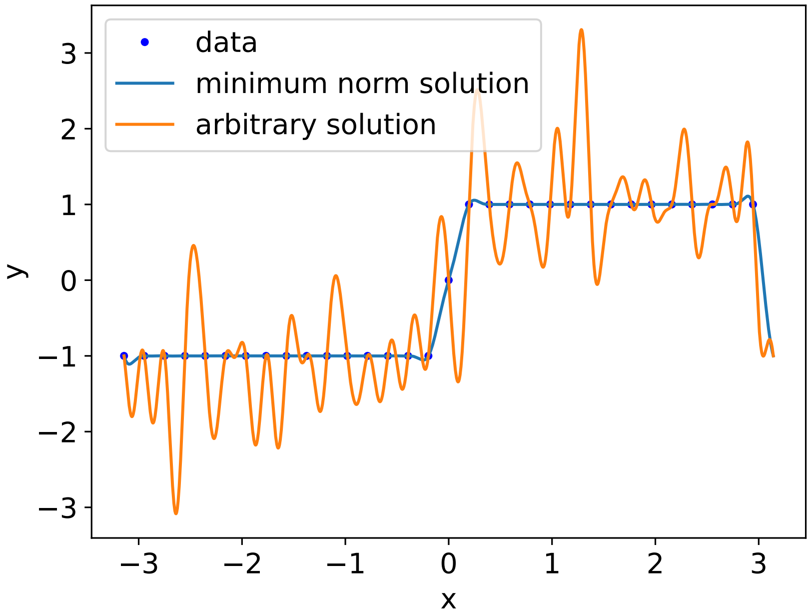
# Regularization

COMS 4771 Fall 2025

## Minimum norm solutions

Normal equations $(A^\mathsf{T} A)w = A^\mathsf{T} b$ can have infinitely-many solutions

$$\varphi(x) = \left(1, \cos(x), \sin(x), \frac{\cos(2x)}{2}, \frac{\sin(2x)}{2}, \ldots, \frac{\cos(32x)}{32}, \frac{\sin(32x)}{32}\right)$$






Norm of $w$ is a measure of "steepness"

$$\underbrace{|w^\mathsf{T}\varphi(x) - w^\mathsf{T}\varphi(x')|}_{\text{change in output}} \leq \|w\| \times \underbrace{\|\varphi(x) - \varphi(x')\|}_{\text{change in input}}$$

(Cauchy-Schwarz inequality)

- Note: Data does not provide a reason to prefer short $w$ over long $w$
- Preference for short $w$ is example of inductive bias (i.e., a preference for one solution over another)

