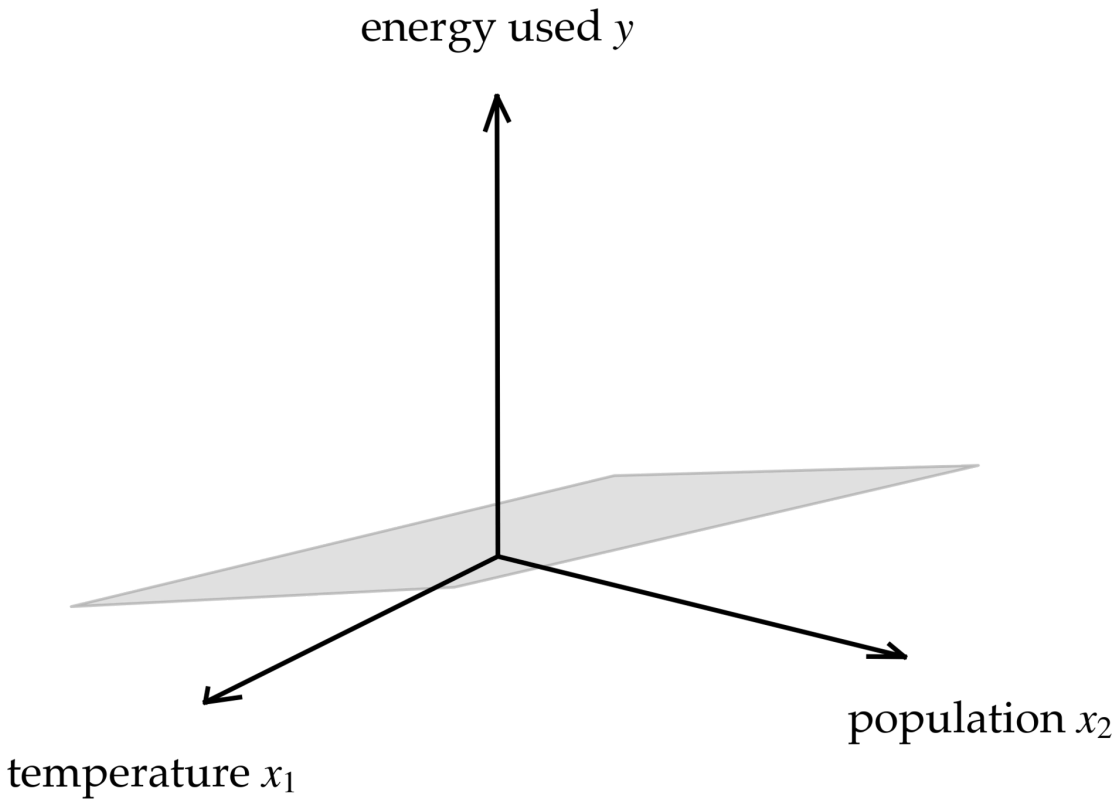
Linear hypothesis class:

$$h(\theta; x) = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

$$= \theta^\top x$$

parameters

features

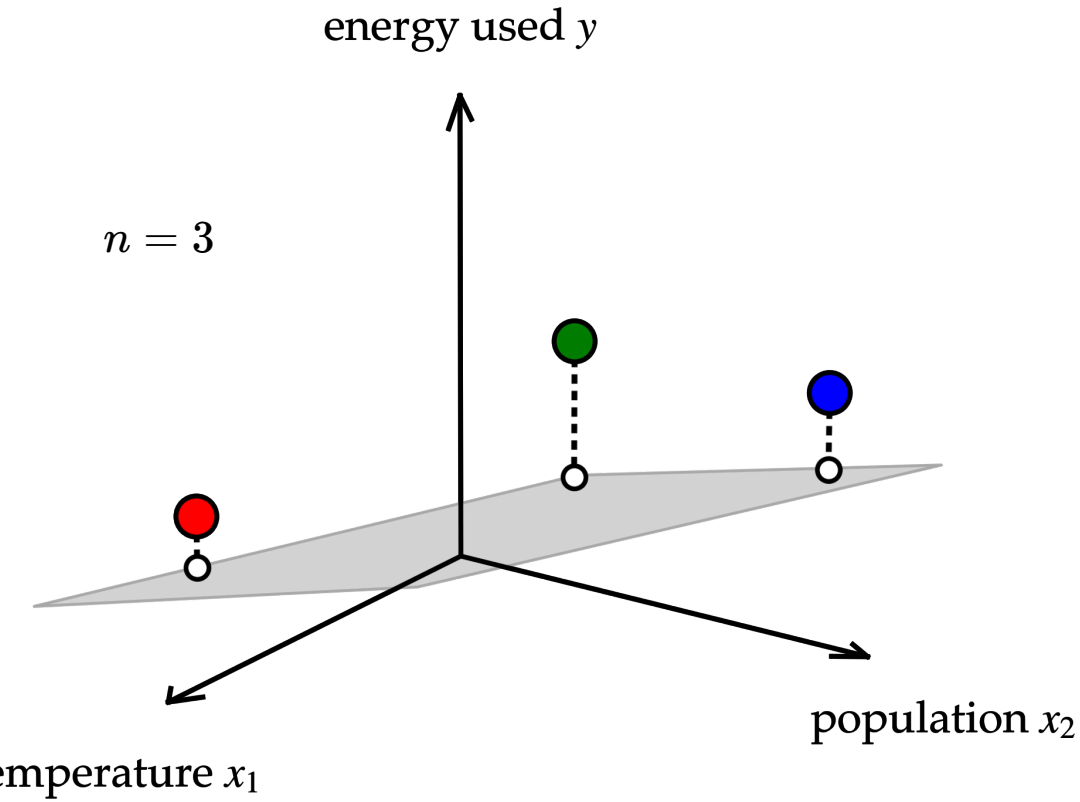


see recitations/labs for how to handle the offset θ_0

Squared loss:

City	Features		Label
	Temperature	Population	Energy Used
Chicago	1	0	2
New York	0	1	3
Boston	1	1	5

