Bias-Variance Tradeoff

Process

n: # of data
x_i: feature $\in \mathbb{R}^d$
y_i: label

Collect a **data set** $\{(x_i, y_i)\}_{i=1}^{n}$

Decide on a **model** function f: $f(x) \approx y$, linear $f(x) = w^T x$

Find the function which fits the data best
- **Choose a loss function** quadratic $(f(x) - y)^2$ ← minimizes
- **Pick the function which minimizes loss on data** find f

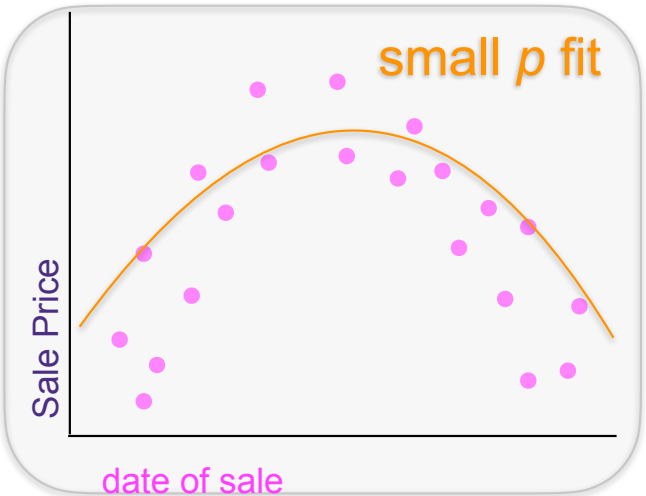
Use function to make prediction on new examples

x_{new}, $f(x_{new}) \approx y_{new}$

The regression problem

Training Data: $\{(x_i, y_i)\}_{i=1}^n$ $x_i \in \mathbb{R}^d$ $y_i \in \mathbb{R}$

$p = 2$



$h_1(x) = x$

$h_2(x) = x^2$

$\vdots$

$h_p(x) = x^p$

Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear in h

$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

Loss: least squares

$$\min_w \sum_{i=1}^n \left(y_i - h(x_i)^T w\right)^2$$

The regression problem

Training Data: $\{(x_i, y_i)\}_{i=1}^n$ $x_i \in \mathbb{R}^d$ $y_i \in \mathbb{R}$

$p = 20$

fit:

training error $= 0$

$h(x_i)^T w = y_i$

Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear in h

$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

Loss: least squares

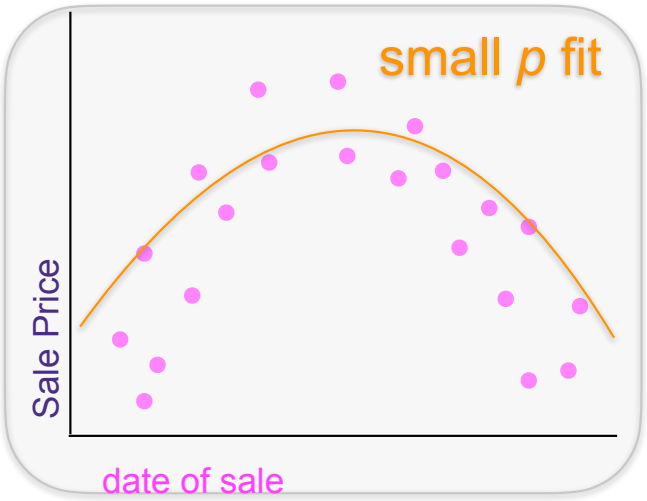
$$\min_w \sum_{i=1}^n \left(y_i - h(x_i)^T w\right)^2$$

Which is better?

