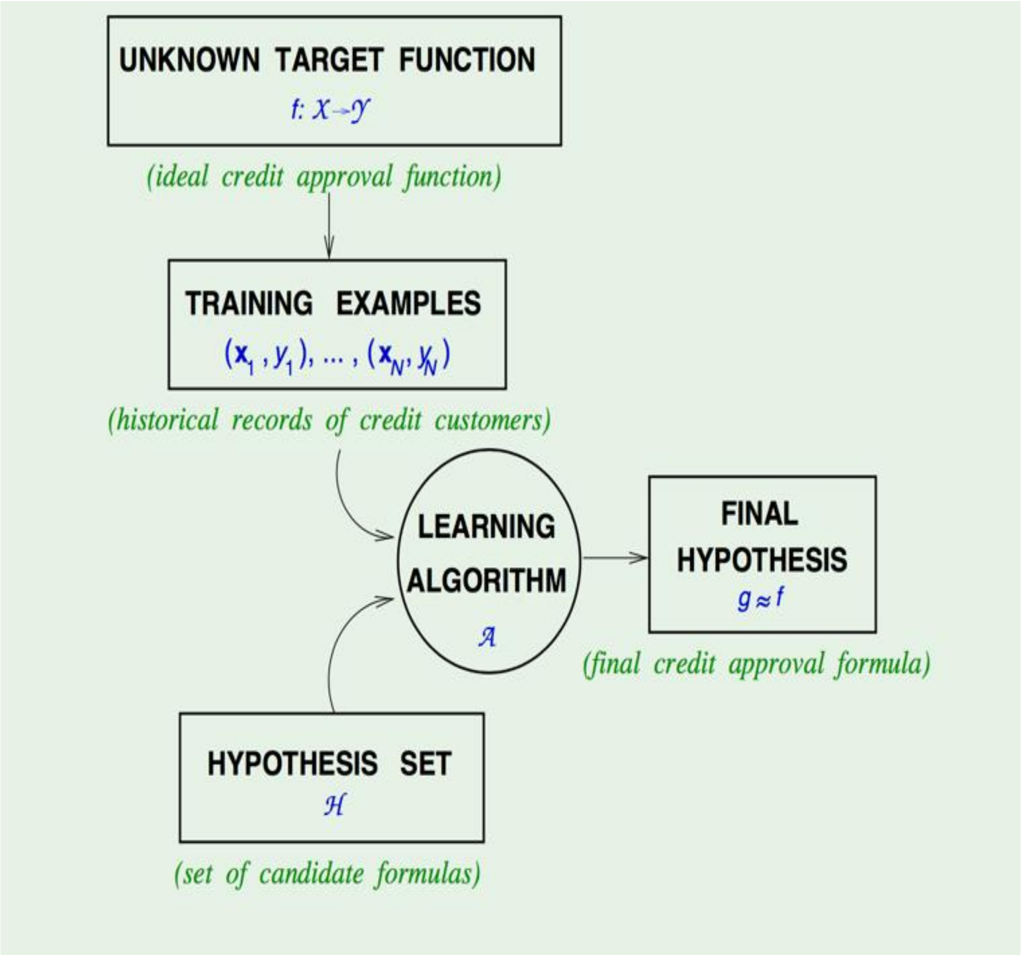
BIAS-VARIANCE TRADEOFF

Sonpvh

OUTLIER

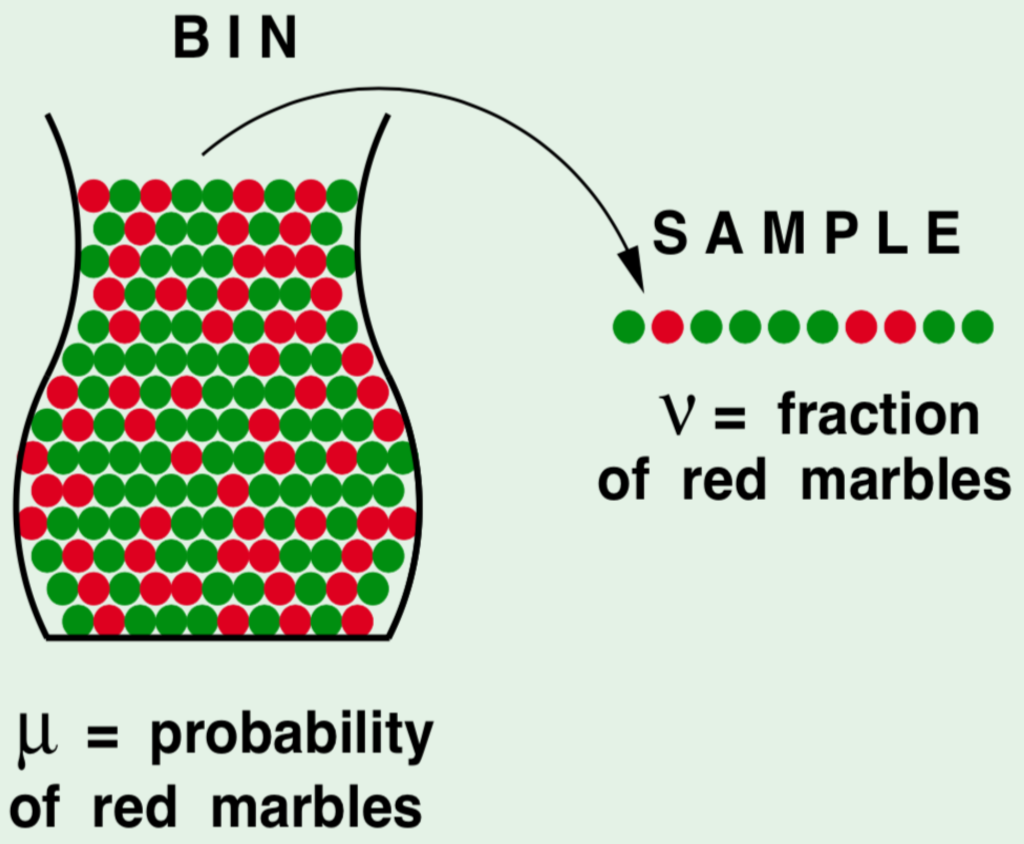
1. Learning from data
2. Error & Noise
3. Approximation vs Generalization
4. Bias – Variance trade-off
5. Learning curve

1. LEARNING FROM DATA [2]



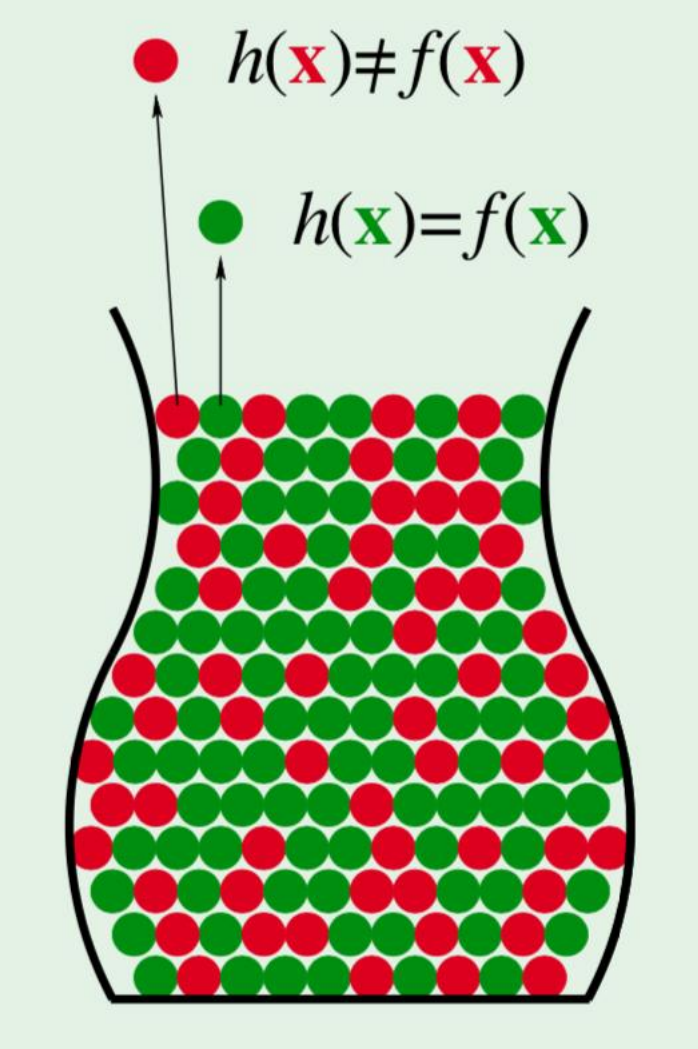
1. **Learning:**
   1. Unknown target function y = f(x)
   2. Dataset $(x_1, y_1), (x_2, y_2)$ …
   3. Learning algorithms pick g ~ f from Hypothesis Set H
2. **Learning components:**
   1. Learning algorithm
   2. Hypothesis set
3. **Purpose:**
   1. g(x) ~ f(x)