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

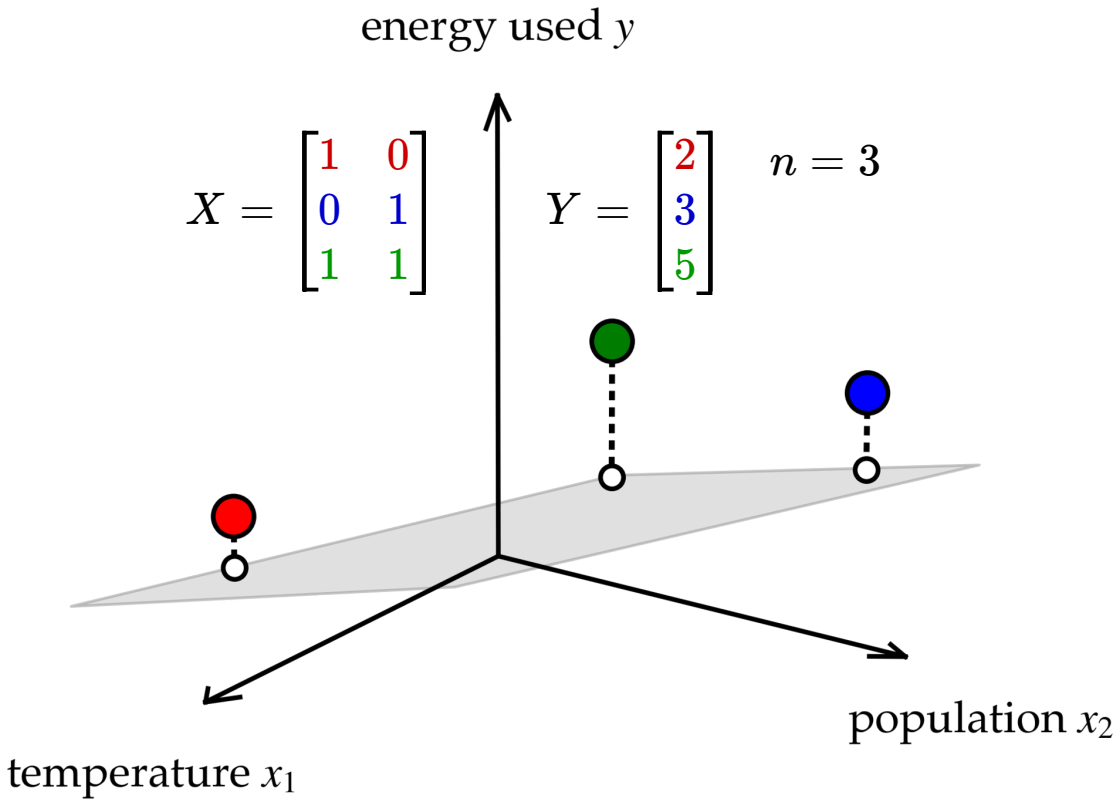


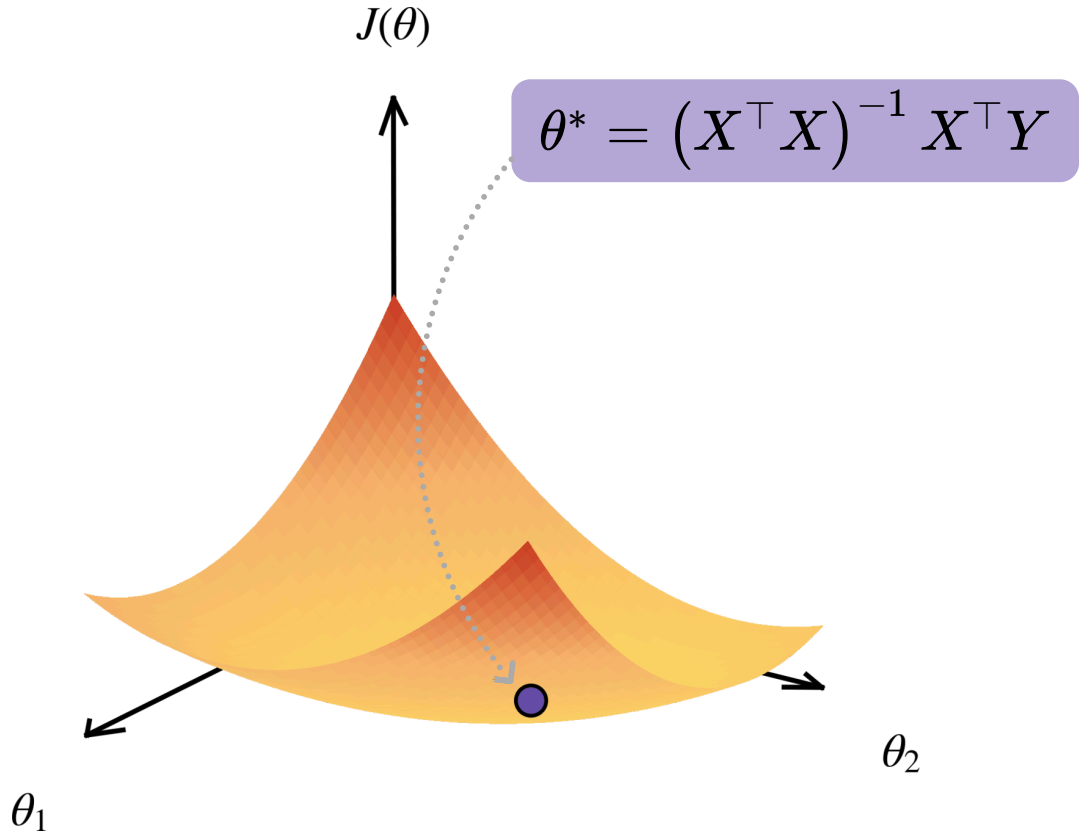
for a given data set, training error only depends on the parameter

$$J(\theta) = \frac{1}{3}\left[(\theta_1 \cdot 1 + \theta_2 \cdot 0 - 2)^2 + (\theta_1 \cdot 0 + \theta_2 \cdot 1 - 3)^2 + (\theta_1 \cdot 1 + \theta_2 \cdot 1 - 5)^2\right]$$

MSE written in matrix-vector form:

$$J(\theta) = \frac{1}{n}(X\theta - Y)^\top (X\theta - Y)$$

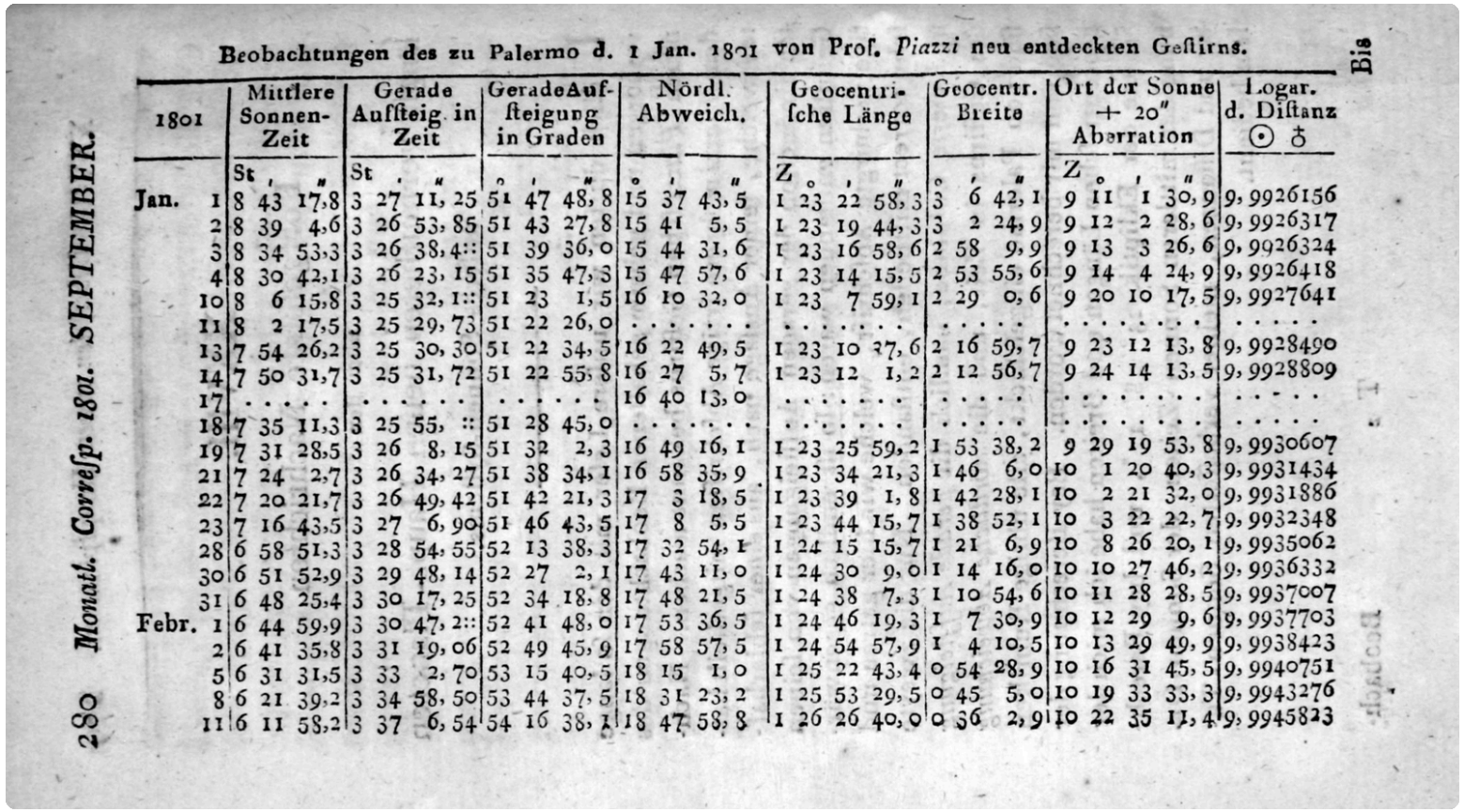




$$J(\theta) = \frac{1}{3}\left[(\theta_1 \cdot 1 + \theta_2 \cdot 0 - 2)^2 + (\theta_1 \cdot 0 + \theta_2 \cdot 1 - 3)^2 + (\theta_1 \cdot 1 + \theta_2 \cdot 1 - 5)^2\right]$$

