Bias - Variance Tradeoff in ML

→ Used to understand the tradeoff between the simplicity of the model and its generalizability.

Bias: Error in model predictions as ML approximates a real-world variable with a simple, presupposed model architecture.

Simpler models introduce higher bias since typically, they have fewer parameters

Variance: Error in prediction due to a model's sensiti small fluctuations in the training dataset, since a model may be too complex and thus captures noise as a signal.

Bias

Leads to underfitting as the model misses important relationships between the target variable & the features.

Variance

Leads to overfitting as the model performs poorly on test data due to an overly complex architecture

Trade-off: To find an optimally complex model

$$\rightarrow \begin{pmatrix}\text{Total}\\ \text{error}\end{pmatrix} = (\text{Bias})^2 + (\text{Variance}) + \begin{pmatrix}\text{Irreducible}\\ \text{error}\end{pmatrix}$$

Let y denote the real-world measurement of the target variable

f denote the true model

ϵ error in the true model $\longrightarrow$ assumed to be normally distributed.

$$y = f + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

estimate of irreducible error that even a true model will incur

Total error $\longrightarrow$ Squared error $E[(y-\hat{f})^2]$

best-fit model using a given dataset.

$$\boxed{\text{Error}} = E[(y-\hat{f})^2]$$

$$= E[y^2 + \hat{f}^2 - 2y\hat{f}]$$

$$= E[y^2] + E[\hat{f}^2] - E[2y\hat{f}]$$

$$\text{Var}(\hat{f}) = E[(\hat{f} - E(\hat{f}))^2]$$

$$= E\left[(E(\hat{f}))^2 + \hat{f}^2 - 2\hat{f}E(\hat{f})\right]$$

$$= (E(\hat{f}))^2 + E(\hat{f}^2) - 2E(\hat{f}E(\hat{f}))$$

$$Var(\hat{f}) = (E(\hat{f}))^2 + E(\hat{f}^2) - 2(E(\hat{f}))^2 = \boxed{E(\hat{f}^2) - (E(\hat{f}))^2}$$

$$\longrightarrow E(\hat{f}^2) = Var(\hat{f}) + (E(\hat{f}))^2$$

Squared Error $= E[y^2] + Var(\hat{f}) + (E(\hat{f}))^2 - E[2y\hat{f}]$

We need $E[y^2]$. Recall that $y = f + \epsilon$ → irr er

$$E[y^2] = E[(f+\epsilon)^2] = E[f^2] + E[2f\epsilon] + E[\epsilon^2]$$

Since the analysis is done for a given data point. Thus,

$$E[y^2] = f^2 + \underline{E(\epsilon)\cdot 2f} + E[\epsilon^2]$$

Recall that $\epsilon \sim N(0, \sigma^2)$

$$E(\epsilon) = 0$$

$$\sigma^2 = E[(\epsilon - \cancel{E(\epsilon)})^2] = E(\epsilon^2)$$

$$\boxed{E[y^2] = f^2 + 0 + \sigma^2 = (f^2 + \sigma^2)}$$

$$E[2y\hat{f}] = 2E[(f+\epsilon)\cdot\hat{f}] = 2[E(f\hat{f}) + \underline{E(\epsilon\hat{f})}]$$

