https://introml.mit.edu/

6.390 Intro to Machine Learning

Lecture 2: Regularization and Cross-validation

Shen Shen

Sept 11, 2025

11am, Room 10-250

Interactive Slides and Lecture Recording

Outline

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

Recall

- Linear hypothesis class:

$$h\left(x;\theta\right) \;=\; \theta^T\, x$$

parameters

features

- Squared loss function: $\mathcal{L}\left(h\left(x^{(i)}\right), y^{(i)}\right) = (\underline{\theta^T x^{(i)}} - y^{(i)})^2$

loss

guess (prediction)

label

See lec1/rec1 for discussion of the offset.

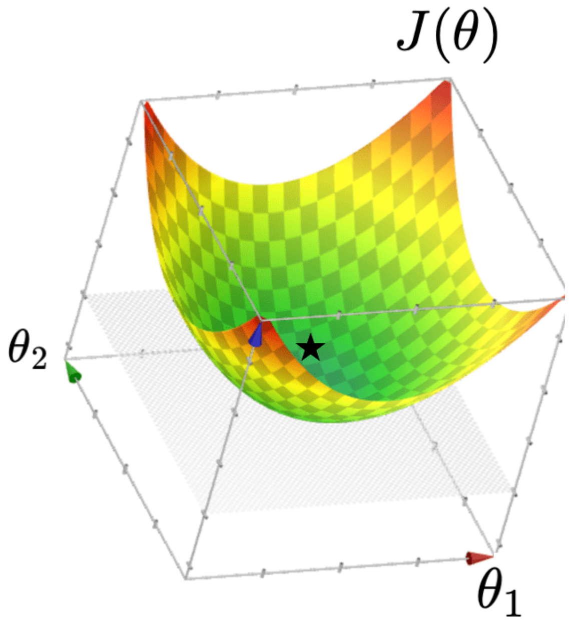
Recall

Let

$$X = \begin{bmatrix} x_1^{(1)} & \dots & x_d^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(n)} & \dots & x_d^{(n)} \end{bmatrix} \in \mathbb{R}^{n\times d} \qquad Y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \in \mathbb{R}^{n\times 1} \qquad \theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} \in \mathbb{R}^{d\times 1}$$

Then

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



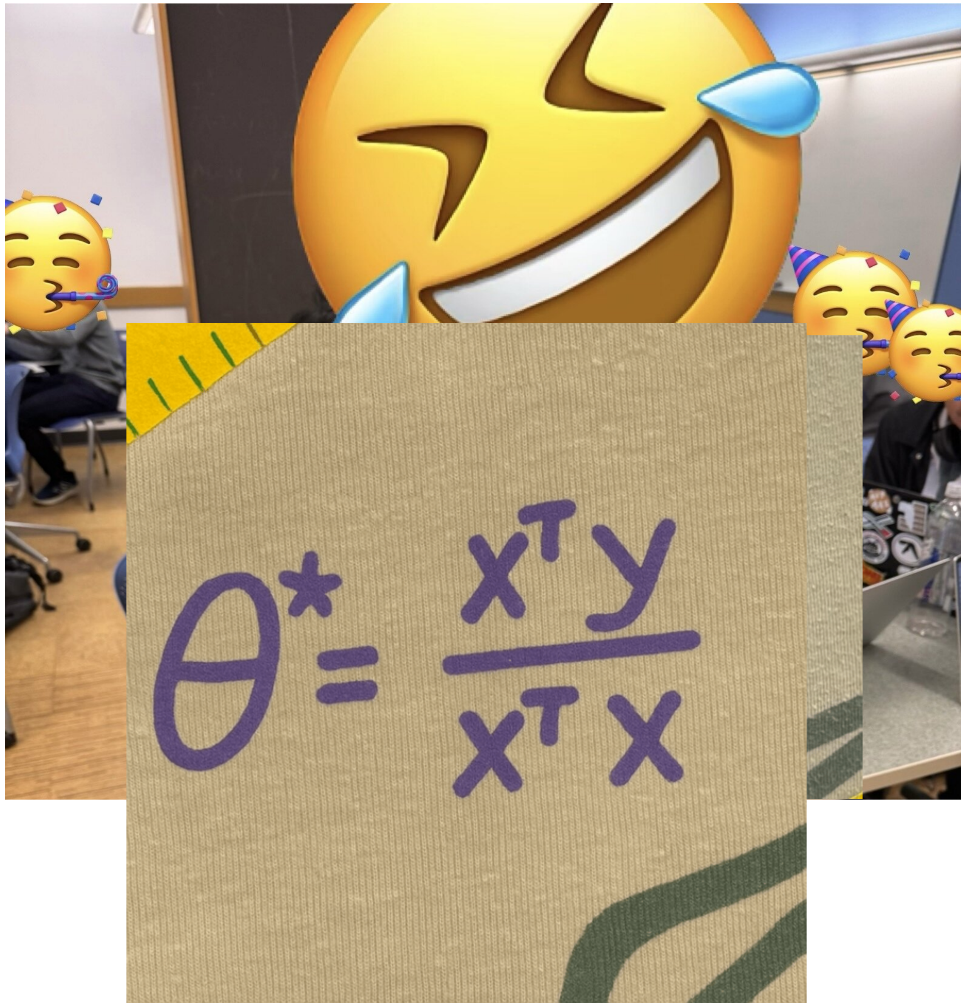
By matrix calculus and optimization

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

$$X^\top X \in \mathbb{R}^{d\times d} \quad X^\top Y \in \mathbb{R}^{d\times 1}$$

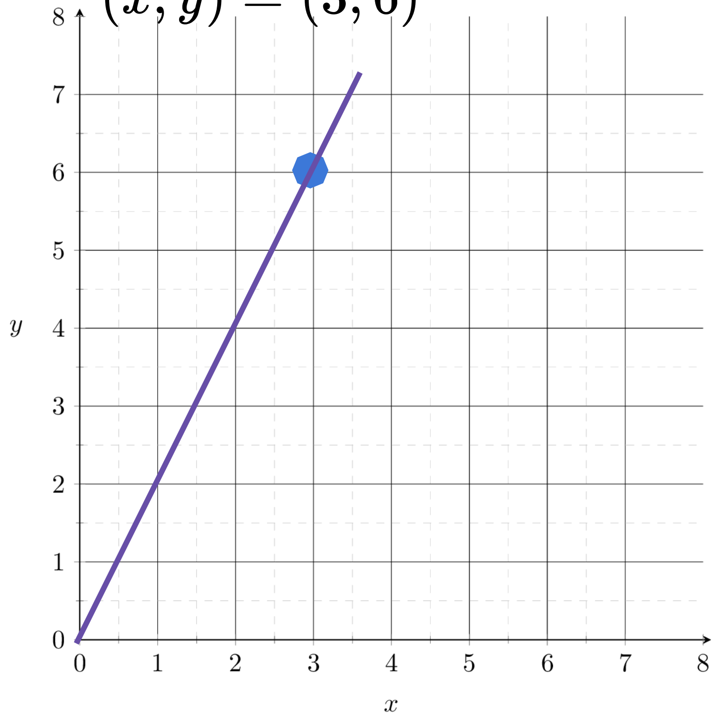
Spotted in lab:

