Lasso Regression:

Regularization for feature selection

CS229: Machine Learning

Carlos Guestrin
Stanford University

Slides include content developed by and co-developed with Emily Fox

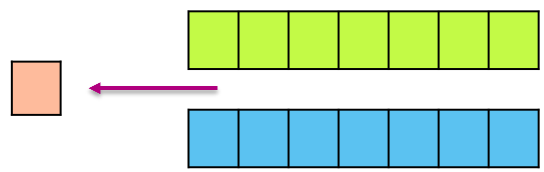
Feature selection task

Why might you want to perform feature selection?

Efficiency:

- If size($\mathbf{w}$) = 100B, each prediction is expensive
- If $\hat{\mathbf{w}}$ sparse , computation only depends on # of non-zeros

$$\hat{y}_i = \sum_{\hat{w}_j \neq 0} \hat{w}_j h_j(\mathbf{x}_i)$$



Interpretability:

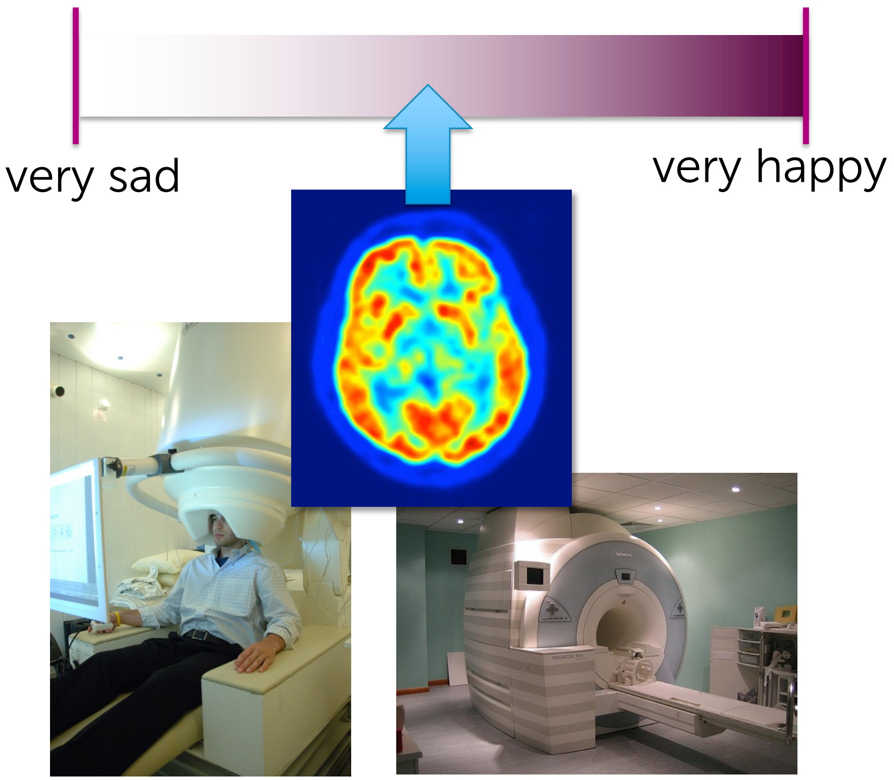
- Which features are relevant for prediction?

Sparsity: Housing application

Lot size
Single Family
Year built
Last sold price
Last sale price/sqft
Finished sqft
Unfinished sqft
Finished basement sqft
floors
Flooring types
Parking type
Parking amount
Cooling
Heating
Exterior materials
Roof type
Structure style

