Finite Sample-Size Considerations

Bias-Variance Decomposition in Linear Regression

Kiran Bagale

June 2025

Overfitting in ML Estimation

- ML/OLS solutions can be unstable or non-unique due to complete reliance on the training data.
- This is often called the **overfitting problem**.
- We model:

$$d = f(\boldsymbol{x}, \boldsymbol{w}) + \varepsilon$$

  where $f(\boldsymbol{x}, \boldsymbol{w})$ is deterministic and ε is the expectational error.
- The purpose of this second model is to encode the empirical knowledge represented by the training sample t, as

$$t \rightarrow \hat{w}$$

Stochastic vs. Physical Models

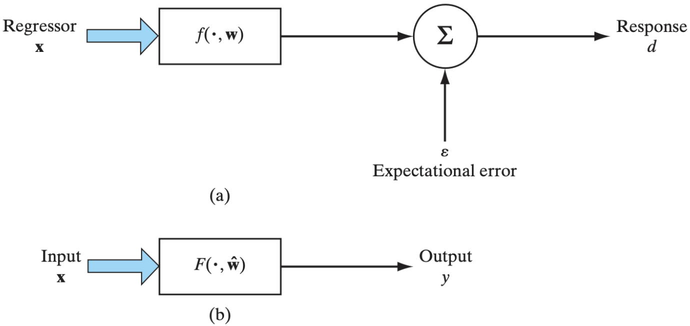


FIGURE 2.4 (a) Mathematical model of a stochastic environment, parameterized by the vector **w**. (b) Physical model of the environment, where $\hat{\mathbf{w}}$ is an estimate of the unknown parameter vector **w**.

- (a): Mathematical model with true parameter $\boldsymbol{w}$ and noise ε
- (b): Physical model with estimated parameter $\hat{\boldsymbol{w}}$
- Output:

$$y = F(\boldsymbol{x}, \hat{\boldsymbol{w}})$$

Cost Function and Approximation

- Cost function:

$$e(\hat{\boldsymbol{w}}) = \frac{1}{2}\sum_{i=1}^{N}(d_i - F(\boldsymbol{x}_i, \hat{\boldsymbol{w}}))^2$$

- Reformulated as:

$$e(\hat{\boldsymbol{w}}) = \frac{1}{2}\mathbb{E}_t\left[(f(\boldsymbol{x}, \boldsymbol{w}) - F(\boldsymbol{x}, t))^2\right] + \frac{1}{2}\mathbb{E}_t[\varepsilon^2]$$

- First term is the key measure:

$$L_{\text{av}} = \mathbb{E}_t\left[(f(\boldsymbol{x}, \boldsymbol{w}) - F(\boldsymbol{x}, t))^2\right]$$

Bias–Variance Decomposition

- Let:

$$f(\boldsymbol{x}, \boldsymbol{w}) = \mathbb{E}[d|\boldsymbol{x}]$$

- Decompose error:

$$L_{\text{av}} = \underbrace{B^2(\hat{\boldsymbol{w}})}_{\text{Bias}^2} + \underbrace{V(\hat{\boldsymbol{w}})}_{\text{Variance}} + \underbrace{\sigma_\epsilon^2}_{\text{Irreducible Error}}$$

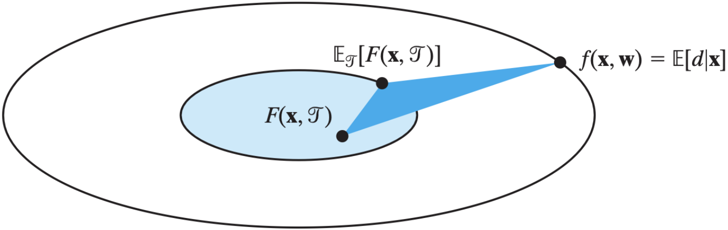
- Where:

$$B(\hat{\boldsymbol{w}}) = \mathbb{E}_t[F(\boldsymbol{x}, t)] - \mathbb{E}[d|\boldsymbol{x}]$$

$$V(\hat{\boldsymbol{w}}) = \mathbb{E}_t\left[(F(\boldsymbol{x}, t) - \mathbb{E}_t[F(\boldsymbol{x}, t)])^2\right]$$

- **Bias**^{2}: How much predicted values differ from true values.
- **Variance**: How predictions made on the same value vary on different realizations of the model.
- **Irreducible Error** (σ_ϵ^2): Noise inherent in the data.