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



1d-feature training data

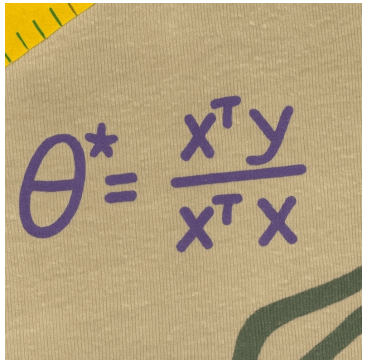
$(x, y) = (3, 6)$



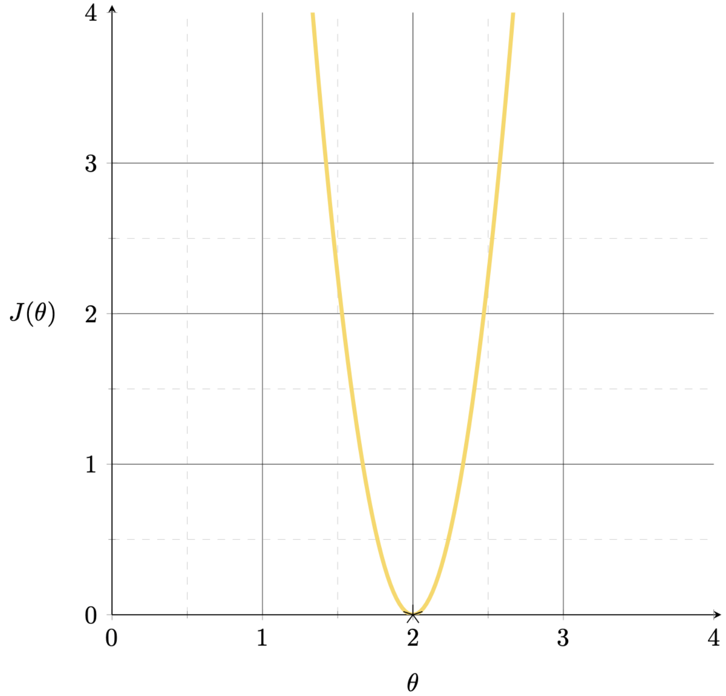
$$\theta^* = \left(X^\top X\right)^{-1} X^\top Y$$

$$X = x = [3]$$

$$Y = y = [6]$$



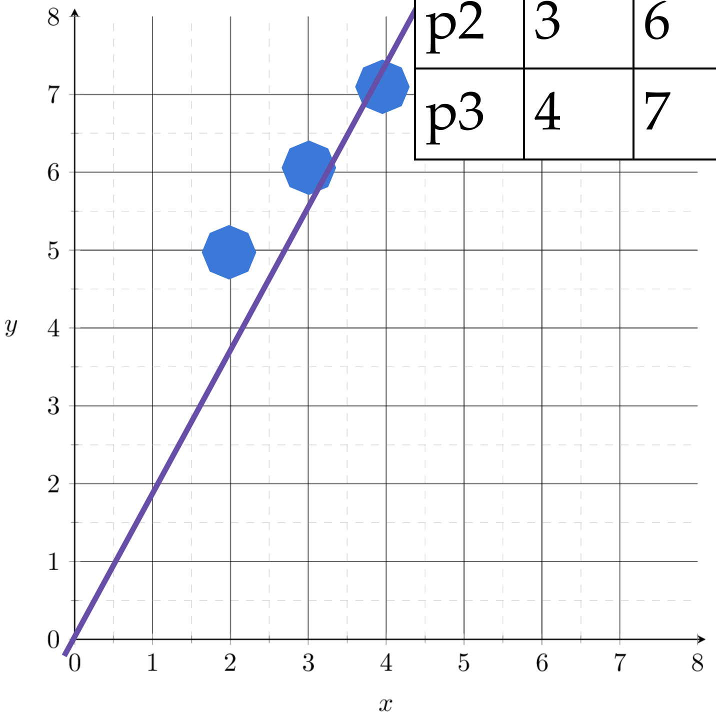
$J(\theta) = (3\theta - 6)^2$



$$\theta^* = (xx)^{-1}(xy) = \frac{xy}{xx} = \frac{y}{x} = \frac{6}{3} = 2$$

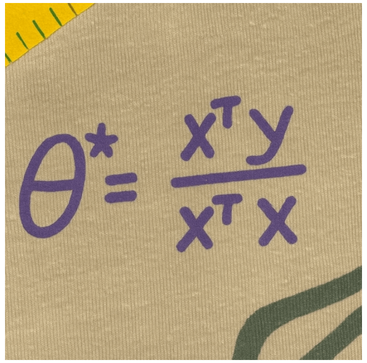
1-d feature
training data set

	x	y
p1	2	5
p2	3	6
p3	4	7

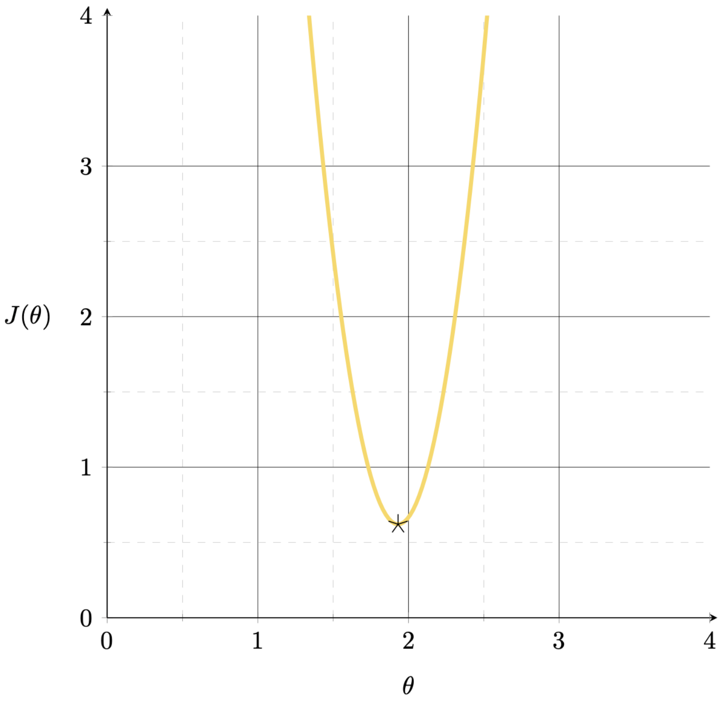


$$\theta^* = \left(X^\top X\right)^{-1} X^\top Y$$

$$X = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \qquad Y = \begin{bmatrix} 5 \\ 6 \\ 7 \end{bmatrix}$$



$$J(\theta) = \frac{1}{3}\left[(2\theta - 5)^2 + (3\theta - 6)^2 + (4\theta - 7)^2\right]$$



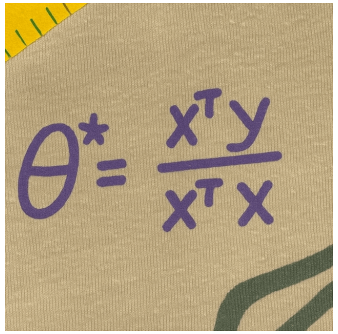
$$\theta^* = \left(\begin{bmatrix} 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}\right)^{-1} \begin{bmatrix} 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \\ 7 \end{bmatrix} = \frac{X^\top Y}{X^\top X} = \frac{56}{29} \approx 1.93$$

