This module walks you through the theory and a few hands-on examples of regularization regressions including ridge, LASSO, and elastic net. You will realize the main pros and cons of these techniques, as well as their differences and similarities.

Learning Objectives:

* Identify the bias-variance trade off and the sources of error
* Recognize regularization as a tool to avoid overfitting of linear models
* Recognize common approaches to regularization, including Ridge, Lasso and Elastic Net regressions
* Assess whether introducing polynomial features improves the error metrics of your linear regression

The Bias-Variance Tradeoff

We want training and test errors to be small

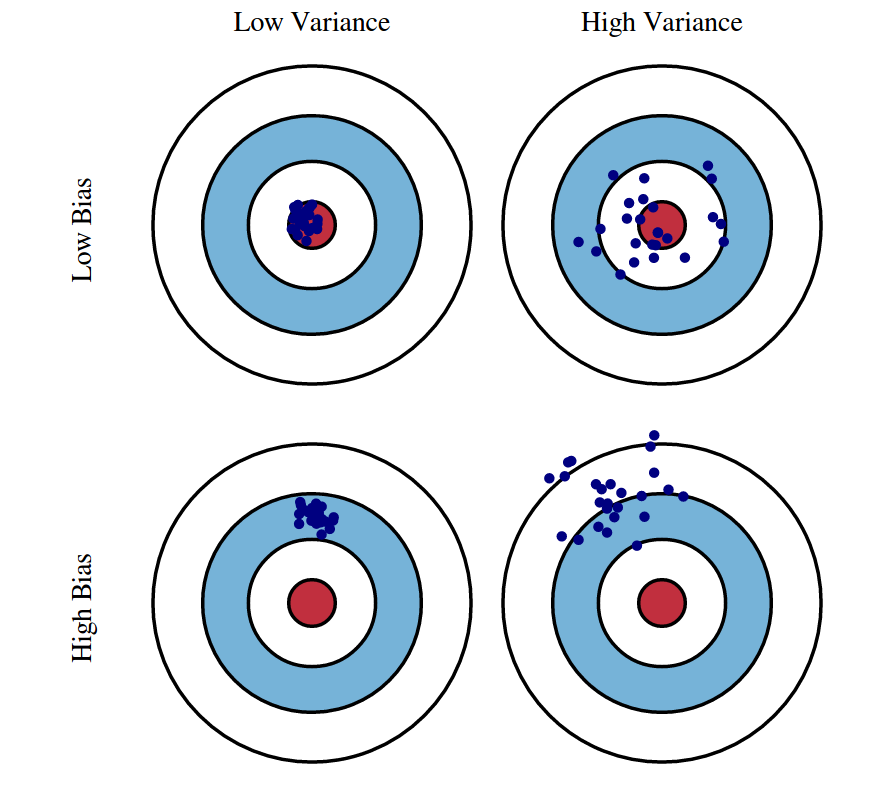
Choosing the level of complexity

* Polynomial degrees

How well does the model generalize

Bias and varianced: inuition

* Bias is a tendency to miss
* Variance is tendency to be inconcistent
* Ideally, we want the top left outcome: highly consistent predictions thata are close to perfecct on average
* Tendency: expectation of out-of-sample behavior over many training set samples.



3 sources of model error

1. Being wrong – bias
2. Being unstable – variance
3. Unavoidable randomness – irreducible error

Bias

Tendency of predictions to miss true values

* Worsened by missing information, over-simplistic assumptions
* Miss real patterns (underfit)

Variance

Tendency of predictions to fluctuate

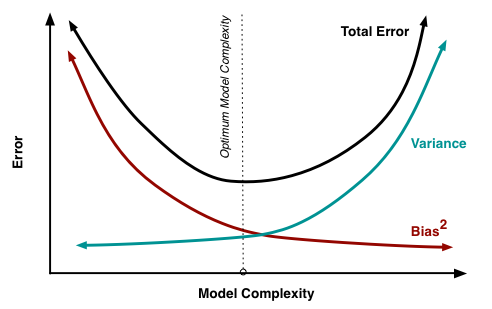
* Characterized by sensistivity or output to small changes in input data
* Often due to overly complex or poorly-fit models.