https://shenshen.mit.edu/demos/ridge/d2-unique-solution.html

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

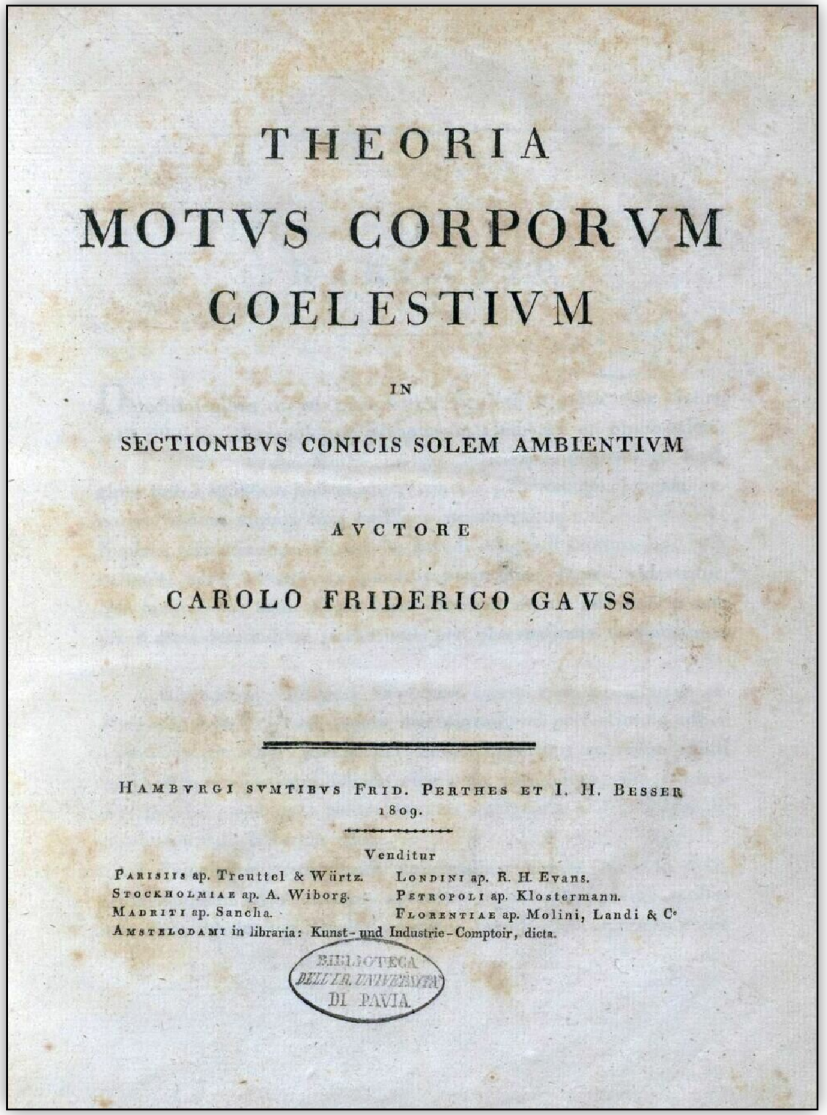


Beobachtungen des zu Palermo d. 1 Jan. 1801 von Prof. *Piazzi* neu entdeckten Geſtirns.

280 *Monatl. Corresp. 1801.* SEPTEMBER.

1801		Mittlere Sonnen-Zeit St ′ ″	Gerade Aufſteig. in Zeit St ′ ″	Gerade Aufſteigung in Graden ° ′ ″	Nördl. Abweich. ° ′ ″	Geocentriſche Länge Z ° ′ ″	Geocentr. Breite ° ′ ″	Ort der Sonne + 20″ Aberration Z ° ′ ″	Logar. d. Diſtanz ☉ ♁
Jan.	1	8 43 17,8	3 27 11,25	51 47 48,8	15 37 43,5	1 23 22 58,3	3 6 42,1	9 11 1 30,9	9,9926156
	2	8 39 4,6	3 26 53,85	51 43 27,8	15 41 5,5	1 23 19 44,3	3 2 24,9	9 12 2 28,6	9,9926317
	3	8 34 53,3	3 26 38,4::	51 39 36,0	15 44 31,6	1 23 16 58,6	2 58 9,9	9 13 3 26,6	9,9926324
	4	8 30 42,1	3 26 23,15	51 35 47,3	15 47 57,6	1 23 14 15,5	2 53 55,6	9 14 4 24,9	9,9926418
	10	8 6 15,8	3 25 32,1::	51 23 1,5	16 10 32,0	1 23 7 59,1	2 29 0,6	9 20 10 17,5	9,9927641
	11	8 2 17,5	3 25 29,73	51 22 26,0	. . . . .	. . . . .	. . . .	. . . . .	. . . .
	13	7 54 26,2	3 25 30,30	51 22 34,5	16 22 49,5	1 23 10 37,6	2 16 59,7	9 23 12 13,8	9,9928490
	14	7 50 31,7	3 25 31,72	51 22 55,8	16 27 5,7	1 23 12 1,2	2 12 56,7	9 24 14 13,5	9,9928809
	17	. . . . .	. . . . .	. . . . .	16 40 13,0	. . . . .	. . . .	. . . . .	. . . .
	18	7 35 11,3	3 25 55,::	51 28 45,0	. . . . .	. . . . .	. . . .	. . . . .	. . . .
	19	7 31 28,5	3 26 8,15	51 32 2,3	16 49 16,1	1 23 25 59,2	1 53 38,2	9 29 19 53,8	9,9930607
	21	7 24 2,7	3 26 34,27	51 38 34,1	16 58 35,9	1 23 34 21,3	1 46 6,0	10 1 20 40,3	9,9931434
	22	7 20 21,7	3 26 49,42	51 42 21,3	17 3 18,5	1 23 39 1,8	1 42 28,1	10 2 21 32,0	9,9931886
	23	7 16 43,5	3 27 6,90	51 46 43,5	17 8 5,5	1 23 44 15,7	1 38 52,1	10 3 22 22,7	9,9932348
	28	6 58 51,3	3 28 54,55	52 13 38,3	17 32 54,1	1 24 15 15,7	1 21 6,9	10 8 26 20,1	9,9935062
	30	6 51 52,9	3 29 48,14	52 27 2,1	17 43 11,0	1 24 30 9,0	1 14 16,0	10 10 27 46,2	9,9936332
	31	6 48 25,4	3 30 17,25	52 34 18,8	17 48 21,5	1 24 38 7,3	1 10 54,6	10 11 28 28,5	9,9937007
Febr.	1	6 44 59,9	3 30 47,2::	52 41 48,0	17 53 36,5	1 24 46 19,3	1 7 30,9	10 12 29 9,6	9,9937703
	2	6 41 35,8	3 31 19,06	52 49 45,9	17 58 57,5	1 24 54 57,9	1 4 10,5	10 13 29 49,9	9,9938423
	5	6 31 31,5	3 33 2,70	53 15 40,5	18 15 1,0	1 25 22 43,4	0 54 28,9	10 16 31 45,5	9,9940751
	8	6 21 39,2	3 34 58,50	53 44 37,5	18 31 23,2	1 25 53 29,5	0 45 5,0	10 19 33 33,3	9,9943276
	11	6 11 58,2	3 37 6,54	54 16 38,1	18 47 58,8	1 26 26 40,0	0 36 2,9	10 22 35 11,4	9,9945823

Observation table: Piazzi's measurements of Ceres (1801).



THEORIA
MOTVS CORPORVM
COELESTIVM
IN
SECTIONIBVS CONICIS SOLEM AMBIENTIVM
AVCTORE
CAROLO FRIDERICO GAVSS

HAMBVRGI SVMTIBVS FRID. PERTHES ET I. H. BESSER
1809.