Outline

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

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

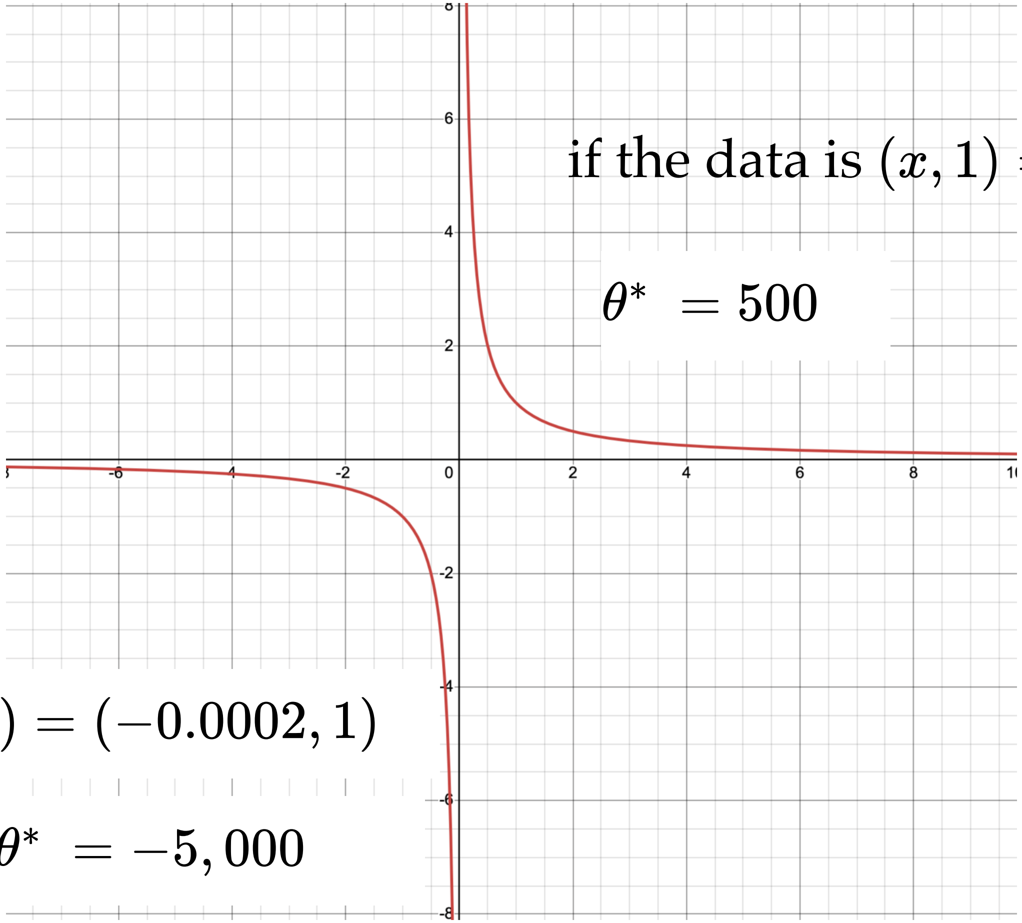
$d = 1$



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

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

most of the time, behaves nicely



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

$\theta^* = 500$

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

$\theta^* = -5,000$

more generally, $d \geq 1$

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

most of the time, behaves nicely

but run into trouble when $\left(X^\top X\right)$ is singular

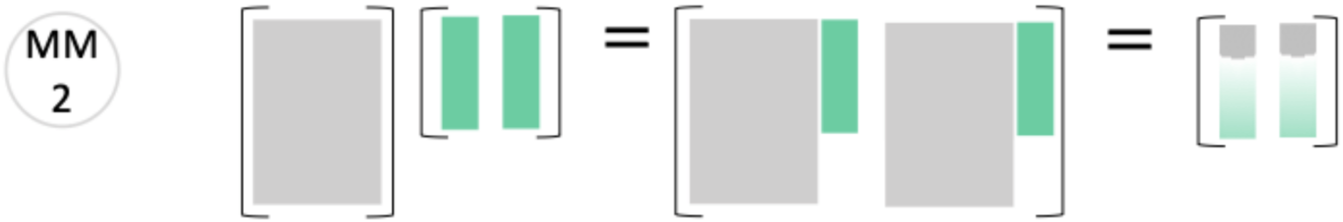
$$\Updownarrow$$

$\left(X^\top X\right)$ has zero eigenvalue(s) $\Leftrightarrow$ the determinant of $\left(X^\top X\right)$ is zero $\Leftrightarrow$ $\left(X^\top X\right)$ is not full rank

$$\Updownarrow$$

X is not full column rank

if X is not full column rank, then $X^\top X$ is singular



$\boldsymbol{Ax}$ and $\boldsymbol{Ay}$ are linear combinations of columns of A.

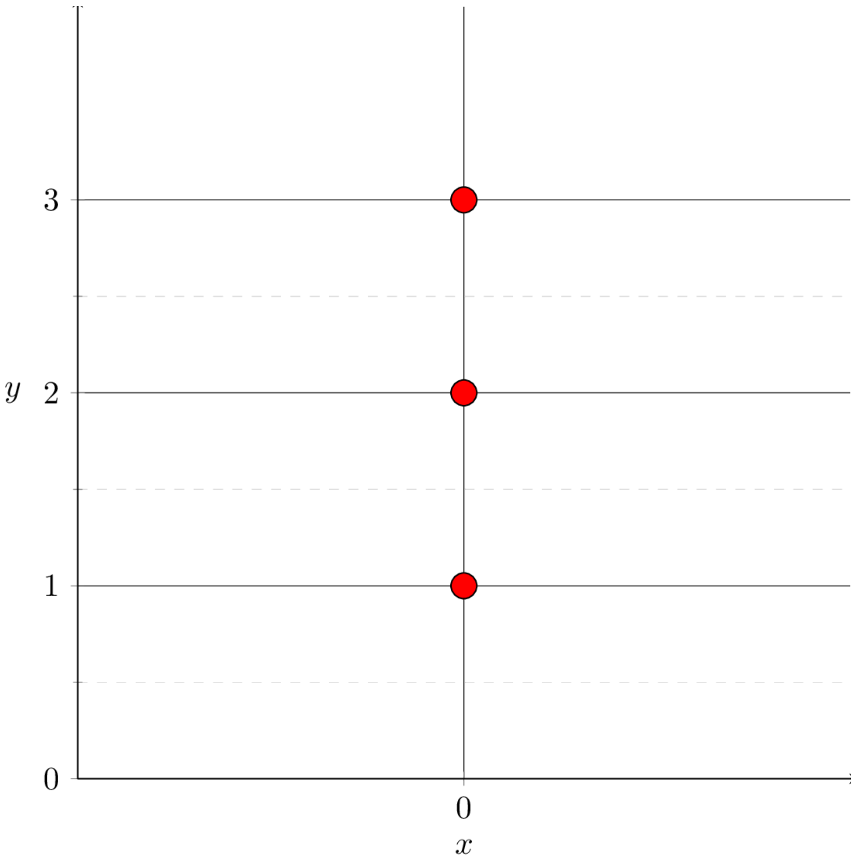
$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix} = A[\boldsymbol{x} \quad \boldsymbol{y}] = [A\boldsymbol{x} \quad A\boldsymbol{y}]$$

