- HW1 out today due Jan 25th Tuesday midnight.

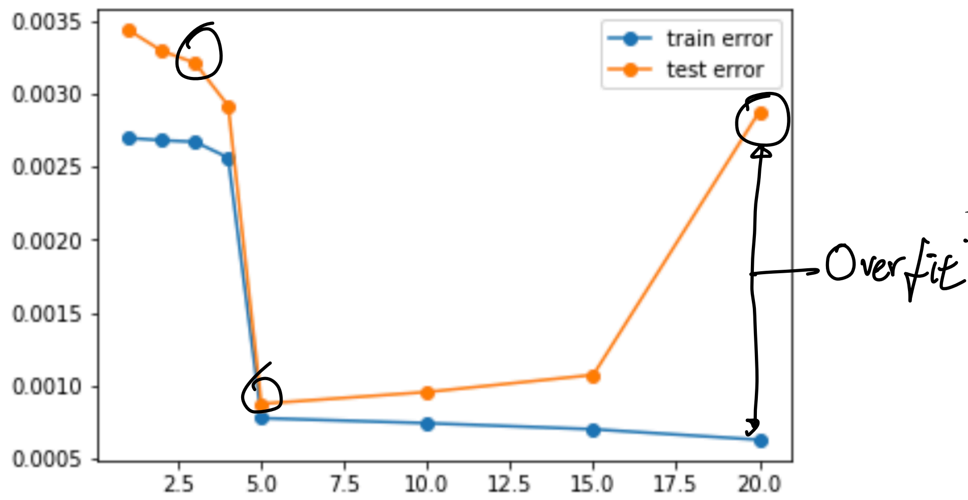
- Demi

Lecture 5:
Bias-Variance Tradeoff

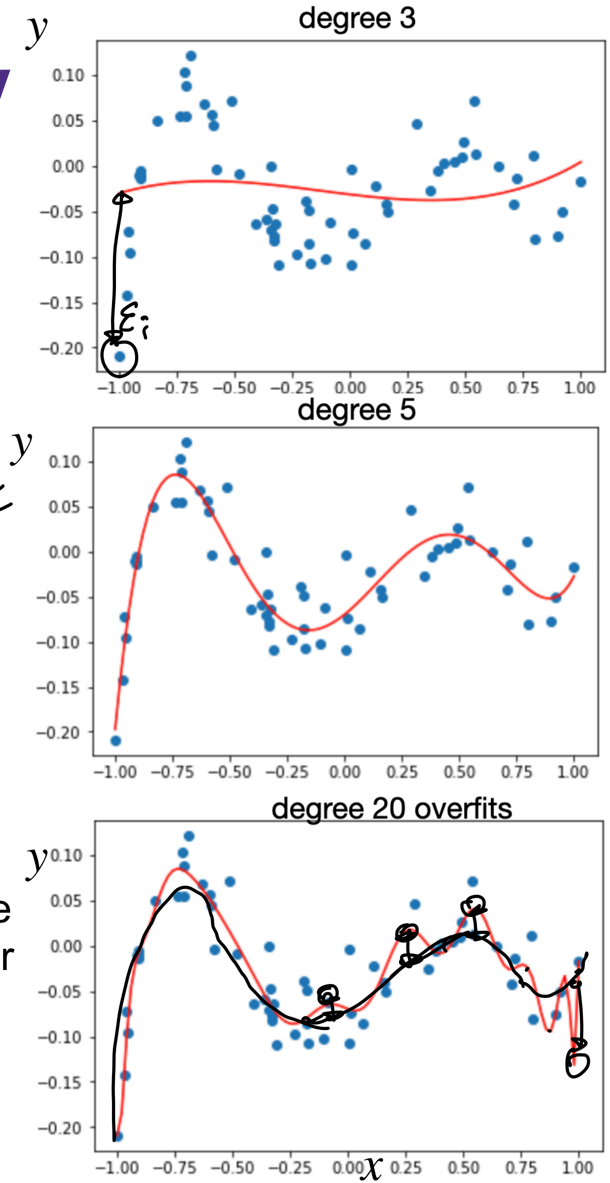
- explaining test error using theoretical analysis