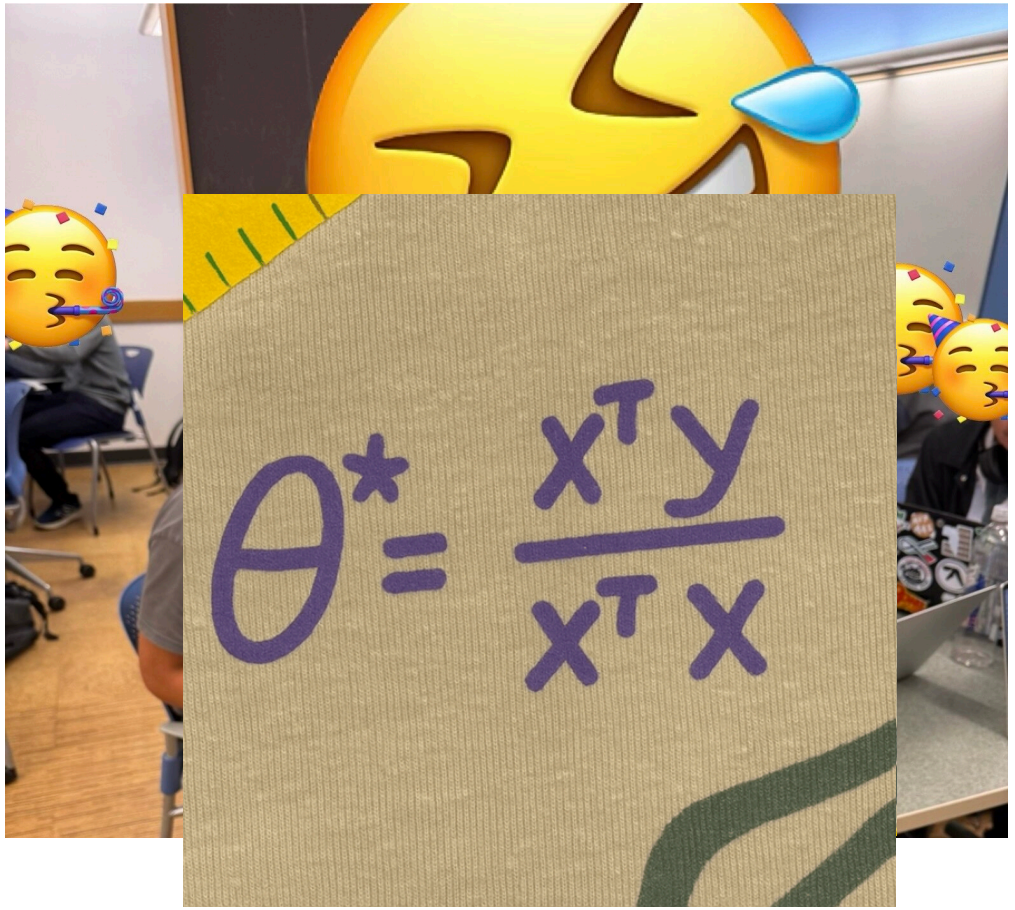
Venditur
PARISIIS ap. Treuttel & Würtz. LONDINI ap. R. H. Evans.
STOCKHOLMIAE ap. A. Wiborg. PETROPOLI ap. Klostermann.
MADRITI ap. Sancha. FLORENTIAE ap. Molini, Landi & C°
AMSTELODAMI in libraria: Kunst- und Industrie-Comptoir, dicta.

Primary source: Gauss, *Theoria motus* (1809).

Still a strong baseline in this era of deep learning and agentic AI

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

Jane street shirt

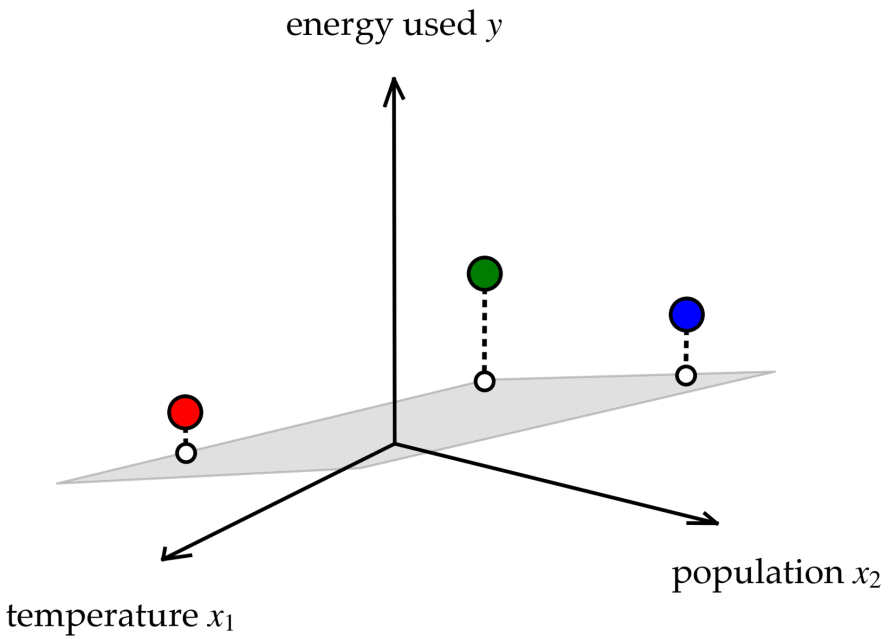


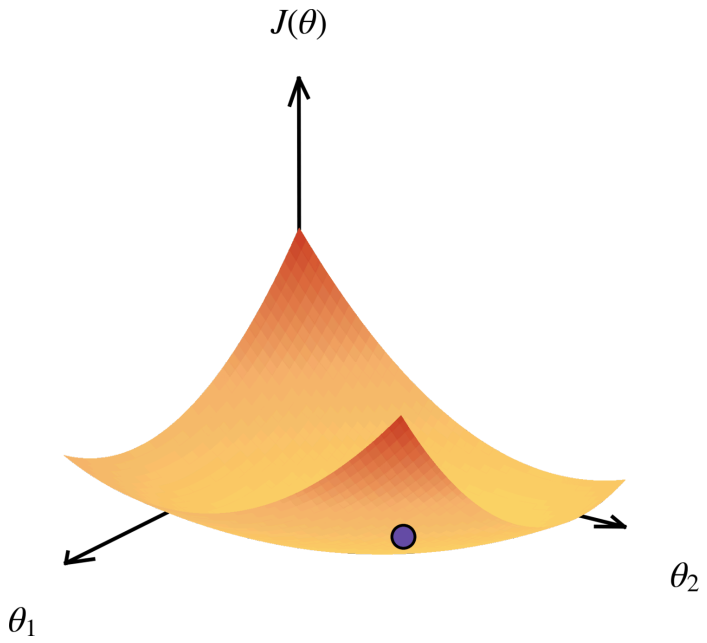
why does the shirt formula look different?

shirt formula assumes one-dimensional feature

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

- We solve it in closed form, so it never feels like training.
- This is a rare case where we get a clean, general solution with a theoretical guarantee.
- When well-defined, θ^* is the unique minimizer of $J(\theta)$.





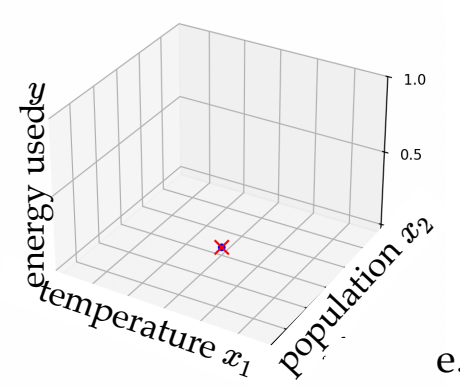
Outline

- The "*trouble*" with the closed-form solution
  - visually, practically, mathematically
- Regularization and ridge regression
  - λ, a hyperparameter
- Validation and cross-validation

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

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

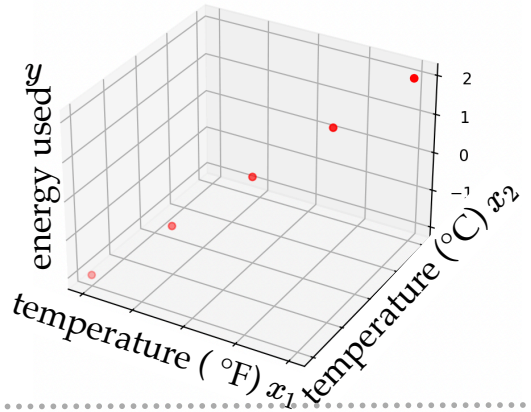
data



(a) $n < d$

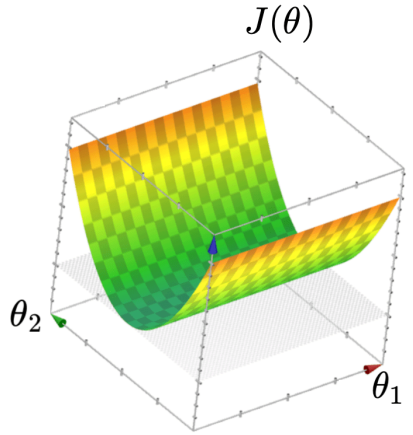
e.g. genomics, NLP

(b) linearly-dependent features (collinear):



e.g. temp °F/°C,
age/birth_year, ...

