Introduction to Machine Learning

Regularization

Instructor: Pat Virtue

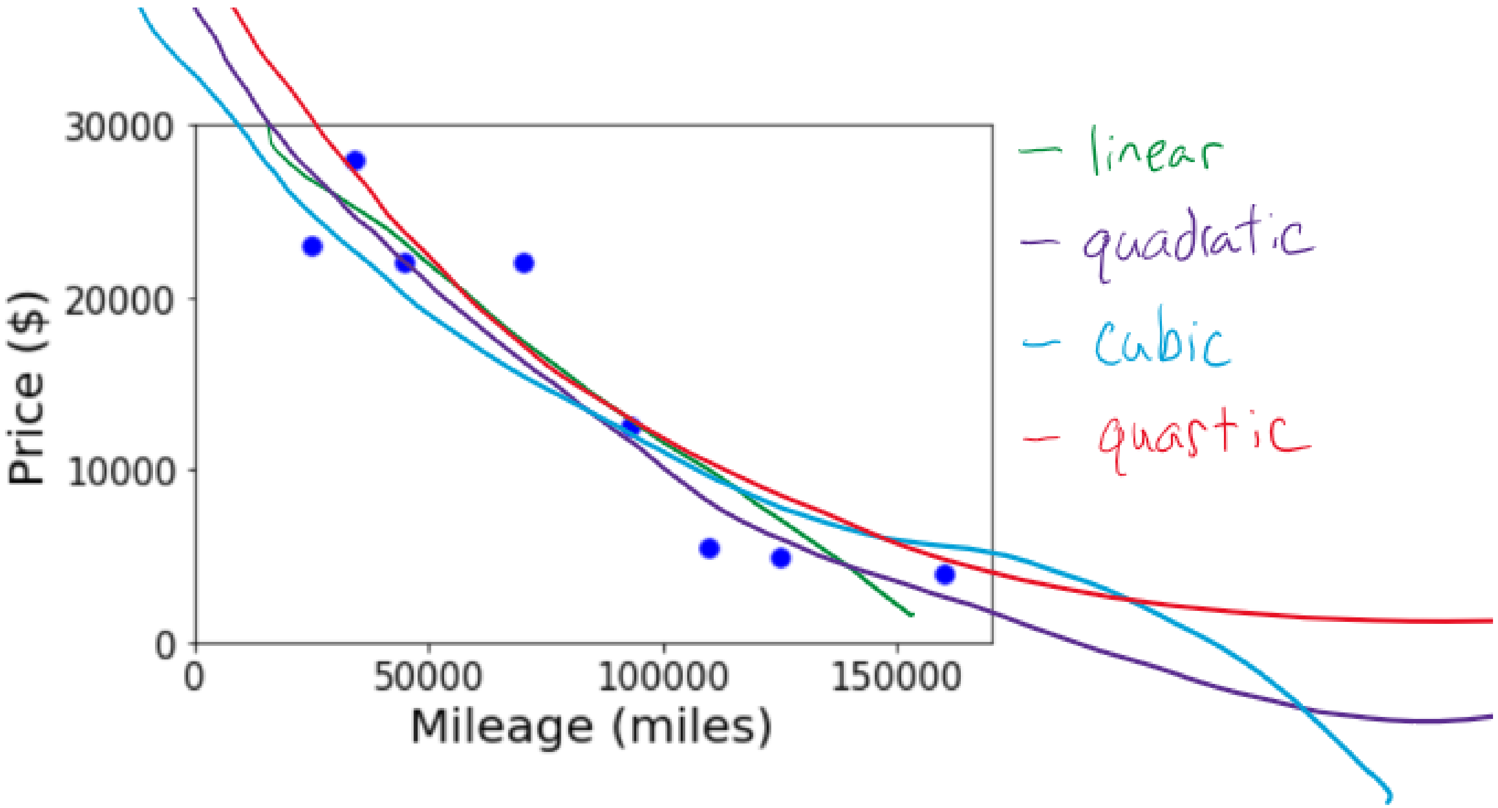
Announcements

Assignments:

