CS 189 Introduction to Machine Learning
Spring 2018 Note 5

1 Bias-Variance Tradeoff

Recall from our previous discussion on supervised learning, that for a fixed input $\mathbf{x}$ the corresponding measurement Y is a noisy measurement of the true underlying response $f(\mathbf{x})$:

$$Y = f(\mathbf{x}) + Z$$

Where Z is a zero-mean random variable, and is typically represented as a Gaussian distribution. Our goal in regression is to recover the underlying model $f(.)$ as closely as possible. We previously mentioned MLE and MAP as two techniques that try to find of reasonable approximation to $f(.)$ by solving a probabilistic objective. We briefly compared the effectiveness of MLE and MAP, and noted that the effectiveness of MAP is in large part dependent on the prior over the parameters we optimize over. One question that naturally arises is: how exactly can we measure the effectiveness of a hypothesis model? In this section, we would like to form a theoretical metric that can exactly measure the effectiveness of a hypothesis function h. Keep in mind that this is only a theoretical metric that cannot be measured in real life, but it can be approximated via empirical experiments — more on this later.

Before we introduce the metric, let's make a few subtle statements about the data and hypothesis. As you may recall from our previous discussion on MLE and MAP, we had a dataset

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$$

In that context, we treated the $\mathbf{x}_i$'s in our dataset $\mathcal{D}$ as *fixed* values. In this case however, we treat the $\mathbf{x}_i$'s as values sampled from **random variables** $\mathbf{X}_i$. That is, $\mathcal{D}$ is a random variable, consisting of random variables $\mathbf{X}_i$ and Y_i. For some arbitrary test input $\mathbf{x}$, $h(\mathbf{x}; \mathcal{D})$ depends on the random variable $\mathcal{D}$ that was used to train h. Since $\mathcal{D}$ is random, we will have a slightly different hypothesis model $h(\mathbf{x}; \mathcal{D})$ every time we use a new dataset. Note that $\mathbf{x}$ and $\mathcal{D}$ are completely independent from one another — $\mathbf{x}$ is a test point, while $\mathcal{D}$ consists of the training data.

1.1 Metric

Our objective is to, for a fixed test point $\mathbf{x}$, evaluate how closely the hypothesis can estimate the *noisy* observation Y corresponding to $\mathbf{x}$. Note that we have denoted $\mathbf{x}$ here as a lowercase letter because we are treating it as a fixed constant, while we have denoted the Y and $\mathcal{D}$ as uppercase letters because we are treating them as random variables. Y and $\mathcal{D}$ as **independent random variables**, because our $\mathbf{x}$ and Y have no relation to the set of $\mathbf{X}_i$'s and Y_i's in $\mathcal{D}$. Again, we can view $\mathcal{D}$ as the training data, and $(\mathbf{x}, Y)$ as a test point — the test point $\mathbf{x}$ is probably not even in the training set $\mathcal{D}$! Mathematically, we express our metric as the expected squared error between

the hypothesis and the observation $Y = f(\mathbf{x}) + Z$:

$$\varepsilon(\mathbf{x}; h) = \mathbb{E}[(h(\mathbf{x}; \mathcal{D}) - Y)^2]$$

The expectation here is over two random variables, $\mathcal{D}$ and Y:

$$\mathbb{E}_{\mathcal{D},Y}[(h(\mathbf{x}; \mathcal{D}) - Y)^2] = \mathbb{E}_{\mathcal{D}}[\mathbb{E}_Y[(h(\mathbf{x}; \mathcal{D}) - Y)^2|\mathcal{D}]]$$

Note that the error is w.r.t the observation Y and not the true underlying model $f(\mathbf{x})$, because we do not know the true model and only have access to the noisy observations from the true model.

