complexity ↑ ⇒ bias ↓ variance ↑
complexity ↓ ⇒ bias ↑ variance ↓

Bias-Variance Tradeoff

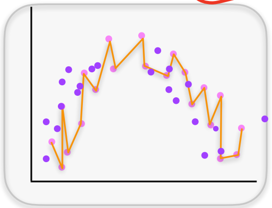
$$\mathbb{E}_{Y|X}[\mathbb{E}_{\mathcal{D}}[(Y - \widehat{f}_{\mathcal{D}}(x))^2]|X = x] = \underbrace{\mathbb{E}_{Y|X}[(Y - \eta(x))^2|X = x]}_{\text{irreducible error}}$$

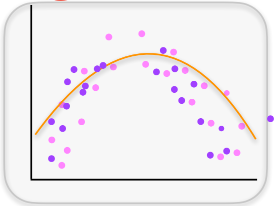
irreducible error

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

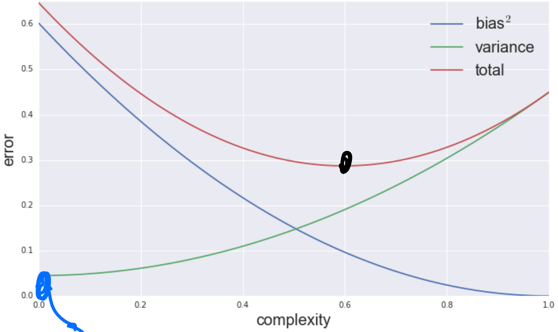
bias squared **variance**

Q: what is variance of a constant predictor $f_n(x) = C$, C is independent of data





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



Example: Linear LS

$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}_{n \times d}$ $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}_n$

$w \in \mathbb{R}^d$

$$\mathbf{Y} = \mathbf{X}w + \epsilon \qquad \varepsilon \in \mathbb{R}^n$$

Assumption $x \in \mathbb{R}^d$

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

$$\eta(x) = \mathbb{E}_{Y|X}[Y \mid X = x] = \mathbb{E}_{Y|X}[x^T w + \varepsilon \mid X = x] = w^T x$$

MLE $\hat{w} = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (Xw + \varepsilon) = w + (X^T X)^{-1} X^T \varepsilon$

new point x_{new} $\hat{f}_D(x_{new}) = x_{new}^T \hat{w} = x_{new}^T w + x_{new}^T (X^T X)^{-1} X^T \varepsilon$

irreducible error

$$\mathbb{E}_{Y|X}[(Y - \eta(x))^2 \mid X = x] = \mathbb{E}_{Y|X}[(w^T x + \varepsilon - w^T x)^2 \mid X = x]$$

$$= \mathbb{E}[\varepsilon^2] = \sigma^2$$

Example: Linear LS: compute bias

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

$D = \{(x_i, y_i)\}_{i=1}^{n}$

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

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

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

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

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

bias squared

x_{new} is fixed

$$\mathbb{E}_D[f_D(x_{new})] = \mathbb{E}_D[x_{new}^T w + x_{new}^T (X^T X)^{-1} X \epsilon]$$

$$= x_{new}^T w + \mathbb{E}_X[x_{new}^T (X^T X)^{-1} X \underbrace{\mathbb{E}_Y[\epsilon]}_{=0}]$$

$$= x_{new}^T w = y(x_{new})$$

$\Rightarrow$ unbiased

Example: Linear LS: compute variance

$tr(I_d) = tr\begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix} = d.$

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

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

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

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

$\hat{f}_D(x) = x^T \hat{w}_{MLE}$

$\mathbb{E}_D[\hat{f}_D(x)] = x^T w$

$x = x_{new}$

$$\underbrace{\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)] - \widehat{f}_{\mathcal{D}}(x))^2]}_{\text{variance}} = \mathbb{E}_D\left[x^T(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}x\right]$$

$$= \mathbb{E}_X\left[x^T(X^TX)^{-1}X^T\mathbb{E}_Y[\epsilon\epsilon^T]X(X^TX)^{-1}x\right]$$