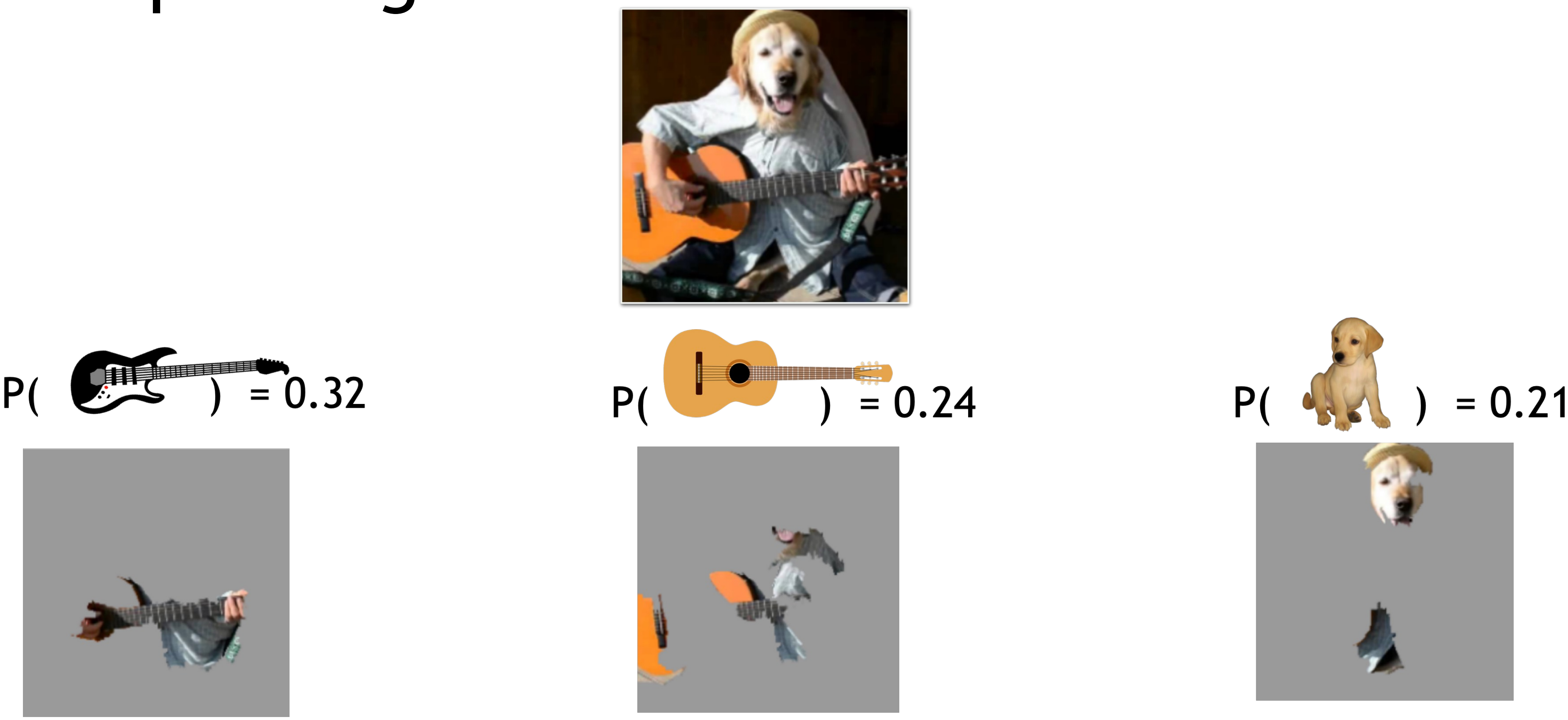
Dishwasher
Garbage disposal
Microwave
Range / Oven
Refrigerator
Washer
Dryer
Laundry location
Heating type
Jetted Tub
Deck
Fenced Yard
Lawn
Garden
Sprinkler System
⋮

Sparsity: Reading your mind



Activity in which brain regions can predict happiness?

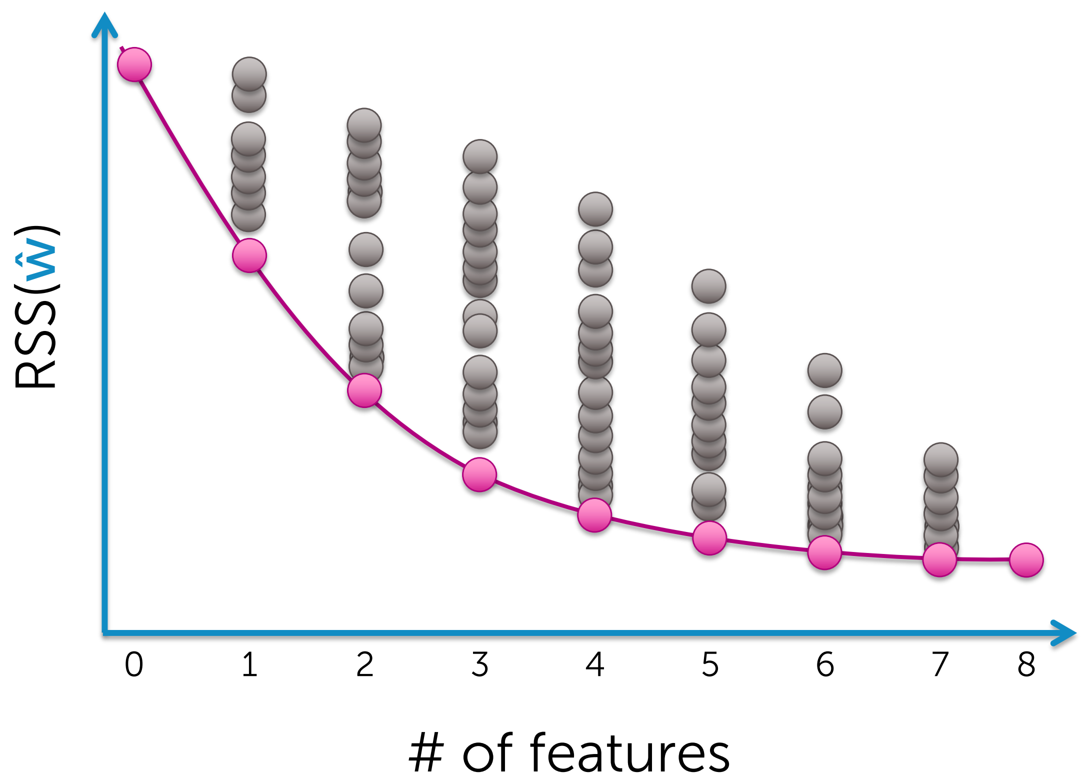
Explaining Predictions



“Why should I trust you?”: Explaining the Predictions of Any Classifier. Ribeiro, Singh & G. KDD 16

Option 1: All subsets or greedy variants

Find best model of for each size



- # bedrooms
- # bathrooms
- sq.ft. living
- sq.ft. lot
- floors
- year built
- year renovated
- waterfront

