Bias-Variance

Features	Train MSE	Test MSE
All	2640	3224
S5 and BMI	3004	3453
S5	3869	4227
BMI	3540	4277
S4 and S3	4251	5302
S4	4278	5409
S3	4607	5419
None	5524	6352

- **test MSE is the primary criteria for model selection**
- Using only 2 features (S5 and BMI), one can get very close to the prediction performance of using all features
- Combining S3 and S4 does not give any performance gain

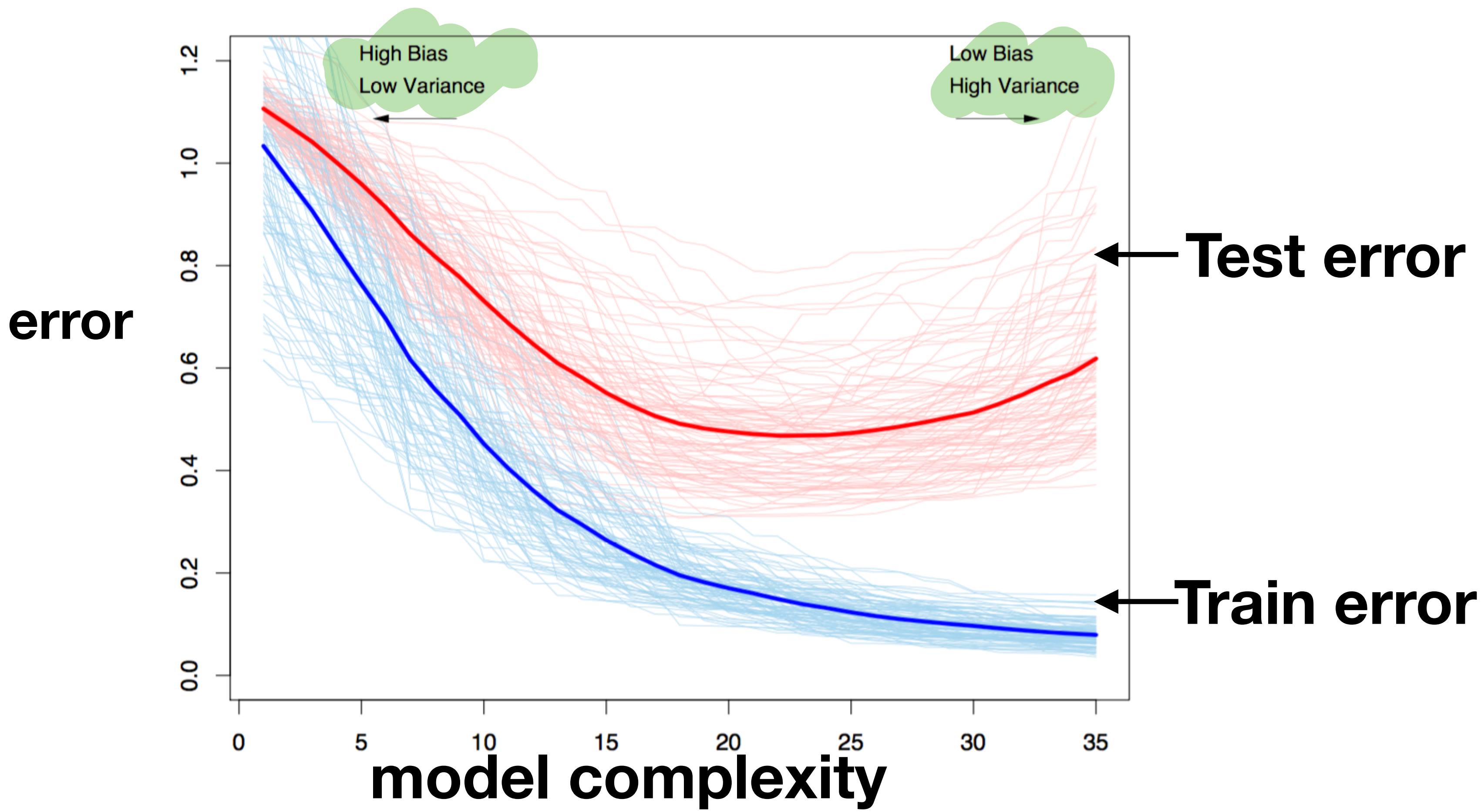
demo3_diabetes.ipynb

What does the bias-variance theory tell us?

