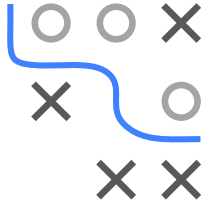
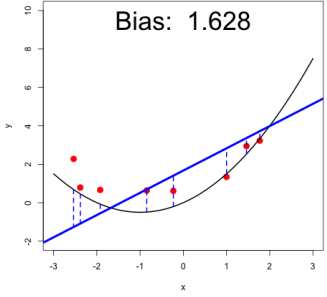
Introduction to Machine Learning



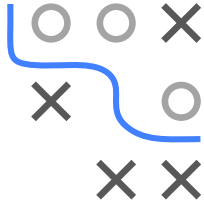
Advanced Risk Minimization
Bias-Variance Decomposition



Learning goals

- Understand how to decompose the generalization error of a learner into
  - bias of the learner
  - variance of the learner
  - inherent noise in the data

BIAS-VARIANCE DECOMPOSITION



Let us take a closer look at the generalization error of a learning algorithm $\mathcal{I}_L$. This is the expected error of an induced model $\hat{f}_{\mathcal{D}_n}$, on training sets of size n, when applied to a fresh, random test observation.

$$GE_n\left(\mathcal{I}_L\right) = \mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}_{xy}^n,(\mathbf{x},y)\sim\mathbb{P}_{xy}}\left(L\left(y, \hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)\right) = \mathbb{E}_{\mathcal{D}_n,xy}\left(L\left(y, \hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)\right)$$

We therefore need to take the expectation over all training sets of size n, as well as the independent test observation.
For the squared loss, there is a nice additive decomposition of $GE_n\left(\mathcal{I}_L\right)$ into three components.*
Hence we assume that the data is generated by

$$y = f_{\text{true}}(\mathbf{x}) + \epsilon\,,$$

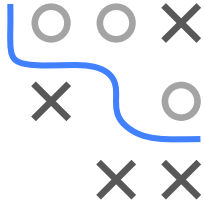
with zero-mean homoskedastic error $\epsilon \sim (0, \sigma^2)$ independent of $\mathbf{x}$.

* Similar decomps also exist for other losses expressible as Bregman divergences (e.g. cross-entropy). One exception is the $0/1$ loss (Brown and Ali 2024)

BIAS-VARIANCE DECOMPOSITION / 2

$$GE_n\left(\mathcal{I}_L\right) =$$

$$\underbrace{\sigma^2}_{\text{Variance of the data}} + \mathbb{E}_{xy}\underbrace{\left[\text{Var}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x}) \mid \mathbf{x}, y\right)\right]}_{\text{Variance of learner at } (\mathbf{x},y)} + \mathbb{E}_{xy}\underbrace{\left[\left(\left(f_{\text{true}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)\right)^2 \mid \mathbf{x}, y\right)\right]}_{\text{Squared bias of learner at } (\mathbf{x},y)}$$



1. The first term expresses the variance of the data. This is pure **noise** in the data. Also called Bayes, intrinsic or irreducible error. No matter what we do, we will never get below this error.
2. The second term expresses, on average, how much $\hat{f}_{\mathcal{D}_n}(\mathbf{x})$ fluctuates around test points if we vary the training data. Expresses also the learner's tendency to learn random things irrespective of the real signal (overfitting).
3. The third term says how much we are "off" on average at test locations (underfitting). Models with high capacity typically have low **bias** and *vice versa*.

