Statistical Learning

$$P_{XY}(X = x, Y = y)$$

Goal: Predict Y given X

Find a function η that minimizes

$$\mathbb{E}_{XY}[(Y - \eta(X))^2]$$

Thus far, we've been using η which is a:
- **Linear functions of X**
- **Degree p polynomials of X**
- **Linear "generalization" of X in p dimensions**

Statistical Learning

$$P_{XY}(X = x, Y = y)$$

Goal: Predict Y given X

Find a function η that minimizes

$$\mathbb{E}_{XY}[(Y - \eta(X))^2] = \mathbb{E}_X\left[\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x]\right]$$

$$\eta(x) = \arg\min_c \mathbb{E}_{Y|X}[(Y - c)^2 | X = x] = \mathbb{E}_{Y|X}[Y | X = x]$$

Under LS loss, optimal predictor: $\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$

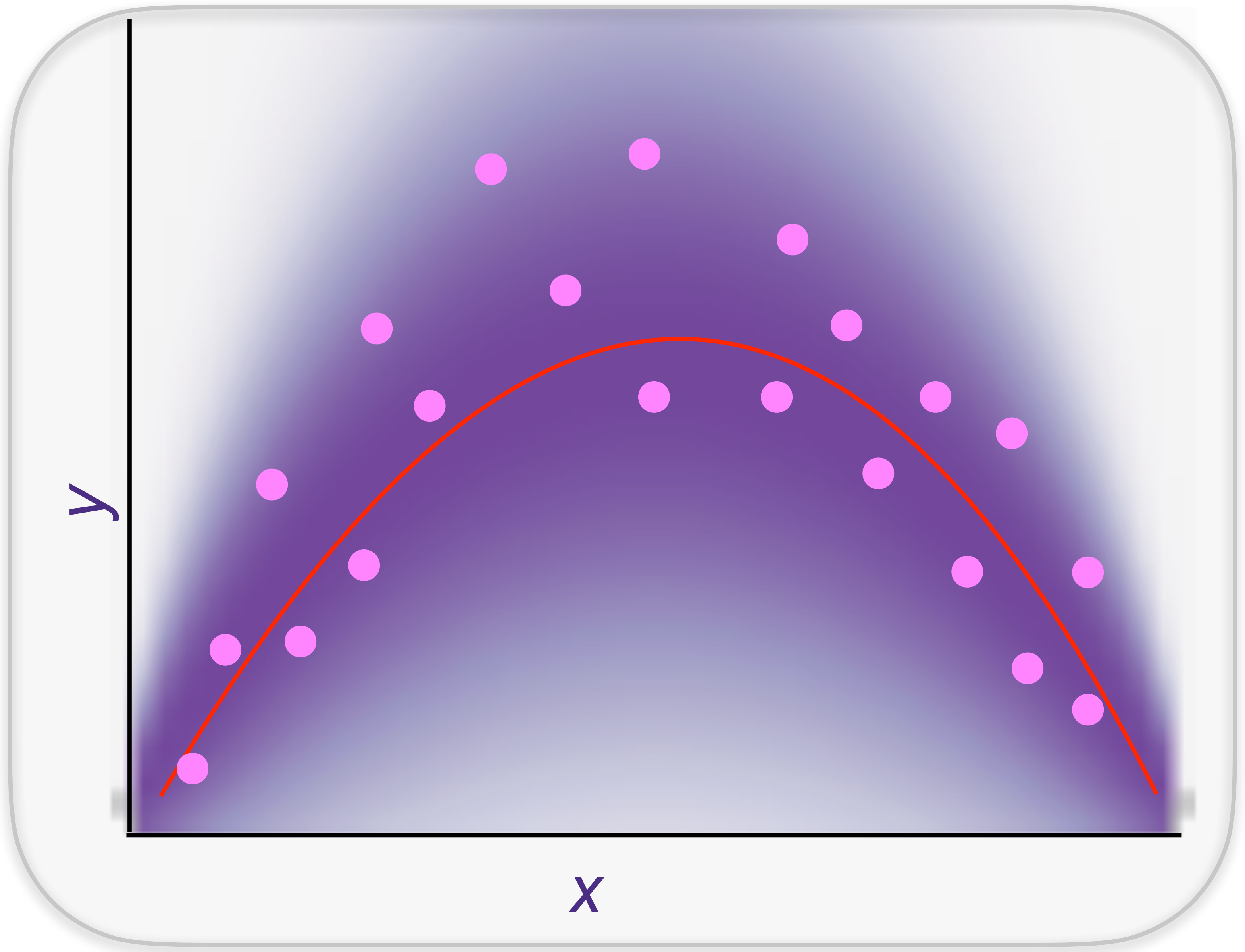
Optimal Prediction

$$\mathbb{E}_{XY}[(Y - \eta(X))^2] = \mathbb{E}_X \left[\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x] \right]$$

