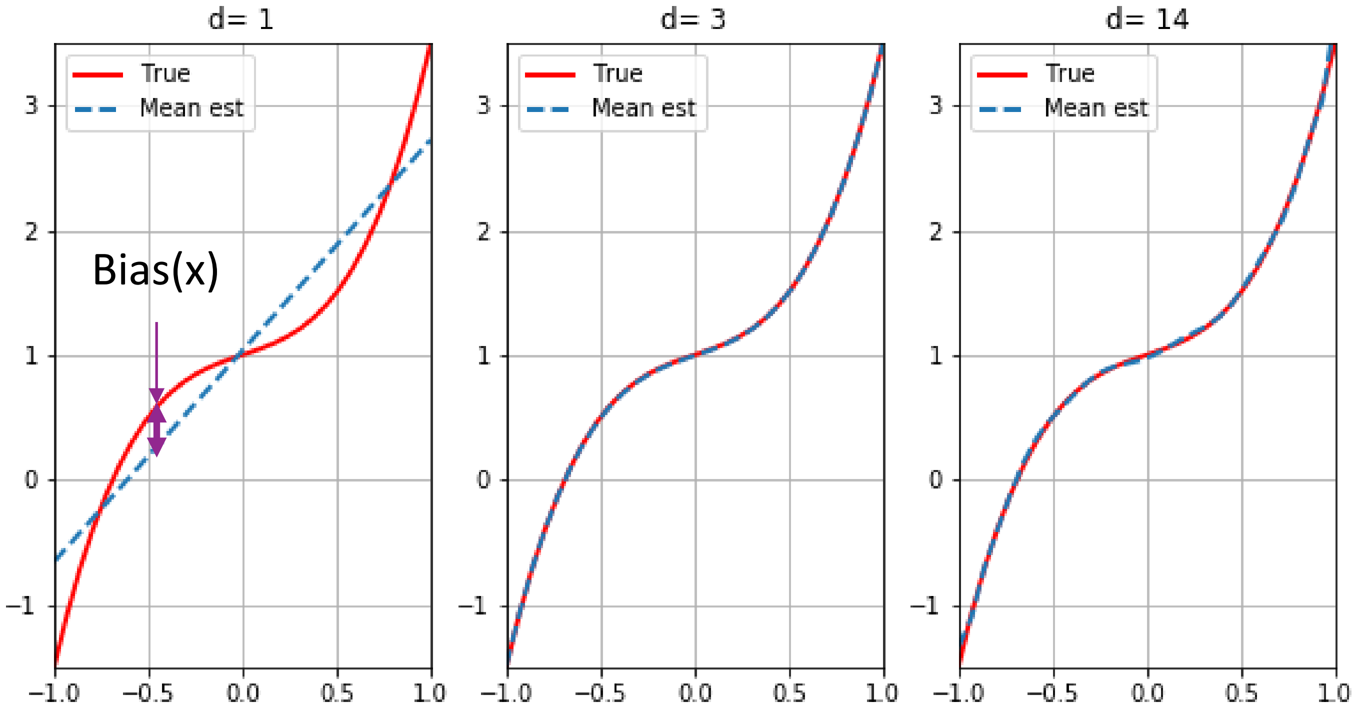
Bias and Variance

- To understand potential problem of using a large model class introduce two key quantities:
- Bias: $Bias(\boldsymbol{x}_{test}) := f_0(\boldsymbol{x}_{test}) - E\left[f\left(\boldsymbol{x}_{test}, \widehat{\boldsymbol{\beta}}\right)\right]$
  - How much the average value of the estimate differs from the true function
- Variance: $Var(\boldsymbol{x}_{test}) \coloneqq E\left[f\left(\boldsymbol{x}_{test}, \widehat{\boldsymbol{\beta}}\right) - E\left[f\left(\boldsymbol{x}_{test}, \widehat{\boldsymbol{\beta}}\right)\right]\right]^2$
  - How much the estimate varies around its average
- Bias and variance are (conceptually) measured as follows:
  - Get many independent training data sets, each with same size N and input values x_i
  - Each dataset has different output values y_i because of independent noise in the training data
  - Obtain $\widehat{\boldsymbol{\beta}}$ for each training data set
  - Bias and variances are computed over the different sets
- Of course, in reality, we have only one training dataset
  - Used to study theoretical averages over different experiments

Bias Illustrated

- Red: True function
- Repeat 100 trials
  - Each trial has independent data
  - Obtain estimate for each trial
- Dashed line:
  Mean estimate among all trials
- Bias=True – Mean estimate
- Conclusions:
  - Low model orders ⇒ bias high
  - High model orders ⇒ bias low