1. IS LEARNING FEASIBLE [2]

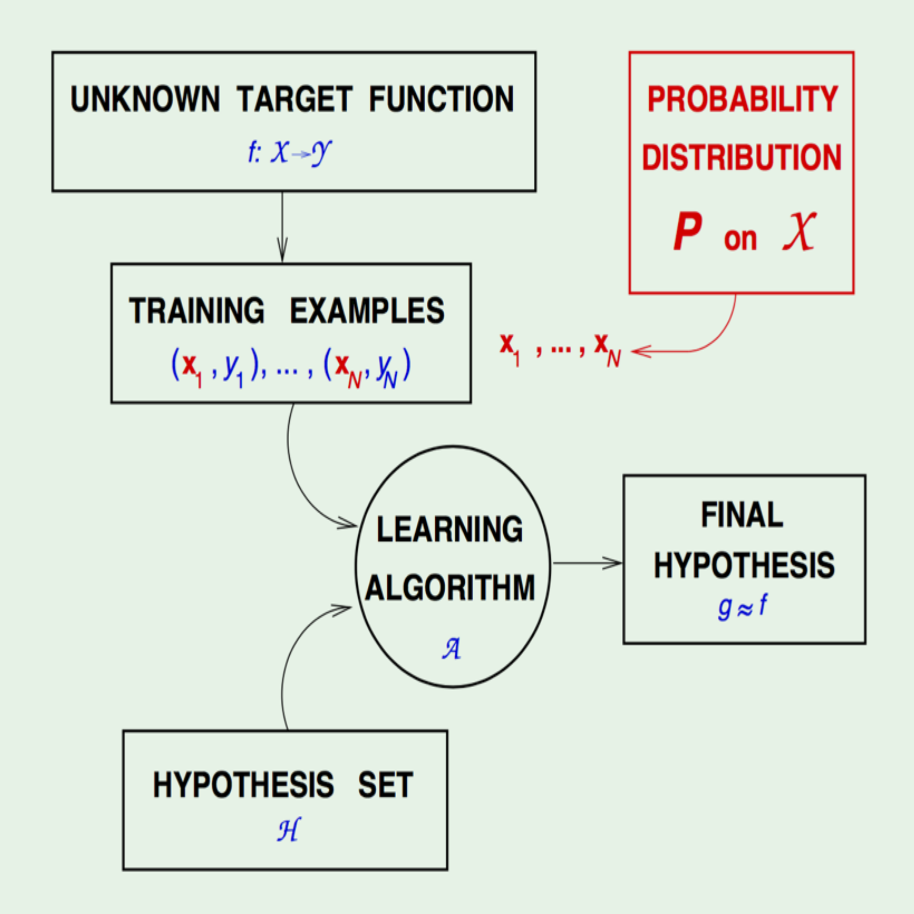


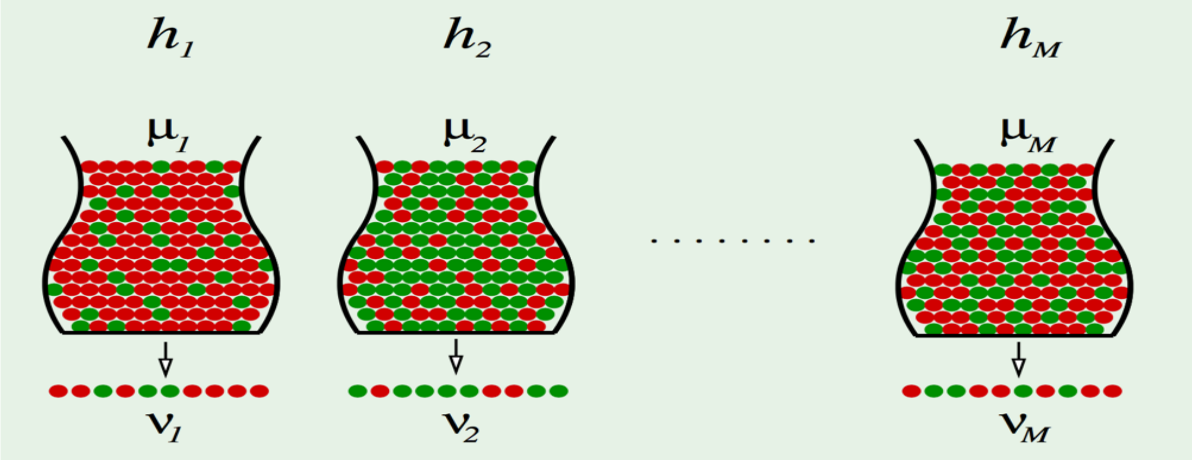
$$\mathbb{P}\left[\,|\nu - \mu| > \epsilon\,\right] \leq 2e^{-2\epsilon^2 N}$$

Hoeffding’s inequality



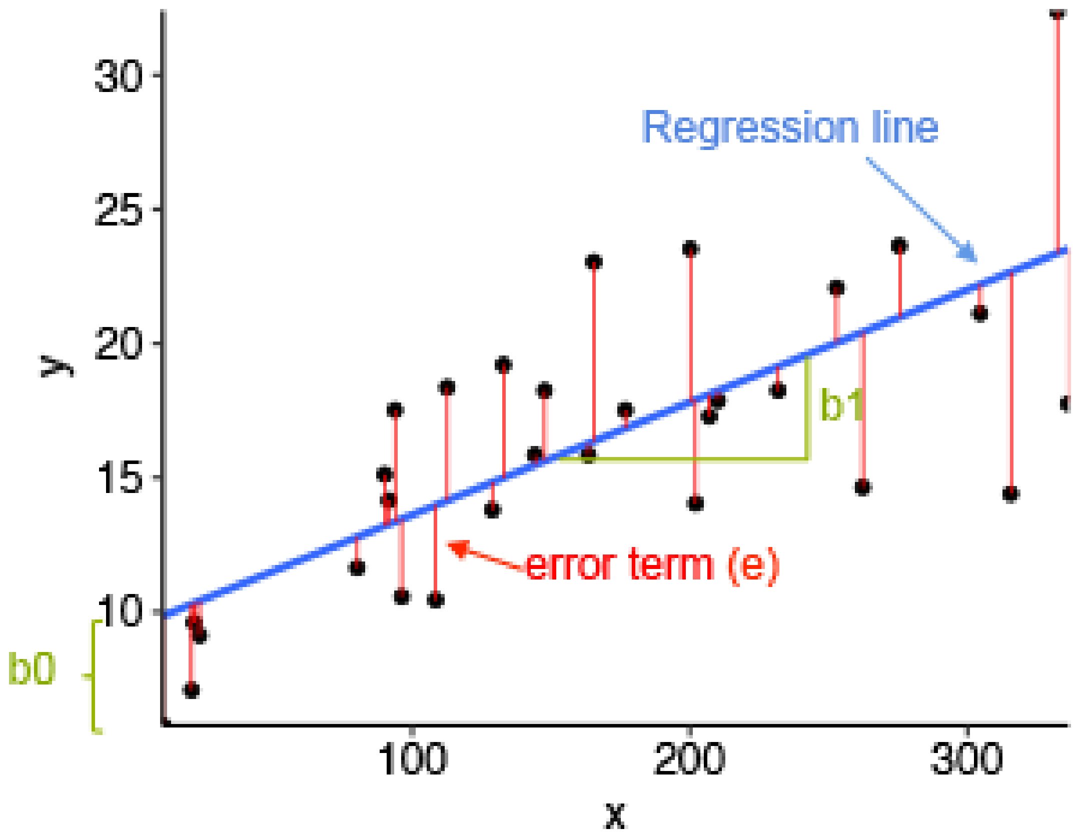
1. IS LEARNING FEASIBLE [2]





$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

2. ERROR [2]



[1]

Learning Purpose: g(x) ~ f(x)

But what the "g ~ f" mean ? E(g,f)

Squared error: $\mathrm{e}\left(h(\mathbf{x}), f(\mathbf{x})\right) = \left(h(\mathbf{x}) - f(\mathbf{x})\right)^2$

Binary error: $\mathrm{e}\left(h(\mathbf{x}), f(\mathbf{x})\right) = [\![h(\mathbf{x}) \neq f(\mathbf{x})]\!]$

