Cross-Validation, Regularization (Reading: 15.3, Ch 16)

Learning goals:

- Learn how to perform K-fold CV and its benefits over a held-out validation set.
- Understand L2 and L1 regularization and how to use regularization to manage the bias-variance tradeoff.

UC Berkeley Data 100 Summer 2019
Sam Lau

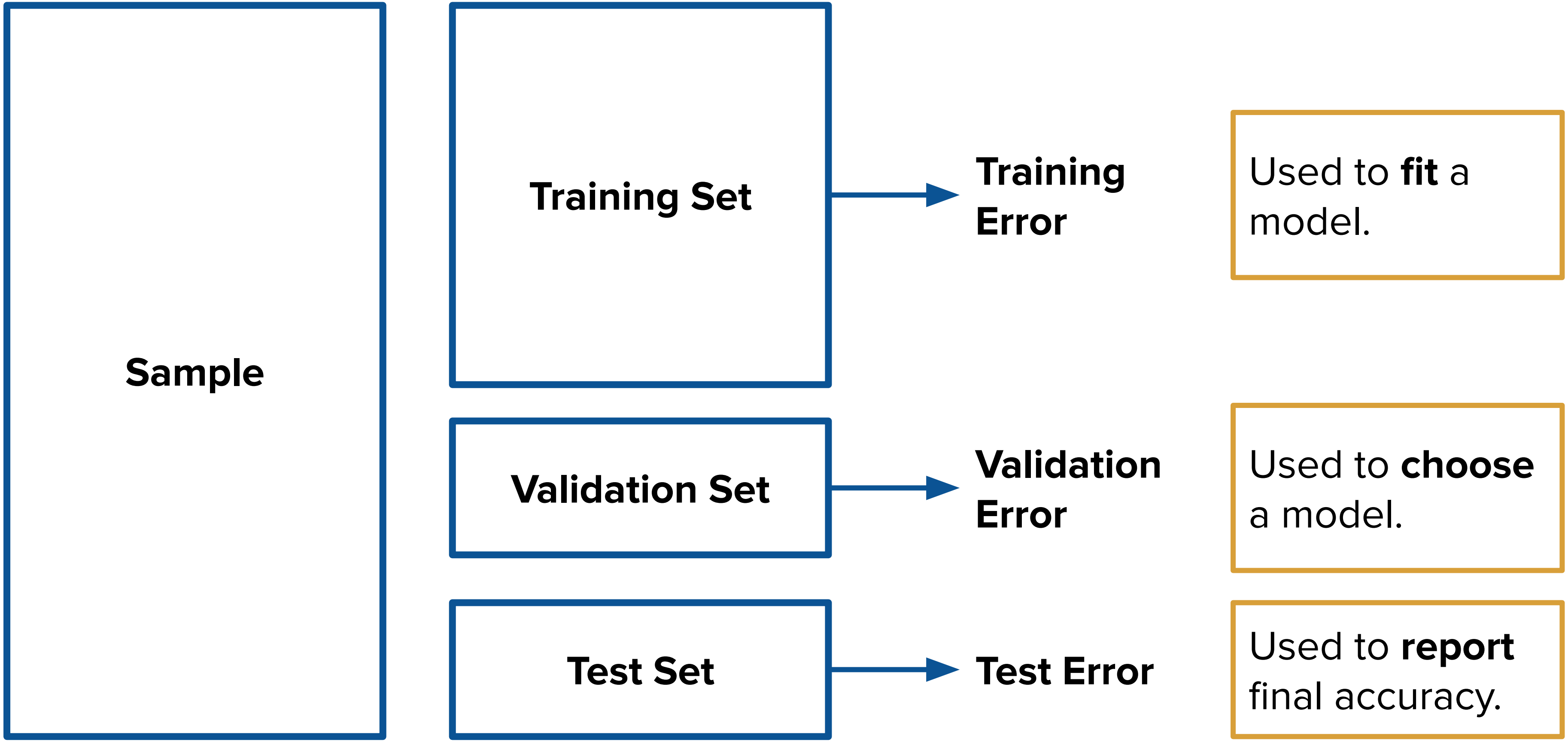
(Slides adapted from Sandrine Dudoit and Joey Gonzalez)

Announcements

- HW5 out, due **tomorrow**
- HW6 out tomorrow, due **Tuesday**
- Screencast yesterday got frozen but audio is there
  - If you leave a comment on the YT video with the slide numbers and times I can update the description, e.g.
  - 00:00 - Slide 1
    01:30 - Slide 2
    etc.

