

Lecture 4. Learning (I): Approximation Theory

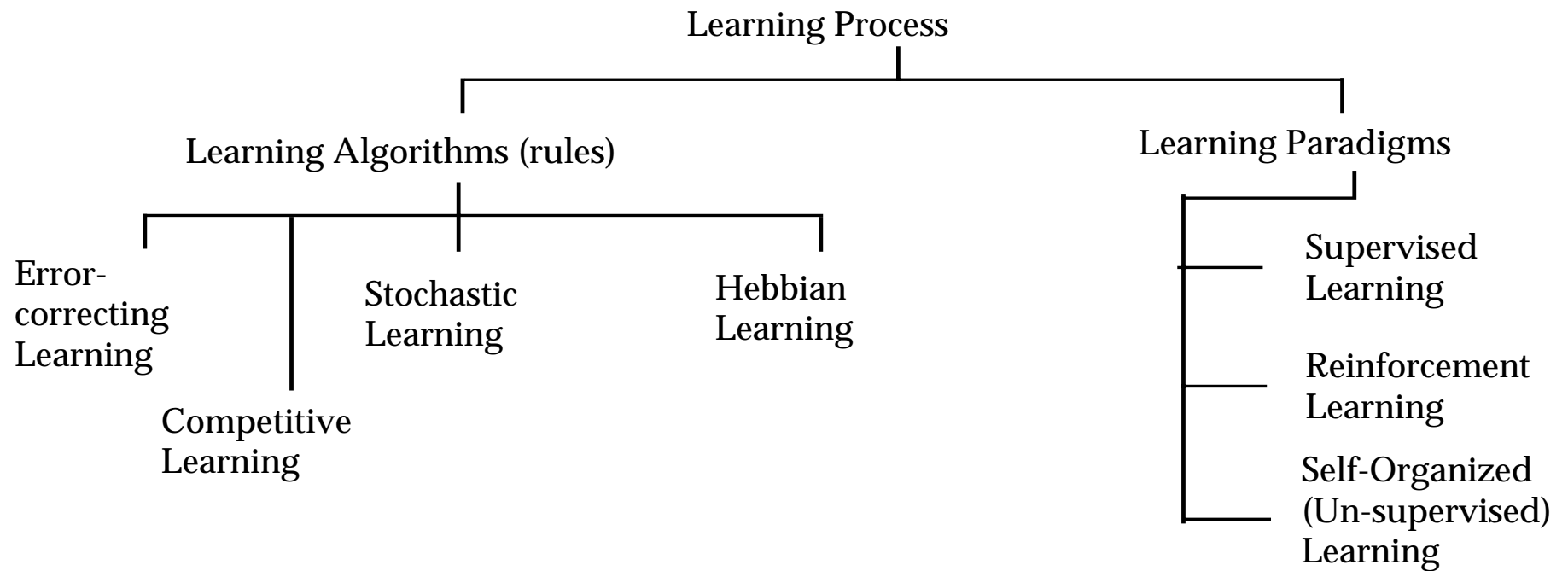
Outline

- What is Learning?
- Supervised Learning and Function Approximation
- Bias and Variance Trade-offs
- Model Selection and Sampling Strategy

What is Learning?

- Learning is a process by which a system modifies its behavior by adjusting its parameters in response to the stimulation by the environment.
- Elements of Learning
 - Stimulation from the environment
 - Parameter adaptation rule
 - Change of behavior of the system

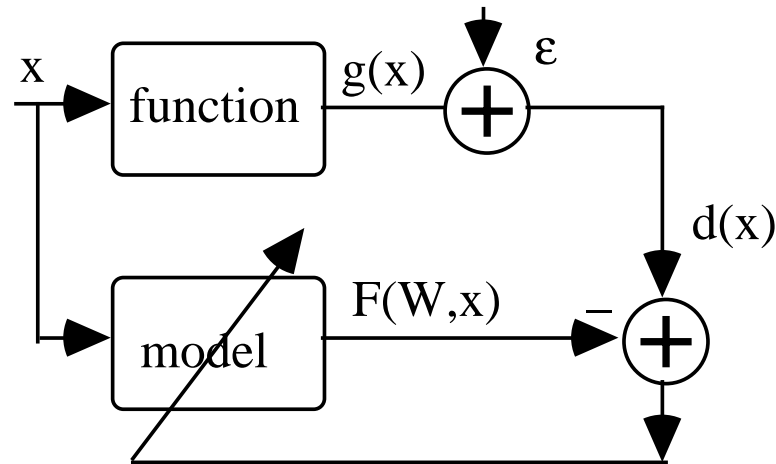
Taxonomy of Learning



Learning Paradigms

- Supervised vs. Un-supervised Learning
- Supervised Learning Rules: A **Teacher** is given.
 - Error-correction learning rule
- Unsupervised Learning Rules: No teacher is available.
 - Hebbian learning rule
 - Stochastic learning rule
 - Competitive learning rule

Learning As Function Approximation



- Observation: $d(x) = g(x) + \varepsilon$
- Model: $F(W, x)$, W : parameters
- Goal: Find W such that $J(w) = E\{\|d(x) - F(W, x)\|^2\}$ is minimized. That is, use a model $F(W, x)$ to approximate an unknown function (mapping) $g(x)$ based on noisy observations.

