Lectures 9

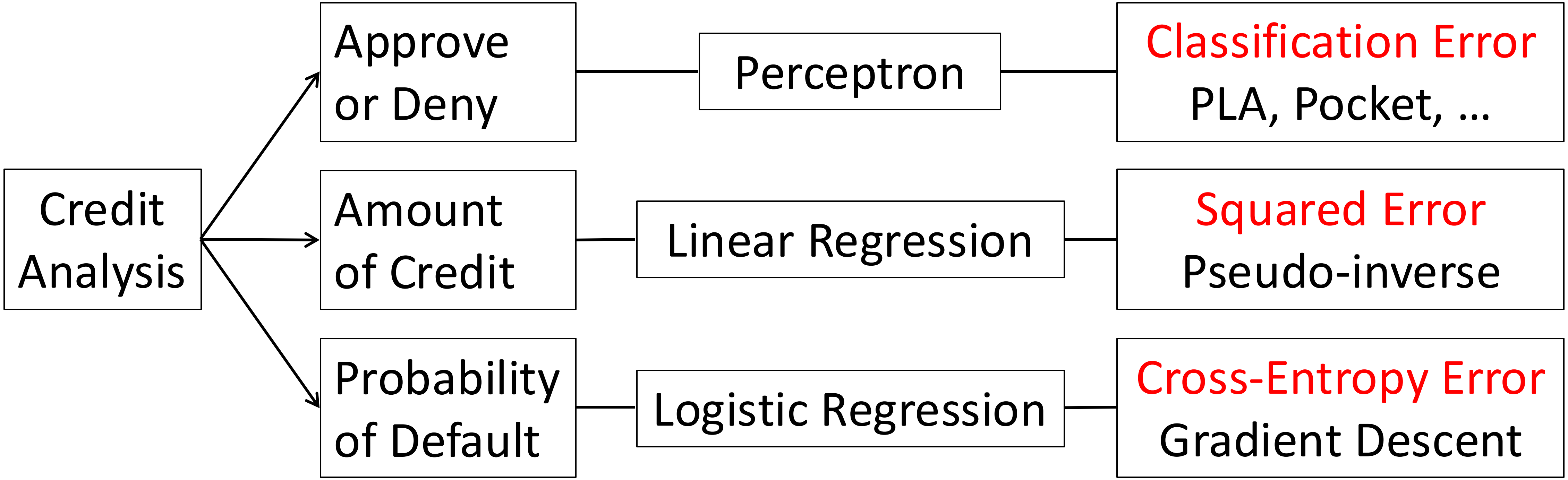
CS436/536: Introduction to Machine Learning

Zhaohan Xi

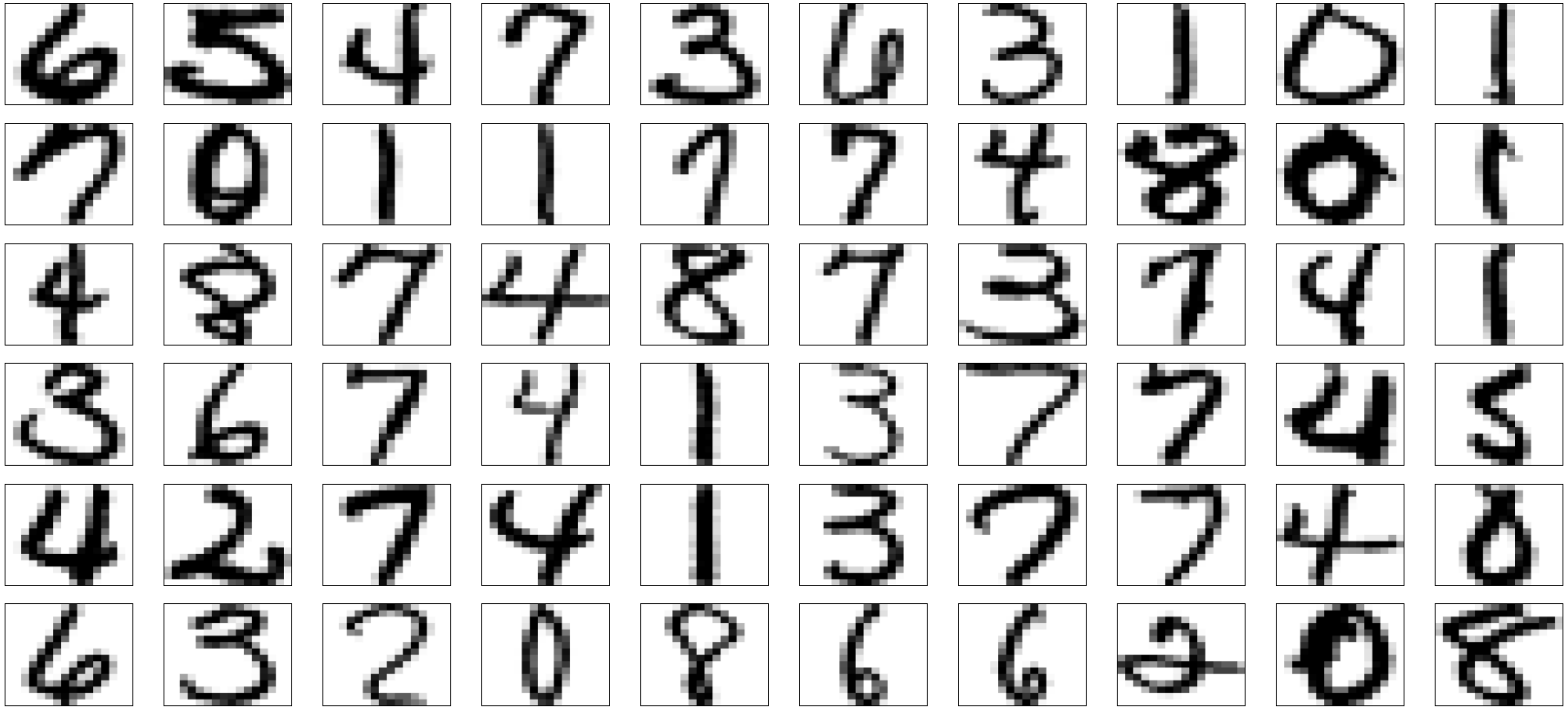
Binghamton University

zxi1@binghamton.edu

Recap: Linear Model for Three Learning Problems



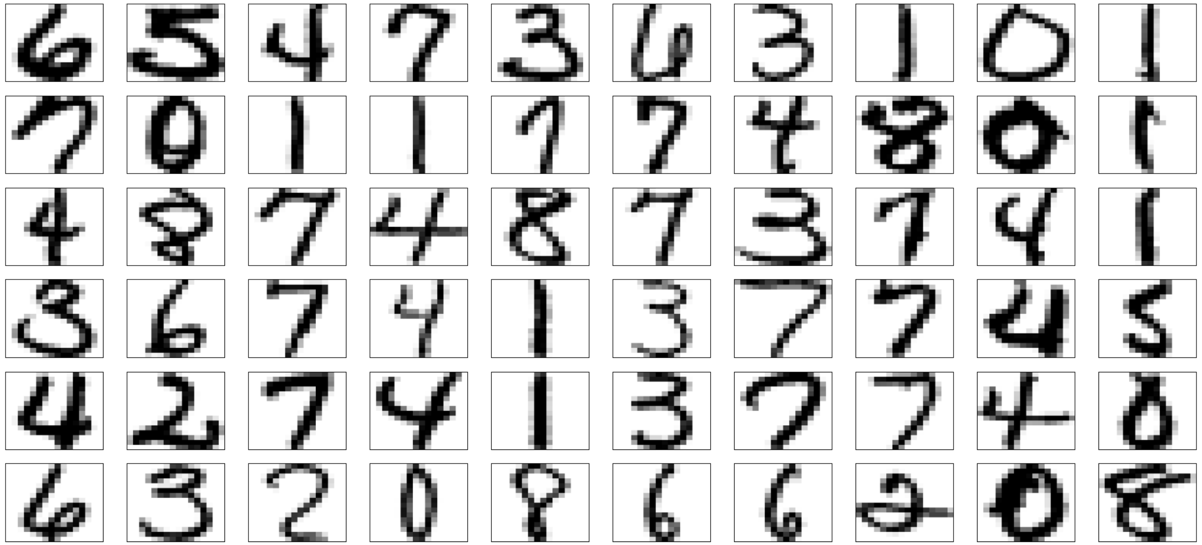
Digits Data

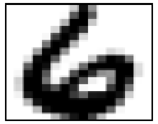


Each example of a digit is a 16×16 image

http://yann.lecun.com/exdb/mnist/

Digits Data





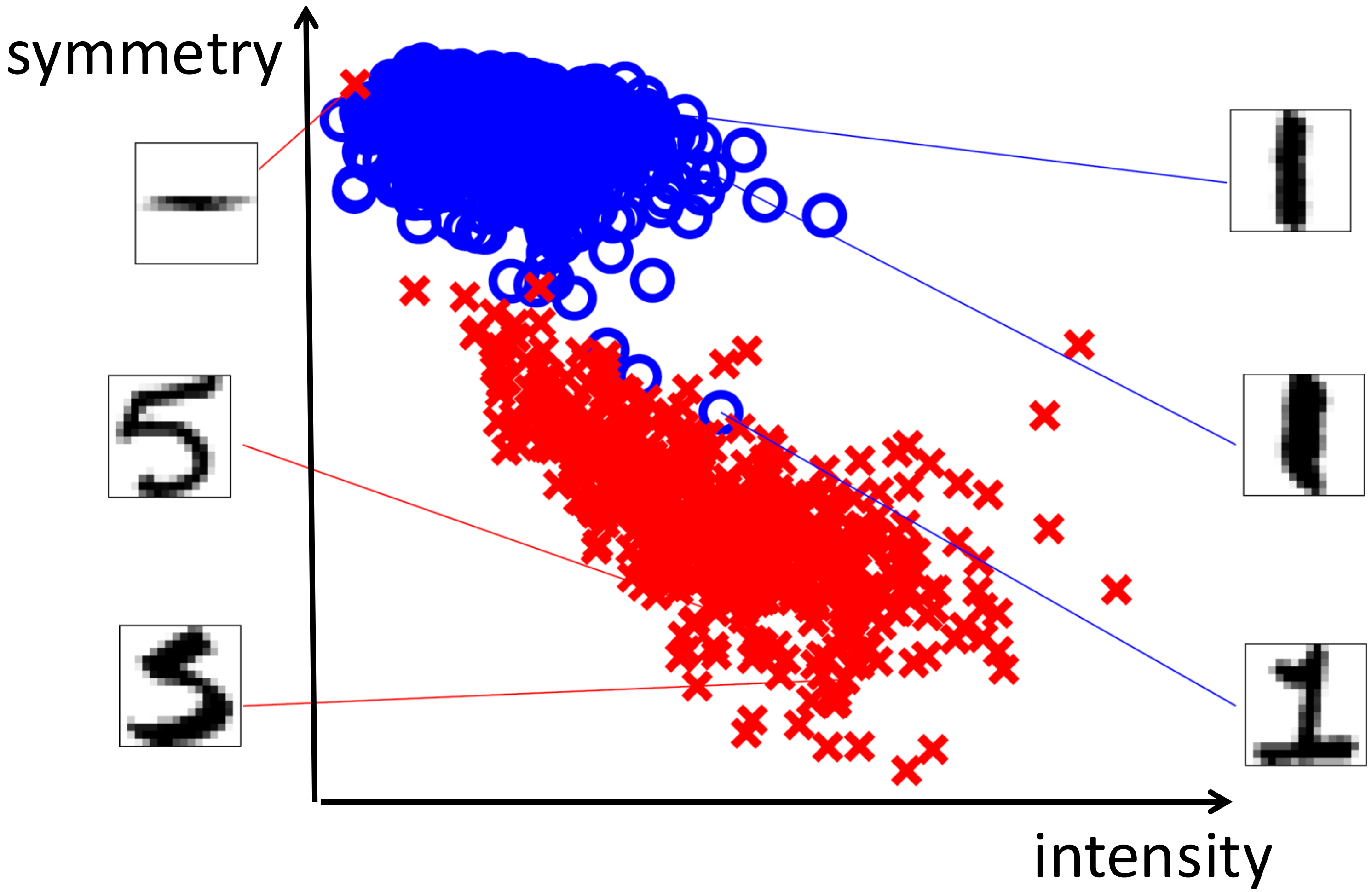
[-1 -1 -1 -1 -1 -1 -1 -0.63 0.86 -0.17 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -0.99 0.3 1 0.31 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -0.41 1 0.99 -0.57 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -0.68 0.83 1 0.56 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -0.94 0.54 1 0.78 -0.72 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 0.1 1 0.92 -0.44 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -0.26 0.95 1 -0.16 -1 -1 -1 -0.99 -0.71 -0.83 -1 -1 -1 -1 -1 -0.8 0.91 1 0.3 -0.96 -1 -1 -0.55 0.49 1 0.88 0.09 -1 -1 -1 -1 0.28 1 0.88 -0.8 -1 -0.9 0.14 0.97 1 1 1 0.99 -0.74 -1 -1 -0.95 0.84 1 0.32 -1 -1 0.35 1 0.65 -0.10 -0.18 1 0.98 -0.72 -1 -1 -0.63 1 1 0.07 -0.92 0.11 0.96 0.30 -0.88 -1 -0.07 1 0.64 -0.99 -1 -1 -0.67 1 1 0.75 0.34 1 0.70 -0.94 -1 -1 0.54 1 0.02 -1 -1 -1 -0.90 0.79 1 1 1 1 0.53 0.18 0.81 0.83 0.97 0.86 -0.63 -1 -1 -1 -1 -0.45 0.82 1 1 1 1 1 1 1 0.13 -1 -1 -1 -1 -1 -1 -0.48 0.81 1 1 1 1 1 1 0.21 -0.94 -1 -1 -1 -1 -1 -1 -1 -0.97 -0.42 0.30 0.82 1 0.48 -0.47 -0.99 -1 -1 -1 -1]