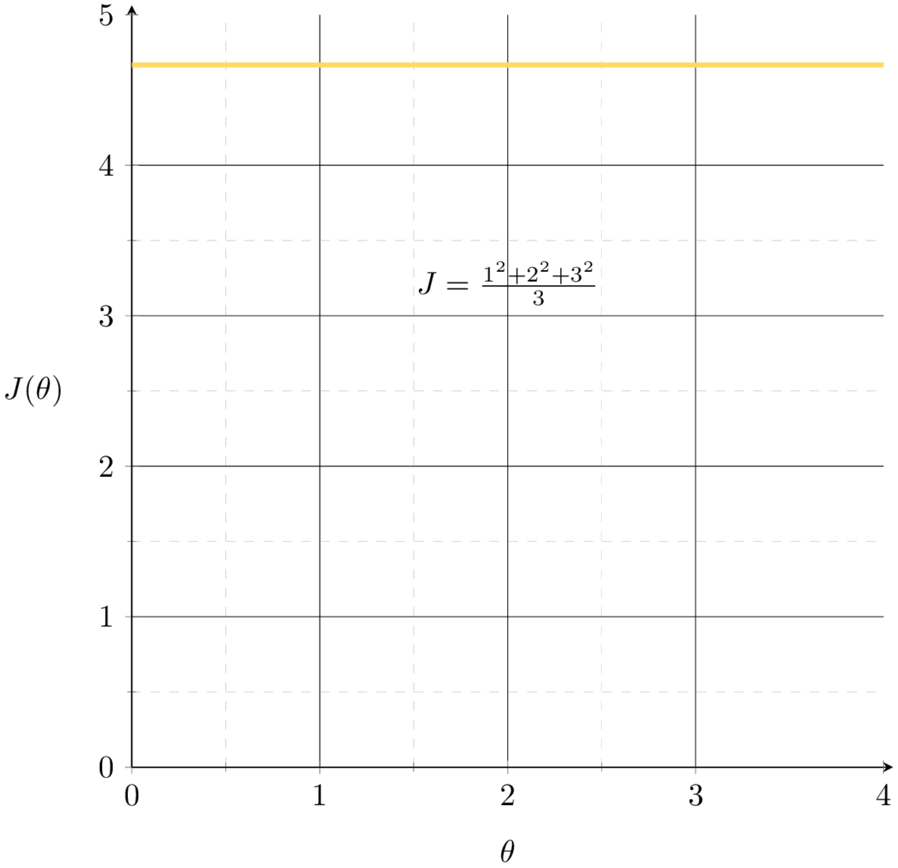
https://github.com/kenjihiranabe/The-Art-of-Linear-Algebra
https://www.3blue1brown.com/topics/linear-algebra

X is not full column rank when:

- a. $d = 1$ and $X \in \mathbb{R}^{n \times 1}$ is simply an all-zero vector, or
- b. $n<d$, or
- c. columns (features) in X are linearly dependent.

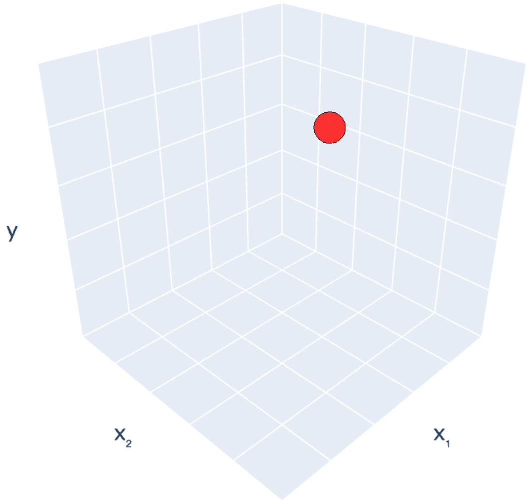
all three cases have similar visual interpretations (a). $d = 1$ and $X \in \mathbb{R}^{n\times 1}$ is simply an all-zero vector





infinitely many optimal θ

(b). $n<d$



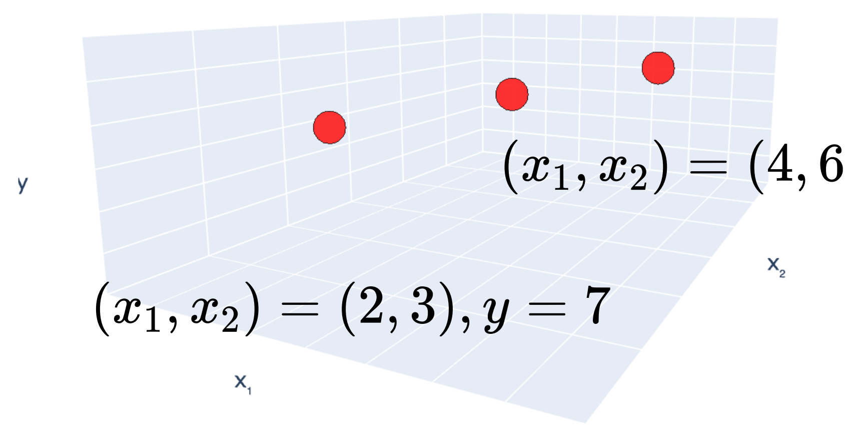
$$(x_1, x_2) = (2, 3), y = 4$$

https://shenshen.mit.edu/demos/ridge/n

infinitely many optimal θ

(c). columns (features) in X are linearly dependent.

$(x_1, x_2) = (6, 9), y = 9$



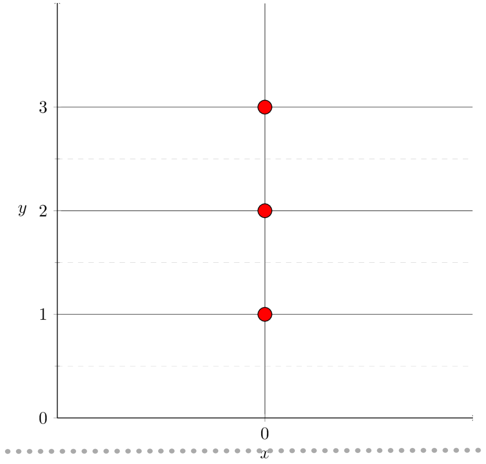
$(x_1, x_2) = (4, 6), y = 8$

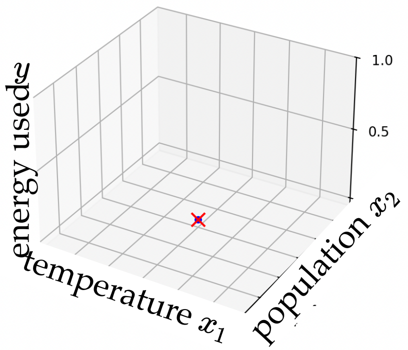
$(x_1, x_2) = (2, 3), y = 7$

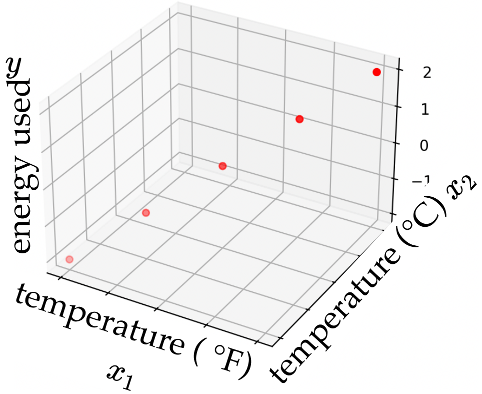
https://shenshen.mit.edu/demos/ridge/colinear_MSE.html