A: large p

non-smooth

B: small p

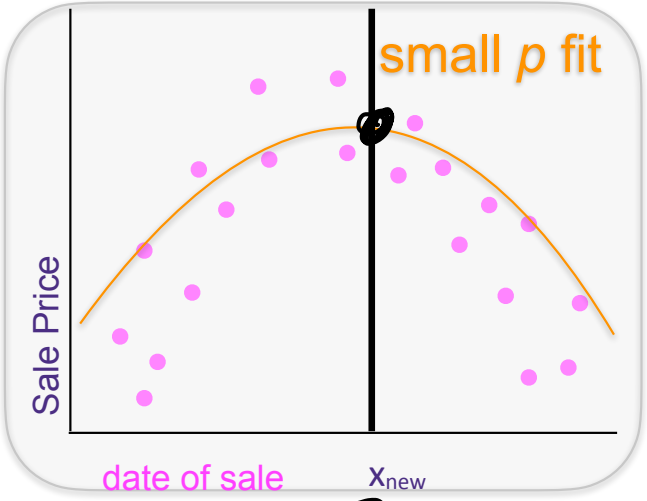


smooth

Predicting sale price for a new house: A vs B

A: large p

B: small p



single

Our goal is to predict prices for new houses

Average Accuracy

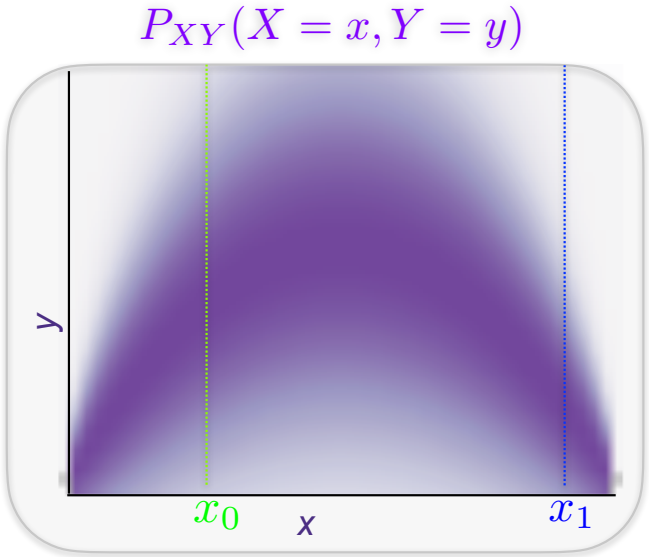
Joint distribution

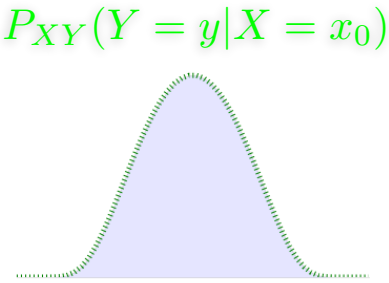
$P_{XY}(X = x, Y = y)$

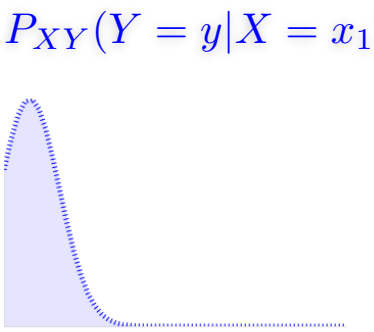
On *average* over a house
drawn from this distribution,
we want to make a good prediction.