- HW3
  - Planned for release tonight
  - Due Tue, 2/11, 11:59 pm

Overfitting with Polynomial Linear Regression

Better fit training data with higher model complexity



$$w \quad \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

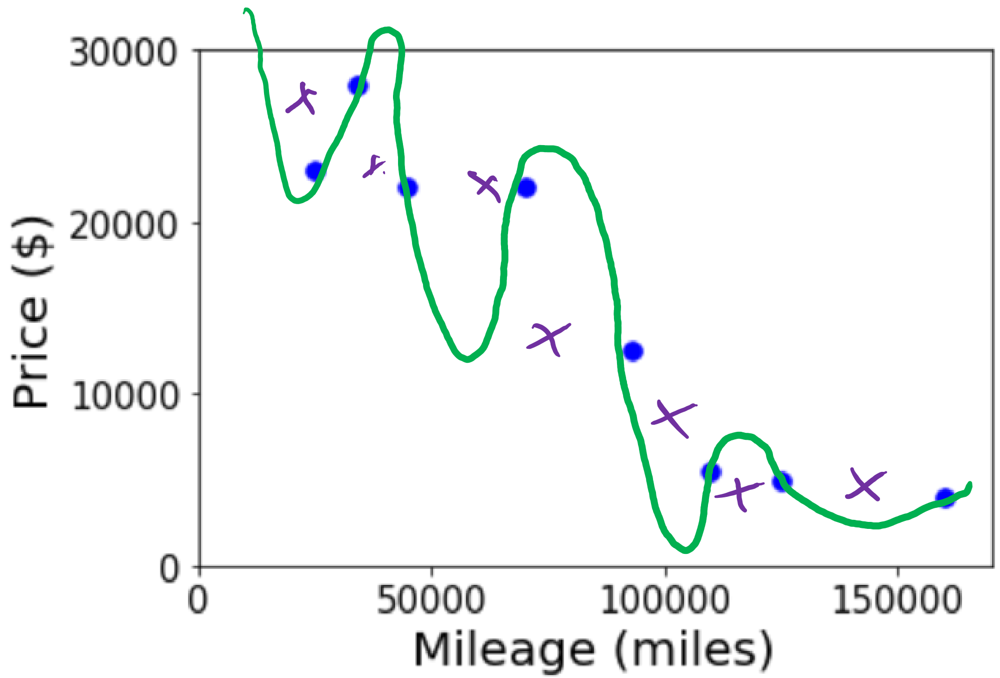
$$X$$

$$\begin{bmatrix} 1 & x^{(1)} & x^{(1)2} & x^3 \\ 1 & x^{(2)} & x^2 & x^3 \\ 1 & x & \vdots^2 & \\ 1 & x^{(N)} & x^2 & x^3 \end{bmatrix}$$