Cross-Validation

Simple Validation



Assessing Model Risk

Training Set →	**Training Error**	Used to **fit** a model.	Minimizes **empirical** risk
Validation Set →	**Validation Error**	Used to **choose** a model.	Estimates **population** risk
Test Set →	**Test Error**	Used to **report** final accuracy.	“Clean” estimate of **population** risk

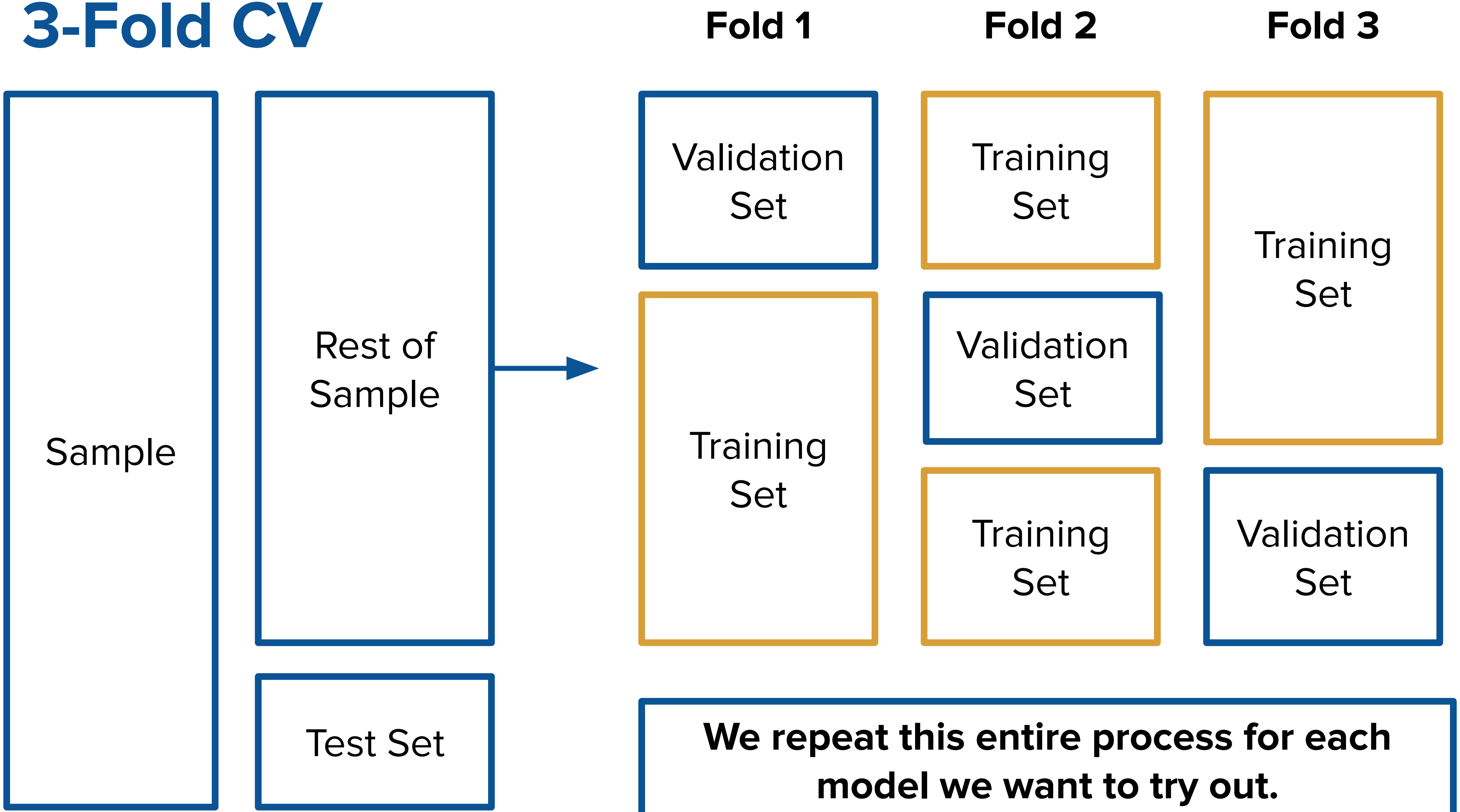
Model Selection

- Given models: $f_\theta^1(\boldsymbol{x}), f_\theta^2(\boldsymbol{x}), \ldots, f_\theta^m(\boldsymbol{x})$
  - E.g. f^1 is linear, f^2 is deg 2 poly, f^3 is linear with fewer features, etc.
- Fit θ for each model by minimizing the training error.
- Compute validation error for each model.
- Pick the model with the **lowest validation error**.
  - This is **model selection**.
- Now, report the test error of chosen model.

K-Fold CV

- Intuition: Validation error will not always be close to true risk. (Sometimes we are just unlucky!)
  - To address, compute **multiple validation errors** for each model.
