Bias-Variance Trade-off, Regularization

Machine Learning

Hamid R Rabiee – Zahra Dehghanian

Spring 2025

Sharif University
of Technology

Model complexity: Bias-variance trade-off

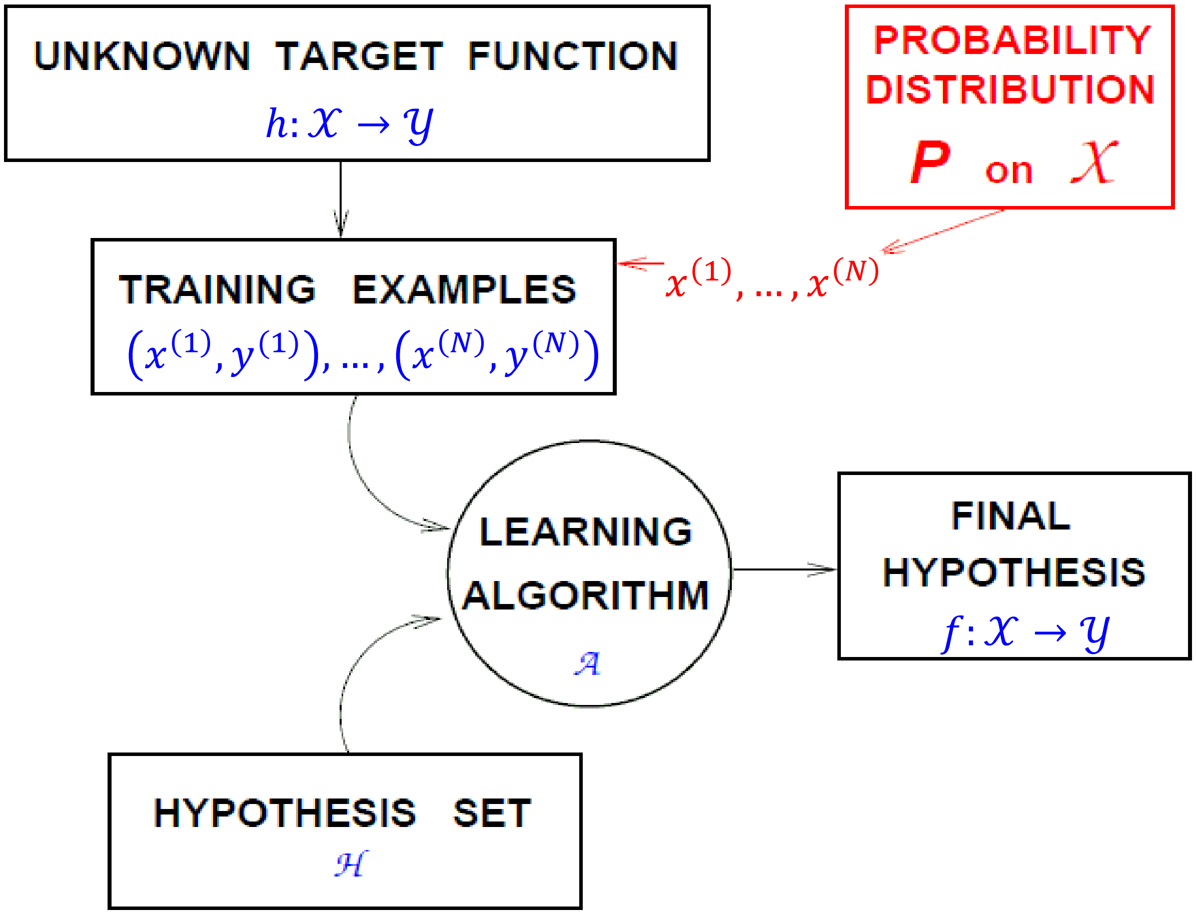
- Least squares can lead to severe over-fitting if complex models are trained using data sets of limited size.

- A frequentist viewpoint of the model complexity issue, known as the *bias-variance* trade-off.

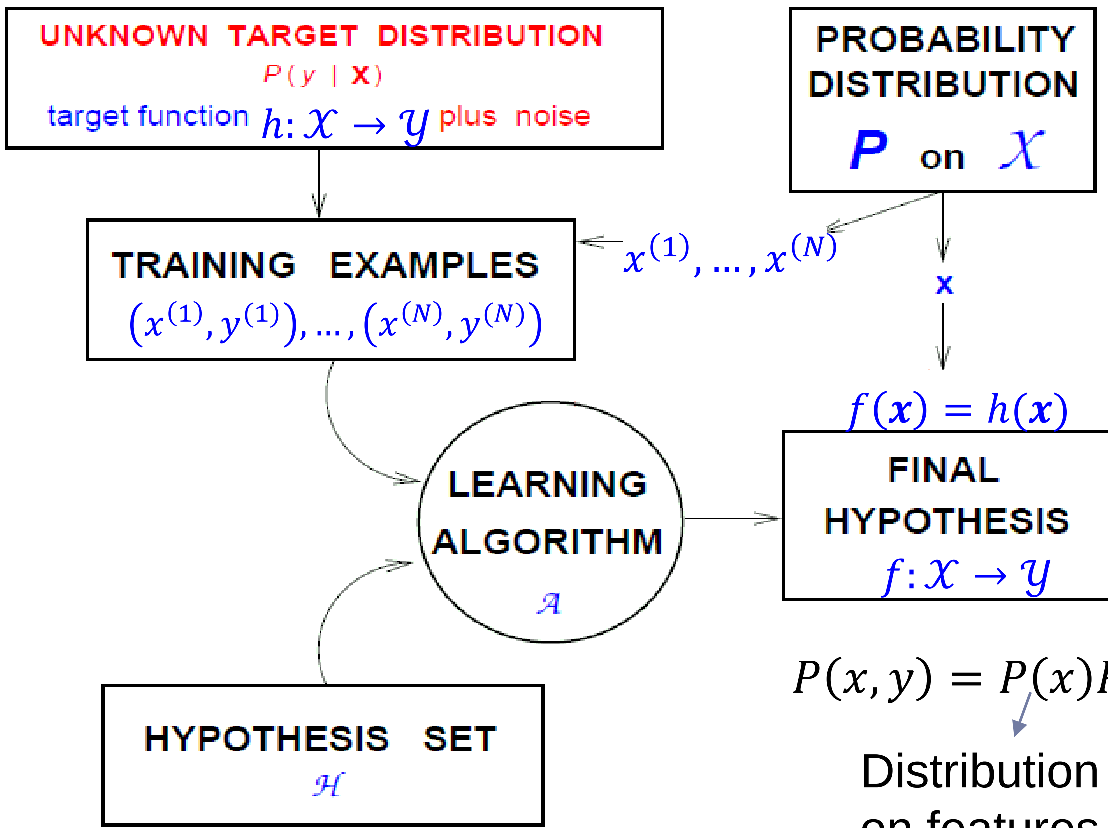
Formal discussion on bias, variance, and noise