MSE



optimal solution

∞ many optimal θ^*

closed-form formula

$\theta^* = (X^\top X)^{-1} X^\top Y$ not well-undefined

Not enough info to pin down a unique solution

mathematically, formula not well-defined when:

$(X^\top X)$ is singular

$\Updownarrow$

X is not full column rank

demo (a)

$n < d$: 1 sample, 2 features

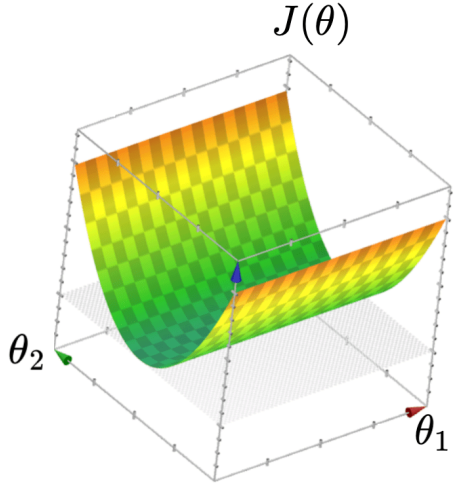
$$X^\top \quad X \quad = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 2 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 6 \\ 6 & 9 \end{bmatrix}$$

demo (b)

collinear: $x_2 = 1.5 \cdot x_1$

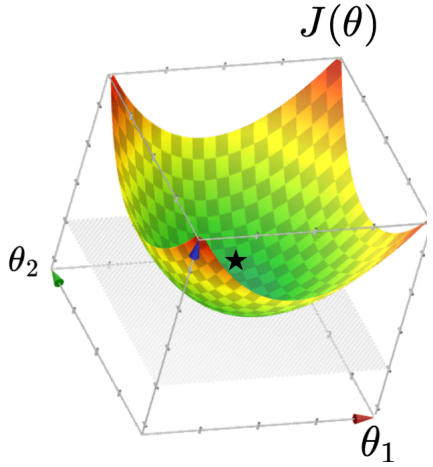
$$X^\top X \quad = \begin{bmatrix} 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 4 & 6 \\ 6 & 9 \end{bmatrix} = \begin{bmatrix} 56 & 84 \\ 84 & 126 \end{bmatrix}$$

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
- That 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

$X^\top X$ singular

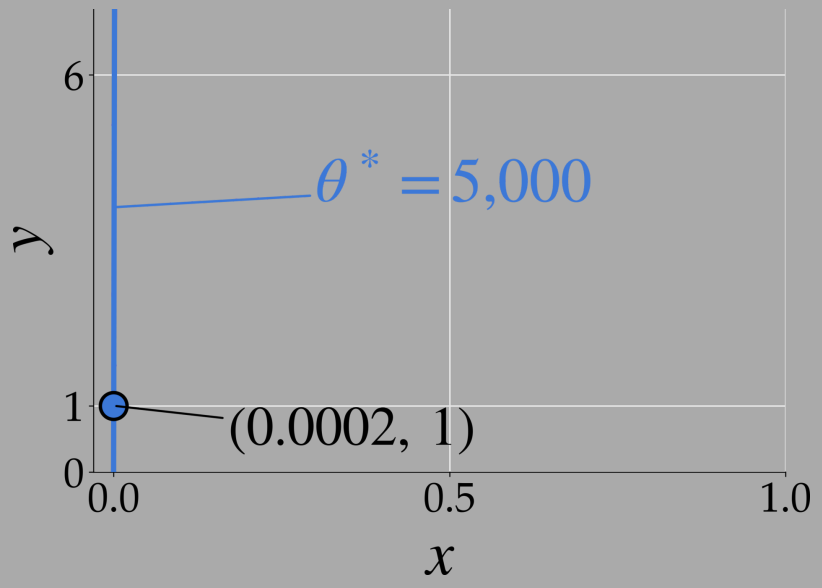
$X^\top X$ "invertibility"

(the formula isn't wrong; data is trouble-making 🥺)

When $X^\top X$ is *almost* singular: $\theta^* = (X^\top X)^{-1} X^\top Y$ technically is well-defined

e.g., single 1d data point (x, y)

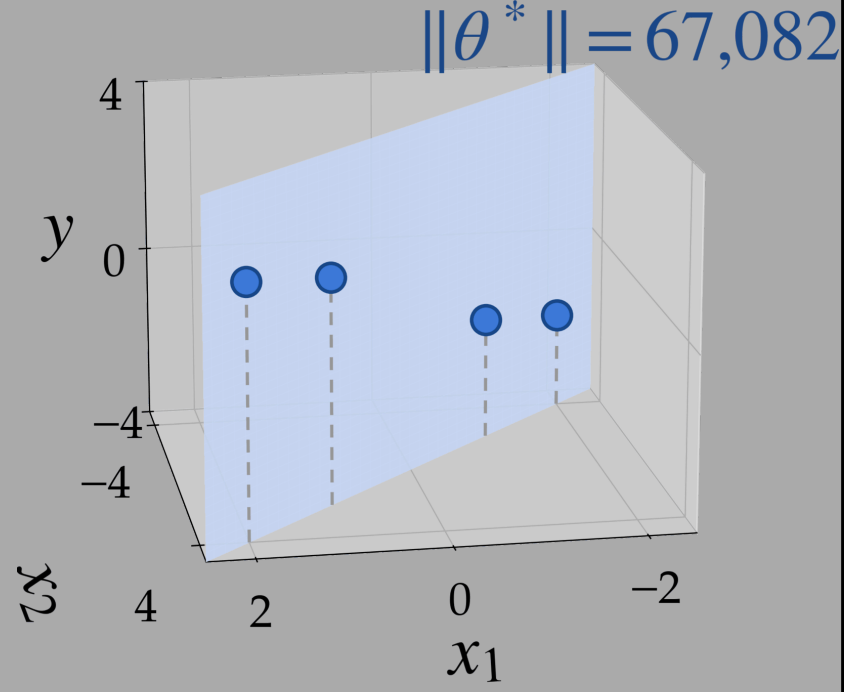
formula simplifies to $\theta^* = \frac{y}{x}$



e.g.,

$$X = \begin{bmatrix} -2 & -4.00006 \\ -1 & -2.00004 \\ 1 & 2.00004 \\ 2 & 4.00006 \end{bmatrix} \quad Y = \begin{bmatrix} -1.8 \\ -1.2 \\ 1.2 \\ 1.8 \end{bmatrix}$$