Under LS loss, optimal predictor: $\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$

Statistical Learning

$$P_{XY}(X = x, Y = y)$$



Ideally, we want to find:

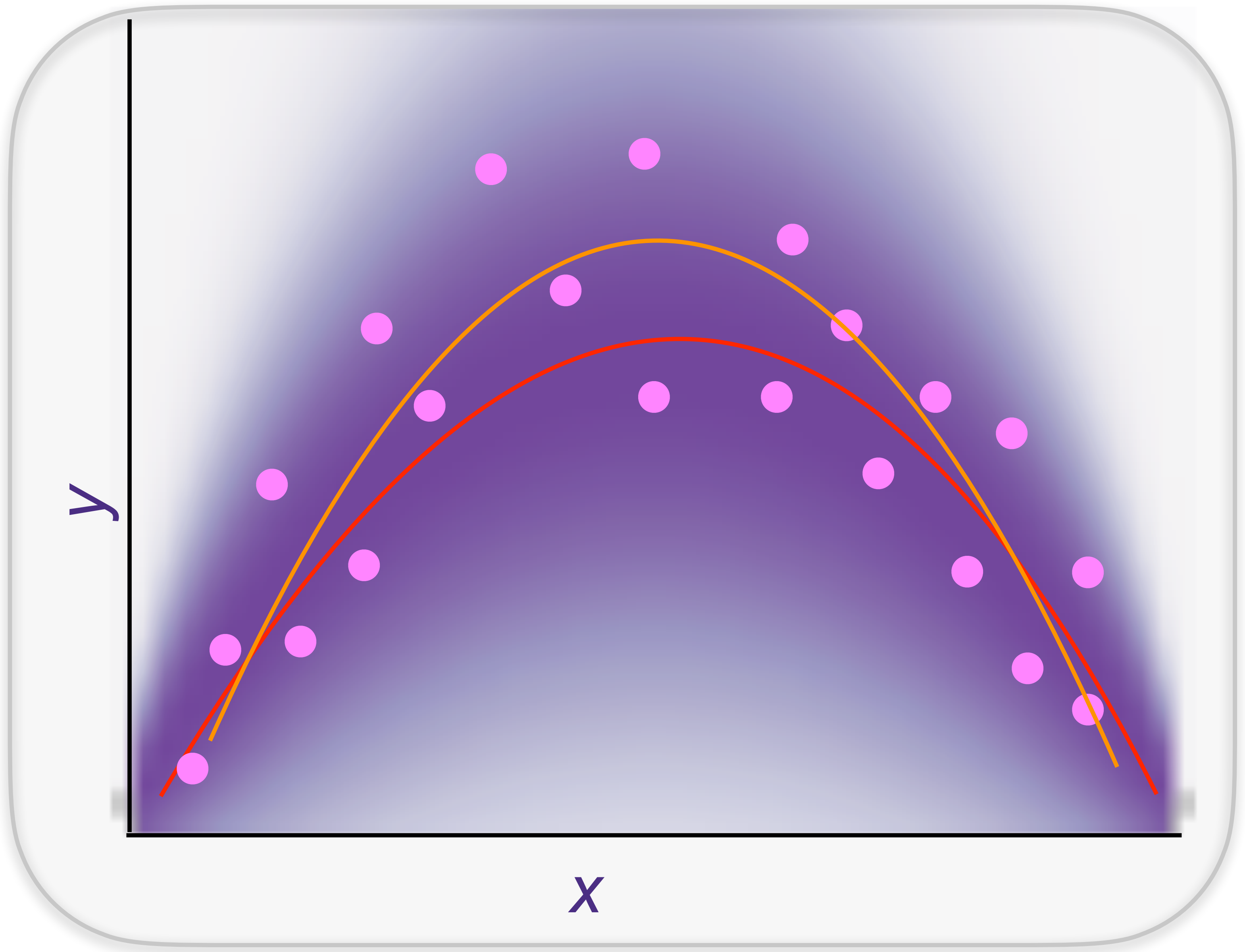
$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

$$(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY} \quad \text{for } i = 1, \dots, n$$

