Training Models 3: Regularization

Glenn Bruns
CSUMB

Learning outcomes

After this lecture you should be able to:

1. Define regularization
2. Explain how regularization relates to the bias/variance tradeoff
3. Explain what regularization has to do with optimization in ML
4. Apply ridge regression, lasso regression, early stopping
5. Explain how regularization relates to feature selection

Recall: bias/variance tradeoff

Variance: sensitivity of a machine learning algorithm to a particular training set

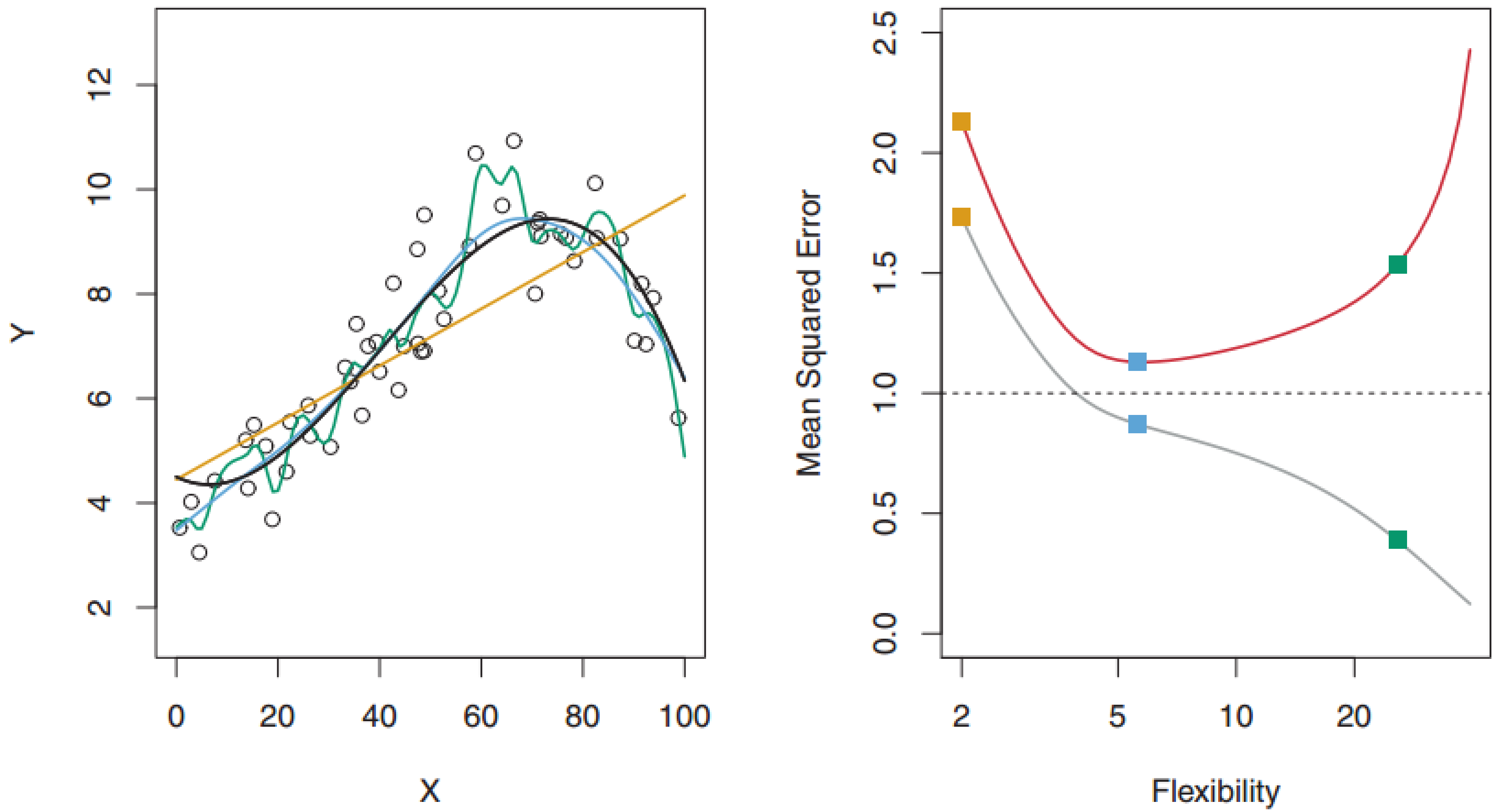
- we don't want small changes in the training set to have a big impact on our model