h	$+1$	
	-1	

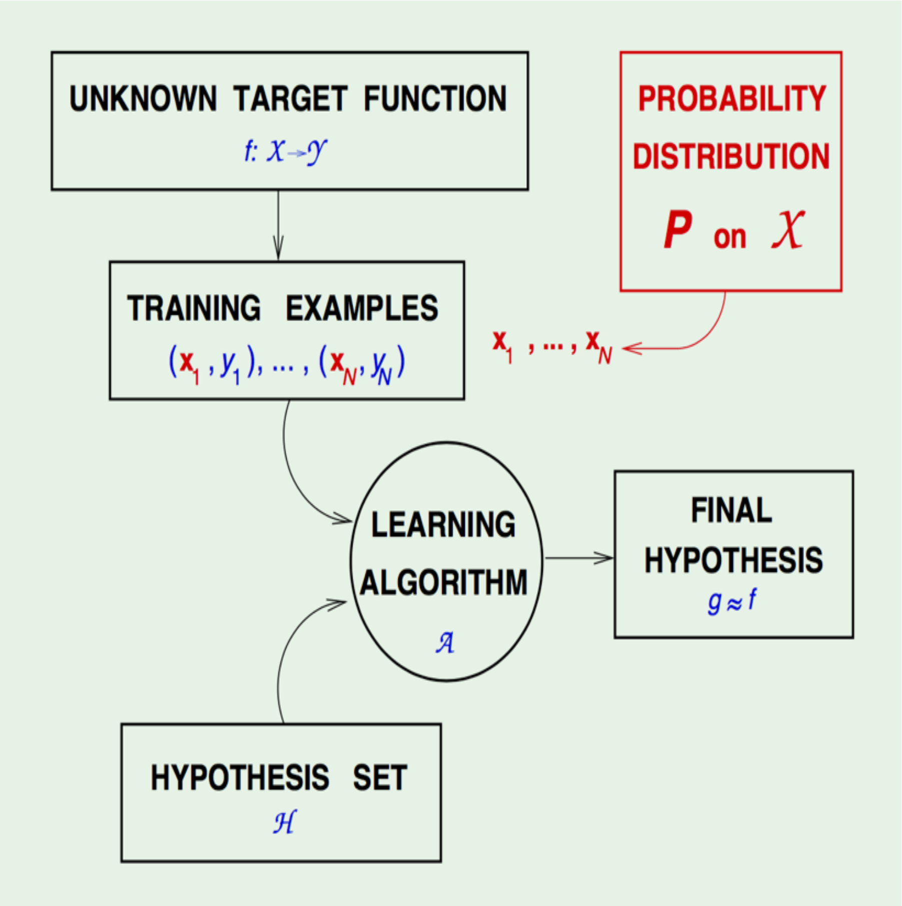
Supper market verify for discount

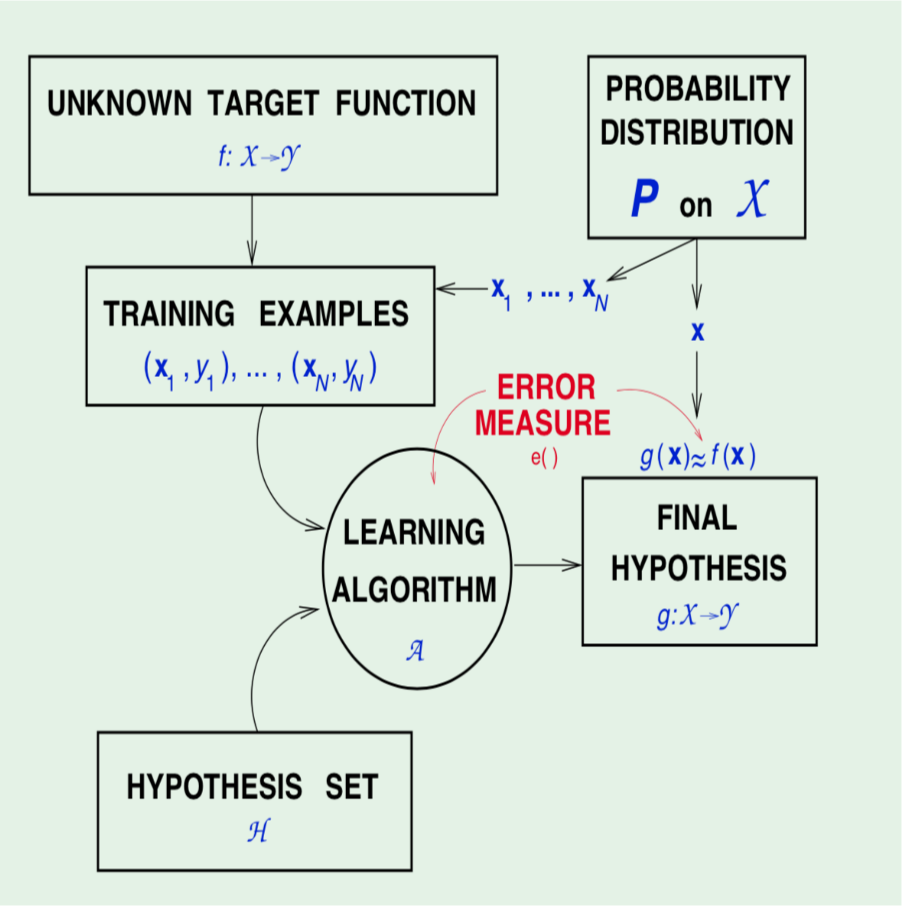
		f	
		$+1$	-1
h	$+1$	0	1
	-1	10	0

		f	
		$+1$	-1
h	$+1$	0	1000
	-1	1	0

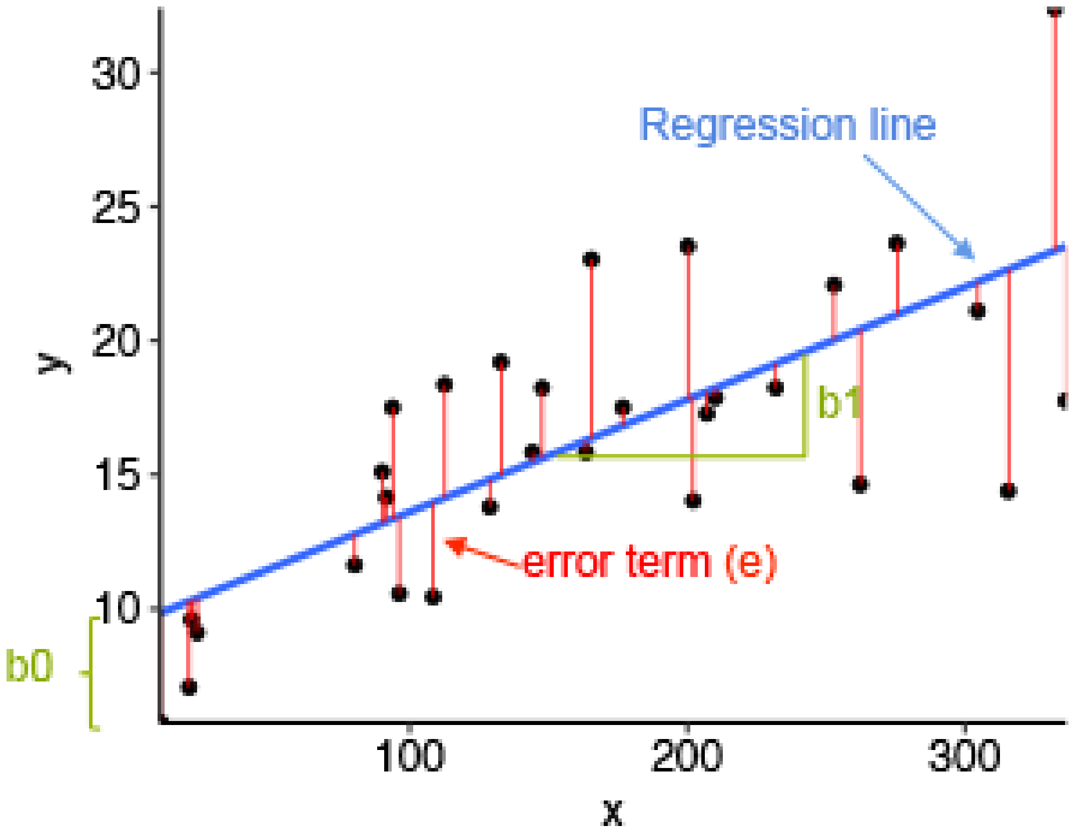
CIA verify for security

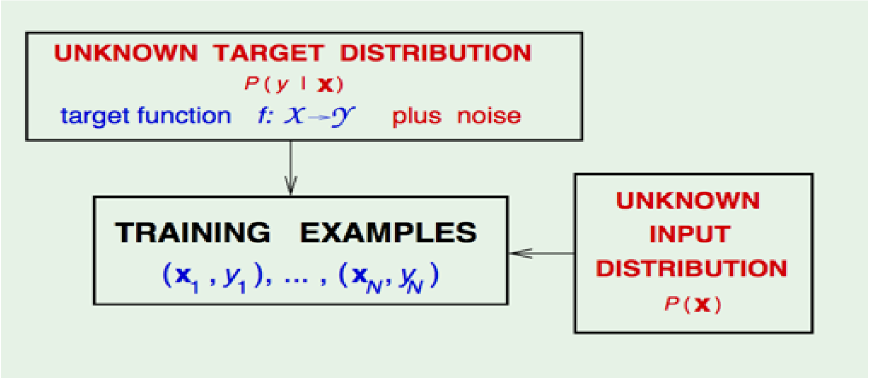
2. ERROR [2]


UNKNOWN TARGET FUNCTION
f: X→Y
PROBABILITY DISTRIBUTION
P on X
TRAINING EXAMPLES
(x1 , y1), ... , (xN, yN)
x1 , ... , xN
LEARNING ALGORITHM
A
FINAL HYPOTHESIS
g ≈ f
HYPOTHESIS SET
H


UNKNOWN TARGET FUNCTION
f: X→Y
PROBABILITY DISTRIBUTION
P on X
x1 , ... , xN
x
TRAINING EXAMPLES
(x1 , y1), ... , (xN, yN)
ERROR MEASURE
e( )
g(x)≈f(x)
LEARNING ALGORITHM
A
FINAL HYPOTHESIS
g:X→Y
HYPOTHESIS SET
H

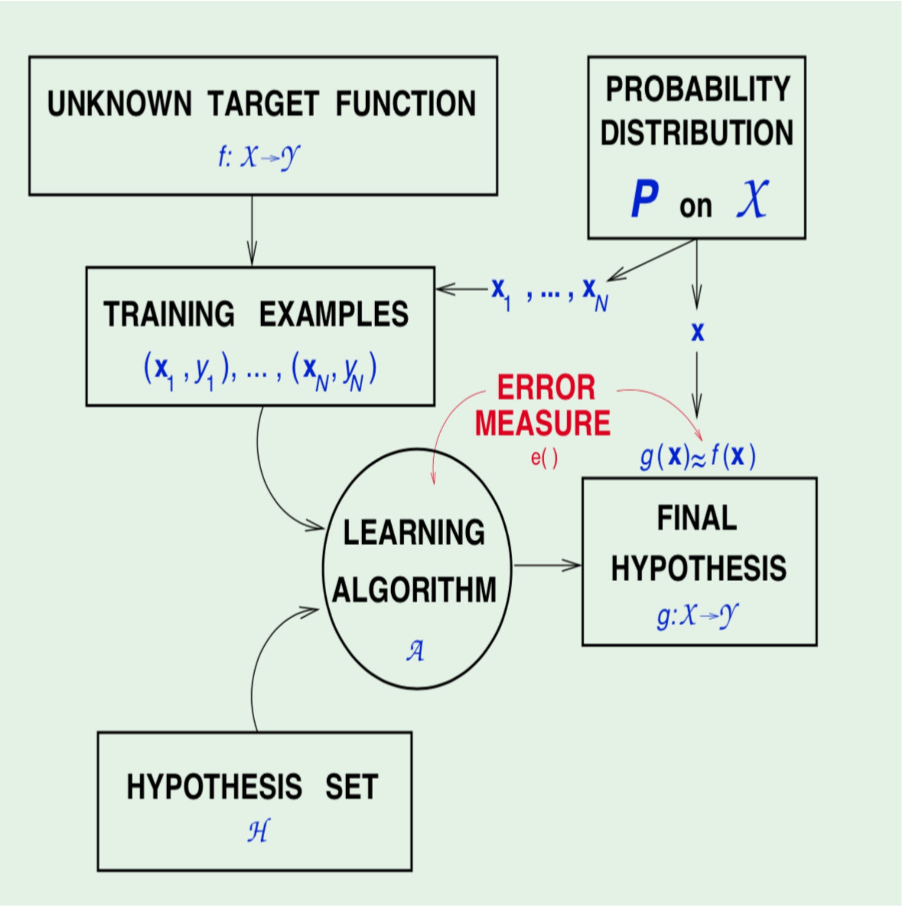
2. NOISE [2]