- **Train error** (random variable, randomness from $\mathcal{D}$)
  - Use $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \sim P_{X,Y}$ to find $\widehat{w}$
  - Train error: $\mathcal{L}_{\text{train}}(\widehat{w}_{\text{LS}}) = \dfrac{1}{|\mathcal{D}|} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i - \widehat{w}^T x_i)^2$
- recall the **test error** is an unbiased estimator of the **true error**
- **True error** (random variable, randomness from $\mathcal{D}$)
  - True error: $\mathcal{L}_{\text{true}}(\widehat{w}) = \mathbb{E}_{(x,y)\sim P_{X,Y}}[(y - \widehat{w}^T x)^2]$
- **Test error** (random variable, randomness from $\mathcal{D}$ and $\mathcal{T}$)
  - Use $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^m \sim P_{X,Y}$
  - Test error: $\mathcal{L}_{\text{test}}(\widehat{w}) = \dfrac{1}{|\mathcal{T}|} \sum_{(x_i,y_i)\in\mathcal{T}} (y_i - \widehat{w}^T x_i)^2$
- theory explains **true error**, and hence expected behavior of the (random) **test error**

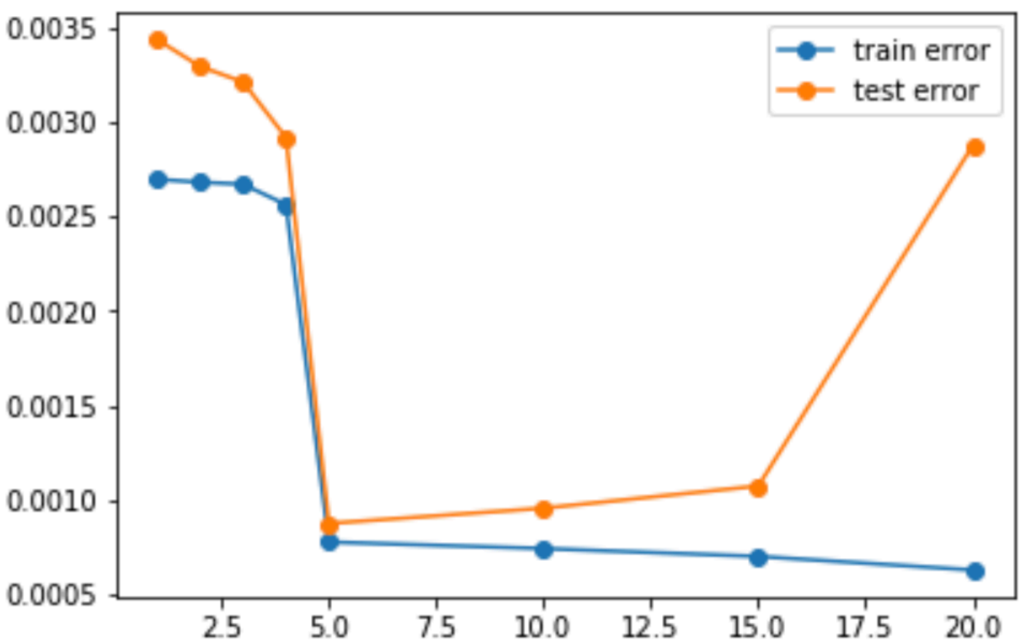
What does bias-variance theory tell us?

- Train error is optimistically biased (i.e. smaller) because the trained model is minimizing the train error
- Test error is unbiased estimate of the true error, if test data is never used in training a model or selecting the model complexity
- Each line is an i.i.d. instance of $\mathscr{D}$ and $\mathscr{T}$

