University of Colorado **Boulder**

Department of Computer Science

CSCI 5622: Machine Learning

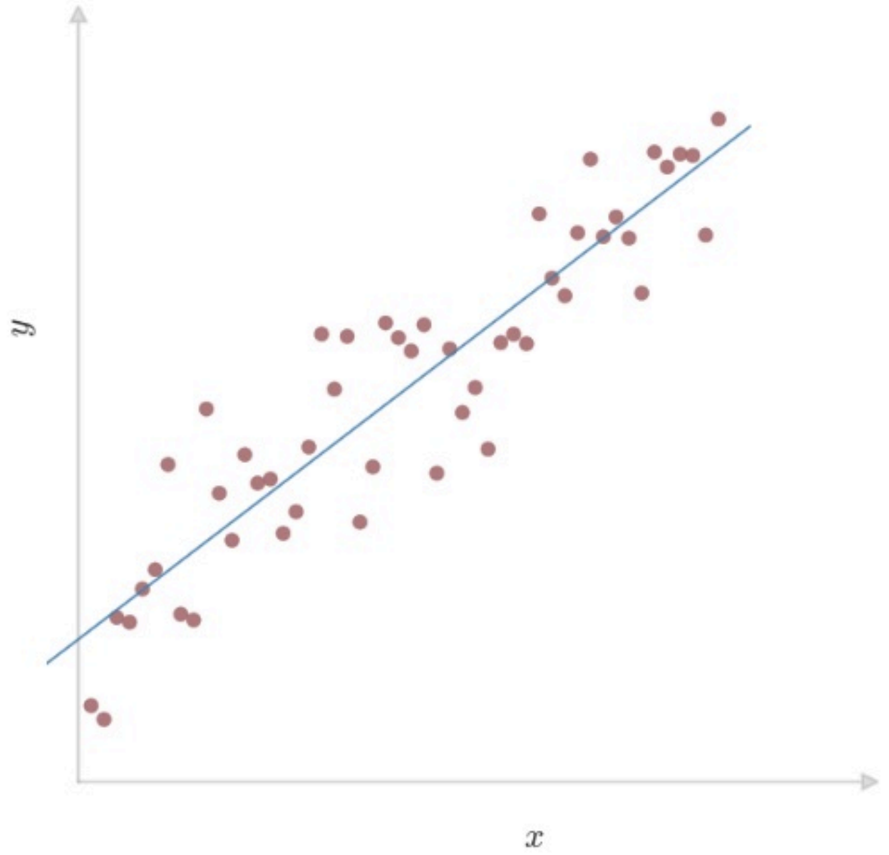
Chris Ketelsen

Lecture 9: Regularization

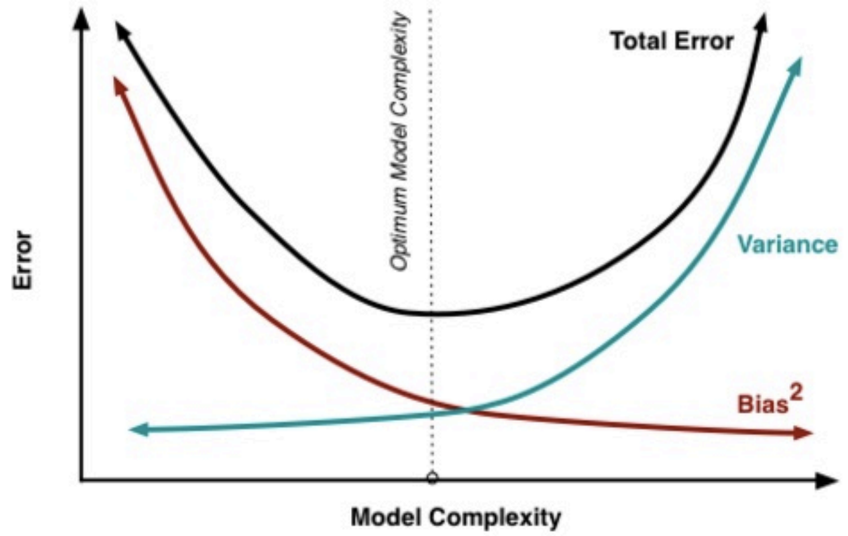
Linear Regression

Data are continuous inputs and outputs $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^m, \mathbf{x}_i \in \mathbb{R}^D$

Want to predict y from $\mathbf{x}$ using a linear model: $y = f(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$



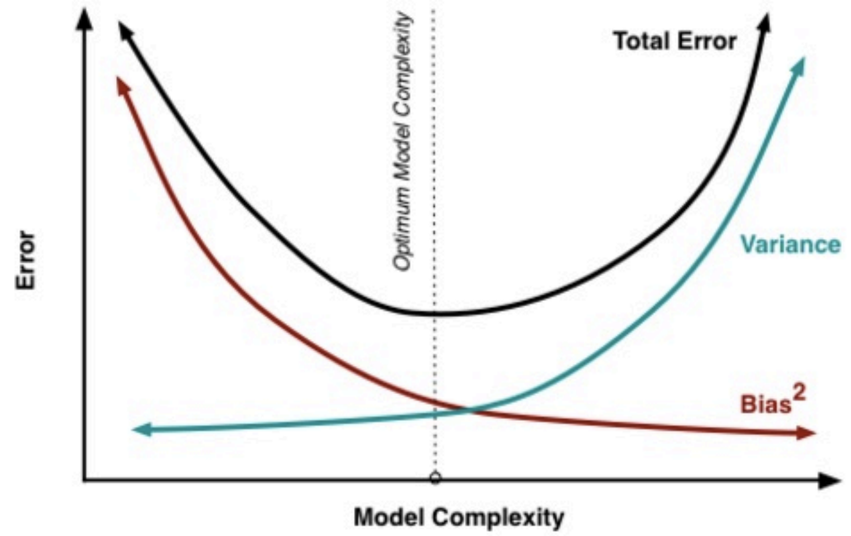
Bias-Variance Trade-Off



Simple models have high **bias** because they can't represent underlying truth

Complicated models have high **variance** because they are sensitive to training data

Bias-Variance Trade-Off

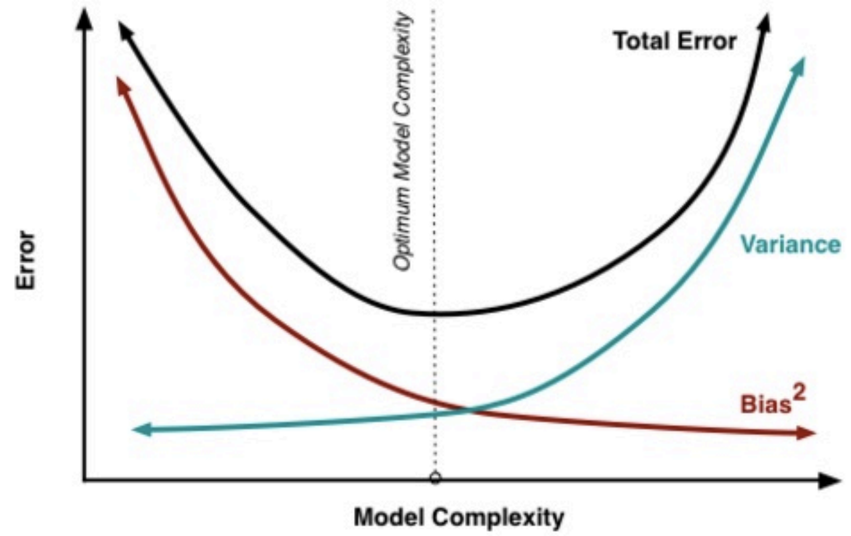


How do we have flexible enough model to capture truth without being sensitive to training data?

Idea: Start with many features and throw some out to reduce variance

But which features do we throw out?

Bias-Variance Trade-Off



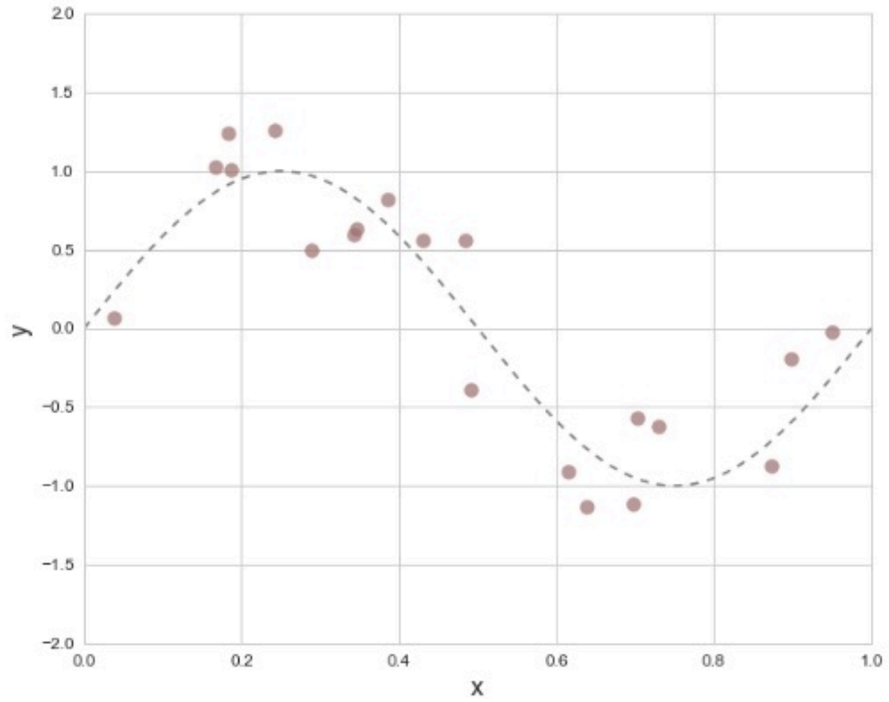
In polynomial regression context it might make sense to throw out higher degree features

But what about multiple linear regression where each feature corresponds to a different predictor?