- **K-Fold cross-validation:**
  - Set aside test set from sample.
  - Split sample into K equal sized partitions
  - Use K - 1 splits to train, last split as validation set.
  - Repeat K times, average of K errors is validation error.

3-Fold CV


Fold 1
Fold 2
Fold 3
Sample
Rest of Sample
Test Set
Validation Set
Training Set
Training Set
Validation Set
Training Set
Training Set
Validation Set
We repeat this entire process for each model we want to try out.

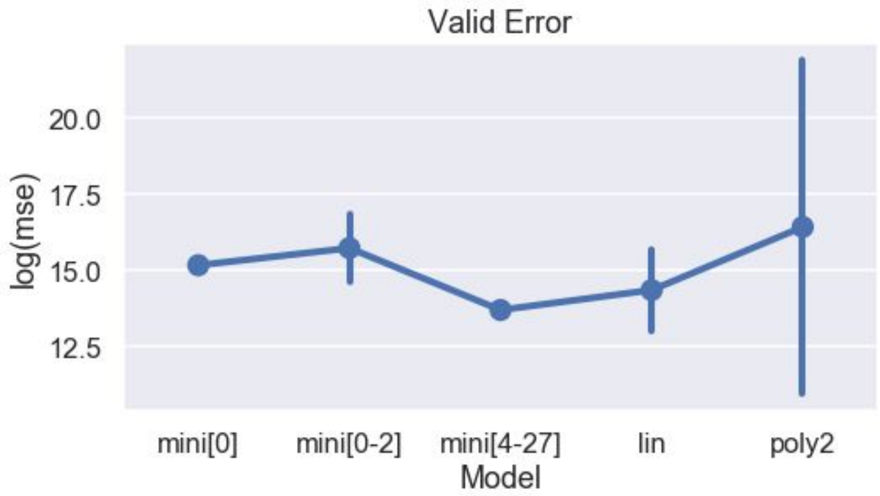
K-Fold CV Analysis

- K usually chosen to be 5 or 10.
- Advantages:
  - Makes use of more data for training (data often scarce)
  - Repeated estimates mitigates variance of splits
  - Can create confidence intervals for validation error
- Disadvantages:
  - More computationally expensive

(Demo)

Estimating Risk, Bias, and Variance

- CV lets us see bias and variance!
- Training errors show model bias
- Validation errors show risk, CIs show model variance



