Discussion 10: Bias-Variance, Regularization, Random Variables

Data 100, Spring 2020

Suraj Rampure

Friday, April 10th, 2020

Agenda

- Random Variables
- Bias-Variance Trade-off
- Regularization

As per usual, everything will be posted at

http://surajrampure.com/teaching/ds100.html

In addition, here's a feedback form:

http://tinyurl.com/feedbacksuraj

Random Variables

The two common discrete probability distributions we use in this class are the Bernoulli and Binomial distributions.

$$E[X] = \sum_{x \in \Omega} x \cdot P(X = x)$$

Bernoulli Distribution

- Models a single trial of some event.
  - For example, a single flip of a coin.
- Single parameter: p. → probability of a success

$$P(X = 1) = p \qquad E[X] = p$$

$$P(X = 0) = 1 - p \qquad Var[X] = p(1-p)$$

Binomial Distribution

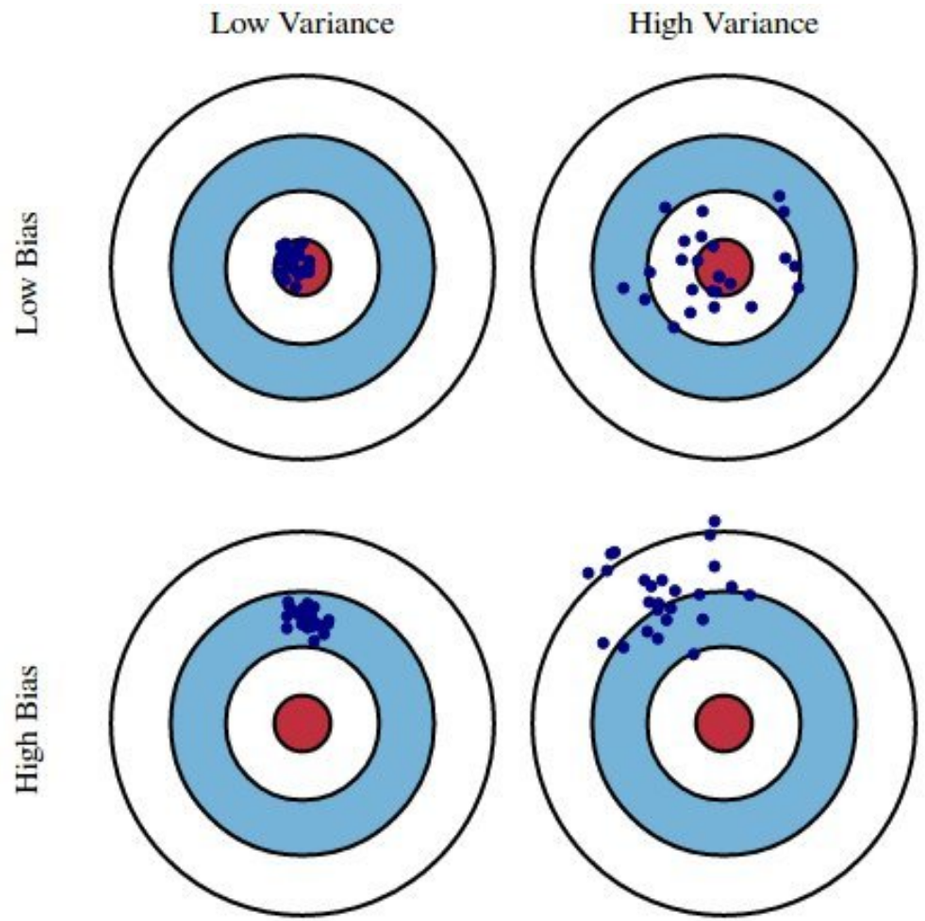
- Models the number of successes in n independent trials of some event, each of which suceed with probability p.
  - Can be thought of as a sum of n indepedent and identically distributed (i.i.d.) Bernoulli random variables.
  - For example: Suppose I flip a coin 12 times. It lands heads each time with probability 0.6. What's the probability I see 7 heads?
- Two parameters: n, p.

parameter p

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

account for all rearrangements of k successes (e.g. "heads"), n-k failures (e.g. "tails")

Bias-Variance



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:

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

true model

- **Our goal is to come up with the best estimate of $g(x)$ possible.**
- To do this, we collect some sample points $\{(x_i, y_i)\}_{i=1}^n$, and fit a model $f_{\hat{\theta}}(x)$.
  - Note, $\hat{Y} = f_{\hat{\theta}}(x)$. $E[(y - \hat{y})^2]$
- We define the model **risk** as $\mathbb{E}[(Y - f_{\hat{\theta}}(x))^2]$.