1.2 Bias-Variance Decomposition

The error metric is difficult to interpret and work with, so let's try to decompose it into parts that are easier to understand. Before we start, let's find the expectation and variance of Y:

$$\mathbb{E}[Y] = \mathbb{E}[f(\mathbf{x}) + Z] = f(\mathbf{x}) + \mathbb{E}[Z] = f(\mathbf{x})$$

$$\text{Var}(Y) = \text{Var}(f(\mathbf{x}) + Z) = \text{Var}(Z)$$

Also, in general for any random variable X, we have that

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \implies \mathbb{E}[X^2] = \text{Var}(X) + \mathbb{E}[X]^2$$

Let's use these facts to decompose the error:

$$\begin{aligned}
\varepsilon(\mathbf{x}; h) &= \mathbb{E}[(h(\mathbf{x}; \mathcal{D}) - Y)^2] = \mathbb{E}[h(\mathbf{x}; \mathcal{D})^2] + \mathbb{E}[Y^2] - 2\mathbb{E}[h(\mathbf{x}; \mathcal{D}) \cdot Y] \\
&= \Big(\text{Var}(h(\mathbf{x}; \mathcal{D})) + \mathbb{E}[h(\mathbf{x}; \mathcal{D})]^2\Big) + \Big(\text{Var}(Y) + \mathbb{E}[Y]^2\Big) - 2\mathbb{E}[h(\mathbf{x}; \mathcal{D})] \cdot \mathbb{E}[Y] \\
&= \Big(\mathbb{E}[h(\mathbf{x}; \mathcal{D})]^2 - 2\mathbb{E}[h(\mathbf{x}; \mathcal{D})] \cdot \mathbb{E}[Y] + \mathbb{E}[Y]^2\Big) + \text{Var}(h(\mathbf{x}; \mathcal{D})) + \text{Var}(Y) \\
&= \Big(\mathbb{E}[h(\mathbf{x}; \mathcal{D})] - \mathbb{E}[Y]\Big)^2 + \text{Var}(h(\mathbf{x}; \mathcal{D})) + \text{Var}(Y) \\
&= \underbrace{\Big(\mathbb{E}[h(\mathbf{x}; \mathcal{D})] - f(\mathbf{x})\Big)^2}_{bias^2 \text{ of method}} + \underbrace{\text{Var}(h(\mathbf{x}; \mathcal{D}))}_{\text{variance of method}} + \underbrace{\text{Var}(Z)}_{\text{irreducible error}}
\end{aligned}$$

Recall that for any two independent random variables $\mathcal{D}$ and Y, $g_1(\mathcal{D})$ and $g_2(Y)$ are also independent, for any functions g_1, g_2. This implies that $h(\mathbf{x}; \mathcal{D})$ and Y are independent, allowing us to express $\mathbb{E}[h(\mathbf{x}; \mathcal{D}) \cdot Y] = \mathbb{E}[h(\mathbf{x}; \mathcal{D})] \cdot \mathbb{E}[Y]$ in the second line of the derivation. The final decomposition, also known as the **bias-variance decomposition**, consists of three terms:

- **Bias^2 of method**: Measures how well the *average* hypothesis (over all possible training sets) can come close to the true underlying value $f(\mathbf{x})$, for a fixed value of $\mathbf{x}$. A low bias means that on average the regressor $h(\mathbf{x})$ accurately estimates $f(\mathbf{x})$.
- **Variance of method**: Measures the variance of the hypothesis (over all possible training sets), for a fixed value of $\mathbf{x}$. A low variance means that the prediction does not change much as the training set varies. An un-biased method (bias = 0) could have a large variance.

- **Irreducible error**: This is the error in our model that we cannot control or eliminate, because it is due to errors inherent in our noisy observation Y.

The decomposition allows us to measure the error in terms of bias, variance, and irreducible error. Irreducible error has no relation with the hypothesis model, so we can fully ignore it in theory when minimizing the error. As we have discussed before, models that are very complex have very little bias because on *average* they can fit the true underlying model value $f(\mathbf{x})$ very well, but have very high variance and are very far off from $f(\mathbf{x})$ on an individual basis.

