Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

$$\text{if} \quad y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

Example: Linear LS: compute bias

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

$$\text{if} \quad y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\widehat{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = w + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

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

$$\widehat{f}_{\mathcal{D}}(x) = \widehat{w}^T x = w^T x + \epsilon^T \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} x$$

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

Example: Linear LS: compute variance

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

$$\text{if} \quad y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\widehat{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = w + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

$$\widehat{f}_{\mathcal{D}}(x) = \widehat{w}^T x = w^T x + \epsilon^T \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} x$$

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

Example: Linear LS

$$\mathbf{Y} = \mathbf{X}w + \epsilon$$

$$\text{if} \quad y_i = x_i^T w + \epsilon_i \quad \text{and} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\widehat{w}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = w + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

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

$$\widehat{f}_{\mathcal{D}}(x) = \widehat{w}^T x = w^T x + \epsilon^T \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x$$

$$\underbrace{\mathbb{E}_{XY}[(Y - \eta(x))^2 | X = x]}_{\text{irreducible error}} = \sigma^2$$

$$\underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])^2}_{\text{bias squared}} = 0$$

$$\mathbb{E}_{X=x}\left[\underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]}_{\text{variance}}\right] = \frac{p\sigma^2}{n}$$

Overfitting

Bias-Variance Tradeoff

- Choice of hypothesis class introduces learning bias
  - More complex class → less bias
  - More complex class → more variance
- But in practice??

Bias-Variance Tradeoff

> Choice of hypothesis class introduces learning bias
  - More complex class → less bias
  - More complex class → more variance

> But in practice??

> Before we saw how increasing the feature space can increase the complexity of the learned estimator:

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \ldots$$

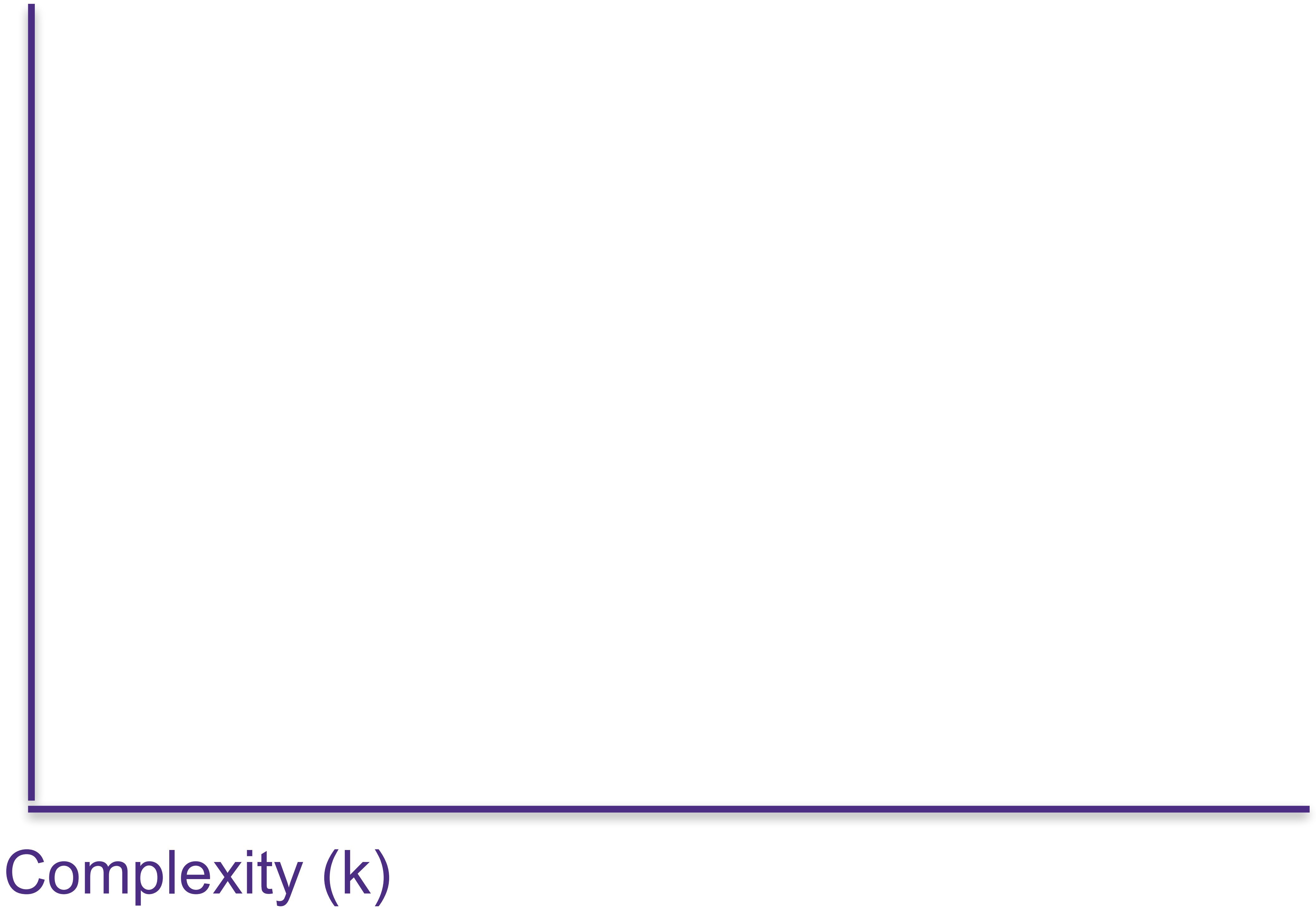
$$\widehat{f}_{\mathcal{D}}^{(k)} = \arg\min_{f \in \mathcal{F}_k} \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - f(x_i))^2$$