Illustration of Bias and Variance



FIG

Decomposition of the natural measure Lav(f(x, w), F(x, wˆ)), into bias and variance terms for linear regression models.

- $\mathbb{E}[d|\boldsymbol{x}]$ is the true regression function.
- $F(\boldsymbol{x}, t)$ is a sample-dependent estimate.
- Bias: distance between true expectation and average model.
- Variance: spread of sample models around their average.

Bias error and Variance error

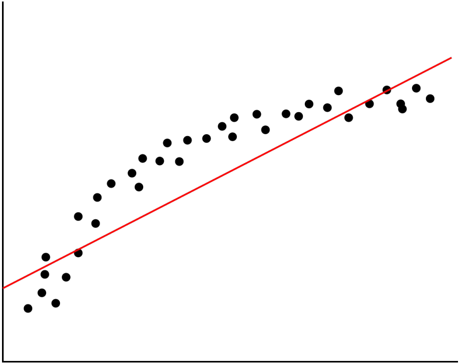


FIG: High bias model(underfitting)

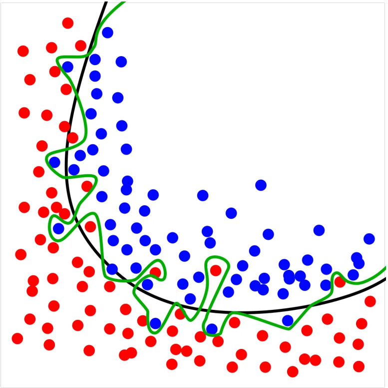


FIG: High variance model(overfitting)

Bias–Variance Dilemma

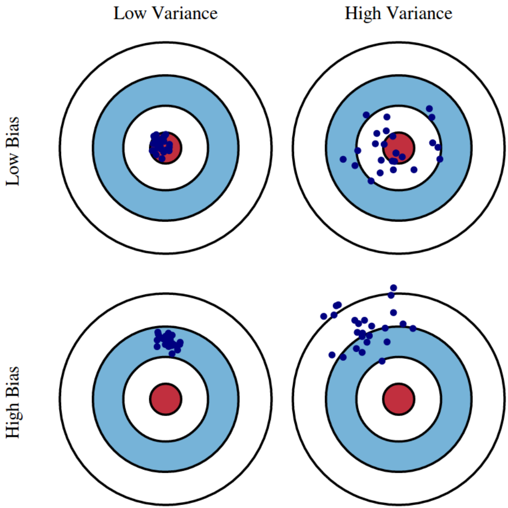


FIG: Graphical illustration of bias and variance.

Bias–Variance Dilemma Cont...

- Small training sets: hard to achieve low bias and low variance.
- Reducing bias $\rightarrow$ higher variance, and vice versa.
- Only with very large samples can both be minimized.
- Regularization or architecture constraints can help reduce variance by introducing a ”harmless” bias.

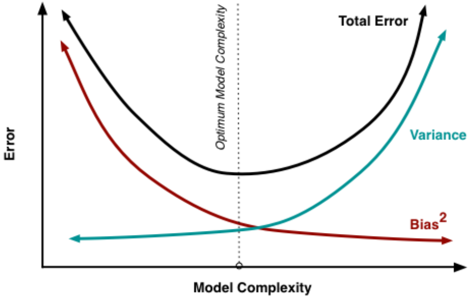


Fig: The variation of Bias and Variance with the model complexity. This is similar to the concept of overfitting and underfitting.

More complex models overfit, while the simplest models under-fit.

Conclusion

- Bias–variance decomposition explains generalization behavior.
- Training set size and model complexity critically affect performance.
- Practical tradeoff:
  - Small bias $\Rightarrow$ large variance
  - Large bias $\Rightarrow$ stable model
- Bias should be purposeful and aligned with the problem.

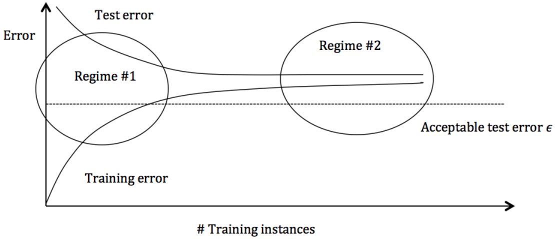


Fig: Test and training error as the number of training instances increases.