Train/test error vs. complexity

Error



degree p of the polynomial regression



- **Model complexity** e.g., degree p of the polynomial model, number of features used in diabetes example
  - Related to the dimension of the model parameter
- **Train error** monotonically decreases with model complexity
- **Test error** has a U shape

Statistical learning

Typical notation:
X denotes a random variable
x denotes a deterministic instance

- Suppose data is generated from a statistical model $(X, Y) \sim P_{X,Y}$
  - and assume we know $P_{X,Y}$ (just for now to explain statistical learning)
- Then **learning** is to find a predictor $\eta : \mathbb{R}^d \to \mathbb{R}$ that minimizes
  - the expected error $\mathbb{E}_{(X,Y)\sim P_{X,Y}}[(Y - \eta(X))^2]$ = True Error
    ↳ True dist P_{XY}
  - think of this random (X, Y) as a new sample you will encounter when you deployed your learned model, and we care about its average performance
- Since, we do not assume anything about the function $\eta(x)$, it can take any value for each $X = x$, hence the optimization can be broken into sum (or more precisely integral) of multiple objective functions, each involving a specific value $X = x$
  - $\mathbb{E}_{(X,Y)\sim P_{X,Y}}[(Y - \eta(X))^2] \;=\; \mathbb{E}_{X\sim P_X}\Big[\mathbb{E}_{Y\sim P_{Y|X}}[(Y - \eta(x))^2 \,|\, X = x]\Big]$

$$= \int \mathbb{E}_{Y\sim P_{Y|X}}[(Y - \eta(x))^2 \,|\, X = x]\, P_X(x)\, dx$$

Or for discrete X, $\quad = \sum_x P_X(x)\, \mathbb{E}_{Y\sim P_{Y|X}}[(Y - \eta(x))^2 \,|\, X = x]$

Where we used the chain rule: $\mathbb{E}_{X,Y}[f(X, Y)] = \mathbb{E}_X\Big[\mathbb{E}_{Y|X}[f(x, Y) \,|\, X = x]\Big]$

Statistical learning

- We can solve the optimization for each $X = x$ separately
  - $\eta(x) \;=\; \arg\min_{a \in \mathbb{R}} \mathbb{E}_{Y \sim P_{Y|X}}[(Y-a)^2 \,|\, X = x]$
- The optimal solution is $\eta(x) = \mathbb{E}_{Y \sim P_{Y|X}}[Y \,|\, X = x]$, which is the best prediction in ℓ_2-loss/Mean Squared Error
- Claim: $\mathbb{E}_{Y \sim P_{Y|X}}[Y \,|\, X = x] \;=\; \arg\min_{a \in \mathbb{R}} \mathbb{E}_{Y \sim P_{Y|X}}[(Y-a)^2 \,|\, X = x]$
- Proof: $\frac{\partial}{\partial a} \mathbb{E}_{Y|X}\left[Y^2 - 2aY + a^2 \,\middle|\, X=x\right] = \frac{\partial \{\mathbb{E}[Y^2 | X=x] - 2a\,\mathbb{E}[Y | X=x] + a^2\}}{\partial a}$