Bias: error introduced by simplifying assumptions of a model

- we want a model that's flexible enough to capture our real-world problem

A model's "generalization error" can be decomposed into bias, variance, and irreducible error.

Visualizing the bias/variance tradeoff



source: An Introduction to Statistical Learning, James et al

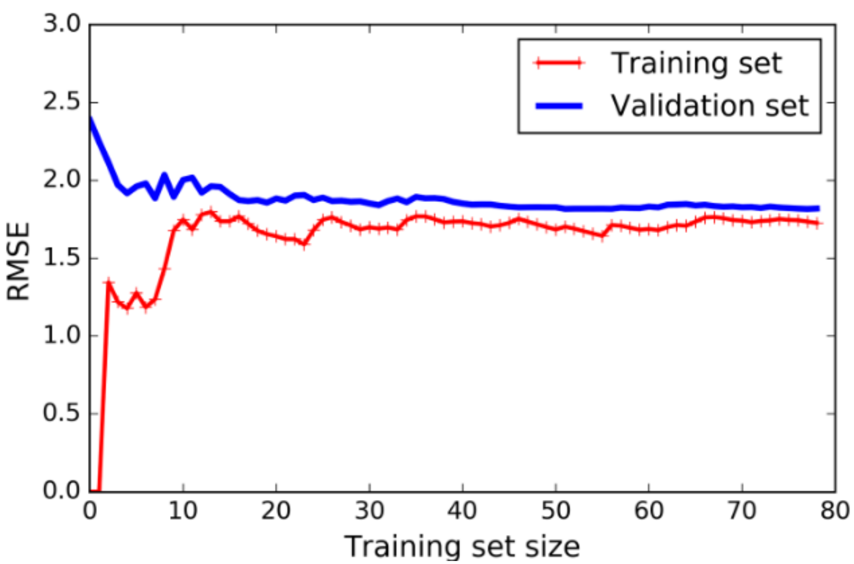
Recall: Learning curves

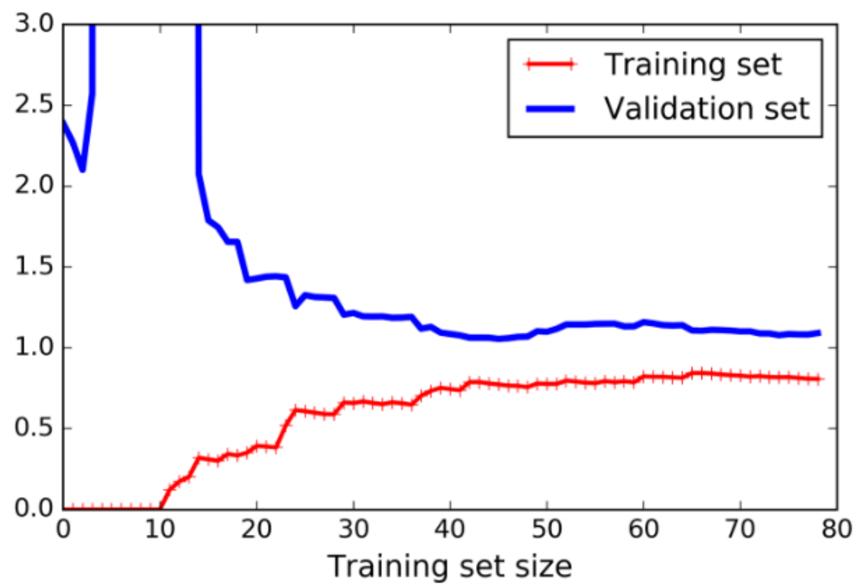
what do we expect in a high bias situation?