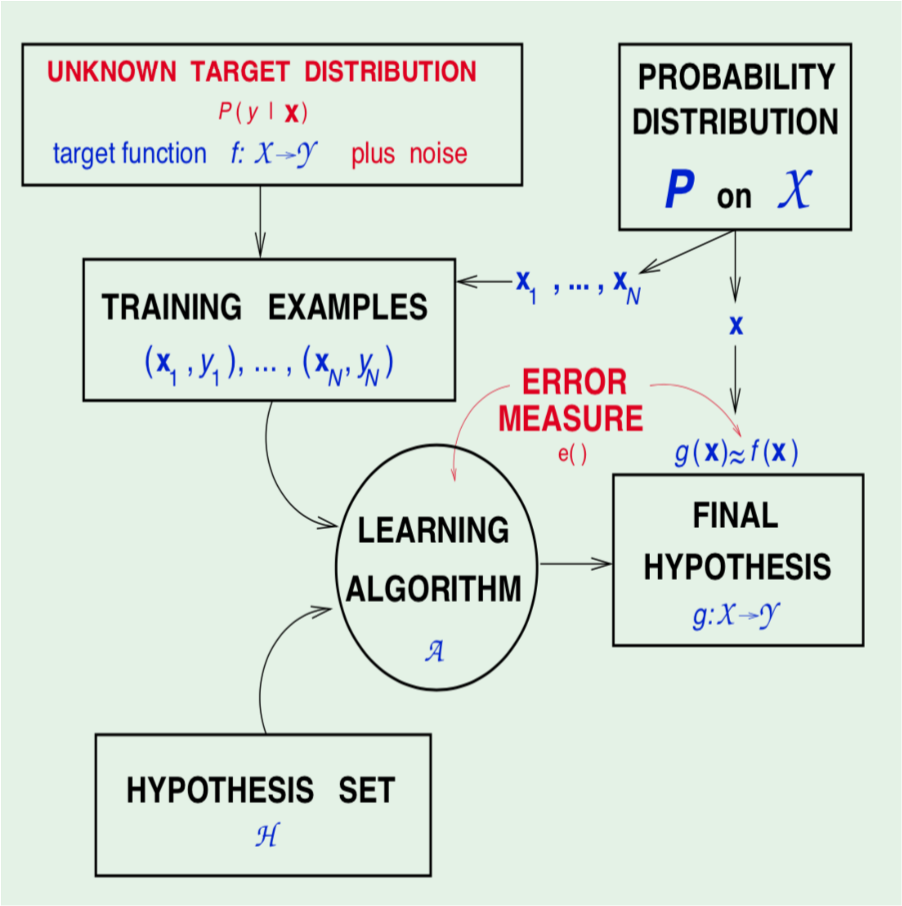




$$y = \hat{y} + \text{noise} = f(x) + \text{noise} = \mathbb{E}\,(y|x) + \text{noise}$$

2. NOISE [2]


UNKNOWN TARGET FUNCTION
$f: \mathcal{X} \to \mathcal{Y}$
PROBABILITY DISTRIBUTION
P on $\mathcal{X}$
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
$\mathbf{x}_1, \ldots, \mathbf{x}_N$
$\mathbf{x}$
ERROR MEASURE
e( )
$g(\mathbf{x}) \approx f(\mathbf{x})$
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g: \mathcal{X} \to \mathcal{Y}$
HYPOTHESIS SET
$\mathcal{H}$


UNKNOWN TARGET DISTRIBUTION
$P(y \mid \mathbf{x})$
target function $f: \mathcal{X} \to \mathcal{Y}$ plus noise
PROBABILITY DISTRIBUTION
P on $\mathcal{X}$
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
$\mathbf{x}_1, \ldots, \mathbf{x}_N$
$\mathbf{x}$
ERROR MEASURE
e( )
$g(\mathbf{x}) \approx f(\mathbf{x})$
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g: \mathcal{X} \to \mathcal{Y}$
HYPOTHESIS SET
$\mathcal{H}$

2. PREAMBLE OF THE THEORY [2]

$E_{out}(g) \approx E_{in}(g)$ *(1)*

$E_{in}(g) \approx 0$ *(2)*

(1)Hoeffding's inequality

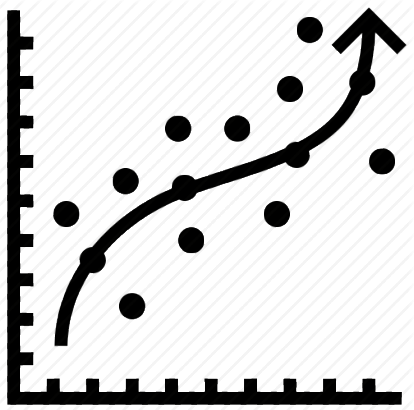
(2)Optimize error

➔ g ~ f

- f(x) – y = (stochastic) noise
- f(x) – g(x) = (deterministic) noise
- y – g(x) = error

3. APPROXIMATION-GENERALIZATION [2]

income



expenditure

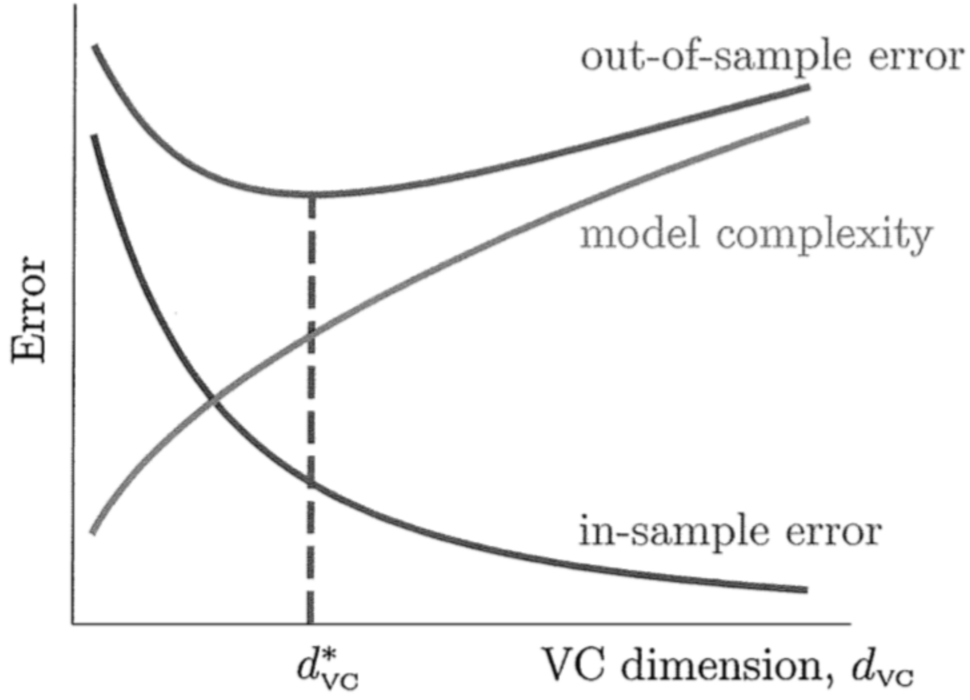
E_{in} → $E_{in}(g) \approx 0$

E_{out} → $E_{out}(g) \approx E_{in}(g)$

Model Complexity

$\sim\mathcal{H}$

3. APPROXIMATION-GENERALIZATION [2]



$d_{VC} \sim \mathcal{H}$

Approximation – generalization trade-off

More complex H ➔ better chance of approximation f

Less complex H ➔ better chance of generalizing out of sample

With probability ≥ 1- δ

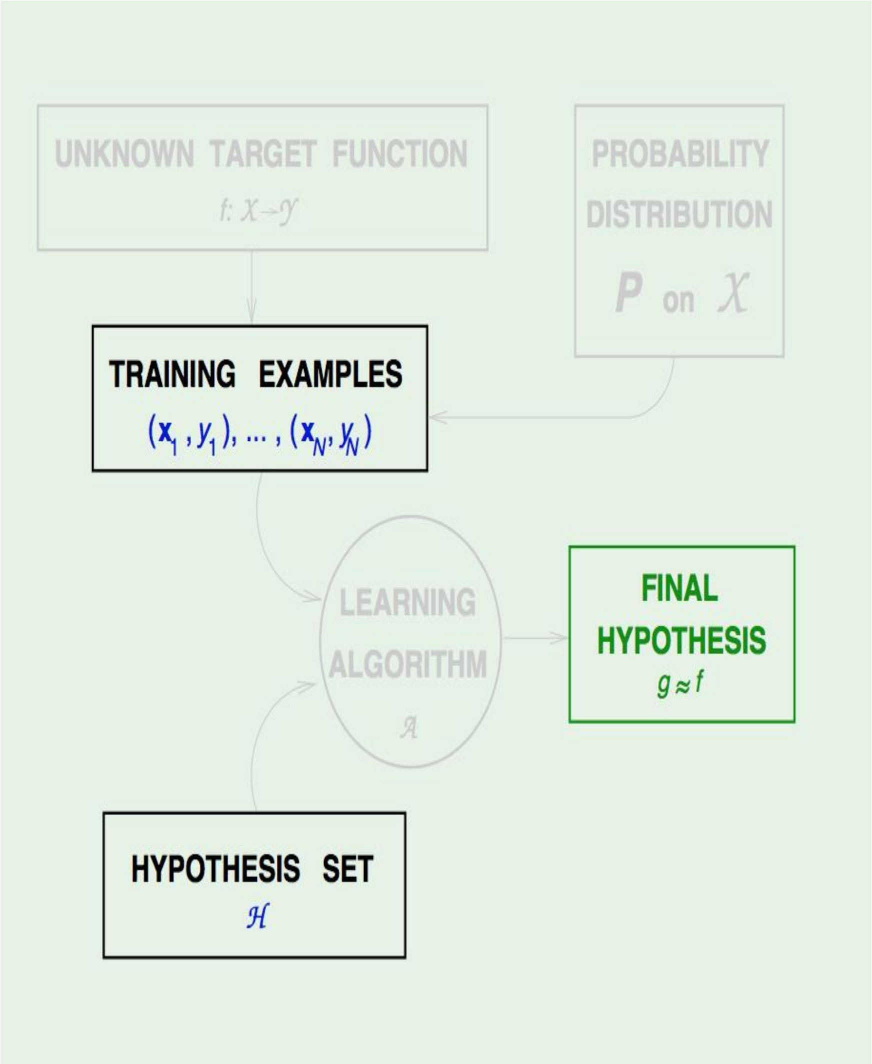
$$E_{out}(g) - E_{in}(g) \leq \Omega(\mathcal{H}, N, \delta)$$