$$= -2\,\mathbb{E}[Y \,|\, X=x] + 2a \Big|_{a=\eta(x)} = 0$$

Opeimal Model Predictor : $\eta(x) = \mathbb{E}[Y \,|\, X=x]$

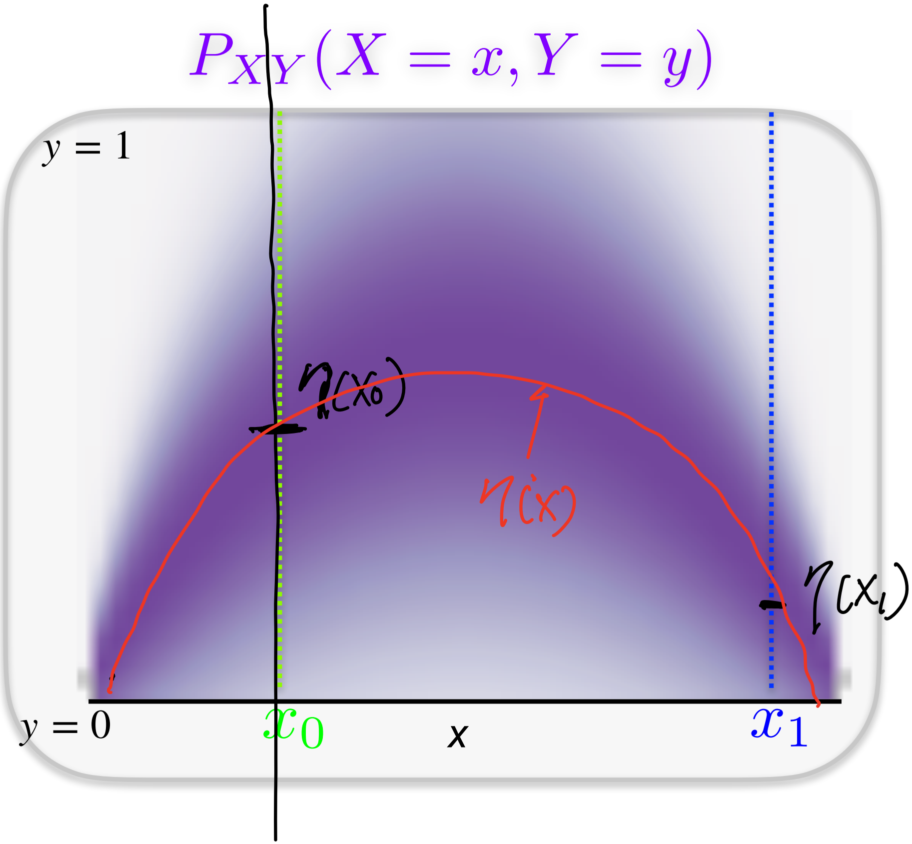
- Note that this optimal statistical estimator $\eta(x) = \mathbb{E}[Y \,|\, X = x]$ cannot be implemented as we do not know $P_{X,Y}$ in practice
- This is only for the purpose of conceptual understanding

Statistical Learning

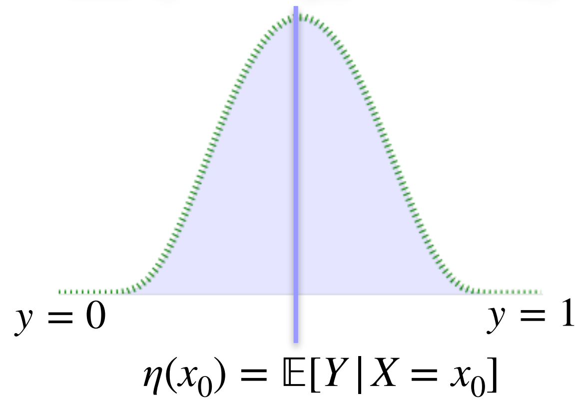
Ideally, we want to find:

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

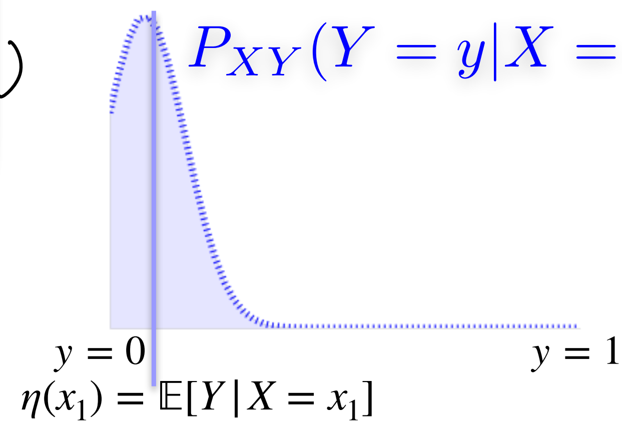
$P_{XY}(X=x, Y=y)$



$P_{XY}(Y=y|X=x_0)$

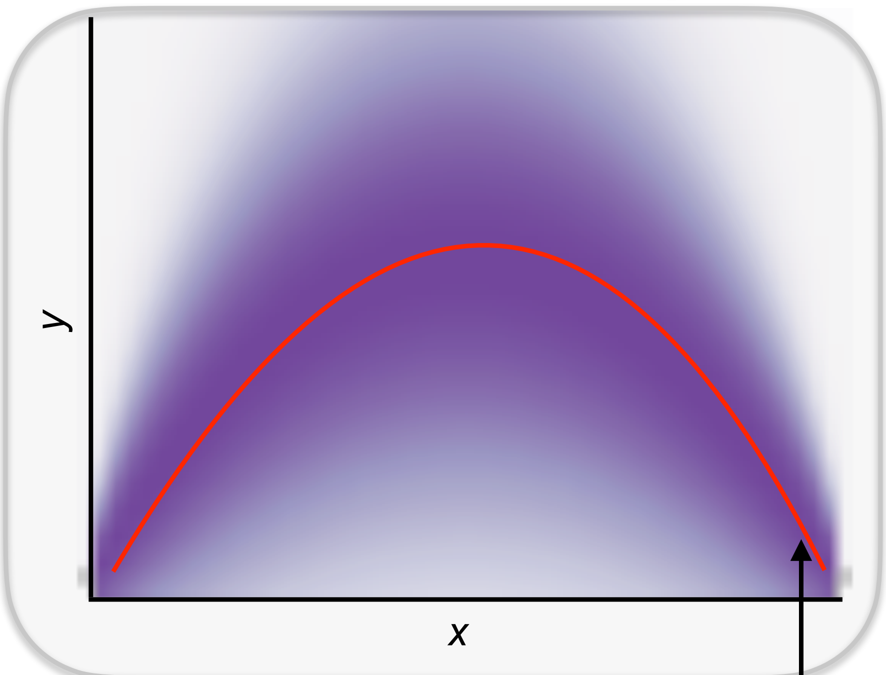


$P_{XY}(Y=y|X=x_1)$



Statistical Learning

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



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

Ideally, we want to find:

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

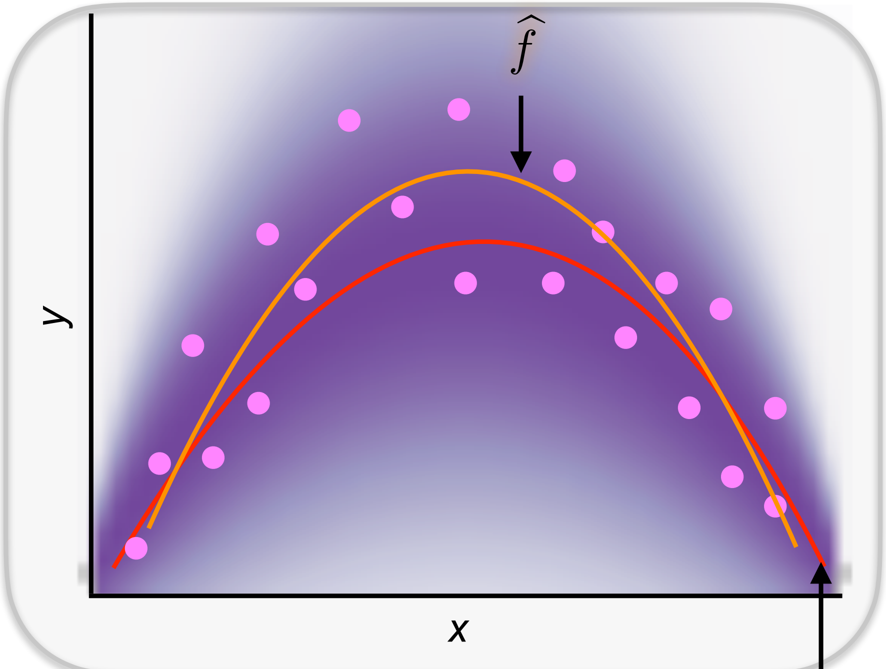
But we do not know $P_{X,Y}$

We only have samples.