formula gives $\theta^* \approx \begin{bmatrix} -60000 \\ 30000 \end{bmatrix}$



if the 2nd feature changes by 0.1, prediction changes 3000

θ^* tends to have huge magnitude and very sensitive to the data

meanwhile, lots of other θs fit the training data very well too

https://shenshen.mit.edu/demos/ridge/ill-conditioned-two-planes.html

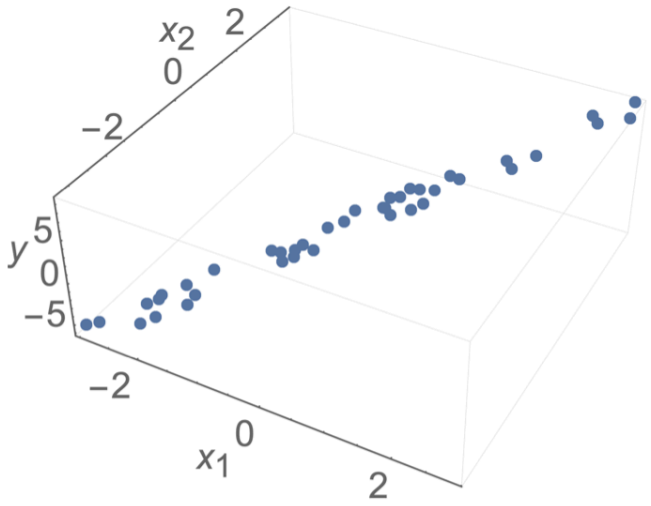
when $X^\top X$ is *almost* singular

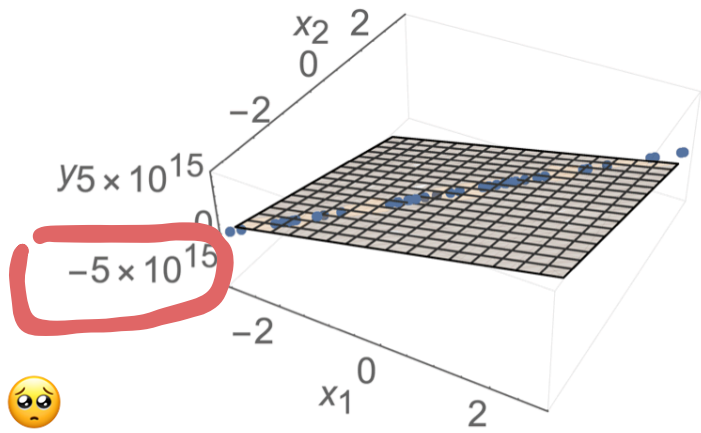
$\theta^* = (X^\top X)^{-1} X^\top Y$ technically exists and is the unique optimal

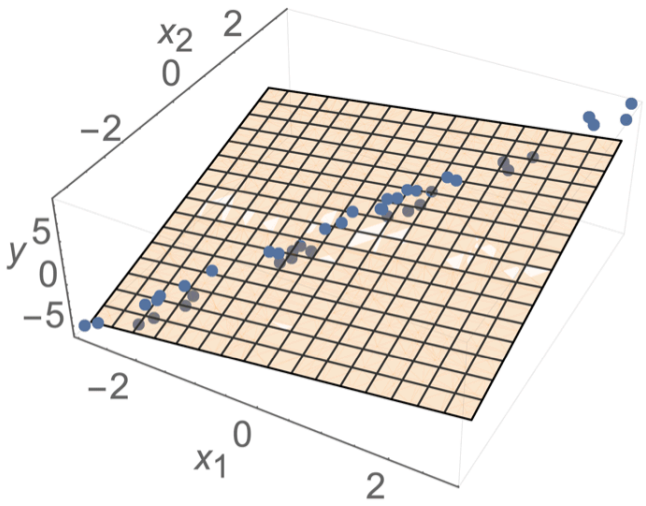
θ^* tends to have huge magnitude and very sensitive to the data

θ^* tends to overly cater to the training data --- overfitting

lots of other θs fit the training data very well too







Outline

- The "*trouble*" with the closed-form solution
- Regularization and ridge regression
  - λ, a hyperparameter
- Validation and cross-validation

Ridge regularization

for the given training data set (so X, Y, n are constants) and a given $\lambda > 0$:

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

still searching for parameters θ

λ controls how heavily we penalize magnitude relative to MSE

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

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

alleviates overfitting by sacrificing MSE

How does λ affect the learned θ?

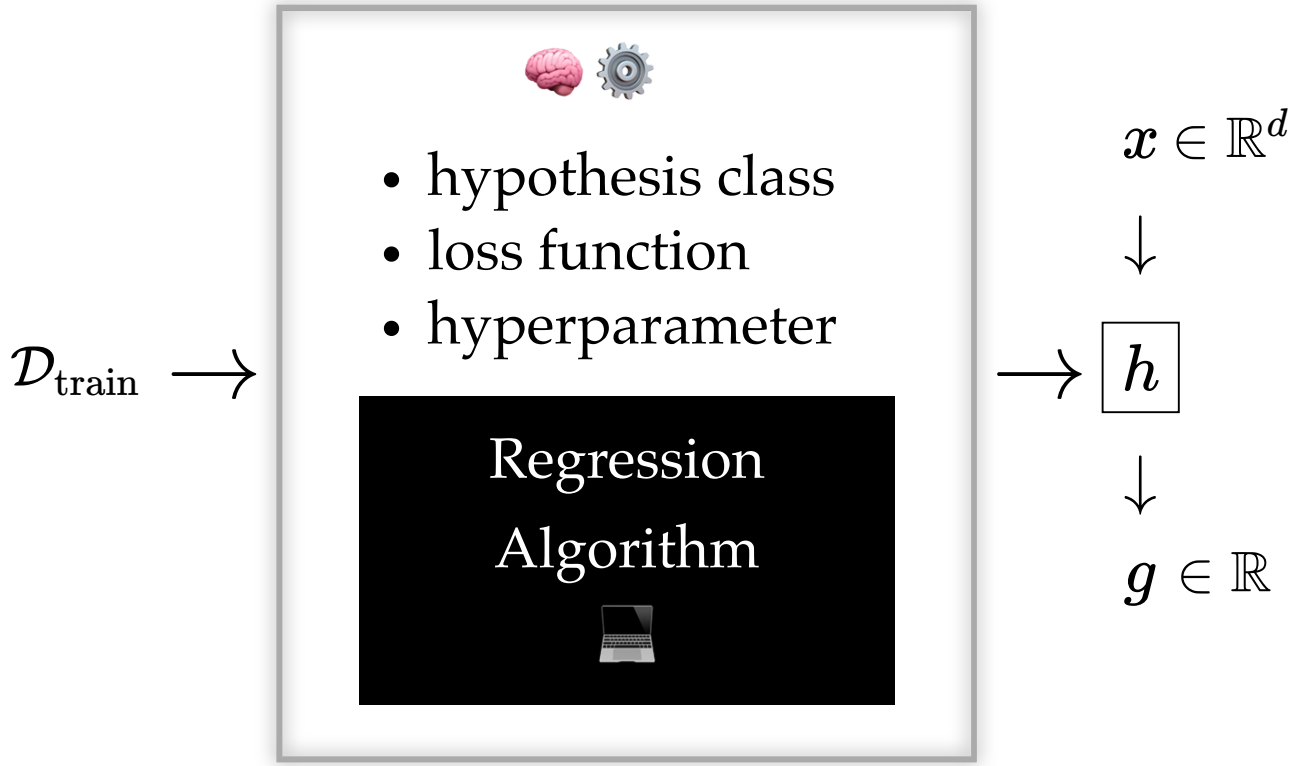
- $\lambda = 0$? — No penalty — reduces to MSE
- $\lambda = 1000$? — Huge penalty — forces $\theta \approx 0$
- $\lambda = -100$? — *Rewards* large θ — counter-productive! [1]

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

1. this is why we require $\lambda > 0$

2. for $\lambda > 0$: θ^*_{ridge} always exists and is unique

λ is a hyperparameter



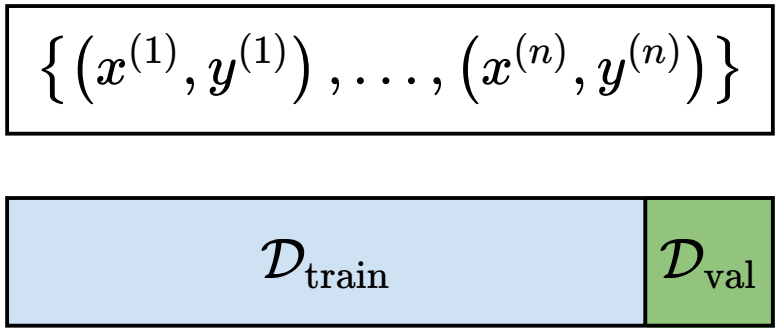
- affects learning outcome, and not learned by minimizing the training error
- we already saw a hyperparameter (the number of random regressor in lab1)