Statistical Learning

$$P_{XY}(X = x, Y = y)$$



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

$$(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY} \quad \text{for } i = 1, \dots, n$$

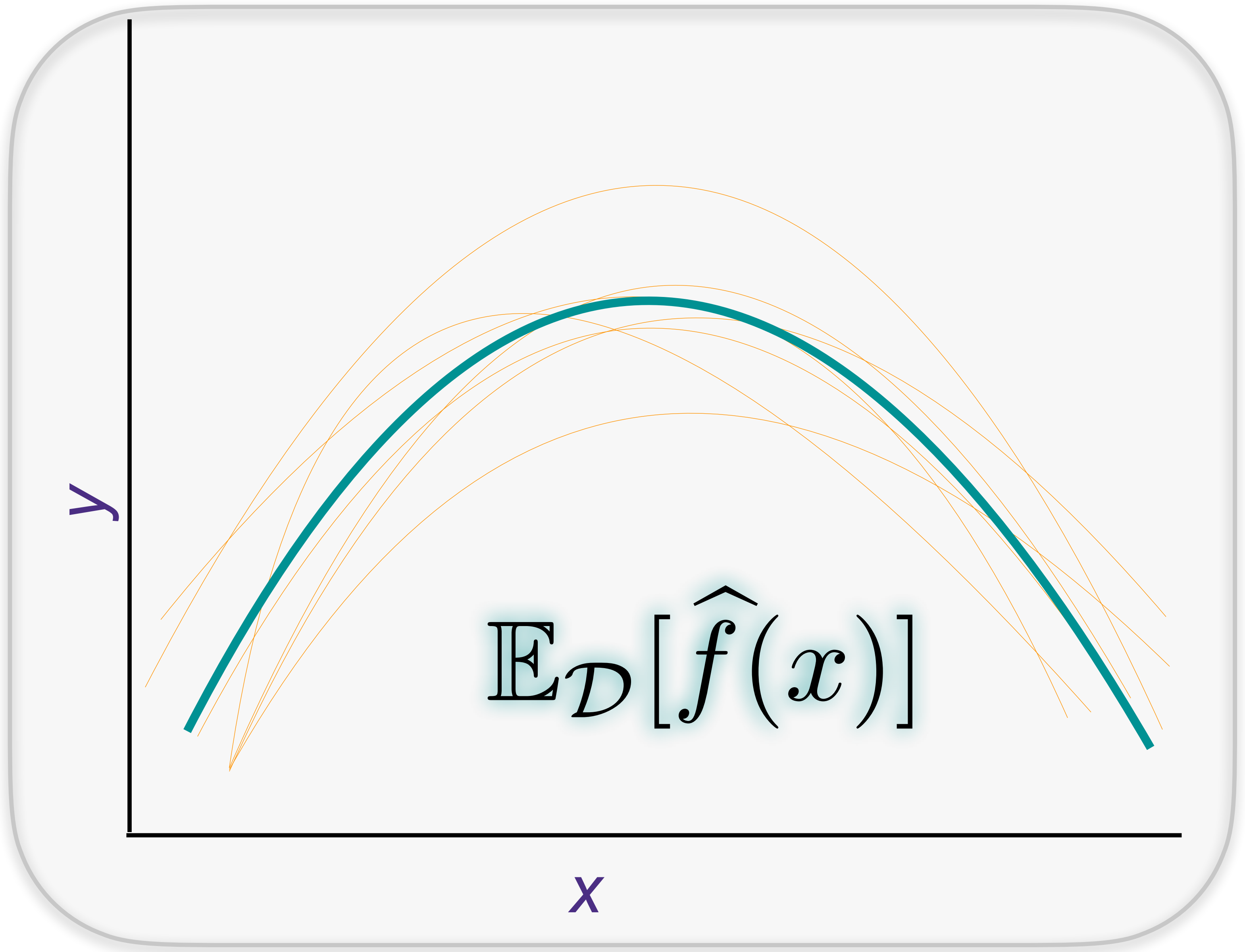
and are restricted to a function class (e.g., linear) so we compute:

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

We care about future predictions: $\mathbb{E}_{XY}[(Y - \widehat{f}(X))^2]$

Statistical Learning

$P_{XY}(X = x, Y = y)$



Ideally, we want to find:

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

$$(x_i, y_i) \overset{i.i.d.}{\sim} P_{XY} \quad \text{for } i = 1, \ldots, n$$

and are restricted to a function class (e.g., linear) so we compute:

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Each draw $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ results in different $\widehat{f}$

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Bias-Variance Tradeoff

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

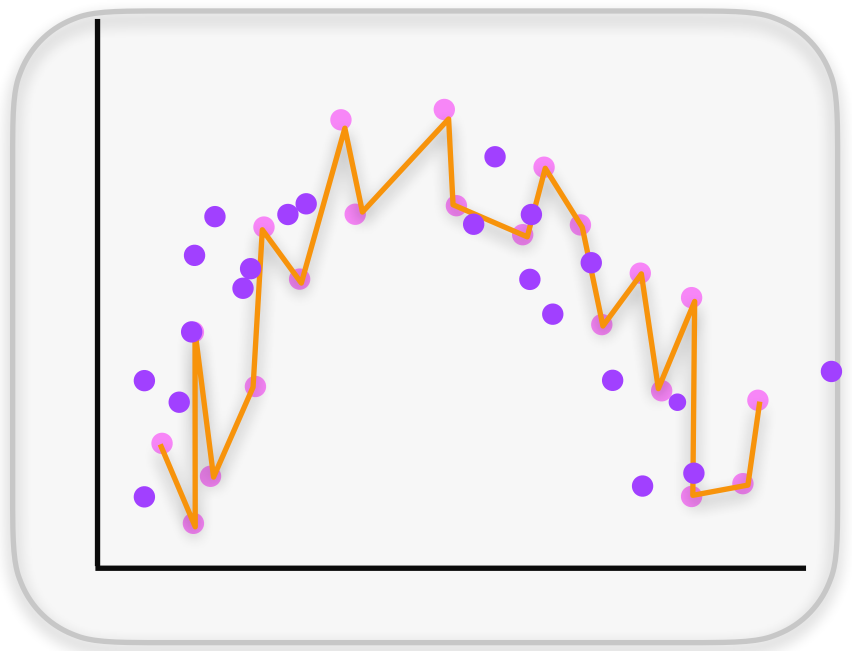
Bias-Variance Tradeoff

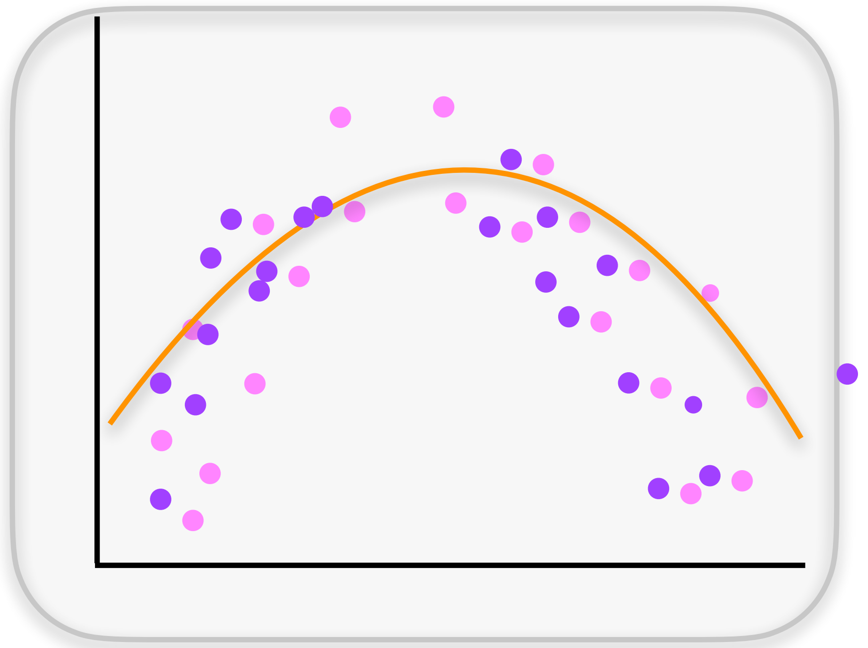
$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \widehat{f}_{\mathcal{D}}(x))^2]\big|X = x] = \underbrace{\mathbb{E}_{Y|X}[(Y - \eta(x))^2\big|X = x]}_{\text{irreducible error}}$$

