Linear Regression, Bias-Variance, Regularization

Data 100 Final Review

Suraj Rampure, Allen Shen

Notation Changes

- In past semesters, Φ was used to represent the feature matrix given data records
- This semester, we are using X to represent the feature matrix
    - In some of the slides and problems, Φ is used, so just replace it with X

$E[y|X] = X^T\beta$

$\|y - X\beta\|_2^2$ ← X is a matrix

X is a vector (test point)

Loss Functions and Risk

A **loss function** is a function that characterizes the cost, error, or loss resulting from a particular choice of model or model parameters.

- **Loss Function:** Measures loss for a particular observation. ← specific point
- **Empirical Risk:** Average loss on the training set of observations. ← entire training set

The choice of loss function **depends on the estimation task**. Questions to consider:

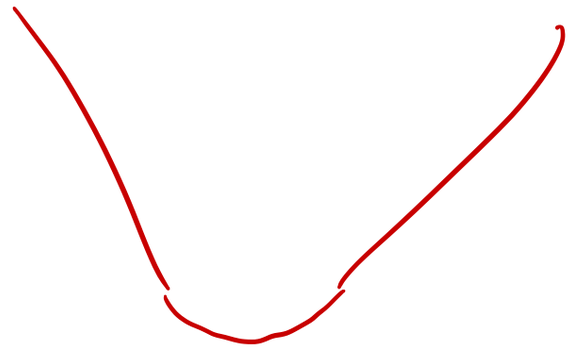
- Qualitative or quantitative predictions?
- Sensitivity to outliers
- The cost of error (how bad are false negatives?)
- Does the loss function have an analytical solution?

Common Loss Functions

Suppose y represents the true value of a variable, and $\hat{y}$ represents our predicted value.

L1 Loss

$$L_1(y, \hat{y}) = |y - \hat{y}|$$



L2 Loss

$$L_2(y, \hat{y}) = (y - \hat{y})^2$$

Huber Loss

$$L_\alpha(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & |y - \hat{y}| \leq \alpha \\ \alpha\left(|y - \hat{y}| - \frac{1}{2}\alpha\right), & \text{else} \end{cases}$$

Minimizing Risk

- Finding the parameter values that minimize the empirical risk on training data
- Generally, we are interested in convex loss functions (as they have global minimums)
- To find the analytical solution, we perform our favorite procedure:

```
1. Take the derivative/gradient
2. Set equal to zero
3. Solve for optimal parameters
```

Problem: Many loss functions do not have an analytical solution! In other words, we cannot solve for our optimal parameters mathematically... what do we do?

Example

Consider the following loss function:

$a=b=1 \Rightarrow |f_\theta(x) - y|$

$$L_{OINK}(\theta, x, y) = \begin{cases} a(f_\theta(x) - y) & f_\theta(x) \geq y \\ b(y - f_\theta(x)) & f_\theta(x) < y \end{cases}$$

with the constant model

$$f_\theta(x) = \theta$$

$L(\theta, x, y) = \begin{cases} a(\theta - y) & y \leq \theta \\ b(y - \theta) & y > \theta \end{cases}$

and empirical risk

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} L_{OINK}(\theta, x_i, y_i)$$

Find the optimal $\hat{\theta}$ that minimizes the risk in each of the following cases:

a) $a = b = 1$
b) $a = 1, b = 5$
c) $a = 3, b = 6$

X	·	·	·	·	·	·	·
y	40	0	50	30	20	60	10

$$L(\theta, x, y) = \begin{cases} a(\theta - y) & y \le \theta \\ b(y - \theta) & y > \theta \end{cases}$$

$$R(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(\theta, x_i, y_i)$$

$$\frac{\partial L}{\partial \theta} = \begin{cases} a & y \le \theta \\ -b & y > \theta \end{cases}$$

$$\frac{\partial R}{\partial \theta} = \frac{1}{n} \left(a \cdot \#\, y \le \theta \; + \; (-b) \#\, y > \theta \right) = 0$$