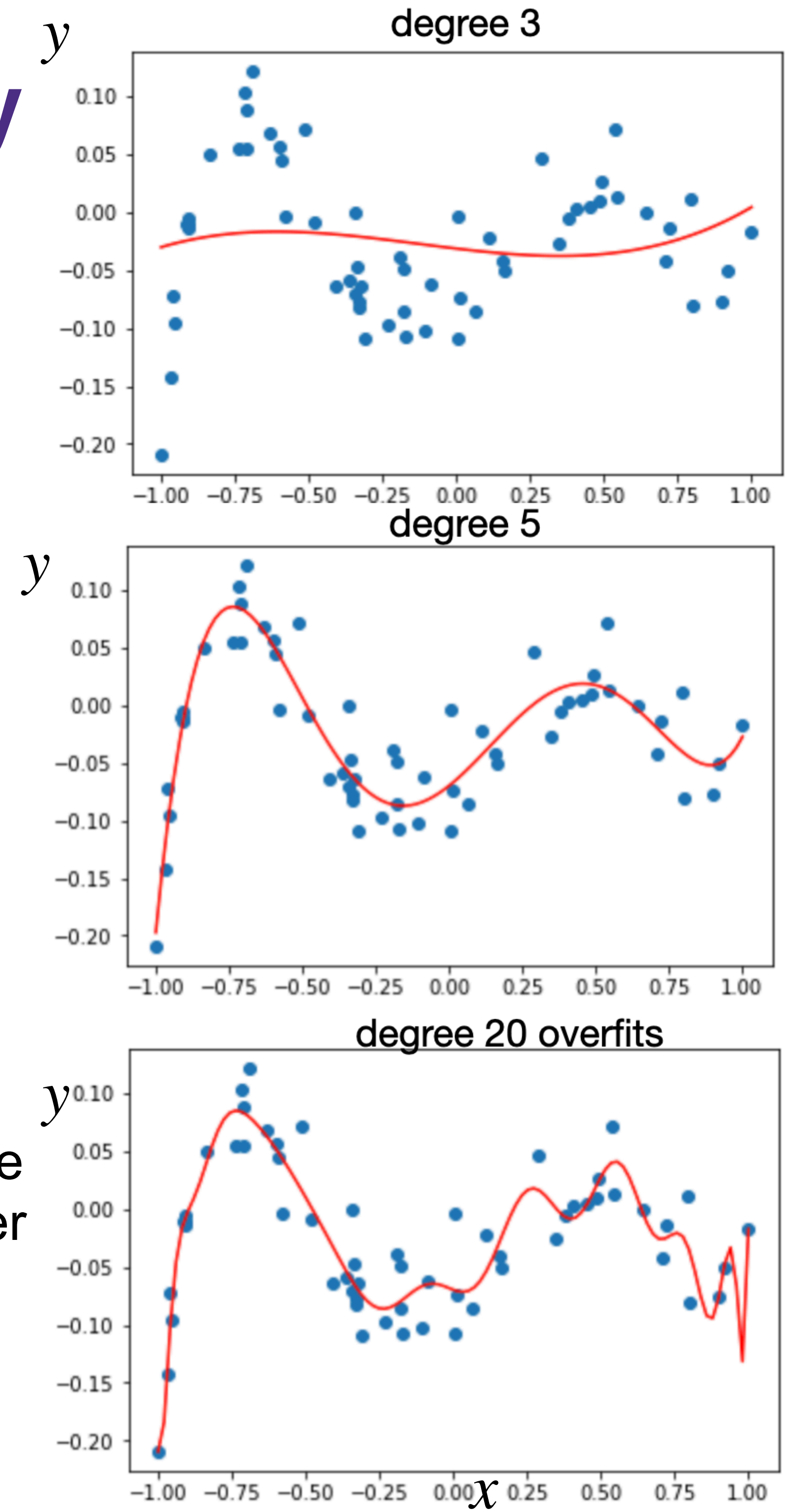


Train/test error vs. complexity

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)
- **learning** aims to find a predictor $\eta : \mathbb{R}^d \to \mathbb{R}$ that minimizes
  - expected error $\mathbb{E}_{(X,Y)\sim P_{X,Y}}[(Y - \eta(X))^2]$
  - think of random (X, Y) as a new sample you will encounter when you deployed your learned model, and we care about its average performance
- We assume the function $\eta(x)$ could be anything
  - it can take any value for each $X = x$
- So the optimization can be done separately for each $X = x$
  - $\mathbb{E}_{(X,Y)\sim P_{X,Y}}[(Y - \eta(X))^2] = \mathbb{E}_{X\sim P_X}\Big[\mathbb{E}_{Y\sim P_{Y|X}}[(Y - \eta(x))^2 \,|\, X = x]\Big]$

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

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

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

- The optimal predictor sets its value for each $X = x$ separately
  - $$\eta(x) = \arg\min_{a \in \mathbb{R}} \mathbb{E}_{Y \sim P_{Y|X}}[(Y-a)^2 \,|\, X = x]$$

$\eta(x)$ min

$(\eta(x) - Y)^2 | X=$

- 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:

$$\arg\min_a \mathbb{E}[Y^2 - 2aY + a^2 \,|\, X=x]$$

$$= \arg\min_a \left(\mathbb{E}[Y^2|X=x] - 2\mathbb{E}[aY|X=x] + \mathbb{E}[a^2|X=x]\right) \text{ (L of Exp)}$$