$f_{\hat{\theta}}(x)$: fitted model (guess of $g(x)$)

(expectation over all possible samples)

The model risk can be decomposed into:

$$\mathbb{E}[(Y - f_{\hat{\theta}}(x))^2] = \sigma^2 + (g(x) - \mathbb{E}[f_{\hat{\theta}}(x)])^2 + \mathbb{E}[(\mathbb{E}[f_{\hat{\theta}}(x)] - f_{\hat{\theta}}(x))^2]$$

obs. var σ^2 | bias^2 $(g(x) - \mathbb{E}[f_{\hat{\theta}}(x)])^2$ | model var. $\mathbb{E}[(\mathbb{E}[f_{\hat{\theta}}(x)] - f_{\hat{\theta}}(x))^2]$

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

Bias and Variance

Note: Both of the following depend on our prediction $f_{\hat{\theta}}(x)$ (and hence, our choice of $\hat{\theta}$).

Bias

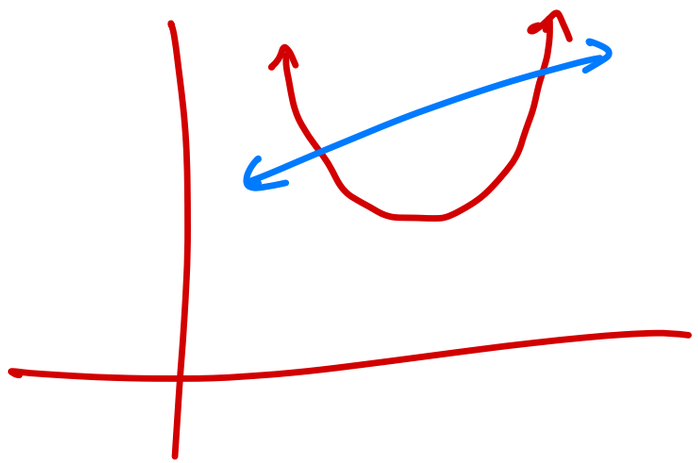
$$g(x) - \mathbb{E}[f_{\hat{\theta}}(x)]$$

- The expected difference between the true value and our prediction.
- High bias typically indicates **underfitting**.
- **Intuitively:** Model may be too basic to capture the underlying relationship.