Complexity of "all subsets"

$$
\left.\begin{array}{ll}
y_i = \varepsilon_i & [0\ 0\ 0\ \ldots\ 0\ 0\ 0] \\
y_i = w_0 h_0(\mathbf{x}_i) + \varepsilon_i & [1\ 0\ 0\ \ldots\ 0\ 0\ 0] \\
y_i = w_1 h_1(\mathbf{x}_i) + \varepsilon_i & [0\ 1\ 0\ \ldots\ 0\ 0\ 0] \\
\vdots & \vdots \\
y_i = w_0 h_0(\mathbf{x}_i) + w_1 h_1(\mathbf{x}_i) + \varepsilon_i & [1\ 1\ 0\ \ldots\ 0\ 0\ 0] \\
\vdots & \vdots \\
y_i = w_0 h_0(\mathbf{x}_i) + w_1 h_1(\mathbf{x}_i) + \ldots + w_D h_D(\mathbf{x}_i) + \varepsilon_i & [1\ 1\ 1\ \ldots\ 1\ 1\ 1]
\end{array}\right\} 2^D
$$

$2^8 = 256$
$2^{30} = 1{,}073{,}741{,}824$
$2^{1000} = 1.071509 \times 10^{301}$
2^{100B} = HUGE!!!!!!

Typically, computationally infeasible

Greedy algorithms

Forward stepwise:
Starting from simple model and iteratively add features most useful to fit

Backward stepwise:
Start with full model and iteratively remove features least useful to fit

Combining forward and backward steps:
In forward algorithm, insert steps to remove features no longer as important

Lots of other variants, too.

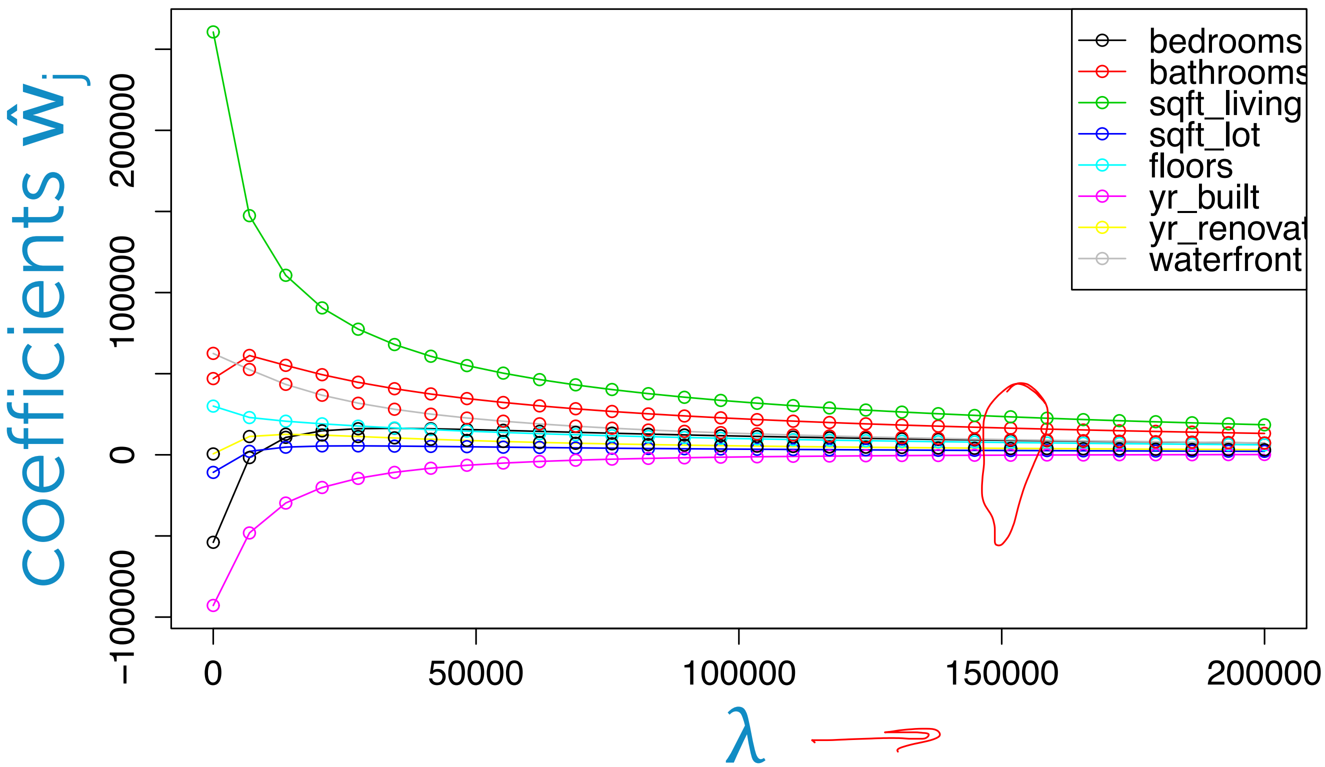
Option 2: Regularize

Ridge regression: L_2 regularized regression

Total cost =

$$\underbrace{\text{measure of fit}}_{\text{RSS}(\mathbf{w})} + \lambda \underbrace{\text{measure of magnitude of coefficients}}_{\|\mathbf{w}\|_2^2 = w_0^2 + \ldots + w_D^2}$$