- high training and test error
- even with lots of training we won't expect low test error

what do we expect in a high variance situation?

- low training error, high test error with a small training set
- with lots of training we'll eventually get low test error





How to combat overfitting?

Constrain the model; known as regularization

- for example, reduce number of polynomial degrees
- in kNN, make k larger
- in linear regression, "shrink" the coefficient estimates towards zero
  - ridge regression and lasso regression
  - these are known as shrinkage methods

Ridge regression

Recall: in linear regression we use this cost function:

$$MSE(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\theta^T \cdot \mathbf{x}^{(i)} - y^{(i)}\right)^2$$

Ridge regression adds a term that penalizes large coefficients:

$$MSE(\theta) + \frac{\alpha}{2}\sum_{i=1}^{n}\theta_i^2$$

α is the regularization parameter, with $\alpha \geq 0$.

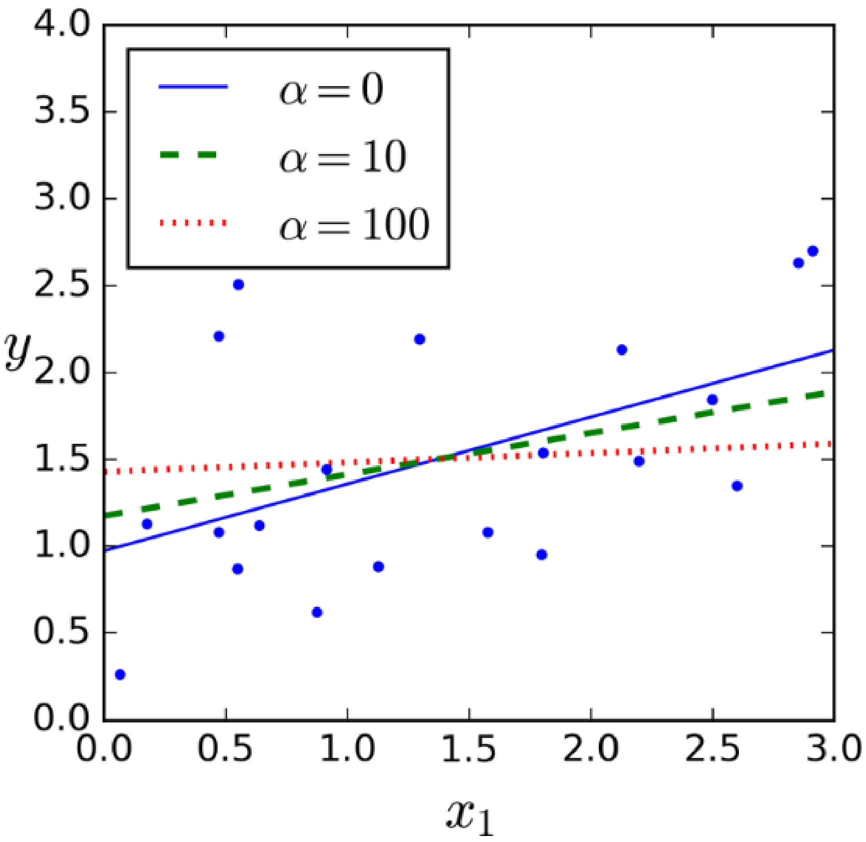
What happens when α is 0?