Outline

- The "*trouble*" with the closed-form solution
- Regularization and ridge regression
  - λ, a hyperparameter
- Validation and cross-validation

We need to choose hyperparameters (like λ)

1. Hold-out some data

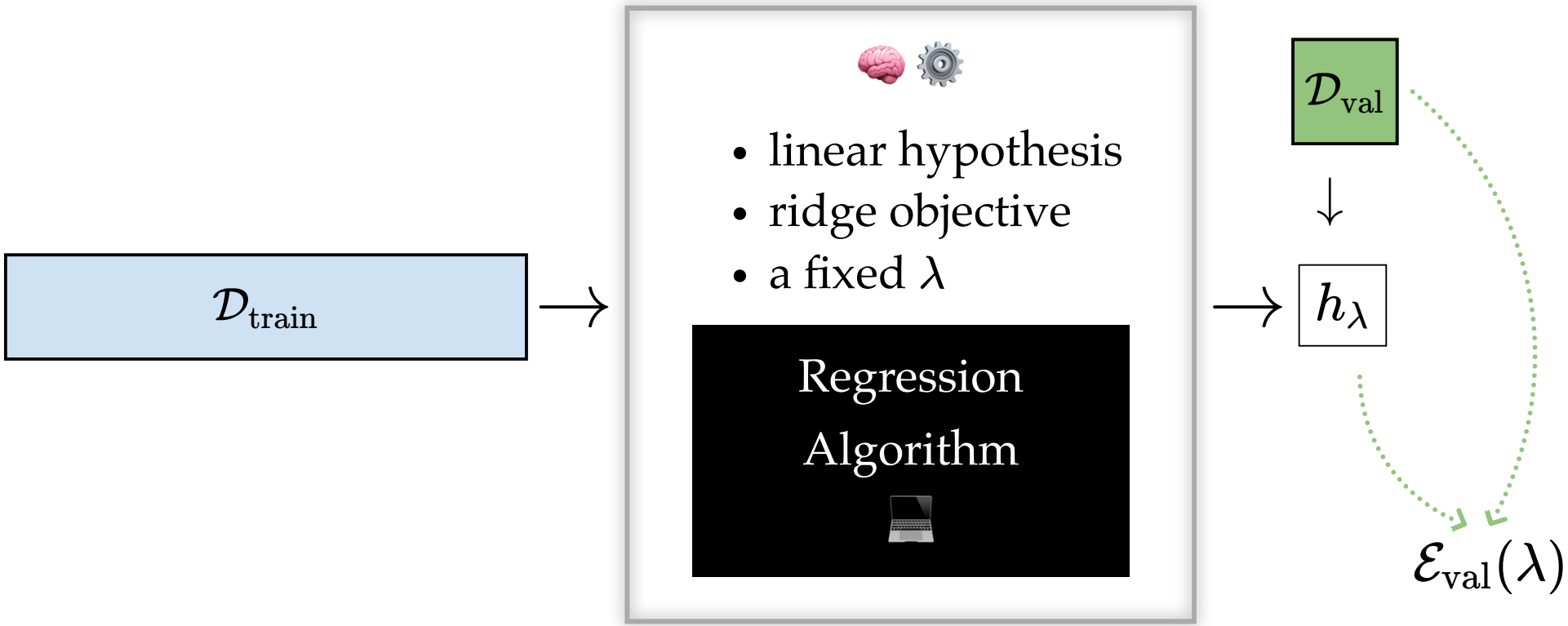


2. For a fixed λ:

    - Use $\mathcal{D}_{\text{train}}$ to train a h_λ
    - Use $\mathcal{D}_{\text{val}}$ on h_λ for that λ's evaluation error

3. Repeat step 2 for a bunch of λ candidates, keep track of their evaluation errors

4. pick the λ candidate that led to the lowest evaluation error



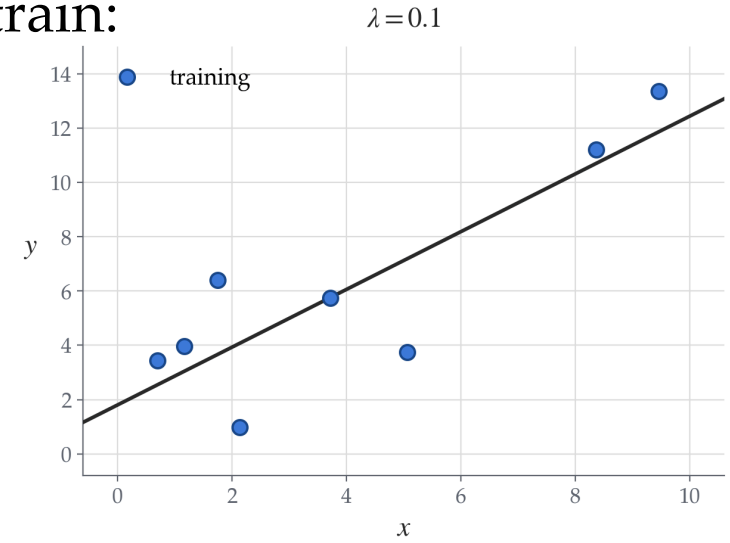
for each λ:

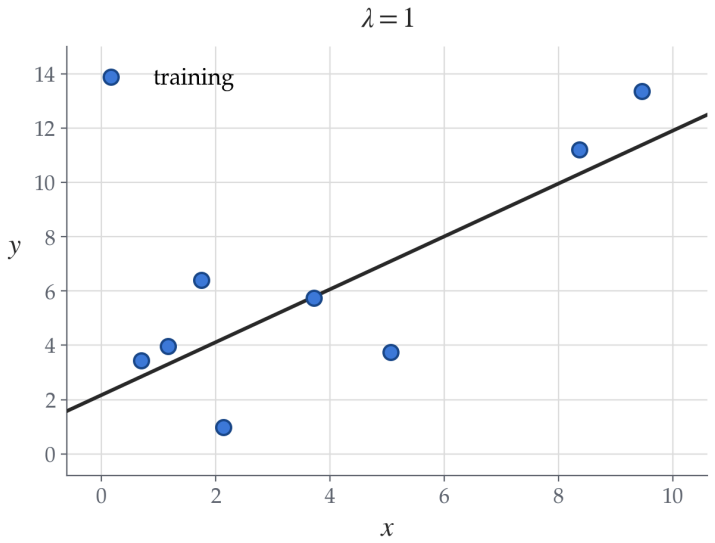
train on $\mathcal{D}_{\text{train}}$ with λ

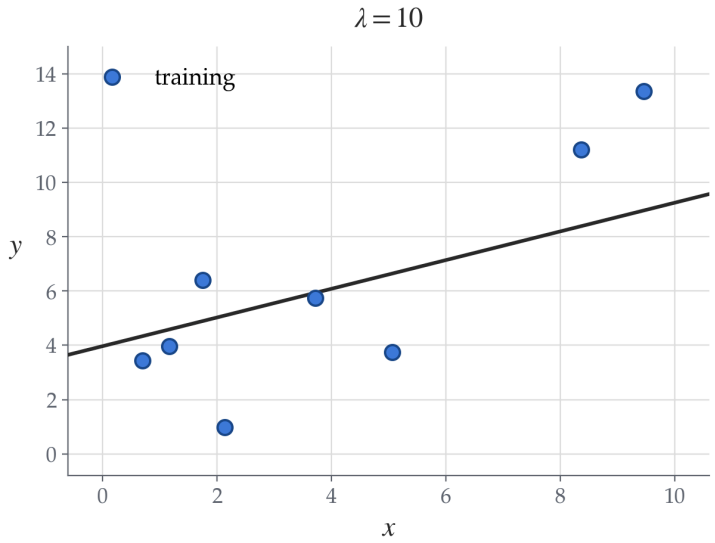
compute $\mathcal{E}_{\text{val}}(\lambda)$ on $\mathcal{D}_{\text{val}}$