$$\left.\begin{array}{l} x = (1, x_1, \dots, x_{256}) \\ w = (w_0, w_1, \dots, w_{256}) \end{array}\right\} d_{VC} = 257$$

Features: Intensity and Symmetry

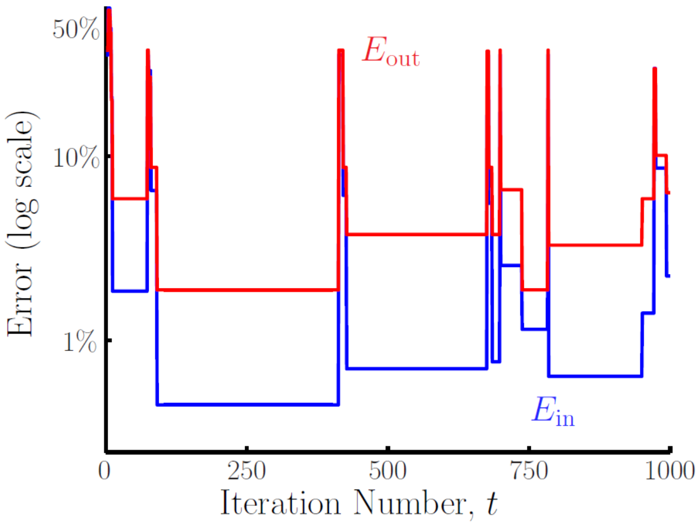
Human-like approach: Summarize image by a few features

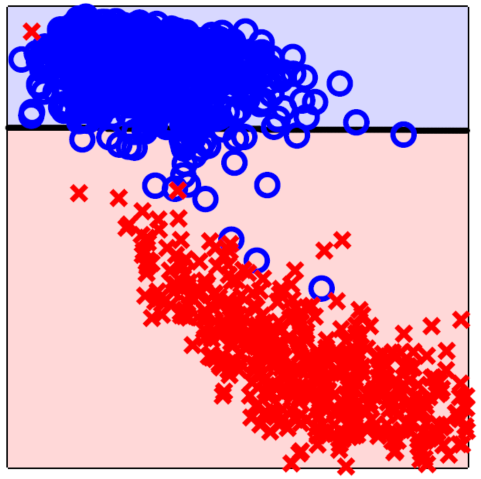
feature: an important property of the input you think is useful for classification



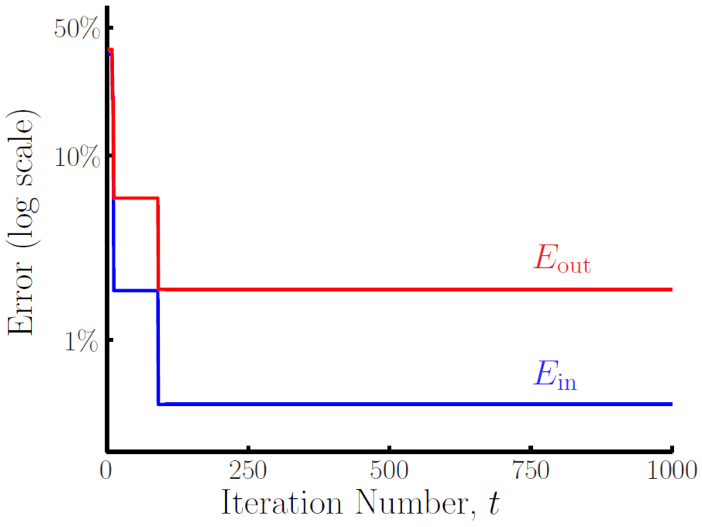
Perceptron Model on Digits Data

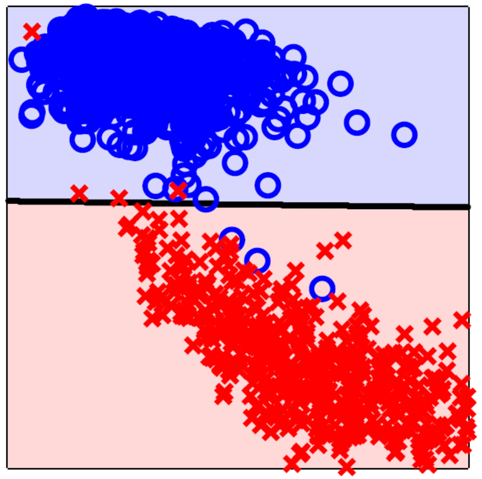
PLA



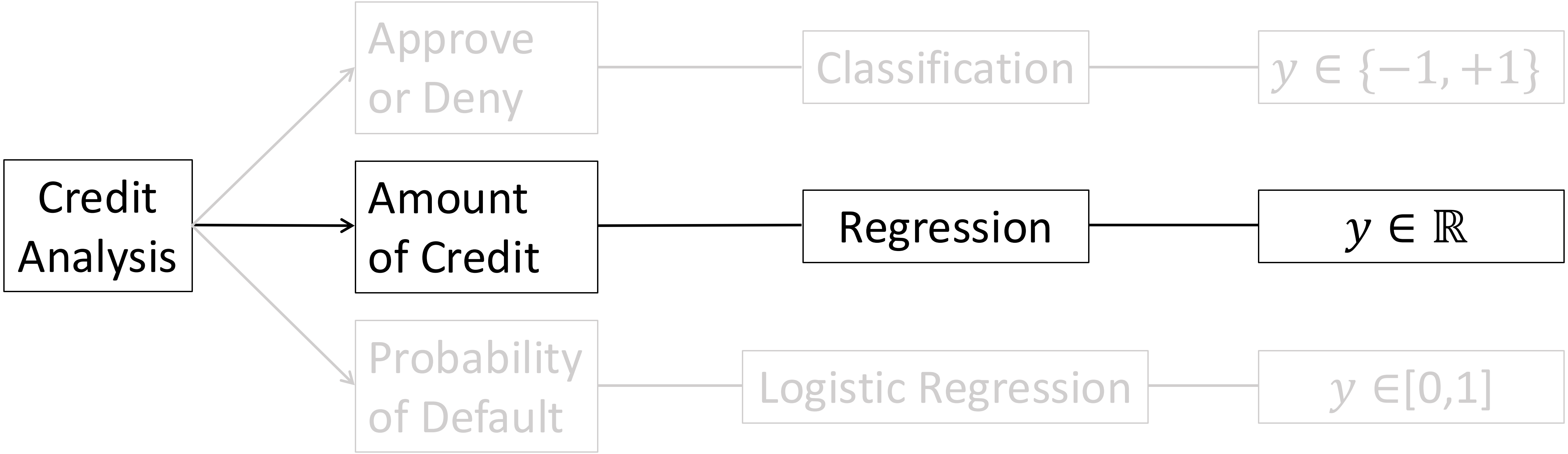


Pocket Algorithm

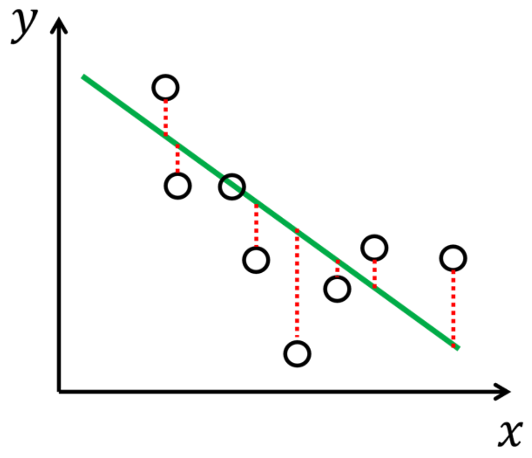


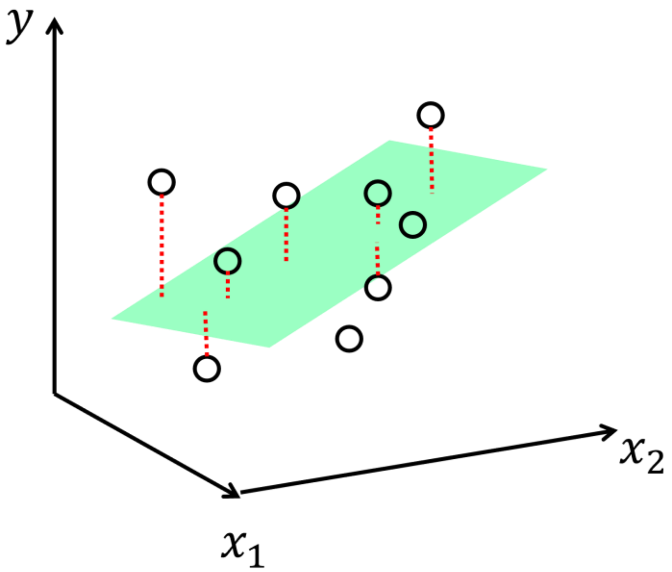


Recap: Linear Models for Credit Analysis



Least Squares Linear Regression

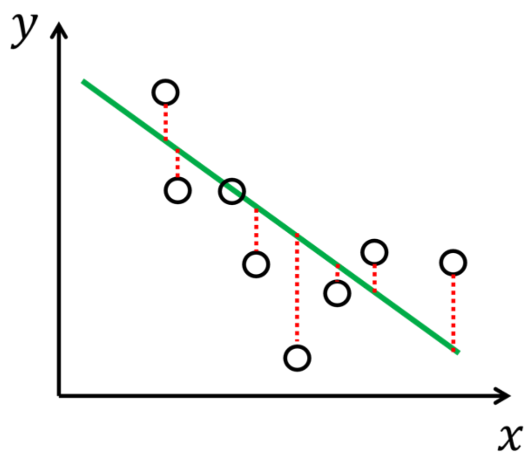


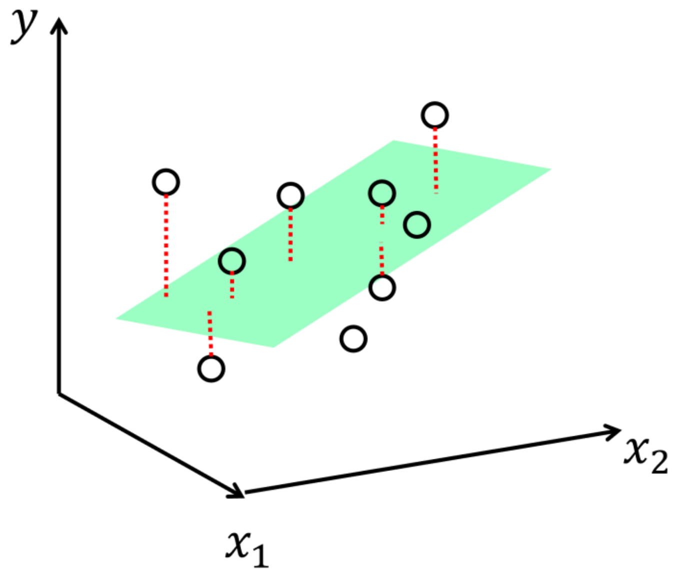


$$error(h(\boldsymbol{x}),\ f(\boldsymbol{x})) = (h(\boldsymbol{x}) - f(\boldsymbol{x}))^2$$