$$\#\, y \le \theta = \frac{b}{a} \cdot \#\, y > \theta$$

0
10
20
30
40
50
60

1) $a = b = 1$

$\Rightarrow \#\, y \le \theta = \#\, y > \theta \Rightarrow \hat{\theta} = \text{median}(y)$

$$\#y \le \theta = \left(\frac{b}{a}\right) \cdot \#y > \theta$$

0 10 20 30 40 50 60

b) $a=1, b=5$ $\frac{b}{a}=5$, $\#y \le \theta = 5 \cdot \#y > \theta$

$\hat{\theta} = 50$

c) $a=4, b=6$ $\hat{\theta} = 40$

Gradients – Review

Suppose $a, x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$.

Recall the gradients of each of the following functions.

$$f(x) = a^T x$$

$$\Rightarrow \nabla f(x) = a$$

$$f(x) = x^T x$$

$$\Rightarrow \nabla f(x) = 2x$$

$$f(x) = x^T A x$$

$$\Rightarrow \nabla f(x) = (A + A^T)x$$

$$= 2Ax \quad (\text{symmetric } A)$$

Linear Regression in 2D

Suppose we're given $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, and want to fit a linear model $y = \beta_1 x + \beta_0$, using MSE (i.e. L2) loss.

Our objective function is

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$$

One way to solve: Take partial derivatives with respect to β_0, β_1. Solve for β_0 and β_1.

$$y_i = \beta_1 x_i + \beta_0$$

Let's try and rewrite this in vector form.

$$L(\beta) = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

We can say the following:

$$\beta = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix}^T$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T$$

$$y = X\beta \quad \text{model}$$

$$L(\beta) = \| y - X\beta \|_2^2$$

$$e$$

Solution to Normal Equation, using Vector Calculus

$$L(\beta) = ||y - X\beta||_2^2 = \left((y - X\beta)^T (y - X\beta)\right)$$

$$= \left(y^T y - y^T X\beta - (X\beta)^T y - \beta^T X^T X\beta\right)$$

$$= y^T y - 2(X^T y)^T \beta - (X\beta)^T (X\beta)$$

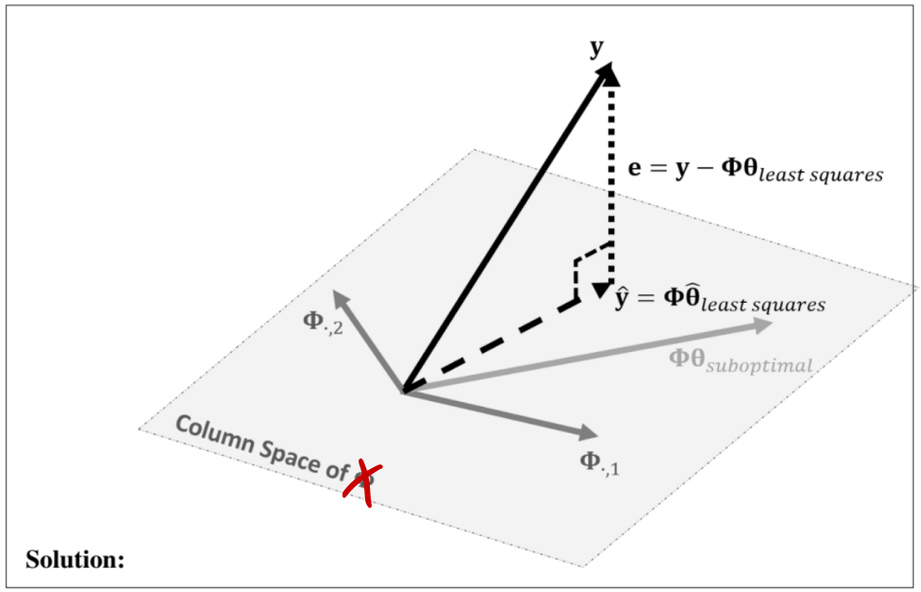
Taking the gradient and setting it equal to 0:

$$\nabla L(\beta) = 0 - 2X^T y - 2X^T X\beta = 0$$

$$\Rightarrow X^T X\beta = X^T y$$

$$\Rightarrow \beta^* = (X^T X)^{-1} X^T y$$

Solution to Normal Equation, using Geometry



error $\perp$ X

$X^T e = 0$

$X^T(y - X\beta) = 0$

$\Rightarrow \beta^* = (X^T X)^{-1} X^T y$

We see that to minimize e, e must be orthogonal to the column space of X (or, in the picture, Φ).