- Best unrestricted regression function
- Noise
- Bias and variance

The learning diagram: deterministic target



The learning diagram including noisy target



$$P(x, y) = P(x)P(y|x)$$

Distribution on features

Target distributio
n

[Y.S. Abou Mostafa, 2012]

Best unrestricted regression function

- If we know the joint distribution $P(\boldsymbol{x}, y)$ and no constraints on the regression function?
  - cost function: mean squared error

$$h^* = \underset{h:\mathbb{R}^d \to \mathbb{R}}{\operatorname{argmin}} \mathbb{E}_{\boldsymbol{x},y} \left[\left(y - h(\boldsymbol{x}) \right)^2 \right]$$

$$h^*(\boldsymbol{x}) = \mathbb{E}_{y|\boldsymbol{x}}[y]$$

Best unrestricted regression function: Proof

$$\mathbb{E}_{\boldsymbol{x},y}\left[\left(y - h(\boldsymbol{x})\right)^2\right] = \iint \left(y - h(\boldsymbol{x})\right)^2 p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy$$

Best unrestricted regression function: Proof

$$\mathbb{E}_{\boldsymbol{x},y}\left[\left(y-h(\boldsymbol{x})\right)^2\right]=\iint\left(y-h(\boldsymbol{x})\right)^2 p(\boldsymbol{x},y)d\boldsymbol{x}dy$$

- For each $\boldsymbol{x}$, separately minimize loss since $h(\boldsymbol{x})$ can be chosen independently for each different $\boldsymbol{x}$:

$$\frac{\delta\mathbb{E}_{\boldsymbol{x},y}\left[\left(y-h(\boldsymbol{x})\right)^2\right]}{\delta h(\boldsymbol{x})}=-\int 2\left(y-h(\boldsymbol{x})\right)p(\boldsymbol{x},y)dy=0$$

Best unrestricted regression function: Proof

$$\mathbb{E}_{\boldsymbol{x},y}\left[\left(y-h(\boldsymbol{x})\right)^2\right] = \iint \left(y-h(\boldsymbol{x})\right)^2 p(\boldsymbol{x},y)d\boldsymbol{x}dy$$

- For each $\boldsymbol{x}$, separately minimize loss since $h(\boldsymbol{x})$ can be chosen independently for each different $\boldsymbol{x}$:

$$\frac{\delta\mathbb{E}_{\boldsymbol{x},y}\left[\left(y-h(\boldsymbol{x})\right)^2\right]}{\delta h(\boldsymbol{x})} = -\int 2\left(y-h(\boldsymbol{x})\right)p(\boldsymbol{x},y)dy = 0$$

$$\Rightarrow h(\boldsymbol{x}) = \frac{\int yp(\boldsymbol{x},y)dy}{\int p(\boldsymbol{x},y)dy} = \frac{\int yp(\boldsymbol{x},y)dy}{p(\boldsymbol{x})} = \int yp(y|\boldsymbol{x})dy = \mathbb{E}_{y|\boldsymbol{x}}[y]$$

Best unrestricted regression function: Proof

$$\mathbb{E}_{\boldsymbol{x},y}\left[\left(y-h(\boldsymbol{x})\right)^2\right] = \iint \left(y-h(\boldsymbol{x})\right)^2 p(\boldsymbol{x},y)d\boldsymbol{x}dy$$

- For each $\boldsymbol{x}$, separately minimize loss since $h(\boldsymbol{x})$ can be chosen independently for each different $\boldsymbol{x}$:

$$\frac{\delta\mathbb{E}_{\boldsymbol{x},y}\left[\left(y-h(\boldsymbol{x})\right)^2\right]}{\delta h(\boldsymbol{x})} = -\int 2\left(y-h(\boldsymbol{x})\right)p(\boldsymbol{x},y)dy = 0$$

$$\Rightarrow h(\boldsymbol{x}) = \frac{\int yp(\boldsymbol{x},y)dy}{\int p(\boldsymbol{x},y)dy} = \frac{\int yp(\boldsymbol{x},y)dy}{p(\boldsymbol{x})} = \int yp(y|\boldsymbol{x})dy = \mathbb{E}_{y|\boldsymbol{x}}[y]$$

$$\Longrightarrow h^*(\boldsymbol{x}) = \mathbb{E}_{y|\boldsymbol{x}}[y]$$

Error decomposition

$(\boldsymbol{x}, y) \sim P$

$h(\boldsymbol{x})$: minimizes the expected loss

$$E_{true}\left(f_{\mathcal{D}}(\boldsymbol{x})\right) = \mathbb{E}_{\boldsymbol{x},y}[(f_{\mathcal{D}}(\boldsymbol{x}) - y)^2]$$

Expected loss

Error decomposition

$(\boldsymbol{x}, y) \sim P$

$h(\boldsymbol{x})$: minimizes the expected loss

$$E_{true}\big(f_{\mathcal{D}}(\boldsymbol{x})\big) = \mathbb{E}_{\boldsymbol{x},y}[(f_{\mathcal{D}}(\boldsymbol{x}) - y)^2]$$ Expected loss

$$= \mathbb{E}_{\boldsymbol{x},y}\,[(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x}) + h(\boldsymbol{x}) - y)^2]$$

Error decomposition

$(\boldsymbol{x}, y) \sim P$

$h(\boldsymbol{x})$: minimizes the expected loss

$$E_{true}\left(f_{\mathcal{D}}(\boldsymbol{x})\right) = \mathbb{E}_{\boldsymbol{x},y}[(f_{\mathcal{D}}(\boldsymbol{x}) - y)^2]$$

Expected loss

$$= \mathbb{E}_{\boldsymbol{x},y}\,[(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x}) + h(\boldsymbol{x}) - y)^2]$$

$$= \mathbb{E}_{\boldsymbol{x}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2\right] + \mathbb{E}_{\boldsymbol{x},y}[(h(\boldsymbol{x}) - \boldsymbol{y})^2]$$
$$+2\mathbb{E}_{\boldsymbol{x},y}\left[\left(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\right)(h(\boldsymbol{x}) - y)\right]$$

Error decomposition

$(\boldsymbol{x}, y) \sim P$

$h(\boldsymbol{x})$: minimizes the expected loss

$$E_{true}\big(f_{\mathcal{D}}(\boldsymbol{x})\big) = \mathbb{E}_{\boldsymbol{x},y}[(f_{\mathcal{D}}(\boldsymbol{x}) - y)^2]$$