BIAS-VARIANCE DECOMPOSITION / 3

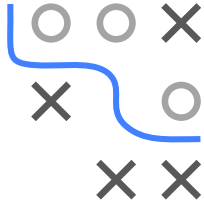
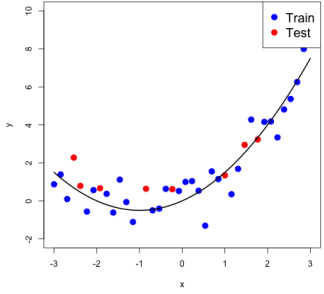


Illustration: Let us consider the following example. We will generate a dataset using the following model :

$$y = x + \frac{x^2}{2} + \epsilon\,, \quad \epsilon \sim N(0,1)$$

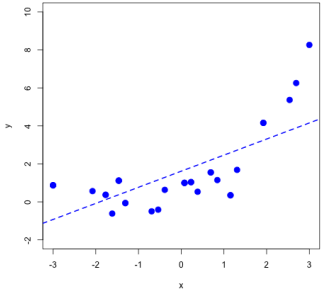
The data is then split into a training set and a test set.

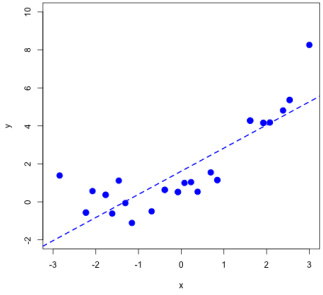


BIAS-VARIANCE DECOMPOSITION / 4

To obtain estimates for the bias and variance, we will train several models by sampling with replacement from the training data. This is commonly known as **bootstrapping**.

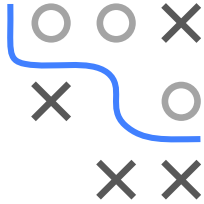
First, we train several (low capacity) linear models (polynomial of degree $d = 1$).

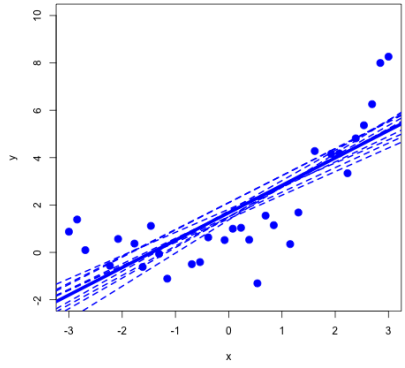




BIAS-VARIANCE DECOMPOSITION / 5

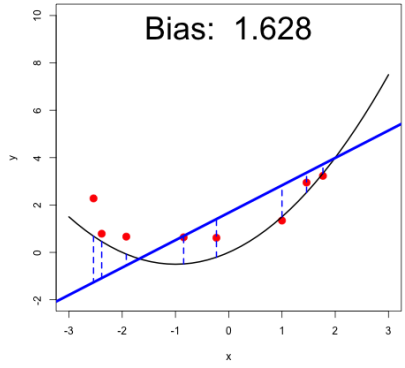
By creating several models, we obtain the average model over different samples of the training dataset.





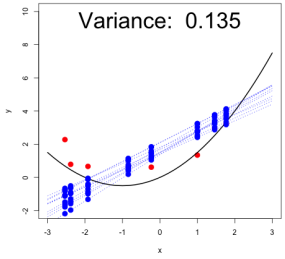
BIAS-VARIANCE DECOMPOSITION / 6

We can now estimate the (squared) bias, by computing the average squared difference between the average model and the true model, at the test point locations.



BIAS-VARIANCE DECOMPOSITION / 7

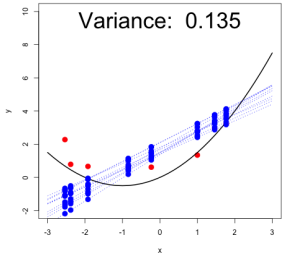
We compute the average variance of the predictions of the models we trained at the test point locations.