infinitely many optimal θ

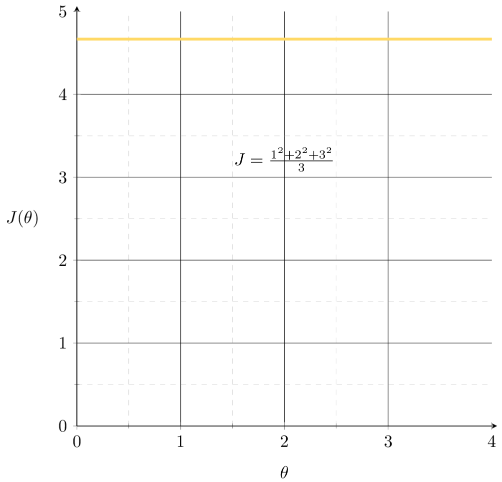
data

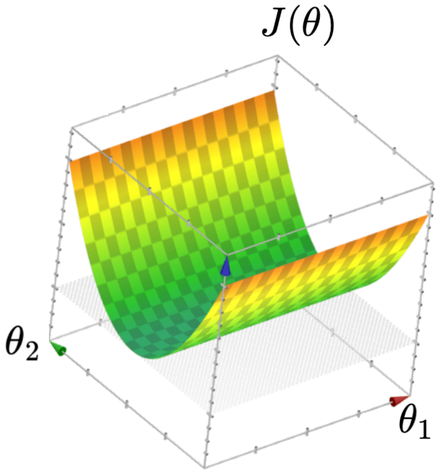






MSE

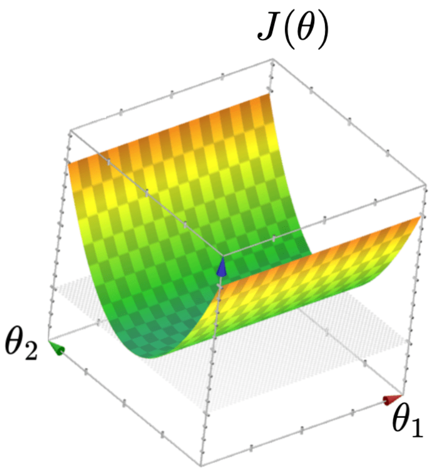




$$\theta^* = \left(X^\top X\right)^{-1} X^\top Y \text{ is not well-defined}$$

infinitely many optimal θ^*

```
Quick Summary:
```

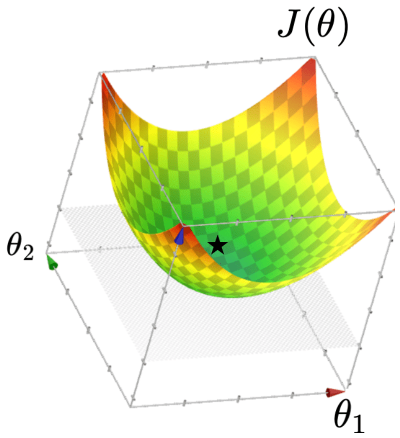


When X is not full column rank

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

🥺

formula isn't wrong, data is trouble-making



Typically, X is full column rank

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

🥰

$X^\top X$ becoming more invertible

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

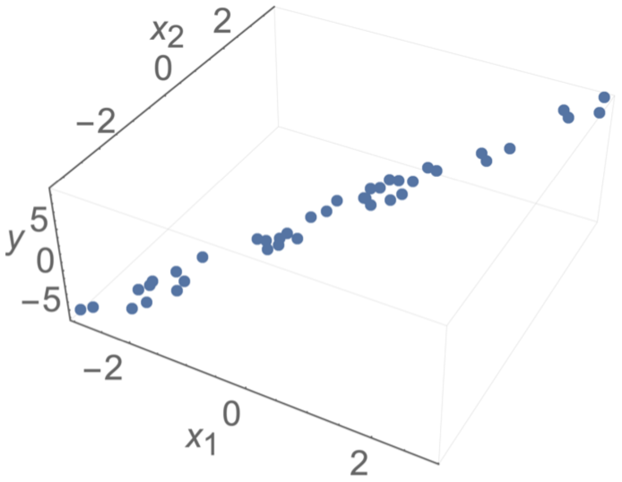
$$\theta^* = \left(X^\top X\right)^{-1} X^\top Y \text{ does exist}$$

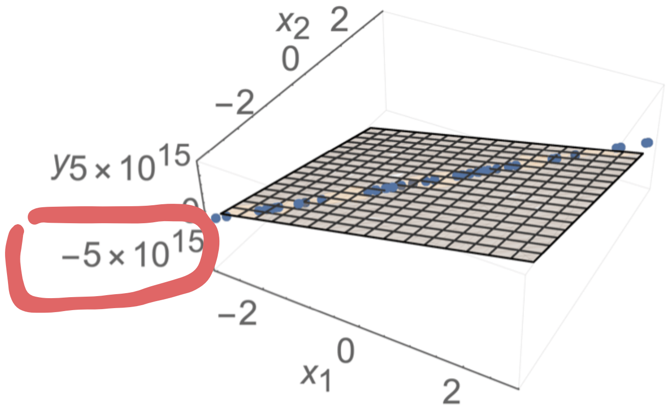
θ^* does give the unique optimal hyperplane