Note that the error above is only for a fixed input $\mathbf{x}$, but in regression our goal is to minimize the average error over all possible values of $\mathbf{X}$. If we know the distribution for $\mathbf{X}$, we can find the effectiveness of a hypothesis model as a whole by taking an expectation of the error over all possible values of $\mathbf{x}$: $\mathbb{E}_{\mathbf{X}}[\varepsilon(\mathbf{x}; h)]$.

1.3 Alternative Decomposition

The previous derivation is short, but may seem somewhat arbitrary. Let's explore an alternative derivation. At its core, it uses the technique that $\mathbb{E}[(Z-Y)^2] = \mathbb{E}[((Z-\mathbb{E}[Z]) + (\mathbb{E}[Z]-Y))^2]$ which decomposes to easily give us the variance of Z and other terms.

$$
\begin{aligned}
\varepsilon(\mathbf{x}; h) &= \mathbb{E}[(h(\mathbf{x};\mathcal{D}) - Y)^2] \\
&= \mathbb{E}\left[\left(h(\mathbf{x};\mathcal{D}) - \mathbb{E}[h(\mathbf{x};\mathcal{D})] + \mathbb{E}[h(\mathbf{x};\mathcal{D})] - Y\right)^2\right] \\
&= \mathbb{E}\left[\left(h(\mathbf{x};\mathcal{D}) - \mathbb{E}[h(\mathbf{x};\mathcal{D})]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}[h(\mathbf{x};\mathcal{D})] - Y\right)^2\right] + 2\mathbb{E}\left[\left(h(\mathbf{x};\mathcal{D}) - \mathbb{E}[h(\mathbf{x};\mathcal{D})]\right) \cdot \left(\mathbb{E}[h(\mathbf{x};\mathcal{D})] - Y\right)\right] \\
&= \mathbb{E}\left[\left(h(\mathbf{x};\mathcal{D}) - \mathbb{E}[h(\mathbf{x};\mathcal{D})]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}[h(\mathbf{x};\mathcal{D})] - Y\right)^2\right] + \cancel{2\mathbb{E}[h(\mathbf{x};\mathcal{D}) - \mathbb{E}[h(\mathbf{x};\mathcal{D})]]} \cdot \mathbb{E}[\mathbb{E}[h(\mathbf{x};\mathcal{D})] - Y] \\
&= \mathbb{E}\left[\left(h(\mathbf{x};\mathcal{D}) - \mathbb{E}[h(\mathbf{x};\mathcal{D})]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}[h(\mathbf{x};\mathcal{D})] - Y\right)^2\right] \\
&= \mathrm{Var}((h(\mathbf{x};\mathcal{D})) + \mathbb{E}\left[\left(\mathbb{E}[h(\mathbf{x};\mathcal{D})] - Y\right)^2\right] \\
&= \mathrm{Var}((h(\mathbf{x};\mathcal{D})) + \mathbb{E}\left[\left(\mathbb{E}[h(\mathbf{x};\mathcal{D})] - \mathbb{E}[Y] + \mathbb{E}[Y] - Y\right)^2\right] \\
&= \mathrm{Var}((h(\mathbf{x};\mathcal{D})) + \mathbb{E}\left[\left(\mathbb{E}[h(\mathbf{x};\mathcal{D})] - \mathbb{E}[Y]\right)^2\right] + \mathbb{E}[(Y - \mathbb{E}[Y])^2] + 2\left(\mathbb{E}[h(\mathbf{x};\mathcal{D})] - \mathbb{E}[Y]\right) \cdot \cancel{\mathbb{E}[\mathbb{E}[Y] - Y]} \\
&= \mathrm{Var}((h(\mathbf{x};\mathcal{D})) + \mathbb{E}\left[\left(\mathbb{E}[h(\mathbf{x};\mathcal{D})] - \mathbb{E}[Y]\right)^2\right] + \mathbb{E}[(Y - \mathbb{E}[Y])^2] \\
&= \mathrm{Var}((h(\mathbf{x};\mathcal{D})) + \left(\mathbb{E}[h(\mathbf{x};\mathcal{D})] - \mathbb{E}[Y]\right)^2 + \mathrm{Var}(Y) \\
&= \mathrm{Var}((h(\mathbf{x};\mathcal{D})) + \left(\mathbb{E}[h(\mathbf{x};\mathcal{D})] - f(\mathbf{x})\right)^2 + \mathrm{Var}(Z) \\
&= \underbrace{\left(\mathbb{E}[h(\mathbf{x};\mathcal{D})] - f(\mathbf{x})\right)^2}_{bias^2 \text{ of method}} + \underbrace{\mathrm{Var}(h(\mathbf{x};\mathcal{D}))}_{\text{variance of method}} + \underbrace{\mathrm{Var}(Z)}_{\text{irreducible error}}
\end{aligned}
$$