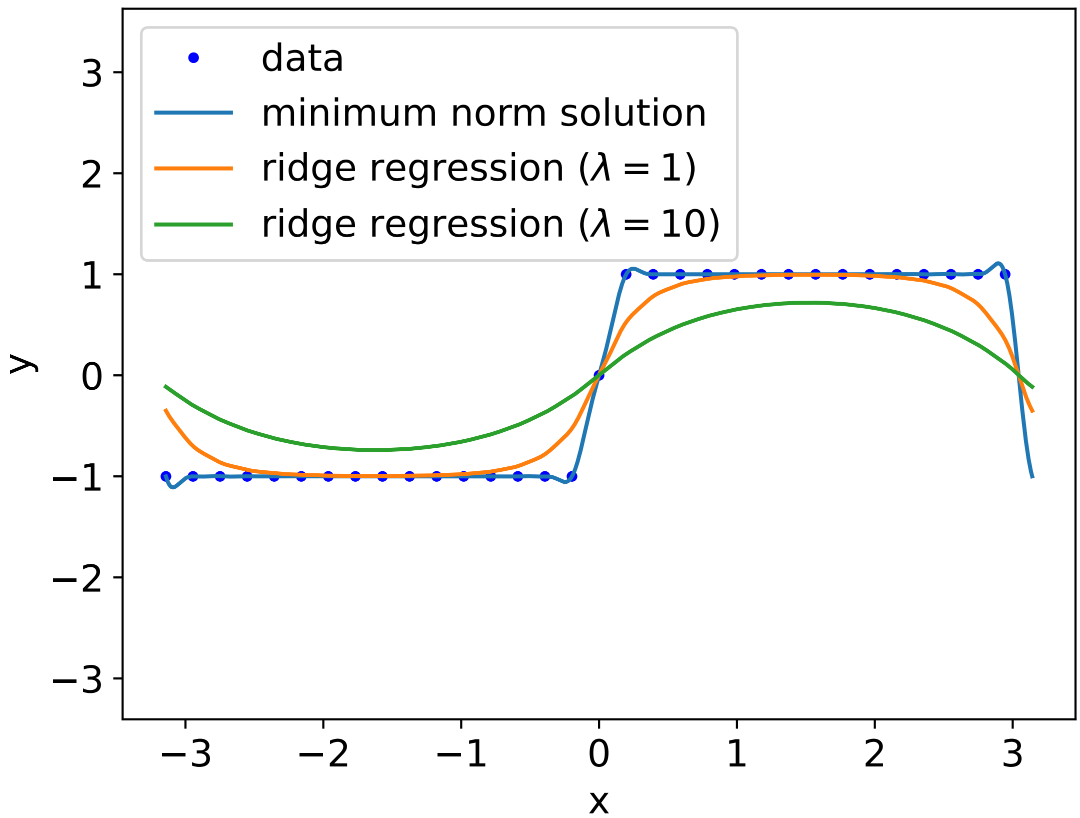
# Ridge regression

Ridge regression: "balance" two concerns by minimizing

$$\|Aw - b\|^2 + \lambda \|w\|^2$$

where $\lambda \geq 0$ is hyperparameter

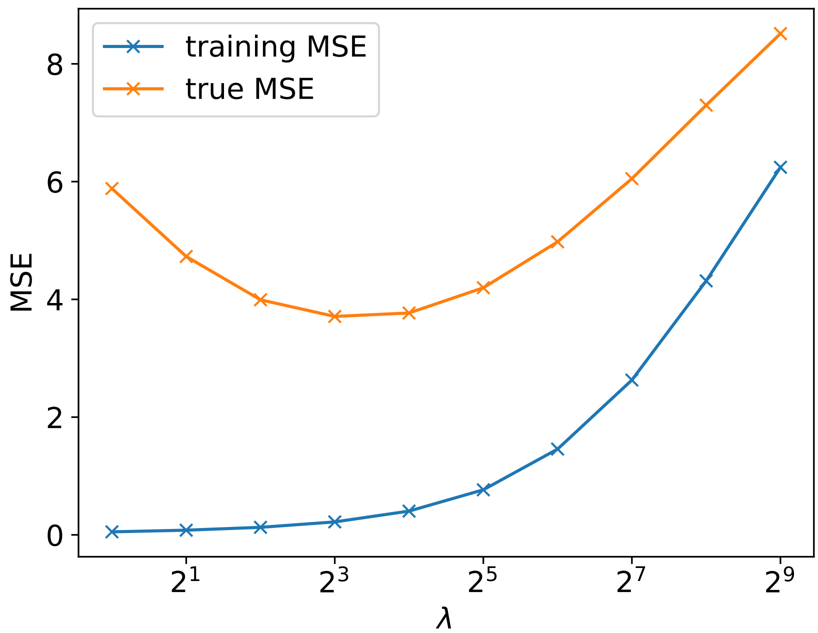
- Concern #1: "data fitting term" $\|Aw - b\|^2$ (involves training data)
- Concern #2: regularizer $\lambda \|w\|^2$ (doesn't involve training data)
- $\lambda = 0$: objective in OLS, might have multiple minimizers
- $\lambda \to 0^+$: minimum norm solution






Example: $n = d = 100$, $((X^{(i)}, Y^{(i)}))_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} (X, Y)$, where $X \sim \mathrm{N}(0, I)$, and conditional distribution of $Y$ given $X = x$ is $\mathrm{N}(\sum_{j=1}^{10} x_j, 1)$

- Normal equations have unique solution, but OLS performs poorly



## Different interpretation of ridge regression objective

$$\begin{aligned} &\|Aw - b\|^2 + \lambda\|w\|^2 \\ &= \|Aw - b\|^2 + \|(\sqrt{\lambda}I)w - 0\|^2 \end{aligned}$$

- Second term is MSE on $d$ additional "fake examples"

$$\begin{aligned} (x^{(n+1)}, y^{(n+1)}) &= \underline{\hspace{4cm}} \\ (x^{(n+2)}, y^{(n+2)}) &= \underline{\hspace{4cm}} \\ &\vdots \\ (x^{(n+d)}, y^{(n+d)}) &= \underline{\hspace{4cm}} \end{aligned}$$



"Augmented" dataset in matrix notation:

$$\widetilde{A} = \begin{bmatrix} \longleftarrow & (x^{(1)})^\mathsf{T} & \longrightarrow \\ & \vdots & \\ \longleftarrow & (x^{(n)})^\mathsf{T} & \longrightarrow \\ \longleftarrow & (x^{(n+1)})^\mathsf{T} & \longrightarrow \\ & \vdots & \\ \longleftarrow & (x^{(n+d)})^\mathsf{T} & \longrightarrow \end{bmatrix}, \quad \tilde{b} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

so

$$\|Aw - b\|^2 + \lambda\|w\|^2 = \|\widetilde{A}w - \tilde{b}\|^2$$

What are "normal equations" for ridge regression objective (in terms of $A$, $b$, $\lambda$)?