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

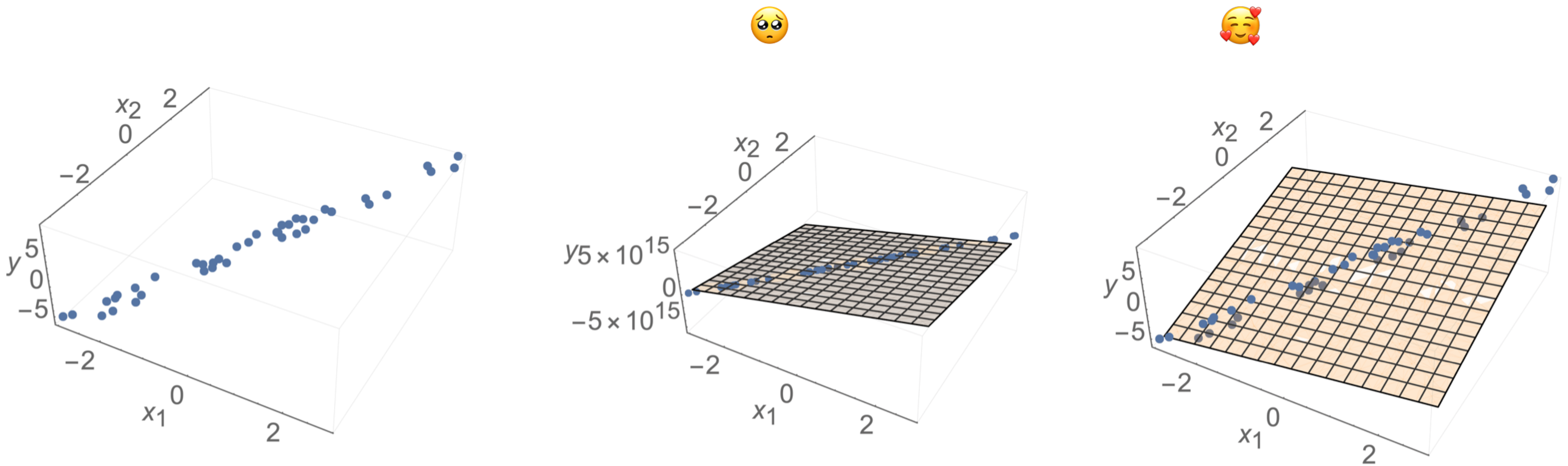
θ^* tends to have huge magnitude

θ^* tends to overfit





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



lots of hypotheses (lots of θs) fit the training data reasonably well

prefer θ with small magnitude (less sensitive prediction when x changes slightly)

Outline

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

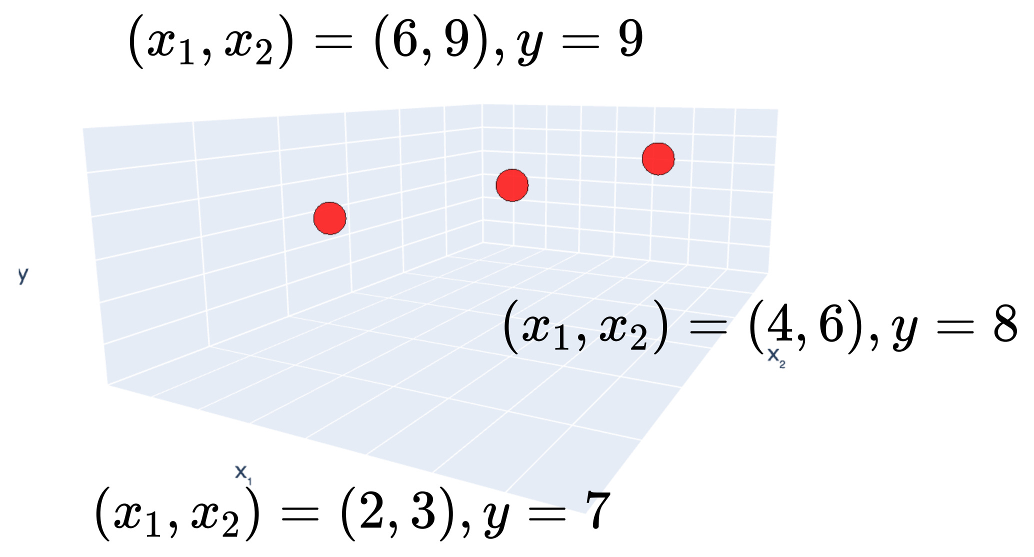
Regularization

- technique to combat overfitting
- at a high-level, it's to sacrifice some training performance, in the hope that testing behaves better
- many ways to regularize (e.g. implicit regularization, drop-out)
- we will look at a particularly simple regularization today, the so-called ridge or $l2$-regularization

Ridge Regression

- Add a square penalty on the magnitude of the parameters
- $J_{\text{ridge}}(\theta) = \frac{1}{n}(X\theta - Y)^\top (X\theta - Y) + \lambda\|\theta\|^2$ $(\lambda > 0)$
- λ is a so-called "hyperparameter" (we've already seen a hyperparameter in lab 1)
- Setting $\nabla_\theta J_{\text{ridge}}(\theta) = 0$ we get $\theta^*_{\text{ridge}} = \left(X^\top X + n\lambda I\right)^{-1} X^\top Y$
- θ^*_{ridge} always exists, and is always the unique optimal parameters.
- (see ex/lab/hw for discussion about the offset.)

case (c) training data set again



https://shenshen.mit.e

Comments on λ

- one that's chosen by users, before we even see the data
- controls the tradeoff between MSE and theta magnitude
- implicitly controls the "richness" of the hypothesis class