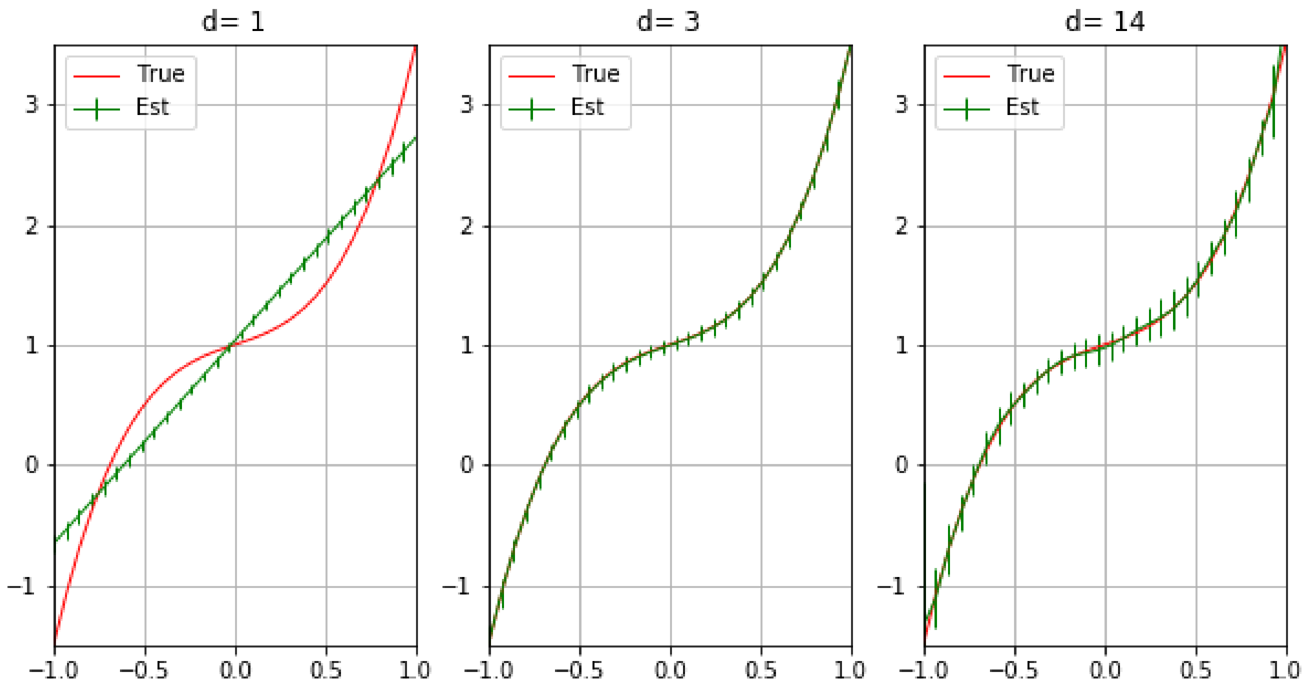
High bias $d < d_{true}$

No bias $d = d_{true}$

No bias $d > d_{true}$

Variance Illustrated

- ❑Red: True function
- ❑Repeat 100 trials
  - Each trial has independent data
  - Obtain estimate for each trial
- ❑Variance=STD around mean
- ❑Conclusions:
  - Low model orders ⇒ low variance
  - High model orders ⇒ high variance