Coefficient path – ridge



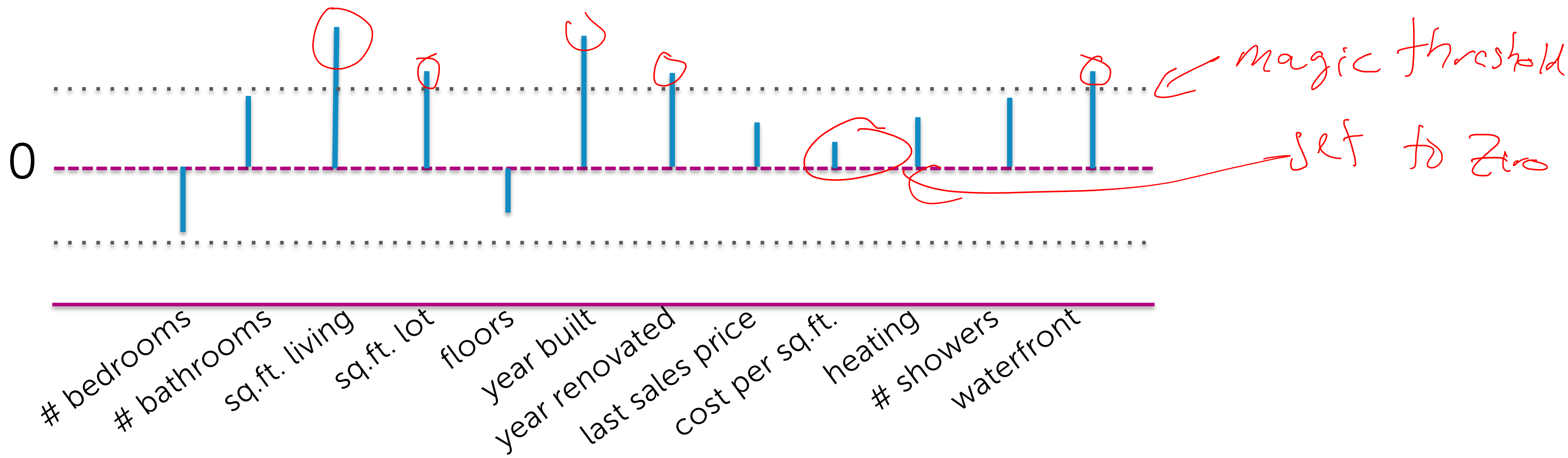
Using regularization for feature selection

Instead of searching over a **discrete** set of solutions, can we use regularization?

- Start with full model (all possible features)
- "Shrink" some coefficients *exactly* to 0
  - i.e., knock out certain features
- Non-zero coefficients indicate "selected" features

Thresholding ridge coefficients?

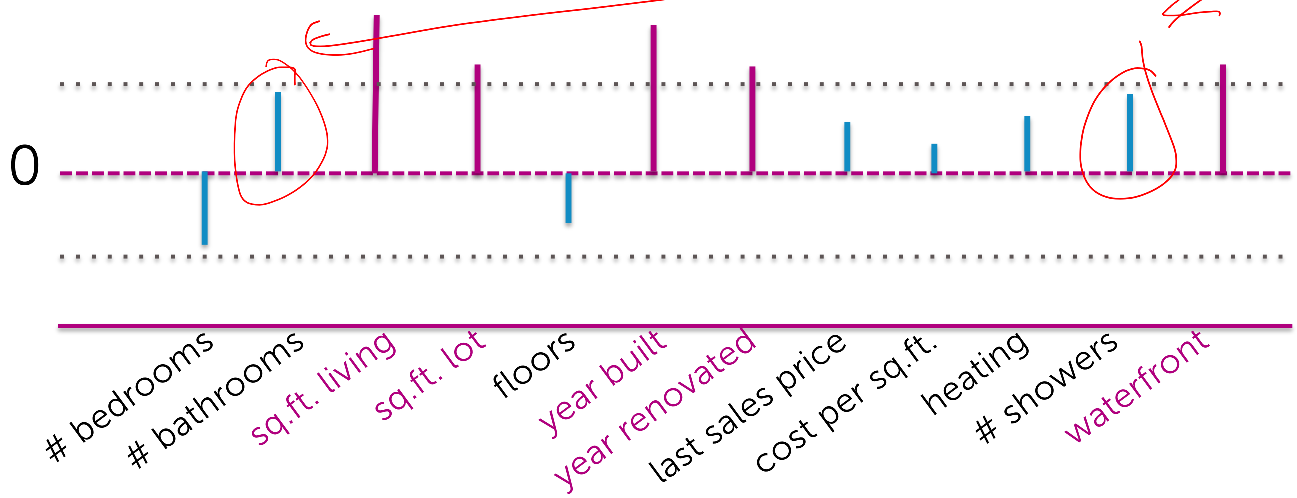
Why don't we just set small ridge coefficients to 0?



Thresholding ridge coefficients?