Expected loss

$$= \mathbb{E}_{\boldsymbol{x},y}\,[(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x}) + h(\boldsymbol{x}) - y)^2]$$

$$= \mathbb{E}_{\boldsymbol{x}}\left[\big(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\big)^2\right] + \mathbb{E}_{\boldsymbol{x},y}[(h(\boldsymbol{x}) - \boldsymbol{y})^2]$$
$$+2\mathbb{E}_{\boldsymbol{x},y}\Big[\big(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\big)(h(\boldsymbol{x}) - y)\Big]$$

$$\mathbb{E}_{\boldsymbol{x}}\Big[\big(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\big)\mathbb{E}_{y|\boldsymbol{x}}[(h(\boldsymbol{x}) - y)]\Big]$$

Error decomposition

$(\boldsymbol{x}, y) \sim P$

$h(\boldsymbol{x})$: minimizes the expected loss

$$E_{true}\big(f_{\mathcal{D}}(\boldsymbol{x})\big) = \mathbb{E}_{\boldsymbol{x},y}[(f_{\mathcal{D}}(\boldsymbol{x}) - y)^2]$$

Expected loss

$$= \mathbb{E}_{\boldsymbol{x},y}\,[(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x}) + h(\boldsymbol{x}) - y)^2]$$

$$= \mathbb{E}_{\boldsymbol{x}}\left[\big(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\big)^2\right] + \mathbb{E}_{\boldsymbol{x},y}[(h(\boldsymbol{x}) - \boldsymbol{y})^2]$$
$$+2\mathbb{E}_{\boldsymbol{x},y}\left[\big(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\big)(h(\boldsymbol{x}) - y)\right]$$

$$\mathbb{E}_{\boldsymbol{x}}\left[\big(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\big)\underbrace{\mathbb{E}_{y|\boldsymbol{x}}[(h(\boldsymbol{x}) - y)]}_{0}\right]$$

Error decomposition

$(\boldsymbol{x}, y) \sim P$

$h(\boldsymbol{x})$: minimizes the expected loss

$$E_{true}\big(f_{\mathcal{D}}(\boldsymbol{x})\big) = \mathbb{E}_{\boldsymbol{x},y}[(f_{\mathcal{D}}(\boldsymbol{x}) - y)^2]$$

$$= \mathbb{E}_{\boldsymbol{x},y}\,[(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x}) + h(\boldsymbol{x}) - y)^2]$$

$$= \mathbb{E}_{\boldsymbol{x}}\left[\big(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\big)^2\right] + \underbrace{\mathbb{E}_{\boldsymbol{x},y}[(h(\boldsymbol{x}) - \boldsymbol{y})^2]}_{\text{noise}}$$

$+ 0$

- Noise shows the irreducible minimum value of the loss function

Expectation of true error

$$E_{true}\left(f_{\mathcal{D}}(\boldsymbol{x})\right) = \mathbb{E}_{\boldsymbol{x},y}[(f_{\mathcal{D}}(\boldsymbol{x}) - y)^2]$$
$$= \mathbb{E}_{\boldsymbol{x}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2\right] + noise$$

Expectation of true error

$$E_{true}\left(f_{\mathcal{D}}(\boldsymbol{x})\right) = \mathbb{E}_{\boldsymbol{x},y}[(f_{\mathcal{D}}(\boldsymbol{x}) - y)^2]$$
$$= \mathbb{E}_{\boldsymbol{x}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2\right] + noise$$

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{x}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2\right]\right]$$
$$= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2\right]\right]$$

We now want to focus on $\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2\right]$.

The average hypothesis

$$\bar{f}(\boldsymbol{x}) \equiv E_{\mathcal{D}}[f_{\mathcal{D}}(\boldsymbol{x})]$$

$$\bar{f}(\boldsymbol{x}) \approx \frac{1}{K}\sum_{k=1}^{K} f_{\mathcal{D}^{(k)}}(\boldsymbol{x})$$

K training sets (of size N) sampled from $P(\boldsymbol{x}, y)$: $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \dots, \mathcal{D}^{(K)}$

Using the average hypothesis

$$\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2\right]$$

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

Using the average hypothesis

$$\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2\right]$$

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

$$= \mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\right)^2 + \left(\bar{f}(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2\right.$$

Using the average hypothesis

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

$$= \mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\right)^2 + \left(\bar{f}(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2\right.$$

Bias and variance

$$\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2\right] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\right)^2\right]}_{\mathrm{var}(\boldsymbol{x})} + \underbrace{\left(\bar{f}(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2}_{\mathrm{bias}(\boldsymbol{x})}$$

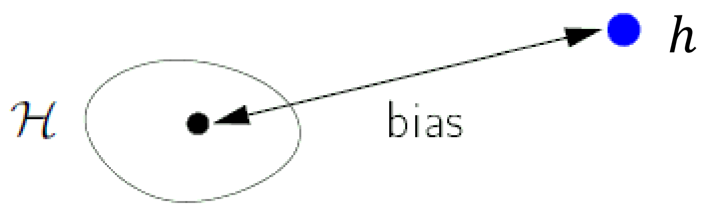
$$\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2\right]\right] = \mathbb{E}_{\boldsymbol{x}}[\mathrm{var}(\boldsymbol{x}) + \mathrm{bias}(\boldsymbol{x})]$$

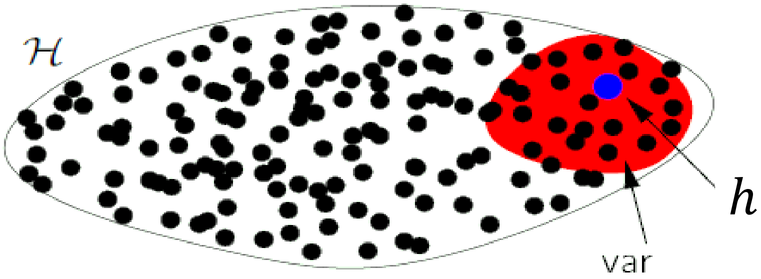
$$= \mathrm{var} + \mathrm{bias}$$

Example: sin target

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

$$\text{bias} = \mathbb{E}_{\boldsymbol{x}}\left[\bar{f}(\boldsymbol{x}) - h(\boldsymbol{x})\right]$$





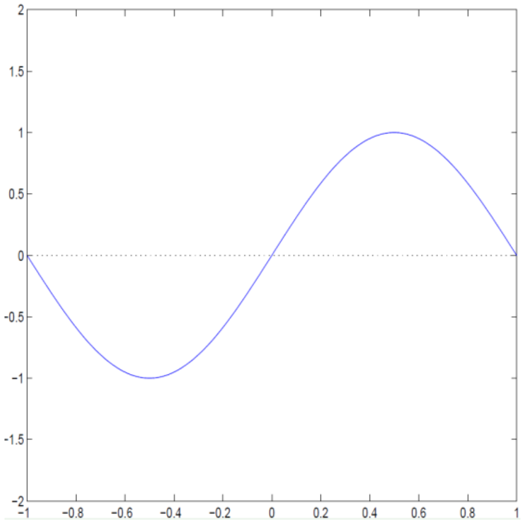
More complex $\mathcal{H}$ $\Rightarrow$ lower bias but higher variance