Complexity grows as k grows

Training set error as a function of model complexity

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \ldots$$

$$\widehat{f}_{\mathcal{D}}^{(k)} = \arg\min_{f \in \mathcal{F}_k} \frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - f(x_i))^2$$



TRAIN error:

$$\mathcal{D} \overset{i.i.d.}{\sim} P_{XY}$$

$$\frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$$

TRUE error:

$$\mathbb{E}_{XY}[(Y - \widehat{f}_{\mathcal{D}}^{(k)}(X))^2]$$

Training set error as a function of model complexity

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \dots$$

$$\widehat{f}_{\mathcal{D}}^{(k)} = \arg\min_{f \in \mathcal{F}_k} \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - f(x_i))^2$$

TRAIN error:

$$\mathcal{D} \overset{i.i.d.}{\sim} P_{XY}$$

$$\frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$$

TRUE error:

$$\mathbb{E}_{XY}[(Y - \widehat{f}_{\mathcal{D}}^{(k)}(X))^2]$$

TEST error:

$$\mathcal{T} \overset{i.i.d.}{\sim} P_{XY}$$

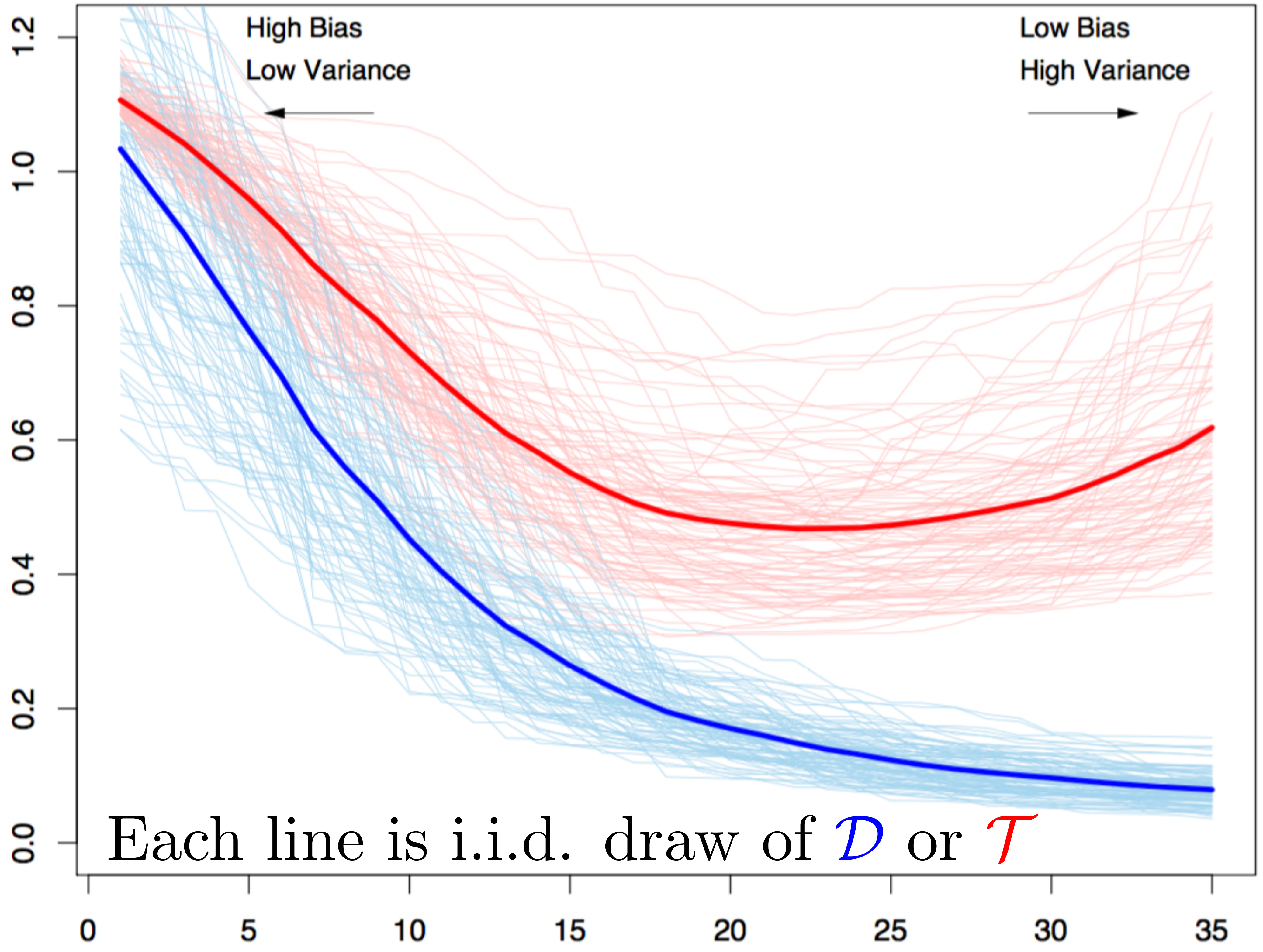
$$\frac{1}{|\mathcal{T}|} \sum_{(x_i, y_i) \in \mathcal{T}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$$

$$\boxed{\text{Important: } \mathcal{D} \cap \mathcal{T} = \emptyset}$$

Training set error as a function of model complexity

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \dots$$

$$\widehat{f}_{\mathcal{D}}^{(k)} = \arg\min_{f \in \mathcal{F}_k} \frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - f(x_i))^2$$



Complexity (k)

Plot from Hastie et al

TRAIN error:

$$\mathcal{D} \overset{i.i.d.}{\sim} P_{XY}$$

$$\frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$$