Need a way to reduce variance while not throwing out potentially interesting features ...

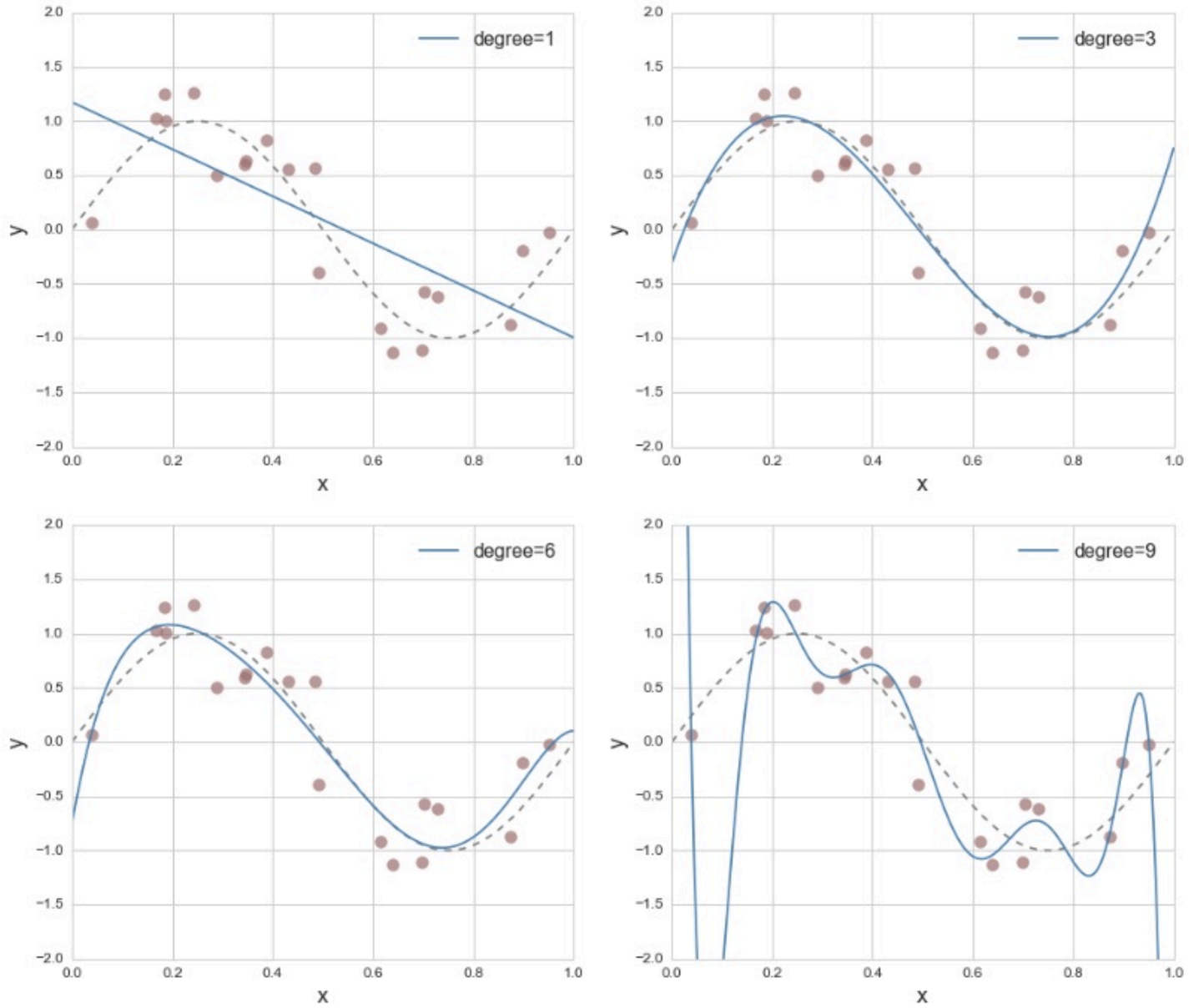
Regularization

Consider polynomial regression on data $f(x) = \sin(2\pi x) + \text{Noise}$



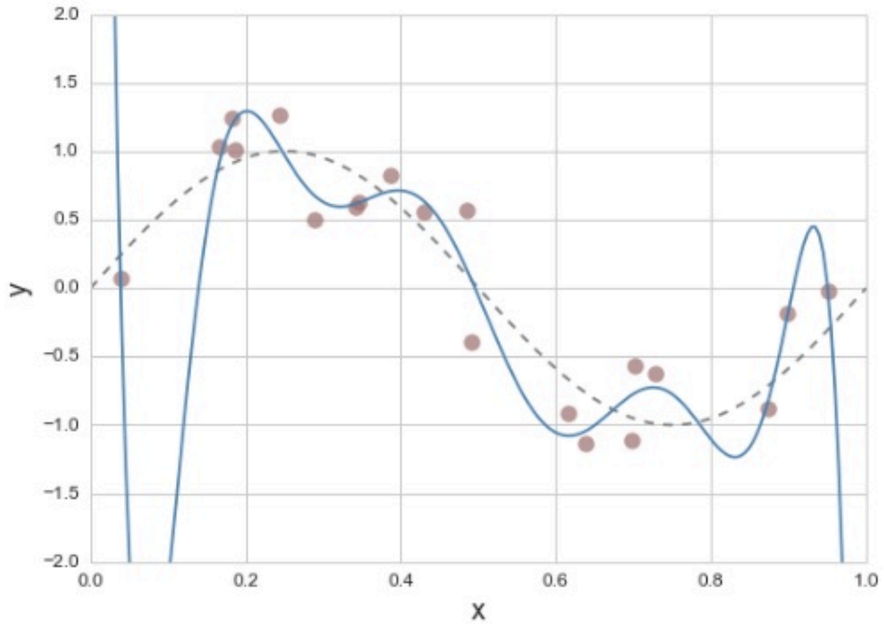
Different degree polynomial models come with different levels of bias and variance

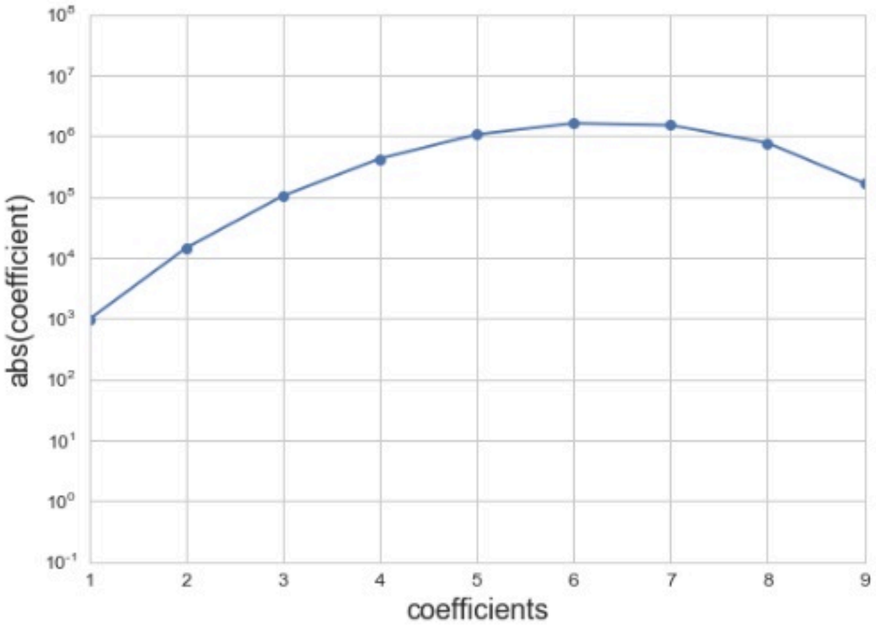
Regularization



Regularization

Examine high variance model of degree 9





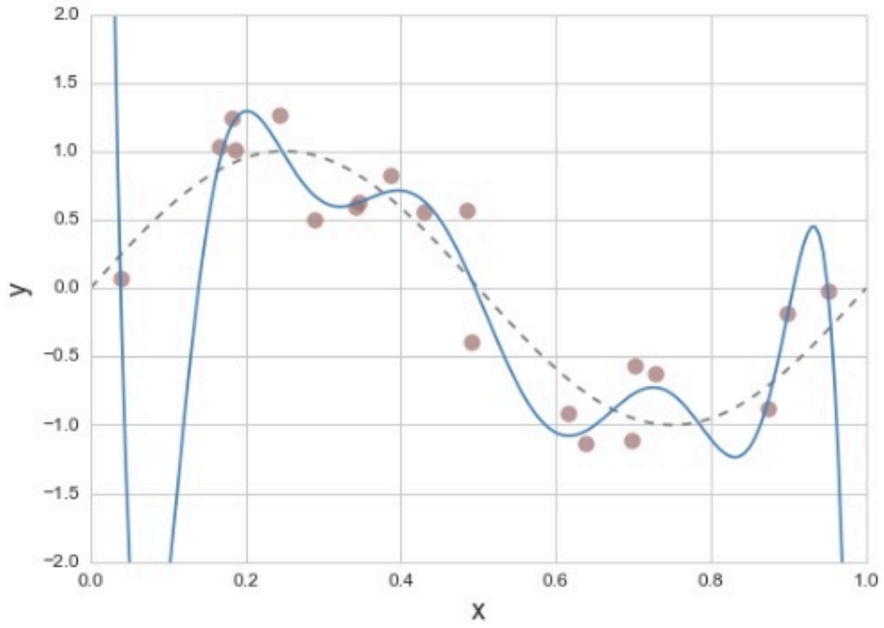
Model wiggles wildly to get close to data

