The Bias–Variance Tradeoff

Shreeyesh Menon

1 Introduction

A central goal in predictive modelling is to build a model that generalizes well: it should perform well not just on the data we used to estimate it, but also on new, unseen data. The *bias–variance tradeoff* provides a powerful lens for understanding why this is often difficult.

Econometricians already know the tension between model simplicity and flexibility: a linear model may be too restrictive, while a very flexible non-parametric model may overfit. The bias–variance tradeoff formalizes this tension and helps justify concepts used later in machine learning, such as cross–validation and regularization.

2 Model "Risk"

Consider a data-generating process

$$y = f(x) + \varepsilon,$$

where f is the true (unknown) function and ε is noise with mean zero and variance σ^2.

Suppose we fit a model $\hat{f}(x)$ using a random training sample. The expected prediction error at a point x_0 is

$$\mathbb{E}\left[(y_0 - \hat{f}(x_0))^2\right].$$

A key result in statistical learning theory states that this error can be decomposed as

$$\underbrace{\big(\mathbb{E}[\hat{f}(x_0)] - f(x_0)\big)^2}_{\text{Bias}^2} + \underbrace{\mathbb{V}\big(\hat{f}(x_0)\big)}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Irreducible noise}}.$$

- **Bias** measures how far the average prediction is from the truth. Simple models (e.g. linear regression when f is nonlinear) tend to have *high bias*.
- **Variance** measures how sensitive the model is to the specific training sample. Very flexible models (e.g. deep trees, high-degree polynomials) tend to have *high variance*.
- **Irreducible noise** is noise inherent in the data, which no model can eliminate.

3 The Tradeoff

The tradeoff arises because reducing bias usually increases variance, and vice versa. For example, moving from a linear model to a flexible nonparametric model reduces bias but increases variance.

Economists will recognize the analogy with misspecification versus overfitting: a model that is too simple fails to capture important structure (bias), while a model that is too flexible tries to capture noise (variance).

4 Illustration

Figure 1 shows a simple conceptual picture. As model complexity increases, bias decreases while variance increases. The optimal model complexity minimizes their sum.

5 Implications

Understanding the bias–variance tradeoff helps motivate:

- **Cross–validation**, which estimates prediction error on unseen data.
- **Regularization** (e.g. ridge, LASSO), which intentionally adds bias to reduce variance.
- **Model selection** procedures that balance flexibility and generalization.

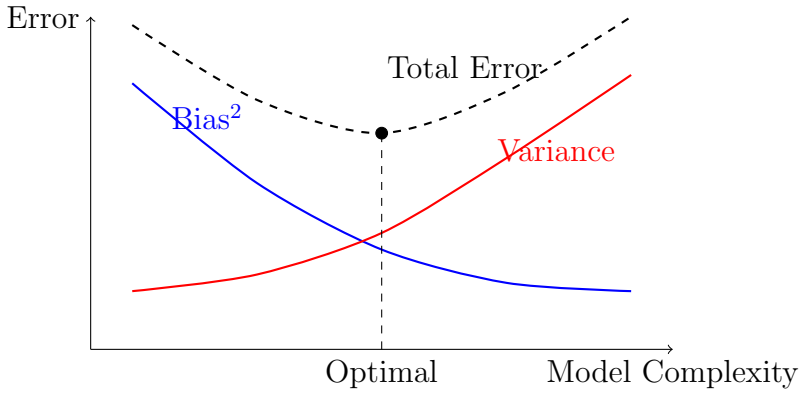

Figure 1: Bias–Variance Tradeoff

6 Beyond the Bias–Variance Tradeoff: The Double Descent Phenomenon

The classical bias–variance tradeoff predicts a U-shaped test error curve: as model complexity increases, the test error first falls (lower bias) and then rises (higher variance). This picture assumes that the model has fewer parameters than data points.

Modern machine learning, however, often operates in *highly overparameterized* regimes. In such cases, empirical results show a striking pattern known as **double descent**, depicted in Figure 2.

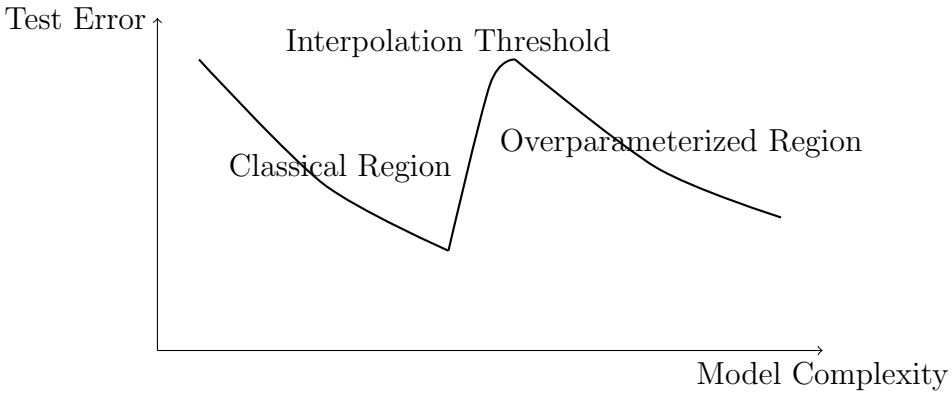

Figure 2: The Double Descent Phenomenon

The Interpolation Threshold

As the number of parameters approaches the number of training samples, the model becomes able to interpolate the data:

$$\hat{f}(x_i) = y_i \quad \text{for all training points.}$$

Training error drops to zero, but test error often spikes due to extremely high variance. This spike marks the *interpolation threshold*.

In our polynomial regression example with 20 training points, this occurs around degree $d = 19$ or 20.

The Second Descent

Strikingly, as model complexity increases even further, test error may fall again. This "second descent" occurs because many modern learning algorithms implicitly select the *minimum-norm* interpolating solution, which tends to generalize better than arbitrary interpolating solutions.

In the polynomial simulation, once $d > 20$, the Moore–Penrose pseudoinverse picks the minimum-norm solution, and test performance can improve again.

Implications

The double descent phenomenon reshapes how we think about model complexity:

- Overparameterization does not automatically cause overfitting.
- The choice of solution (e.g., minimum-norm) matters as much as the number of parameters.
- Many modern models (e.g., deep neural networks) operate far to the right of the interpolation threshold.

For economists entering data science, double descent highlights that high-dimensional models can generalize surprisingly well, challenging classical intuitions based solely on parameter counts.