Learning Model

- ε : observation noise due to measurement error, lack of exact knowledge of $g(x)$, etc. Assume ε is zero mean (i.e. $E\{\varepsilon\} = 0$), and independent of x (i.e. $P\{\varepsilon|x\} = P\{\varepsilon\}$),

$$\begin{aligned} E\{d(x)|x\} &= E\{g(x) + \varepsilon | x\} = E\{g(x)|x\} + E\{\varepsilon|x\} \\ &= g(x) + E\{\varepsilon\} = g(x) \end{aligned}$$

- Also, ε is assumed to be independent of $F(W,x)$. Thus,

$$\begin{aligned} E\{ [d(x) - g(x)][g(x) - F(W,x)]|x\} &= E\{[\varepsilon][g(x) - F(W,x)]|x\} \\ &= E\{\varepsilon|x\} E\{[g(x) - F(W,x)]\} = E\{\varepsilon\} E\{[g(x) - F(W,x)]\} = 0. \end{aligned}$$

Optimality of $F(W,x)$

- Question: We want to use $F(W,x)$ to approximate $g(x)$, not $d(x)$!
But, $J(W)$ minimizes $\|d(x) - F(W,x)\|$?
- $J(W) = E\{\|d(x) - F(W,x)\|^2\} =$
 $= E\{\|d(x) - g(x) + g(x) - F(W,x)\|^2\}$
 $= E\{\|d(x) - g(x)\|^2\} + E\{\|g(x) - F(W,x)\|^2\}$
 $\quad + 2E\{[d(x) - g(x)][g(x) - F(W,x)]\}$
 $= E\{\|d(x) - g(x)\|^2\} + E\{\|g(x) - F(W,x)\|^2\}$
 $= \sigma_\varepsilon^2 + E\{\|g(x) - F(W,x)\|^2\} \geq \sigma_\varepsilon^2.$
- Answer: Since σ_ε^2 is independent of W , the optimal W^* which minimizes $J(W)$ will also minimize $E\{\|g(x) - F(W,x)\|^2\}$

Model Selection Problem

- Assume $F_n(W,x) = w_0 + w_1x + w_2x^2 + \dots + w_nx^n$ is a n-th order polynomial model. The true function $g(x) = g_0 + g_1x + g_2x^2$.
- Since $g(x)$ is unknown, one may choose $F_n(W,x)$ to have different values of n instead of 2. This leads to a wrong model order.
- If there is infinite amount of training samples x , one may try to fit each model order, and if $g(x)$ can be best approximated by a finite order polynomial, then it is possible to find the best model order.
- However, in practice, only finite number of training samples are available. Depending on specific training data set D , one may lead to a specific solution W_D and corresponding cost $J(W_D)$.
- Hence we want to find the $J(W)$ over all possible selection of D and use it to determine which model order to use.

Bias vs. Variance

- Denote
 - $D = \{(x(i), d(i)); 1 \leq i \leq K\}$: an arbitrarily sampled training set out of the population.
 - W_D to be the parameters obtained by training on D .
 - $E_D\{F(W_D, x)\}$: the expected $F(W, x)$ over all possible choices of D .
- Then $J(W) = \sum_{i \in D} \|d(i) - F(W, x(i))\|^2 = \sigma_\varepsilon^2 + E_D\{\|g(x) - F(W, x)\|^2\}$
- Now

$$\begin{aligned}
 & E_D\{\|g(x) - F(W, x)\|^2\} \\
 &= E_D\{\|g(x) - E_D\{F(W, x)\} + E_D\{F(W, x)\} - F(W, x)\|^2\} \\
 &= E_D\{\|g(x) - E_D\{F(W, x)\}\|^2\} + E_D\{\|E_D\{F(W, x)\} - F(W, x)\|^2\} \\
 &\quad + E_D\{[g(x) - E_D\{F(W, x)\}] \cdot [E_D\{F(W, x)\} - F(W, x)]\}
 \end{aligned}$$

Bias vs Variance (2)