$$y - Xw$$

Overfitting with Polynomial Linear Regression

Better fit training data with higher model complexity



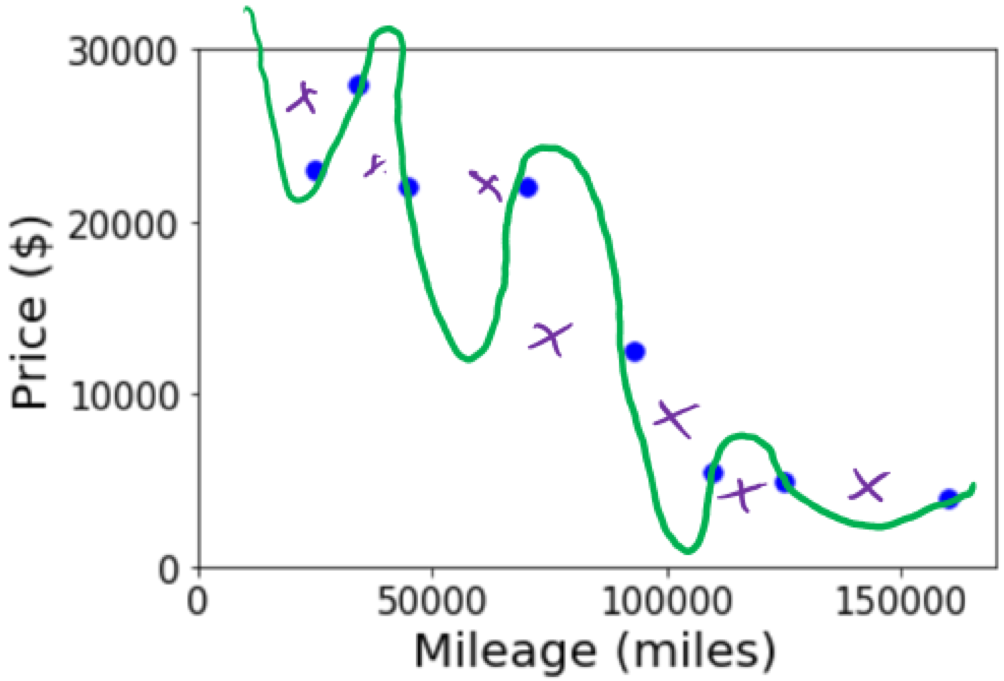
$$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \dots + w_{20} x^{20}$$

How can we deal with overfitting? Use validation. More training data

What are some symptoms of overfitting? Huge weights!

Overfitting with Polynomial Linear Regression

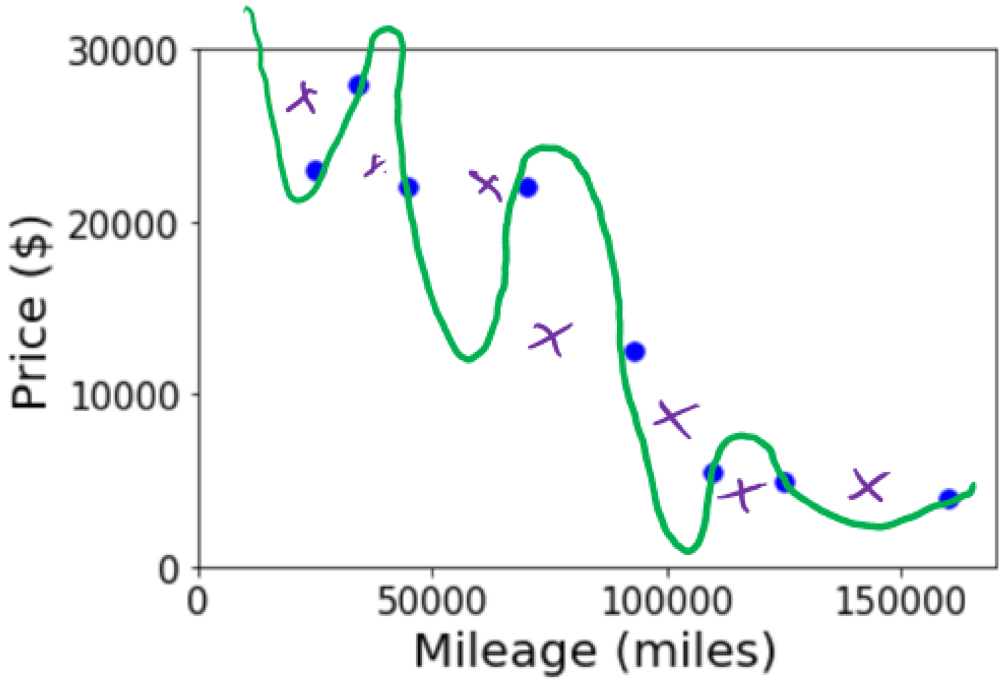
How can we deal with overfitting?