$$Var(X) = E\left[(X - E[X])^2\right]$$

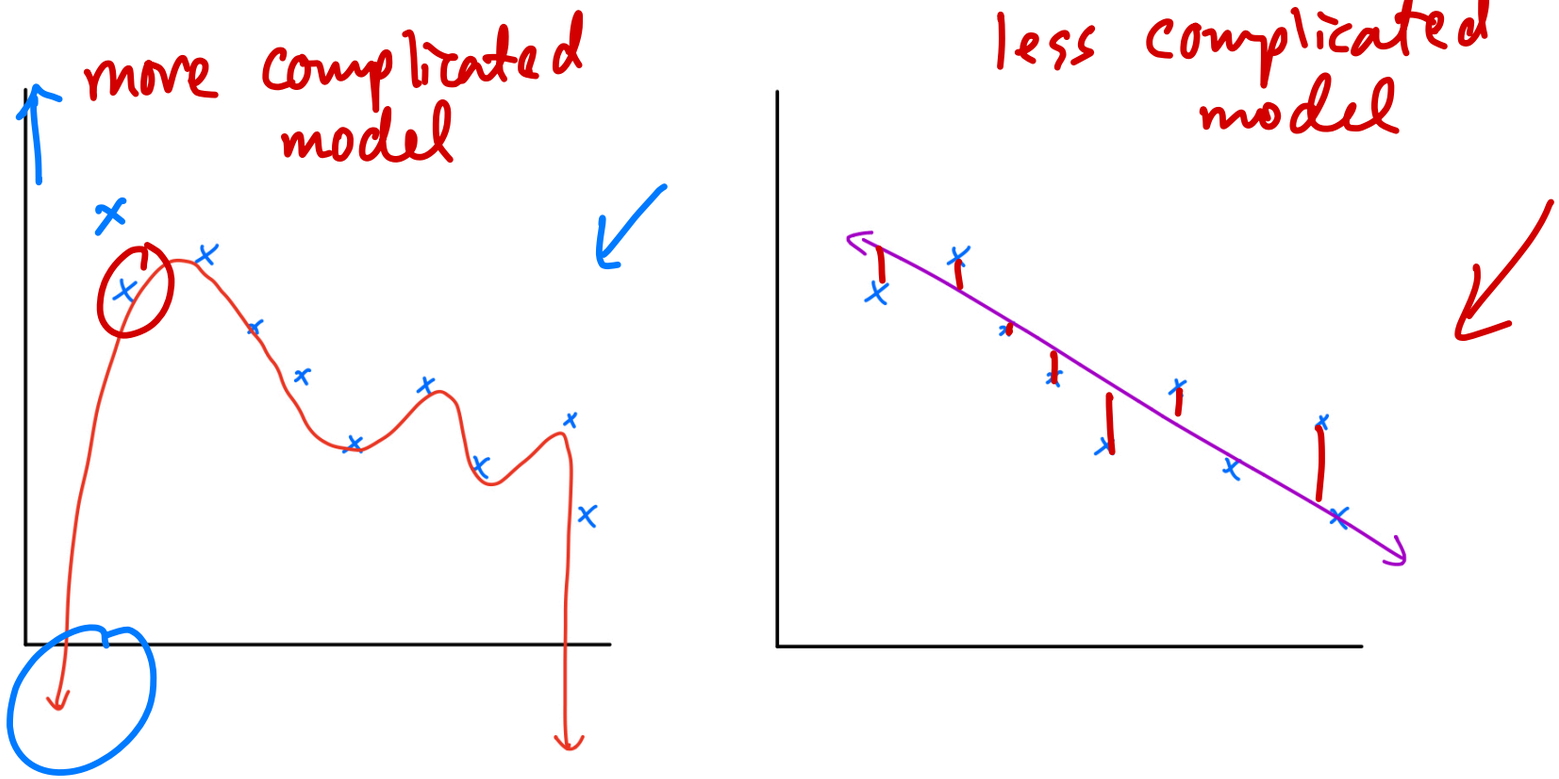
Model Variance:

$$\mathbb{E}[(\mathbb{E}[f_{\hat{\theta}}(x)] - f_{\hat{\theta}}(x))^2]$$



- Variance of $f_{\hat{\theta}}(x)$, our prediction.
- **Intuitively:** How much our predictions vary, given unseen data.
- High variance indicates **overfitting** to training data.

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.

Model Complexity

Observation: We can make our training error arbitrarily close to 0, by adding more and more features.

Why don't we do this?

Overfitting to our training data likely means we won't be able to generalize to unseen data!

Pitfalls of Ordinary Least Squares

In **Ordinary Least Squares**, our goal was to find the vector θ that minimizes the following empirical risk:

$$R(\theta) = \frac{1}{n}||y - X\theta||_2^2$$

$y = X\theta$

MSE

The **optimal value of** θ (i.e. $\hat{\theta}$) is given by

$$\hat{\theta} = (X^T X)^{-1} X^T y$$

Issues with OLS:

- Solution doesn't always exist (if X is not full-rank, $X^T X$ will not be full rank).
- Potential overfitting to training set – model can be too complex.

Pitfalls of Ordinary Least Squares

Solution: **Add penalty on magnitude of θ.**

Now, our optimization problem is to find the θ that minimizes

$$R(\theta) = \frac{1}{n}||y - X\theta||_2^2 + \lambda S(\theta)$$

- If $S(\theta) = \sum_{i=1}^{p} \theta_i^2 = ||\theta||_2^2$, we are performing L_2 regularization, called **ridge regression**.
- If $S(\theta) = \sum_{i=1}^{p} |\theta_i| = ||\theta||_1$, we are performing L_1 regularization, called **LASSO**.

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

Ridge Regression

When we use the L_2 vector norm for the penalty term, our objective function becomes

$$R(\theta) = \frac{1}{n}||y - X\theta||_2^2 + \lambda||\theta||_2^2$$

Solution can also be determined using vector calculus.

$$\Rightarrow \hat{\theta}_{ridge} = (X^T X + n\lambda I)^{-1} X^T y$$

- λ represents the penalty on the size of our model. It is a **hyperparameter**, in that we get to choose it as opposed to learn it from our data. We will discuss this more shortly.
- Unlike OLS, Ridge Regression always has a unique solution!