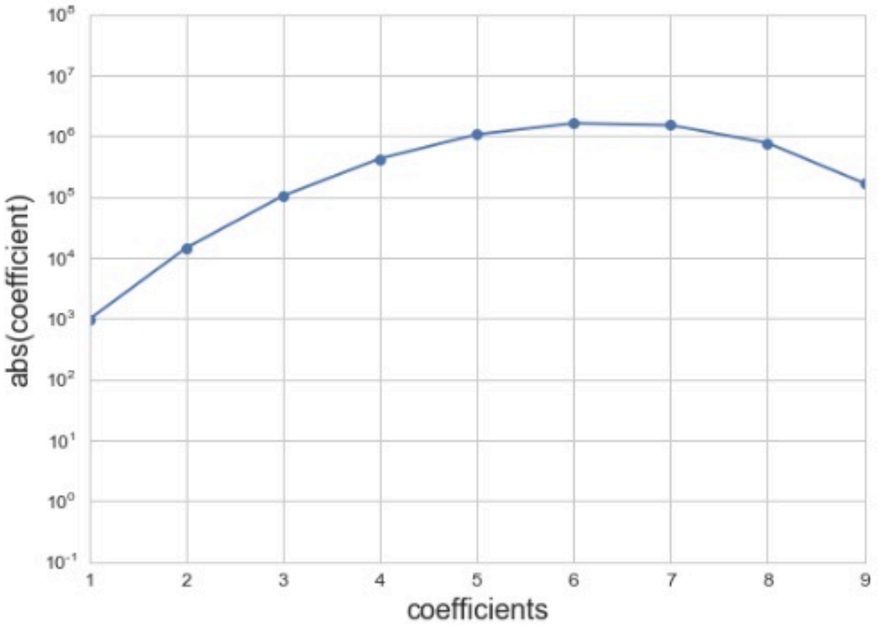
To get big swings, model coefficients are very large

$$w_{6:9} \sim 10^5 - 10^6$$

Regularization

Examine high variance model of degree 9





Idea: Keep all the features, but force the coefficients to be smaller

This is called **Regularization** or sometimes **shrinkage**

Regularization

Add penalty term to RSS objective function

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \sum_{k=1}^{D} \text{penalty}(w_k)$$

RSS

Balance between small RSS **and** small coefficients

Regularization coefficient λ is tuning parameter

$\lambda \to 0$ leads to Ordinary Least-Squares

$\lambda \to \infty$ leads to zero-coefficients (i.e. constant model)

Usually set λ via cross-validation

Regularization

Ridge Regression or ℓ_2-Regularization:

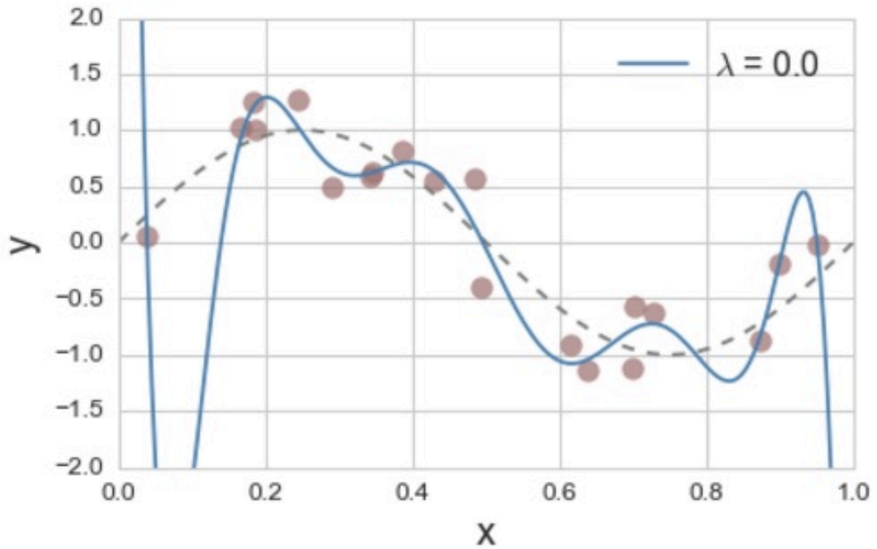
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \sum_{k=1}^{D} w_k^2$$

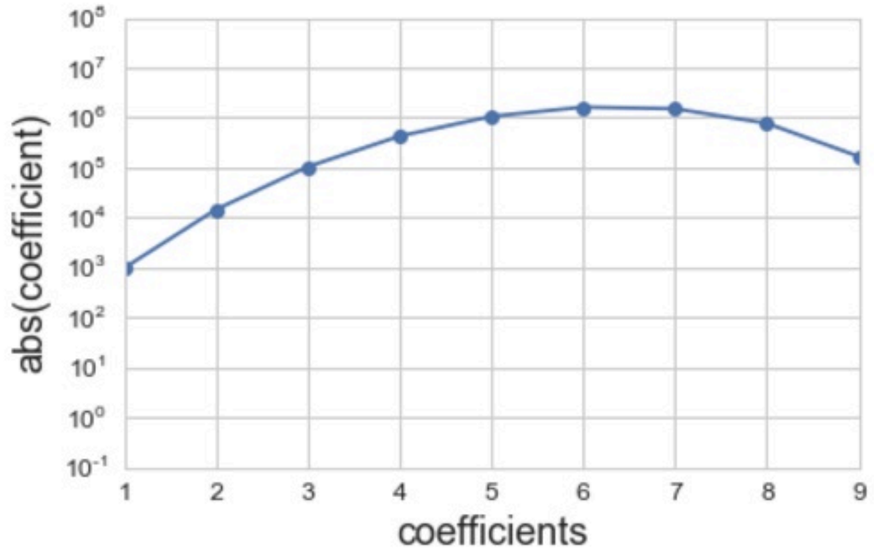
Lasso Regression or ℓ_1-Regularization:

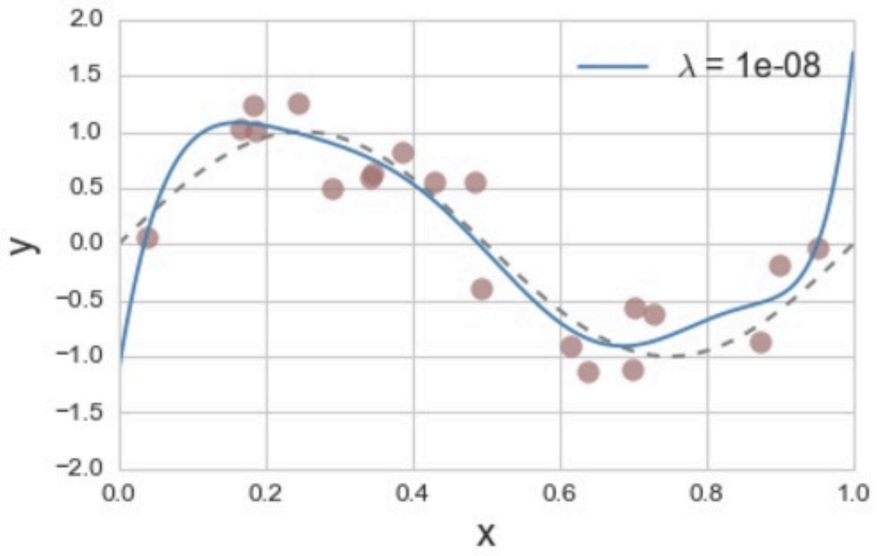
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \sum_{k=1}^{D} |w_k|$$