$\mathcal{H}$ ~ model complexity

N: sample size

$1 - \delta$: confidence requirement

3. APPROXIMATION-GENERALIZATION [2]



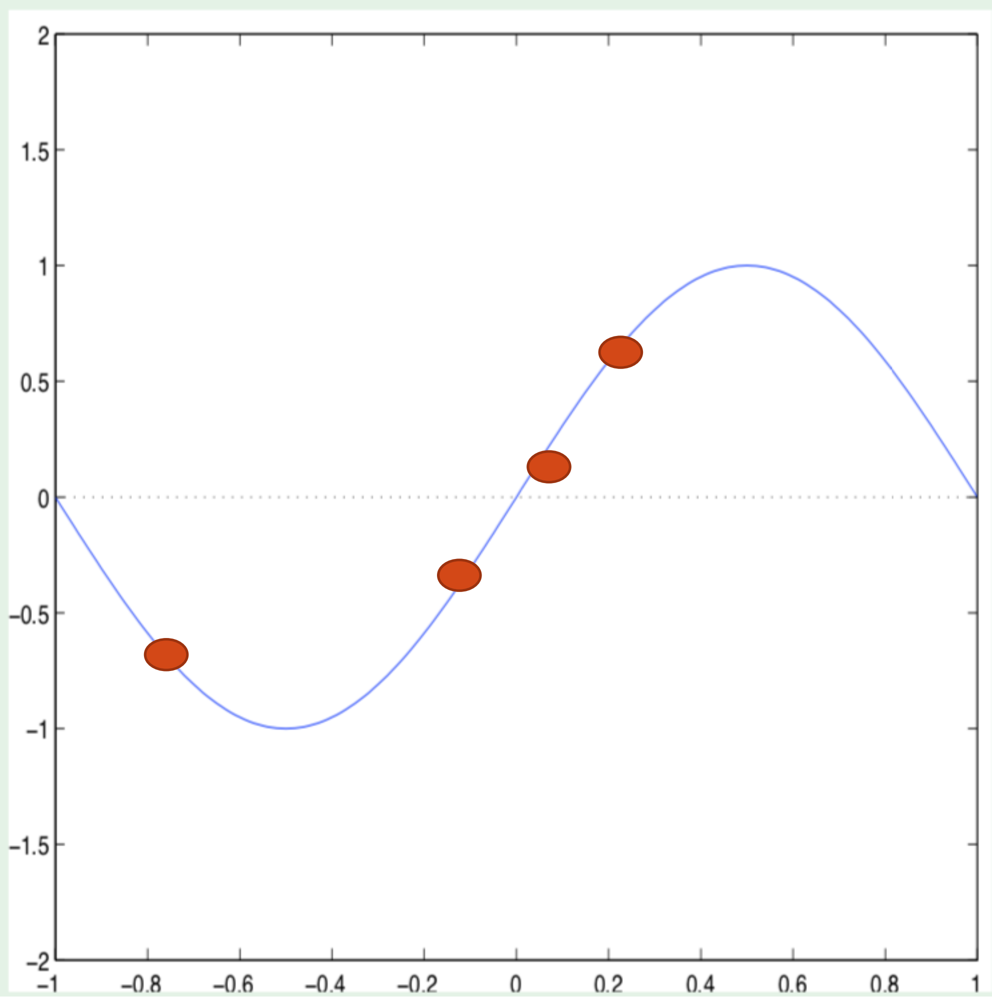
$$\mathbb{P}[|E_{\text{out}} - E_{\text{in}}| > \epsilon] \leq \underbrace{4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}}_{\delta}$$

VC Dimension (1960 – 1990)
"fundamental theory of learning"
Vladimir Vapnik - Alexey Chervonenkis

$$\Omega(N, \mathcal{H}, \delta) = \sqrt{\frac{8}{N} \ln\left(\frac{4m_{\mathcal{H}}(2N)}{\delta}\right)}$$

Generalization bound

4. BIAS – VARIANCE TRADEOFF



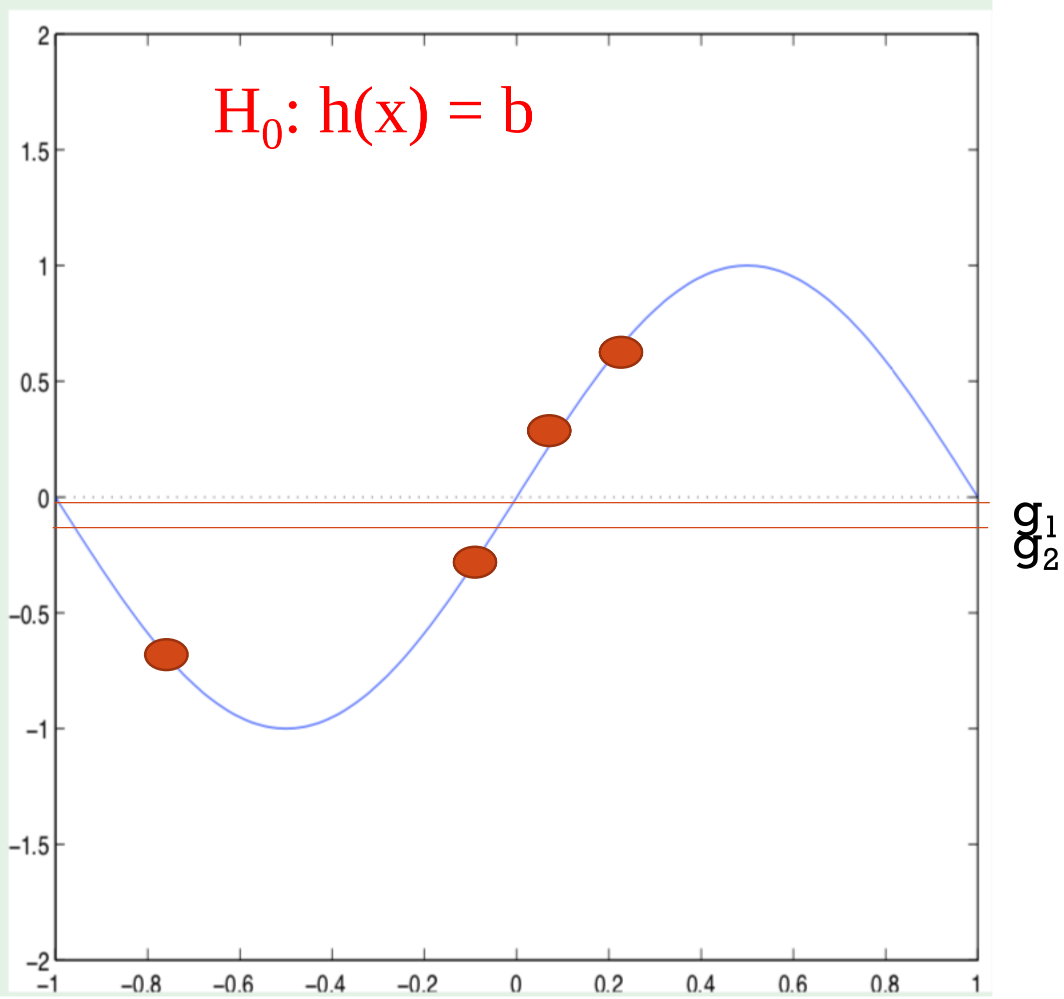
$y = f(x) = \sin(\pi x).$

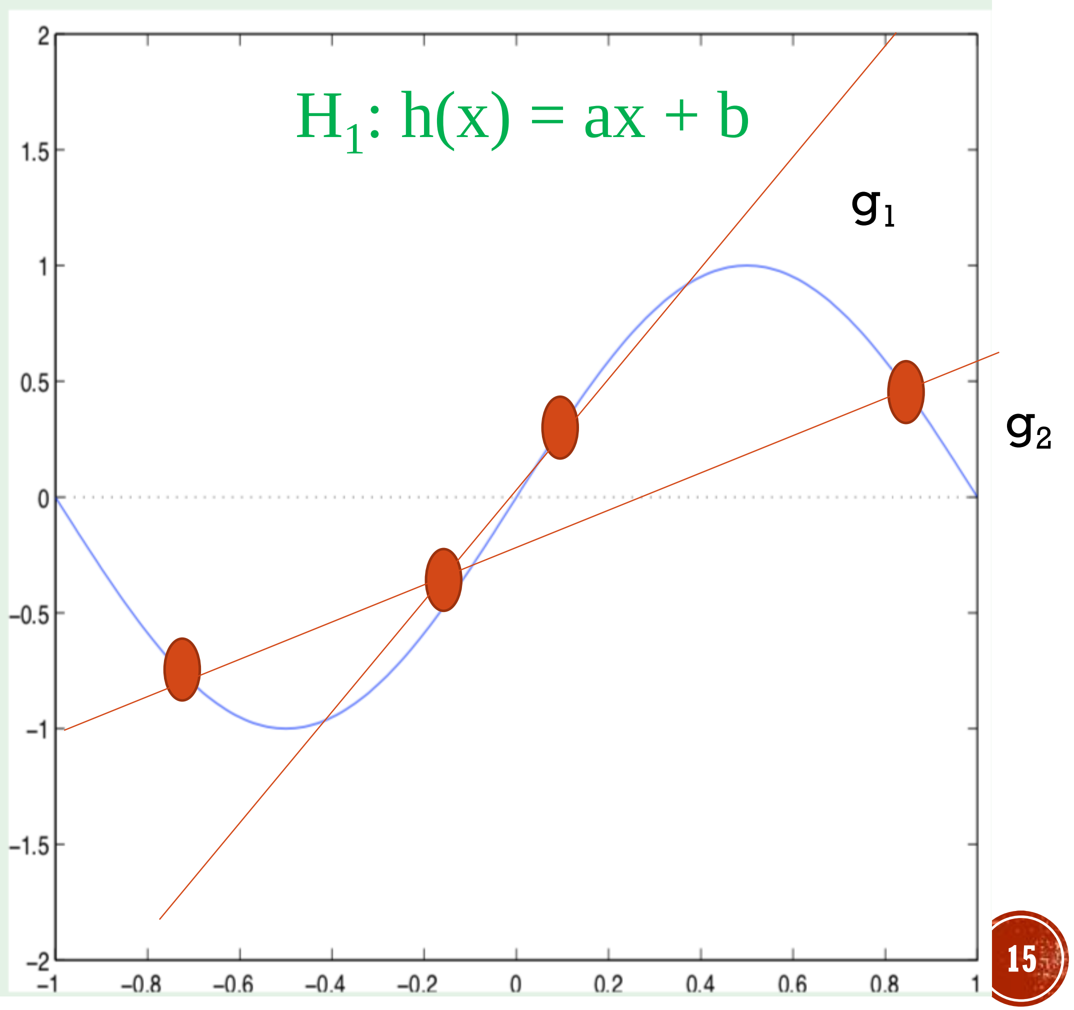
H_0: $h(x) = b$ vs H_1: $h(x) = ax + b$

Which is better?