prediction actual

Least Squares Linear Regression





$y = f(\boldsymbol{x}) + \epsilon$ $\quad\quad\quad\quad$ $\sim P(y|\boldsymbol{x})$

$$E_{in}(h) = \frac{1}{N}\sum_{n=1}^{N}\left(h(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2$$

$$E_{out}(h) = \mathbb{E}_{\boldsymbol{x}}\left[\left(h(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2\right]$$

Recap: Ordinary Least Squares: Minimizing E_{in}

$$E_{in}(\boldsymbol{w}) = \frac{1}{N}(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y})$$

Differentiable

Let $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X}$ with dimensions $(d+1) \times (d+1)$

Let $\boldsymbol{b}$ = $\boldsymbol{X}^T\boldsymbol{y}$ with dimensions $(d+1) \times 1$

Let $c = \boldsymbol{y}^T\boldsymbol{y}$

$$E_{in}(\boldsymbol{w}) = \frac{1}{N}(\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{b} + c)$$

$$\nabla_{\boldsymbol{w}} E_{in}(\boldsymbol{w}) = \frac{1}{N}[(\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{w} - 2\boldsymbol{b}]$$

Useful gradient identities:

- $\nabla_z(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}) = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{z}$
- $\nabla_z(\boldsymbol{z}^T\boldsymbol{b}) = \boldsymbol{b}$

Recap: Ordinary Least Squares: Minimizing E_{in}

$$E_{in}(\boldsymbol{w}) = \frac{1}{N}(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y})$$

Let $\boldsymbol{A} = X^TX$, $\boldsymbol{b} = X^Ty$, $c = y^Ty$

$$\nabla_{\boldsymbol{w}}E_{in}(\boldsymbol{w}) = \frac{1}{N}[(\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{w} - 2\boldsymbol{b}]$$

$$(\boldsymbol{X}^T\boldsymbol{X})^T = (\boldsymbol{X})^T(\boldsymbol{X}^T)^T = \boldsymbol{X}^T\boldsymbol{X}$$

$$\nabla_{\boldsymbol{w}}E_{in}(\boldsymbol{w}) = \frac{2}{N}[\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{X}^T\boldsymbol{y}]$$

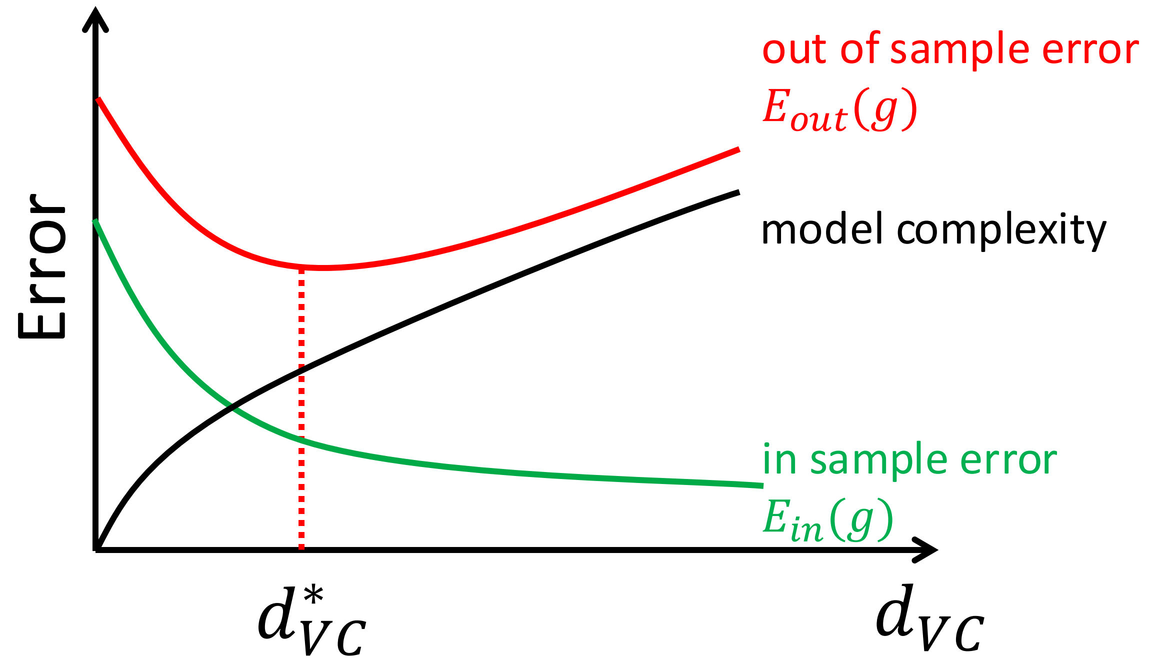
Solve for $\boldsymbol{w}$:

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$$

$$\boldsymbol{w}_{lin} = \boldsymbol{X}^{\dagger}\boldsymbol{y}$$

where $\boldsymbol{X}^{\dagger} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$

Flashback: VC Analysis, Approximation vs. Generalization



$d_{VC} \uparrow \Rightarrow E_{in} \approx 0$: Higher chance of approximating target function on data

$d_{VC} \downarrow \Rightarrow E_{in} \approx E_{out}$: Higher chance of generalizing to out of data

Depends only on $\mathcal{H}$; Independent of $f, P(\boldsymbol{x}), \mathcal{A}$ (the learning algorithm)

Bias – Variance Analysis

- An alternate view of the approximation-generalization tradeoff

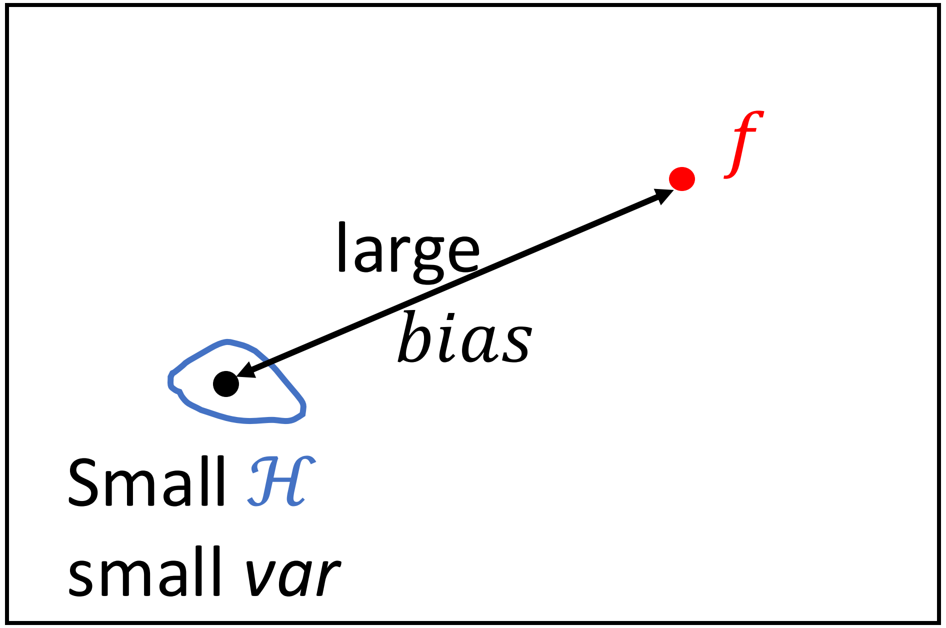
for squared error measures (e.g. regression)

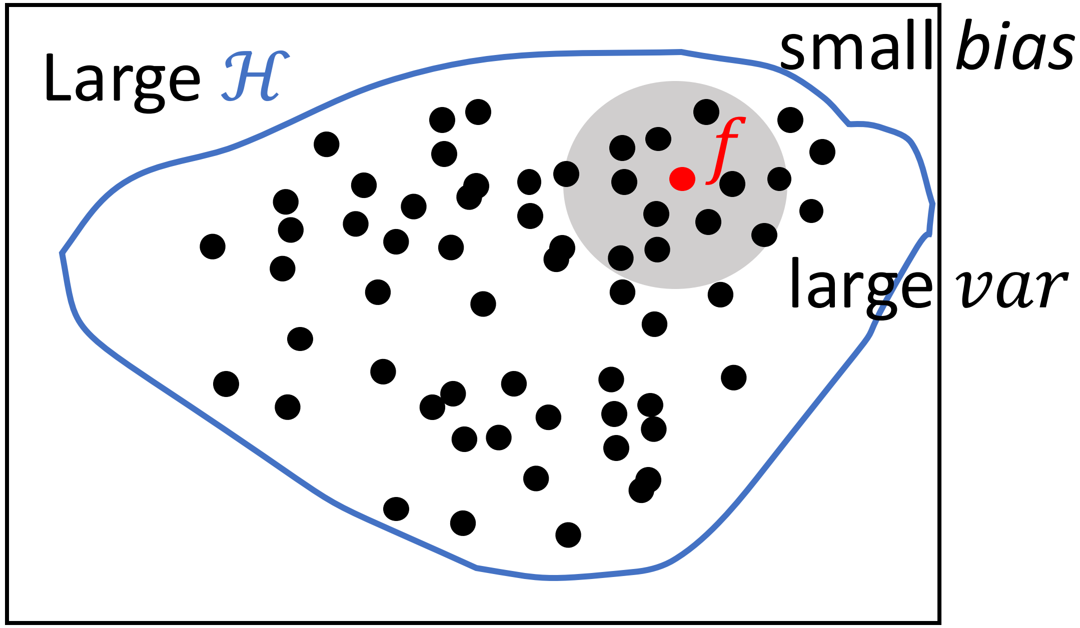
1. How well ***can*** a hypothesis in $\mathcal{H}$ approximate f?

2. How close can we get to this using a finite dataset?

Bias-Variance Tradeoff

bias: How well can $\mathcal{H}$ fit target f?
var: How often can we find a good approximation?