$$= \sigma^2\mathbb{E}_X\left[x^T(X^TX)^{-1}X^TX(X^TX)^{-1}x\right]$$

$$= \sigma^2\mathbb{E}_X\left[x^T(X^TX)^{-1}x\right]$$

$\mathbb{E}[\epsilon\epsilon^T] = \sigma^2 \cdot I$

$\mathbb{E}[\epsilon\epsilon^T]_{ij} = \mathbb{E}[\epsilon_i \cdot \epsilon_j]$

$X^TX = \sum_{i=1}^n x_i x_i^T \xrightarrow[n\to\infty]{} n\cdot\Sigma$

central limit theorem

$\Sigma: \mathbb{E}_x[xx^T]$

$$\mathbb{E}_x[\mathbb{E}_D[(\mathbb{E}_D[\hat{f}_D(x)] - \hat{f}_D(x))^2]]$$

$$\xrightarrow{n\to\infty} \mathbb{E}_x[\sigma^2 \quad [x^T(n\Sigma)^{-1}x]]$$

by Trace formula

$= \mathbb{E}_x[\sigma^2 tr((n\Sigma)^{-1}xx^T)]$

$= \sigma^2 tr((n\Sigma)^{-1}\Sigma)$

$= \frac{\sigma^2}{n} tr(I)$

$= \frac{\sigma^2 d}{n}$

Example: Linear LS

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

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

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

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

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

$$\underbrace{\mathbb{E}_{XY}[(Y - \eta(x))^2 | X = x]}_{\text{irreducible error}} = \sigma^2 \qquad \underbrace{(\eta(x) - \mathbb{E}_{\mathcal{D}}[\widehat{f}_{\mathcal{D}}(x)])^2}_{\text{bias squared}} = 0$$

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

as $d \uparrow \;\rightarrow\; \uparrow$ variance

Overfitting

Bias-Variance Tradeoff

> Choice of hypothesis class introduces learning bias
  - More complex class → less bias $\exists f \in \mathcal{F}$ $g \approx f$
  - 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:

linear quadratic cubic . . .

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

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

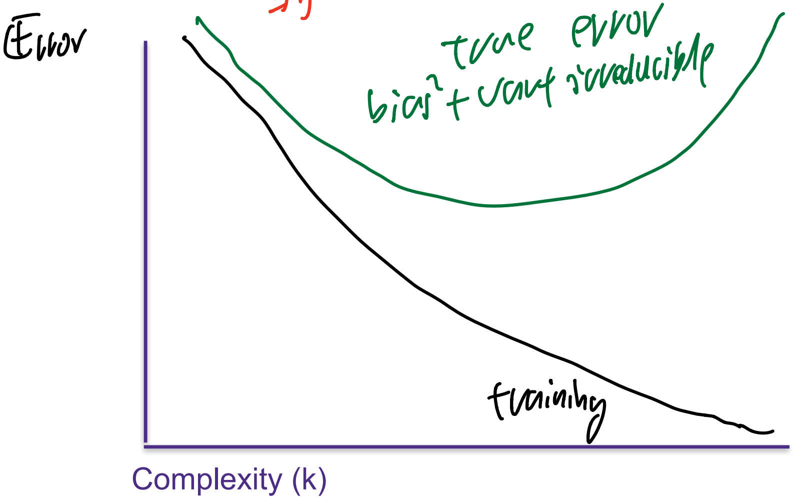
Complexity grows as k grows

Training set error as a function of model complexity

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

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

$\exists f$



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

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

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

all data $\mathcal{D} \cup \mathcal{T} \sim P_{XY}$

test set

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:

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

$O\left(\frac{1}{\sqrt{|\mathcal{T}|}}\right)$

TEST error:

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

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

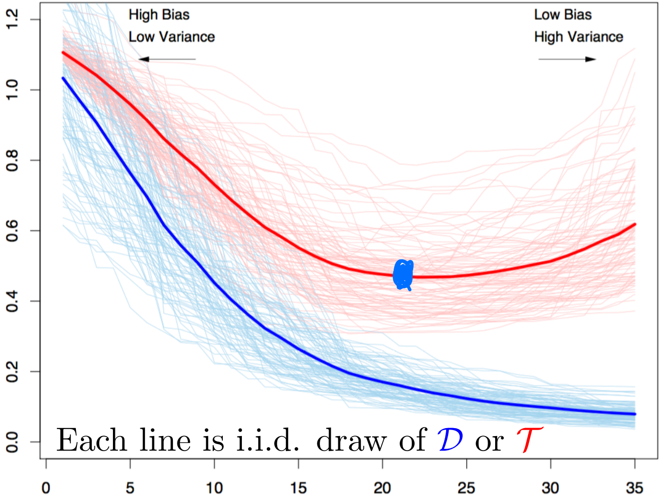
$\boxed{\text{Important: } \mathcal{D} \cap \mathcal{T} = \emptyset}$

Training set error as a function of model complexity

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

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

error



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:

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

plug in

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

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

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:

$\mathbb{E}_{\mathcal{T}}$[test error] = True error

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

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

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

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)