Q. How is $\hat{f}(x)$ related to optimal predictor $\eta(x)$

Statistical Learning

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



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

Ideally, we want to find:

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

But we only have samples:

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

So we need to restrict our predictor to a function class (e.g., linear, degree-p polynomial) to avoid overfitting:

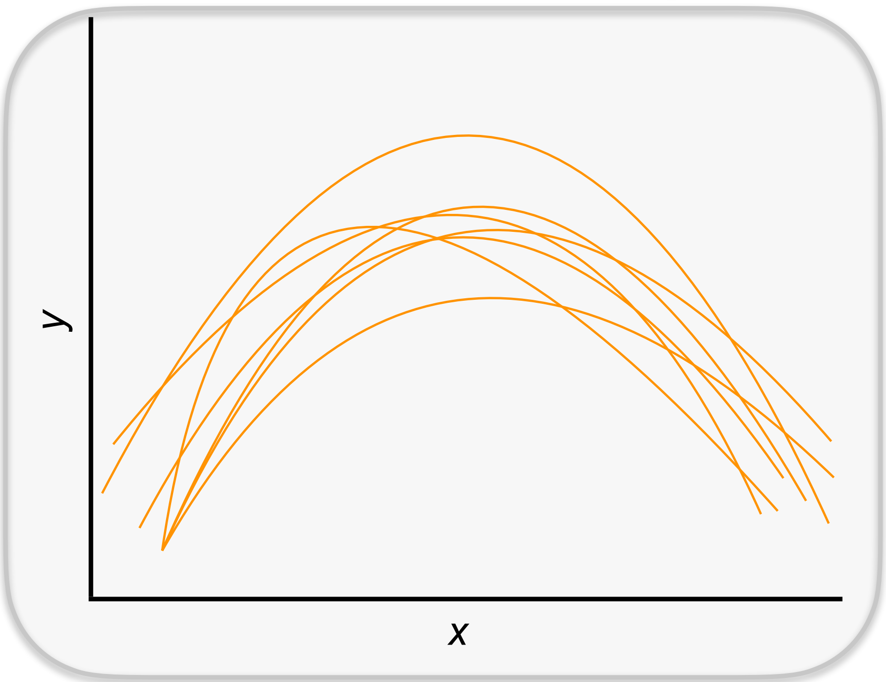
$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

We care about how our predictor performs on future unseen data

True Error of $\hat{f}$: $\mathbb{E}_{X,Y}[(Y - \hat{f}(X))^2]$

Future prediction error $\mathbb{E}_{X,Y}[(Y - \hat{f}(X))^2]$ is random because $\hat{f}$ is random (whose randomness comes from training data $\mathscr{D}$)

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



True Error: $\mathbb{E}_{X,Y}[(Y - \hat{f}(x))^2]$

Random

Average True Error: $\mathbb{E}_{D}\left[\mathbb{E}_{X,Y}[(Y - \hat{f}(x))^2]\right]$

*Mayura

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

Bias-variance tradeoff

Notation:
I use predictor/model/estimate, interchangeably

Ideal predictor

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

Learned predictor

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

- We are interested in the **True Error** of a (random) learned predictor:

$$\mathbb{E}_{X,Y}[(Y - \hat{f}_{\mathcal{D}}(X))^2] \quad \leftarrow \text{Random}$$

- But the analysis can be done for each $X = x$ separately, so we analyze the **conditional true error**:

$$\mathbb{E}_{Y|X}[(Y - \hat{f}_{\mathcal{D}}(x))^2 \,|\, X = x] \quad \leftarrow \text{Random}$$

- And we care about the **average conditional true error**, averaged over training data:

$$\mathbb{E}_{\mathcal{D}}\Big[\mathbb{E}_{Y|X}[(Y - \hat{f}_{\mathcal{D}}(x))^2 \,|\, X = x]\Big] \quad \leftarrow \text{deterministic}$$

written compactly as $= \mathbb{E}[(Y - \hat{f}_{\mathcal{D}}(x))^2]$

Bias-variance tradeoff

Ideal predictor

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

Learned predictor

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

- **Average conditional true error**:

$$\mathbb{E}_{\mathscr{D},Y|x}[(Y - \hat{f}_{\mathscr{D}}(x))^2] = \mathbb{E}_{\mathscr{D},Y|x}[(Y \underbrace{- \eta(x) + \eta(x)}_{0} - \hat{f}_{\mathscr{D}}(x))^2]$$