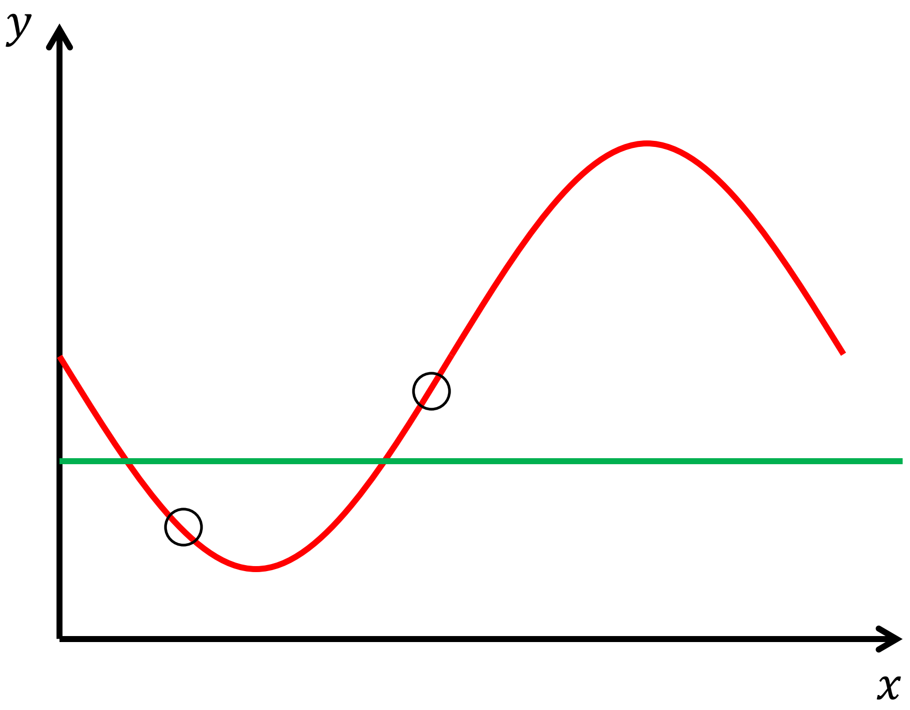


A Simple Learning Problem $f(x) = \sin \pi x$

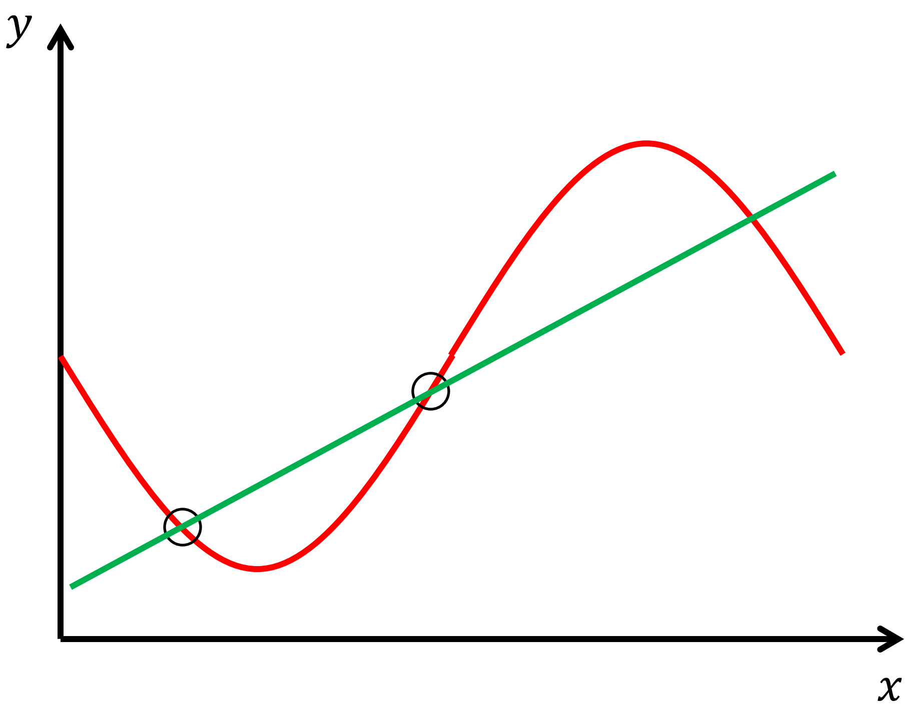
2 Data points

2 Hypothesis sets:

Flat lines $\mathcal{H}_0$: $h(\boldsymbol{x}) = b$



All lines $\mathcal{H}_1$: $h(\boldsymbol{x}) = a\boldsymbol{x} + b$

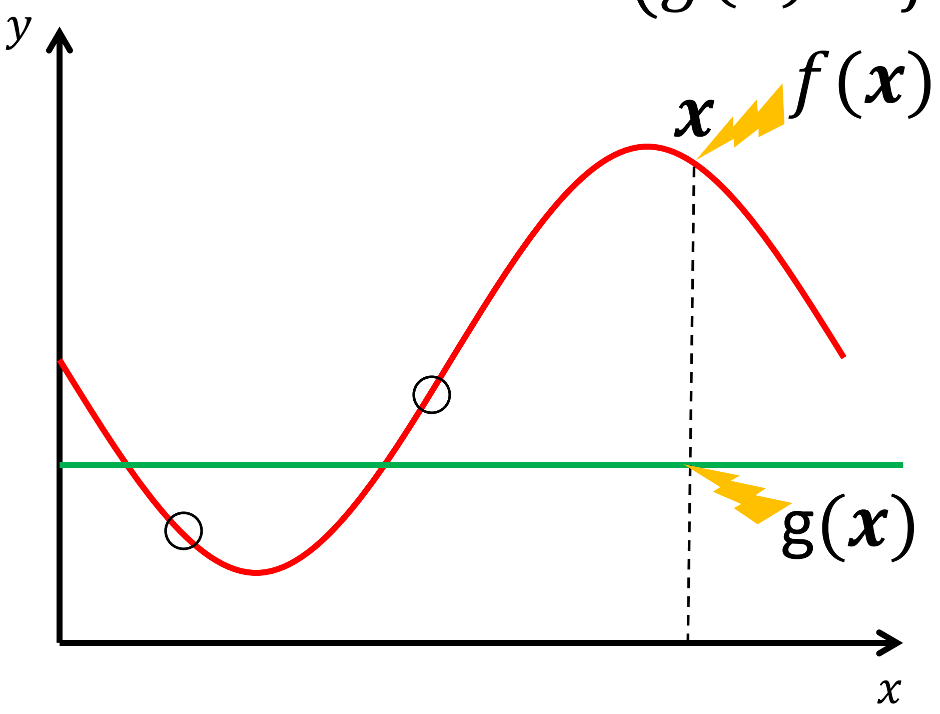


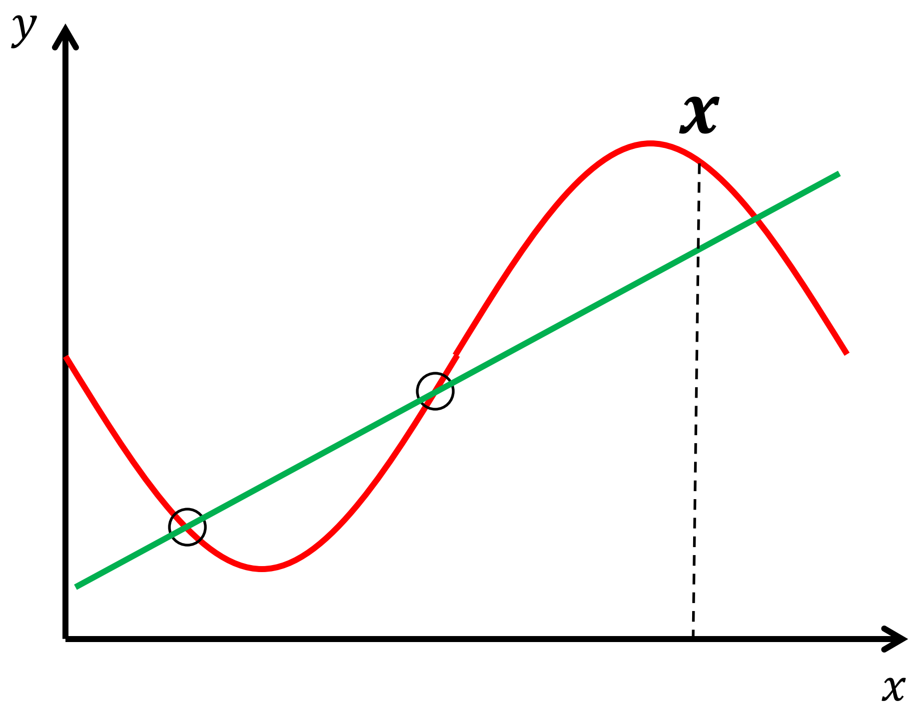
A Simple Learning Problem $f(x) = \sin \pi x$