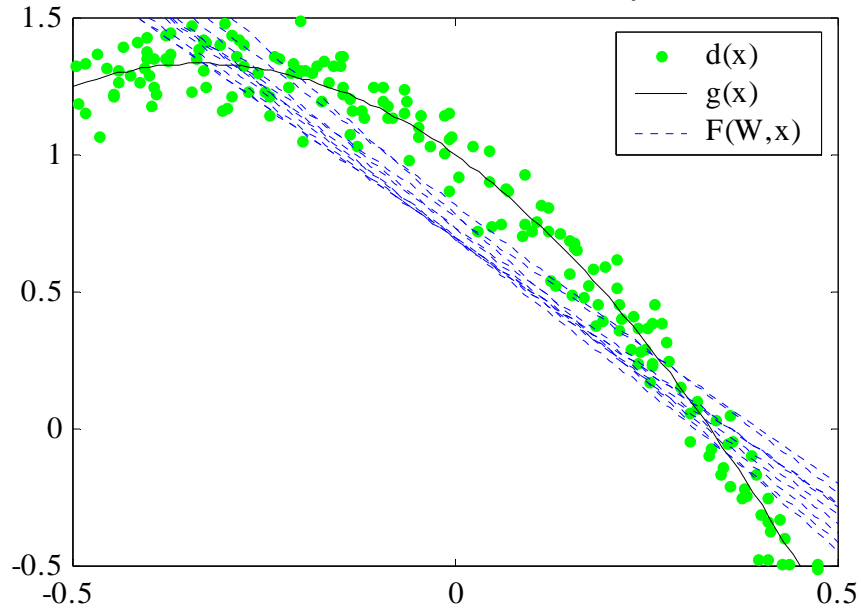
- Since $E_D \{ [g(x) - E_D \{F(W, x)\}] \cdot [E_D \{F(W, x)\} - F(W, x)] \}$
 $= [g(x) - E_D \{F(W, x)\}] \cdot [E_D \{F(W, x)\} - E_D \{F(W, x)\}] = 0$
- Hence $E_D \{ \|g(x) - F(W, x)\|^2 \}$
 $= \underbrace{\|g(x) - E_D \{F(W, x)\}\|^2}_{\text{bias}} + \underbrace{E_D \{ \|E_D \{F(W, x)\} - F(W, x)\|^2 \}}_{\text{variance}}$
- Bias: due to difference between $g(x)$ and the ensemble average of $F(W, x)$ realized on different data set D .
 - Relates more to the modeling error between $F(W, x)$ and $g(x)$.
- Variance: due to the variation of $F(W, x)$ on different data sets D .
 - Relates more to the particular data set D .

An Example

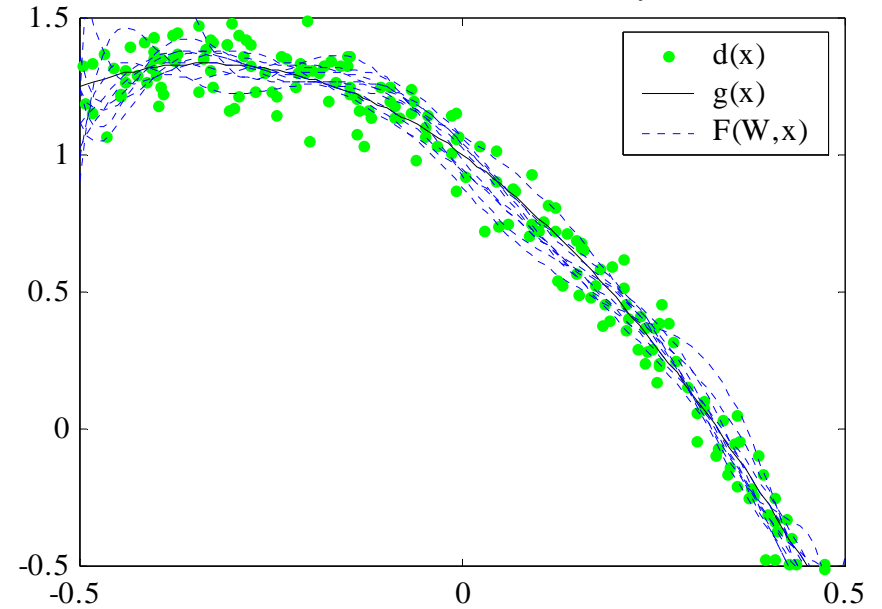
- Consider a true function $g(x) = 1$ for $x \geq 0$; and 0 for $x < 0$.
 $F(W,x) = [1 \ x \ x^2 \ x^3 \ \dots \ x^N]$ W : Nth order polynomial.
 $\varepsilon \sim N(0,0.1)$ — Normal distributed observation noise.
 $d(x(i)) = g(x(i)) + \varepsilon$;
 Population: $\{x(i); i \in [1,200], x(i) \in [-0.5, 0.5], \text{uniformly distributed}\}$
- Experiment:
 - Each run, 10 points (D) are sampled. $F(W,x)$ is computed via least square polynomial fitting using D as a training set.
 - After N runs, compute $E_D\{F(W,x)\}$ by averaging all $F(W,x)$ point-wise over a set of uniform spaced points $\in [-0.5, 0.5]$.
 - Compute bias and variance for different polynomial orders.