Goal: predict future sale prices

Two sources randomness
1) X is random
2) given X
y is random







Statistical Learning

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

unknown

η : learned function

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

performance measure

Goal: Predict Y given X

Find a function η that minimizes

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

average P_{XY}

metric

- quadratic
- l_1 - loss
- logistic
- 0/1

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

Suppose Know

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

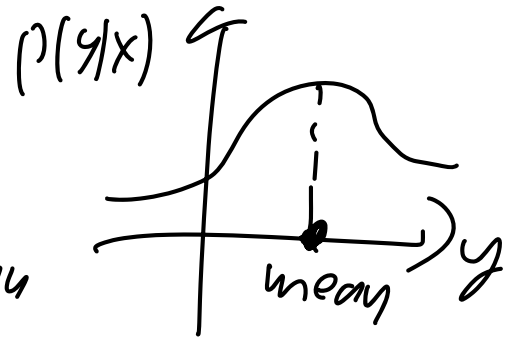
Goal: Predict Y given X

Find a function η that minimizes

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

conditional mean

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



Theorem

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

Optimal Prediction

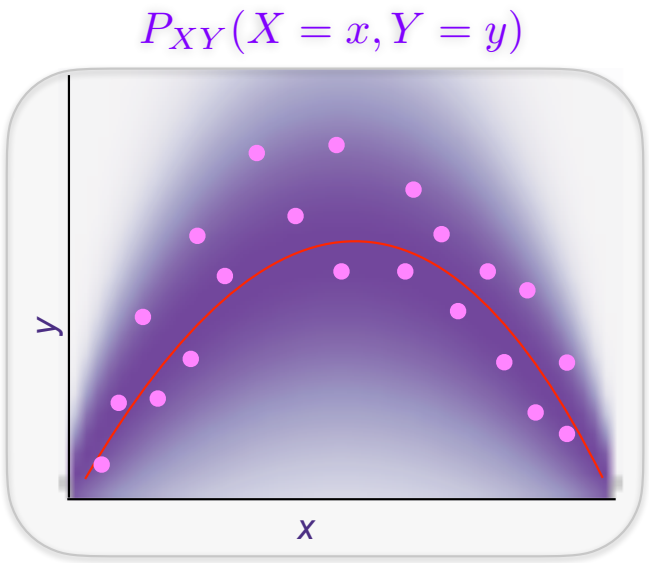
$$\mathbb{E}_{XY}[(Y-\eta(X))^2] = \mathbb{E}_X\left[\mathbb{E}_{Y|X}[(Y-\eta(x))^2|X=x]\right]$$

fix x
view $\eta(x)$
as a variable

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

Pf:
$$0 = \frac{d}{d\eta(x)} \mathbb{E}_{Y|X}\left[(Y-\eta(x))^2 \,\middle|\, X=x\right]$$
$$= \mathbb{E}_{Y|X}\left[\frac{d}{d\eta(x)}(Y-\eta(x))^2 \,\middle|\, X=x\right]$$
$$= \mathbb{E}_{Y|X}\left[-2(Y-\eta(x)) \,\middle|\, X=x\right]$$
$$= -2\,\mathbb{E}_{Y|X}[Y|X=x] + 2\eta(x)$$
$$\Rightarrow \eta(x) = \mathbb{E}_{Y|X}[Y|X=x] \qquad \square$$

Statistical Learning



Ideally, we want to find:

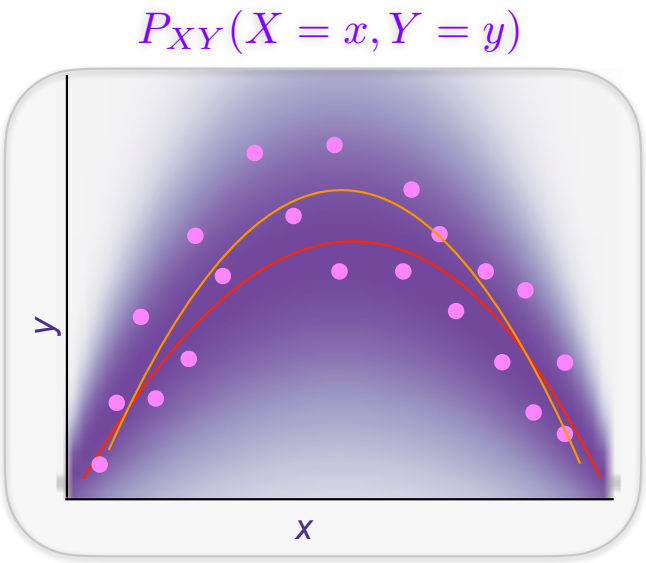
$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

But we only have samples:

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

→ estimate $\eta(x)$

Statistical Learning



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:

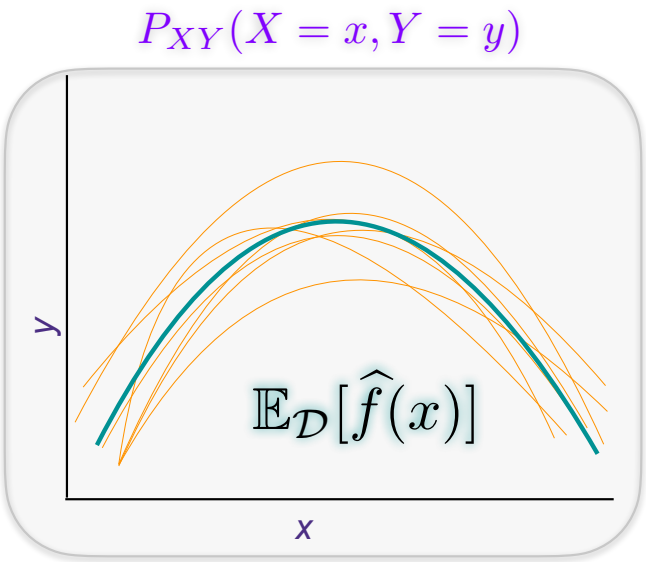
$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

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

F: . linear
. quadratic
. poly
. neural net

Q: is f random or deterministic?

Statistical Learning



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

$D = \{(x_i, y_i)\}_{i=1}^{M} \overset{i.i.d.}{\sim} P_{XY}$

$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$ $\quad$ $\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$

what we care

$\mathbb{E}_{XY}\left[(Y - \hat{f}(x))^2\right]$

$\Rightarrow \mathbb{E}_{XY}\left[\mathbb{E}_D (Y - \hat{f}_D(x))^2\right]$

$= \mathbb{E}_X\left[\mathbb{E}_{Y|X} \mathbb{E}_D (Y - \hat{f}_D(x))^2 \mid X = x\right]$

$= \mathbb{E}_X\left[\mathbb{E}_{Y|X} \mathbb{E}_D (Y - \eta_\theta(x) + \eta_\theta(x) - \hat{f}_D(x))^2 \mid X = x\right]$

$= \mathbb{E}_X\left[\mathbb{E}_{Y|X} \mathbb{E}_D ((Y - \eta_\theta(x))^2 + 2(Y - \eta_\theta(x))(\eta_\theta(x) - \hat{f}_D(x)) + (\eta_\theta(x) - \hat{f}_D(x))^2 \mid X = x\right]$

$= \mathbb{E}_{XY}\left[(Y - \eta_\theta(x))^2\right] + \mathbb{E}_X \mathbb{E}_{Y|X} \mathbb{E}_D \left[(\eta_\theta(x) - \hat{f}_D(x))^2 \mid X = x\right]$

for any x

$\mathbb{E}_{Y|X}\left[(Y - \eta_\theta(x))(\mathbb{E}_D(\eta_\theta(x) - \hat{f}_D(x)) \mid X = x\right]$

$= \mathbb{E}_{Y|X}\left[(Y - \eta_\theta(x) \mid X = x\right] \cdot \mathbb{E}_D(\eta_\theta(x) - \hat{f}_D(x))$

$\eta_\theta(x) :=$ conditional mean

$= 0$

Bias-Variance Tradeoff

$X = 0$

$Y = X$

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \qquad \widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

(1) $\mathbb{E}_{XY}\left[(Y - \eta_{\hat{f}}(X))^2\right]$: irreducible error
independent of data

(2) $\mathbb{E}_X \mathbb{E}_{Y|X}\left[\mathbb{E}_D (\eta_{\hat{f}}(X) - f_D(Y))^2 \,\middle|\, X = Y\right]$

$= \mathbb{E}_X \mathbb{E}_D\left[(\eta_{\hat{f}}(X) - f_D(X))^2\right]$

learning error
. depend on D

Bias-Variance Tradeoff

$D = \{(x_i, y_i)\}_{i=1}^{n} \overset{i.i.d}{\sim} P_{XY}$

$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x] \qquad \widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Fix x:

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

$$= \mathbb{E}_D\left[(\eta(x) - \mathbb{E}_D[\hat{f}_D(x)] + \mathbb{E}_D[\hat{f}_D(x)] - \hat{f}_D(x))^2\right]$$

$$= \mathbb{E}_D\left[(\eta(x) - \mathbb{E}_D[\hat{f}_D(x)])^2 + 2(\eta(x) - \mathbb{E}_D[\hat{f}_D(x)])(\mathbb{E}_D[\hat{f}_D(x)] - \hat{f}_D(x)) + (\mathbb{E}_D[\hat{f}_D(x)] - \hat{f}_D(x))^2\right]$$

$$= (\eta(x) - \mathbb{E}_D[\hat{f}_D(x)])^2 + \mathbb{E}_D\left[(\mathbb{E}_D[\hat{f}(x)] - \hat{f}_D(x))^2\right]$$

bias squar — variance due to D

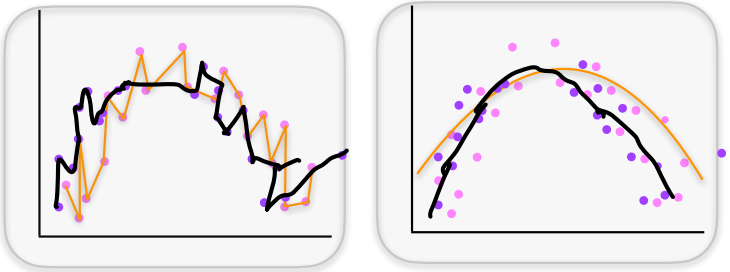
Bias-Variance Tradeoff

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

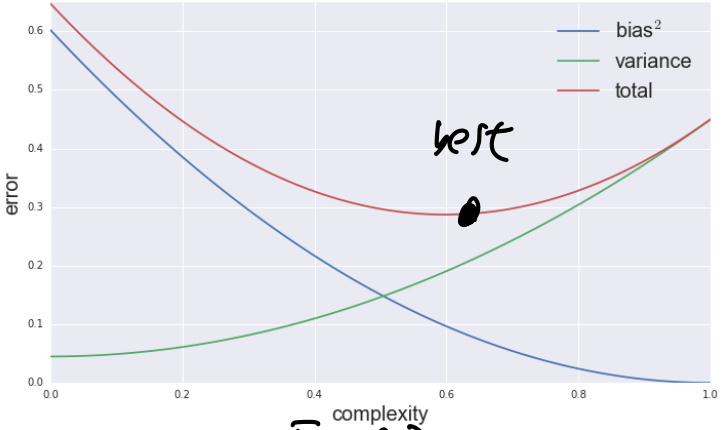
irreducible error

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

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?