- Use validation set to detect overfitting
- Collect more training data
- Reduce model complexity
  - Lower degree polynomial
  - But then we might underfit ☹
- Try fitting to many different degrees
  - Use validation data to decide which level of model complexity to use
- Penalize the weights

Overfitting with Polynomial Linear Regression

What are symptoms of overfitting?



- Poor validation score
- HUGE weights!

Regularization

Combine original objective with penalty on parameters

$$\min_w J(w) \quad + \quad \min_w \text{weight}$$

$$\min_w J(w) + f(w)$$

$$f_1(w) = \sum_m w_m \qquad -9000 + 8900$$

$$f_2(w) = \sum_m |w_m|$$

$$f_3(w) = \sum_m w_m^2$$

Piazza Poll 1:

Given the optimization of our new objective:

$$\hat{\boldsymbol{w}} = \min_{\boldsymbol{w}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 \ + \ \lambda\|\boldsymbol{w}\|_2^2$$

Select ALL that are true:

I. As $\lambda \to 0$, $\hat{\boldsymbol{w}} \to$ point A

II. As $\lambda \to 0$, $\hat{\boldsymbol{w}} \to$ point B

III. As $\lambda \to \infty$, $\hat{\boldsymbol{w}} \to$ point A

IV. As $\lambda \to \infty$, $\hat{\boldsymbol{w}} \to$ point B

V. None of the above

VI. I have no clue

Regularization

Ridge Regression: Linear regression with ℓ_2 penalty on weights

Ridge Regression

Linear regression with ℓ_2 penalty on weights

$$J(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{1}{2}\lambda\|\boldsymbol{w}\|_2^2$$

$$= \frac{1}{2}[\boldsymbol{y}^T\ \boldsymbol{y}\ - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + \lambda\boldsymbol{w}^T\boldsymbol{w}]$$

Compute gradient

$$\nabla J(\boldsymbol{w}) = -\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + \lambda\boldsymbol{w}$$

Closed form solution:

$$-\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + \lambda\boldsymbol{w} = 0$$

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + \lambda\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$$

$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda)\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$$ Not quite $(\boldsymbol{A} + 7)\boldsymbol{z} \neq \boldsymbol{A}\boldsymbol{z} + 7\boldsymbol{z}$

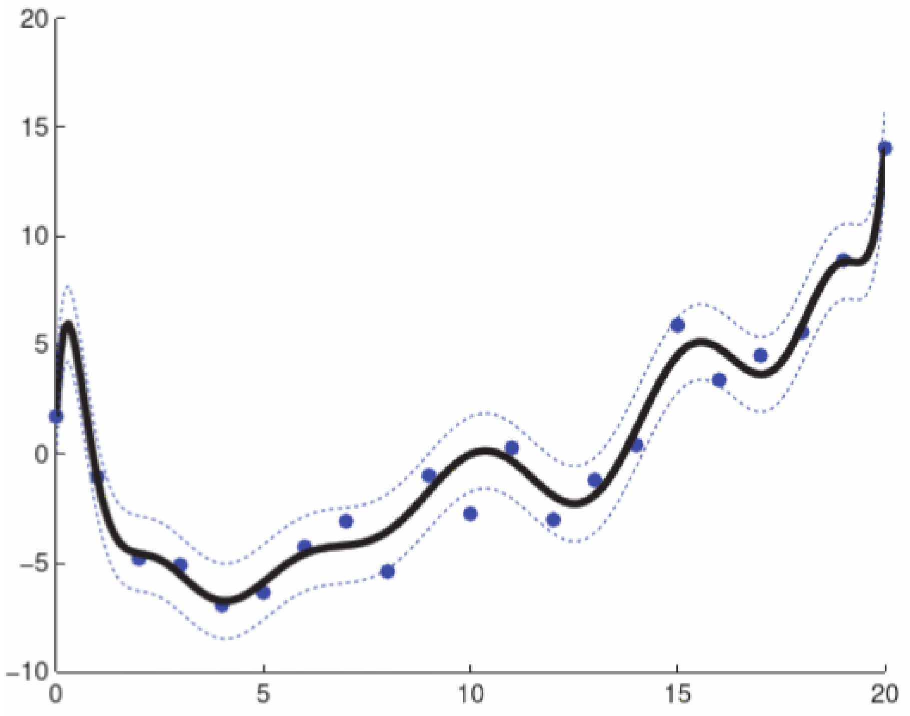
$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$$