Sample Results

Ensemble of $F(W,x)$ over 10 sets of D_s . Polynomial order = 1



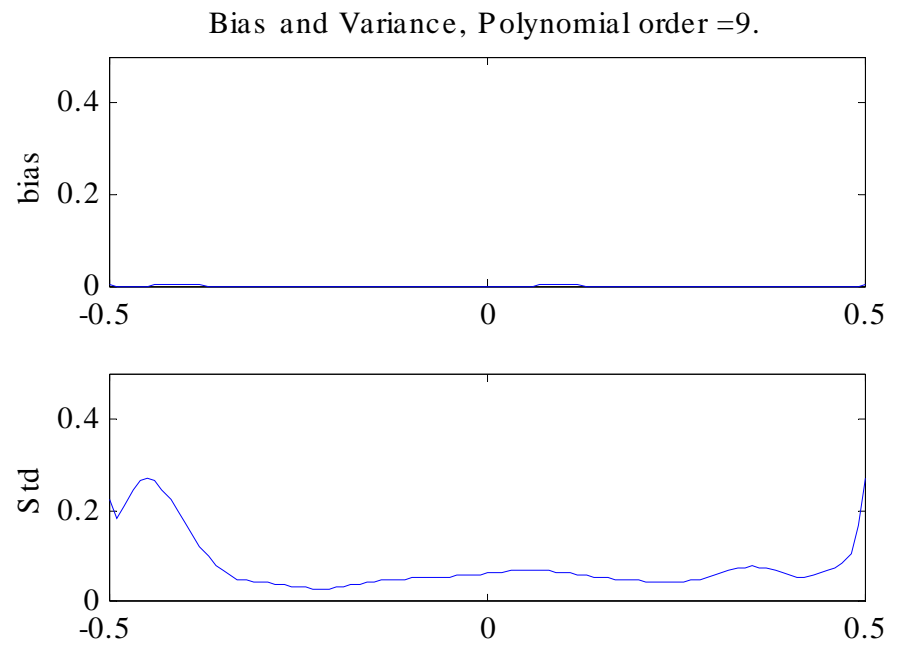
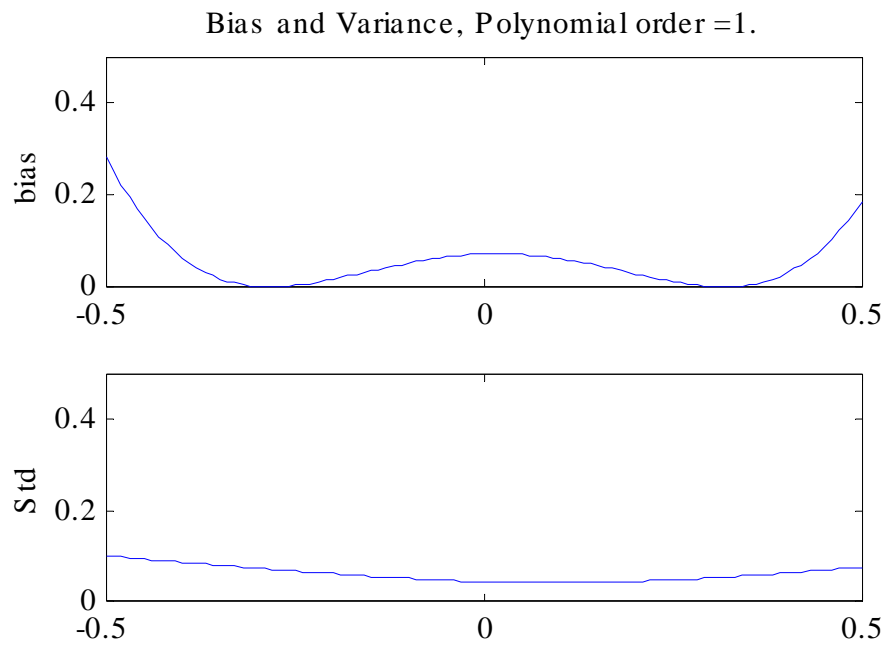
Ensemble of $F(W,x)$ over 10 sets of D_s . Polynomial order = 9



Results are averaged over 10 trials. Sample size $|D| = 30$
 Population size = 200. $\sigma_\varepsilon = 0.1$

Demonstrate Matlab program: [bv.m](#)

Comparing Bias/Variance



When polynomial orders increases from 2 to 9, bias reduces while Variance increases