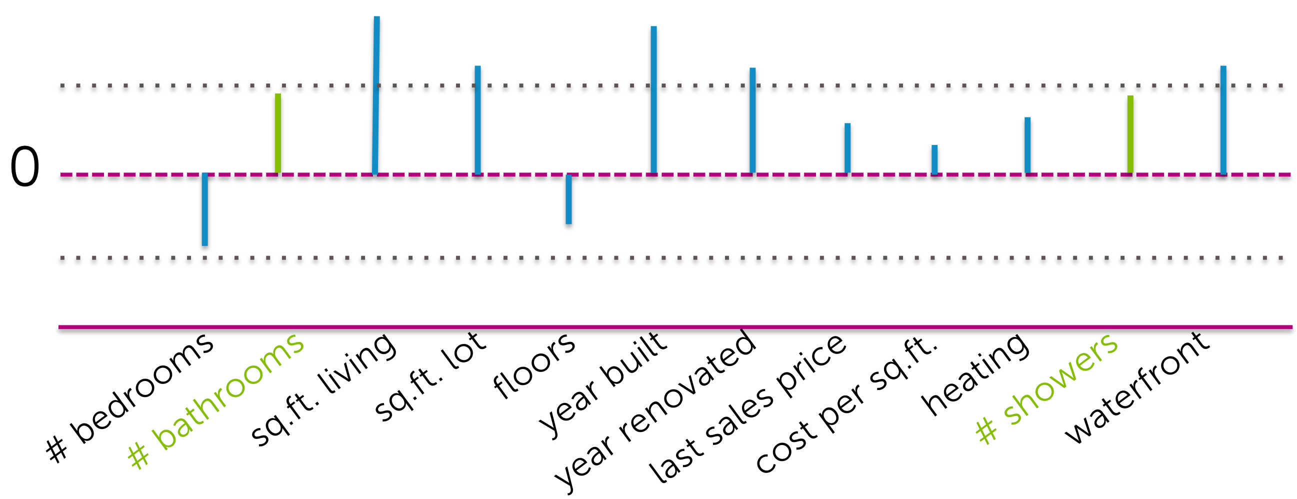
Selected features for a given threshold value

bathrooms
Seems
important



Thresholding ridge coefficients?

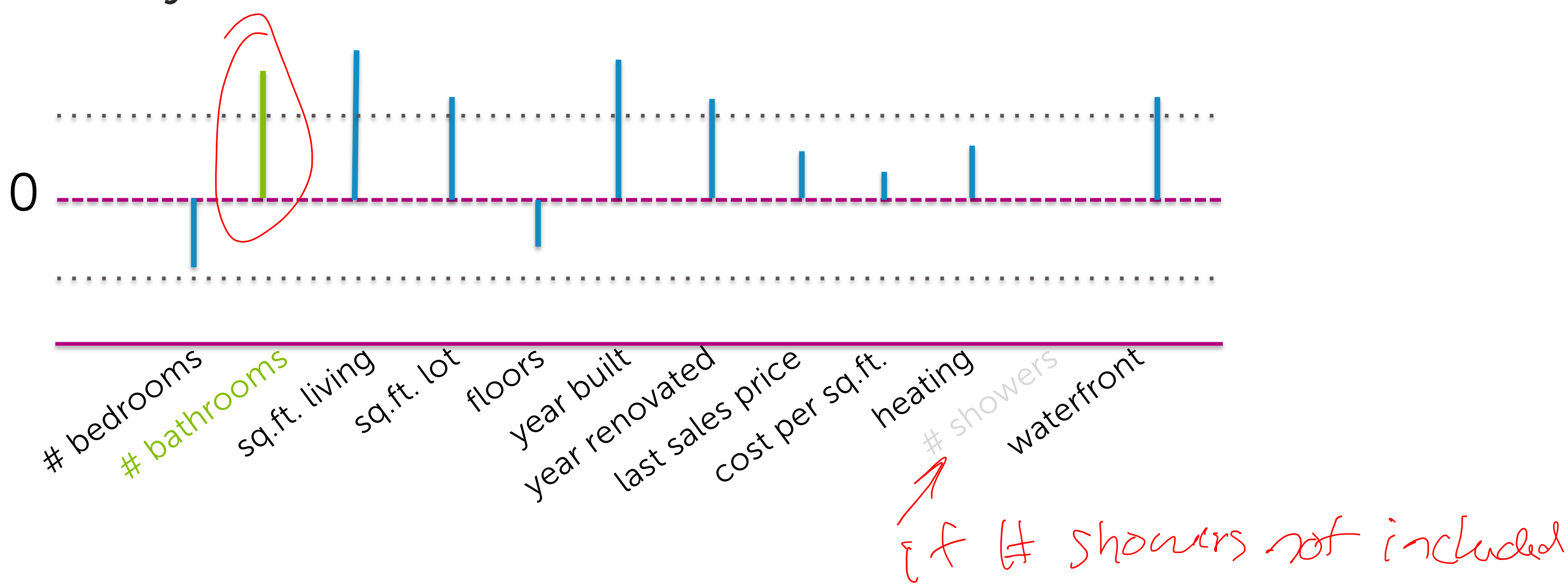
Let's look at two related features…



Nothing measuring bathrooms was included!