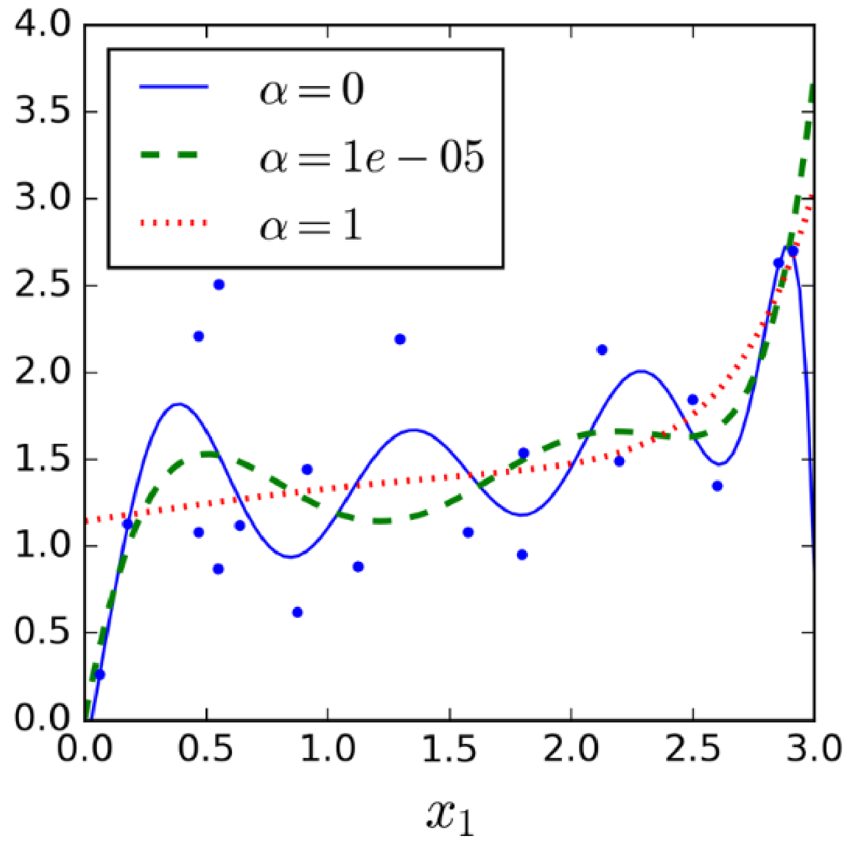
What happens when it is very large?