Flat lines $\mathcal{H}_0$: $h(\boldsymbol{x}) = b$

All lines $\mathcal{H}_1$: $h(\boldsymbol{x}) = a\boldsymbol{x} + b$

$$error = \left(g(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2$$



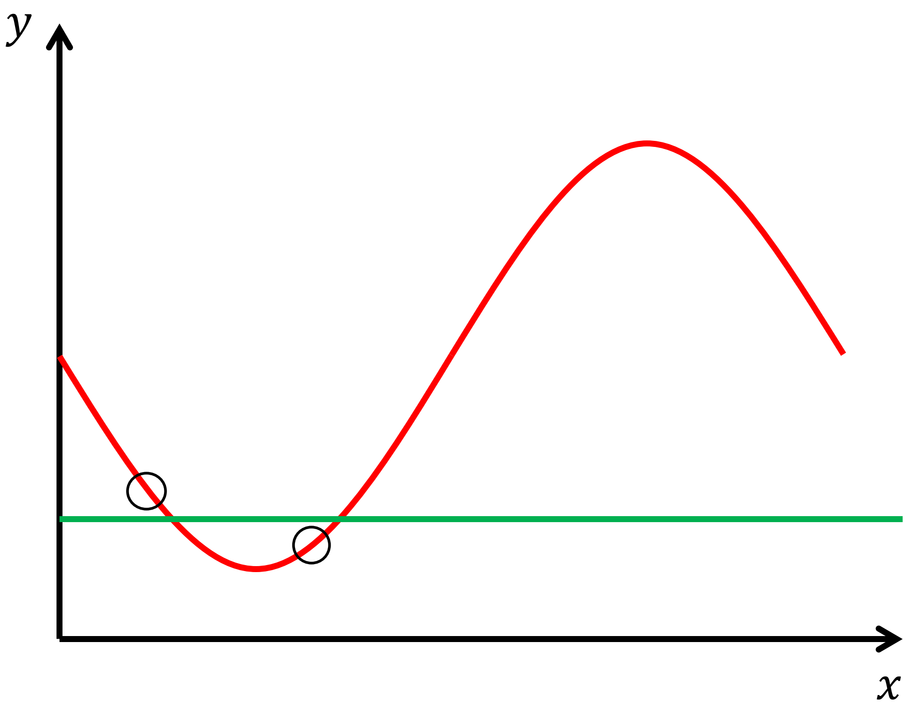


A Simple Learning Problem $f(x) = \sin \pi x$

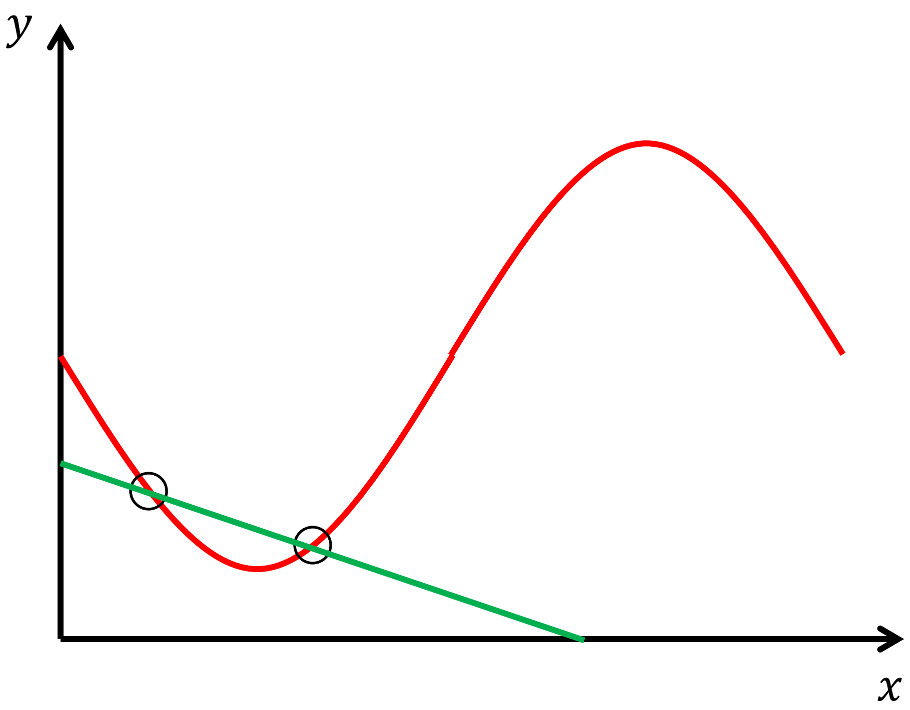
2 Data points

2 Hypothesis sets:

Flat lines $\mathcal{H}_0$: $h(\boldsymbol{x}) = b$



All lines $\mathcal{H}_1$: $h(\boldsymbol{x}) = a\boldsymbol{x} + b$

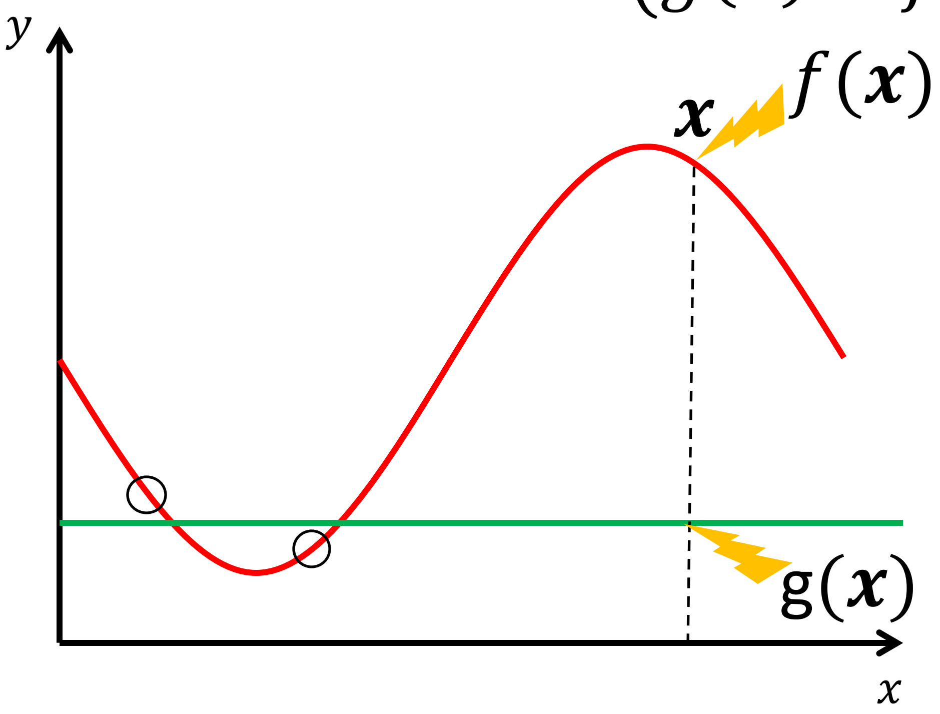


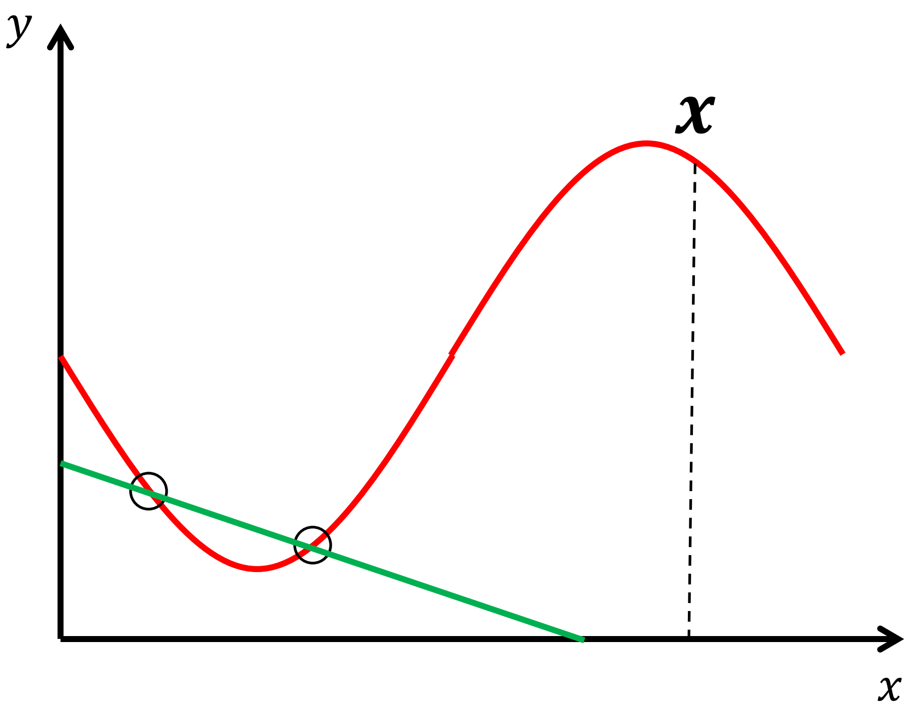
A Simple Learning Problem $f(x) = \sin \pi x$

Flat lines $\mathcal{H}_0$: $h(\boldsymbol{x}) = b$

All lines $\mathcal{H}_1$: $h(\boldsymbol{x}) = a\boldsymbol{x} + b$

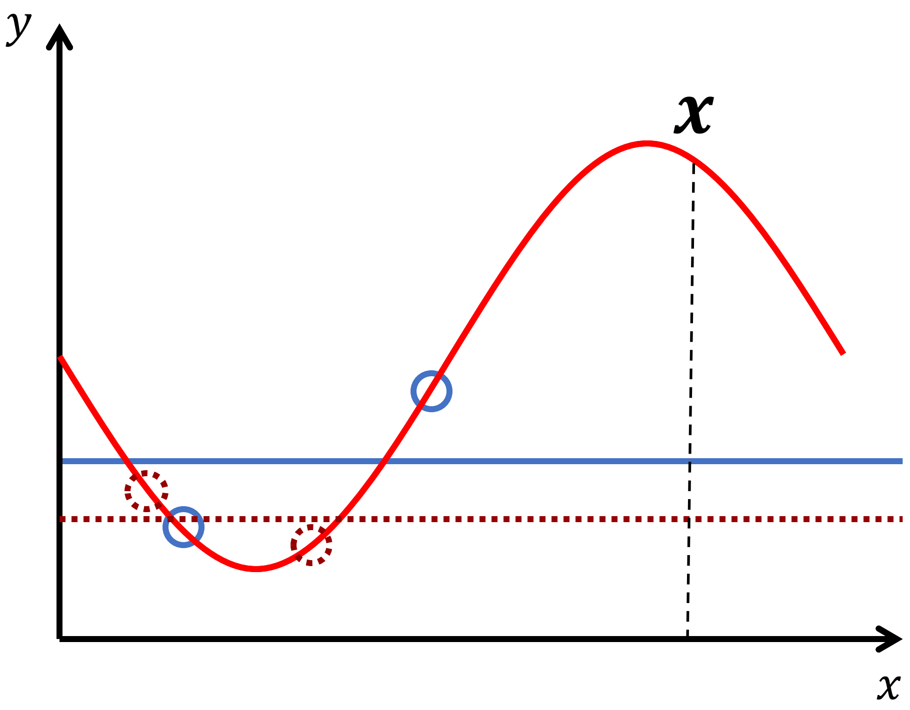
$$error = \big(g(\boldsymbol{x}) - f(\boldsymbol{x})\big)^2$$

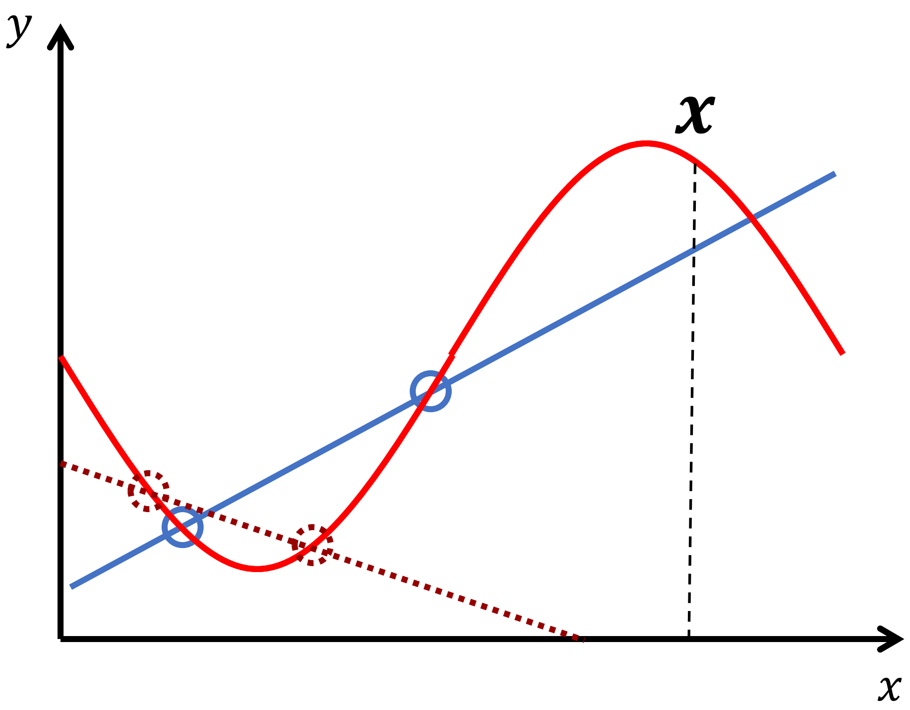




Different Dataset $\mathcal{D}$, Different Output $g^{\mathcal{D}}$

- For a fixed test point $\boldsymbol{x}$,
  $g^{\mathcal{D}}(\boldsymbol{x})$ is a random value that depends on $\mathcal{D}$

