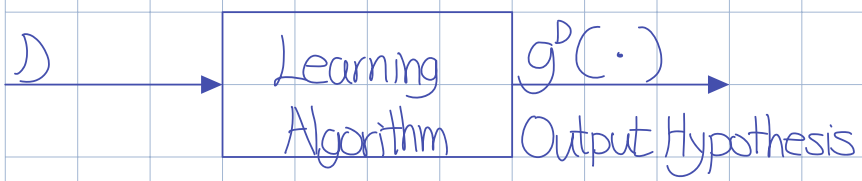
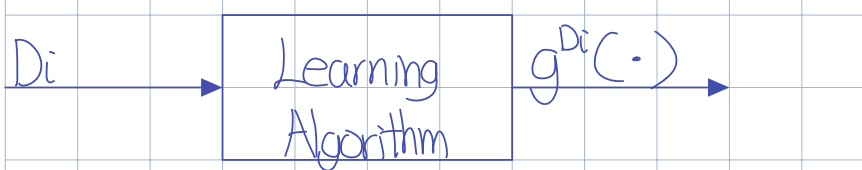
Bias Variance Tradeoff (Chapter 2)

Training set $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$



$D \sim P_D(\cdot)$ Distribution that generate D

Let $D_1, \ldots, D_K$ be K independent dataset drawn from $P_D(\cdot)$



Average hypothesis

$$\frac{1}{K}\sum_{i=1}^{K} g^{D_i}(x)$$

$$\overline{g}(x) = \lim_{K\to\infty} \frac{1}{K}\sum_{i=1}^{K} g^{D_i}(x)$$

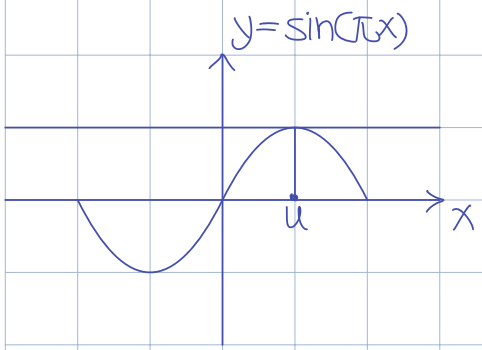
$$= \mathbb{E}_D[g^D(x)]$$

$$= \int_D g^D(x)\, p(x)\, dD$$

Example,

Unknown function $y = f(x) = \sin(\pi x)$

$d = 1$, $x \in \{-1, 1\}$



$\mathcal{D} = \{(u, r)\}$ $N = 1$

$u \sim \text{unif}(-1, 1)$

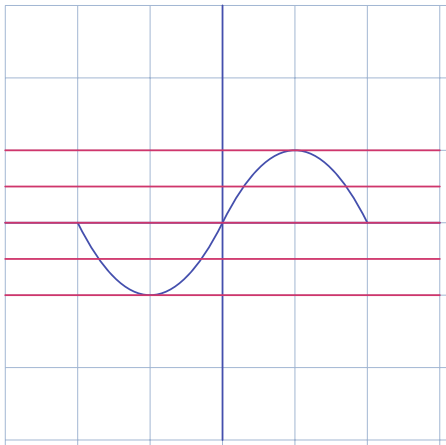
$r = f(u) = \sin(\pi u)$

$\mathcal{H} \equiv$ Set of constant hypothesis

$g^{\mathcal{D}}(x) =$ Constant

Output Hypothesis

$g^{u}(x) = r = \sin(\pi u)$



$$\overline{g}(x) = \mathbb{E}_D(g^D(x))$$
$$= \mathbb{E}_D(\sin(\pi w))$$
$$= \int_{-1}^{1} \sin(\pi w) f(w)\,dw = 0$$

$\overline{g}(x)$ can be interpreted in 2 different ways

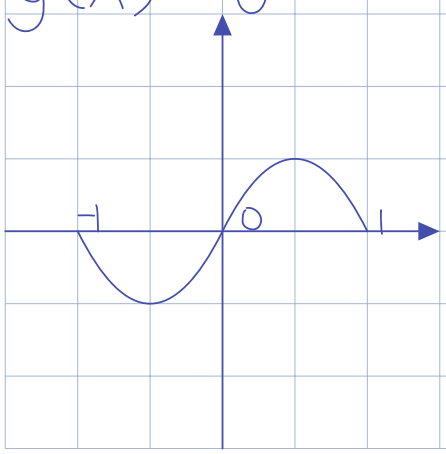
1. Output hypothesis in $\mathcal{H}$ with infinite amount of training data
2. Best approximation to $f(x)$ in $\mathcal{H}$