- The Discussion 7 walkthrough talks about this in significant detail.

Regularization

$$\beta^*_{OLS} = \underbrace{(X^TX)^{-1}}_{\text{rank}(X) = \text{rank}(X^TX)} X^Ty$$

Issues with OLS:

- Solution doesn't always exist (if X is not full-rank, X^TX will not be full rank)
- Numerical issues with inversions
- Potential overfitting to training set – model can be too complex

To fix: Add penalty on magnitude of β. However, we could use either the L2 norm, or L1 norm!

Using L2 (Ridge): $L(\beta) = ||y - X\beta||_2^2 + \lambda||\beta||_2^2$

Using L1 (LASSO): $L(\beta) = ||y - X\beta||_2^2 + \lambda||\beta||_1$

$$||\beta||_2^2 = \beta_0^2 + \beta_1^2 + \dots \beta_p^2$$

$$||\beta||_1 = |\beta_0| + |\beta_1| + \dots + |\beta_p|$$

Note: $||x||_2 = \sqrt{x_1^2 + x_2^2 + ... + x_n^2}$, and $||x||_1 = |x_1| + |x_2| + ... + |x_n|$.

Regularization – Ridge Regression

When we use the L2 vector norm for the penalty term, our objective function becomes

$$\min_{\beta} ||y - X\beta||_2^2 + \lambda ||\beta||_2^2$$

This is called "ridge regression."

Solution can also be determined using vector calculus.

$$\Rightarrow \beta^*_{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

- λ represents the penalty on the size of our model. We will discuss this more later in the review.
- Unlike OLS, Ridge Regression always has a unique solution!

Regularization – LASSO Regression

When we use the L1 vector norm for the penalty term, our objective function becomes

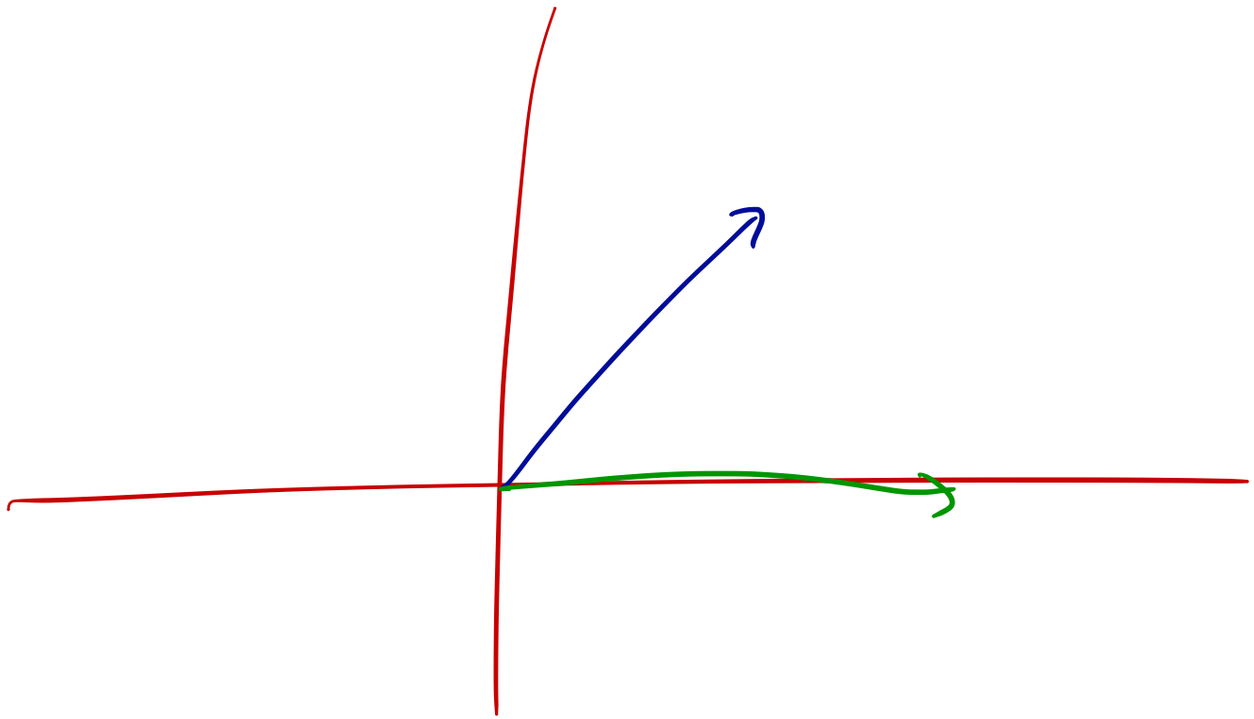
$$\min_{\beta} \left|\left|y - X\beta\right|\right|_2^2 + \lambda\left|\left|\beta\right|\right|_1$$