$$= \arg\min_a \sum_{Y=y} \Pr[Y=y \,|\, X=x]\left[y^2 - 2ay + a^2\right]$$

$$\nabla_a = \sum_{Y=y} \Pr[Y=y \,|\, X=x](-2y + 2a) = 0 \text{ \& solve}$$

$$a = \sum \Pr[Y=y|X=x] \cdot y$$

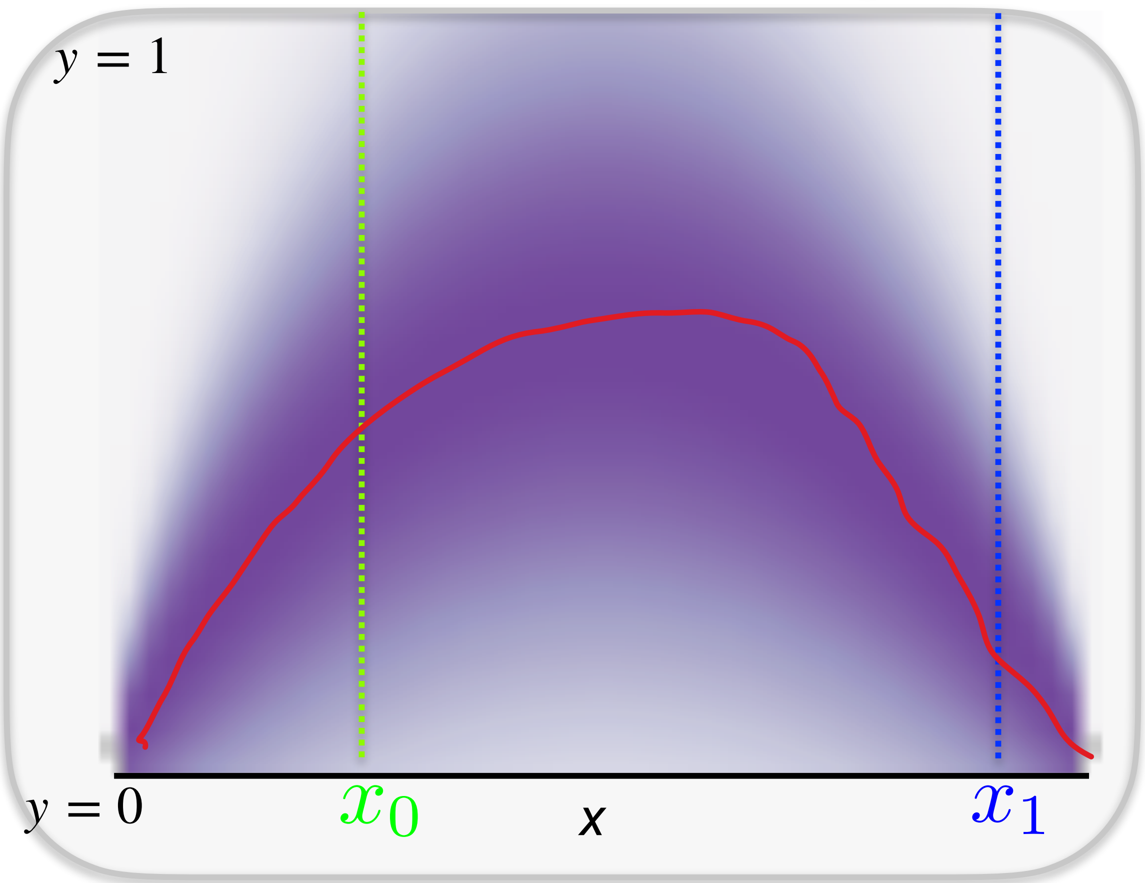
- Can't implement optimal statistical estimator $\eta(x) = \mathbb{E}[Y \,|\, X = x]$
  - as we do not know $P_{X,Y}$ in practice
- This is only for the purpose of conceptual understanding

Statistical Learning

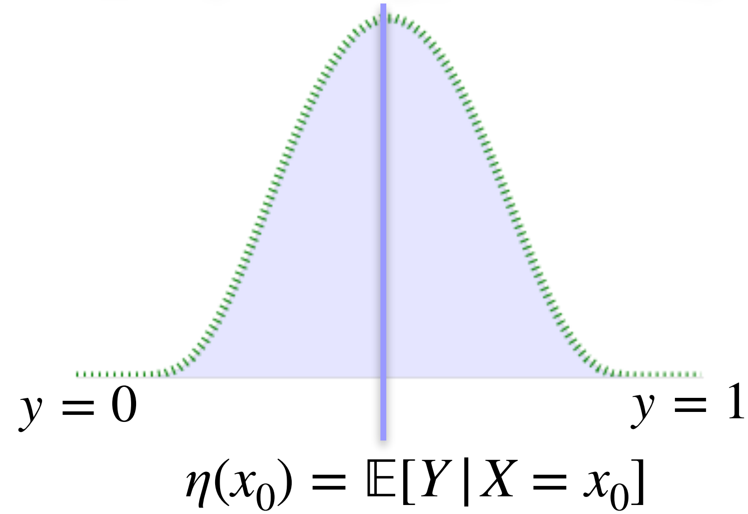
Ideally, we want to find:

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

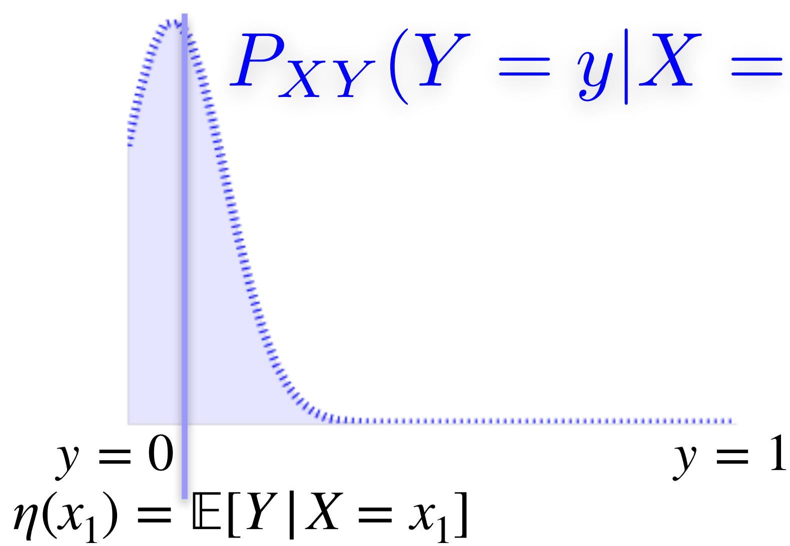
$$P_{XY}(X=x, Y=y)$$



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

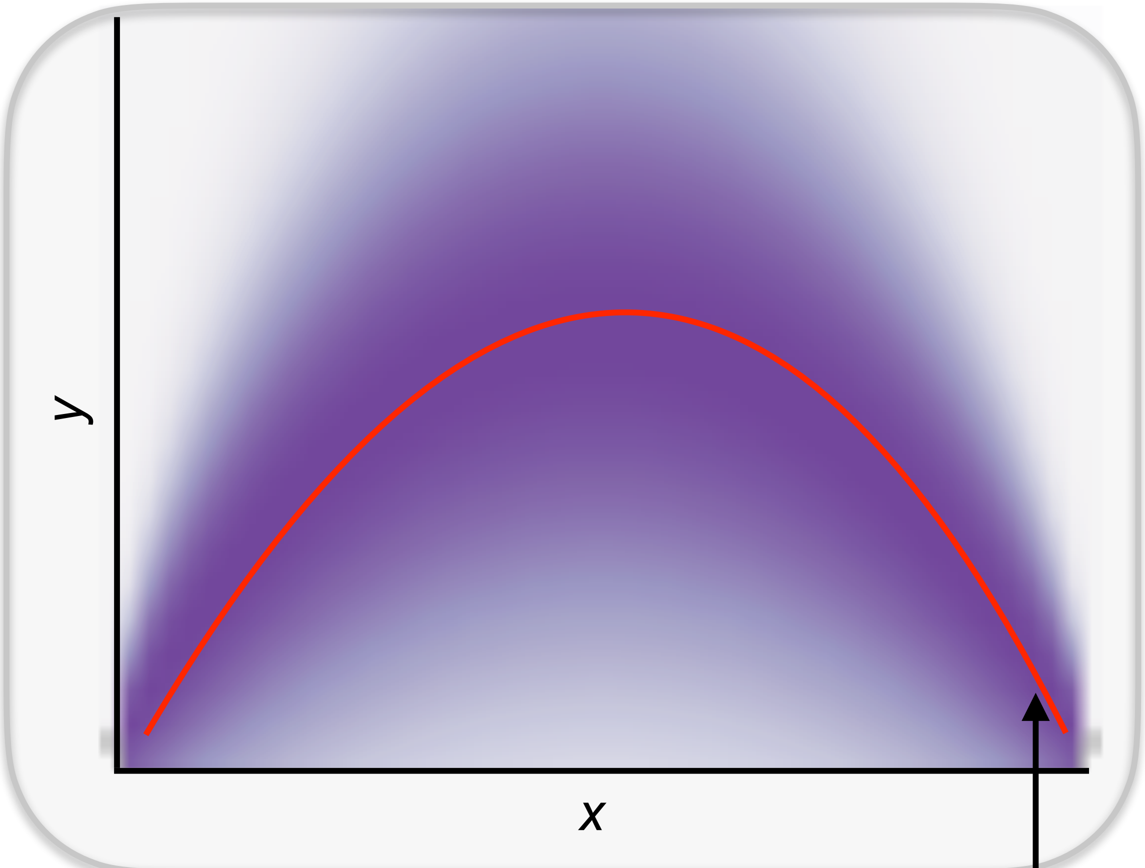


$$P_{XY}(Y=y|X=x_1)$$



Statistical Learning

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



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

Ideally, we want to find:

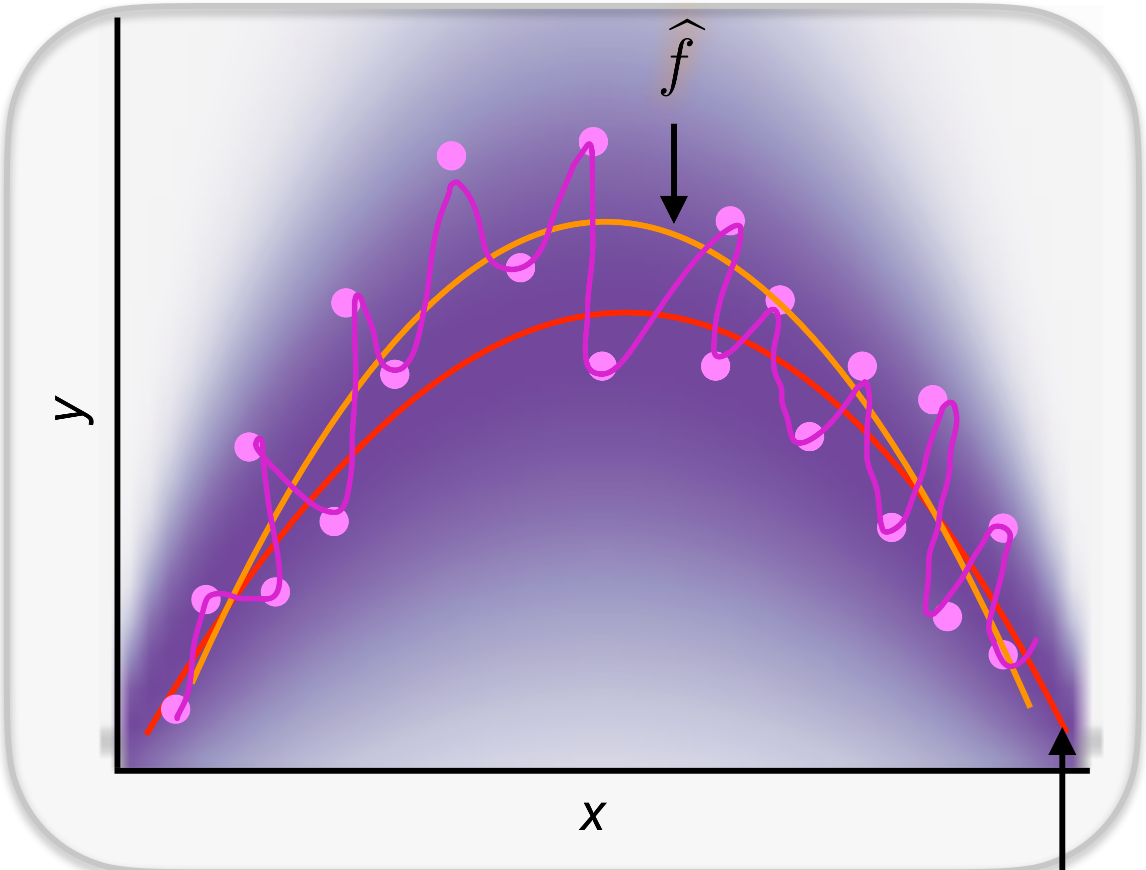
$$\eta(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

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

We only have samples.

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:

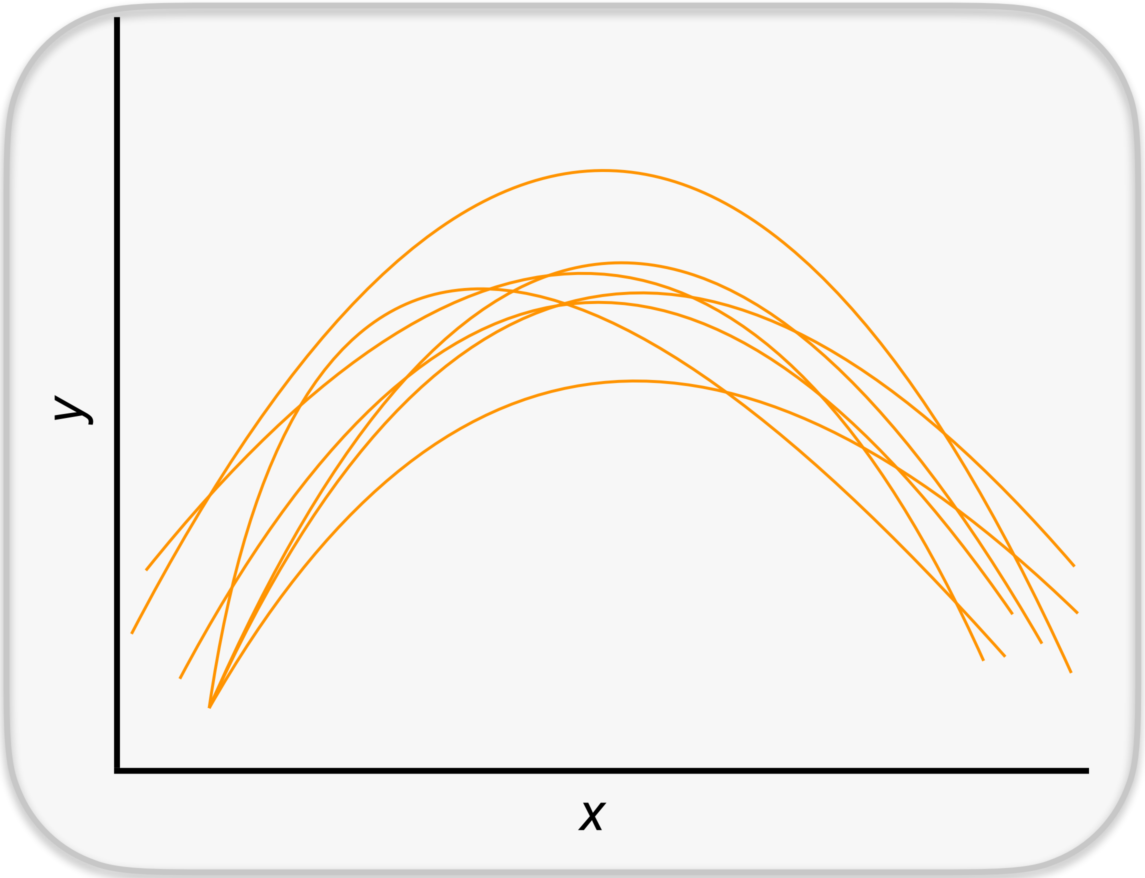
$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