Ridge regression example

Simple linear regression

Polynomial regression, with degree up to 10

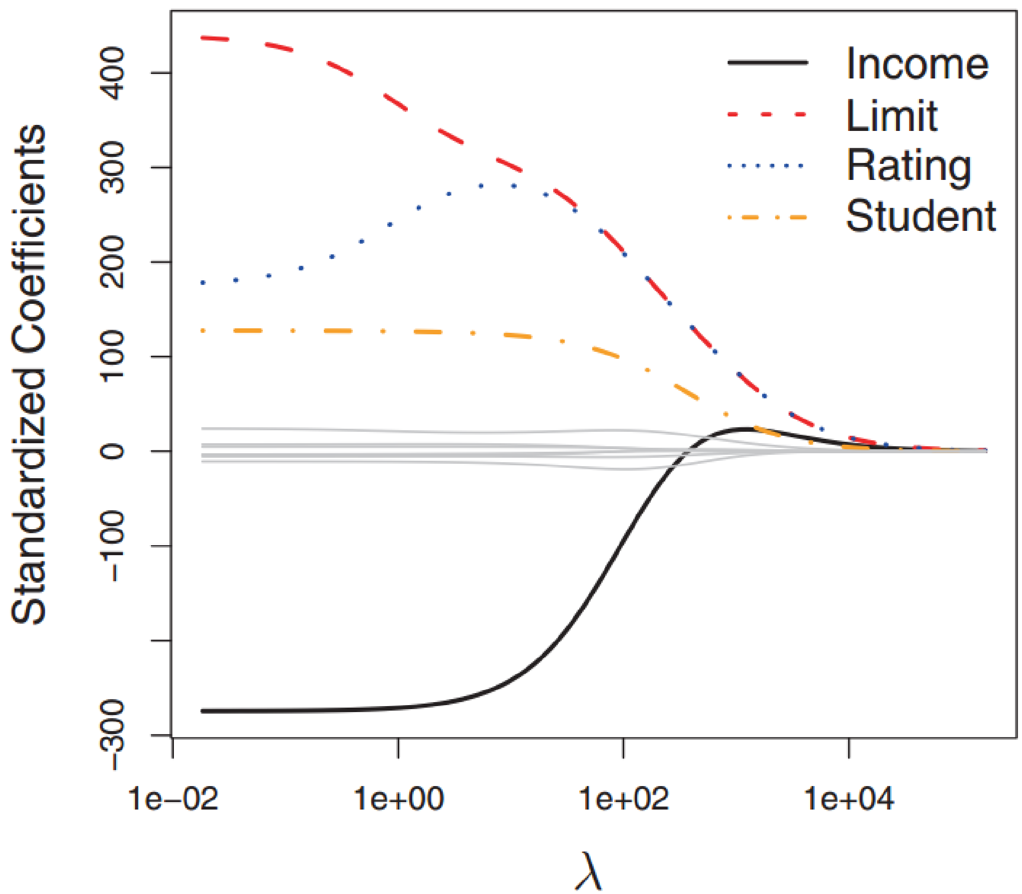




source: Géron

Another example

Ridge regression on the Credit data set.



Question:
What do you observe here?

source: Intro to Statistical Learning, James et al

Lasso regression

Very similar to ridge regression. This is the ridge regression cost function:

$$MSE(\theta) + \frac{\alpha}{2} \sum_{i=1}^{n} \theta_i^2$$

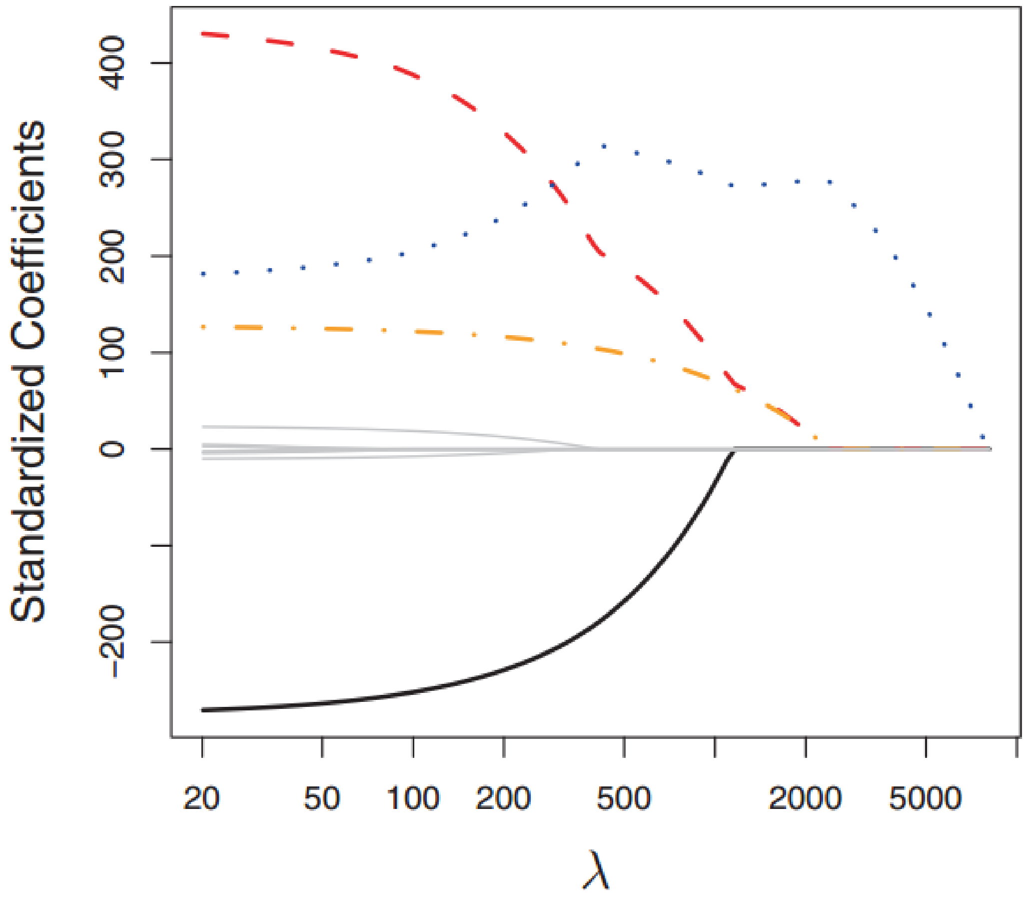
This is the lasso regression cost function:

$$MSE(\theta) + \frac{\alpha}{2} \sum_{i=1}^{n} |\theta|$$

The change is small. Surprisingly, lasso tends to zero the weights of less important features.