This is called "LASSO regression." *Fun fact: LASSO stands for Least Absolute Shrinkage and Selection Operator.*

Unlike OLS and Ridge Regression, there is (in general) no closed form solution. Need to use a numerical method, such as gradient descent.

- LASSO regression encourages sparsity, that is, it sets many of the entries in our β vector to 0. LASSO effectively selects features for us, and also makes our model less complex (many weights set to 0 ——> less features used ——> less complex)
- Again, λ represents the penalty on the size of our model.

L_1 vs. L_2 vector norms: $(3, 4)$ vs. $(5, 0)$



$L_1(5, 0) = |5| + |0|$

$= 5$

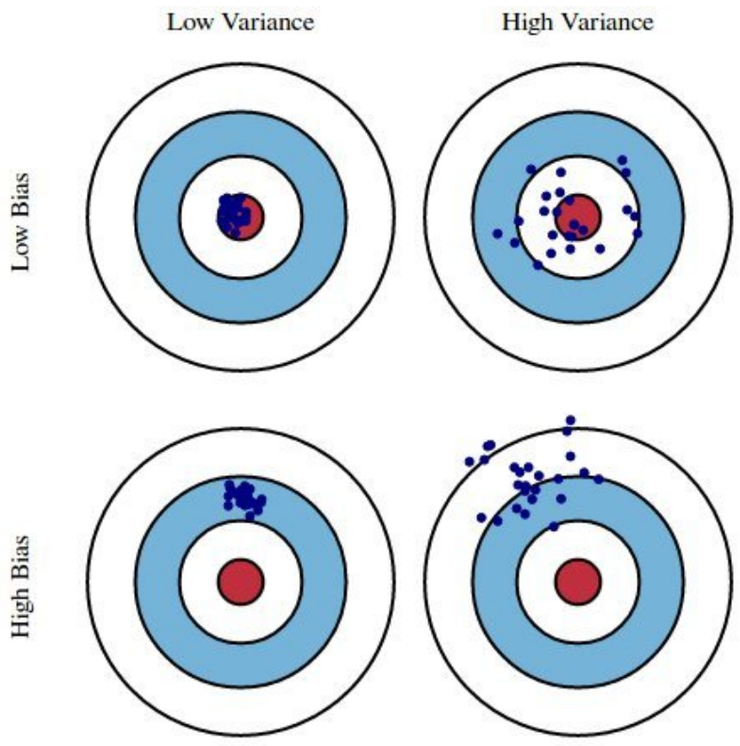
$L_2(5, 0) = \sqrt{5^2 + 0^2}$

$= 5$

$L_1(3, 4) = |3| + |4| = 7$

$L_2(3, 4) = \sqrt{3^2 + 4^2} = 5$

Bias-Variance



Intuitively speaking:

- **Low bias and high variance** means that your predictions will vary wildly – depending on the dataset, they may be very close, or very far off
- **High bias and low variance** means your predictions will be a decent amount wrong, no matter what dataset you through at your model

This semester: ignore ϵ

Bias-Variance Decomposition

Suppose ϵ is some random variable such that $\mathbb{E}[\epsilon] = 0$ and $var[\epsilon] = \sigma^2$. Also, suppose we have Y generated as follows:

$$Y = h(x) + \epsilon$$

We collect some sample points $\{(x_i, y_i)\}_{i=1}^n$, and want to fit a model $f_\beta(x)$. We define the model **risk** as $\mathbb{E}[(Y - f_\beta(x))^2]$.

$$\mathbb{E}[(Y - f_\beta(x))^2] = (h(x) - \mathbb{E}[f_\beta(x)])^2 + \mathbb{E}(\mathbb{E}[f_\beta(x)] - f_\beta(x))^2] + \sigma^2$$

bias^2 model variance obs. error

This is sometimes referred to as the **bias-variance decomposition.**

Let's analyze the objective function for ridge regression (however, the analysis is the same for

$$\min_{\theta} \left|\left|y - X\beta\right|\right|_2^2 + \lambda\left|\left|\beta\right|\right|_2^2$$

As λ increases, model complexity decreases. This is because increasing λ increases the penalty on the magnitude of β. Since we are trying to minimize the objective, if λ increases, $\left|\left|\beta\right|\right|_2^2$ must decrease.

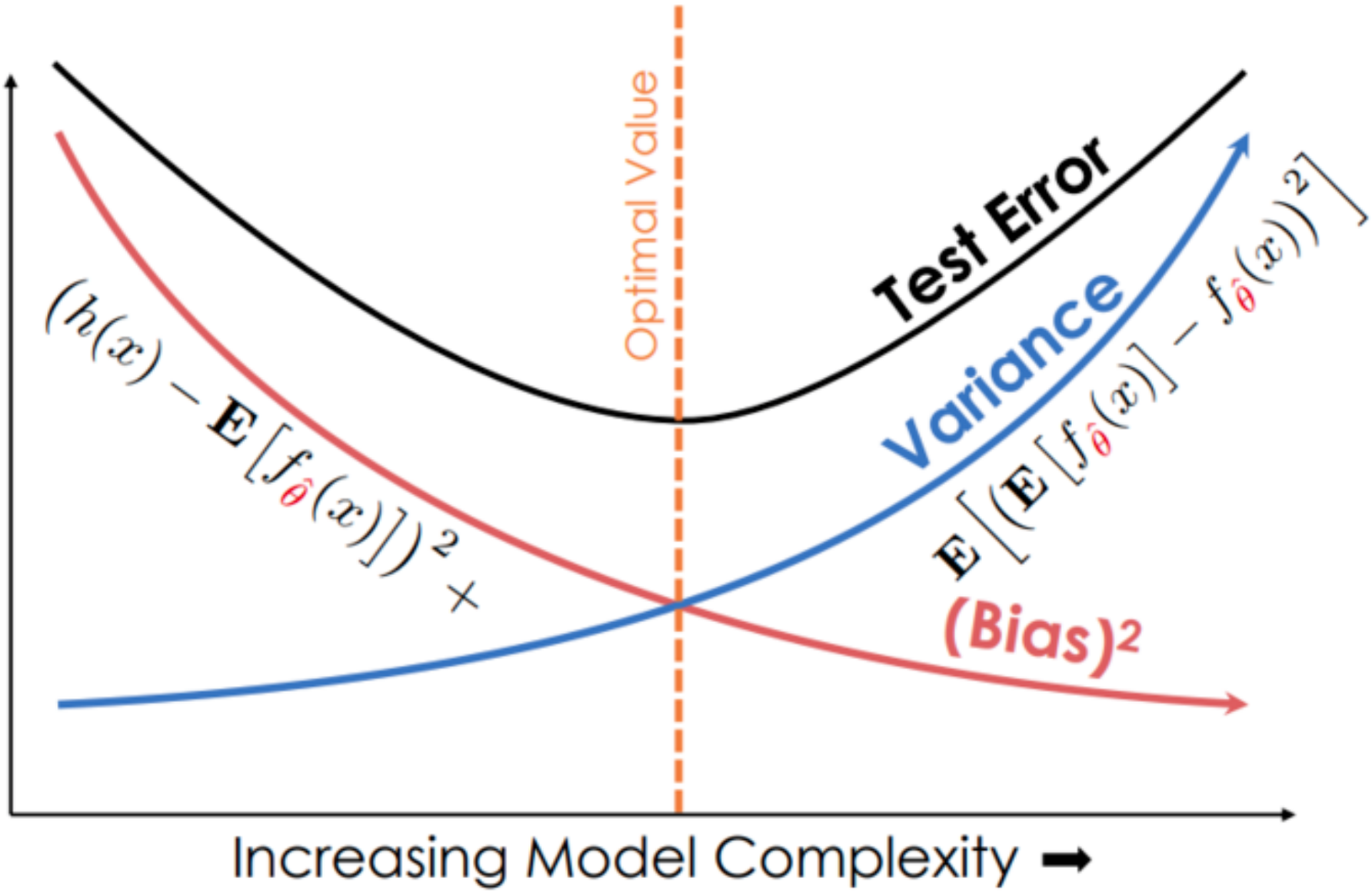
As a result, as λ increases, model bias increases, and variance decreases.