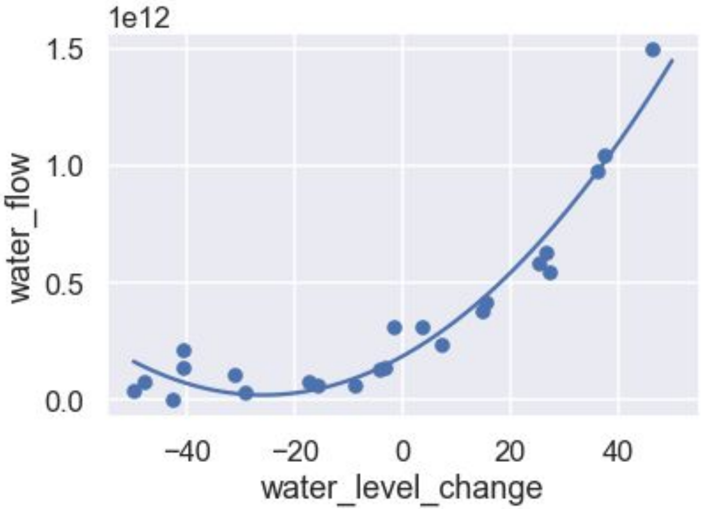
Break!
Fill out Attendance:
http://bit.ly/at-d100

Regularization

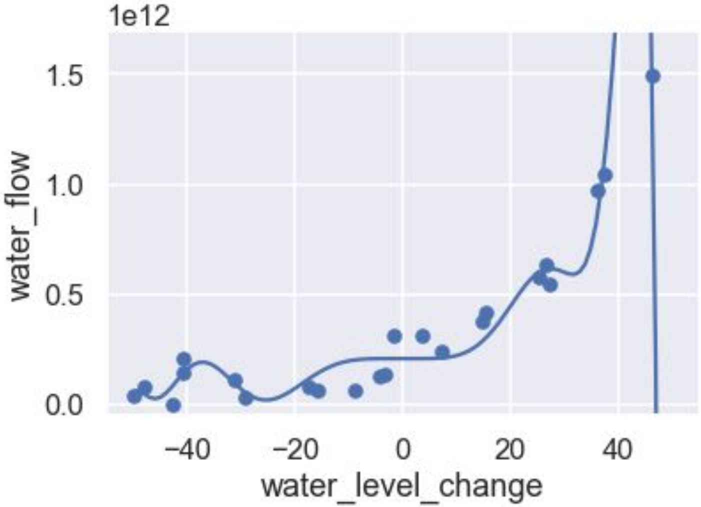
Weighty Issues

Large model weights create complicated models.

Idea: Prevent large weights to make simpler models.



deg	coef
0	186358885440.08
1	12848065664.58
2	247652389.75



deg	coef
0	124.36
1	-257721.06
2	32990.58
3	-79440.10
4	4648550.09
5	137009.22
6	-8829.80
7	-287.46
8	4.97

Regularization

- **Regularization** (aka shrinkage) adds a penalty for model weights to the loss function.
- MSE loss with L2 regularization:

$$L(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) = \boxed{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(\boldsymbol{X_i}))^2} + \boxed{\lambda \sum_{j=1}^{p} \theta_j^2}$$

Same ol' loss as usual

Penalty for θ values

λ: Regularization parameter (non-negative)

Ridge and Lasso Regression

- **Ridge regression**: linear model with L2 regularization

$$L(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{X_i} \cdot \boldsymbol{\theta})^2 + \boxed{\lambda \sum_{j=1}^{p} \theta_j^2}$$

L_2 norm

- **Lasso regression**: linear model with L1 regularization

$$L(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{X_i} \cdot \boldsymbol{\theta})^2 + \boxed{\lambda \sum_{j=1}^{p} |\theta_j|}$$

L_1 norm

(Demo)

Regularization Parameter

L2

$$L(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(\boldsymbol{X_i}))^2 + \lambda\sum_{j=1}^{p}\theta_j^2$$

L1

$$L(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(\boldsymbol{X_i}))^2 + \lambda\sum_{j=1}^{p}|\theta_j|$$

- λ is the regularization parameter.
- Higher values penalize model weights more.
- Discuss:
  - What happens when λ = 0?
  - What happens when λ = ∞?
  - Does this change between L2 and L1 regularization?