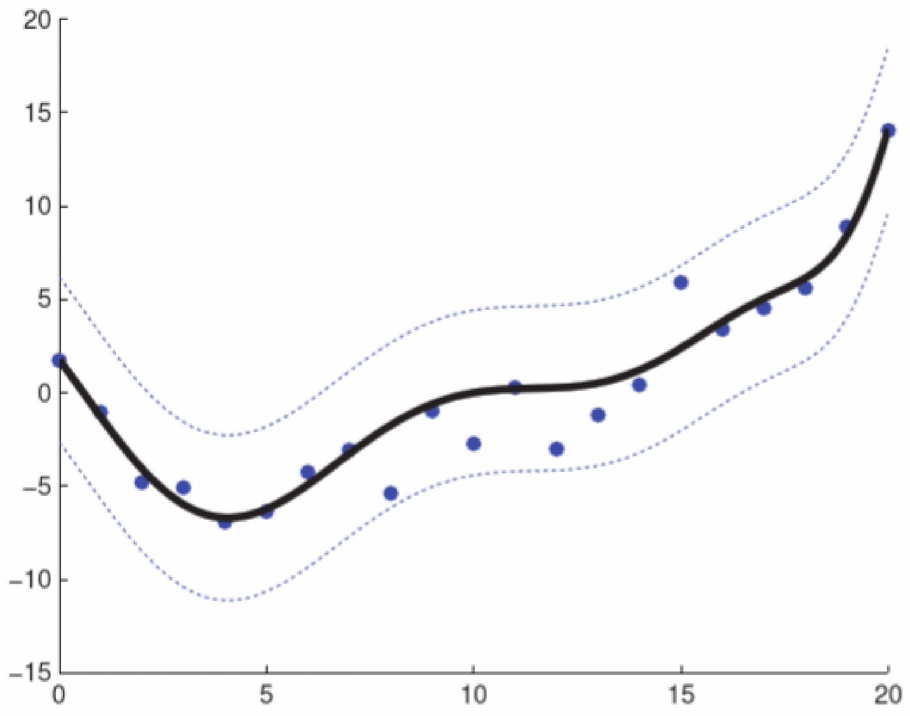
$$\boldsymbol{w} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

A robust solution to make $\boldsymbol{X}^T\boldsymbol{X}$ invertible

Regularization

But how do we choose λ?





Figures: Murphy, Ch 7.5

Probabilistic Interpretation

What assumptions are we making about our parameters?

MLE and MAP

Coin Flipping Example

Piazza Poll 2:

$$p(\theta \mid \mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

$$p(\theta \mid \mathcal{D}) \propto \prod p(\mathcal{D}^{(n)}|\theta)\, p(\theta)$$

As the number of data points increases, which of the following are true?

Select ALL that apply

A. The MAP estimate approaches the MLE estimate
B. The posterior distribution approaches the prior distribution
C. The likelihood distribution approaches the prior distribution
D. The posterior distribution approaches the likelihood distribution
E. The likelihood has a lower impact on the posterior
F. The prior has a lower impact on the posterior