Approximation & generalization

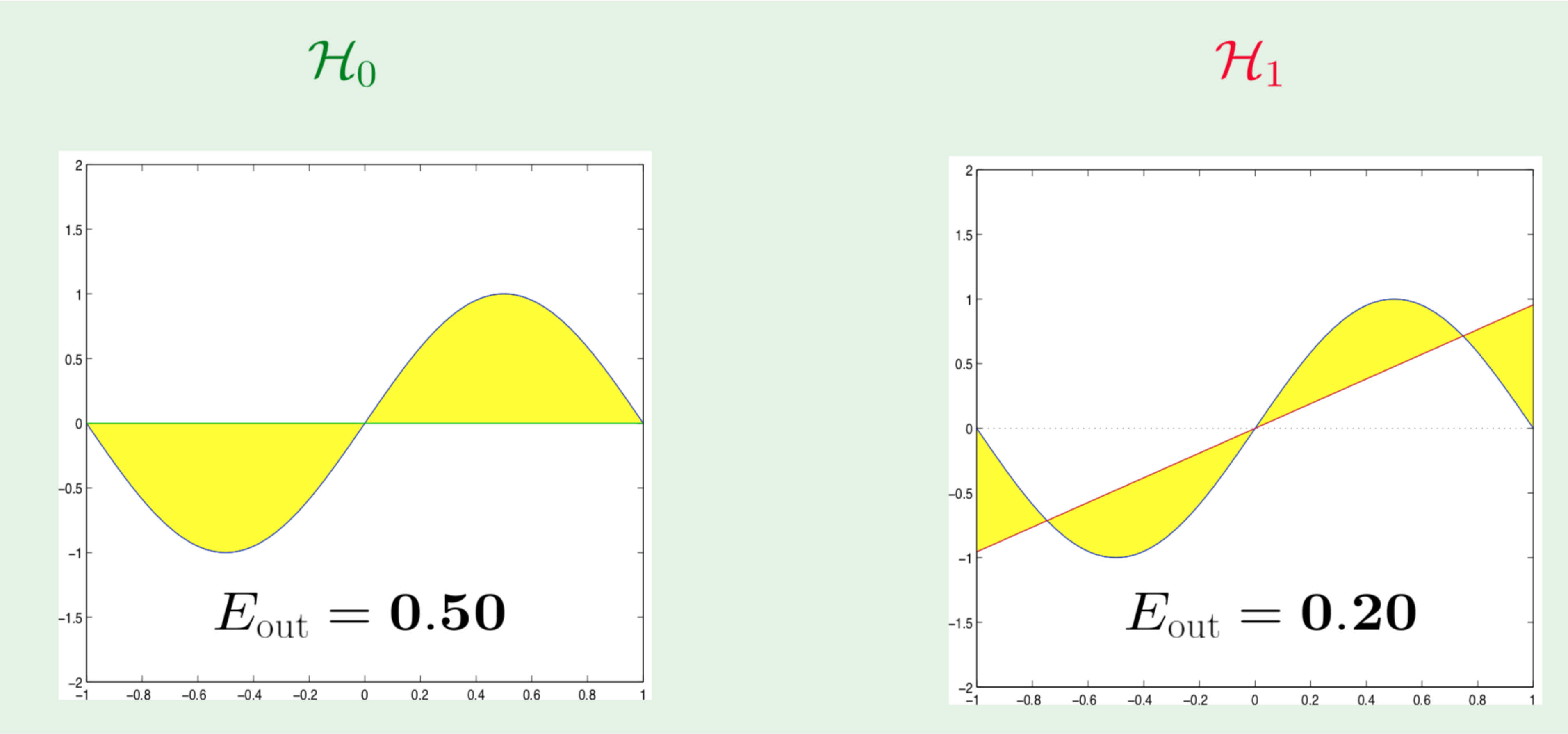
4. BIAS – VARIANCE TRADEOFF





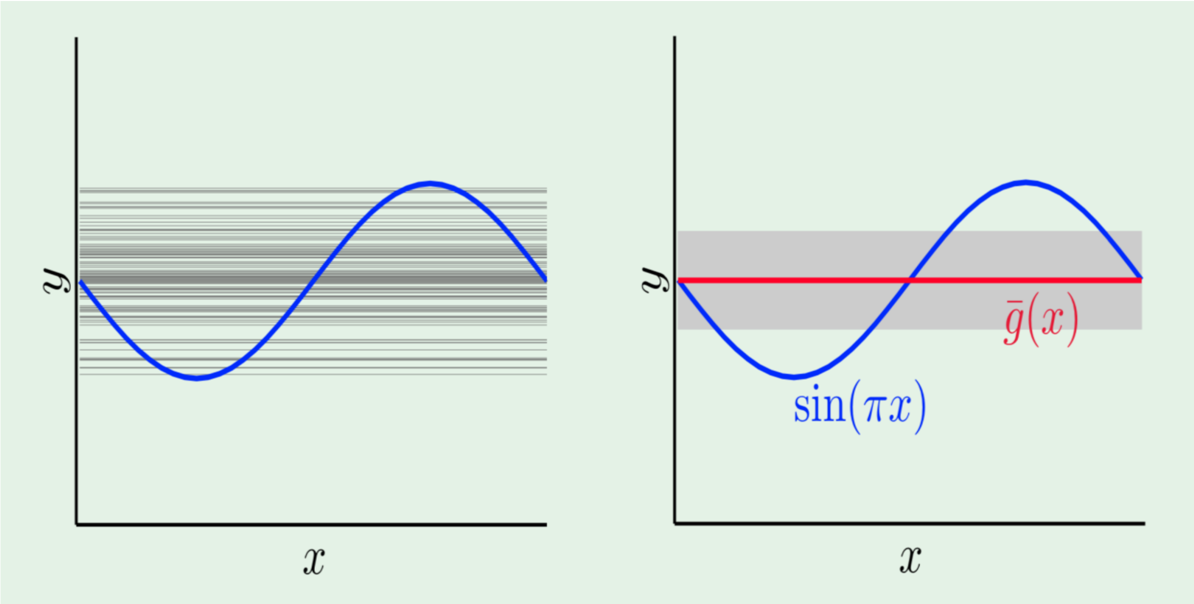
4. BIAS – VARIANCE TRADEOFF

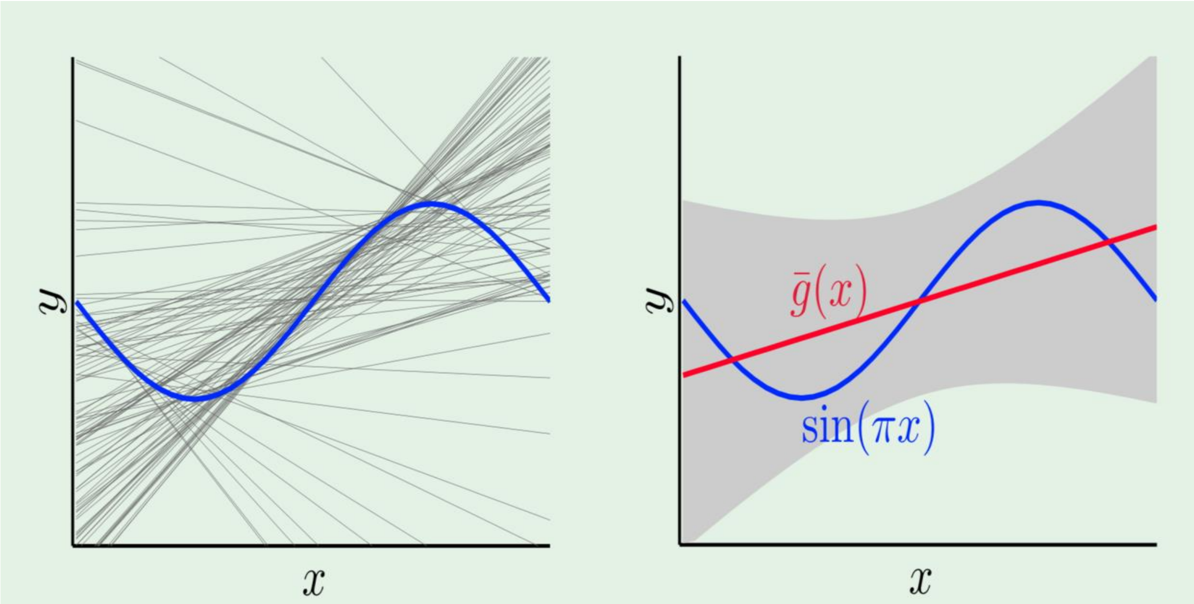
"Approximation" - bias



4. BIAS – VARIANCE TRADEOFF

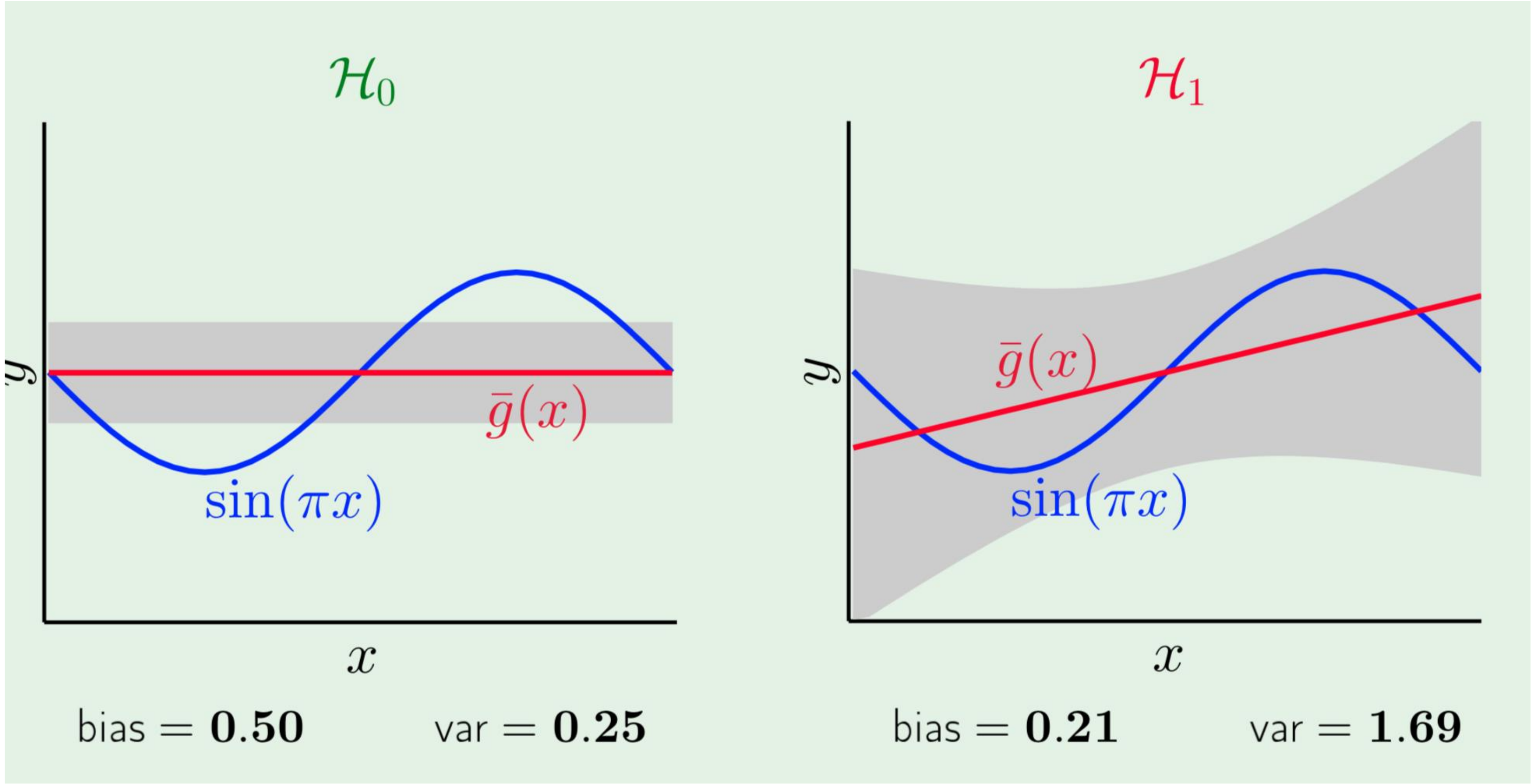
"Generalization" - Variance



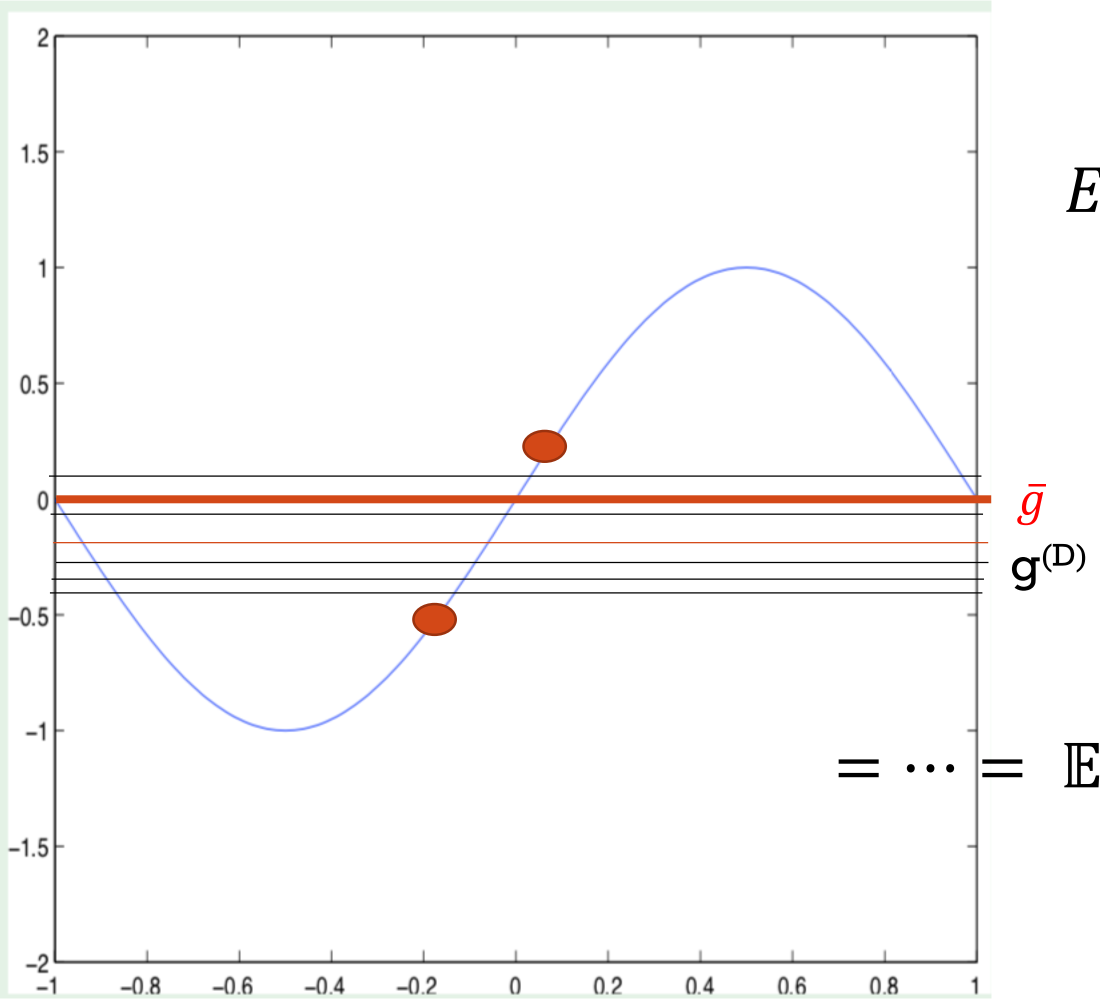


4. BIAS – VARIANCE TRADEOFF [2][4]

Bias – Variance – who win ?



4. BIAS – VARIANCE DECOMPOSITION [2]



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

$$= \mathbb{E}_D\left[\left(g^{(D)}(x) - \bar{g}(\mathrm{x}) + \bar{g}(\mathrm{x}) - f(x)\right)^2\right]$$