$\mathcal{D}_{\text{train}} \longrightarrow$

$\left\{\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right\}$

🧠

- hypothesis class
- loss function
- hyperparameter

Regression Algorithm

💻

$\longrightarrow$

$x \in \mathbb{R}^d$

$\downarrow$

$\boxed{h}$ hypothesis

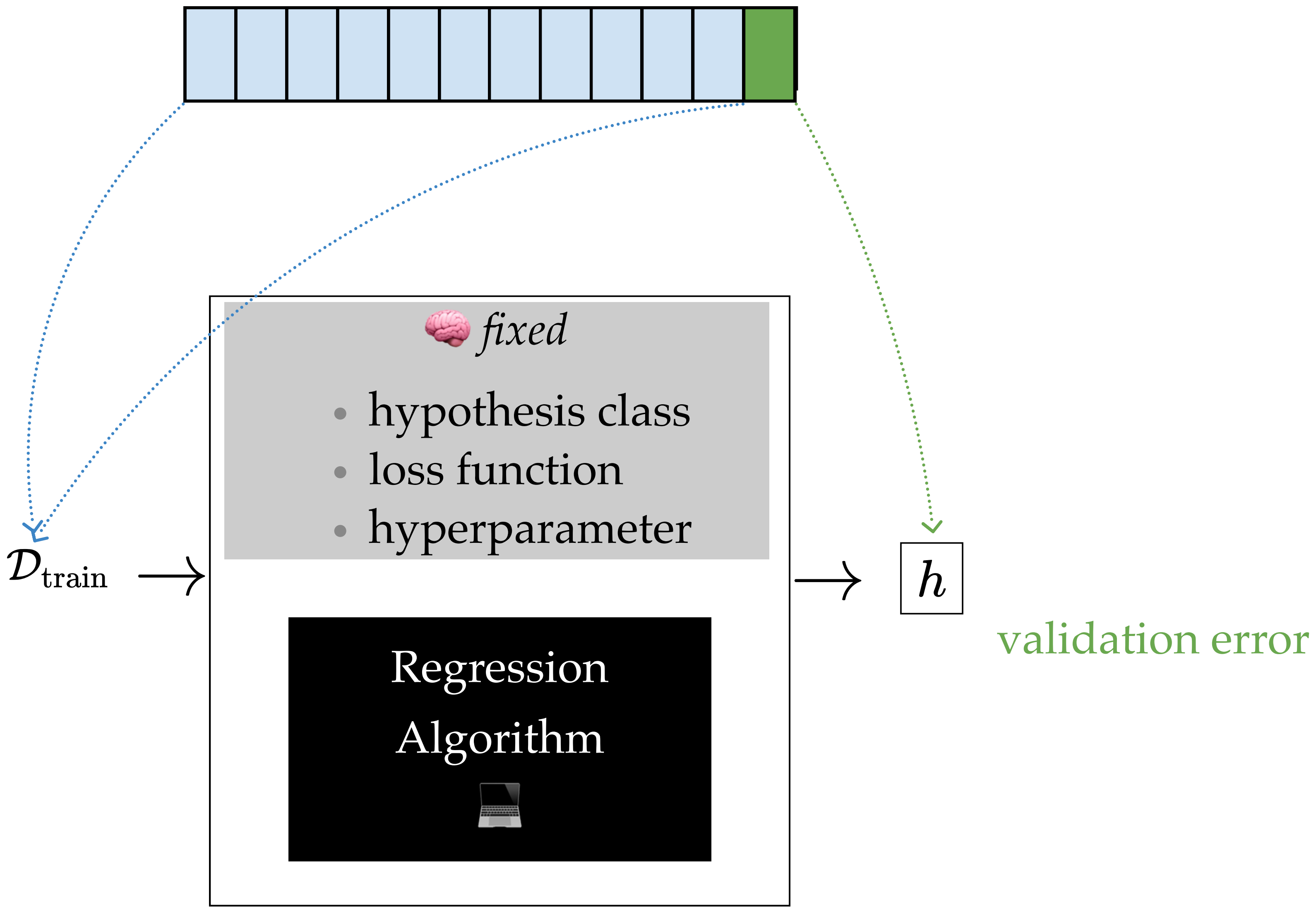
$\downarrow$

$y \in \mathbb{R}$

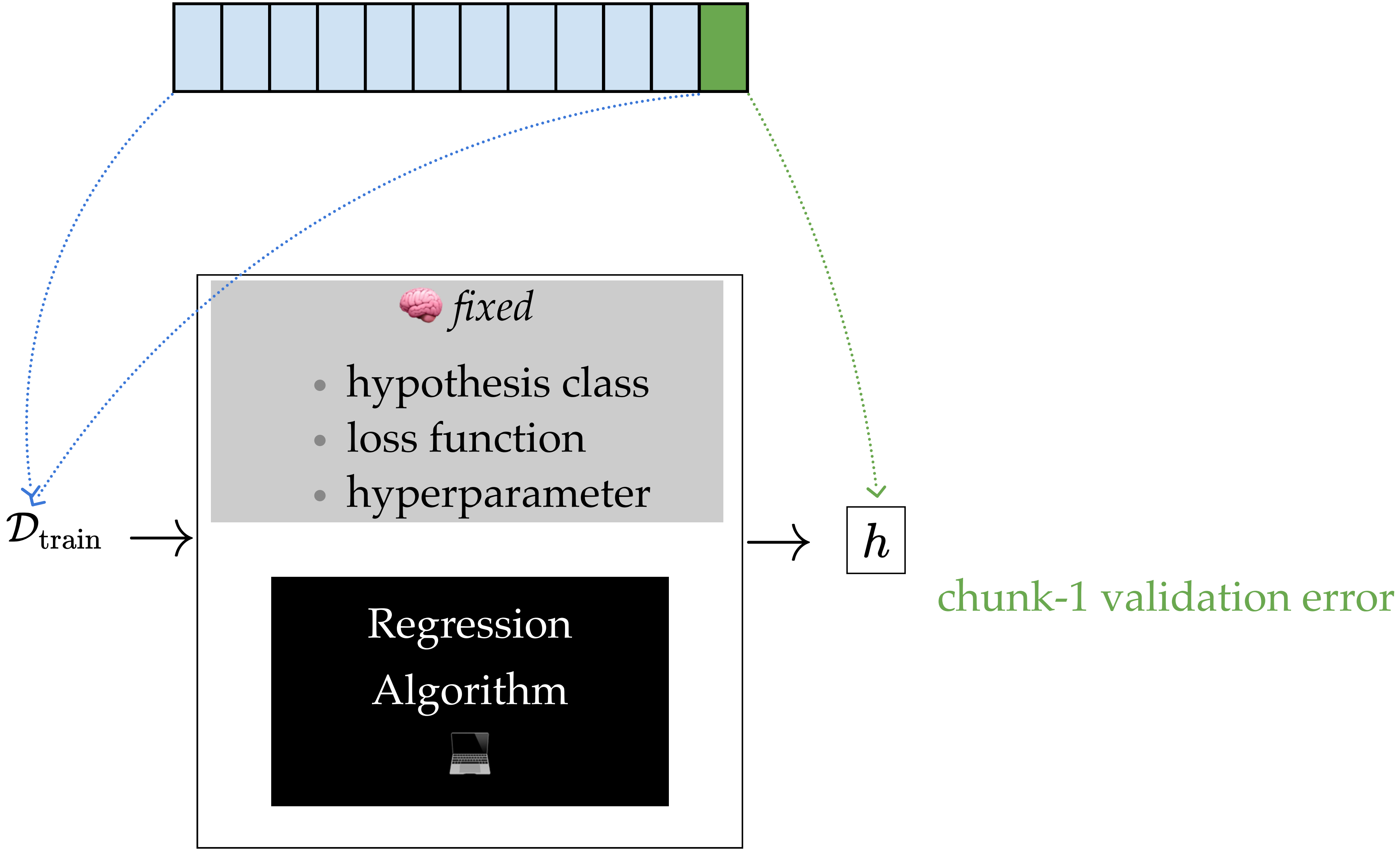
Outline

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

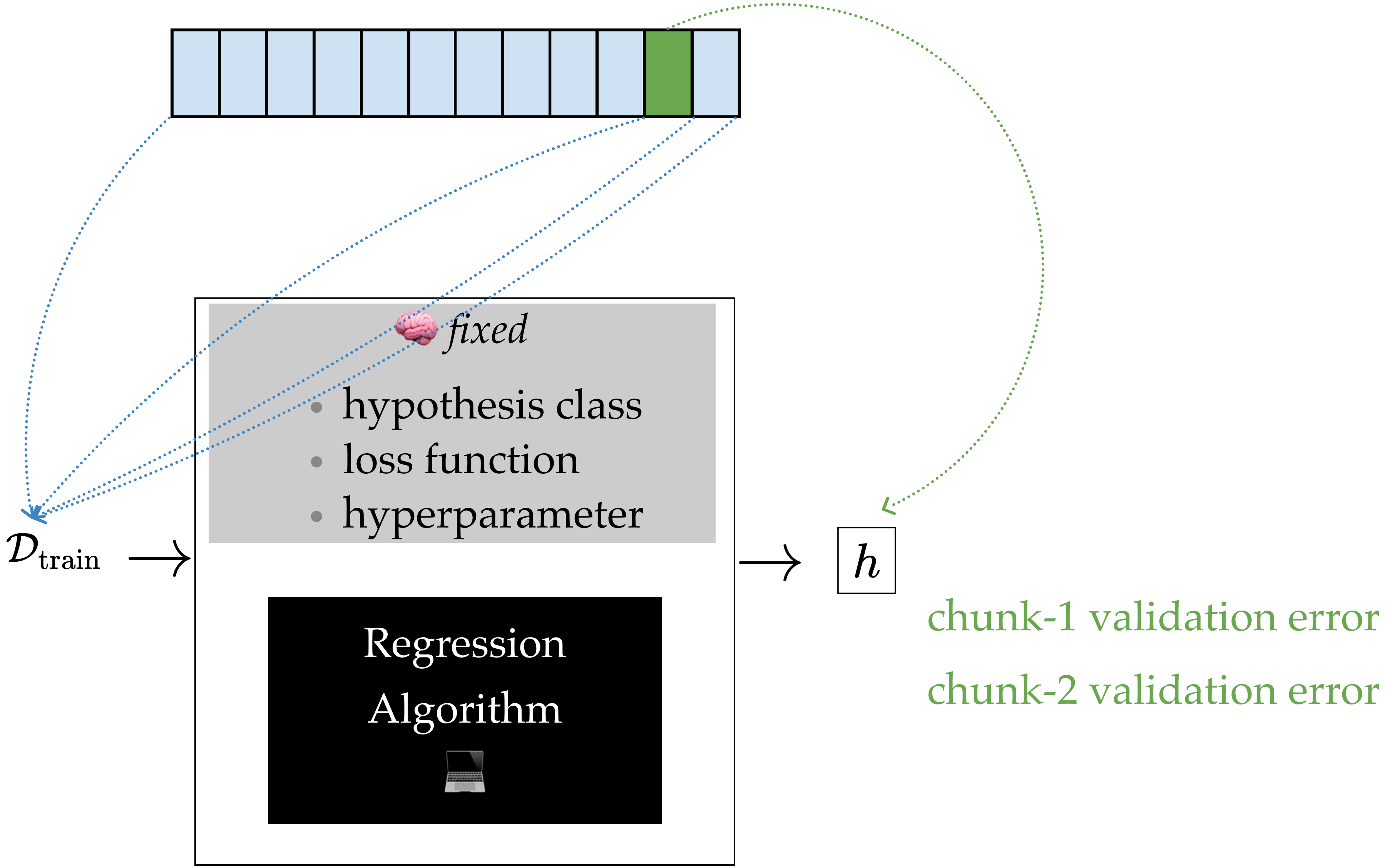
Validation


fixed
hypothesis class
loss function
hyperparameter
$\mathcal{D}_{\text{train}}$
h
validation error
Regression
Algorithm

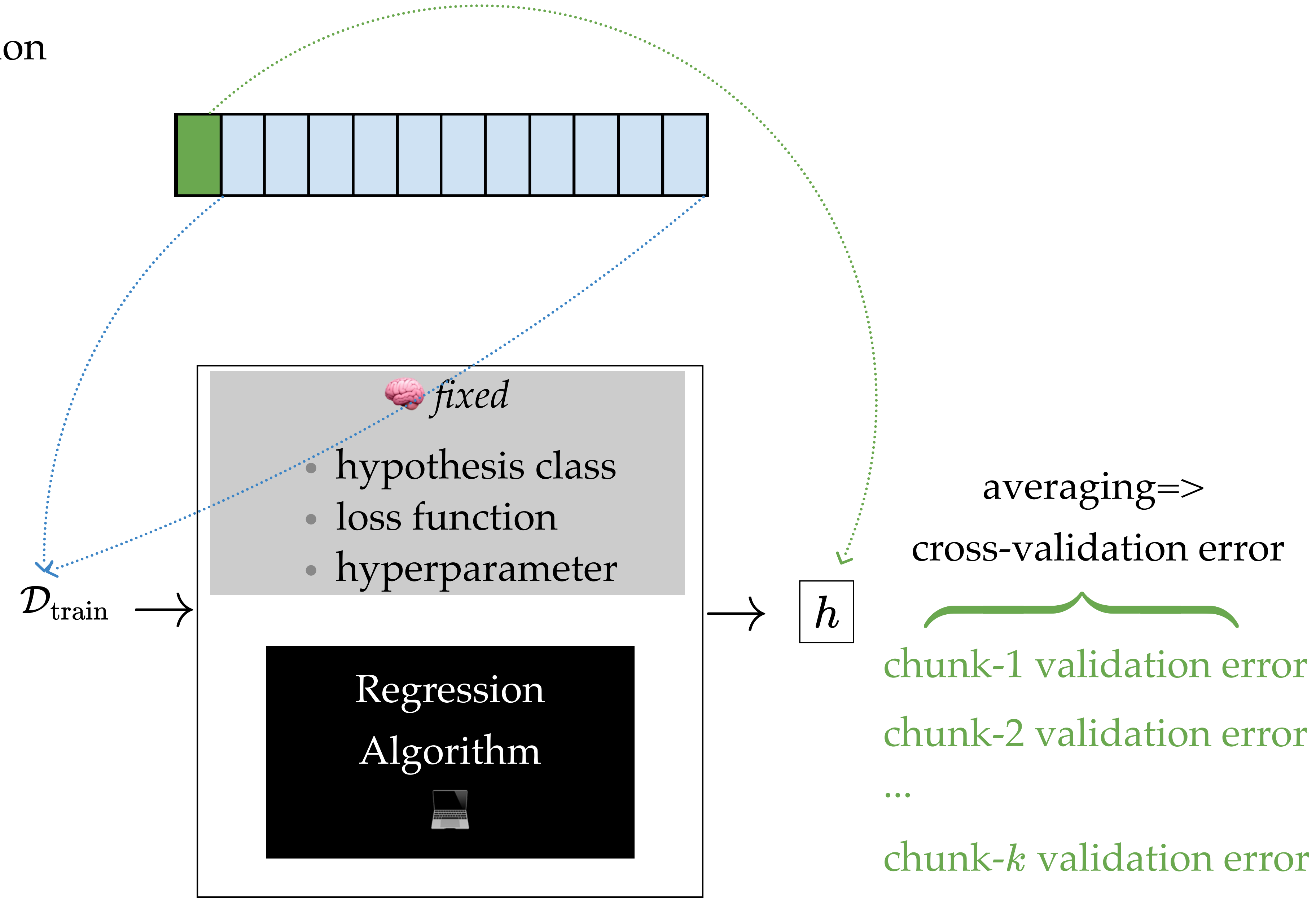
Cross-validation



Cross-validation



Cross-validation



Comments on cross-validation

- good idea to shuffle data first
- a way to "reuse" data
- cross-validation is more "reliable" than validation (less sensitive to chance)
- it's not to evaluate a hypothesis (testing error is)
- rather, it's to *evaluate* learning algorithm (e.g. hypothesis class choice, hyperparameter choice)
- Can have an outer loop for *picking* good hyperparameter or hypothesis class

Summary

- Closed-form formula for OLS is not well-defined when $X^T X$ is singular, and we have infinitely many optimal θ^*.
- Even in scenarios where $X^T X$ is just ill-conditioned, we get sensitivity issues, many almost-as-good solutions, while the absolutely best θ^* is overfitting to the data.
- We need to indicate our preference somehow, and also fight overfitting.
- Regularization helps battle overfitting -- by constructing a new optimization problem that implicitly prefers small-magnitude θ.
- Least-squares regularization leads to the ridge-regression formulation. (Good news: we can still solve it analytically!)
- λ trades off training MSE and regularization strength, it's a hyperparameter.
- Validation/cross-validation are a way to choose (regularization) hyperparameters.

https://docs.google.com/forms/d/e/1FAIpQLSftMB5hSccgAbIAFmP_LuZt95w6KFx0x_R3uuzBP8WwjSzZeQ/viewform?embedded=true

We'd love to hear your thoughts.

Thanks!