[Y.S. Abou Mostafa, 2012]

Example: sin target

- Only two training example $N = 2$

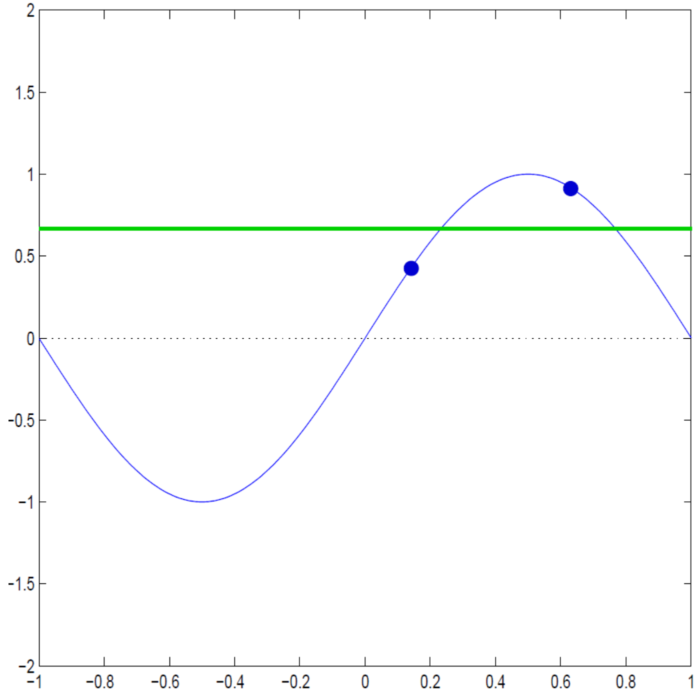
- Two models used for learning:
  - $\mathcal{H}_0: f(x) = b$
  - $\mathcal{H}_1: f(x) = ax + b$



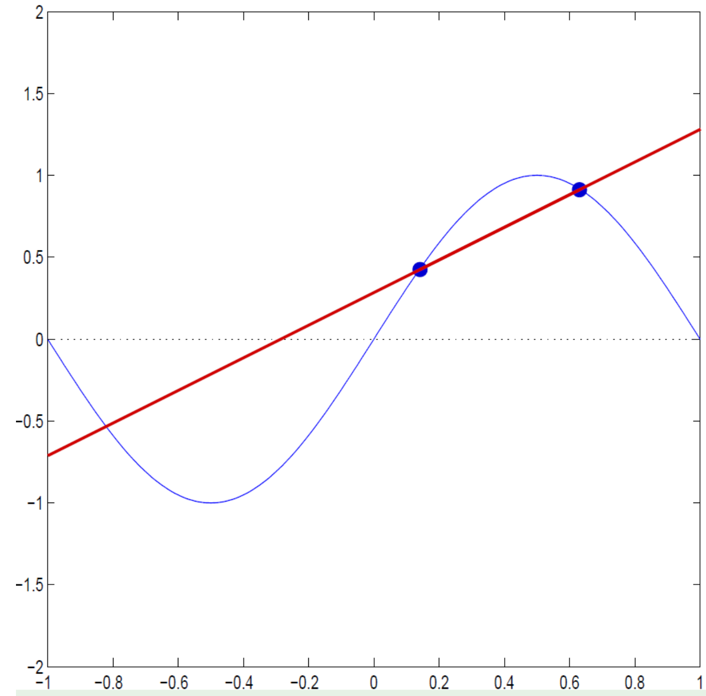
- Which is better $\mathcal{H}_0$ or $\mathcal{H}_1$?