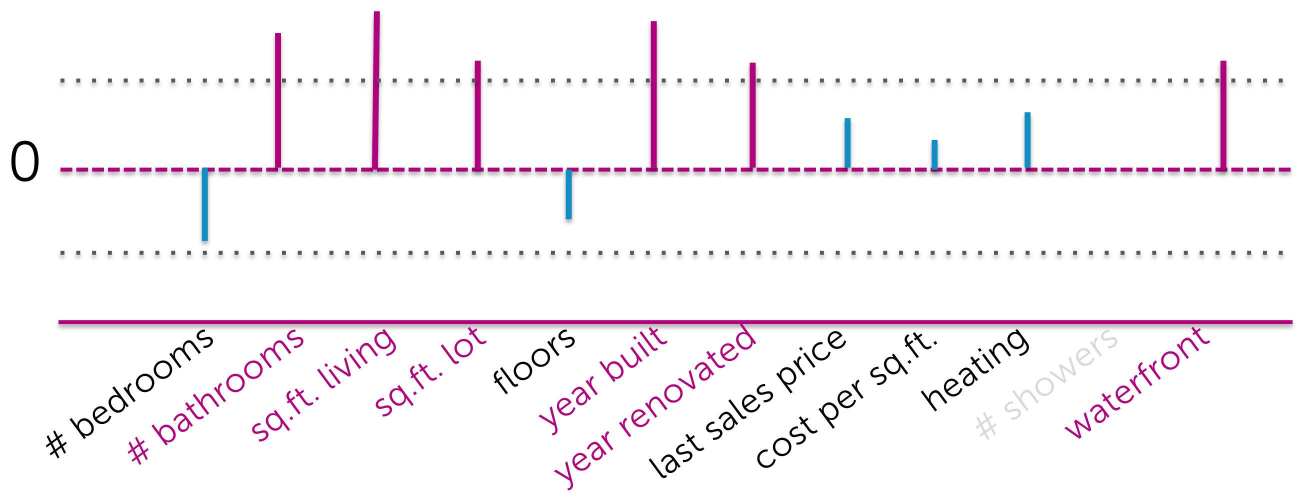
Thresholding ridge coefficients?

If only one of the features had been included…



Thresholding ridge coefficients?

Would have included bathrooms in selected model



Can regularization lead directly to sparsity?

Try this cost instead of ridge…

Total cost =

$$\underbrace{\text{measure of fit}}_{\text{RSS}(\mathbf{w})} + \lambda \underbrace{\text{measure of magnitude of coefficients}}_{||\mathbf{w}||_1 = |w_0| + \ldots + |w_D|}$$

Leads to **sparse** solutions!

Lasso regression
(a.k.a. L_1 regularized regression)

Lasso regression: L_1 regularized regression

Just like ridge regression, solution is governed by a continuous parameter λ

$$\text{RSS}(\mathbf{w}) + \lambda\|\mathbf{w}\|_1$$

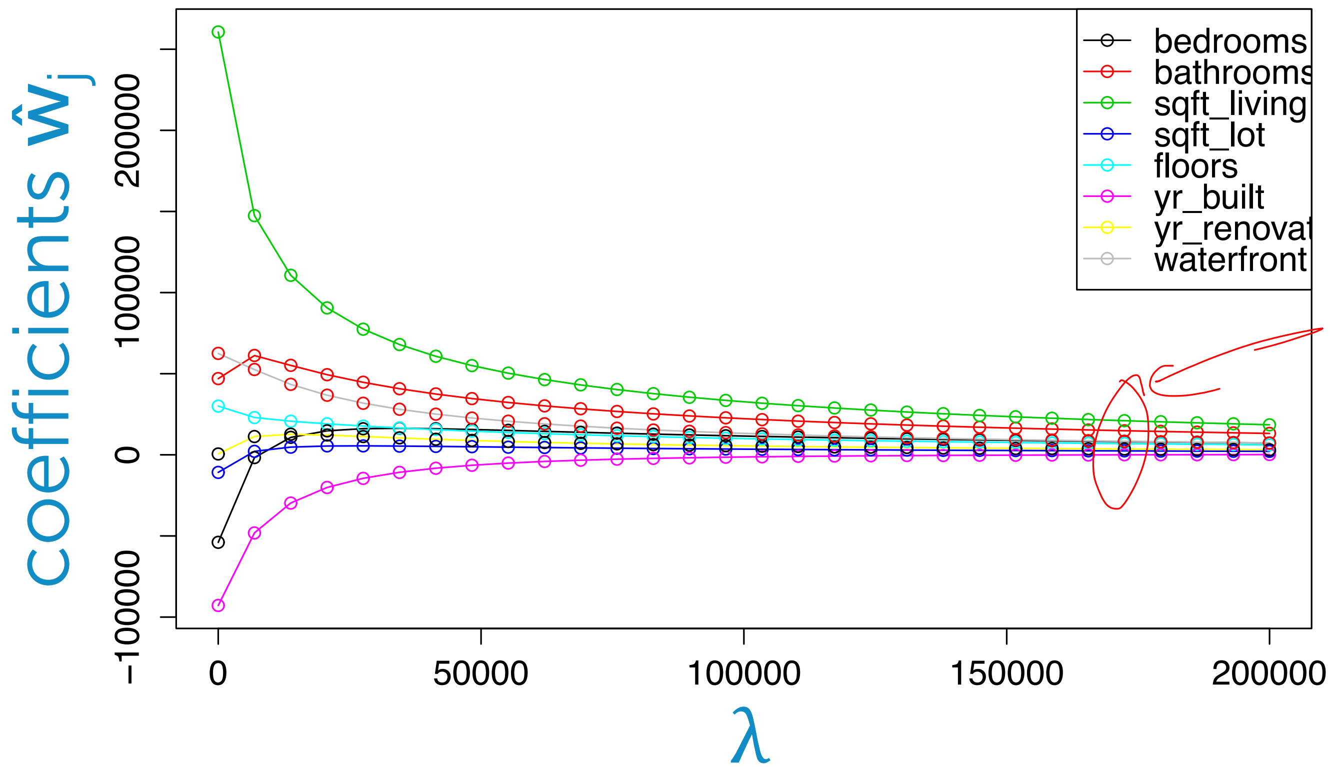
tuning parameter = balance of fit and **sparsity**