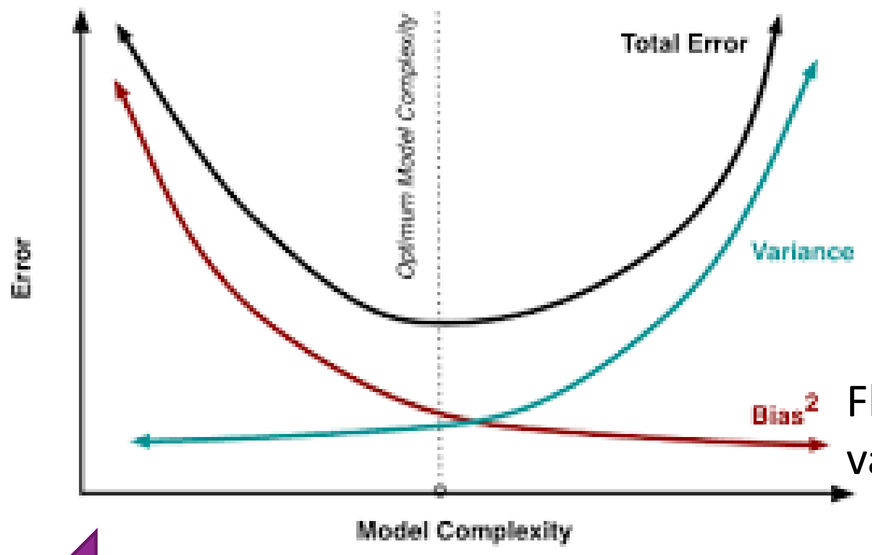
High bias | No bias | No bias

Low variance | Med variance | High variance

Bias-Variance Formula

- Recall definitions:
  - Function MSE: $\mathrm{MSE}_f(\boldsymbol{x}_{test}) \coloneqq E\left[f_0(\boldsymbol{x}_{test}) - f\left(\boldsymbol{x}_{test}, \widehat{\boldsymbol{\beta}}\right)\right]$:
  - Bias: $Bias(\boldsymbol{x}_{test}) := f_0(\boldsymbol{x}_{test}) - E\left[f\left(\boldsymbol{x}_{test}, \widehat{\boldsymbol{\beta}}\right)\right]$
  - Variance: $Var(\boldsymbol{x}_{test}) \coloneqq E\left[f\left(\boldsymbol{x}_{test}, \widehat{\boldsymbol{\beta}}\right) - E\left[f\left(\boldsymbol{x}_{test}, \widehat{\boldsymbol{\beta}}\right)\right]\right]^2$
- Bias-Variance formula : $\boxed{MSE_f(\boldsymbol{x}_{test}) = Bias(\boldsymbol{x}_{test})^2 + Var(\boldsymbol{x}_{test})}$
  - Will be proved below
- Bias-Variance tradeoff
- Bias due to under-modeling
  - Reduced with high model order
- Variance is due to noise in training data and number of parameters to estimate
  - Increases with higher model order

Bias-Variance Tradeoff



Floor is the noise variance

Simpler models
Less parameters
Under-fitting

Richer models
More parameters
Over-fitting

- Bias:
  - Due to under-modeling
  - Reduced with high model order
- Variance:
  - Increases with noise in training data
  - Increase with high model order
- Optimal model order depends on:
  - Amount of samples available
  - Underlying complexity of the relation

Bias-Variance Formula Proof

- Define $\bar{f}(\boldsymbol{x}_{test}) = E[f(\boldsymbol{x}_{test}, \widehat{\boldsymbol{\beta}})]$ = average value of estimated function
- $\mathrm{MSE}_f(\boldsymbol{x}_{test}) = E[f_0(\boldsymbol{x}_{test}) - f(\boldsymbol{x}_{test}, \widehat{\boldsymbol{\beta}})]^2 = E\left[f_0(\boldsymbol{x}_{test}) - \bar{f}(\boldsymbol{x}_{test}) + \bar{f}(\boldsymbol{x}_{test}) - f(\boldsymbol{x}_{test}, \widehat{\boldsymbol{\beta}})\right]^2$
- Three components: $\mathrm{MSE}_f(\boldsymbol{x}_{test}) = M_1 + M_2 - 2M_3$
  - $M_1 = E\left[f_0(\boldsymbol{x}_{test}) - \bar{f}(\boldsymbol{x}_{test})\right]^2 = \left[f_0(\boldsymbol{x}_{test}) - \bar{f}(\boldsymbol{x}_{test})\right]^2 = Bias(x_{test})$
  - $M_2 = E\left[f(\boldsymbol{x}_{test}, \widehat{\boldsymbol{\beta}}) - \bar{f}(\boldsymbol{x}_{test})\right]^2 = Var(x_{test})$
  - $$\begin{aligned} M_3 &= E\left[(f_0(\boldsymbol{x}_{test}) - \bar{f}(\boldsymbol{x}_{test}))(f(\boldsymbol{x}_{test}, \widehat{\boldsymbol{\beta}}) - \bar{f}(\boldsymbol{x}_{test}))\right] \\ &= (f_0(\boldsymbol{x}_{test}) - \bar{f}(\boldsymbol{x}_{test}))E\left[f(\boldsymbol{x}_{test}, \widehat{\boldsymbol{\beta}}) - \bar{f}(\boldsymbol{x}_{test})\right] \\ &= (f_0(\boldsymbol{x}_{test}) - \bar{f}(\boldsymbol{x}_{test}))(\bar{f}(\boldsymbol{x}_{test}) - \bar{f}(\boldsymbol{x}_{test})) = 0 \end{aligned}$$

Summary of Results for Linear Models

- Suppose model is linear with $N =$ num samples, p = num parameters
- Can show the following results (need some math, see next section)
- Result 1: When $N < p$, linear estimate is not unique
  - Need at least as many samples as parameters
- Now assume that $N \geq p$ and parameter estimate is unique
- Result 2: When there is no under-modeling, estimate is unbiased

$$E\left[f\left(\boldsymbol{x}_{test}, \widehat{\boldsymbol{\beta}}\right)\right] = f_0\left(\boldsymbol{x}_{test},\right).$$

- Result 3: If test point drawn randomly from the training data:

$$Var = \frac{p}{N}\sigma_\epsilon^2$$

  - Variance increases linearly with number of parameters and inversely with number of samples