What happens when...

- $\lambda = 0$?
  - No regularization, back to linear model
- $\lambda = \infty$?
  - Flat line, all model weights = 0
- Does this change between L2 and L1 regularization?
  - No

Don’t regularize the bias

- Notice that we don’t regularize the bias term!

$$f_\theta(\boldsymbol{x}) = \boxed{\theta_0} + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p$$

$$L(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(\boldsymbol{X_i}))^2 + \lambda \sum_{\boxed{j=1}}^{p} \theta_j^2$$

- Discuss: why not?
    - Bias term doesn’t add complexity to model

Normalize Data Before Using Regularization

$$L(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(\boldsymbol{X_i}))^2 + \lambda\sum_{j=1}^{p}\theta_j^2$$

- Before using regularization, **normalize** data
  - Subtract mean and scale data to lie between -1 and 1.
- Discuss: what happens if we don't do this?
  - Artificial penalty on features with small numbers

Exercise to take home:

- Prove that the stochastic gradient descent update rule for ridge regression is:

$$\boldsymbol{\theta}^{(t+1)} = (1 - 2\lambda\alpha)\boldsymbol{\theta}^{(t)} + 2\alpha(y_i - \boldsymbol{\theta} \cdot \boldsymbol{x})(\boldsymbol{x})$$

- (Lasso is a bit tricker but also doable.)

Why two kinds of regularization?

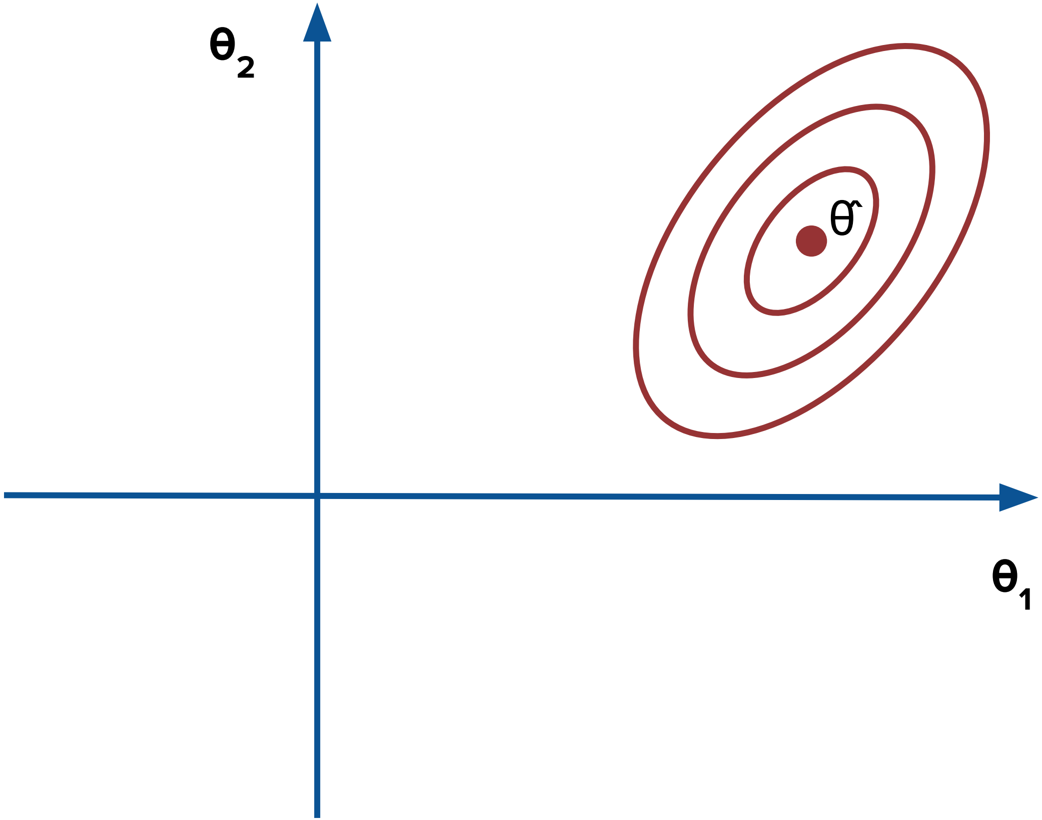
- Intuitive, hand-wavy explanation:
- L2 regularization typically has all non-zero weights.
  - Makes sense when we think many small factors contribute to outcome.
- L1 regularization will set some model weights = 0 depending on how big λ is.
  - L1 regularization lets us perform **feature selection**.
  - Makes sense when we think a few major factors contribute to outcome.

