Bias variance trade-off

Monday, 9 February 2026 12:29 PM

Bias–Variance Tradeoff

In supervised learning, the goal of a model is to learn an underlying true function $f(x)$ from data and make accurate predictions on unseen samples. However, prediction error arises due to multiple sources, primarily **bias** and **variance**.

Error Decomposition

For a regression problem with squared loss, the expected prediction error at a point x can be decomposed as:

$$\mathbb{E}[(y - \hat{f}(x))^2] = \underbrace{\text{Bias}^2}_{\text{systematic error}} + \underbrace{\text{Variance}}_{\text{sensitivity to data}} + \underbrace{\sigma^2}_{\text{irreducible noise}}$$

where σ^2 represents noise inherent in the data that no model can eliminate.

Bias

Bias measures the error introduced by approximating a real-world problem with a simplified model.

$$\text{Bias}(x) = \mathbb{E}[\hat{f}(x)] - f(x)$$

Characteristics

- High bias models make **strong assumptions**
- Often too simple to capture complex patterns
- Leads to **underfitting**
- Example: Linear regression on highly nonlinear data

Variance

Variance measures how much the model's predictions change when trained on different datasets.

$$\text{Variance}(x) = \mathbb{E}_{\downarrow}[\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2]$$

Characteristics

- High variance models are **overly complex**
- Highly sensitive to training data
- Leads to **overfitting**
- Example: High-degree polynomial fitting noisy data

Bias–Variance Tradeoff

Bias and variance have an **inverse relationship**:

- Increasing model complexity ↓ bias but ↑ variance
- Decreasing model complexity ↑ bias but ↓ variance

The objective is **not** to minimize bias or variance individually, but to find an **optimal balance** that minimizes total generalization error.

Regularization

Regularization is a technique used to control model complexity by adding a penalty term to the loss function. It discourages overly large model parameters, helping reduce variance and improve generalization.

Ordinary Least Squares (OLS)

$$\mathcal{L}_{OLS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

OLS minimizes training error only, which can lead to overfitting.

L2 Regularization (Ridge Regression)

$$\mathcal{L}_{Ridge} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} w_j^2$$

$w_1^2 + w_2^2 + w_3^2$

tuning knob.

Effect

- Penalizes **squared magnitude** of weights
- Shrinks coefficients smoothly toward zero
- Reduces variance without eliminating features
- Increases bias slightly to lower total error

L1 Regularization (Lasso Regression)

$$\mathcal{L}_{Lasso} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |w_j|$$

cost func

Effect

- Encourages **sparse solutions**
- Forces some weights exactly to zero
- Performs **feature selection**
- Can increase bias more than Ridge

Elastic Net Regularization

$$\mathcal{L}_{ElasticNet} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \left(\alpha \sum |w_j| + (1 - \alpha) \sum w_j^2 \right)$$

Elastic Net Regularization

$$\mathcal{L}_{ElasticNet} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \left(\alpha \sum |w_j| + (1 - \alpha) \sum w_j^2 \right)$$

Effect

- Combines strengths of L1 and L2
- Handles correlated features well
- Balances sparsity and stability