- Bias increases because our model becomes less complex, and thus more general.
- Variance decreases because, again, our model becomes more general.

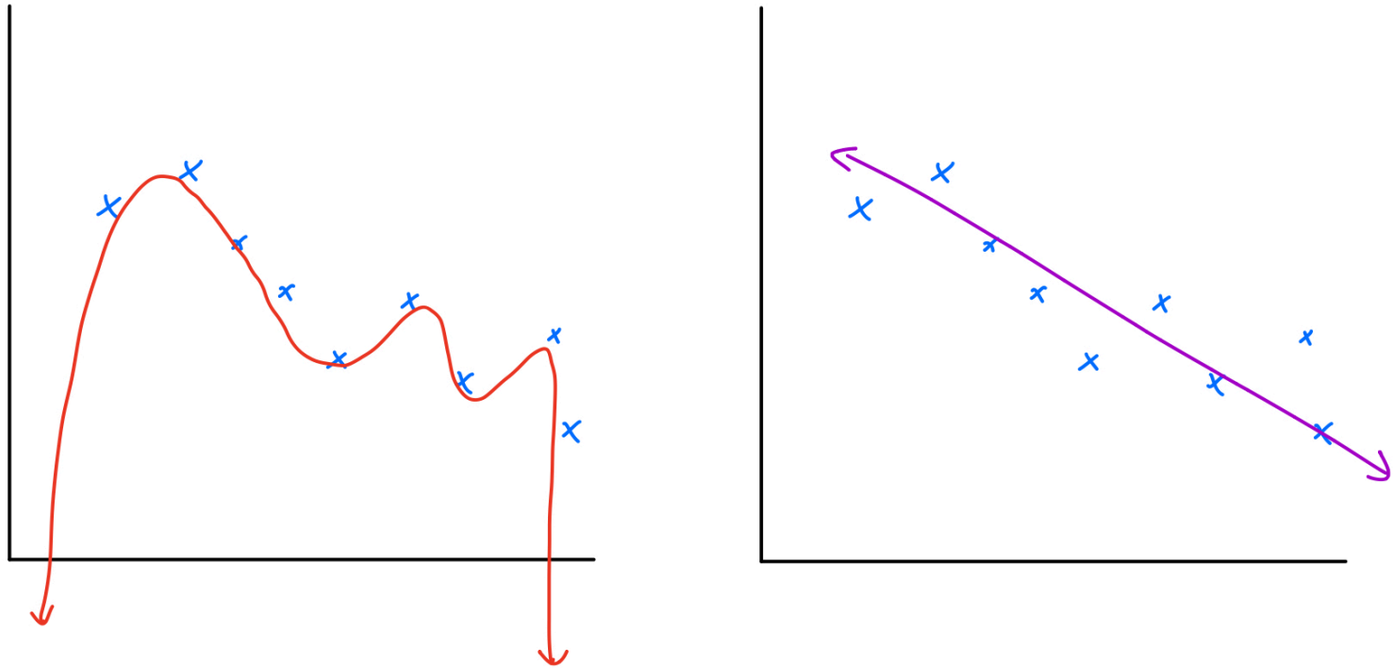
as $\lambda \to \infty$, $\beta \to 0$

← underfitting

overfitting →



Polynomial regression with large d, small d:



The high degree polynomial model has lower bias, but higher variance, than the model on the right.

One way to interpret variance: In the model on the left, if we were to introduce a new point, our polynomial model would change significantly. However, on the right, introducing a new point is unlikely to change our model by much.

Practice: Regularization, B-V (T/F)

True/False: L_1 regularization can help us select a subset of the features that are important.

True/False: After regularization, we expect the training accuracy to increase and the test accuracy to decrease.

True/False: In ridge regression, if we let $\lambda \to \infty$, our model will become more and more complex.