# Other forms of regularization

Regularization using **domain-specific data augmentation**

Create "fake examples" from existing data by applying transformations that do not change appropriateness of corresponding label, e.g.,

- Image data: rotations, rescaling
- Audio data: change playback rate
- Text data: replace words with synonyms

Functional penalties (e.g., norm on $w$)

- Ridge: (squared) $\ell^2$ norm

$$\|w\|^2$$

- Lasso: $\ell^1$ norm

$$\|w\|_1 = \sum_{j=1}^{d} |w_j|$$

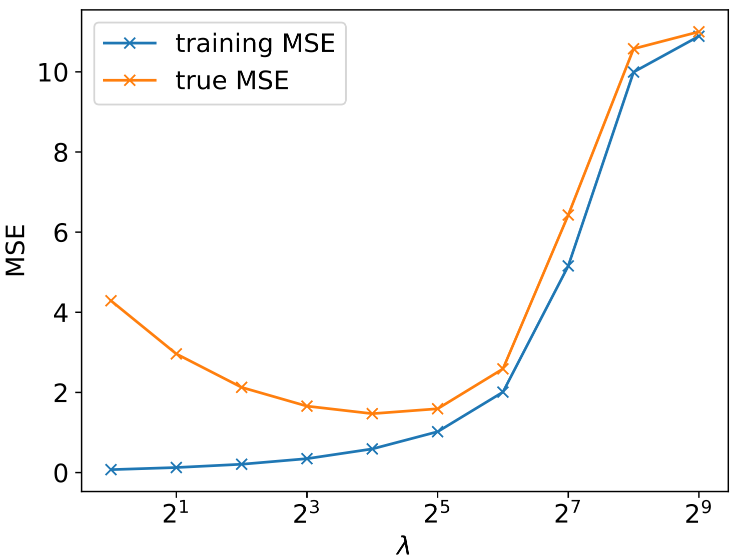
- Sparse regularization: $\ell^0$ "norm" (not really a norm)

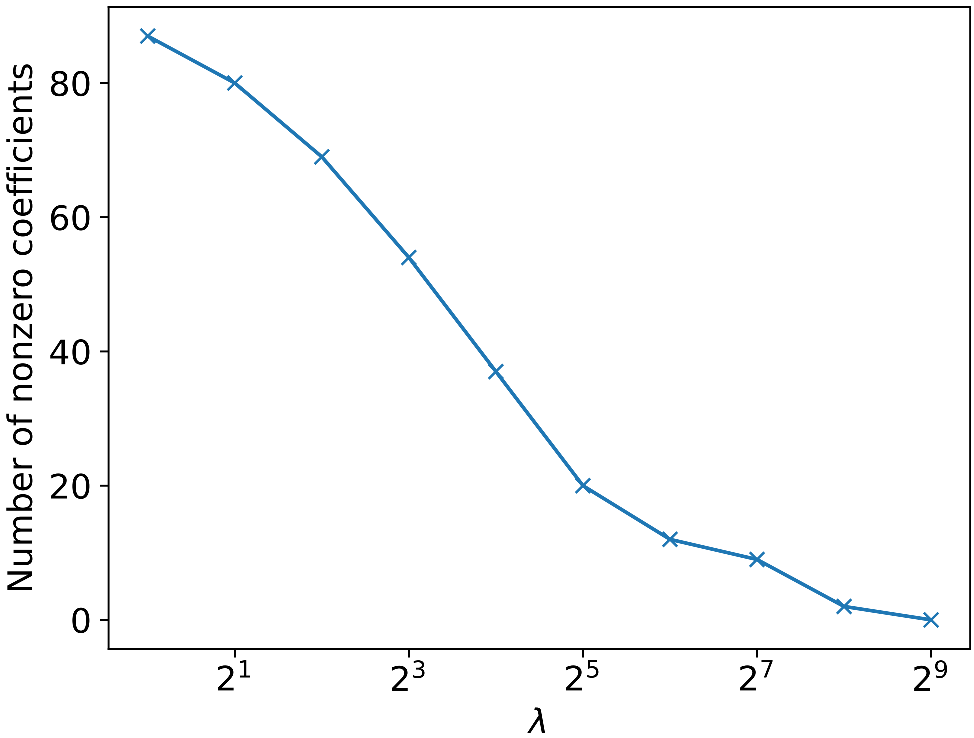
$$\|w\|_0 = \#\text{ coefficients in } w \text{ that are non-zero}$$