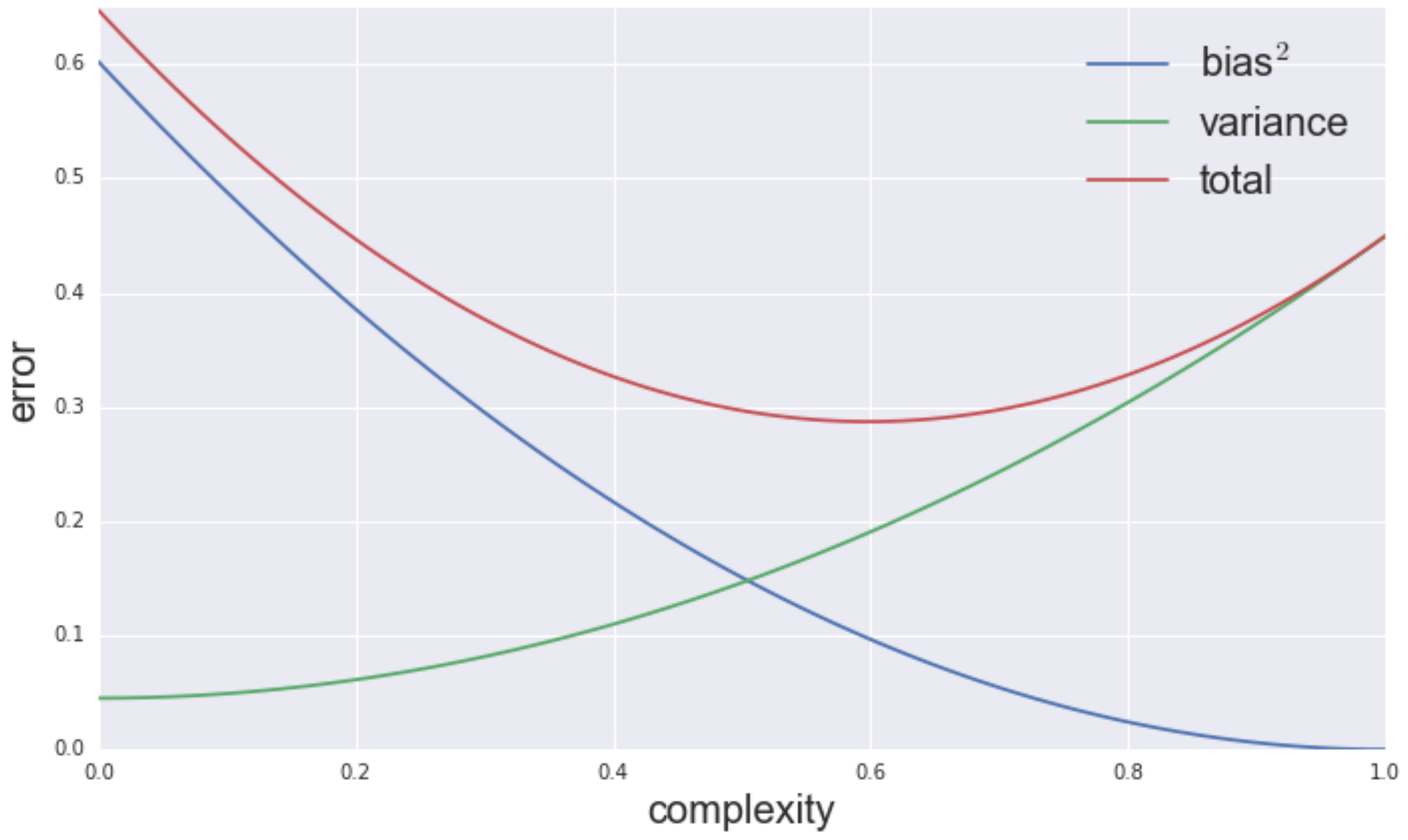
irreducible error

$$+\underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])^2}_{\text{bias squared}} + \underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]}_{\text{variance}}$$

bias squared **variance**







If we re-drew our data, what the LS training error estimator look like for generalized linear functions in small p/large p dimensions?