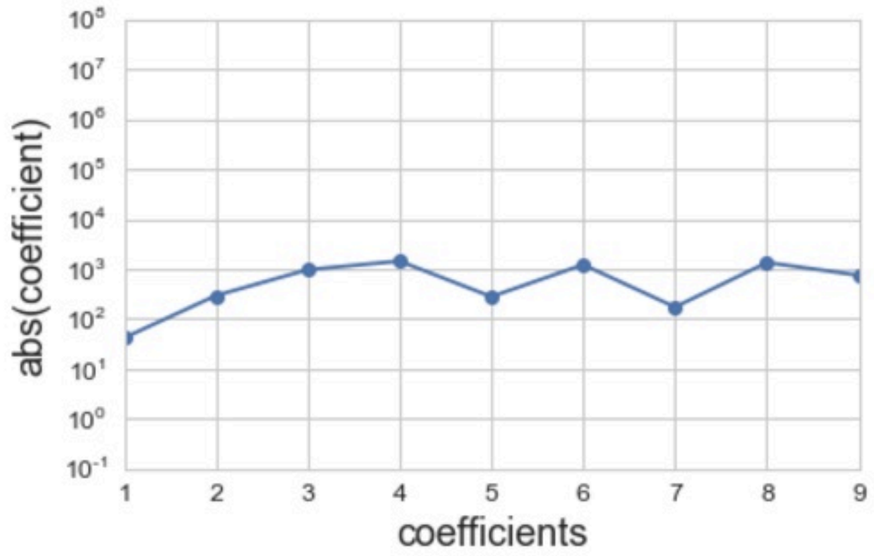
Different penalty terms lead to different character of models

Ridge Regression

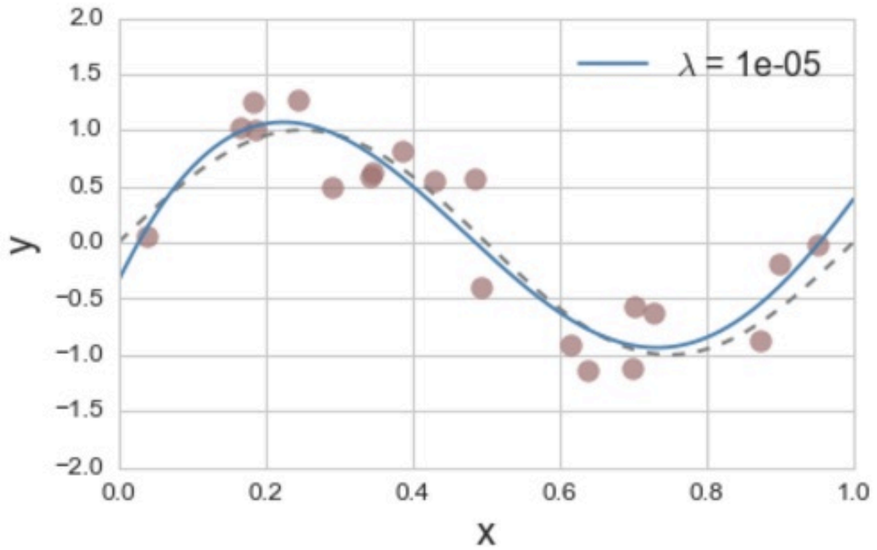


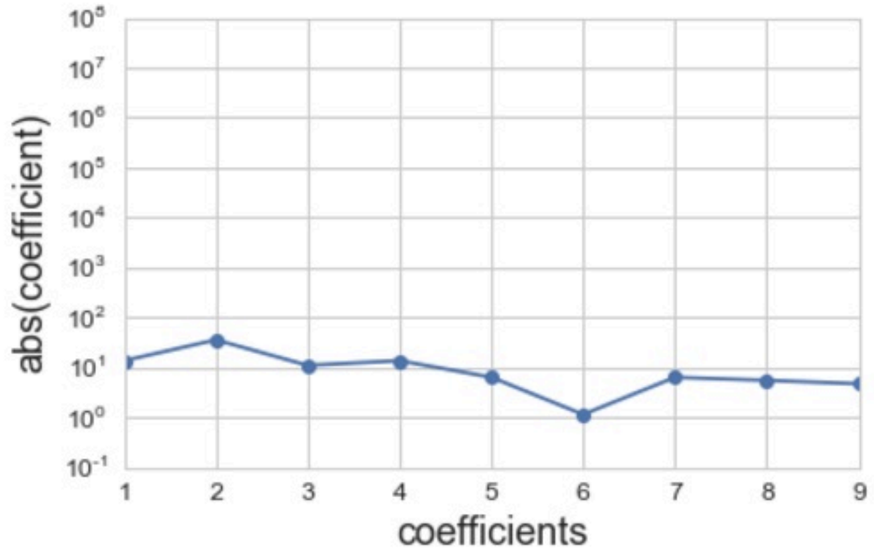


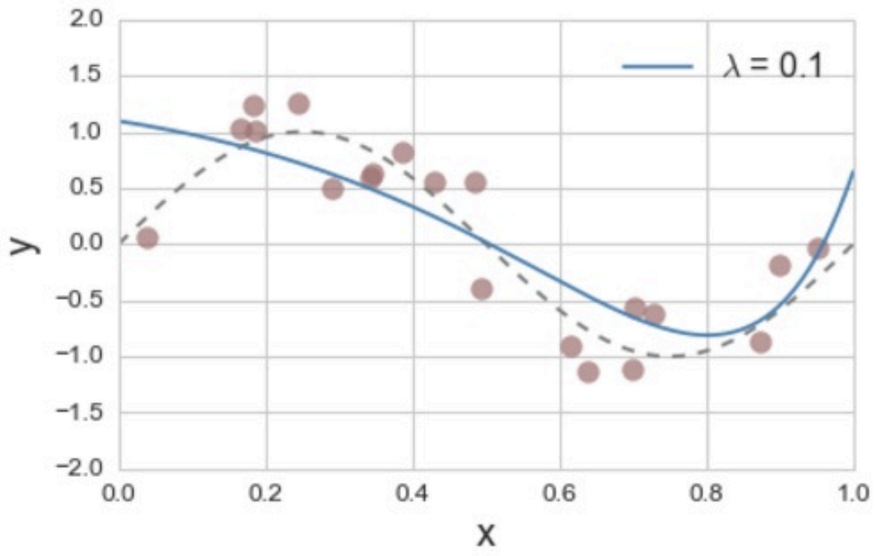


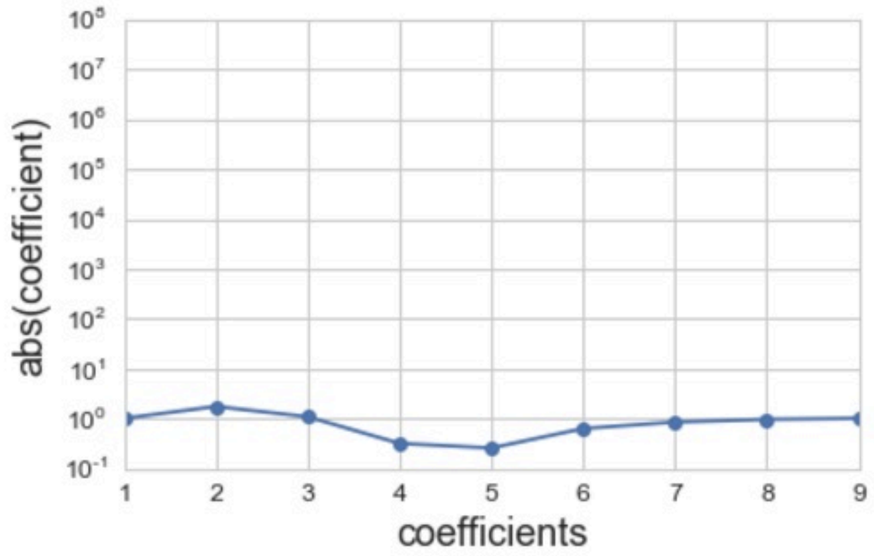


Ridge Regression









Ridge Regression

Bias vs Variance vs λ

As $\lambda \to \infty$, **variance** goes down

Shrinking coefficients prevents large swings in model, which prevents overfitting

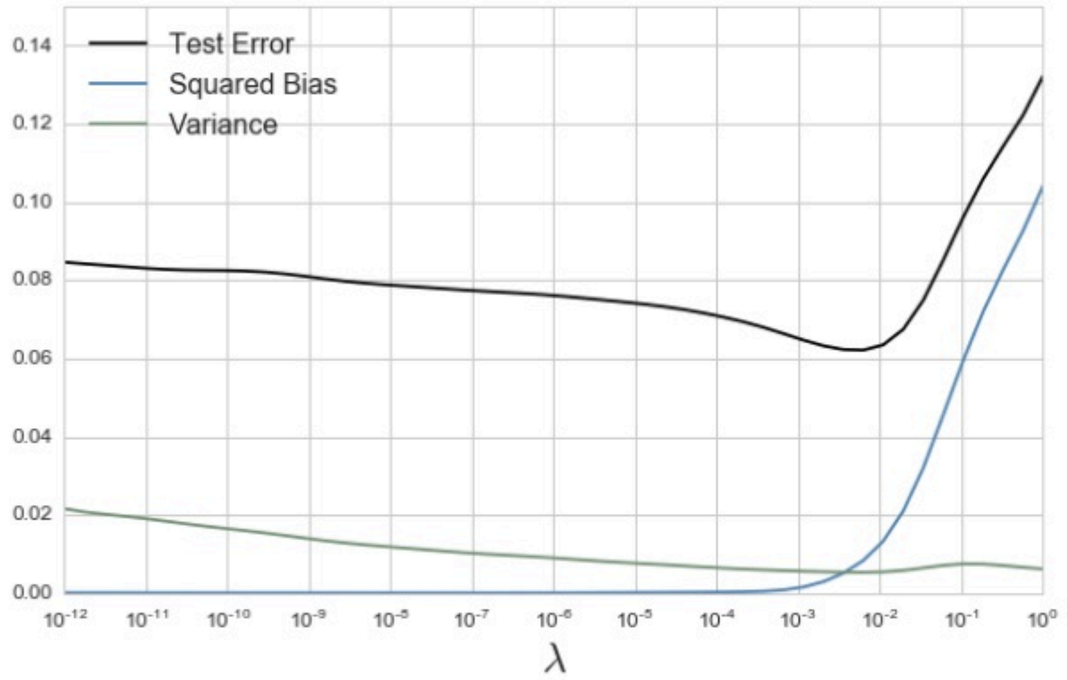
... but **bias** also goes up

If your true function is wiggly, a large λ will prevent even high degree models from wiggling too much