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 $\mathscr{D}$)

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



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

Bias-variance tradeoff

Notation:
I use predictor/model/estimate, interchangeably

Ideal predictor

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

Learned predictor

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

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

$$\mathbb{E}_{X,Y}[(Y - \hat{f}_{\mathscr{D}}(X))^2]$$

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

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

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

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

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

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 conditional true error**:

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

$$\mathbb{E}[A^2] + \mathbb{E}[B^2] + 2\,\mathbb{E}[AB]$$

$$\mathbb{E}[Y - \eta(x) \mid X = x]\;\mathbb{E}[\eta(x) - \hat{f}_{D}(x)]$$

Bias-variance tradeoff

Ideal predictor

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

Learned predictor

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

- **Average conditional true error**:

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

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

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

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

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

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**:

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

Bias-variance tradeoff

Ideal predictor

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

Learned predictor

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

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 learning error**:

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 learning error**:

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

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 learning error**:

$$\mathbb{E}_{\mathscr{D}}[(\eta(x) - \hat{f}_{\mathscr{D}}(x))^2] \;=\; \mathbb{E}_{\mathscr{D}}\Big[\Big(\overbrace{\eta(x) - \mathbb{E}_{\mathscr{D}}[\hat{f}_{\mathscr{D}}(x)]}^{A} + \overbrace{\mathbb{E}_{\mathscr{D}}[\hat{f}_{\mathscr{D}}(x)] - \hat{f}_{\mathscr{D}}(x)}^{B}\Big)^2\Big]$$

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

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 learning error**:

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

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

$$+ (\mathbb{E}_{\mathscr{D}}[\hat{f}_{\mathscr{D}}(x)] - \hat{f}_{\mathscr{D}}(x))^2\Big]$$

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**:

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

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

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

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**:

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

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

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

Bias-variance tradeoff

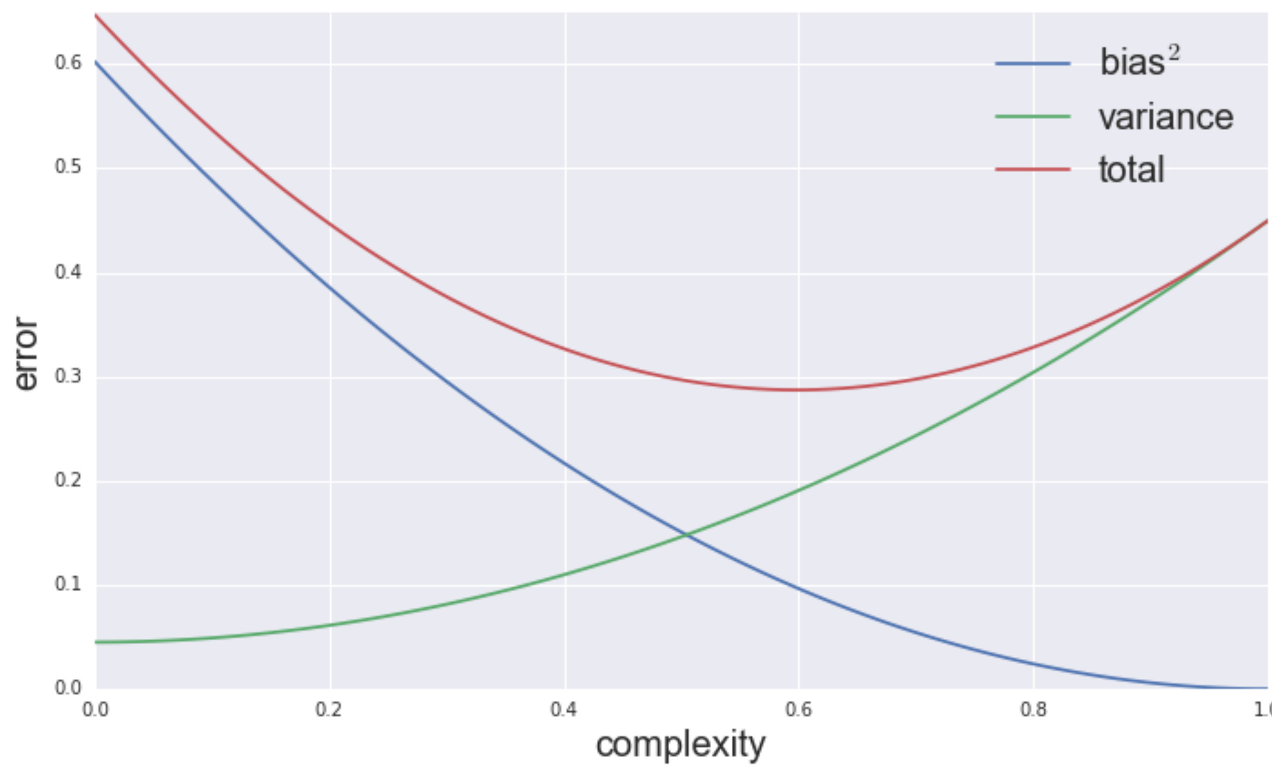
- **Average conditional true error**:

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

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

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



Regularization

Sensitivity: how to detect overfitting

- For a linear model,
  $y \simeq b + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$
  if $|w_j|$ is large then the prediction is sensitive to small changes in x_j
- Large sensitivity leads to overfitting and poor generalization, and equivalently models that overfit tend to have large weights
- Note that b is a constant and hence there is no sensitivity for the offset b

- In **Ridge Regression,** we use a regularizer $||w||_2^2$ to measure and control the sensitivity of the predictor
- And optimize for small loss and small sensitivity, by adding a **regularizer** in the objective (assume no offset for now)

$$\widehat{w}_{ridge} = \arg\min_w \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2 + \lambda ||w||_2^2$$

Use *k*-fold cross validation

> Randomly divide training data into *k* equal parts
  - $D_1, \ldots, D_k$

$\mathscr{D} = \quad \mathscr{D}_1 \quad \mathscr{D}_2 \quad \mathscr{D}_3 \quad \mathscr{D}_4 \quad \mathscr{D}_5$

$f_{\mathscr{D} \setminus \mathscr{D}_3}$	Train	Train	Validation	Train	Train

> For each *i*
  - Learn model $f_{\mathscr{D} \setminus \mathscr{D}_i}$ using data point not in $\mathscr{D}_i$
  - Estimate error of $f_{\mathscr{D} \setminus \mathscr{D}_i}$ on validation set $\mathscr{D}_i$:

$$\text{error}_{\mathcal{D}_i} = \frac{1}{|\mathcal{D}_i|} \sum_{(x_j, y_j) \in \mathcal{D}_i} (y_j - f_{\mathcal{D} \setminus \mathcal{D}_i}(x_j))^2$$

> k-fold cross validation error is average over data splits:

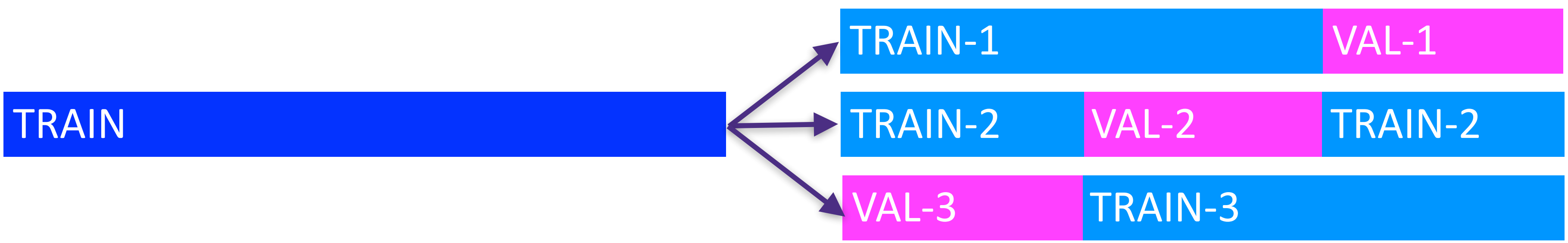
$$\text{error}_{k-\text{fold}} = \frac{1}{k} \sum_{i=1}^{k} \text{error}_{\mathscr{D}_i}$$

> k-fold cross validation properties:
  - Much faster to compute than LOO as $k \ll n$
  - More (pessimistically) biased – using much less data, only $n - \frac{n}{k}$
  - Usually, k = 10

Recap

> Given a dataset, begin by splitting into

> Model selection: Use k-fold cross-validation on TRAIN to train predictor and choose hyper-parameters such as λ



> Model assessment: Use TEST to assess the accuracy of the model you output

- Never ever ever ever ever train or choose parameters based on the test data