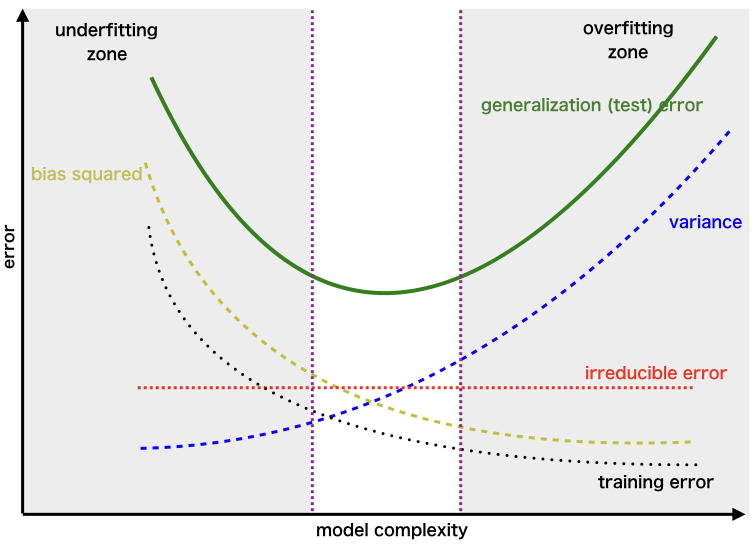
Irreducible error

Tendency to intrinsic uncertainty/randomness.

* Present in even the best possible model.

Bias-Variance Tradeoff





Summary of bias-variance tradeoff:

* Model adjustments that decrease bias often increase variance, and vice versa.
* The bias-variance tradeoff is analogous to a complexity tradeoff.
* Finding the best model means choosing the right level of complexity
* Want a model elaborate enough to not underfit, but not so exceedingly elaborate that it overfits

the higher the degree of a polynomial regression, the more complex the model (lower bias, higher variance).

At lower degrees, we can see visual signs of bias: predictions are too rigid to capture the curve pattern in the data

Regularization and Model Selection

Can we tune with more granularity than choosing polynomial degrees?

* Yes, by using regularization

What does regularization accomplish

M(w) + λR(w)

Adjusted cost funtion

M{w): model error

R(w): function of estimated paramter(s)

Λ: regularization strength parameters

The regularization strength parameter lambda allows us to manage the complexity tradeoff:

* More regularization introduces a simpler model or more bias.
* Less regularization makes the model more complex and increaes variance.

If our model is overfit (variance too high), regularization can improve the generalization error and reduce variance.

Regularization performs feature selection by shrinking the contribution of features.

For L1-regularization, this is accomplished by driving some coefficients to zero.

Feature selection can also be performed by removing features.

Reducing the number of features can prevent overfitting.

For some models, fewer features can improve fitting time and/or results.

Identifying most critical features can improve model interpretability.

Ridge Regression

Fit model by minimizing the function.

Keep in mind that scale matters.

Ridge regression:

* The complexity penalty lambda is applied proportionally to squared coefficient values.
  + The penalty term has the effect of “shrinking” coefficients toward 0.
  + This imposes bias on the model, but also reduces variance.
  + We can select the best regularization strength lambda via cross-validation.
  + It's best practice to scale features (I.e. using StandardScaler) so penalties aren’t impacted by variable scale.

Penalty shrinks magnitude of all coefficients

Larger coefficients strongly penalized because of the squaring

Shrinkage effect as regularization strength increases

Complexity tradeoff: variance reduction may outpace increase in bias, leading to a better model fit.

Alternative: LASSO Regression

Aside: penalties are closely related to L1/L2 norms, that measure vector length.

In lasso regression: the complexity penalty lambda is proprtional to the absolute value of coefficients

* Lasso: least absolute shrikage and slelection operator
* Similar effect to ridge in terms of complexit ytradeoff. Increasing lambda raises bias but lowers variance
* LASSO is more likely than ridge to perform feature selection, in that for a fixed lambda, lasso is more likely to result in coefficients being set to zero.

Lasso regression (l1)

Penalty selectively shrinks some coefficients

Can be used for feature selection.

Slower to converge than ridge regression.

Lasso regression in action

Shrinkage and selection effect as regularzation strength increases: some features drop to 0

Complexity tradeoff: variance reduction may outpace increases in bias, leading to a better model fit.

Elastic net (hybrid approach)

Validatoin gives us an empriical method for selecting between different models

Lasso's feature selection property yields an interpretability bonus, but may underperform if the target truly depnds on many of the features.

Elastic net, an alternative hybrid approach, introduces a new parameter alpha that determines a weighted average of l1 and l2 penalties

Elastic net combines penalties from oth ridge and lasso regression

It requires tuning of an additional parameter that determines emphasis of l1 vs. L1 regularization.

