# Derivation of Bias-Variance Decomposition Theorem

## Department of Electrical and Computer Engineering

**The material contained in this note is completely optional and will not be tested.**

The note seeks to derive the bias-variance decomposition theorem with lots of details. There are portions, which I took liberally from Wikipedia.

Suppose $y = f(x) + \epsilon$, where $f$ is assumed to be deterministic and $\epsilon$ is random with mean 0 and variance $\sigma^2$. Note that $\epsilon$ does not need to be Gaussian distributed.

Suppose we are given a training set consisting of $D = \{(x_1, y_1), (x_2, y_2) \cdots (x_N, y_N)\}$, where sample $(x_n, y_n)$ is randomly sampled from some distribution $p(x, y) = p(x)p(y|x)$. Because $D$ is randomly sampled, there is a probability distribution $p(D)$ associated with $D$.

Using the training data $D$, we want to find a function $\hat{f}_D(x)$ that predicts $y$ well, for a new test sample $x$. In this note, we will show that for this new test sample $x$,

$$E_{p(D)}\left[\left(y(x) - \hat{f}_D(x)\right)^2\right] = \left(\text{Bias}_D(\hat{f}_D(x))\right)^2 + \text{Var}_D(\hat{f}_D(x)) + \sigma^2, \tag{1}$$

where

$$\text{Bias}_D(\hat{f}_D(x)) = E_{p(D)}(\hat{f}_D(x)) - f(x), \tag{2}$$

and

$$\text{Var}_D(\hat{f}_D(x)) = E_{p(D)}\left(\hat{f}_D^2(x)\right) - \left(E_{p(D)}(\hat{f}_D(x))\right)^2 \tag{3}$$

$$= E_{p(D)}\left[\left(\hat{f}_D(x) - E_{p(D)}(\hat{f}_D(x))\right)^2\right] \tag{4}$$

The expectation ranges over different possible training sets $D$ all sampled from the same joint distribution $p(x, y)$.

**Interpretation.** According to wikipedia (which I am quoting), the three terms can be interpreted as:

1. The square of the bias of the learning method, which can be thought of as the error caused by the simplifying assumptions built into the method, e.g., when approximating a non-linear function $f(x)$ using a learning method for linear models, there will be error in the estimates $\hat{f}_D(x)$.

2. The variance of the learning method, or, intuitively, how much the learning method $\hat{f}_D(x)$ will move around its mean.

3. The irreducible error $\sigma^2$

Since all three terms are non-negative, this forms a lower bound on the expected error on unseen samples. The more complex the model $\hat{f}(x)$, the more data points it will capture, and the lower the bias will be. However, complexity will make the model "move" more to capture the data points, and hence its variance will be larger.

**Important subtleties not discussed in Wikipedia:** It is not the case that complex models will always lead to lower bias! This is because $\hat{f}_D(x)$ is dependent on $D$. So if the model is very complex and $D$ is small, then the bias can actually be very big!

**Proof**. Note that we will simplify the notation, such that $E_{p(D)}$ will be written as $E_D$. Wikipedia removes "$D$" completely, which I found confusing. I will also make the dependence on the new test sample $x$ explicit (unlike Wikipedia). This is actually quite important for one of the steps.

$$E_D\left[\left(y(x)-\hat{f}_D(x)\right)^2\right] \tag{5}$$

$$= E_D\left[\left(f(x)+\epsilon(x)-\hat{f}_D(x)\right)^2\right] \tag{6}$$

$$= E_D\left[\left(f(x)+\epsilon(x)-\hat{f}_D(x)+E_D\left[\hat{f}_D(x)\right]-E_D\left[\hat{f}_D(x)\right]\right)^2\right] \tag{7}$$

$$= E_D\left[\left(f(x)-E_D\left[\hat{f}_D(x)\right]+\epsilon(x)+E_D\left[\hat{f}_D(x)\right]-\hat{f}_D(x)\right)^2\right] \quad \text{rearranging the terms} \tag{8}$$

$$\begin{aligned}
&= E_D\left[\left(f(x)-E_D\left[\hat{f}_D(x)\right]\right)^2\right]+E_D\left[\epsilon^2(x)\right]+E_D\left[\left(E_D\left[\hat{f}_D(x)\right]-\hat{f}_D(x)\right)^2\right]\\
&+2E_D\Big[\left(f(x)-E_D\left[\hat{f}_D(x)\right]\right)\epsilon(x)\Big]+2E_D\Big[\epsilon(x)\left(E_D\left[\hat{f}_D(x)\right]-\hat{f}_D(x)\right)\Big]\\
&+2E_D\Big[\left(f(x)-E_D\left[\hat{f}_D(x)\right]\right)\left(E_D\left[\hat{f}_D(x)\right]-\hat{f}_D(x)\right)\Big]
\end{aligned} \tag{9}$$