Coin Flipping Example

Housing Price Example

Predict housing price from several features

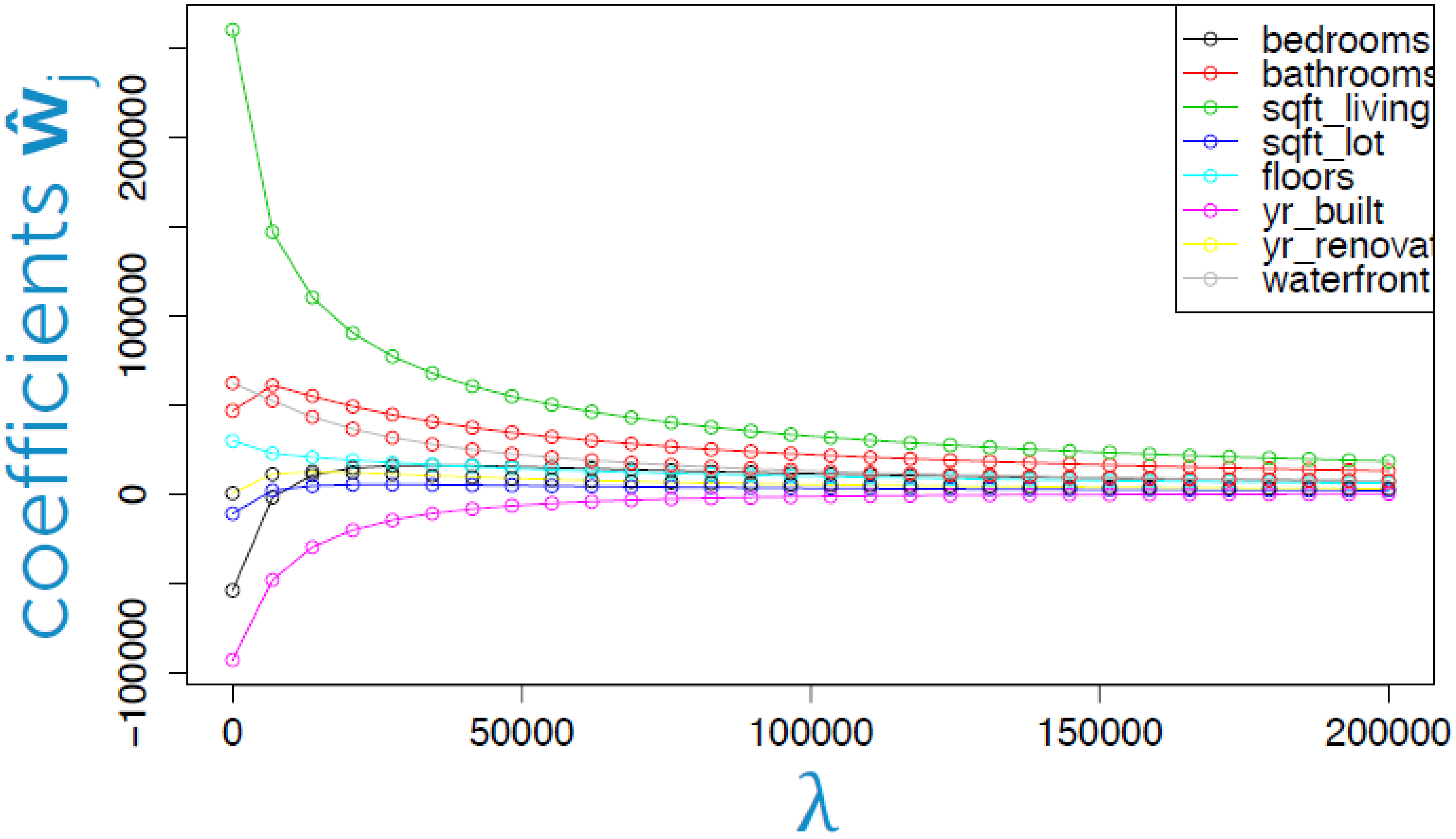


Figure: Emily Fox, University of Washington

Housing Price Example

Predict housing price from several features

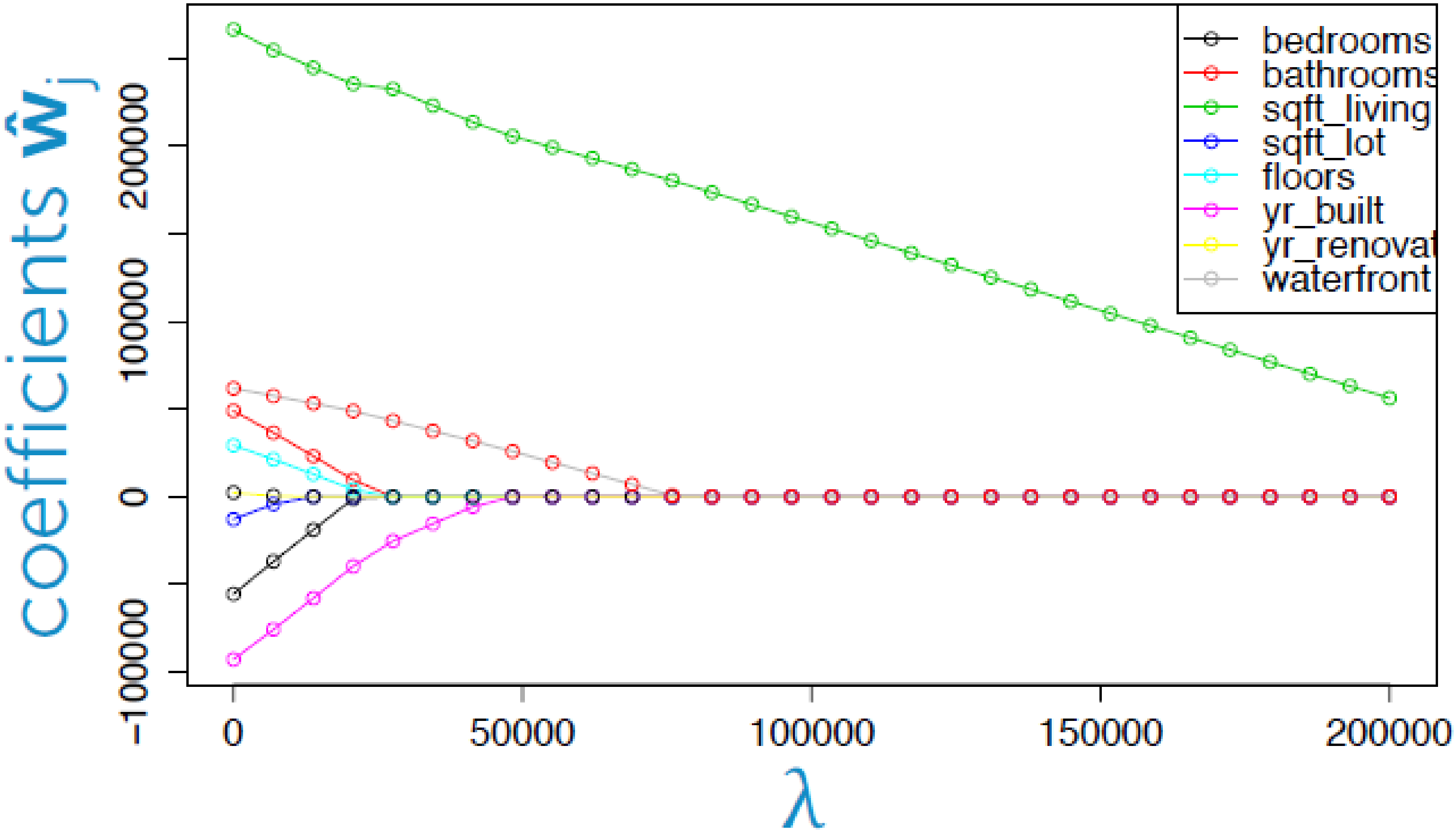


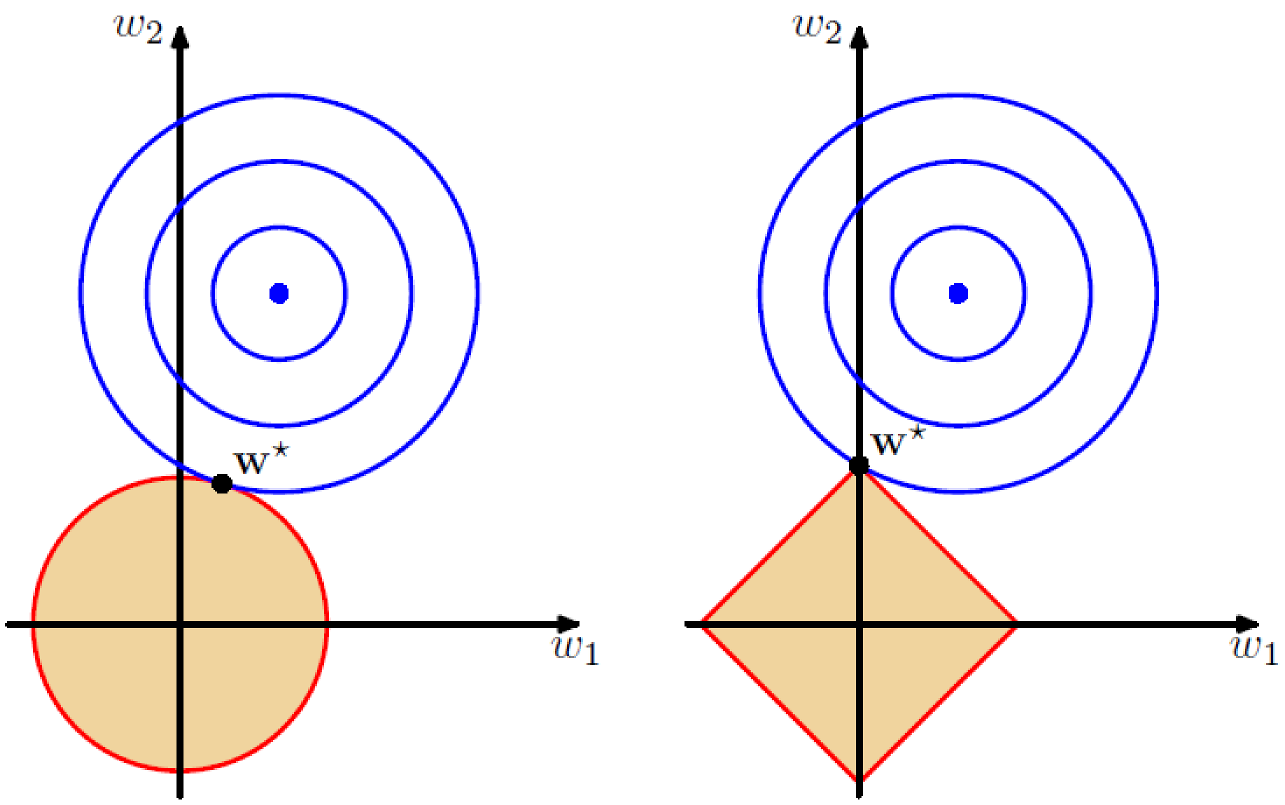
Figure: Emily Fox, University of Washington

Regularization

Combine original objective with penalty on parameters

Regularization

Combine original objective with penalty on parameters



Figures: Bishop, Ch 3.1.4

LASSO

Linear regression with ℓ_1 penalty on weights

LASSO

Linear regression with ℓ_1 penalty on weights

$$J(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{1}{2}\lambda\|\boldsymbol{w}\|_1$$

$$= \frac{1}{2}[\boldsymbol{y}^T\ \boldsymbol{y}\ - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + \sum_m |\boldsymbol{w}_m|]$$

Probabilistic interpretation

Laplace prior on weights

$$w \sim Laplace(\mu = 0, b)$$

$$f(\,w \mid b\,) = \frac{1}{2b} e^{\left(-\frac{|x|}{b}\right)}$$