$$= \cdots = \mathbb{E}_D\left[\left(g^{(D)}(x) - \bar{g}(\mathrm{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{g}(\mathrm{x}) - f(x)\right)^2\right]$$

3. BIAS - VARIANCE DECOMPOSITION [2]

$$E_{out}\left(g^{(D)}\right) = \mathbb{E}_D\left[\left(g^{(D)}(x) - \bar{g}\,(\mathrm{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{g}\,(\mathrm{x}) - f(x)\right)^2\right]$$

Bias

Variance

Bias – variance decomposition E_{out} to:

- How well H can approximate f
- How well we can zoom in on a good h of H

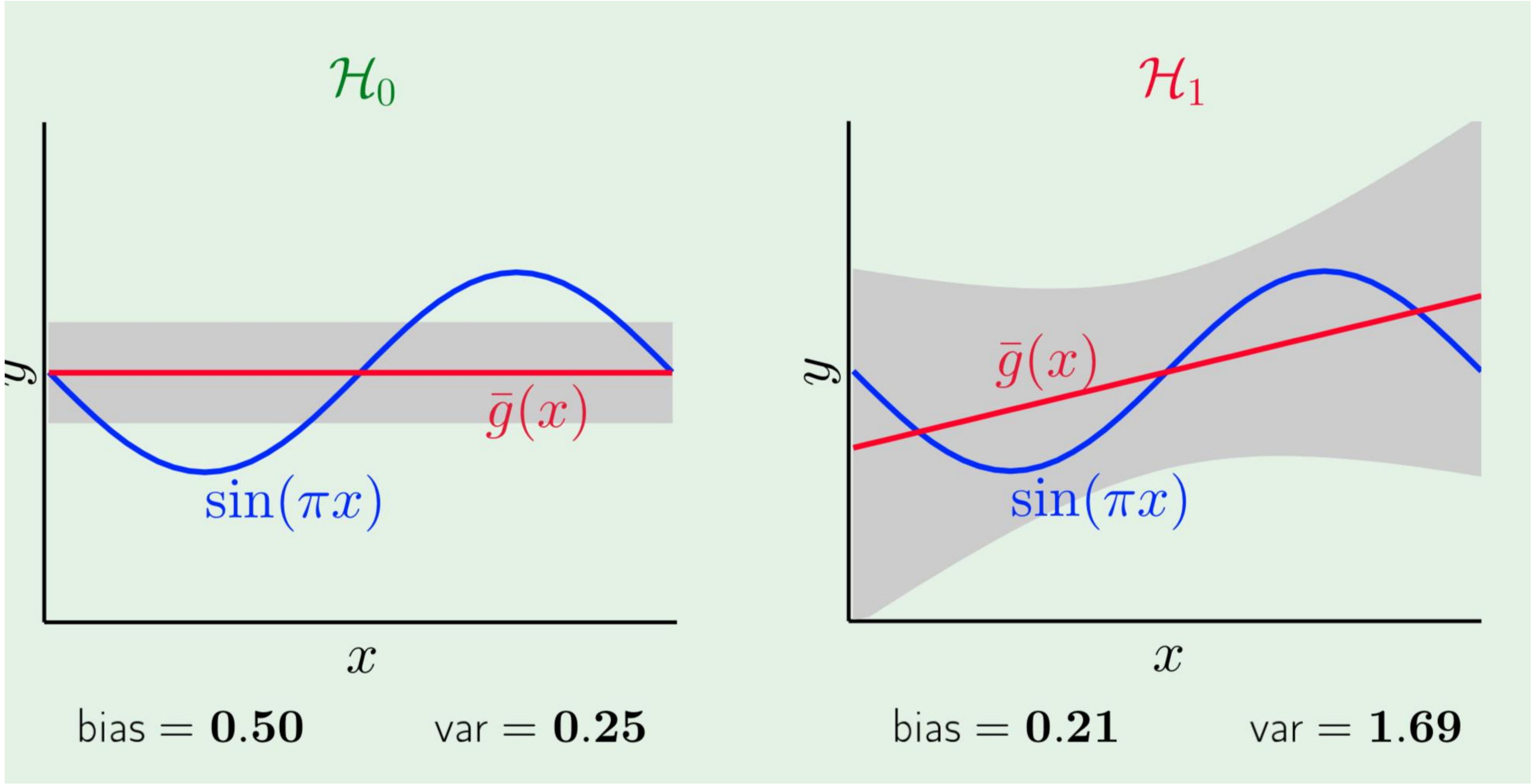
Imagine **many** data sets $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_K$

$$\bar{g}(\mathbf{x}) \approx \frac{1}{K}\sum_{k=1}^{K} g^{(\mathcal{D}_k)}(\mathbf{x})$$