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

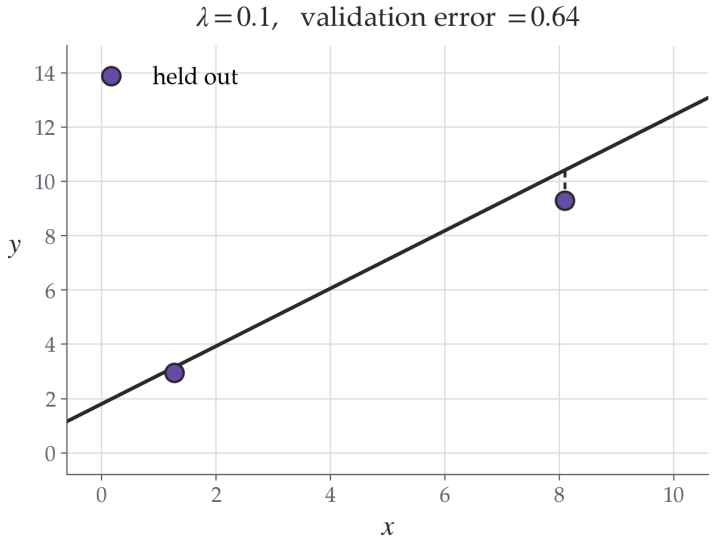
train:

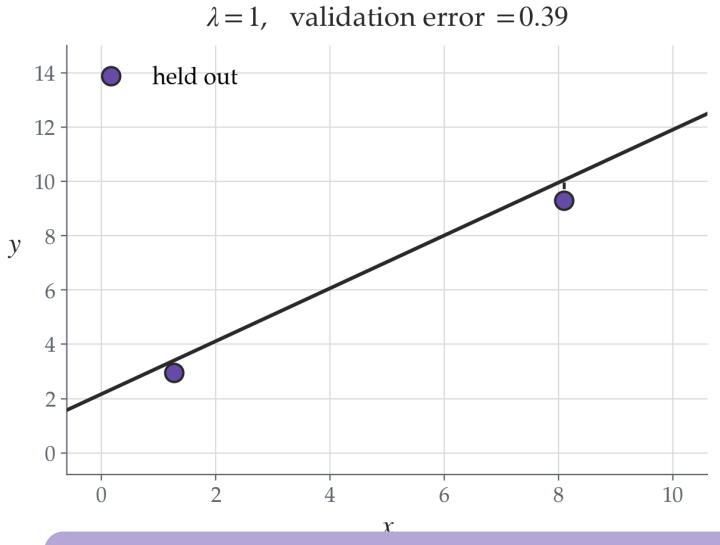


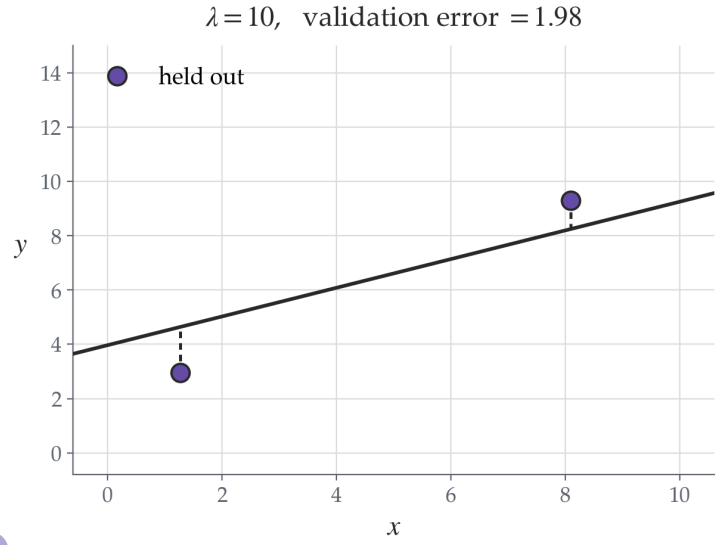




validate:







$\lambda = 1$ "wins" as λ^*,

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

using all 10 data points

K-fold cross-validation

e.g., 5-fold, and 3 λ candidates:

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

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

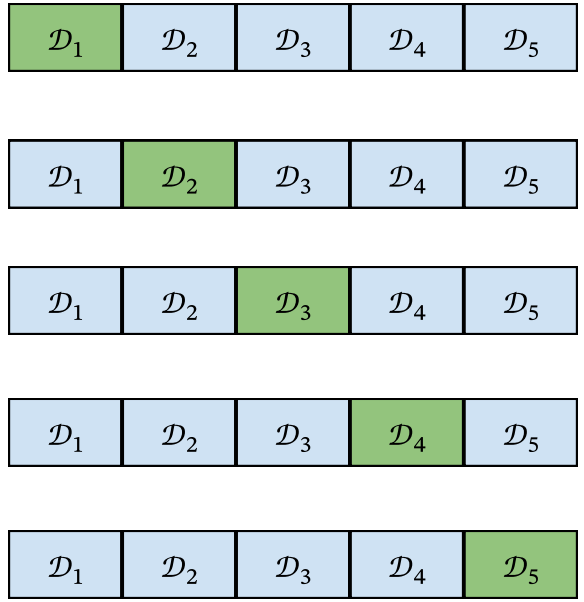
    train h_i on $\mathcal{D} \setminus \mathcal{D}_i$ with λ

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

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

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

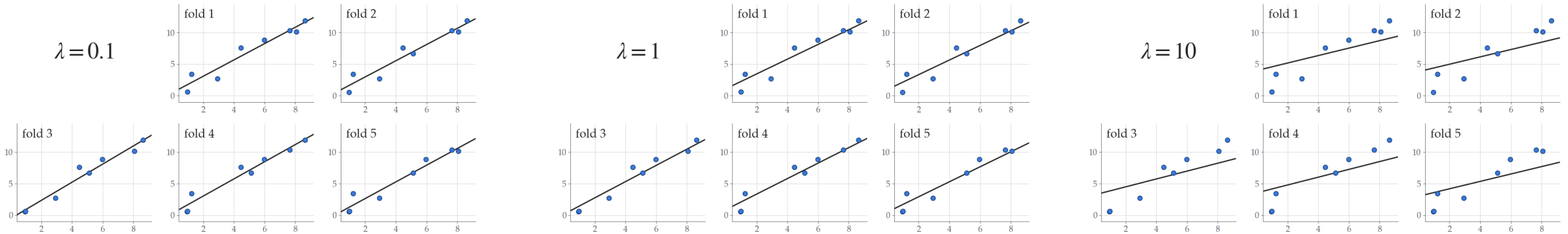
outer loop of $\lambda \in \{0.1, 1, 10\}$:



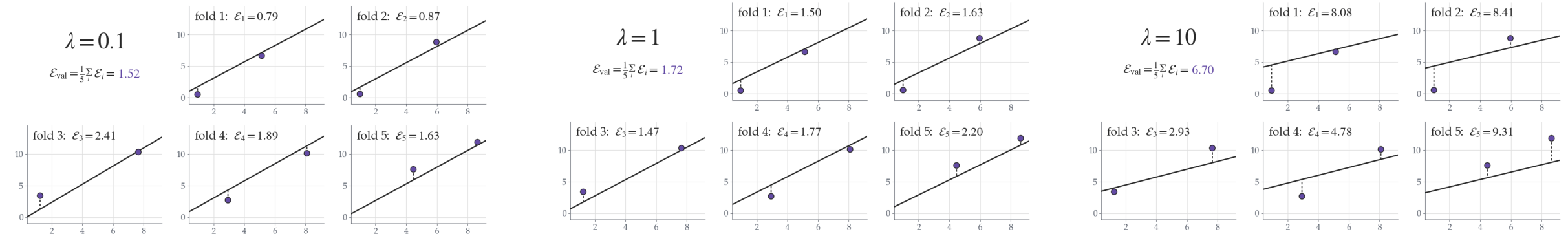
$\mathcal{E}_{\text{val}}(\lambda) = (\mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 + \mathcal{E}_4 + \mathcal{E}_5)/5$

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

train:



validate:



$\lambda = 0.1$ "wins" as λ^*,

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

using all 10 data points

Summary

full column rank | singular | regularization | ridge regression | overfitting

hyperparameter | validation data | cross-validation | K-fold

- When $X^\top X$ is singular, OLS has multiple minimizers.
- When $X^\top X$ is nearly singular (ill-conditioned), fitted parameters and predictions are sensitive to small changes in data.
- 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.