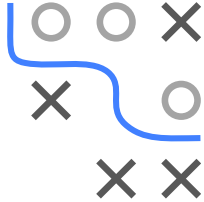


$$GE_n\left(\mathcal{I}_L\right) \approx 1 + 1.628 + 0.135 = 2.763$$

- The biggest component of the generalization error is the bias.
- Computing the MSE in the usual way for each model, via L2 loss, and then averaging over models gives rise to nearly the same value, as expected

BIAS-VARIANCE DECOMPOSITION / 8

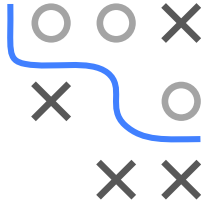


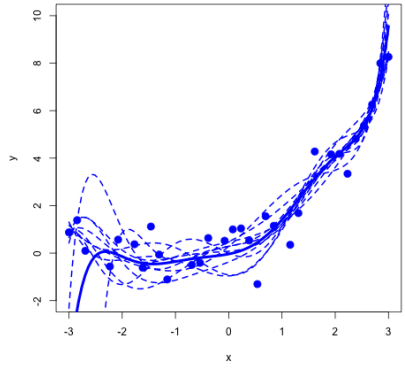


- We can now check whether this alternative computation of the GE is correct
- So, we simply compute the MSE in the standard fashion for each model
- So for each model we compute the L2 loss at each data point, then average
- Then we average these MSEs over all models
- Result = 2.72, would be closer if we average over more models and test points

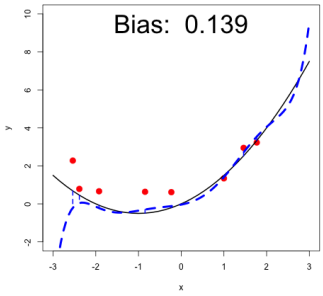
BIAS-VARIANCE DECOMPOSITION / 9

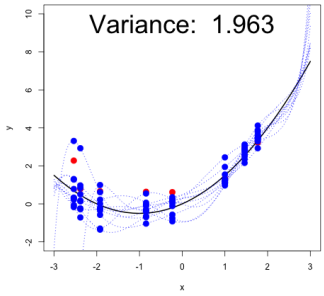
We will repeat the same procedure, but use a high-degree polynomial ($d = 7$) with more capacity.

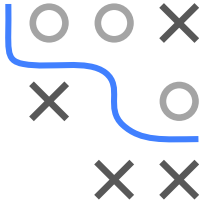




BIAS-VARIANCE DECOMPOSITION / 10





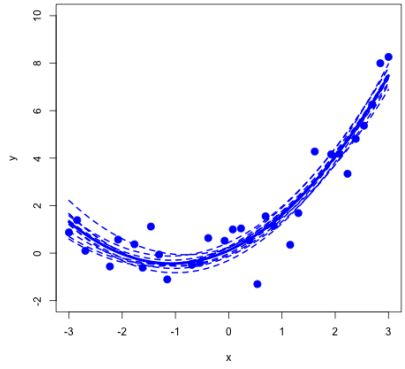


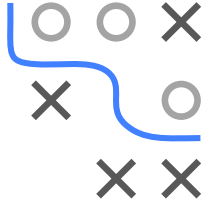
$$GE_n(\mathcal{I}_L) \approx 1 + 0.139 + 1.963 = 3.102$$

- The generalization error is higher than before
- Even though the bias is lower, the variance of the learner is higher.

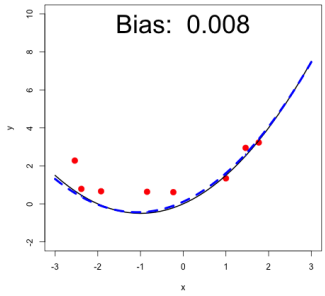
BIAS-VARIANCE DECOMPOSITION / 11

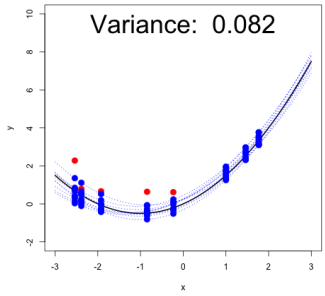
What happens if we use a model with the same complexity as the true model (quadratic polynomial)?