Example: $n = d = 100$, $((X^{(i)}, Y^{(i)}))_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} (X, Y)$, where $X \sim \mathrm{N}(0, I)$, and conditional distribution of $Y$ given $X = x$ is $\mathrm{N}(\sum_{j=1}^{10} x_j, 1)$

- Minimize $\|Aw - b\|^2 + \lambda\|w\|_1$ (Lasso)








Weighted (squared) $\ell^2$ norm:

$$\sum_{i=1}^{d} c_i\, w_i^2$$

for some "costs" $c_1, \ldots, c_d \geq 0$

- Motivation: make it more "costly" (in regularizer) to use certain features
- Where do costs come from?

Example:

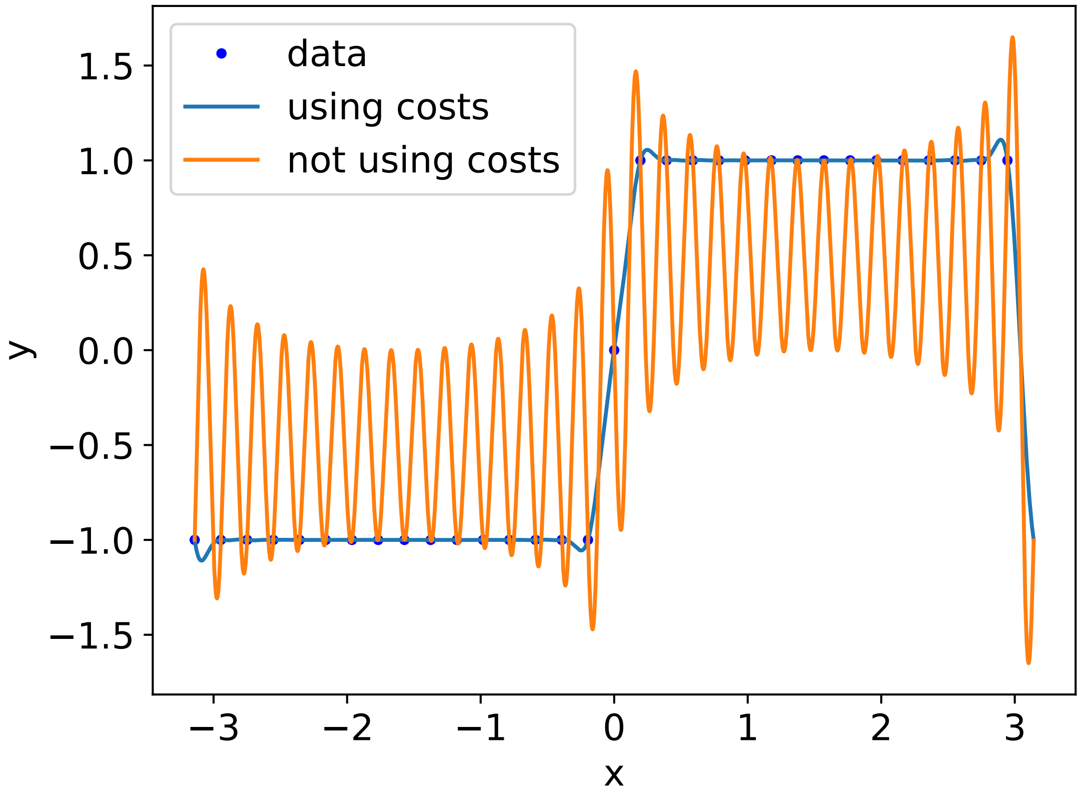
$$\varphi(x) = (1, \cos(x), \sin(x), \cos(2x), \sin(2x), \ldots, \cos(32x), \sin(32x))$$

with regularizer on $w = (w_0, w_{\cos,1}, w_{\sin,1}, \ldots, w_{\cos,32}, w_{\sin,32})$

$$w_0^2 + \sum_{j=1}^{32} j^2 \times \left(w_{\cos,j}^2 + w_{\sin,j}^2\right)$$

(More expensive to use "high frequency" features)





**Regularization elsewhere:**

- Limit size/depth of decision tree
- Restrict flexibility of covariance matrices in normal generative model
- Increasing $K$ in $K$-nearest neighbor (so the predictor averages/votes over more neighbors)
- Bagging / model averaging
- ...



Question: Can effect of costs be achieved using (original) ridge regularization by changing $\varphi$?