Ridge Regression

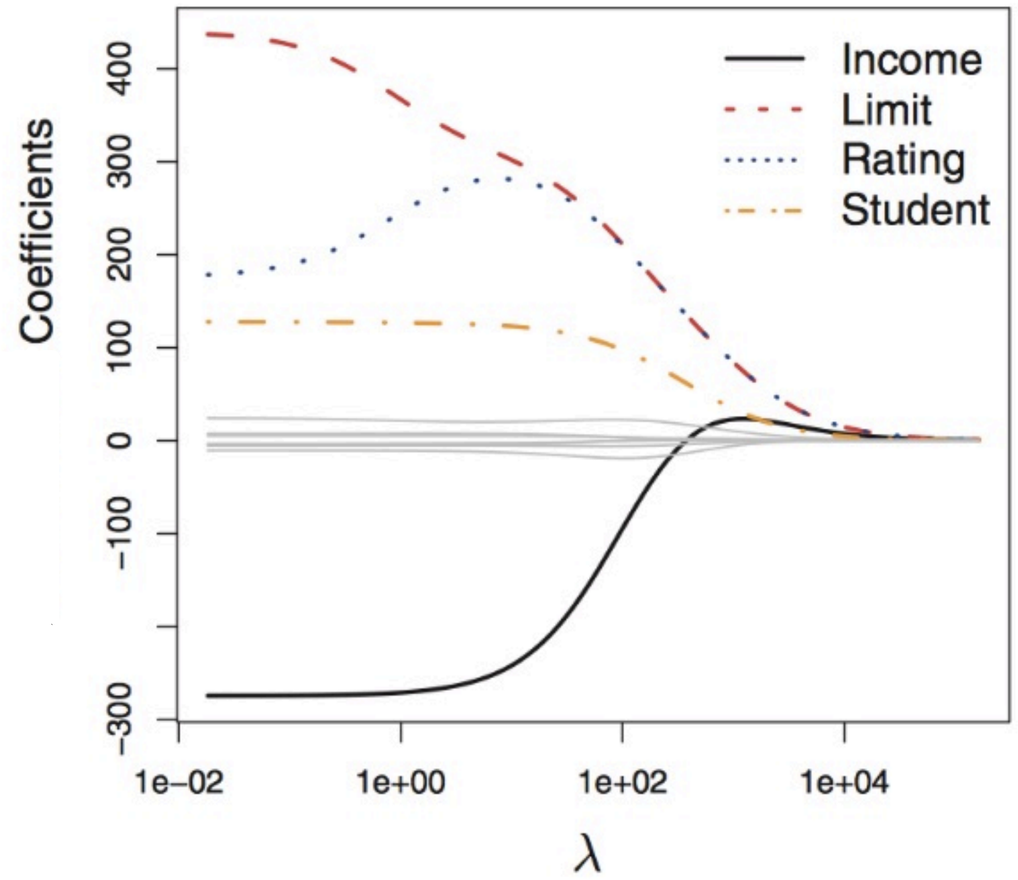
Bias vs Variance vs λ

Same simulated $\sin(2\pi x)$ data



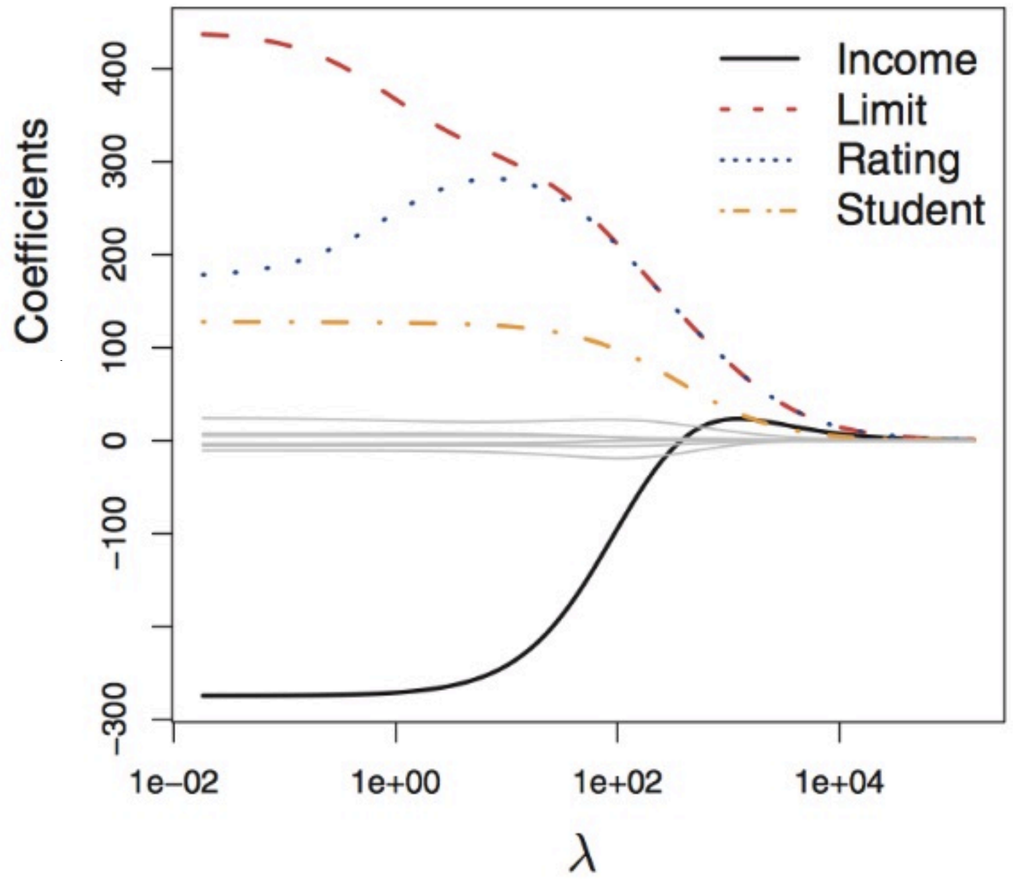
Ridge Regression

How do the coefficients behave as λ increases?



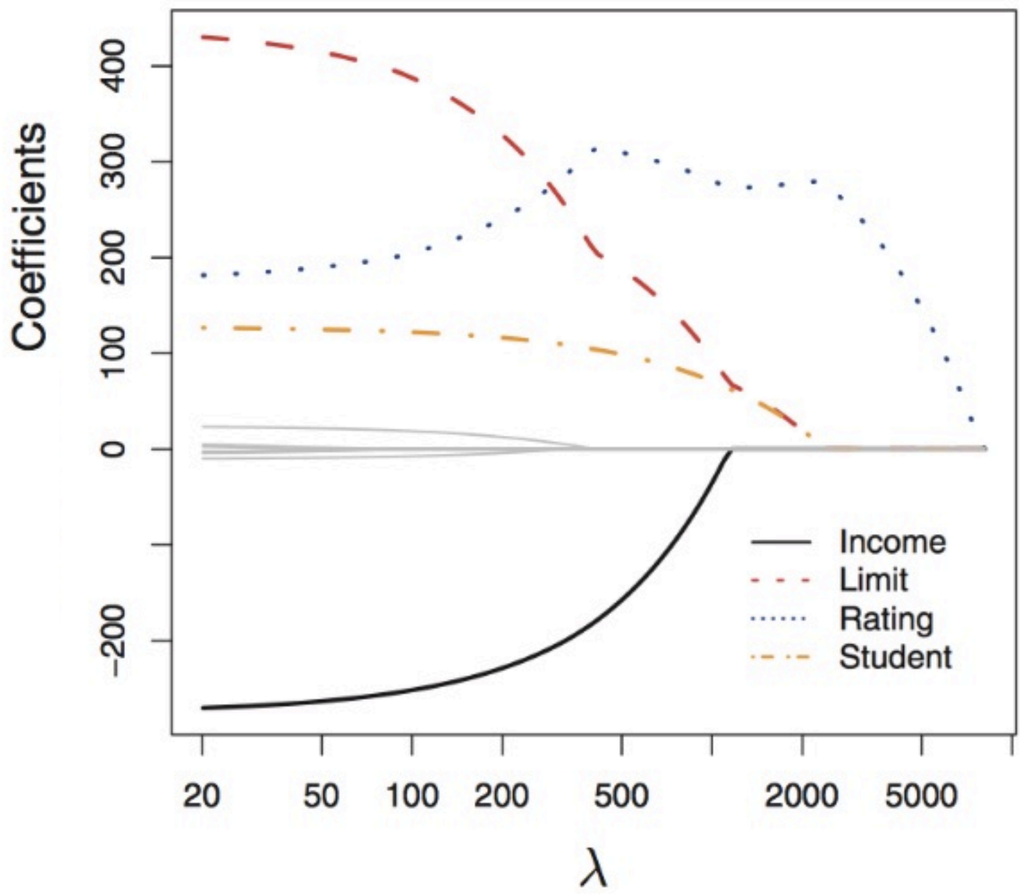
Ridge Regression

They shrink to zero uniformly smoothly.