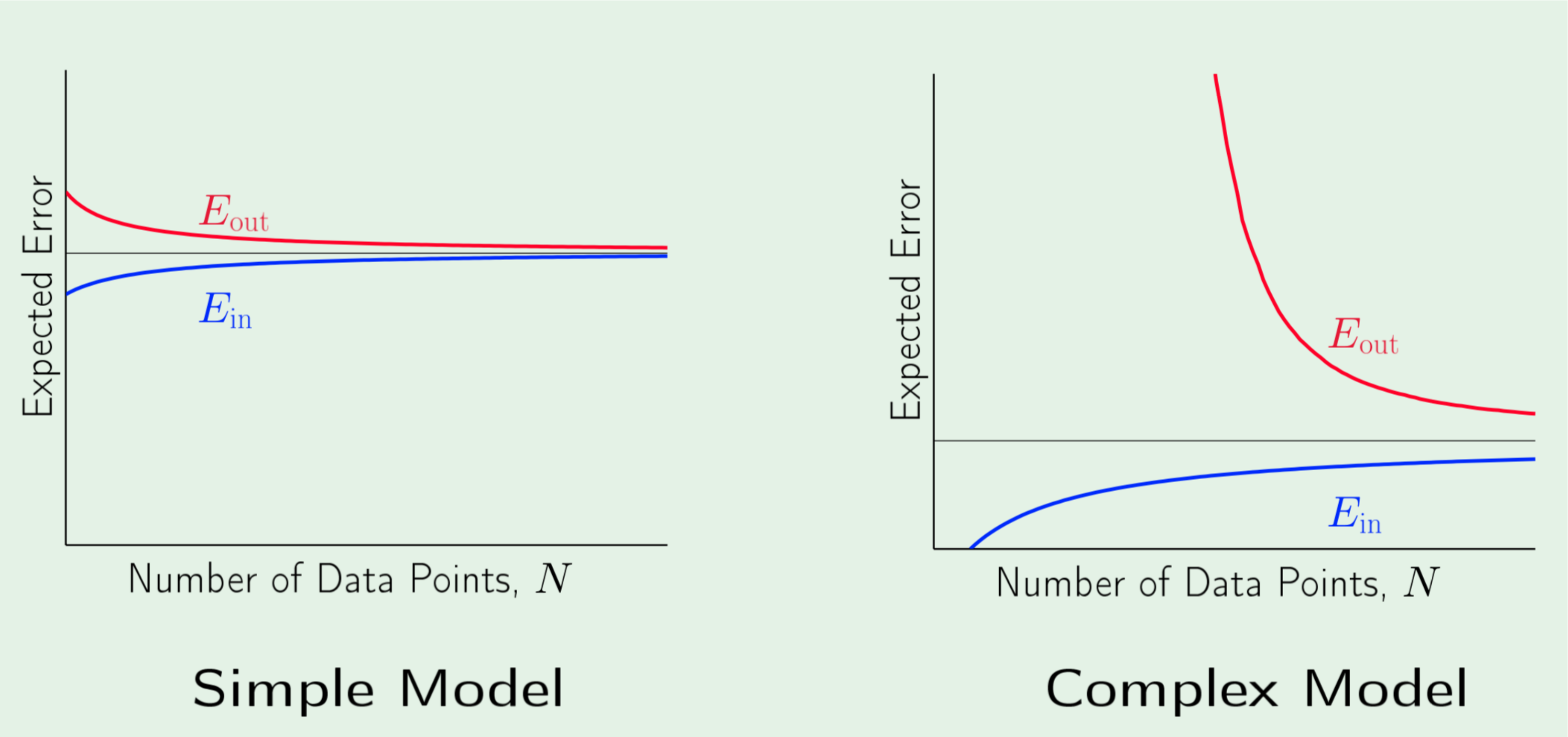
4. BIAS – VARIANCE TRADEOFF

WHO WON … ?

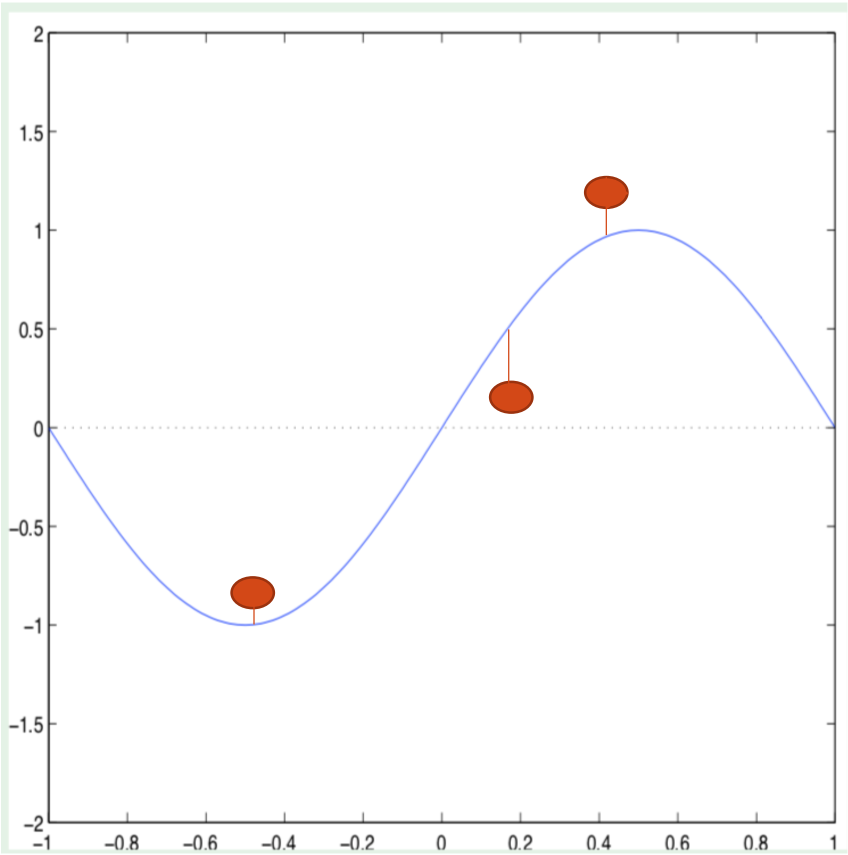
Congratulation $\mathcal{H}_0$



5. THE LEARNING CURVE



BUT NOISE ....



REFERENCES:

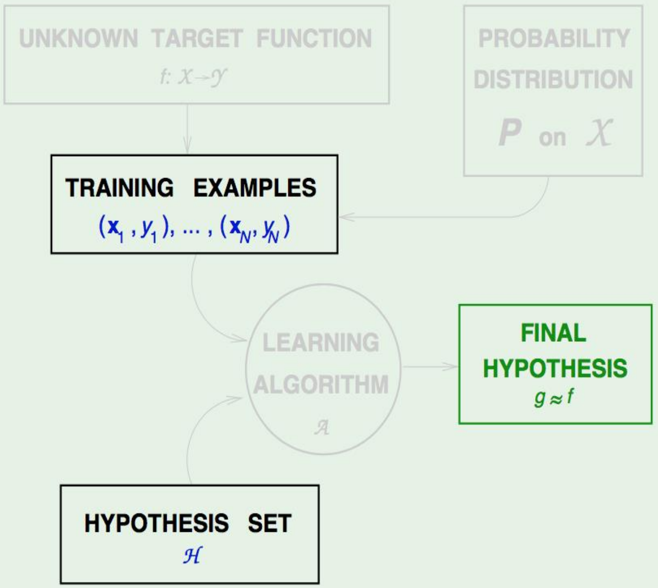
1. Pic: http://www.sthda.com/english/articles/40-regression-analysis/165-linear-regression-essentials-in-r/
2. Yaser S. Abu-Mostafa, Malik Magdon-Ismail, Hsuan-Tien Lin-Learning From Data. A short course-AMLBook (2012)
3. https://medium.com/@mp32445/understanding-bias-variance-tradeoff-ca59a22e2a83
4. Bishop - Pattern Recognition And Machine Learning - Springer 2006

3. GENERALIZATION [2]

VC Dimension (1960 – 1990)
"fundamental theory of learning"
Vladimir Vapnik - Alexey Chervonenkis

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

$$\mathbb{P}[\ |E_{in}(g) - E_{out}(g)| > \epsilon\ ] \leq 4\ m_{\mathcal{H}}(2N)\ e^{-\frac{1}{8}\epsilon^2 N}$$



$$\mathbb{P}[|E_{out} - E_{in}| > \epsilon] \leq \underbrace{4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}}_{\delta}$$

$$\delta = 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N} \implies \epsilon = \underbrace{\sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}}_{\Omega}$$

With probability $\geq 1 - \delta$, $\quad |E_{out} - E_{in}| \leq \Omega(N, \mathcal{H}, \delta)$

BIAS-VARIANCE VS VC.DIMENSION