Learning from a training set

$\mathcal{H}_0$

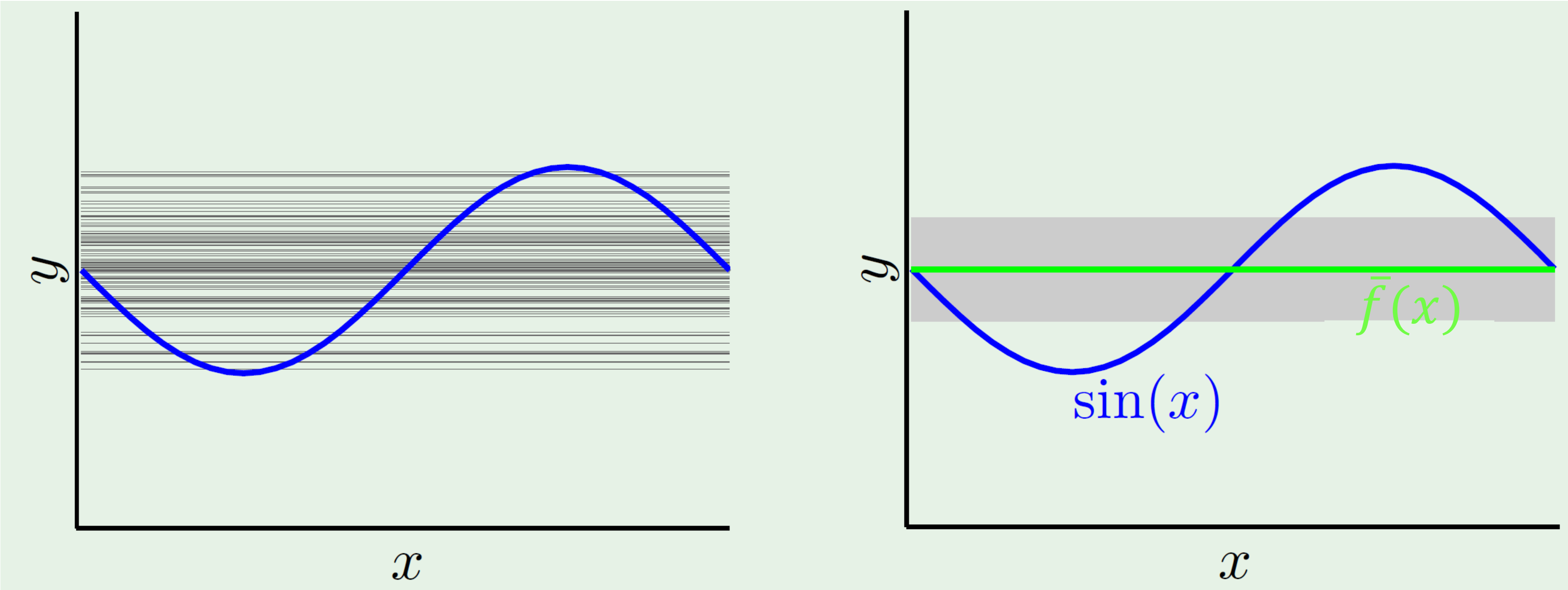


$\mathcal{H}_1$



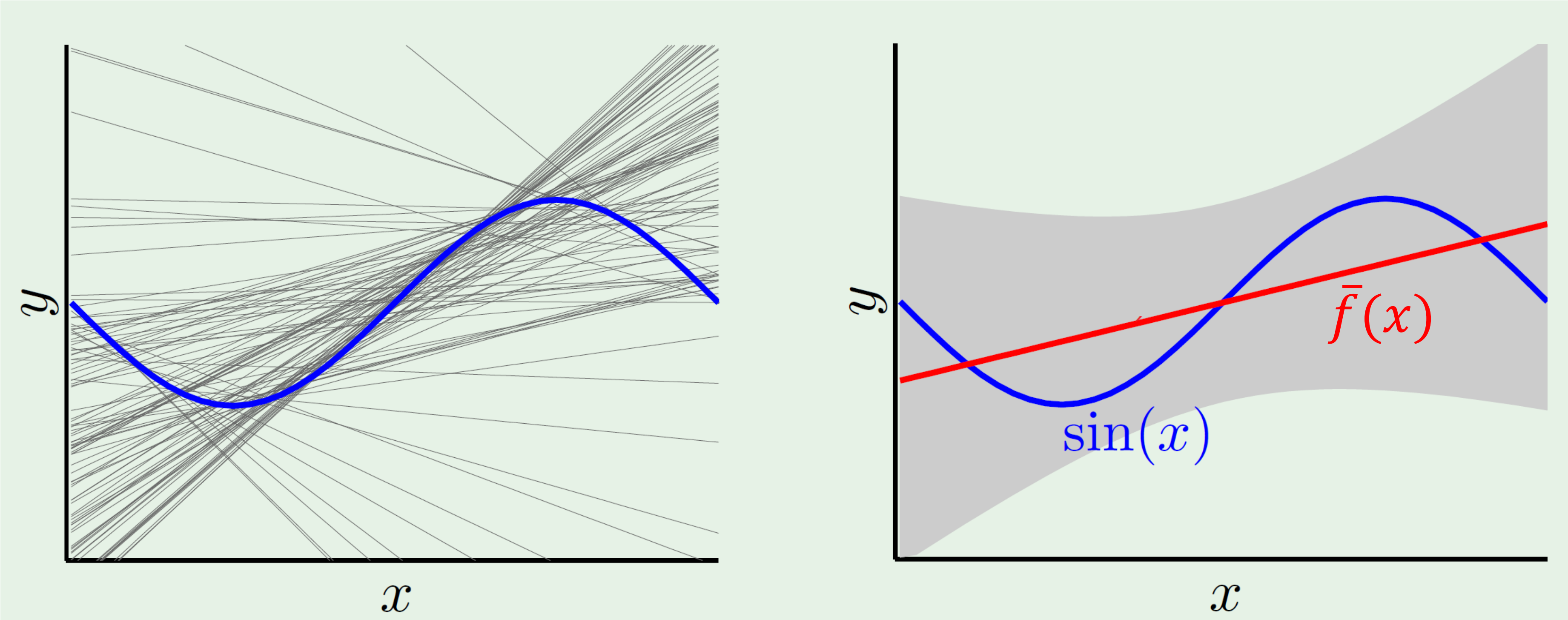
[Y.S. Abou Mostafa, 2012]

Variance $\mathcal{H}_0$



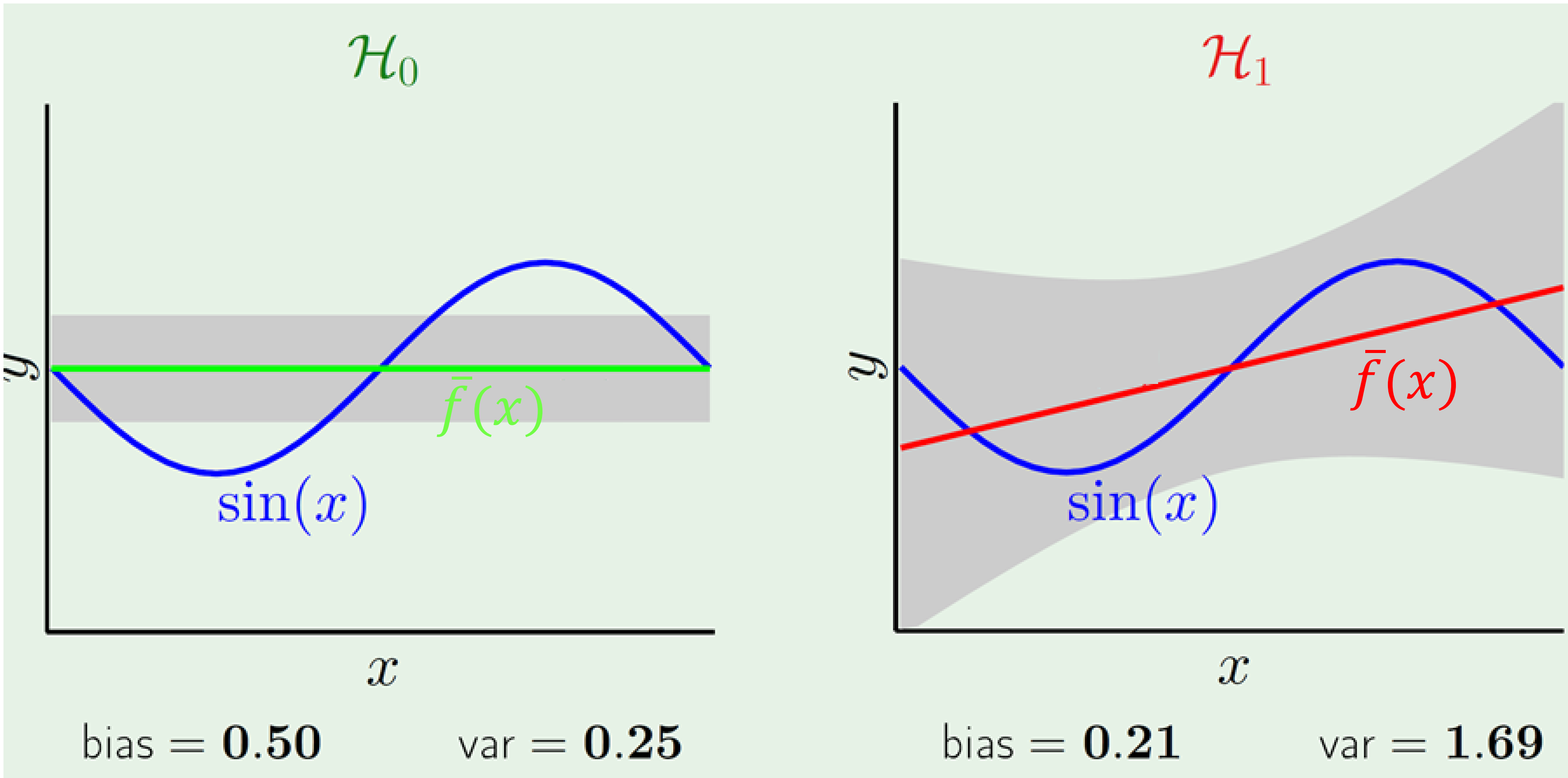
[Y.S. Abou Mostafa, et. al]

Variance $\mathcal{H}_1$



[Y.S. Abou Mostafa, et. al]

Which is better?



