Data Science II

(P8106)

Department of Biostatistics
Mailman School of Public Health
Columbia University

Spring 2024

Bias-Variance trade-off

- Fit a model $\widehat{f}(x)$ to some training data
- Let (x_0, y_0) be a test observation drawn from the population
- $\mathrm{Bias}(\widehat{f}(x_0)) = E(\widehat{f}(x_0)) - f(x_0)$
- For a given x_0, expected test MSE
  $E(y_0 - \widehat{f}(x_0))^2 = \mathrm{Var}(\widehat{f}(x_0)) + [\mathrm{Bias}(\widehat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$

$$\Big(\underbrace{y_0 - f(x_0)}_{\varepsilon_0} + \underbrace{f(x_0) - E\hat{f}(x_0)}_{-\text{bias}} + \underline{E\hat{f}(x_0) - \hat{f}(x_0)}\Big)^2$$

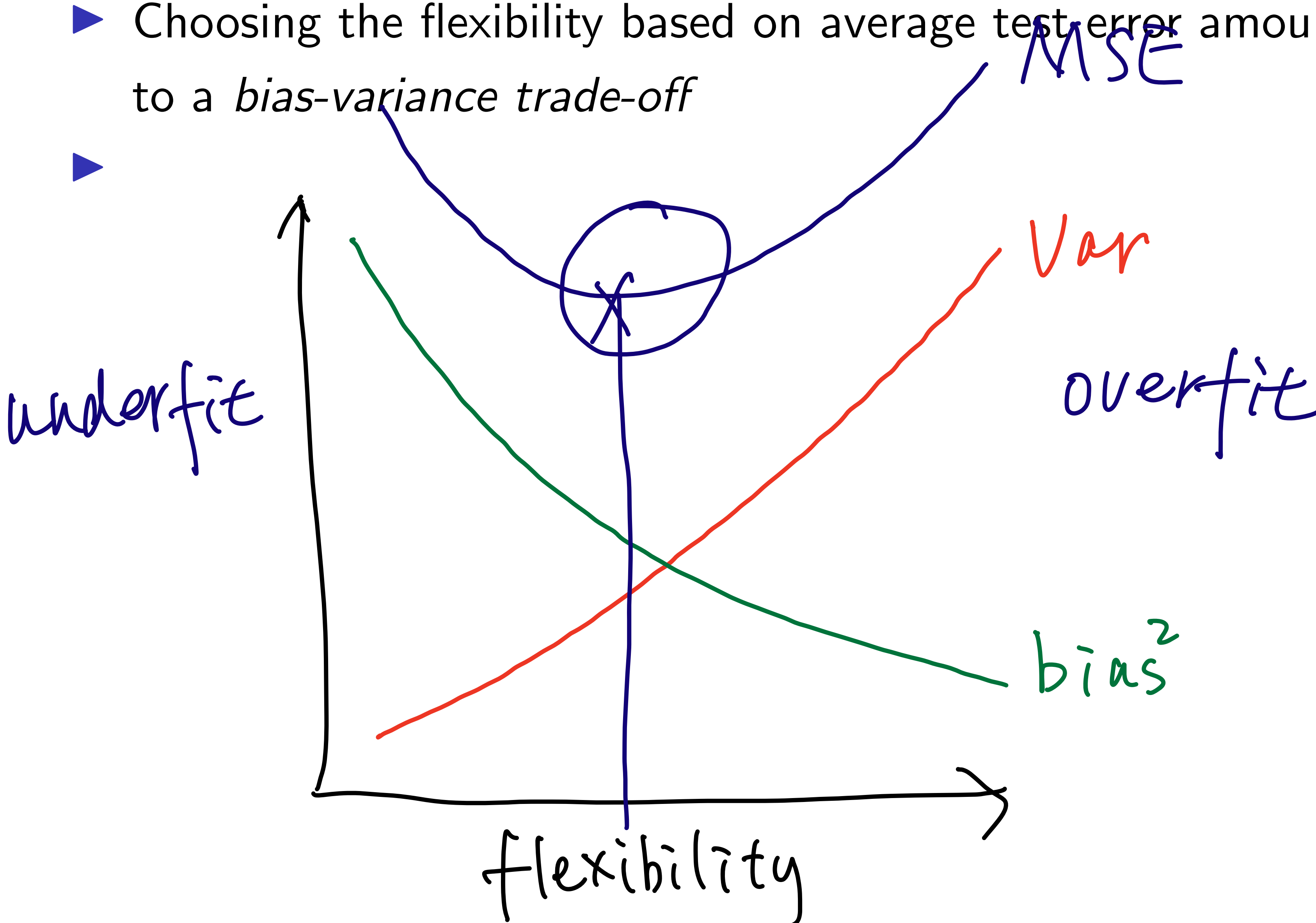
$$(a+b+c)^2 = a^2 + b^2 + c^2 + \boxed{2ab + 2bc + 2ac}$$