If λ=0: $\hat{w}_{Lasso} = \hat{w}_{RSS}$

If λ=∞: $\hat{w}_{Lasso} = 0$

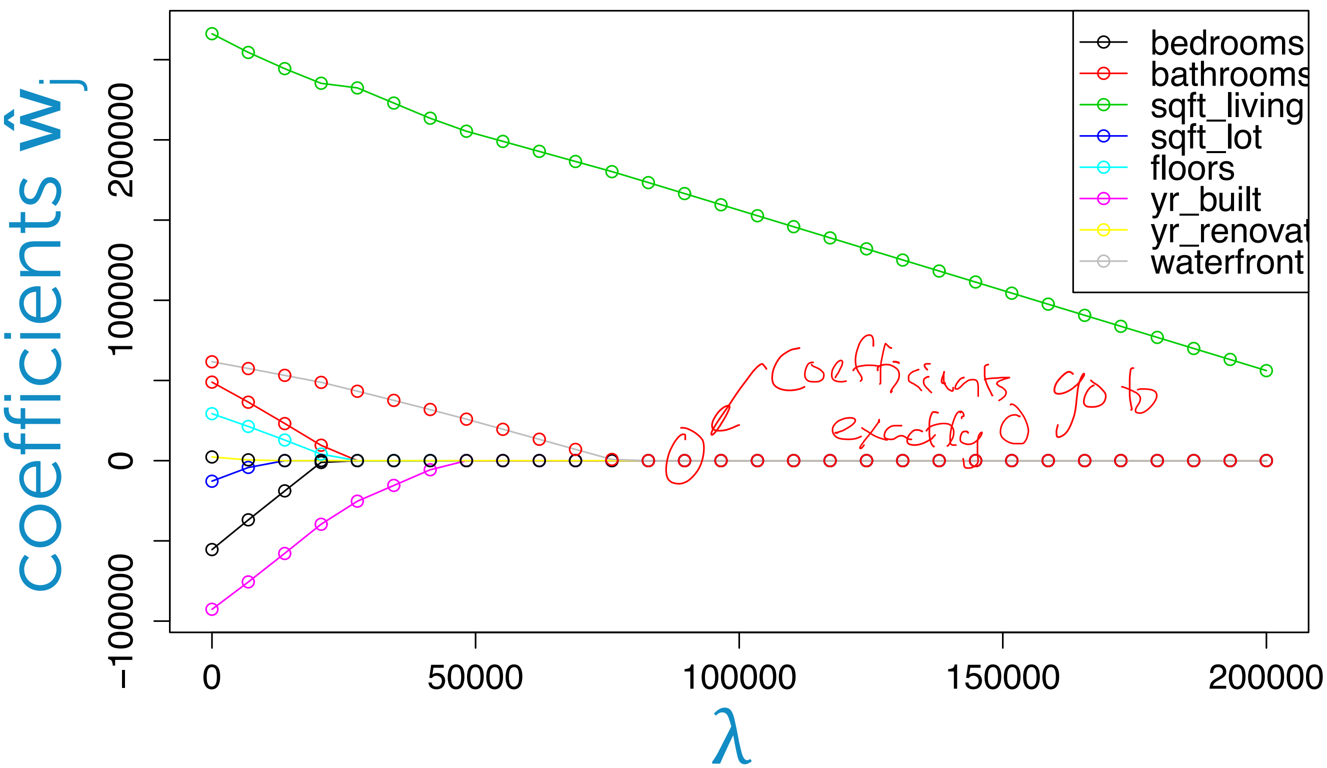
If λ in between: $\|\hat{w}_{Lasso}\|_1 < \|\hat{w}_{RSS}\|_1$