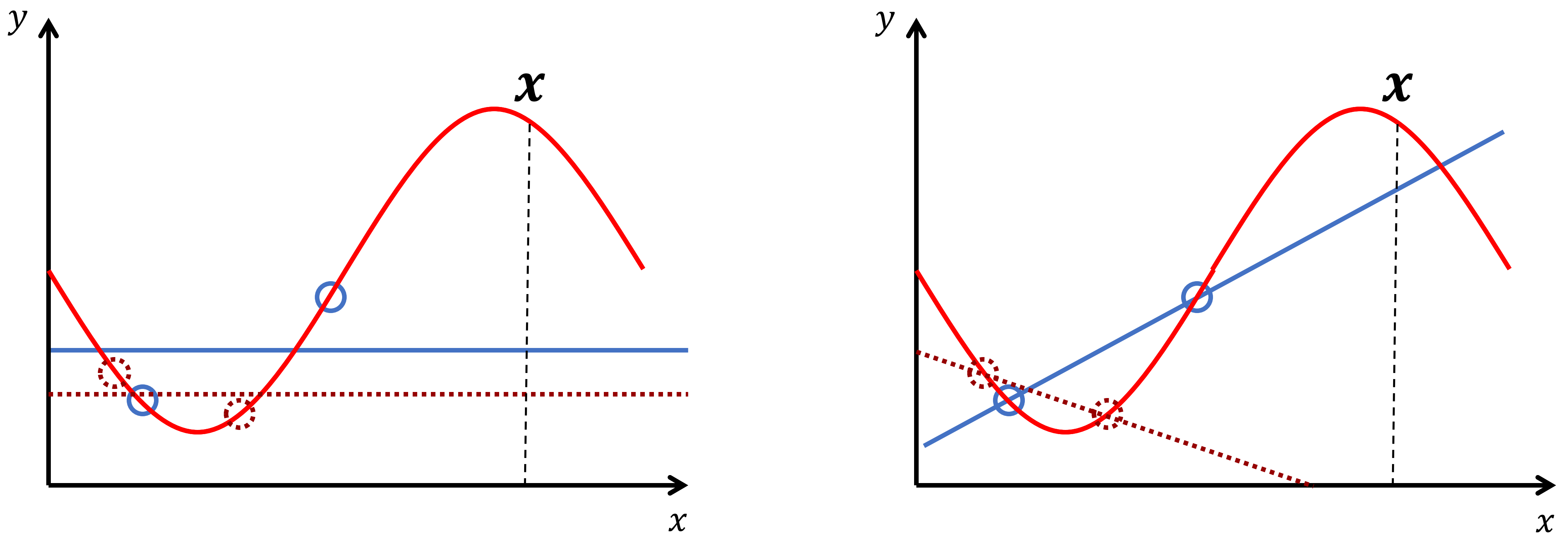




We can’t pick our data set!

We must analyze the entire process:

- Sample a data set
- Fit it (pick g from $\mathcal{H}$ using $\mathcal{A}$)
- Measure E_{in}



Expected Behavior w.r.t. Data

Dataset		Fit		Test
$\mathcal{D}_1$	$\rightarrow$	$g^{\mathcal{D}_1}$	$\rightarrow$	$g^{\mathcal{D}_1}(\boldsymbol{x})$
$\mathcal{D}_2$	$\rightarrow$	$g^{\mathcal{D}_2}$	$\rightarrow$	$g^{\mathcal{D}_2}(\boldsymbol{x})$
…				
$\mathcal{D}_K$	$\rightarrow$	$g^{\mathcal{D}_K}$	$\rightarrow$	$g^{\mathcal{D}_K}(\boldsymbol{x})$

Average (expected) prediction:

$$\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\boldsymbol{x})] = \bar{g}(\boldsymbol{x}) \approx \frac{1}{K}\left(g^{\mathcal{D}_1}(\boldsymbol{x}) + g^{\mathcal{D}_2}(\boldsymbol{x}) + \cdots + g^{\mathcal{D}_K}(\boldsymbol{x})\right)$$

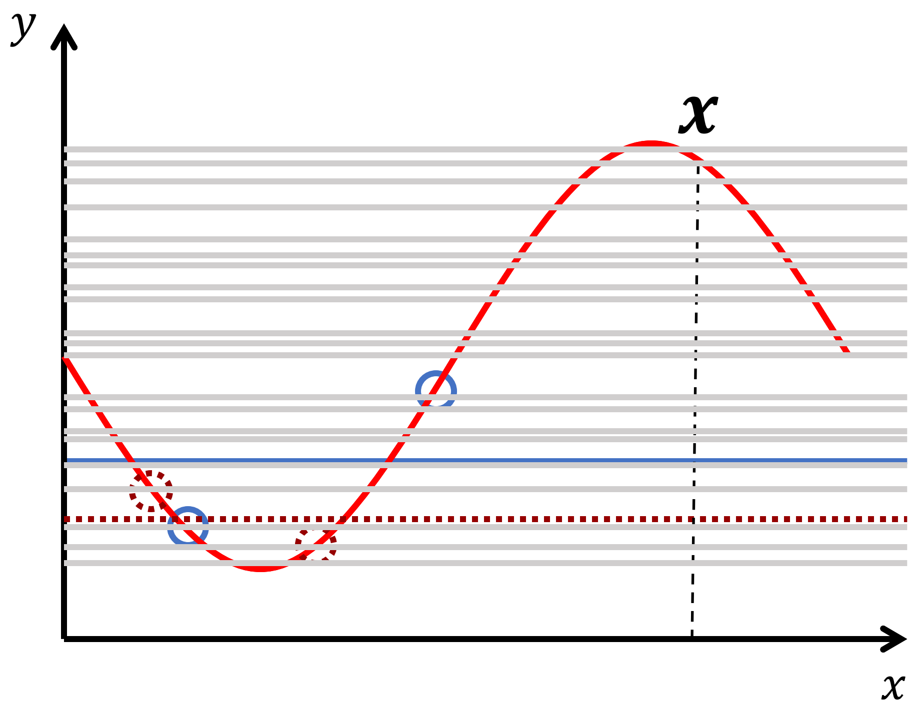
Variance of prediction: $var(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(x) - \bar{g}(x)\right)^2\right]$

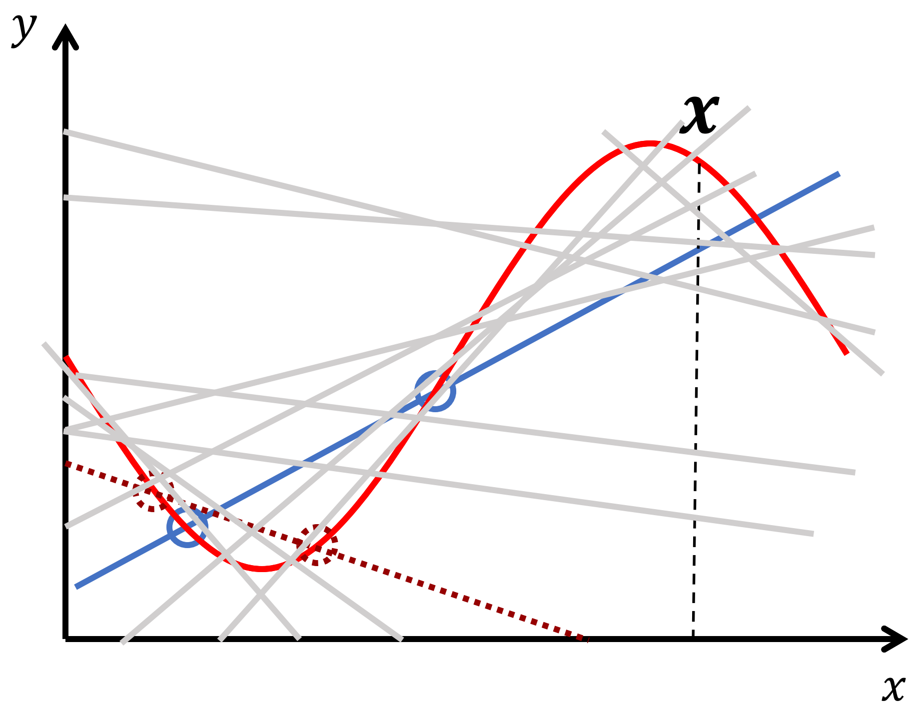
Different Dataset $\mathcal{D}$, Different Output $g^{\mathcal{D}}$

$g^{\mathcal{D}}(\boldsymbol{x})$ is a random value depending on $\mathcal{D}$ (randomly generated data)

$\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\boldsymbol{x})] = \bar{g}(\boldsymbol{x}) \approx \frac{1}{K}\left(g^{\mathcal{D}_1}(\boldsymbol{x}) + g^{\mathcal{D}_2}(\boldsymbol{x}) + \cdots + g^{\mathcal{D}_K}(\boldsymbol{x})\right)$, average prediction

$var(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(x) - \bar{g}(x)\right)^2\right]$, variance of prediction

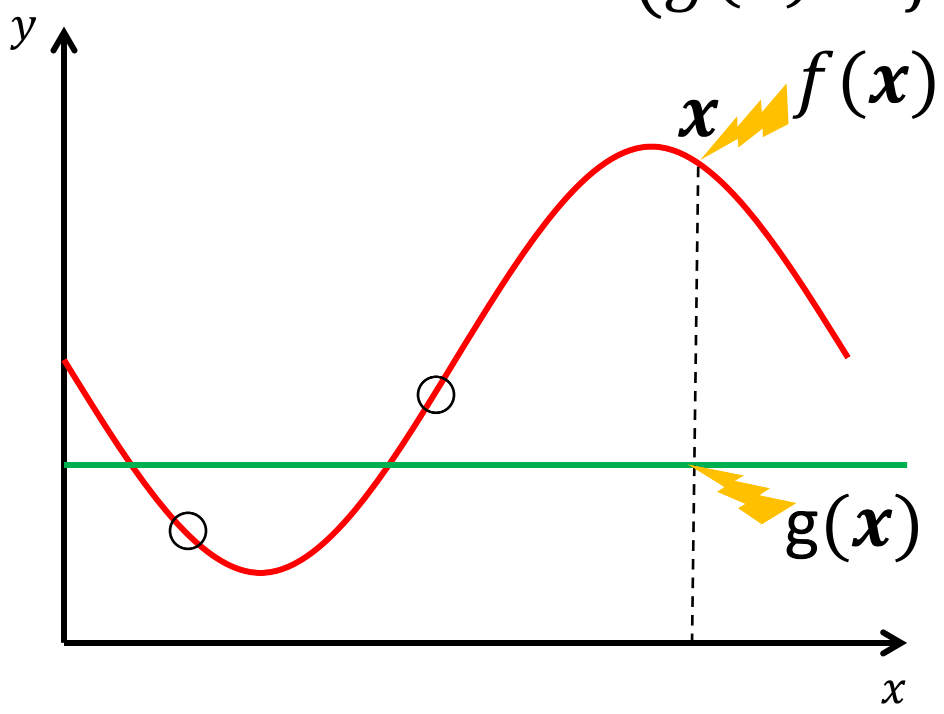


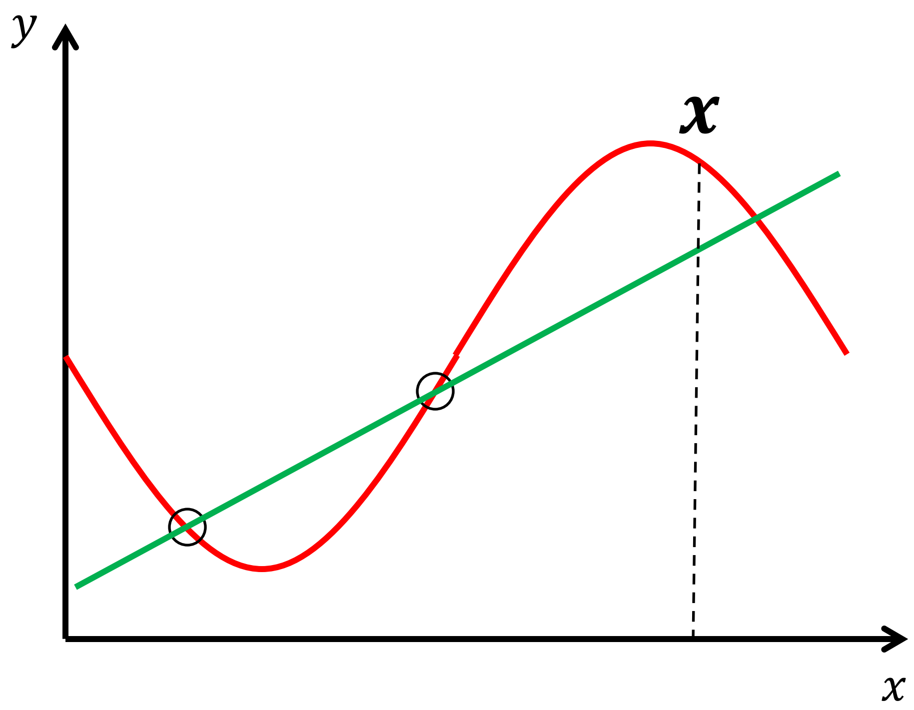


Out of Sample Error: E_{out} on test point $\boldsymbol{x}$ for data $\mathcal{D}$