Recusrive feature elimination (rfe) is an approach that combines:

* A model or estimation approach
* A desired number of features

Rfe then repeatdely applies the model, measures feature importance, and recursively removes less important features.

Import the class containing the feature selection method

From sklearn.feature\_selection import aRFE

Create an instance of the class

RfeMod = RFE(est, n\_features\_to\_select=5)

(est = an instance of the mode to use)

F is the final number offeatures

Fit the instance on the data and then predict the expected value

RfeMod = rfeMod.fit(X\_train, y\_train)

Y\_predict=rfeMod.predict(X\_test)

The RFECV class will perform feature elimination using cross validation.

**Quiz:**

Which of the following statements about model complexity is TRUE?

* Higher model complexity leads to a higher chance of overfitting.

Which of the following statements about model errors is TRUE?

* Underfitting is characterized by higher errors in both training and test samples.

Which of the following statements about regularization is TRUE?

* Regularization decreases the likelihood of overfitting relative to training data.

BOTH Ridge regression and Lasso regression

* add a term to the loss function proportional to a regularization parameter.

Compared with Lasso regression (assuming similar implementation), Ridge regression is:

* less likely to set feature coefficients to zero.

**Polynomial Features and Regularization Demo**

We will generate the ground truth model

We’ll plot the data points with the underlying noise.

What is os.sep?

Np.linspace( first value, second value, and number of values between the first and second value)

Data.head() will be our random data points

Create a plot using data.plot() and ax.plot()

Part 2

Pf = PolynomialFeatures(degree)

* Above ‘degree’ is a hyperparameter

Use polynomial features to fit the model

After overfitting we want to bring in some normalization

Part 3

Absolute value of the different coefficients

Applymap will apply ‘it’ to all of our...

Part4

Every single column except sales price

Part 5

Root mean squared error

Combined a lot of steps which previously we had separated out

Part 6

Ridge lasso elastic net

Part 7

Gradient descent might take longer

Further details of regulatizatoin

Having worked through ssome regularization examples, lets examine intuitively how these thechniques (Ridge, lasso, elastic net) interact with modeling.

There are several approaches to this interpretation:

* The analytic view
* The geometric view
* The probabilistic view

The analytic view:

Increases L2/L2 penalties force coefficients to be smaller, restricting their plausible range.

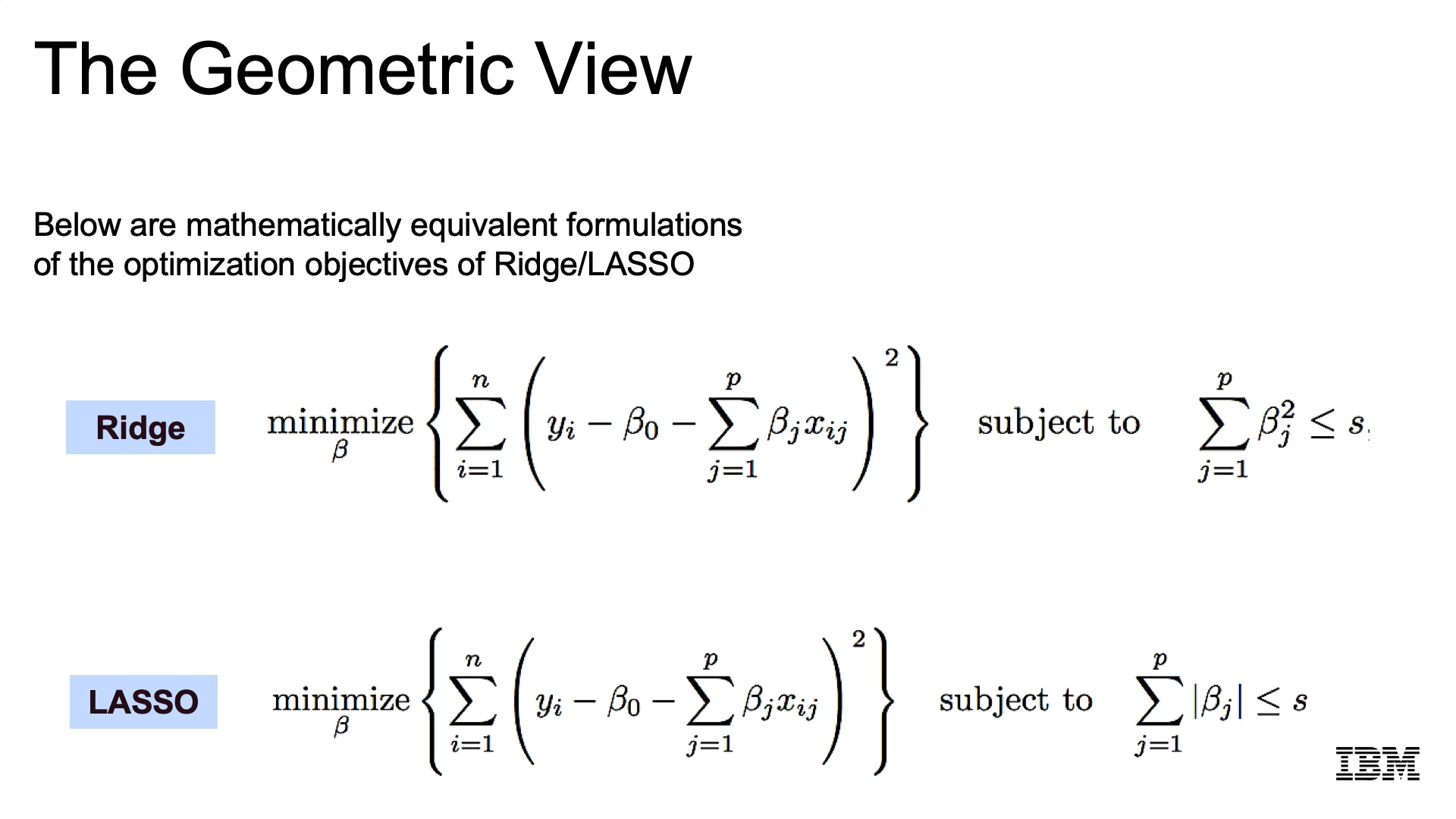
A smaller range for coefficients must be simpler/lower va rangeiance than a model with an infintite possible coefficients