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

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

$Y|X$ and D are independent.

$$= \mathbb{E}_{Y|X}[Y - \eta(x)] \cdot \mathbb{E}[\eta(x) - \hat{f}_D(x)] = 0$$

$$\underbrace{\mathbb{E}[Y|X=x]}_{=\eta(x)} - \eta(x) = 0$$

$$= \underbrace{\mathbb{E}_{Y|X}\left[(Y - \eta(x))^2 \,|\, X = x\right]}_{\text{Irreducible Error}} + \underbrace{\mathbb{E}_D\left[(\eta(x) - \hat{f}_D(x))^2\right]}_{\text{Expected Learning Error}}$$

Bias-variance tradeoff

any function = Largest function class

*Matthew W.

*

Ideal predictor

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

best predictor you get if you have infinite samples

Learned predictor

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

smaller class

finite samples

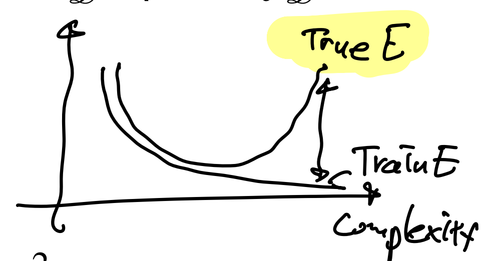
- **Average conditional true error**:

$$\mathbb{E}_{\mathcal{D},Y|x}[(Y-\hat{f}_{\mathcal{D}}(x))^2] \;=\; \mathbb{E}_{\mathcal{D},Y|x}[(Y-\eta(x)+\eta(x)-\hat{f}_{\mathcal{D}}(x))^2]$$

$$= \mathbb{E}_{\mathcal{D},Y|x}\Big[(Y-\eta(x))^2 + 2(Y-\eta(x))(\eta(x)-\hat{f}_{\mathcal{D}}(x)) + (\eta(x)-\hat{f}_{\mathcal{D}}(x))^2\Big]$$

$$= \mathbb{E}_{Y|x}[(Y-\eta(x))^2] + 2\underbrace{\mathbb{E}_{\mathcal{D},Y|x}[(Y-\eta(x))(\eta(x)-\hat{f}_{\mathcal{D}}(x))]}_{=0} + \mathbb{E}_{\mathcal{D}}[(\eta(x)-\hat{f}_{\mathcal{D}}(x))^2]$$

(this follows from independence of $\mathcal{D}$ and (X,Y) and $\mathbb{E}_{Y|x}[Y-\eta(x)] = \mathbb{E}[Y|X=x]-\eta(x) = 0$)



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

Irreducible error

(a) Caused by stochastic label noise in $P_{Y|X=x}$

(b) cannot be reduced

Average learning error

Caused by

(a) either using too “simple” of a model or

(b) not enough data to learn the model accurately

Bias-variance tradeoff

Ideal predictor

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

Learned predictor

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

- **Average learning error**:

$$\mathbb{E}_{\mathcal{D}}[(\eta(x) - \hat{f}_{\mathcal{D}}(x))^2] = \mathbb{E}_{\mathcal{D}}\Big[\Big(\eta(x) \underbrace{- \overbrace{\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]}^{\bar{f}(x)} + \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]}_{=0} - \hat{f}_{\mathcal{D}}(x)\Big)^2\Big]$$

$$= (\eta(x) - \bar{f}(x))^2 + \cancel{2 \cdot \mathbb{E}\big[\underbrace{(\eta(x) - \bar{f}(x))}_{\text{Det.}}(\bar{f}(x) - \hat{f}_D(x))\big]}^{\,0} + \mathbb{E}\big[(\bar{f}(x) - \hat{f}_D(x))^2\big]$$

$$= \mathbb{E}[\bar{f}(x) - \hat{f}_D(x)]$$

$$= \bar{f}(x) - \underbrace{\mathbb{E}[\hat{f}_D(x)]}_{=\bar{f}(x)} = 0$$

$$= \underbrace{(\eta(x) - \bar{f}(x))^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\big[(\bar{f}(x) - \hat{f}_D(x))^2\big]}_{\text{Variance of } \hat{f}_D(x)}$$