A more sophisticated explanation

Suppose we have a linear model with two parameters and no intercept term.

As we tweak the two parameters, loss changes.

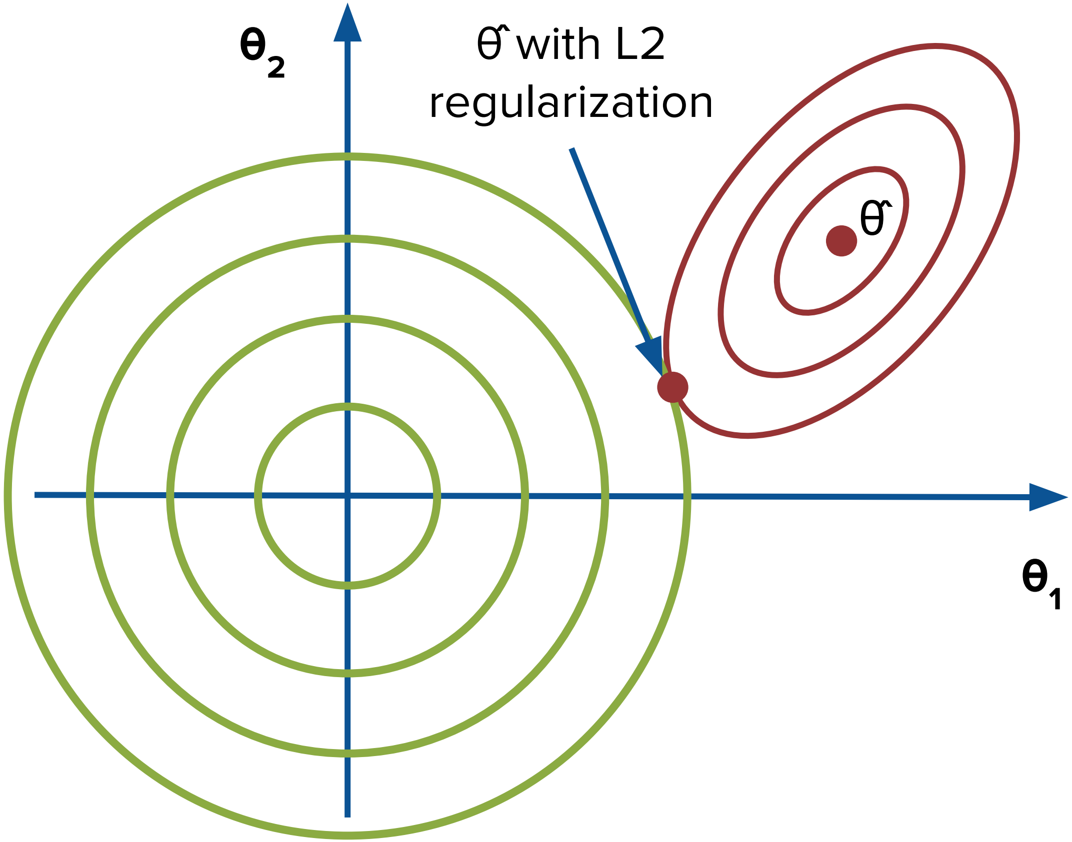
Without regularization, we just pick $\hat{\theta}$.