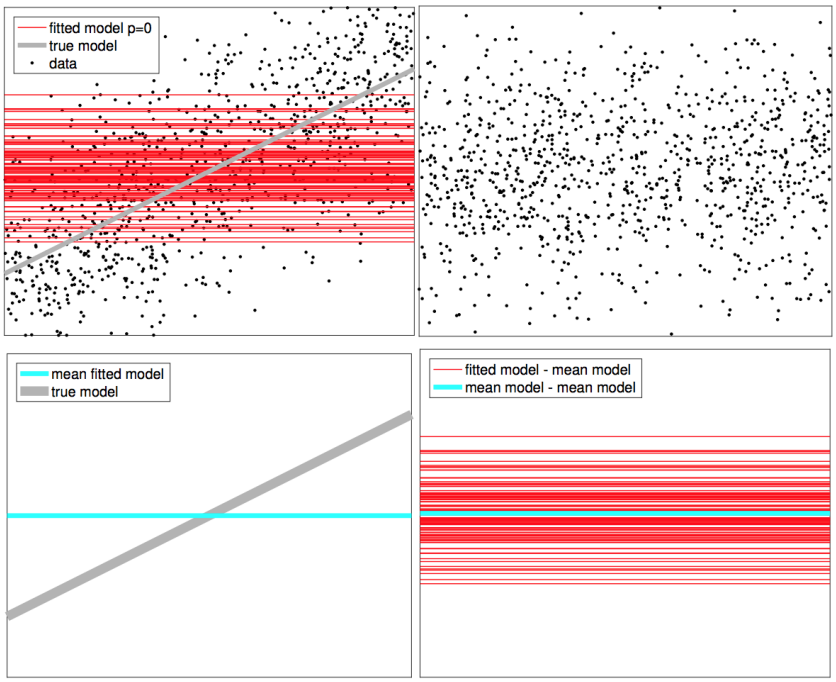
1.4 Experiments

Let's confirm the theory behind the bias-variance decomposition with an empirical experiment that measures the bias and variance for polynomial regression with 0 degree, 1st degree, and 2nd degree polynomials. In our experiment, we will repeatedly fit our hypothesis model to a random training set. We then find the expectation and variance of the fitted models generated from these training sets.

Let's first look at a 0 degree (constant) regression model. We repeatedly fit an optimal constant line to a training set of 10 points. The true model is denoted by gray and the hypothesis is denoted by red. Notice that at each time the red line is slightly different due to the different training set used.

Let's combine all of these hypotheses together into one picture to see the bias and variance of our model.