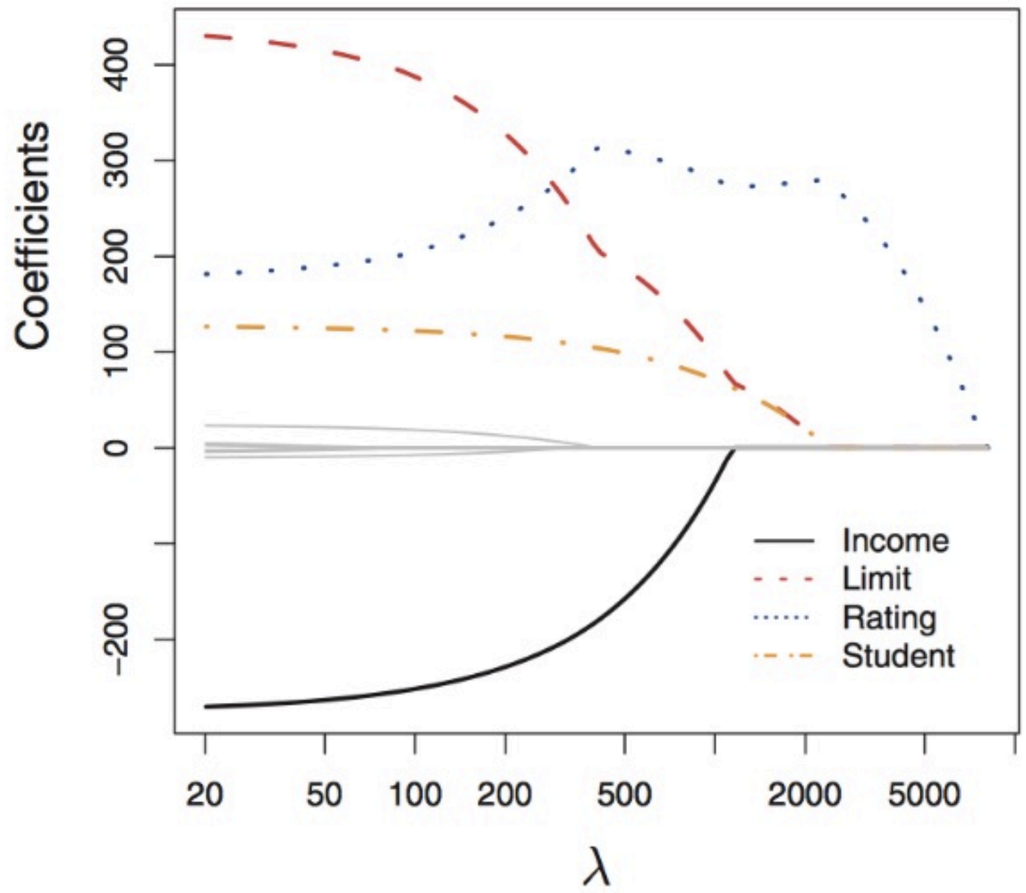
Lasso Regularization

Some coefficients shrink to zero very fast



Lasso Regularization

Feature Selection: Throw out features that shrink to zero fastest!



Ridge vs Lasso Regularization

Why does the choice between the two types of regularization lead to very different behavior?

Several ways to look at it.

- Constrained Minimization
- Look at a simplified case of data
- Prior probabilities on parameters

Ridge vs Lasso: Intuition #1

Consider the minimizer of

$$\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \sum_{k=1}^{D} w_k^2 \quad \text{or} \quad \min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda \sum_{k=1}^{D} |w_k|$$

For each objective function, can show that for a given λ there is an equivalent s such that the usual solution also solves

$$\text{Ridge:} \quad \min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 \quad \text{s.t.} \quad \sum_{k=1}^{D} w_k^2 \leq s$$

$$\text{Lasso:} \quad \min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 \quad \text{s.t.} \quad \sum_{k=1}^{D} |w_k| \leq s$$

Ridge vs Lasso: Intuition #1

Think of the constraint as a budget on the size of the parameters

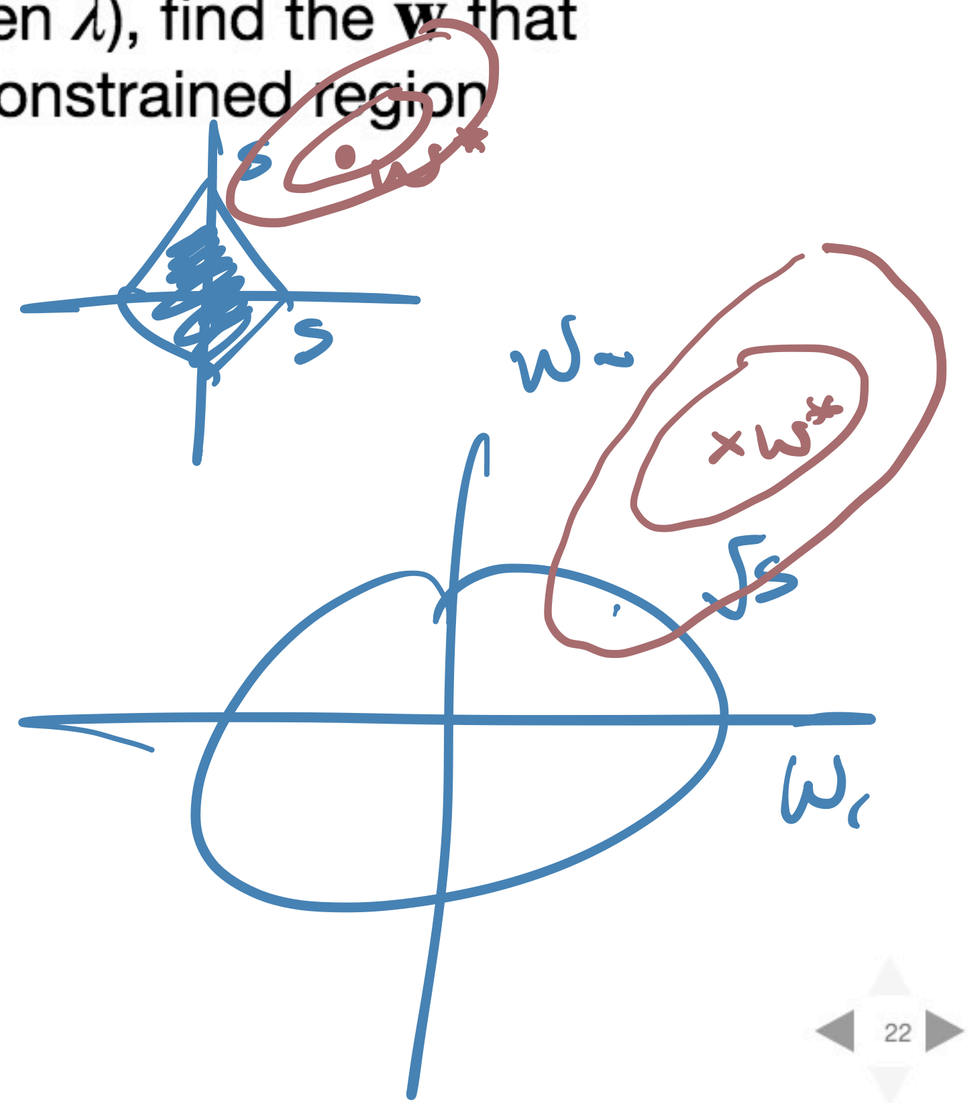
For a given budget s (corresponding to a given λ), find the $\mathbf{w}$ that minimizes the RSS while staying inside the constrained region