TRUE error:

$$\mathbb{E}_{XY}[(Y - \widehat{f}_{\mathcal{D}}^{(k)}(X))^2]$$

TEST error:

$$\mathcal{T} \overset{i.i.d.}{\sim} P_{XY}$$

$$\frac{1}{|\mathcal{T}|} \sum_{(x_i,y_i)\in\mathcal{T}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$$

$$\boxed{\text{Important: } \mathcal{D} \cap \mathcal{T} = \emptyset}$$

Training set error as a function of model complexity

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \dots$$

$$\widehat{f}_{\mathcal{D}}^{(k)} = \arg\min_{f \in \mathcal{F}_k} \frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - f(x_i))^2$$

TRAIN error is **optimistically biased** because it is evaluated on the data it trained on. **TEST error** is **unbiased** only if *T* is never used to train the model or even pick the complexity k.

TRAIN error:

$$\mathcal{D} \overset{i.i.d.}{\sim} P_{XY}$$

$$\frac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$$

TRUE error:

$$\mathbb{E}_{XY}[(Y - \widehat{f}_{\mathcal{D}}^{(k)}(X))^2]$$

TEST error:

$$\mathcal{T} \overset{i.i.d.}{\sim} P_{XY}$$

$$\frac{1}{|\mathcal{T}|} \sum_{(x_i,y_i)\in\mathcal{T}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$$

$$\boxed{\text{Important: } \mathcal{D} \cap \mathcal{T} = \emptyset}$$

How many points do I use for training/testing?

- Very hard question to answer!
  - Too few training points, learned model is bad
  - Too few test points, you never know if you reached a good solution
- More on this later the quarter, but still hard to answer
- Typically:
  - If you have a reasonable amount of data 90/10 splits are common
  - If you have little data, then you need to get fancy (e.g., bootstrapping)