Lasso example

Lasso regression on the Credit data set.



Features "drop out" as the regularization parameter is increases.

Beautiful idea: the lasso is a cheap way to perform feature selection.

source: Intro to Statistical Learning, James et al

Scikit-Learn

Ridge:

```
from sklearn.linear_model import Ridge
```

- Ridge class for closed-form solution, or
- SGDRegressor for stochastic gradient descent (with "l2" penalty)

Lasso:

```
from sklearn.linear_model import Lasso
```

- Lasso class for closed-form solution, or
- SGDRegressor for stochastic gradient descent (with "l1" penalty)

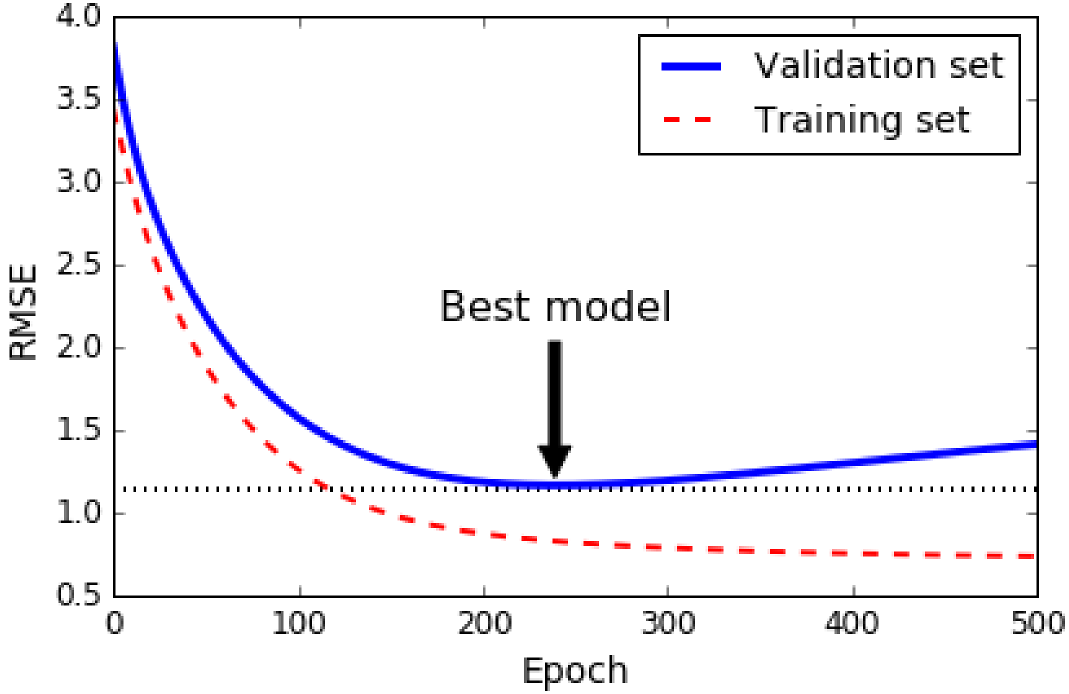
See text for more details.

Early stopping

Idea: use gradient descent, but stop when validation error (aka test error) is minimized.

Pseudocode:

```
for epoch in range(num_iterations):
    update theta
    compute MSE (or other cost measure)
    if this MSE is minimal:
        record theta
```



Summary

1. Regularization is a method to reduce overfitting
2. We explore three methods:
   - Ridge regression
   - Lasso regression
   - Early stopping

   'elastic net' is a combination of these two
3. Ridge and Lasso work by changing the optimization problem to be solved
4. Lasso regression is a feature selection method, too!