- $E_{out}^{\mathcal{D}}(\boldsymbol{x}) = \left(g^{\mathcal{D}}(\boldsymbol{x}) - f(x)\right)^2$, squared error depending on random $\mathcal{D}$
- E_{out} before seeing the data: $E_{out}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}\left[E_{out}^{\mathcal{D}}(\boldsymbol{x})\right]$

$$error = \left(g(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2$$





Bias-Variance Decomposition: Expected Error on Test Point

$$E_{out}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}\left[E_{out}^{\mathcal{D}}(\boldsymbol{x})\right] = \mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2\right]$$

$$= \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\boldsymbol{x})^2 - 2g^{\mathcal{D}}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2]$$

$[\mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\boldsymbol{x})] = \bar{g}(\boldsymbol{x})]$

$$= \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\boldsymbol{x})^2] - 2\bar{g}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2$$

$$= \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\boldsymbol{x})^2] + \bar{g}(\boldsymbol{x})^2 - \bar{g}(\boldsymbol{x})^2 - 2\bar{g}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2$$

$$= \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(\boldsymbol{x})^2] - \bar{g}(\boldsymbol{x})^2 \quad + \quad \left(\bar{g}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2$$

$\mathbb{E}_{\mathcal{D}}[-2g^{\mathcal{D}}(\boldsymbol{x})\bar{g}(\boldsymbol{x}) + \bar{g}(\boldsymbol{x})^2] = -\bar{g}(\boldsymbol{x})^2$

$$= \mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(\boldsymbol{x}) - \bar{g}(\boldsymbol{x})\right)^2\right] \quad + \quad \left(\bar{g}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2$$

$$= \quad var(\boldsymbol{x}) \quad + \quad bias(\boldsymbol{x})$$

The Bias-Variance Decomposition

- Fact: For any random variable X, (and also for any $X = \hat{\theta} - \theta$),

$\mathbb{E}[X^2] = (\mathbb{E}[X])^2 + var(X)$

estimator unknown, fixed parameter

Theorem: $\mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2\right] = \left(\bar{g}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 + var(\boldsymbol{x})$

$= \quad bias(\boldsymbol{x}) \quad + var(\boldsymbol{x})$

$=$ expected out of sample error at test point $\boldsymbol{x}$

Here, $var(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(\boldsymbol{x}) - \bar{g}(\boldsymbol{x})\right)^2\right]$

E_{out}: Average over $\boldsymbol{x}$

$$E_{out} = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{\mathcal{D}}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2\right]\right] = \mathbb{E}_{\boldsymbol{x}}\left[\left(\bar{g}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 + var(\boldsymbol{x})\right]$$

$$= \mathbb{E}_{\boldsymbol{x}}[bias(\boldsymbol{x}) + var(\boldsymbol{x})]$$

$$= \quad bias \quad + \quad var$$

generalization

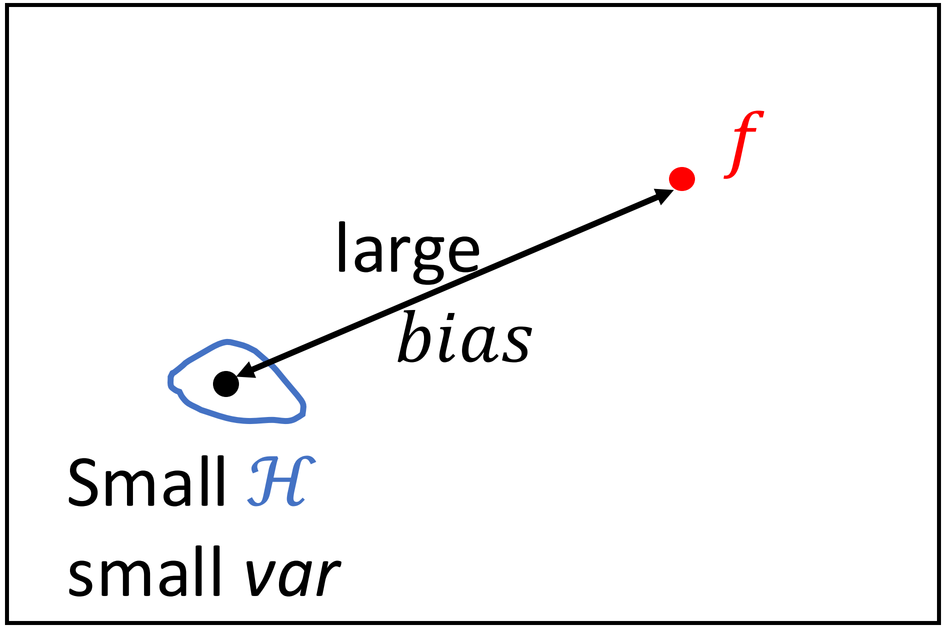
approximation

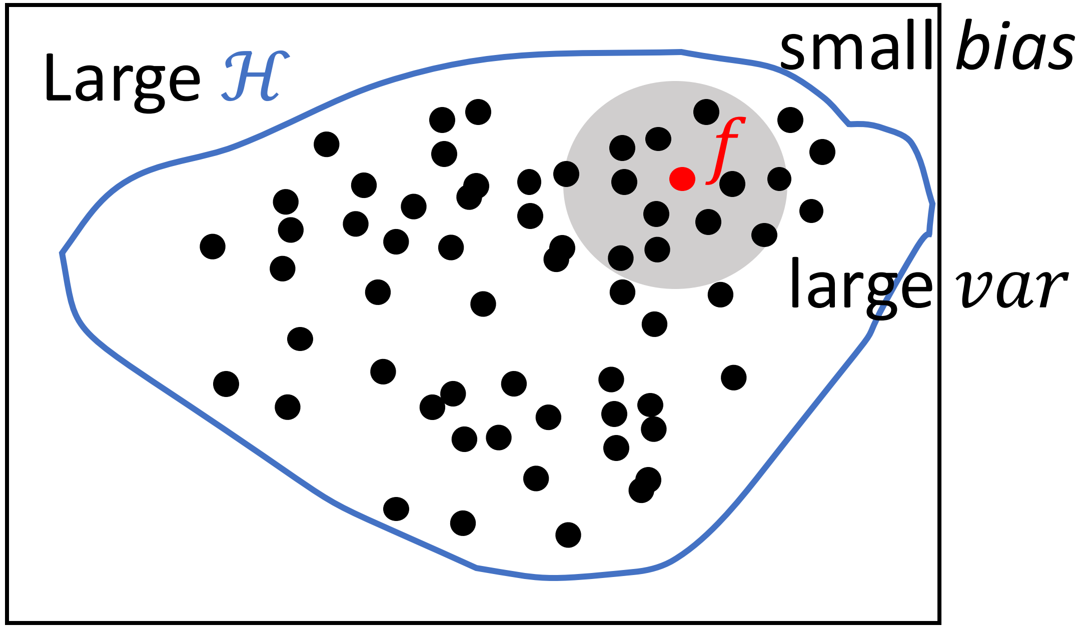
How close is average learned hypothesis to target function?

How often can we find a good hypothesis?

Bias-Variance Tradeoff

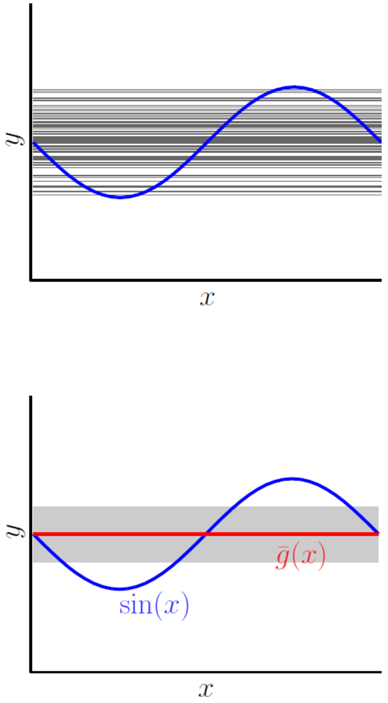
bias: How well can $\mathcal{H}$ fit target f?
var: How often can we find a good approximation?





Match Learning Power to Data, Not the Target f

2 Data points

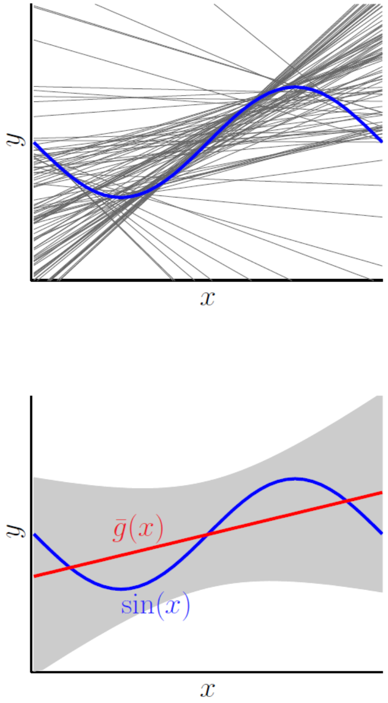


$\mathcal{H}_0$

$bias = 0.5$

$var = 0.25$

✓ $E_{out} = 0.75$



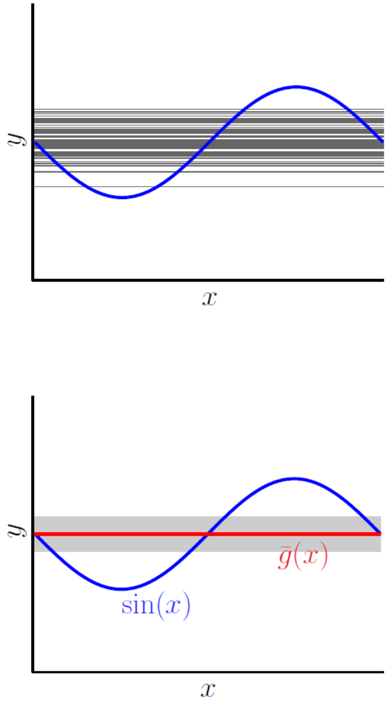
$\mathcal{H}_1$

$bias = 0.21$

$var = 1.69$

$E_{out} = 1.90$

5 Data points

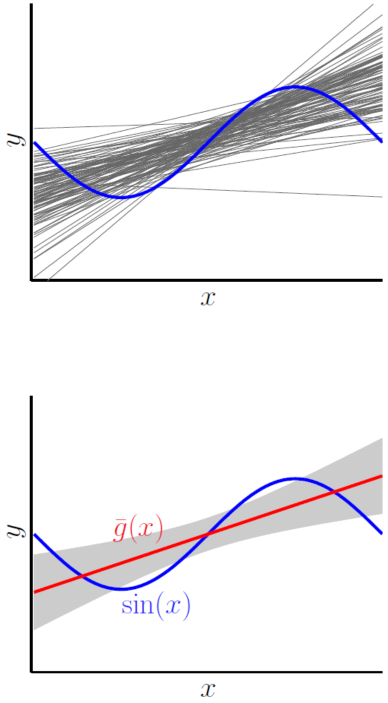


$\mathcal{H}_0$

$bias = 0.5$

$var = 0.1$

$E_{out} = 0.6$



$\mathcal{H}_1$

$bias = 0.21$

$var = 0.21$

✓ $E_{out} = 0.42$

Bias-Variance Analysis: A Useful Conceptual Tool

- Depends on f, $P(\boldsymbol{x})$ – both **unknown**
- Depends on $\mathcal{A}$