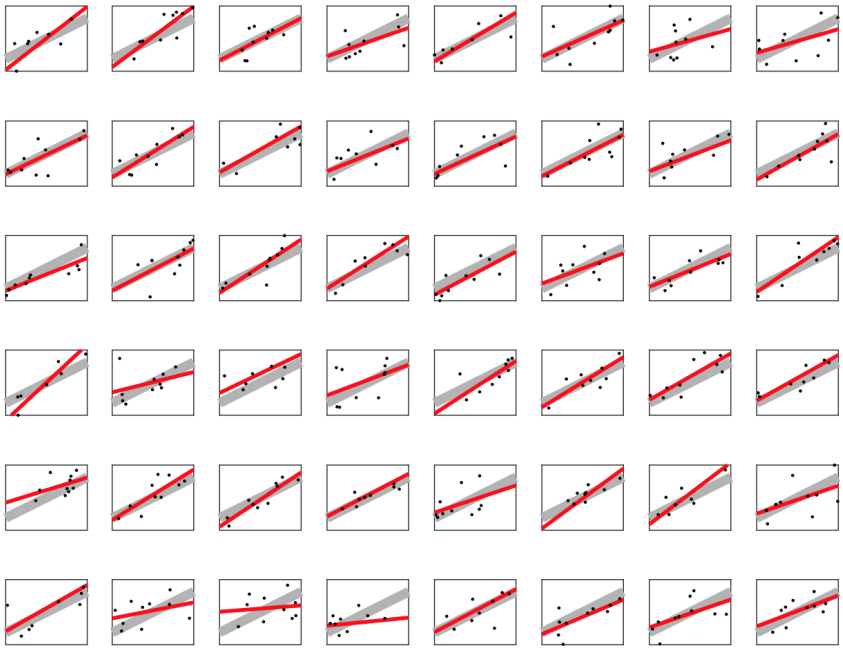
Bias and Variance in Model Selection: $p = 0$



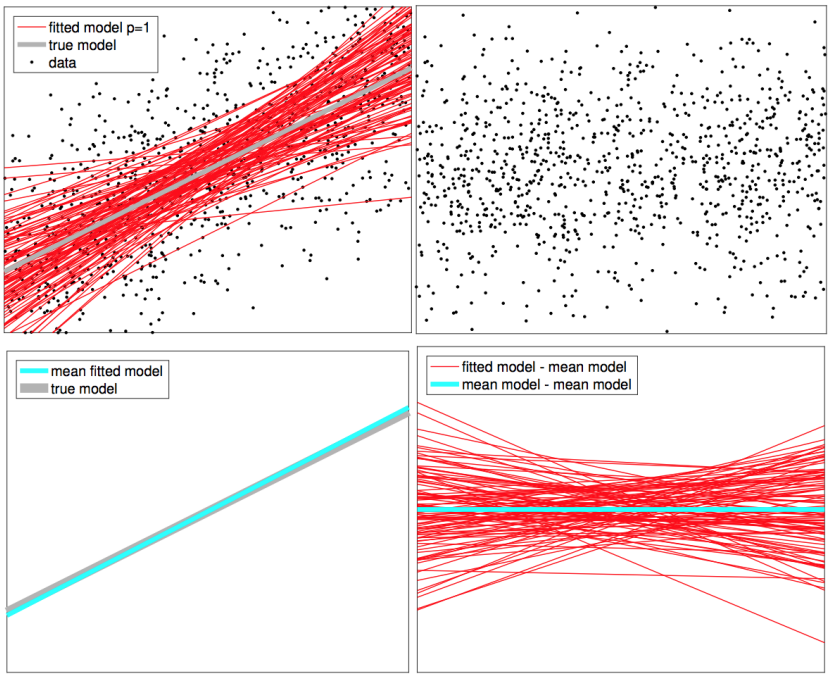
On the top left diagram we see all of our hypotheses and all training sets used. The bottom left diagram shows the average hypothesis in cyan. As we can see, this model has low bias for $\mathbf{x}$'s in the center of the graph, but very high bias for $\mathbf{x}$'s that are away from the center of the graph. The diagram in the bottom right shows that the variance of the hypotheses is quite high, for all values of $\mathbf{x}$.

Now let's look at a 1st degree (linear) regression model.