$$= 2[f E(\hat{f}) + \cancel{E(\epsilon)} E(\hat{f})]$$

error for a datapoint and model prediction are independent

$$\boxed{\cancel{E}[2y\hat{f}] \;=\; 2f\,E(\hat{f})}$$

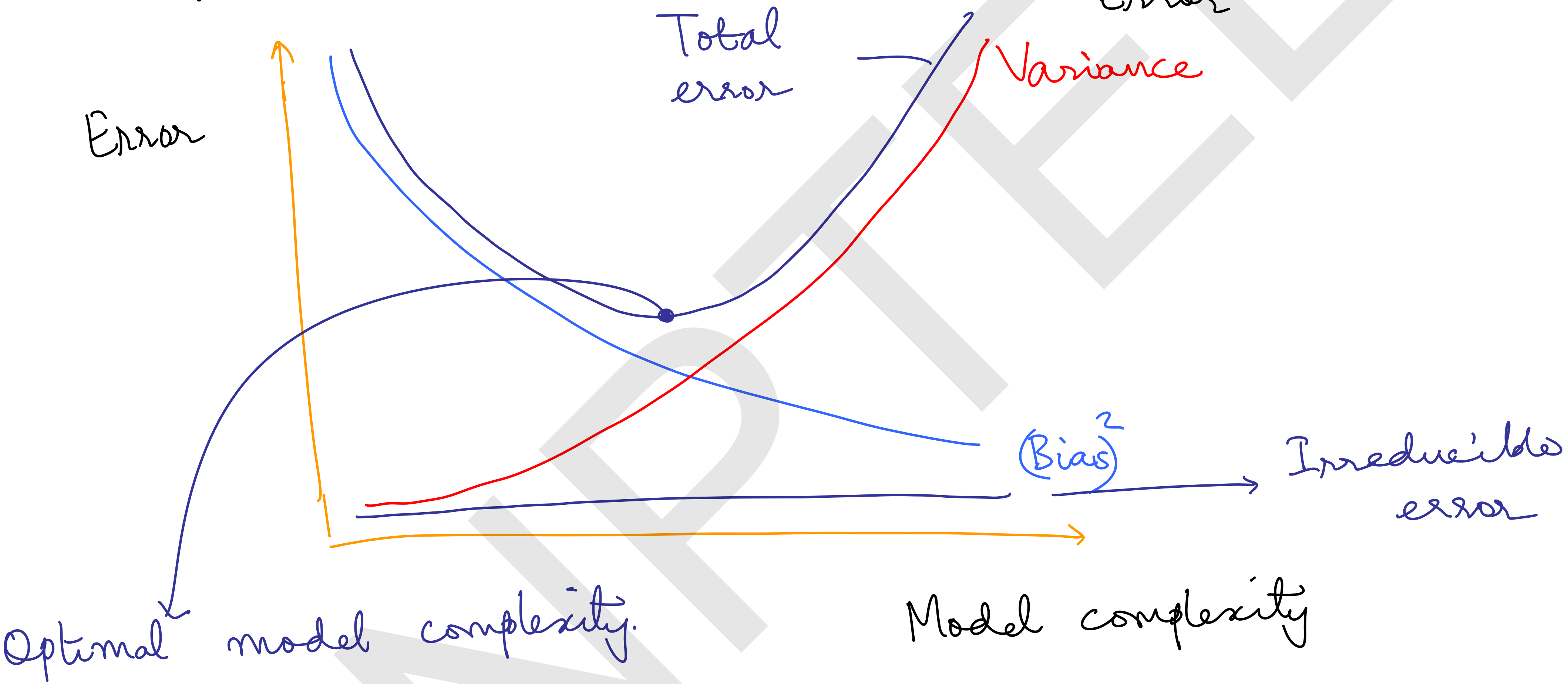
$$\text{Squared error} = E[y^2] + \mathrm{Var}(\hat{f}) + (E(\hat{f}))^2 - 2E[y\hat{f}]$$

$$= (f^2 + \sigma^2) + \mathrm{Var}(\hat{f}) + (E(\hat{f}))^2 - 2f\,E(\hat{f})$$

$$= \underbrace{(E[f-\hat{f}])^2}_{\text{Bias}^2} + \mathrm{Var}(\hat{f}) + \sigma^2$$

$$\boxed{\text{Squared error} = (\text{Bias}(\hat{f}))^2 + \mathrm{Var}(\hat{f}) + \sigma^2}$$

$$\text{Total error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$



To decrease bias: Increase model complexity

To decrease variance: Employing dimensionality reduction, regularization, or feature selection