The geometric view

Below are mathematically equivalent formulations of the optimation objectives of ridge/lasso

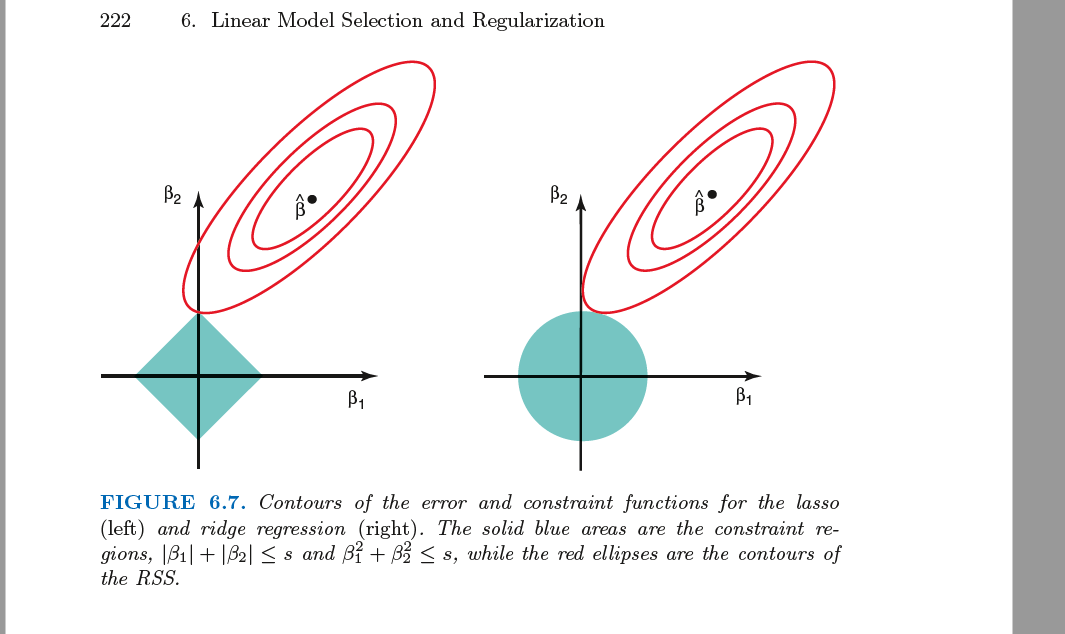
Ridge

Lasso



Under this geometric formulation, the cost function minimum is found at the intersection of the penalty boundary and a contour of the tradional ols cost function surface.

The geometry reveals the selection effect of lasso (intersection at a corner/axis zeroes out coefficients).