Fitting A Model over Multiple Datasets: $p = 1$

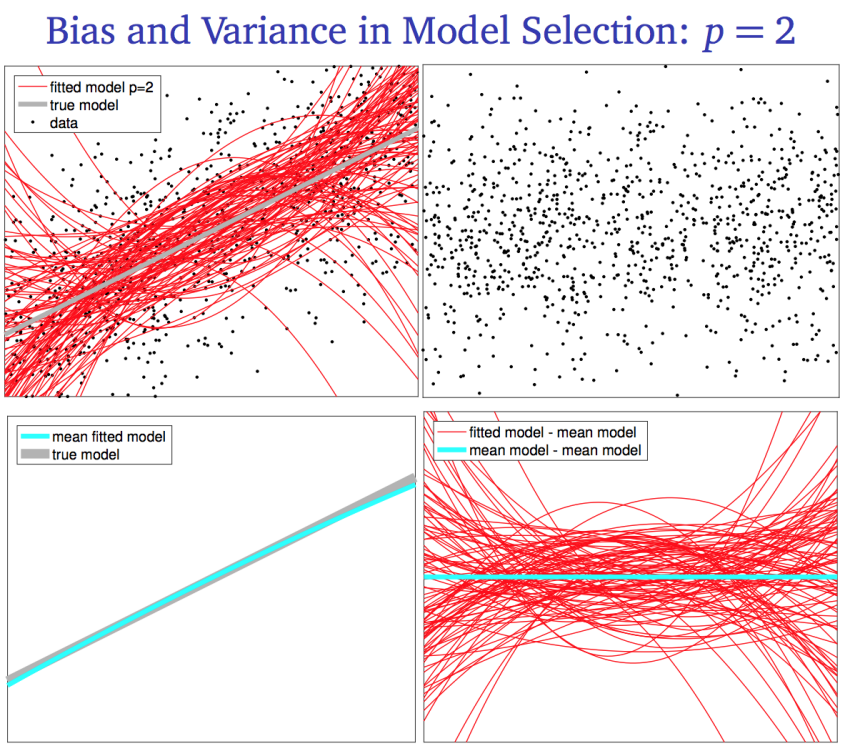


Bias and Variance in Model Selection: $p = 1$



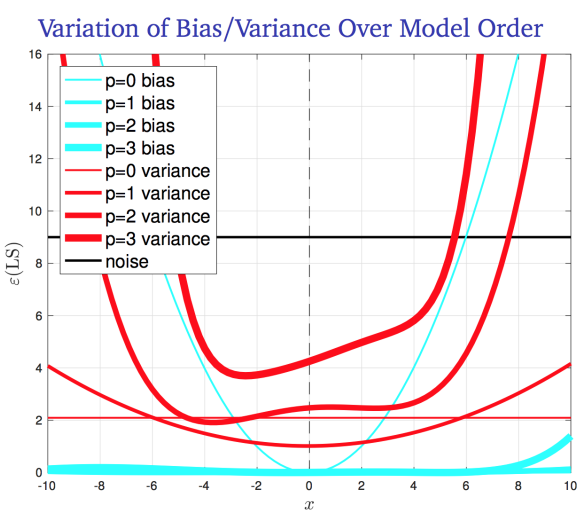
The bias is now very low bias for all $\mathbf{x}$'s. The variance is low for $\mathbf{x}$'s in the middle of the graph, but higher for $\mathbf{x}$'s that are away from the center of the graph.

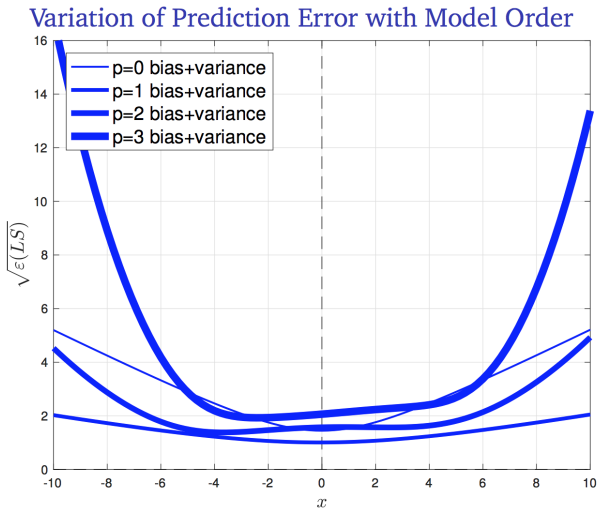
Finally, let's look at a 2nd degree (quadratic) regression model.