[Y.S. Abou Mostafa, 2012]

Lesson

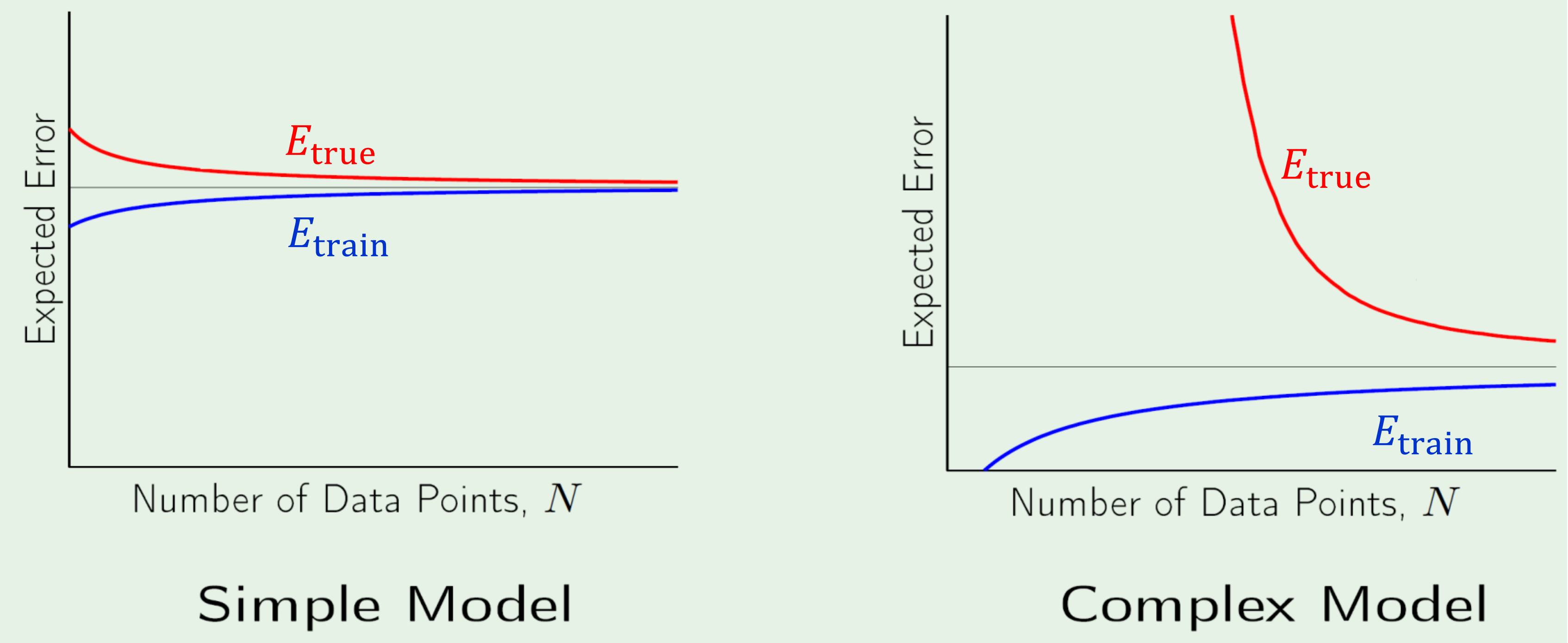
Match the **model complexity**

to the data sources

not to the complexity of the target function.

Expected training and true error curves

- Errors vary with the number of training samples

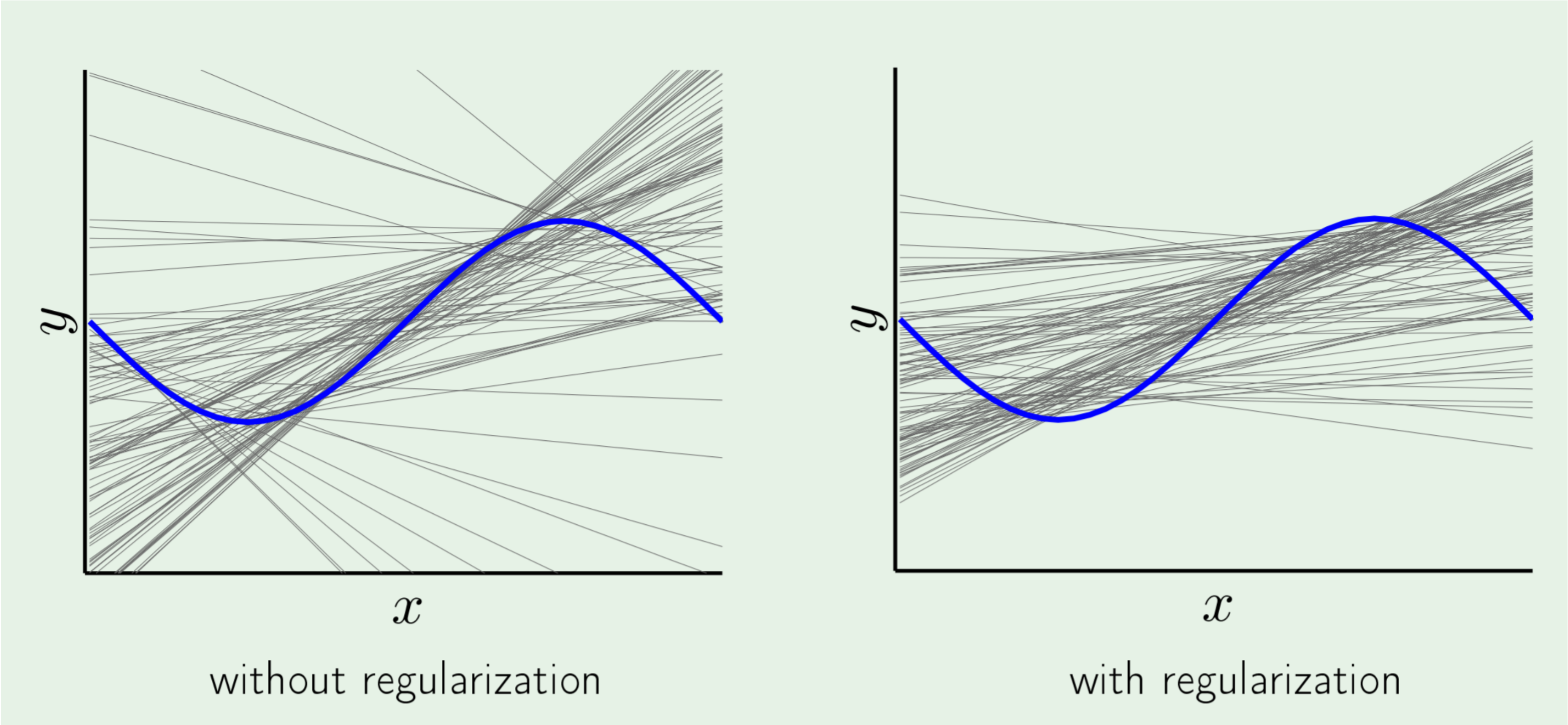


expected true error: $\mathbb{E}_{\mathcal{D}}\left[E_{true}\left(f_{\mathcal{D}}(\boldsymbol{x})\right)\right]$