Lasso Region for Two Features: Diamond

$$\sum_{k} |w_k| = |w_1| + |w_2| \leq s$$

Ridge Region for Two Features: Circle

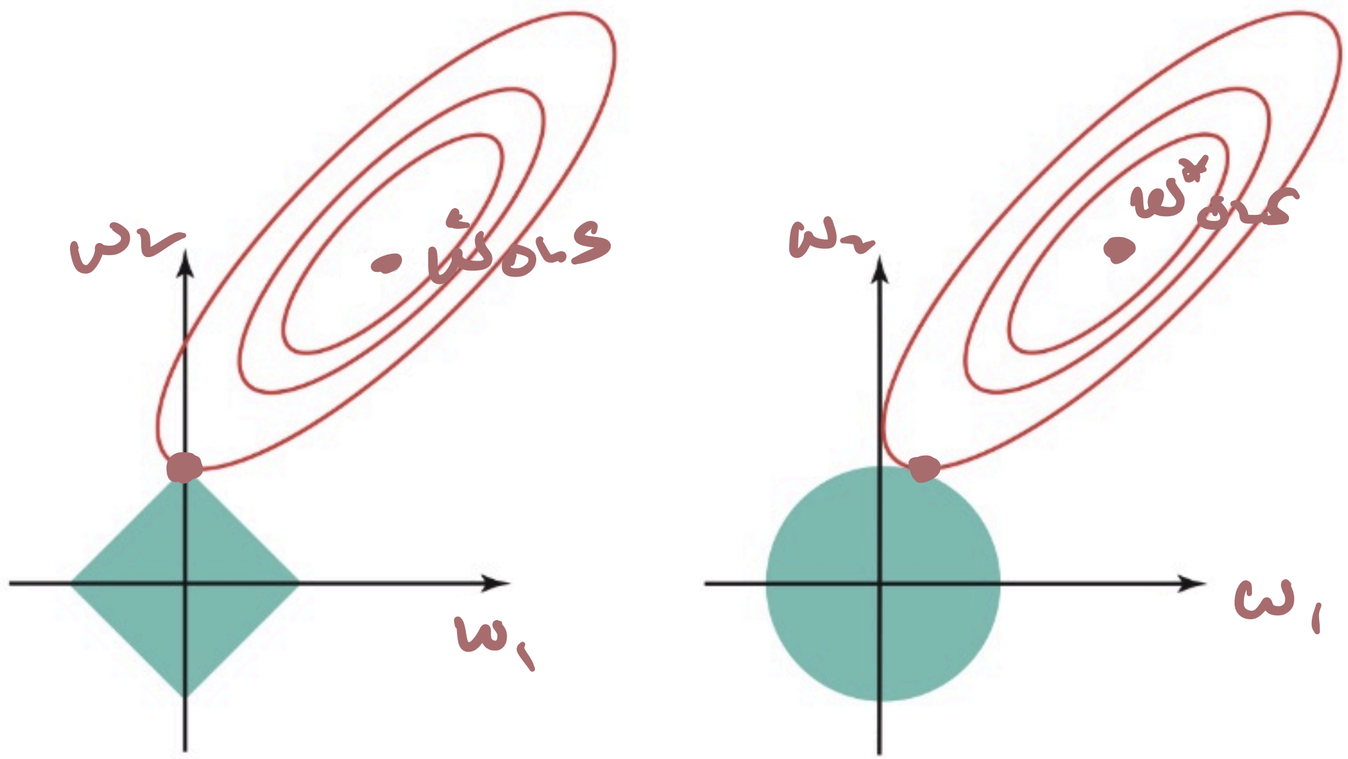
$$\sum_{k=1}^{p} w_k^2 = w_1^2 + w_2^2 \leq s$$



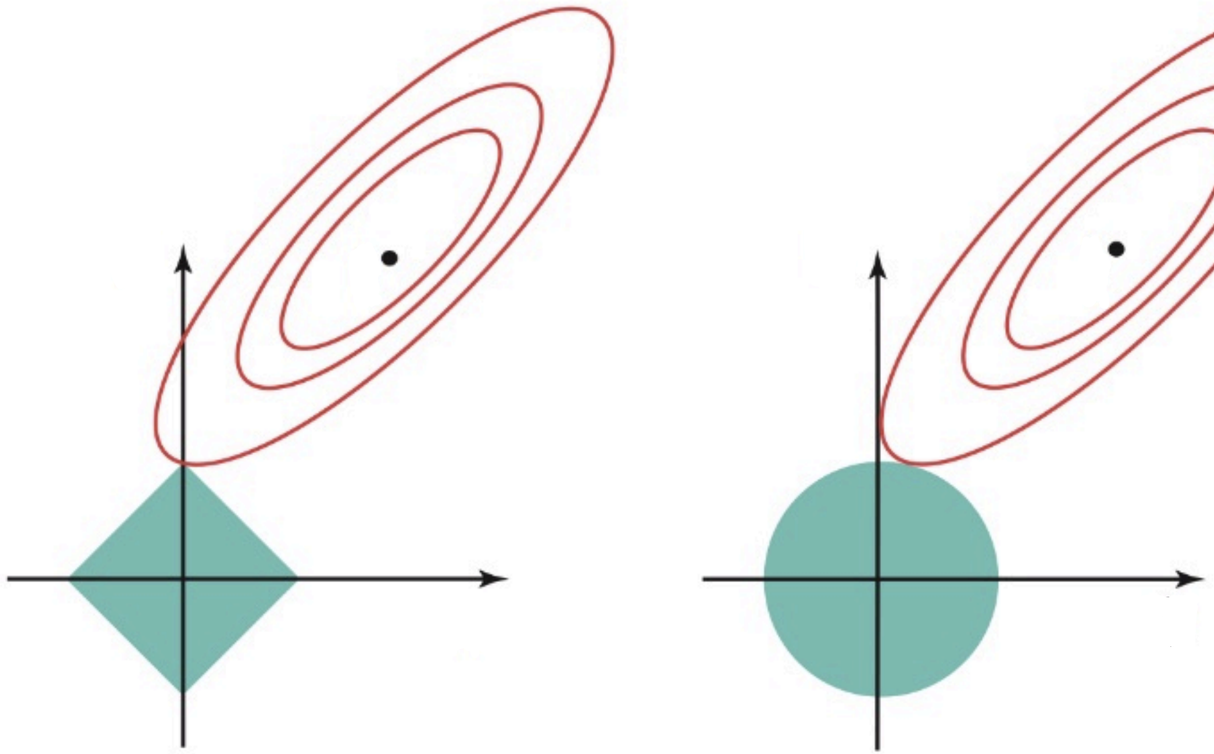
Ridge vs Lasso: Intuition #1

Think of the constraint as a budget on the size of the parameters

For a given budget s (corresponding to a given λ), find the $\mathbf{w}$ that minimizes the RSS while staying inside the constrained region



Ridge vs Lasso: Intuition #1



Minimum more likely to be at point of diamond with Lasso, causing some feature weights to be set to zero.

Ridge vs Lasso: Intuition #2

Consider silly example where $m = D$ and $\mathbf{X}$ is a diagonal matrix with 1's along the diagonal and 0's everywhere else, and assume regression without the bias term

regularized problem reduces to finding $w_1, w_2, \ldots, w_D$ s.t.

$$\text{Ridge}: \quad \sum_{k=1}^{D}(y_k - w_k)^2 + \lambda \sum_{k=1}^{D} w_k^2$$

$$\text{Lasso}: \quad \sum_{k=1}^{D}(y_k - w_k)^2 + \lambda \sum_{k=1}^{D} |w_k|$$

is minimized

Ridge vs Lasso: Intuition #2

$$\text{Ridge}: \quad \sum_{k=1}^{D}(y_k - w_k)^2 + \lambda \sum_{k=1}^{D} w_k^2$$

$$\text{Lasso}: \quad \sum_{k=1}^{D}(y_k - w_k)^2 + \lambda \sum_{k=1}^{D} |w_k|$$

Can show that optimal solutions are

$$\hat{w}_k^R = \frac{y_k}{1+\lambda} \quad \text{and} \quad \hat{w}_k^L = \begin{cases} y_k - \lambda/2 & \text{if } y_k > \lambda/2 \\ y_k + \lambda/2 & \text{if } y_k < -\lambda/2 \\ 0 & \text{if } |y_k| \leq \lambda/2 \end{cases}$$

Ridge vs Lasso: Intuition #2

Can show that optimal solutions are

$$\hat{w}_k^R = \frac{y_k}{1+\lambda} \quad \text{and} \quad \hat{w}_k^L = \begin{cases} y_k - \lambda/2 & \text{if } y_k > \lambda/2 \\ y_k + \lambda/2 & \text{if } y_k < -\lambda/2 \\ 0 & \text{if } |y_k| \leq \lambda/2 \end{cases}$$