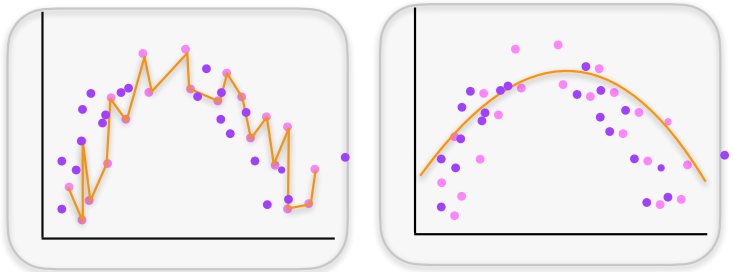
Bias-Variance Tradeoff

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

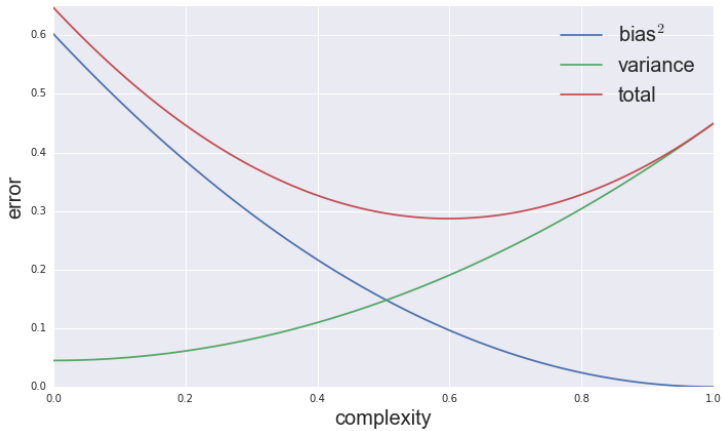
irreducible error

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

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?



Example: Linear LS

$x_i : \mathbb{R}^d$, $w \in \mathbb{R}^d$

data set $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$, $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ $n \times 1$

$$\mathbf{Y} = \mathbf{X}w + \epsilon \quad n \times d$$

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

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

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

MLE:

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

irreducible error:

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

Example: Linear LS: compute bias

$D = (X, Y)$

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

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

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

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

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

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

bias squared

$$\mathbb{E}_D[\hat{f}_D(x)] = \mathbb{E}_D[x^T w + x^T (X^T X)^{-1} X\epsilon] \quad // \; 0$$

$$= x^T w + \mathbb{E}_X[x^T (X^T X)^T X] \cdot \mathbb{E}_Y[\epsilon]$$

$$= x^T w$$

Example: Linear LS: compute variance

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

$n \times 1$

$I = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$

$n \times 4$

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

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

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

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

$w^T x$

variance

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

$\mathbb{E}[\epsilon_i \epsilon_j] = \begin{cases} \sigma^2 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$

$$= \mathbb{E}_{\mathcal{D}}[x^T(X^TX)^{-1}X^T\epsilon\epsilon^T X(X^TX)^{-1}x]$$

$$= \mathbb{E}_X[x^T(X^TX)^{-1}X^T \mathbb{E}_{Y|X}[\epsilon\epsilon^T] X(X^TX)^{-1}x]$$

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

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

Example: Linear LS: compute variance

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

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

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

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

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

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

$$\approx \sigma^2 \mathbb{E}_X\left[x^T\left(n\Sigma\right)^{-1}x\right]$$

$$= \frac{\sigma^2}{n}\mathbb{E}_X\left[Tr\left(xx^T(\Sigma)^{-1}\right)\right]$$

$$= \frac{\sigma^2}{n} \cdot Tr\left(I_d\right) = \frac{d\sigma^2}{n}$$

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_X$

$\mathbb{E}_{P_X}[XX^T] = \Sigma \in \mathbb{R}^{d \times d}$

$X^TX = \sum_{i=1}^{n} x_i x_i^T$

as $n \to \infty$

by Central Limit Theorem

$\rightarrow n \cdot \Sigma$

$Tr(A) = \sum A_{ii}$

Example: Linear LS

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

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

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

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

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

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

irreducible error

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

bias squared

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

variance

$n \uparrow$ var $\downarrow$

$d \uparrow$ var $\uparrow$

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:

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

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

Complexity grows as k grows

Training set error as a function of model complexity

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

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

TRAIN error:

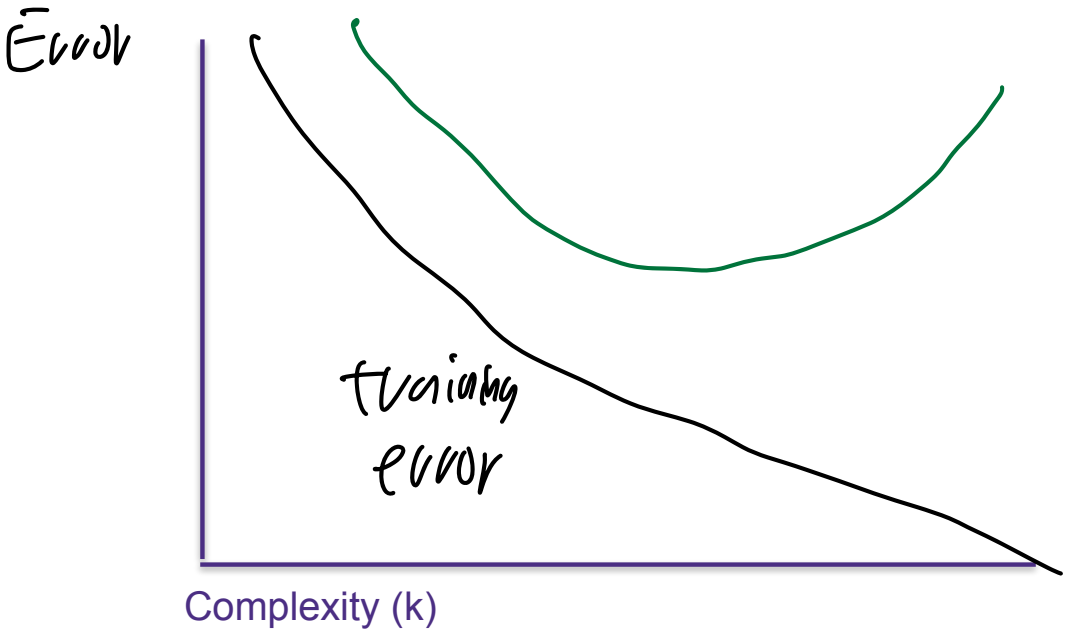
$$\mathcal{D} \overset{i.i.d.}{\sim} P_{XY}$$

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

TRUE error:

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

Cannot observe



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 = $\mathcal{D} \cup \mathcal{T} \sim P_{XY}$

$\mathcal{D}$: training, $\mathcal{T}$: test

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:

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

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

TEST error:

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

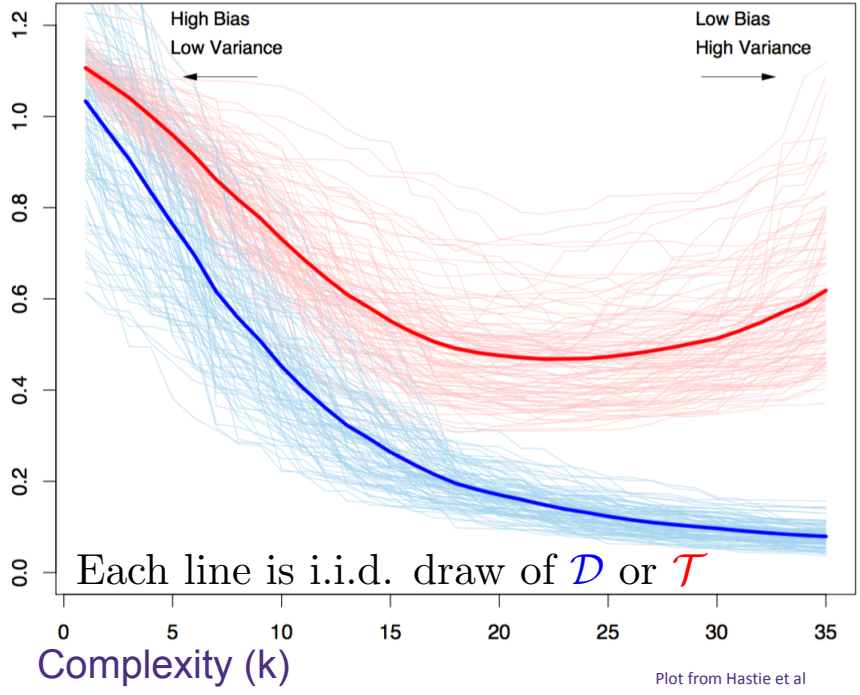
$$\frac{1}{|\mathcal{T}|} \sum_{(x_i, y_i) \in \mathcal{T}} (y_i - \widehat{f}_{\mathcal{D}}^{(k)}(x_i))^2$$

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

Training set error as a function of model complexity

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

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



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:

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

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

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:

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

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

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)**