Bias-Variance trade-off

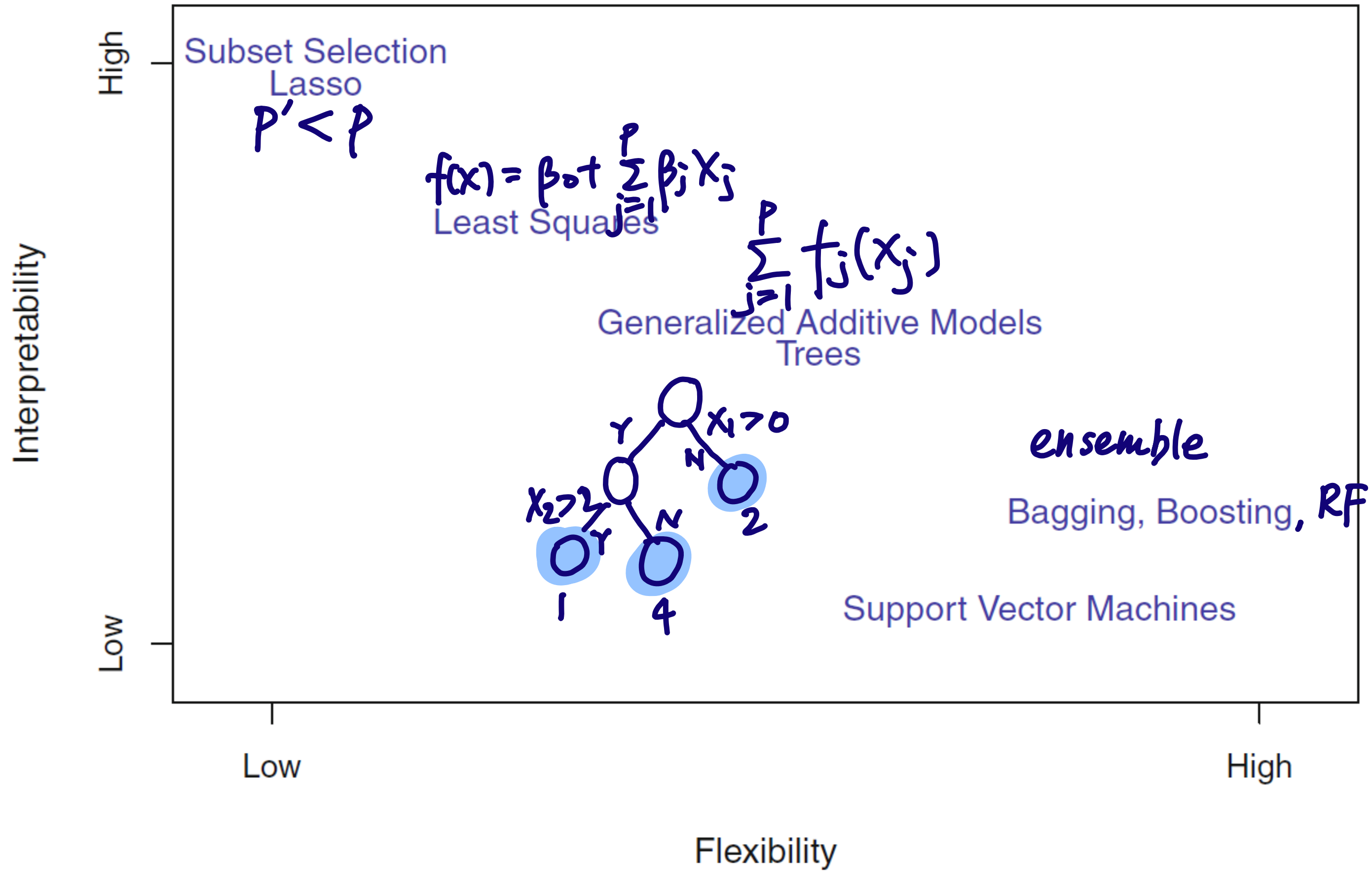
- **Variance** refers to the amount by which $\widehat{f}$ would change if we estimated it using a different training data set
- **Bias** refers to the error that is introduced by approximating a real-life problem by a much simpler model

Bias-Variance trade-off

- As the flexibility of $\widehat{f}$ increases, its variance increases, and its bias decreases
- Choosing the flexibility based on average test error amounts to a *bias-variance trade-off*
-



Trade-off between flexibility and interpretability



[ISL] Figure 2.7

Some trade-offs

- Flexibility versus interpretability
  - – Linear models are easy to interpret
  - – High order polynomials?
- Good fit versus over-fit or under-fit
  - – How do we know when the fit is just right?

CV

- Parsimony versus black-box
  - – We often prefer a simpler model involving fewer variables
  - – When the goal is prediction, which do you prefer?

There is no free lunch in statistics!

Modeling process

tuning para.