LASSO Regression

When we use the L_1 vector norm for the penalty term, our objective function becomes

$$R(\theta) = \frac{1}{n}\left|\left|y - X\theta\right|\right|_2^2 + \lambda\left|\left|\theta\right|\right|_1$$

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

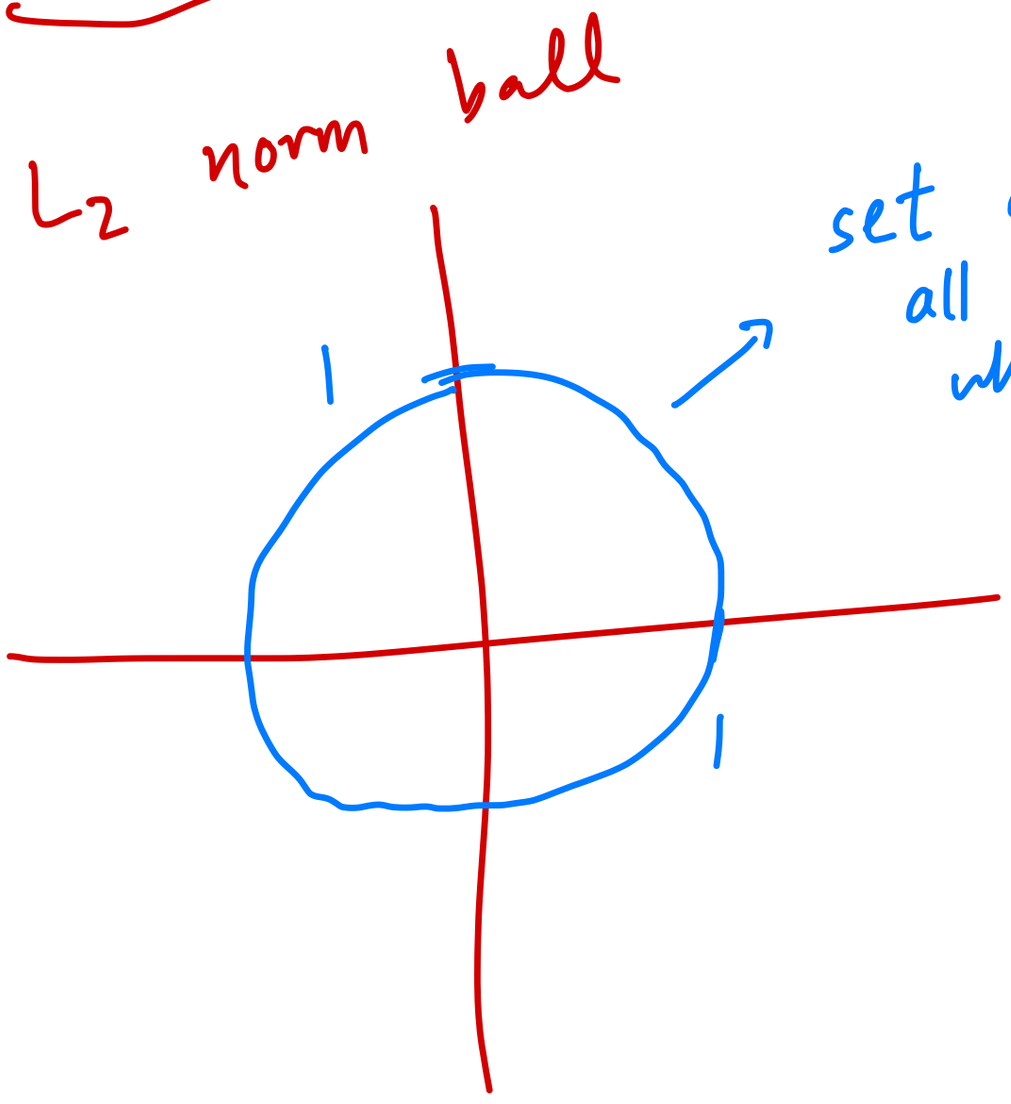
- Again, λ represents the penalty on the size of our model.
- LASSO regression encourages sparsity, that is, it sets many of the entries in our θ vector to 0. LASSO effectively selects features for us, and also makes our model less complex (many weights set to 0 ——> less features used ——> less complex)