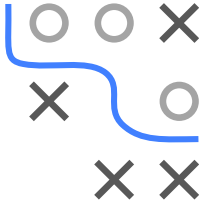




BIAS-VARIANCE DECOMPOSITION / 12



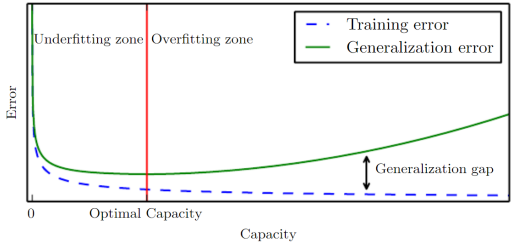




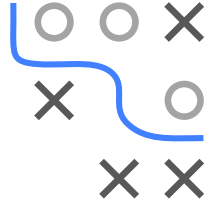
$$GE_n\left(\mathcal{I}_L\right) \approx 1 + 0.008 + 0.082 = 1.091$$

- The generalization error is the lowest at this complexity.
- The variance of the data acts as a lower bound.

CAPACITY AND OVERFITTING

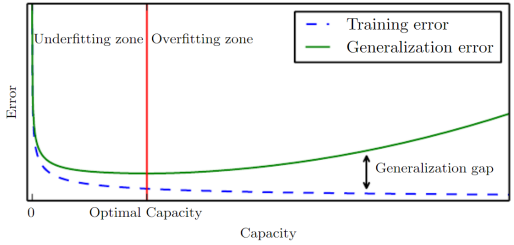


Credit: Ian Goodfellow

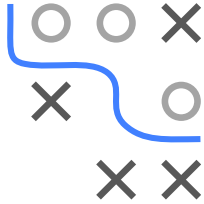


- The performance of a learner depends on its ability to
  1. **fit** the training data well
  2. **generalize** to new data
- Failure of the first point is called **underfitting**
- Failure of the second item is called **overfitting**

CAPACITY AND OVERFITTING / 2



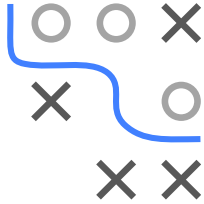
Credit: Ian Goodfellow



- The tendency of a model to underfit/overfit is a function of its capacity, determined by the type of hypotheses it can learn.
- Usually, low bias means high capacity, which in turn means a higher chance of overfitting
- Low-bias models usually have also higher variance
- For such models, regularization (we discuss later) is essential
- Even for correctly specified models, the generalization error is lower-bounded by the irreducible noise σ^2

APPROXIMATION AND ESTIMATION

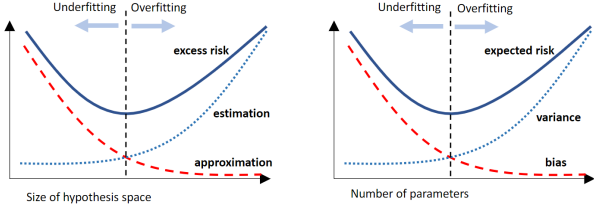
▸ Brown and Ali 2024



The Bias-Variance decomp is often confused or equated with the related (but different) decomp of **excess risk** into **estimation** and **approximation** error.

$$\underbrace{\mathcal{R}(\hat{f}_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}_{all}})}_{\text{excess risk}} = \underbrace{\mathcal{R}(\hat{f}_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}})}_{\text{estimation error}} + \underbrace{\mathcal{R}(f^*_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}_{all}})}_{\text{approx. error}}$$

Both are commonly described using the same figure and analogies

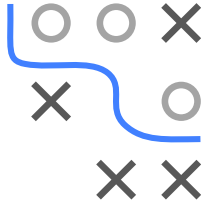


Credit: Brown and Ali 2024

NB: It should be noted that the bias-variance decomp. only holds for certain losses, while the above decomposition is universal.