Let's simplify the many terms in Eq.(9)

1. The first term

$$E_D\left[\left(f(x)-E_D\left[\hat{f}_D(x)\right]\right)^2\right] \tag{10}$$

$$= \left(f(x)-E_D\left[\hat{f}_D(x)\right]\right)^2 \quad \text{because } f(x) \text{ \& } E_D\left[\hat{f}_D(x)\right] \text{ are both non-random} \tag{11}$$

$$= \left(\text{Bias}_D(\hat{f}_D(x))\right)^2 \quad \text{from the definition in Eq. 2} \tag{12}$$

2. The second term $E_D\left[\epsilon^2(x)\right] = \sigma^2$ because $\epsilon(x)$ has a mean of 0 and variance of $\sigma^2$). For any random variable $Z$, $\mathrm{Var}(Z) = E\left[(Z - E(Z))^2\right] = E[Z^2] - (E(Z))^2$, so $E[Z^2] = \mathrm{Var}(Z) + (E(Z))^2$. Therefore $E(\epsilon^2(x)) = \mathrm{Var}(\epsilon) = \sigma^2$

3. The third term

$$E_D\left[\left(E_D\left[\hat{f}_D(x)\right] - \hat{f}_D(x)\right)^2\right] = \mathrm{Var}_D(\hat{f}_D(x)) \text{ from the definition in Eq. 4} \tag{13}$$

4. The fourth term is 0 because

$$2E_D\Big[\left(f(x) - E_D\left[\hat{f}_D(x)\right]\right)\epsilon(x)\Big] = 2\left(f(x) - E_D\left[\hat{f}_D(x)\right]\right)E_D\left[\epsilon(x)\right] \tag{14}$$

because $\left(f(x) - E_D\left[\hat{f}_D(x)\right]\right)$ is non-random so we can just pull it out of the expectation. Since $E_D\left[\epsilon(x)\right] = 0$, so the fourth term is just 0.

5. The fifth term is also 0. Note that $E(AB) = E(A)E(B)$ when $A$, $B$ are independent.

$$2E_D\left[\epsilon(x)\left(E_D\left[\hat{f}_D(x)\right] - \hat{f}_D(x)\right)\right] = 2E_D\left[\epsilon\right]E_D\left[E_D\left[\hat{f}_D(x)\right] - \hat{f}_D(x)\right] \tag{15}$$

because the noise in the new test sample $x$ (i.e., $\epsilon(x)$) and $\hat{f}_D(x)$ are independent. Note that since $\hat{f}$ is estimated from the training data, so it is dependent on the noise in the training data, BUT it is independent from the noise from the new test sample $x$ (i.e., $\epsilon(x)$). This is why I have made the dependence on $x$ explicit here.

6. The sixth term is also 0.

$$2E_D\Big[\left(f(x) - E_D\left[\hat{f}_D(x)\right]\right)\left(E_D\left[\hat{f}_D(x)\right] - \hat{f}_D(x)\right)\Big] \tag{16}$$

$$= 2\left(f(x) - E_D\left[\hat{f}_D(x)\right]\right)E_D\Big[\left(E_D\left[\hat{f}_D(x)\right] - \hat{f}_D(x)\right)\Big] \tag{17}$$

because $\left(f(x) - E_D\left[\hat{f}_D(x)\right]\right)$ is non-random, so we can pull it out of the expectation. The second term in the above expression:

$$E_D\Big[\left(E_D\left[\hat{f}_D(x)\right] - \hat{f}_D(x)\right)\Big] = E_D\left[\hat{f}_D(x)\right] - E_D\left[\hat{f}_D(x)\right] = 0, \tag{18}$$

so the sixth term is also 0.

Therefore we have proven the bias-variance decomposition theorem: $E_{p(D)}\left[\left(y(x) - \hat{f}_D(x)\right)^2\right] = \left(\mathrm{Bias}_D(\hat{f}_D(x))\right)^2 + \mathrm{Var}_D(\hat{f}_D(x)) + \sigma^2$

**MSE for infinite number of new samples.** Note that the above equation is for one single new sample $x$. Suppose we have a very, very big test set sampled from $x \sim p(x)$, then the mean squared error (MSE) of this test set would be

$$\text{MSE} = E_{p(x)}\Bigg(\Big(\text{Bias}_D(\hat{f}_D(x))\Big)^2 + \text{Var}_D(\hat{f}_D(x))\Bigg) + \sigma^2 \tag{19}$$