The probabilistic view

Bayes: regularization imposes certain priors on the regression coefficients.

Letting ‘f’ be the likelihood (probability of target given parameter vector beta), and p(beta) the prior distribution of beta, we can calculate the posterior of beta.

P(beta) is derived from independent draws of a prior coefficient density function g that we choose when regularizing.

L2(ridge) regularization imposes a gaussian prior on the coefficients. While L1 (lasso) regularization imposes a Laplacian prior.

Visualizing these prior distributions again reveal the difference in behavior between Ridge and LASSO: the laplacian distribution has peaked density at 0, explaing its tendency to zero out some coefficients.

data:image/png;base64,

Regularization recap

Complexity tradeoff

Reduce complexity by penalizing it in cost function

Increaes bias, but reduces variance (may be worth the trade-off)

Options: L2, L1, can validate the choice and strength.

Regularization

Optimizing predictive models is about finding the right bias/variance tradeoff

We need models that are sufficiently complex to capture patterns in data, but not so complex that they overfit

How it works

Analytically: penalty constrains the coefficient range

Geometrically: L1/L2 imposes bounded regions.

Probabilistically: imposes prior on coefficients.

**Details of Regularization**

Details of regularization demo

Can only fit scaling on training set

Lasso takes longer because of gradient descent

## End of module review: Regression with Regularization Techniques

### **Regularization Techniques**

Three sources of error for your model are: bias, variance, and, irreducible error.  
  
Regularization is a way to achieve building simple models with relatively low error. It helps you avoid overfitting by penalizing high-valued coefficients. It reduces parameters and shrinks the model.  
  
Regularization adds an adjustable regularization strength parameter directly into the cost function.

Regularization performs feature selection by shrinking the contribution of features, which can prevent overfitting.

In Ridge Regression, the complexity penalty λ is applied proportionally to squared coefficient values.

* The penalty term has the effect of “shrinking” coefficients toward 0.
* This imposes bias on the model, but also reduces variance.
* We can select the best regularization strength lambda via cross-validation.
* It’s a best practice to scale features (i.e. using StandardScaler) so penalties aren’t impacted by variable scale.

In LASSO regression: the complexity penalty λ (lambda) is proportional to the absolute value of coefficients. LASSO stands for : Least Absolute Shrinkage and Selection Operator.

* Similar effect to Ridge in terms of complexity tradeoff: increasing lambda raises bias but lowers variance.
* LASSO is more likely than Ridge to perform feature selection, in that for a fixed λ, LASSO is more likely to result in coefficients being set to zero.

Elastic Net combines penalties from both Ridge and LASSO regression. It requires tuning of an additional parameter that determines emphasis of L1 vs. L2 regularization penalties.  
LASSO’s feature selection property yields an interpretability advantage, but may underperform if the target truly depends on many of the features.  
  
Elastic Net, an alternative hybrid approach, introduces a new parameter α (alpha) that determines a weighted average of L1 and L2 penalties.  
  
Regularization techniques have an analytical, a geometric, and a probabilistic interpretation.

**Quiz**

### Question 1

The variance of a model is determined by the degree of irreducible error.

* False

### Question 2

As more variables are added to a model, both its complexity and its variance generally increase.

* True

### Question 3

Model adjustments that decrease bias also decrease variance, leading to a bias-variance tradeoff.

* False

### Question 4

Which of the following statements about scaling features prior to regularization is TRUE?

A. The scale or features must be the same to implement L1 or L2 regularization.

B. Features should rarely or never be scaled prior to implementing regularization.

C. The larger a feature’s scale, the more likely its estimated impact will be influenced by regularization.

D. The smaller a feature’s scale, the more likely its estimated impact will be influenced by regularization.

Answer: C

### Question 5

Which of the following statements about model complexity is TRUE?

A. Higher model complexity leads to a lower chance of overfitting.

B. Higher model complexity leads to a higher chance of overfitting.

C. Reducing the number of features while adding feature interactions leads to a lower chance of overfitting.

D. Reducing the number of features while adding feature interactions leads to a higher chance of overfitting.

Answer: B

### Question 6

A model with high variance is characterized by sensitivity to small changes in input data.

* True

### Question 7

Which of the following statements about Elastic Net regression is TRUE?

A. Elastic Net combines L1 and L2 regularization.

B. Elastic Net does not use L1 or L2 regularization.

C. Elastic Net uses L2 regularization, as with Ridge regression.

D. Elastic Net uses L1 regularization, as with Ridge regression.

Answer: A