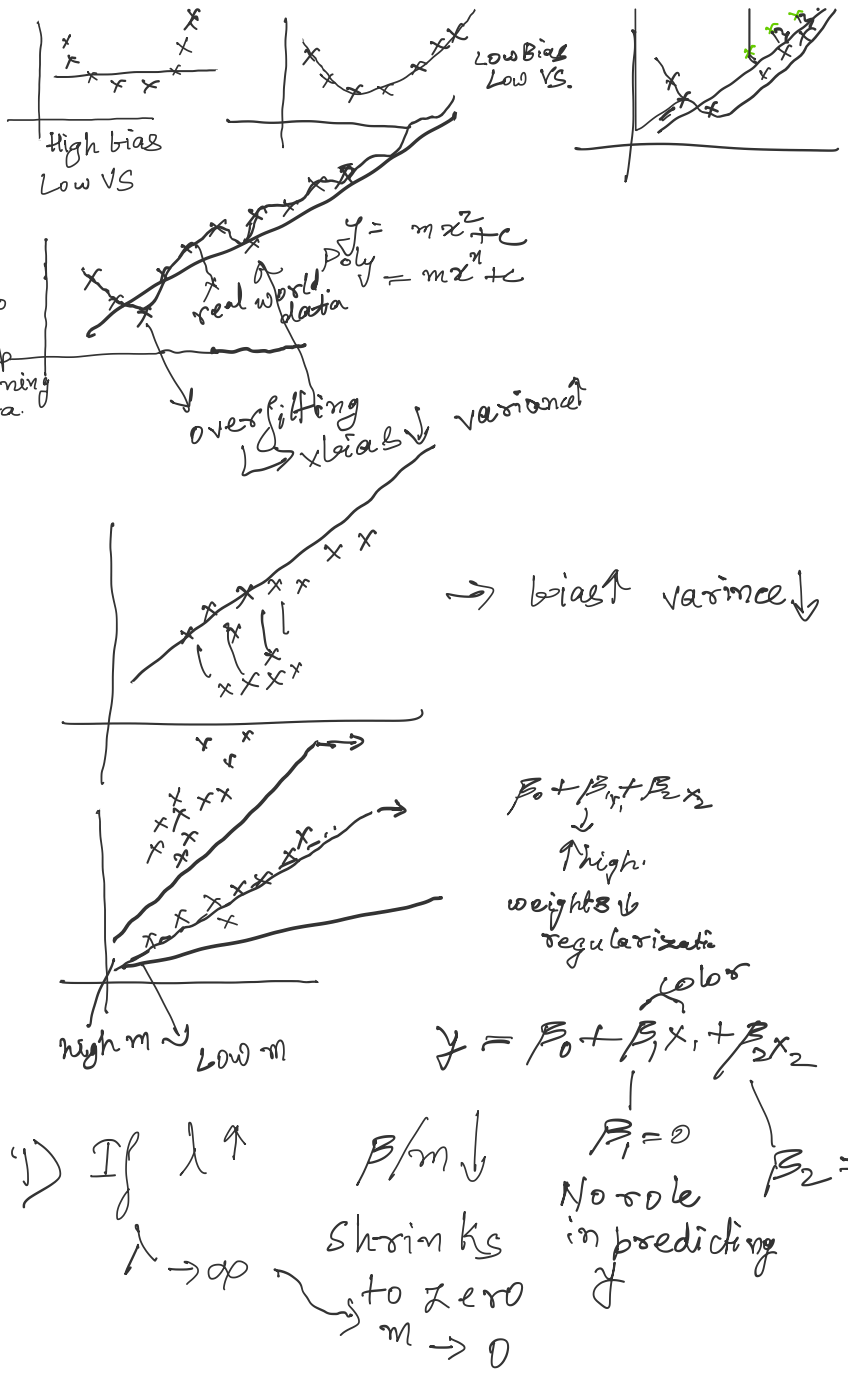
Impact of Regularization on Bias and Variance

Model	Bias	Variance
No regularization	Low	High
Ridge (L2)	Slightly higher	Lower
Lasso (L1)	Higher	Much lower
Elastic Net	Balanced	Balanced

Regularization **intentionally increases bias** to achieve a **larger reduction in variance**, resulting in better generalization performance.

Key Takeaways

- Bias–variance tradeoff is fundamental to machine learning
- Overfitting and underfitting are direct consequences of imbalance
- Regularization is a principled way to control complexity
- The best model minimizes **total expected error**, ↓ot training error



2) Higher coefficients are impacted more

3) Bias Variance trade off

Elastic Net

Ridge

Lasso

$$L = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \lambda |w|^2 \qquad L = \sum_{i=1} (Y_i - Y_i) + \lambda |w|$$

when we donknow if all the features are important or not.

EN

$$\Rightarrow L = \sum (Y_i - \hat{Y}_i)^2 + \alpha |w|^2 + b |w|$$

$$\begin{cases} \lambda = a + b \\ L\_ratio = \dfrac{a}{a+b} \\ L_1 = \dfrac{a}{\lambda} \end{cases}$$

$a = L_1 \times \lambda$

$b = \lambda - a$

Also used in multicolinearity.