APPROX./ESTIMATION ERROR

Brown and Ali 2024



- The approx. error is a structural property of $\mathcal{H}$.
- The estimation error is random due to dependence on data in $\hat{f}$
- Estimation error arises because we choose $f \in \mathcal{H}$ with limited training data using $\mathcal{R}_{\text{emp}}$ instead of $\mathcal{R}$

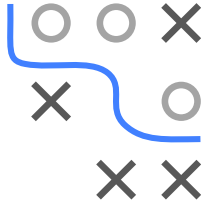
Knowing $\hat{f}_{\mathcal{H}} \in \arg\inf_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f)$ assumes we found a global minimizer of $\mathcal{R}_{\text{emp}}$, which is often impossible (e.g. DNNs).

In practice, optimizing $\mathcal{R}_{\text{emp}}$ gives us our 'best guess' $\tilde{f}_{\mathcal{H}} \in \mathcal{H}$ of $\hat{f}_{\mathcal{H}}$. We can now decompose its excess risk finer as

$$\underbrace{\mathcal{R}(\tilde{f}_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}_{all}})}_{\text{excess risk}} = \underbrace{\mathcal{R}(\tilde{f}_{\mathcal{H}}) - \mathcal{R}(\hat{f}_{\mathcal{H}})}_{\text{optim. error}} + \underbrace{\mathcal{R}(\hat{f}_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}})}_{\text{estimation error}} + \underbrace{\mathcal{R}(f^*_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}_{all}})}_{\text{approx. error}}$$

Note that the optimization error can be negative, even though $\mathcal{R}_{\text{emp}}(\tilde{f}_{\mathcal{H}}) \geq \mathcal{R}_{\text{emp}}(\hat{f}_{\mathcal{H}})$ always holds.

APPROX./ESTIMATION ERROR ▸ Brown and Ali 2024 / 2



We can further decompose estimation error more finely by defining the *centroid* model or 'systematic' model part.

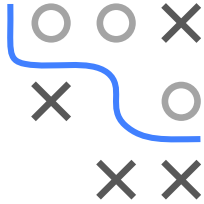
Given $\hat{f}_{\mathcal{H}} \in \arg\min_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f)$ the centroid model under squared loss is the mean prediction at each x over all $\mathcal{D}_n$, $f^{\circ}_{\mathcal{H}} := \mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}^n_{xy}}[\hat{f}_{\mathcal{H}}]$,.

With $f^{\circ}_{\mathcal{H}}$ we can decompose the expected estimation error as

$$\underbrace{\mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}^n_{xy}}\left[\mathcal{R}(\hat{f}_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}})\right]}_{\text{expected estimation error}} = \underbrace{\mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}^n_{xy}}\left[\mathcal{R}(\hat{f}_{\mathcal{H}}) - \mathcal{R}(f^{\circ}_{\mathcal{H}})\right]}_{\text{estimation variance}} + \underbrace{\mathcal{R}(f^{\circ}_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}})}_{\text{estimation bias}}$$

- estimation bias measures distance of centroid model to risk minimizer over $\mathcal{H}$
- estimation variance measures spread of ERM around centroid model induced by randomness due to $\mathcal{D}_n$

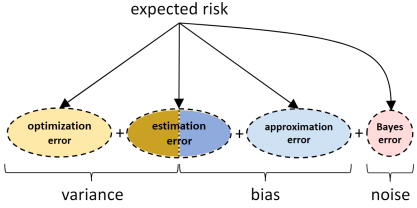
APPROX./ESTIMATION ERROR ▸ Brown and Ali 2024 / 3



We can now connect the derived quantities back to bias and variance and see how they differ. As we see, bias is not only approx. error and variance is not estimation error.

$$\text{bias} = \text{approximation error} + \text{estimation bias}$$

$$\text{variance} = \text{optimization error} + \text{estimation variance}$$



Credit: Brown and Ali 2024

NB: For special case of LM and squared loss (OLS), we have zero optim. error and estimation bias. Hence both decompositions agree there.