Bias-variance tradeoff

Ideal predictor

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

Learned predictor

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

$(a-b+b-c)^2 = (a-b)^2 + 2(a-b)(b-c) + (b-c)^2$

- **Average learning error**:

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

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

$$= \underbrace{\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2}_{\textbf{biased squared}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)\right)^2\right]}_{\textbf{variance}}$$

Bias-variance tradeoff

Model class $\tilde{\mathcal{F}}$
Traing Data Size

× Doniza

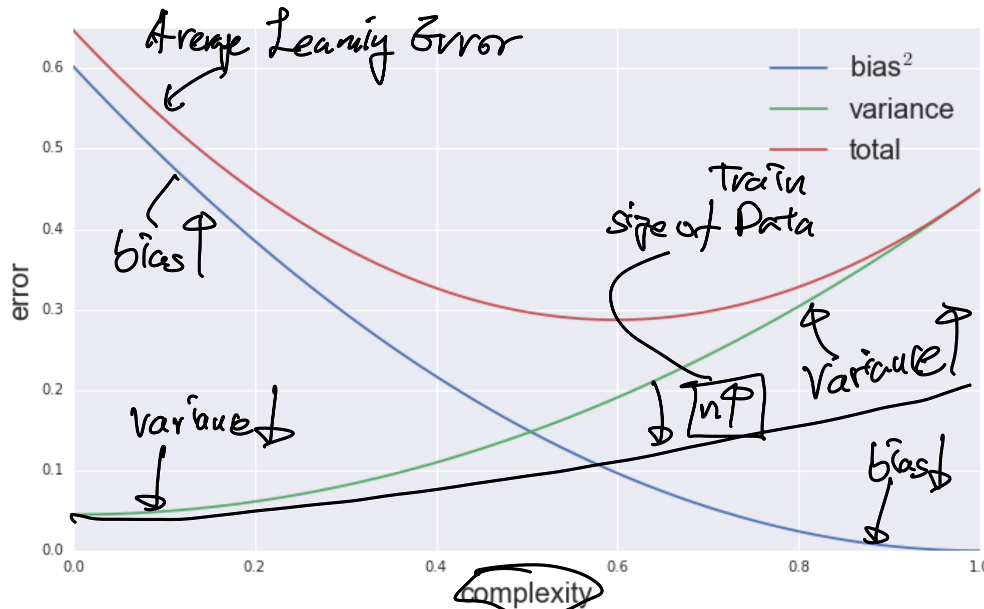
- **Average conditional true error**:

$$\mathbb{E}_{\mathcal{D},Y|x}[(Y - \hat{f}_{\mathcal{D}}(x))^2] = \underbrace{\mathbb{E}_{Y|x}\left[(Y - \eta(x))^2\right]}_{\textbf{irreducible error}}$$

$$+ \underbrace{\left(\eta(x) - \mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]\right)^2}_{\textbf{biased squared}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)] - \hat{f}_{\mathcal{D}}(x)\right)^2\right]}_{\textbf{variance}}$$

Bias squared:
measures how the predictor is mismatched with the best predictor in expectation

variance:
measures how the predictor varies each time with a new training datasets



Questions?

Model $\overset{\text{class } \mathcal{F}}{\text{Complexity}}$

$$\underset{f \in \mathcal{F}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - f(x_i)\right)^2$$

$$\underset{\text{simple}}{\mathcal{F}_1} \subseteq \underset{\text{complex}}{\mathcal{F}_2}$$

Constant $\subseteq$ Linear $\subseteq$ Quadratic $\subseteq$ . . .