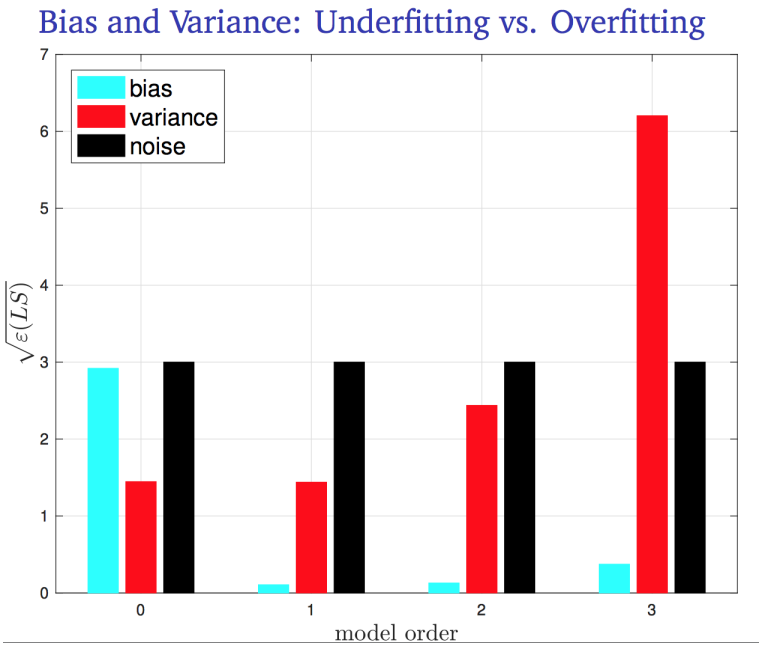


The bias is still very low for all $\mathbf{x}$'s. However, the variance is much higher for all values of $\mathbf{x}$.

Let's summarize our results. We find the bias and the variance empirically and graph them for all values of $\mathbf{x}$, as shown in the first two graphs. Finally, we take an expectation over the bias and variance over all values of $\mathbf{x}$, as shown in the third graph.







The bias-variance decomposition confirms our understanding that the true model is linear. While a quadratic model achieves the same theoretical bias as a linear model, it overfits to the data, as indicated by its high variance. On the other hand a constant model underfits the data, as indicated by its high bias. In the process of training our model, we can tell that a constant model is a poor choice, because its high bias is reflected in poor training error. However we cannot tell that a quadratic model is poor, because its high variance is not reflected in the training error. This is the reason why we use validation data and cross-validation as a means to measure the performance of our hypothesis model on unseen data.

1.5 Takeaways

Let us conclude by stating some implications of the Bias-Variance Decomposition:

1. Underfitting is equivalent to high bias; most overfitting correlates to high variance.

2. Training error reflects bias but not variance. Test error reflects both. In practice, if the training error is much smaller than the test error, then there is overfitting.

3. Variance $\to 0$ as $n \to \infty$.

4. If f is in the set of hypothesis functions, bias will decrease with more data. If f is not in the set of hypothesis functions, then there is an underfitting problem and more data won't help.

5. Adding good features will decrease the bias, but adding a bad feature rarely increase the bias. (just set the coefficient of the feature to 0)

6. Adding a feature usually increase the variance, so a feature should only be added if it decreases bias more than it increases variance.

7. Irreducible error can not be reduced.

8. Noise in the test set only affects $\mathrm{Var}(Z)$, but noise in the training set also affects bias and variance.

9. For real-world data, f is rarely known, and the noise model might be wrong, so we can't calculate bias and variance. But we can test algorithms over synthetic data.