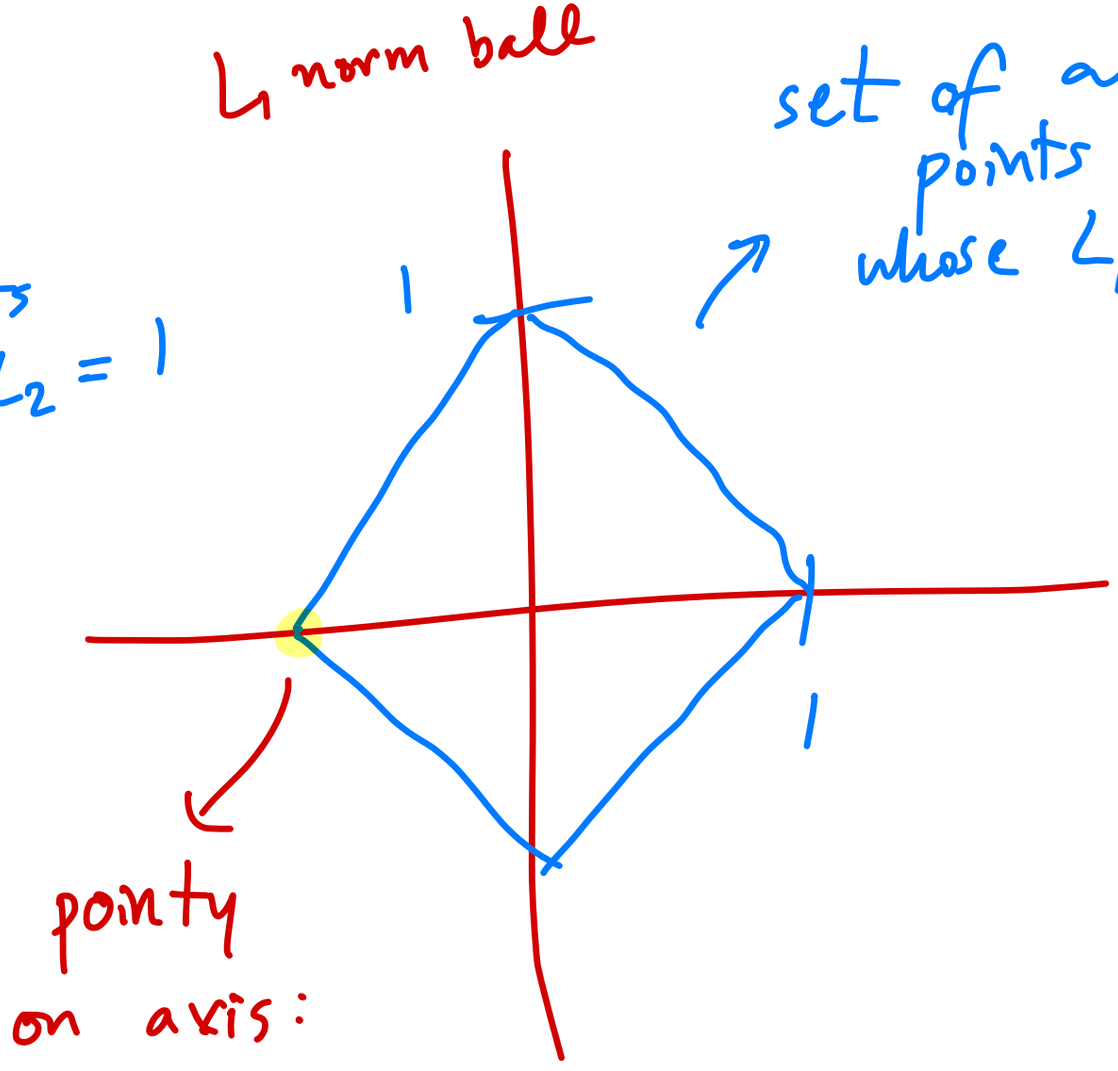
L_1 vs. L_2 vector norms: $(3, 4)$ vs. $(5, 0)$