True/False: As we improve our model to reduce bias, we often run the risk of under-fitting.

True/False: Training error is typically larger than test error.

Practice: Regularization

Consider the following general loss formulation.

$$\arg\min_{\theta} \left[\sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{p=1}^{d} \beta^2 \right]$$

a) How many data points are there? n

b) What dimension is our data, i.e. how many features are we using? d

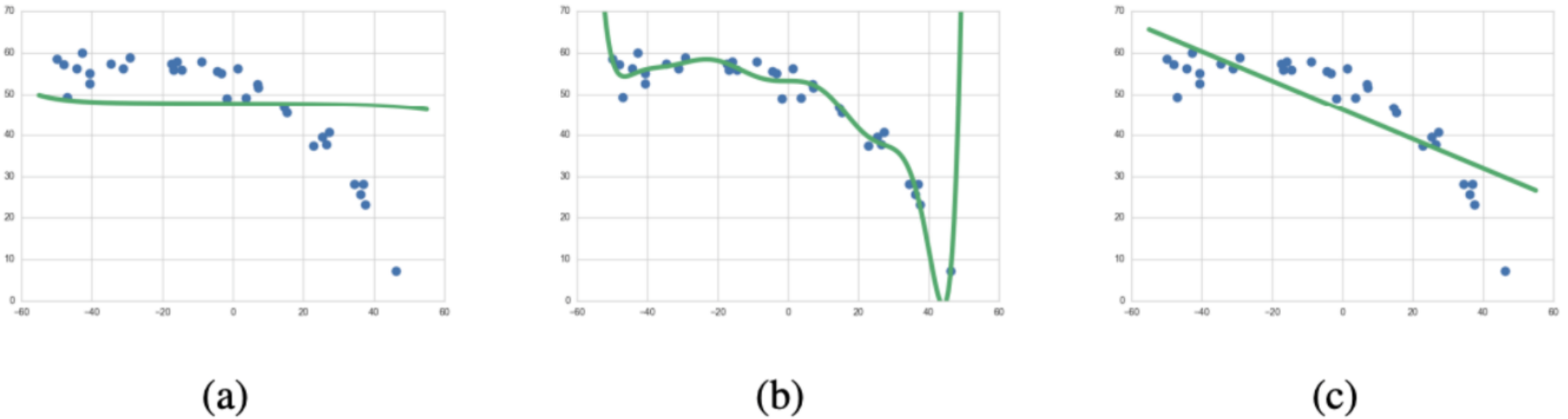
c) Is this a classification or regression problem? regression

d) What type of regularization is being used? L2

e) As λ increases, what will happen to bias? increases

f) As λ increases,what will happen to variance? decreases

Practice: Identifying types of regression



(a) (b) (c)

Determine the plot above that best represents each of the following models.

i. Linear regression C

ii. Regularized linear regression, using polynomial features and a large λ A

iii. Linear regression, using degree 10 polynomial features B