A more sophisticated explanation

Regularization balances loss with the regularization penalty.

For L2 regularization, we have circular contours for the penalty. Why?

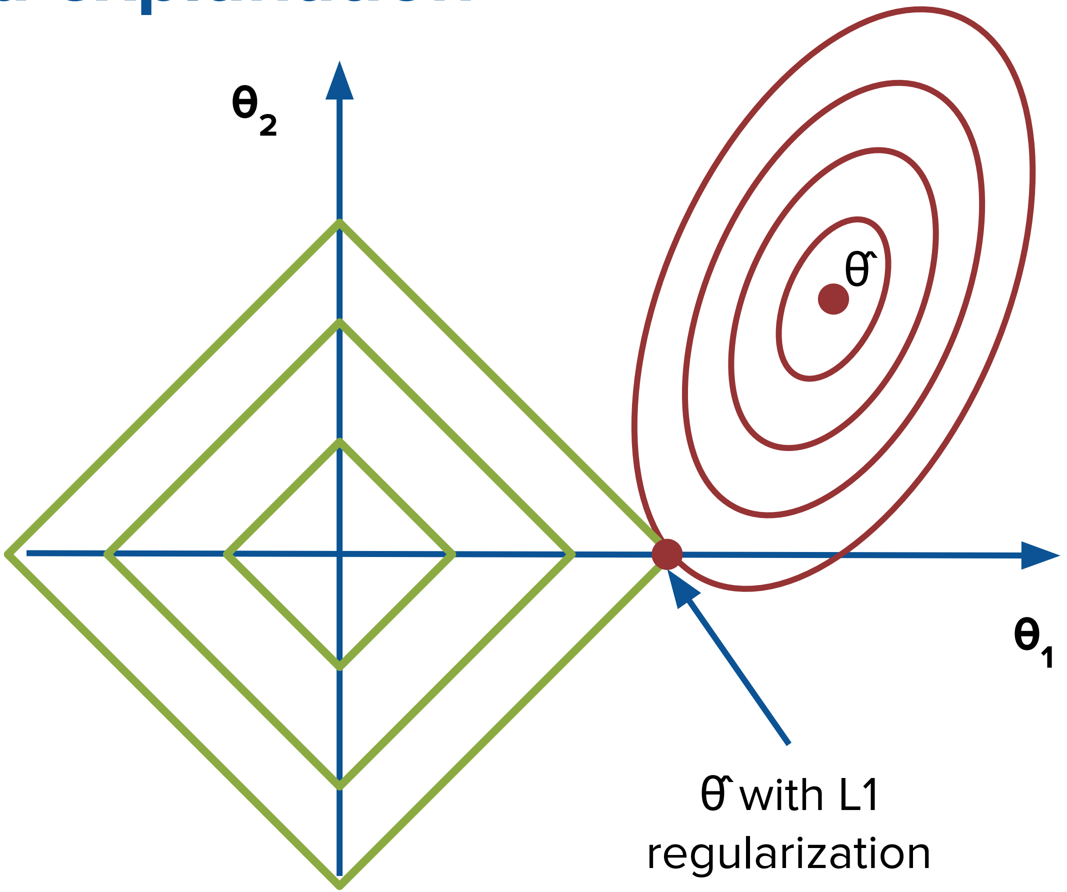


A more sophisticated explanation

For L1 regularization, we have diamond-shaped contours for the penalty. Why?

Notice that this sets one parameter = 0!

This idea extends to multiple dimensions.



A tuning knob for bias-variance

- Regularization gives us yet another way to manage the bias-variance tradeoff.
  - Increase λ = more bias, less variance
  - Decrease λ = less bias, more variance
- How do we pick λ?
  - Cross-validation!

Summary

- K-Fold cross-validation lets us estimate model bias, model variance, and overall risk.
  - We use CV to perform model and feature selection.
- Regularization gives us a way to add complexity to our models while avoiding overfitting.
  - We use CV to tune the regularization amount.