$L_1(3,4) = 7$ $\quad$ $L_1(5,0) = 5$

$L_2(3,4) = 5$ $\quad$ $L_2(5,0) = 5$

→ same L_2 norm, but the sparser one has the smaller L_1 norm





Regularization and Bias / Variance

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

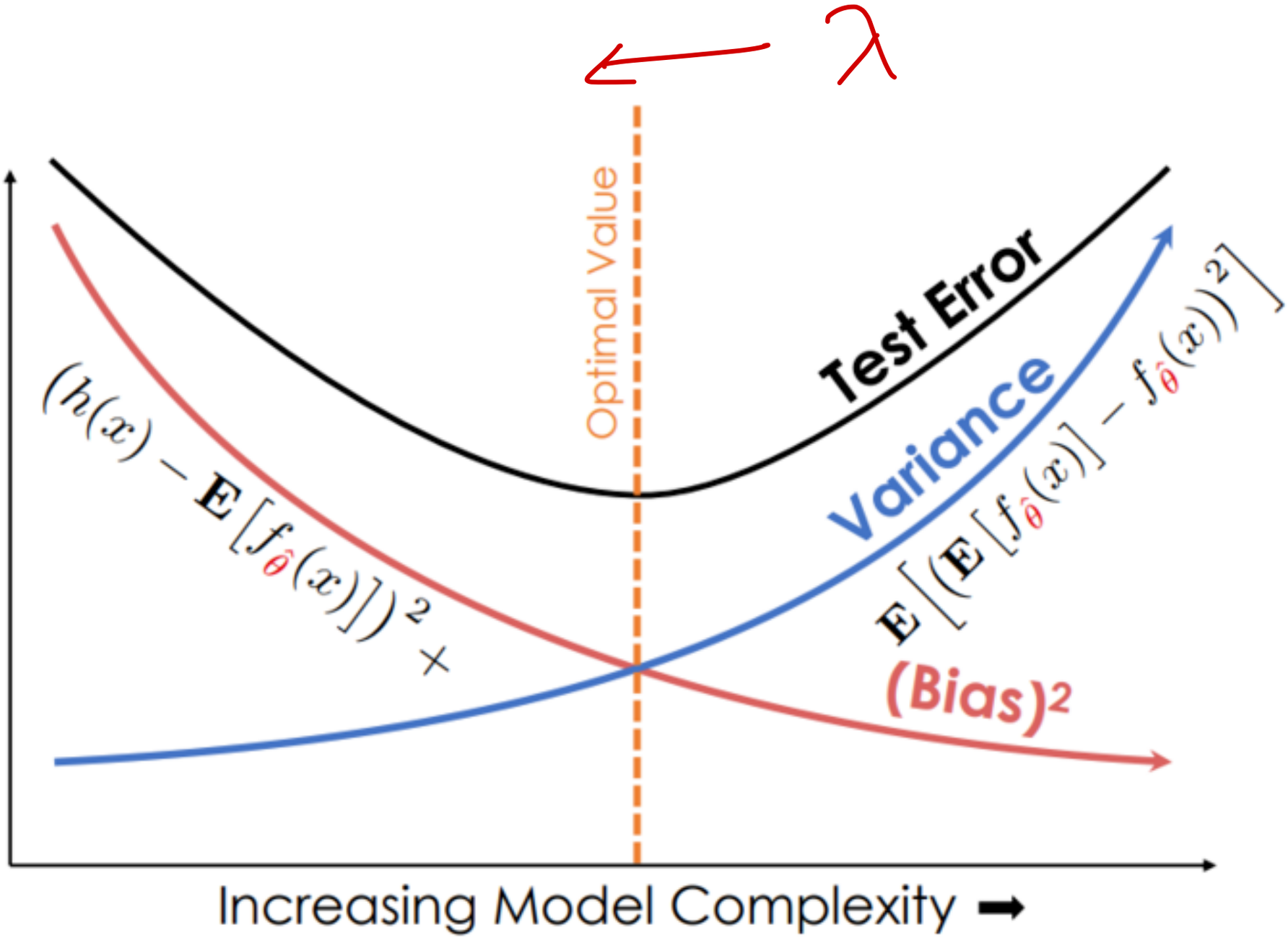
$$R(\theta) = \frac{1}{n}\left|\left|y - X\theta\right|\right|_2^2 + \lambda\left|\left|\theta\right|\right|_2^2$$

As λ increases, model complexity decreases.

- This is because increasing λ increases the penalty on the magnitude of θ.
- Since we are trying to minimize the objective, if λ increases, $\left|\left|\theta\right|\right|_2^2$ must decrease.

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

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



Reminder: *Model complexity and λ are inversely related!*