expected training error: $\mathbb{E}_{\mathcal{D}}\left[E_{train}\left(f_{\mathcal{D}}(\boldsymbol{x})\right)\right]$

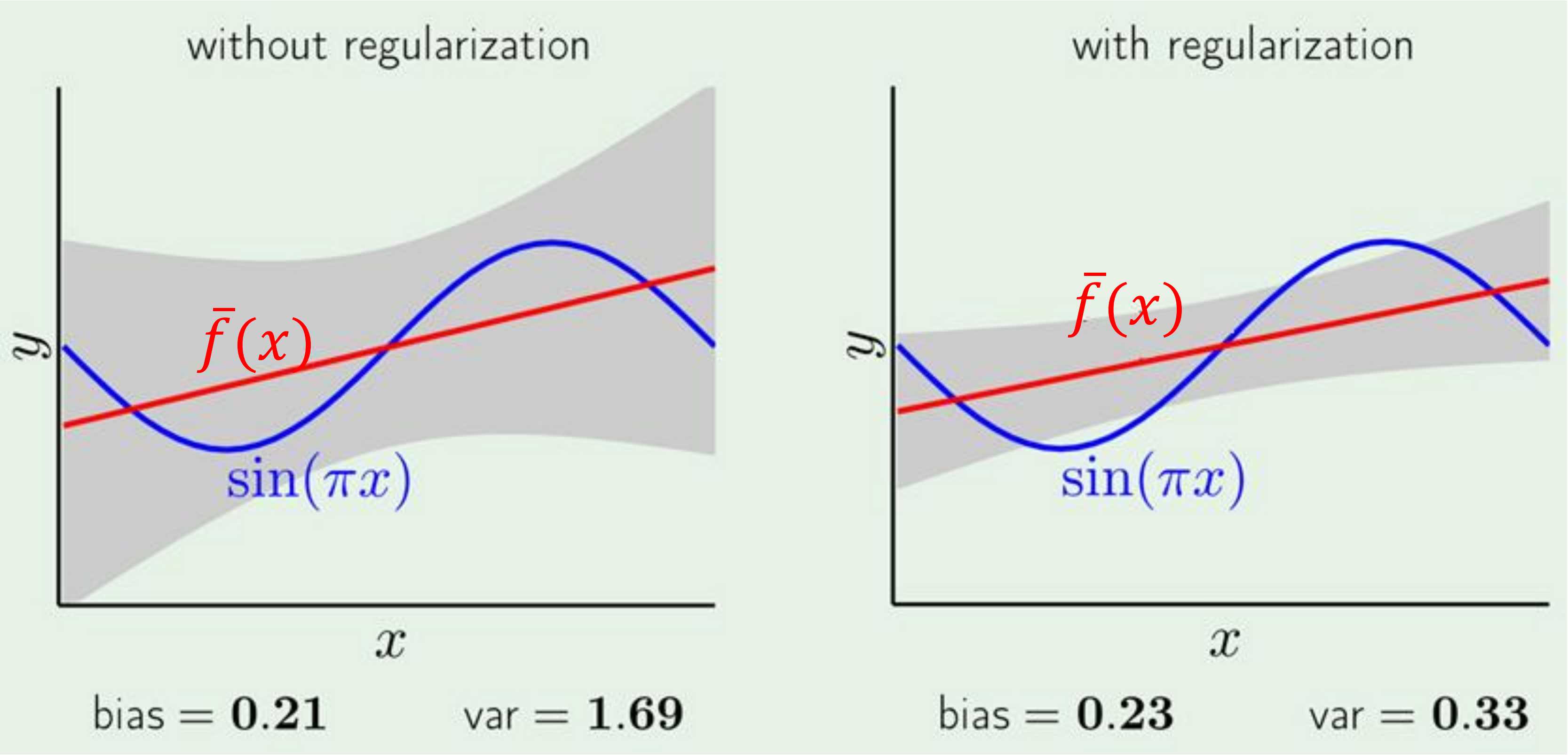
[Y.S. Abou Mostafa, 2012]

Regularization



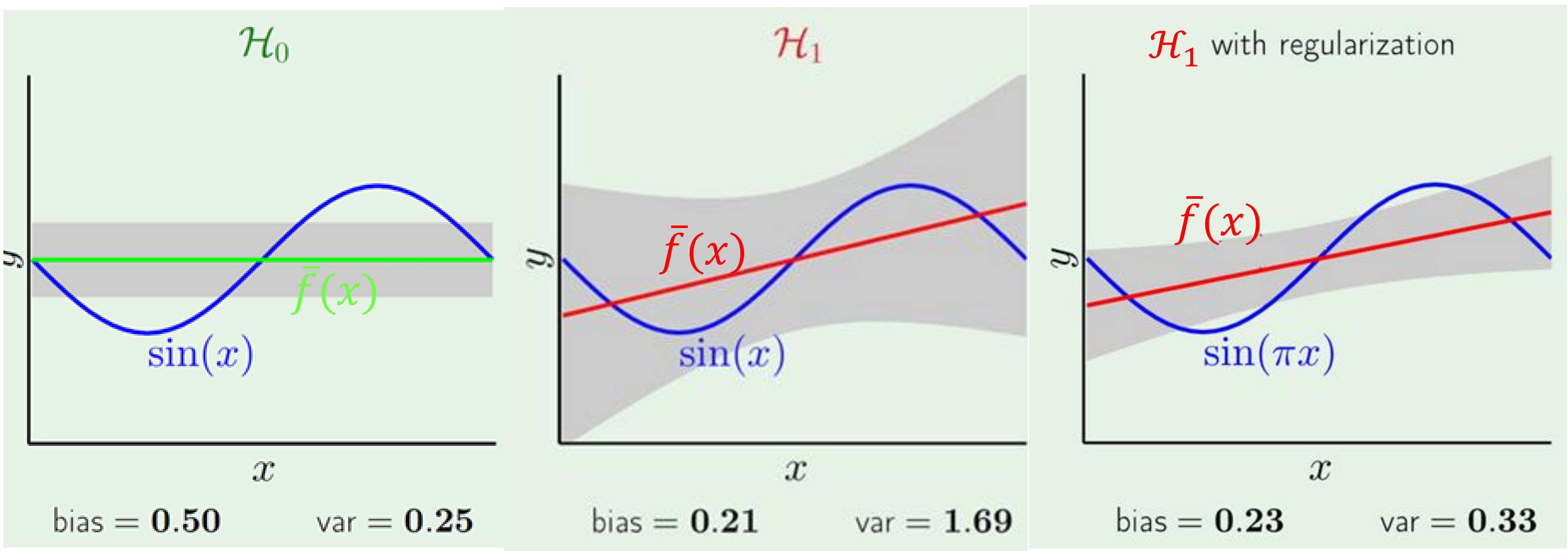
[Y.S. Abou Mostafa, 2012]