The objective of $\mathcal{A}$ is to minimize squared error

But for Bias-Variance analysis, we will use the squared error of g selected by $\mathcal{A}$

- Developing a model:
    - Lower variance without increasing bias
    - Lower bias without increasing variance

    Techniques discussed later in this course

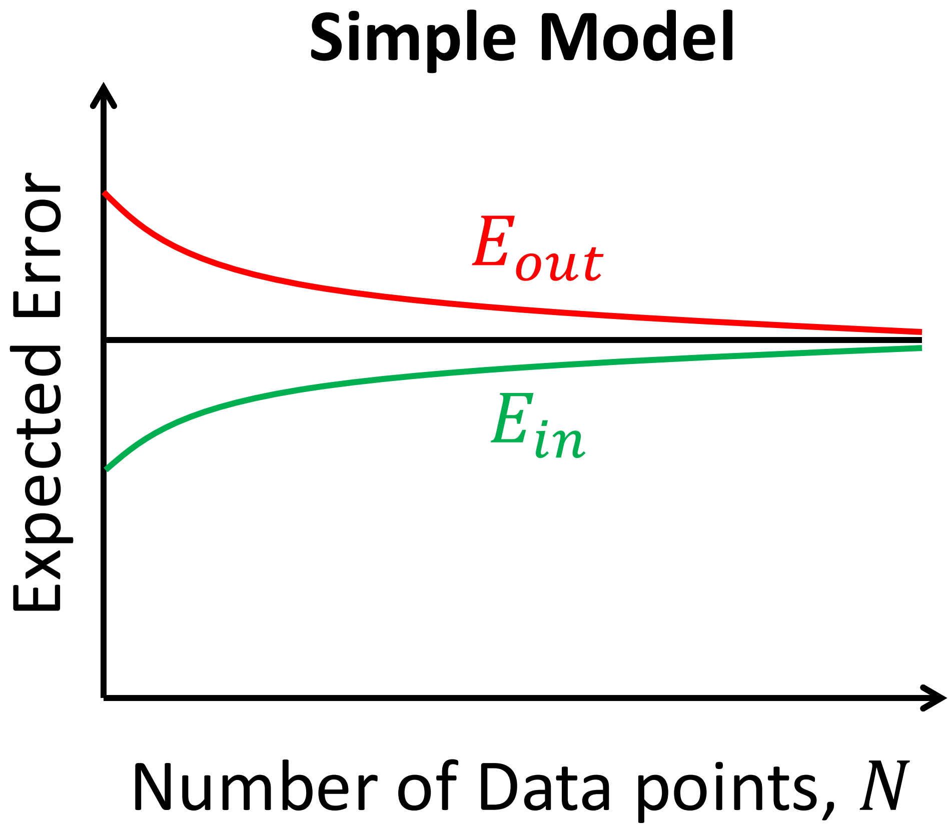
Summary: The Learning Process

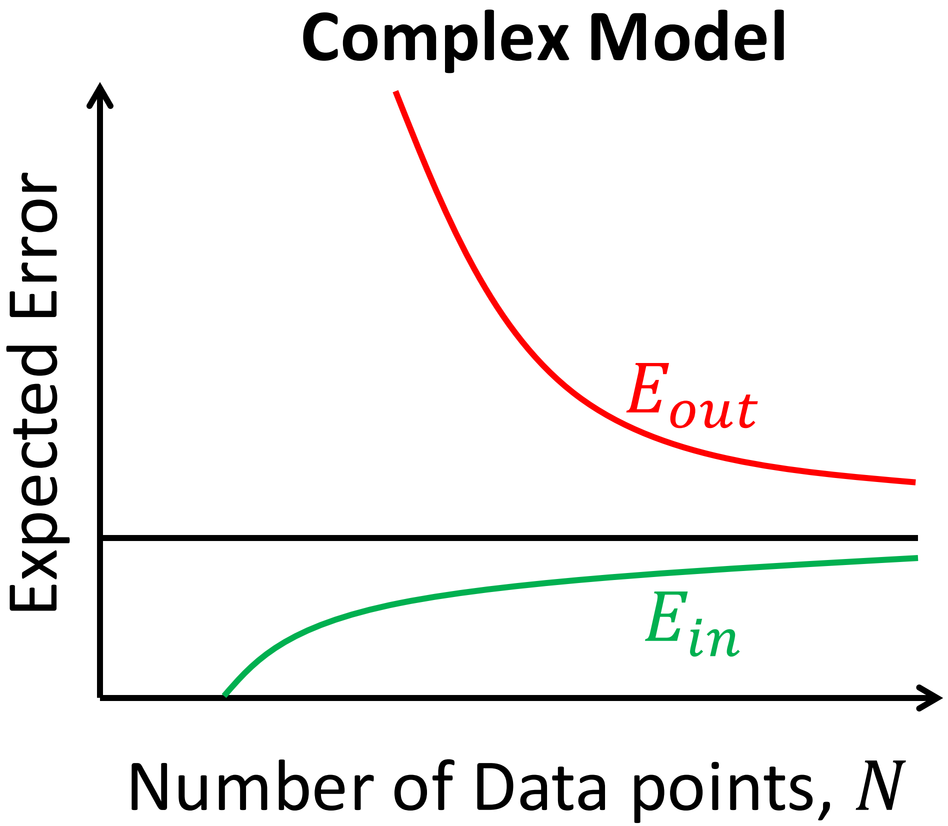
The Learning Process: For unknown $f(\boldsymbol{x})$ [or $P(y|\boldsymbol{x})$] and $P(\boldsymbol{x})$

- Fix $\mathcal{H}$
- Draw N data points $\mathcal{D}$ from $\boldsymbol{\mathcal{X}}$ i.i.d. at random according to $P(\boldsymbol{x})$
- Pick $g^{\mathcal{D}}$ from $\mathcal{H}$
  - Which has $E_{in}\left(g^{\mathcal{D}}\right)$ [measured on $\mathcal{D}$] and $E_{out}(g^{\mathcal{D}})$

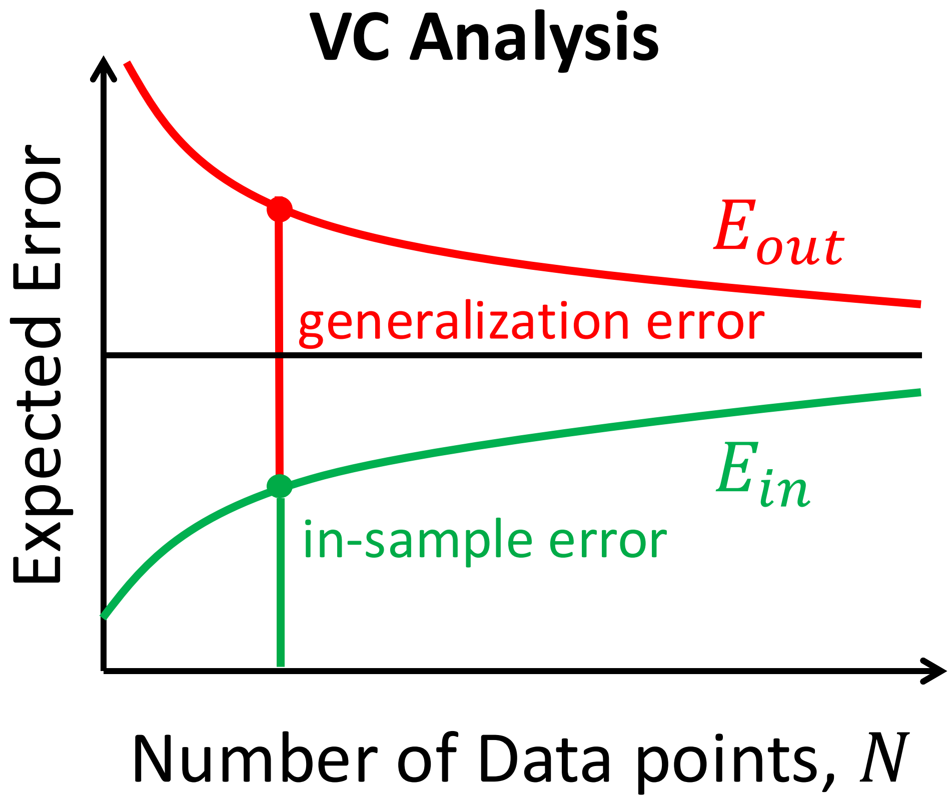
- Expected error of learning process: Expectation over all $\mathcal{D}$
  - $\mathbb{E}_{\mathcal{D}}[E_{in}(g^{\mathcal{D}})]$
  - $\mathbb{E}_{\mathcal{D}}[E_{out}(g^{\mathcal{D}})]$

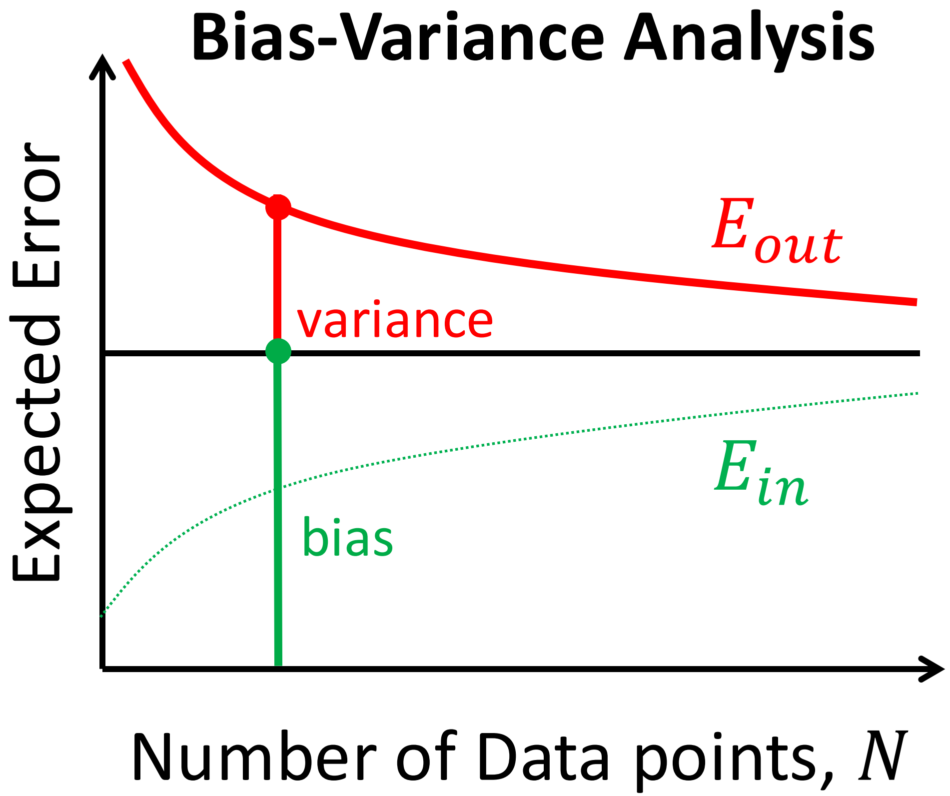
Summary: The Learning Curve





Decomposing the Learning Curve





Pick $\mathcal{H}$ to:

1. Generalize well, i.e., ensure $E_{out} \approx E_{in}$
2. Fit $\mathcal{D}$, i.e., get small E_{in}

Pick $\mathcal{H}, \mathcal{A}$ to:

1. Approximate f
2. Not vary too wildly with $\mathcal{D}$