Coefficient path – ridge



Small but not zero

Coefficient path – lasso

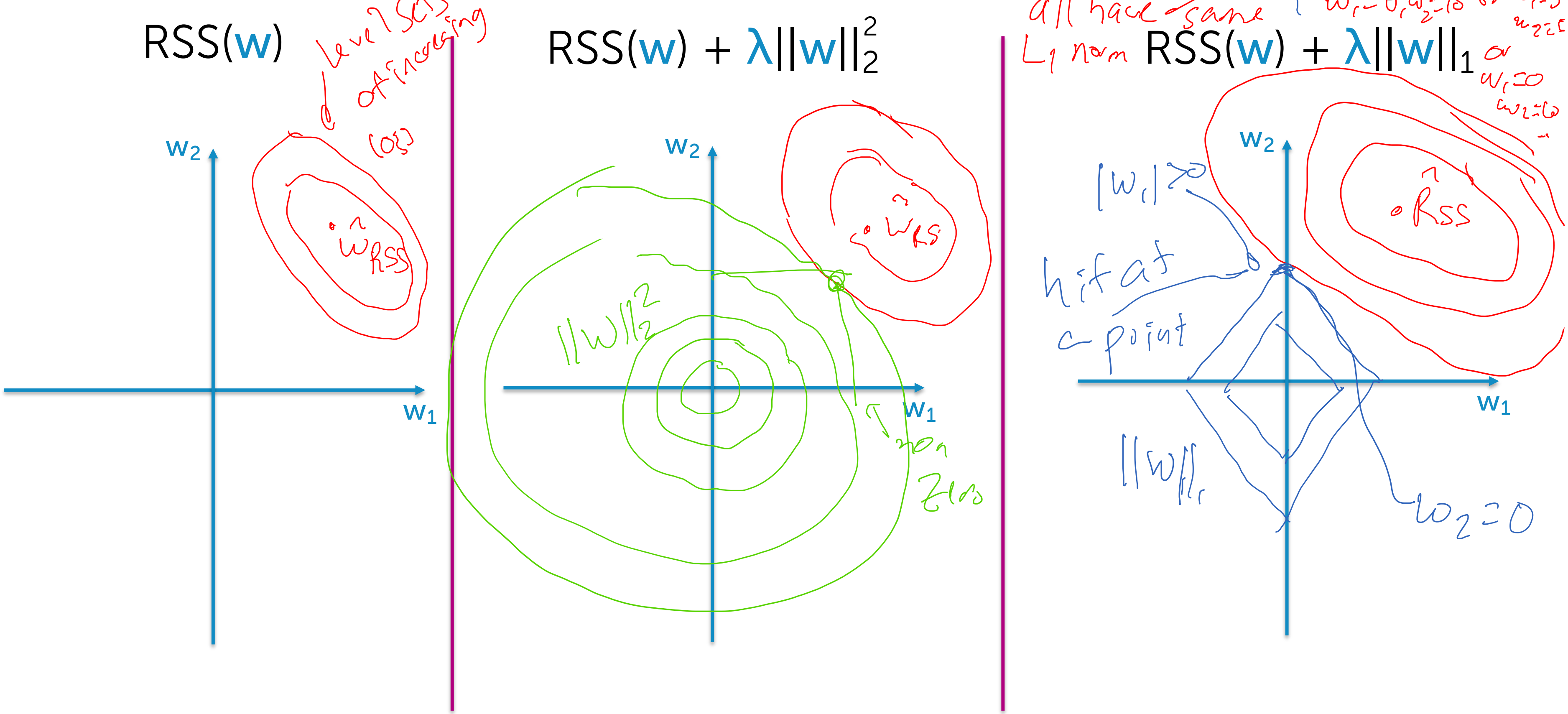


what happens, say, with identical features $x^{(1)} = x^{(2)}$?

Solution is indifferent: if weight is 10, could be $\{w_1=0, w_2=10$ or $w_1=5, w_2=5$ or $w_1=10, w_2=0$

all have same L_1 norm

Intuitive difference between Lasso and Ridge



Practical concerns with lasso

Debiasing lasso

Lasso shrinks coefficients relative to LS solution

→ more bias, less variance

Can reduce bias as follows:

1. Run lasso to select features
2. Run least squares regression with only selected features

"Relevant" features no longer shrunk relative to LS fit of same reduced model

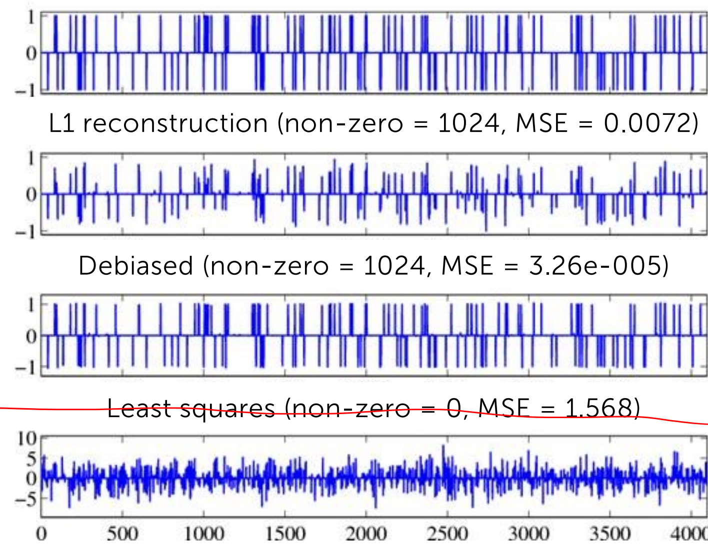


Figure used with permission of Mario Figueiredo (*captions modified to fit course*)

true

L1 reconstruction (smaller than truth, because of penalty)

ridge solution not sparse

Issues with standard lasso objective

1. With group of highly correlated features, lasso tends to select amongst them arbitrarily
   - Often prefer to select all together

2. Often, empirically ridge has better predictive performance than lasso, but lasso leads to sparser solution

Elastic net aims to address these issues
- hybrid between lasso and ridge regression
- uses L_1 and L_2 penalties

See Zou & Hastie '05 for further discussion

Summary for feature selection and lasso regression

Impact of feature selection and lasso

Lasso has changed machine learning, statistics, & electrical engineering

But, for feature selection in general, be careful about interpreting selected features

- selection only considers features included
- sensitive to correlations between features
- result depends on algorithm used
- there are theoretical guarantees for lasso under certain conditions

What you can do now…

- Describe "all subsets" and greedy variants for feature selection
- Analyze computational costs of these algorithms
- Formulate lasso objective
- Describe what happens to estimated lasso coefficients as tuning parameter λ is varied
- Interpret lasso coefficient path plot
- Contrast ridge and lasso regression
- Implement K-fold cross validation to select lasso tuning parameter λ