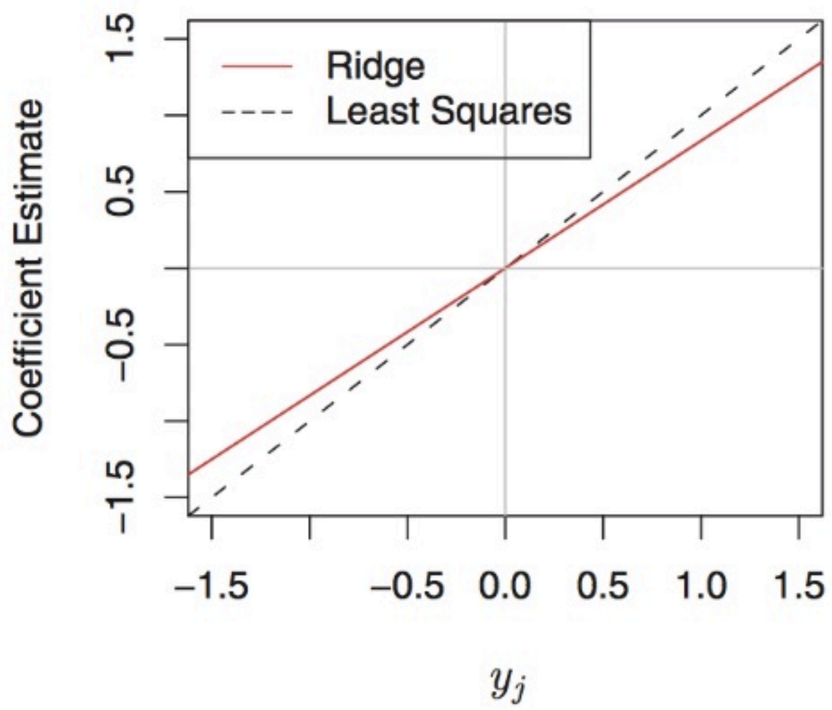
Very different shrinking behavior

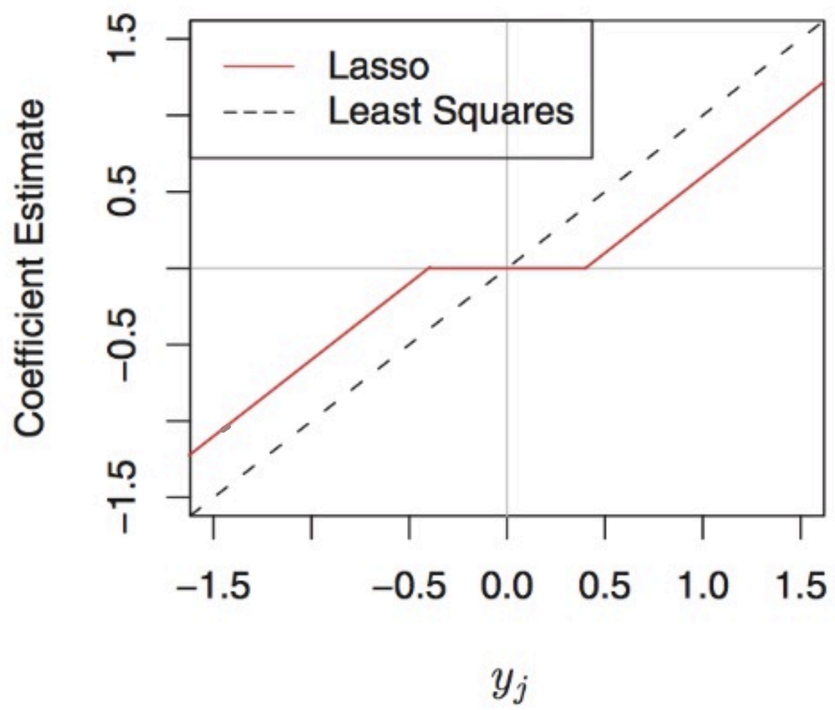
- Ridge regression shrinks every parameter by a constant factor
- Lasso shrinks each coefficient towards zero by constant amount, $\lambda/2$. Coefficients that are less than $\lambda/2$ in absolute value get set to zero.

Ridge vs Lasso: Intuition #2

Can show that optimal solutions are

$$\hat{w}_k^R = \frac{y_k}{1+\lambda} \quad \text{and} \quad \hat{w}_k^L = \begin{cases} y_k - \lambda/2 & \text{if } y_k > \lambda/2 \\ y_k + \lambda/2 & \text{if } y_k < -\lambda/2 \\ 0 & \text{if } |y_k| \leq \lambda/2 \end{cases}$$





Ridge vs Lasso: Intuition #3

Can also get to regularization by putting prior beliefs on parameters

$$p(\mathbf{w} \mid D) \propto p(D \mid \mathbf{w})p(\mathbf{w})$$

Then MAP estimate for $\mathbf{w}$ is $\hat{\mathbf{w}}$ which maximizes posterior

Ridge: Assume Gaussian prior $p(w_j) = \mathcal{N}(w_j \mid 0, \tau^2)$

Lasso: Assume Laplace prior $p(w_j) = \text{Lap}(w_j \mid 0, 1/\lambda) \propto e^{-\lambda|w_j|}$

The Laplace distribution is characterized by a sharp spike at it's mean and heavier tails than the Gaussian distribution

Ridge vs Lasso: Intuition #3

Ridge Regression Details:

$$p(\mathbf{w} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \mathbf{w}) p(\mathbf{w})$$

becomes

$$\begin{aligned} L(\mathbf{w}) &= \mathcal{N}(y \mid \mathbf{w}^T \mathbf{x}, \sigma^2) \mathcal{N}(\mathbf{w} \mid 0, \tau^2) \\ &\propto \prod_{i=1}^{m} \exp\left[\frac{1}{-2\sigma^2}(y_i - \mathbf{w}^T \mathbf{x}_i)^2\right] \prod_{k=1}^{D} \exp\left[\frac{1}{-2\tau^2} w_k^2\right] \end{aligned}$$

Ridge vs Lasso: Intuition #3

Ridge Regression Details:

$$
\begin{aligned}
L(\mathbf{w}) &= \mathcal{N}(y \mid \mathbf{w}^T\mathbf{x}, \sigma^2)\mathcal{N}(\mathbf{w} \mid 0, \tau^2) \\
&\propto \prod_{i=1}^{m} \exp\left[\frac{1}{-2\sigma^2}(y_i - \mathbf{w}^T\mathbf{x}_i)^2\right] \cdot \prod_{k=1}^{D} \exp\left[\frac{1}{-2\tau^2} w_k^2\right]
\end{aligned}
$$

Then the negative log-likelihood is

RSS

RIDGE

$$
\begin{aligned}
NLL(\mathbf{w}) &\propto \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \mathbf{w}^T\mathbf{x})^2 + \frac{1}{2\tau^2}\sum_{k=1}^{D} w_k^2 \\
&\propto \sum_{i=1}^{m}(y_i - \mathbf{w}^T\mathbf{x})^2 + \frac{\sigma^2}{\tau^2}\sum_{k=1}^{D} w_k^2 \\
&= \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda\sum_{k=1}^{D} w_k^2
\end{aligned}
$$

Ridge vs Lasso: Intuition #3

Lasso Regression Details:

$$
\begin{aligned}
L(\mathbf{w}) &= \mathcal{N}(y \mid \mathbf{w}^T\mathbf{x}, \sigma^2)\mathrm{Lap}(\mathbf{w} \mid 0, 1/\lambda) \\
&\propto \prod_{i=1}^{m} \exp\left[\frac{1}{-2\sigma^2}(y_i - \mathbf{w}^T\mathbf{x}_i)^2\right] \prod_{k=1}^{D} \exp[-\lambda\,|w_k|\ ]
\end{aligned}
$$

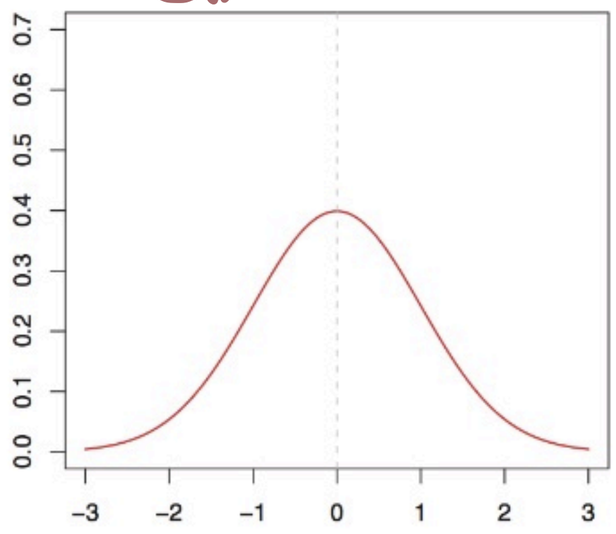
Then the negative log-likelihood is

$$
\begin{aligned}
NLL(\mathbf{w}) &\propto \tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^{m}(y_i - \mathbf{w}^T\mathbf{x})^2 + \lambda\textstyle\sum_{k=1}^{D}|w_k| \\
&\propto \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda\textstyle\sum_{k=1}^{D}|w_k|
\end{aligned}
$$

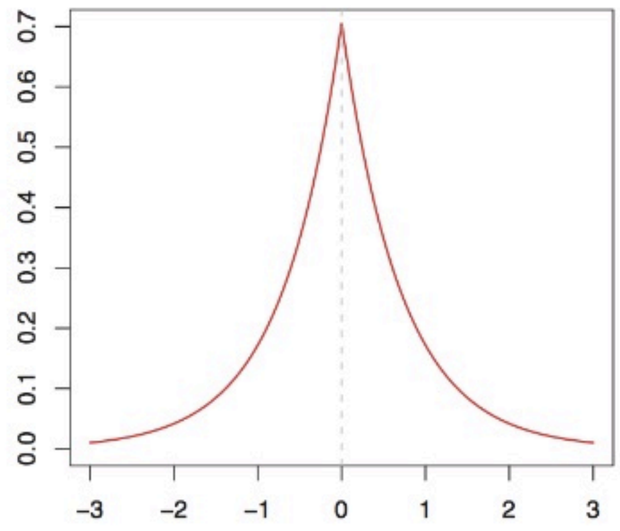
Ridge vs Lasso: Intuition #3

Comparing the Priors:

GAUSS



LAPLACE



Lasso's prior peaked at 0 means expect many params to be zero

Ridge's prior flatter and fatter around 0 means we expect many coefficients to be smallish

Wrap-Up

Question: When should you use Regularization with Regression?

Answer: Always! It helps fight variance and overfitting

In particular if you have $D > m$ you should definitely use Regularization.

Standard LR does not have unique solution. Regularization enforces unique solution.

Further, if $D > m$ it seems weird to expect interaction between response and *all* features.

Regularization sets unimportant parameters to zero (or close to it)

Next Time

- Learning Theory!

Acknowledgements

Many of the slides in this presentation were adopted from Jordan Boyd-Graber and Lauren Hannah

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

In Class

In Class

In Class

In Class