Regularization: bias and variance



[Y.S. Abou Mostafa, 2012]

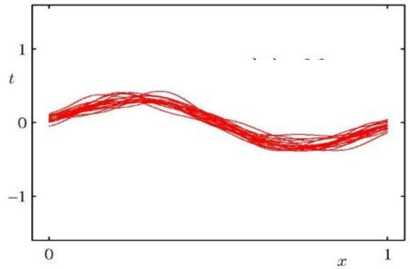
ResourcWinner of $\mathcal{H}_0$, $\mathcal{H}_1$, and $\mathcal{H}_1$ with regularization es

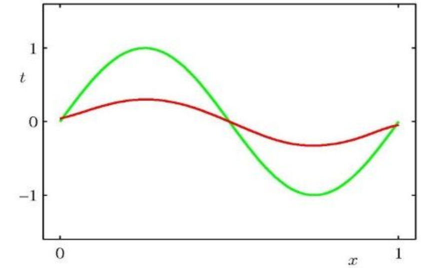


[Y.S. Abou Mostafa, 2012]

Regularization and bias/variance

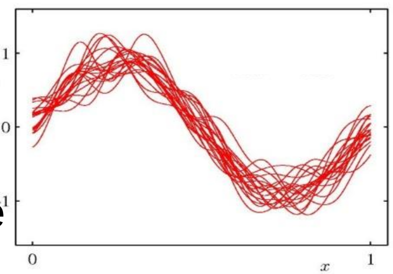
λ is large

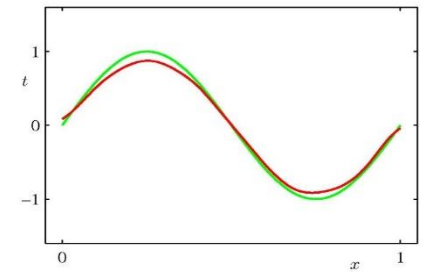




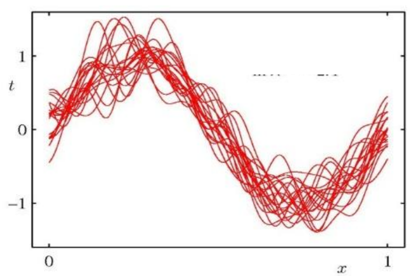
$L = 100$ data sets
$N = 25$
$m = 25$

λ is
intermediate

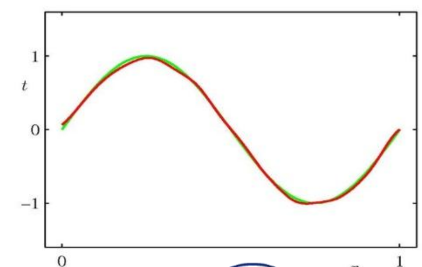




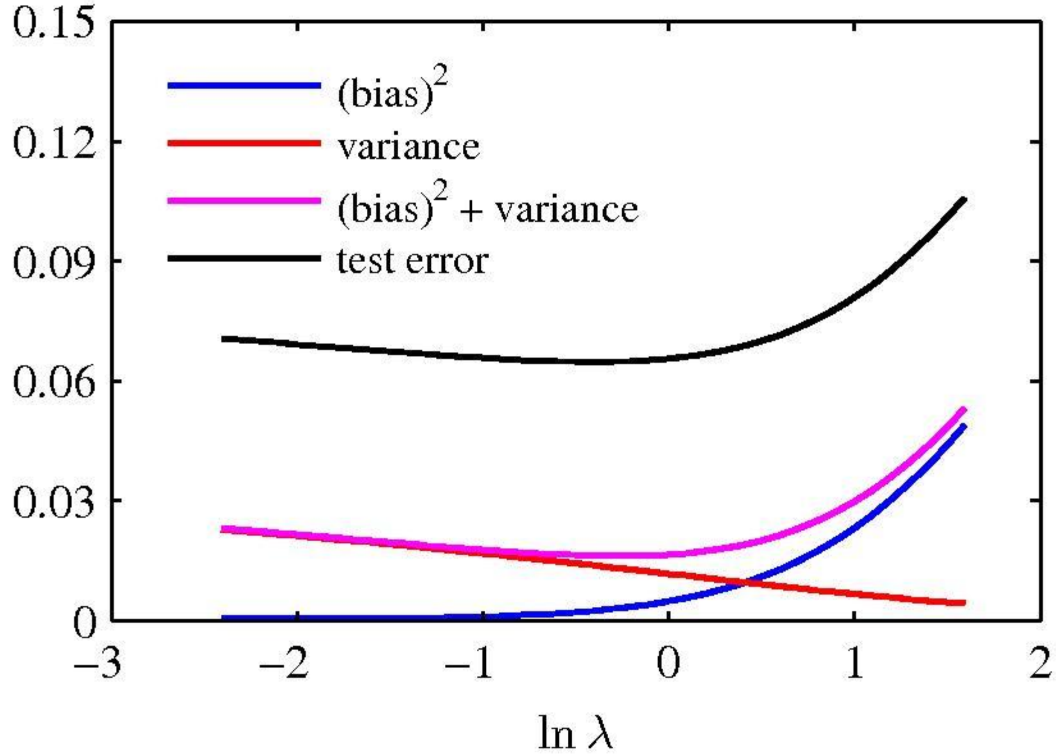
λ is
sma
ll



[Bishop]



Learning curves of bias, variance, and noise



[Bishop]

Bias-variance decomposition: summary

- The noise term is unavoidable.
- The terms we are interested in are bias and variance.
- The approximation-generalization trade-off is seen in the bias-variance decomposition.

Resources

- C. Bishop, “Pattern Recognition and Machine Learning”, Chapter 3.2.
- Yaser S. Abu-Mostafa, Malik Maghdon-Ismail, and Hsuan Tien Lin, “**Learning from Data**”, Chapter 2.3, 3.2, 3.4.
- Course CE-717, Dr. M.Soleyman