Bias Function

$$b(x) = (\overline{g}(x) - f(x))^2$$

$f(x) = \sin(\pi x)$ $\quad$ $b(x) = (\sin(\pi x))^2$

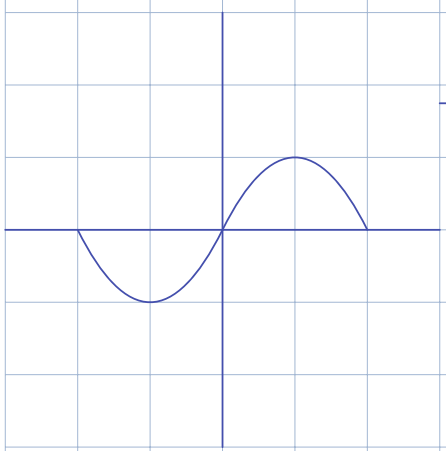
$\overline{g}(x) = 0$



Variance

$$Var(x) = \mathbb{E}_D\left(g^D(x) - \overline{g(x)}\right)^2$$

$$g^D(x) \triangleq g^u(x) = \sin(\pi u)$$



$$\overline{g(x)} = 0$$

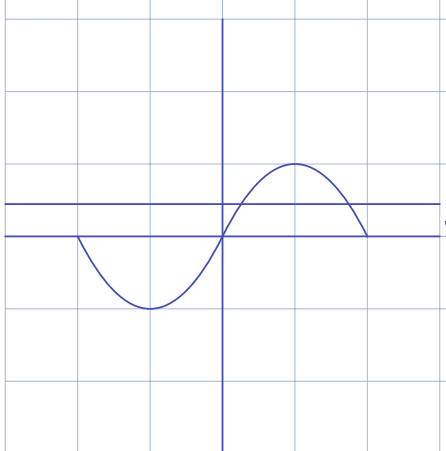
$$Var(x) = \mathbb{E}_u(\sin^2(\pi u))$$

pdf of u

$$= \int_{-1}^{1} \sin^2(\pi u) p(u)\, du$$

$$= \frac{1}{2}\int_{-1}^{1} \sin^2(\pi u)\, du$$

$$= \frac{1}{2}$$



Average test error

$$\triangle(x) = \mathbb{E}_D\left(\left(g^D(x) - f(x)\right)^2\right)$$

$$g^D(x) = g^u(x) = \sin(\pi u)$$
$$f(x) = \sin(\pi x)$$

$$\triangle(x) = \mathbb{E}_u\left[(\sin(\pi u) - \sin(\pi x))^2\right]$$

$$= \mathbb{E}_w[\sin^2(\pi w) + \sin^2(\pi x) - 2\sin(\pi w)\sin(\pi x)]$$

$$= \mathbb{E}_w[\sin^2(\pi w)] + \mathbb{E}_w[\sin^2(\pi x)] - \mathbb{E}_w[2\sin(\pi w)\sin(\pi x)]$$

$$= \frac{1}{2} + \sin^2(\pi x) - 2\sin(\pi x)\mathbb{E}_w[\sin(\pi w)]$$

$$= \frac{1}{2} + \sin^2(\pi x)$$

$$= \text{Var}(x) + \text{bias}(x)$$

Key Observation

Average Test Error $= \underline{\triangle(x)}$

$= \text{bias}(x) + \text{var}(x)$

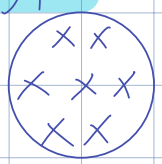
Bias: How well can we approximate $y = f(x)$ using hypothesis class H, complexity of class

Variance: How much $g^D(\underline{x})$ fluctuates around $\overline{g(x)}$ as we vary D

Small H

Large bias small variance

Large H



$f(x)$

Large variance small bias

Recall

$$\triangle(x) = \mathbb{E}_D\left(g^D(x) - f(x)\right)^2$$

$$b(x) = \left(\bar{g}(x) - f(x)\right)^2$$

$$var(x) = \mathbb{E}_D\left(g^D(x) - \bar{g}(x)\right)^2$$

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

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

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

Zero

$$= var(x) + bias(x)$$

$$\overline{E_{out}} = \int \underbrace{\Delta(x)}_{\text{Average of Test Error}} p(x)\,dx$$

$$= \int (\text{bias}(x) + \text{var}(x))\,p(x)\,dx$$

$$= \int \text{bias}(x)\,p(x)\,dx + \text{var}(x)\,p(x)\,dx$$

$$= \overline{\text{bias}(x) + \text{var}(x)}$$

Chapter 4, Validation

$\mathcal{D}$ (N Sample) $\rightarrow$ $\mathcal{D}_{train}$ (N−K Samples)

$\mathcal{D}_{train}$ (N−K Samples) $\rightarrow$ $\mathcal{D}_{val}$ (K Samples)

$\mathcal{D}_{train}$ (N−K Samples) $\rightarrow$ [Learning Algorithm] $\rightarrow$ $\bar{g}(\cdot)$ (Output Hypothesis)

$$E_{val}(\bar{g}) = \frac{1}{K} \sum_{(x_n, y_n) \in \mathcal{D}_{val}} e(\bar{g}(x_n), y_n)$$

How well does $E_{val}(\bar{g})$ approximate $E_{out}(\bar{g}) = \mathbb{E}_x(e(\bar{g}(x), y_n))$ (Test Error)

As $k \longrightarrow \infty$

$E_{val}(g^-) = E_{out}(\overline{g})$

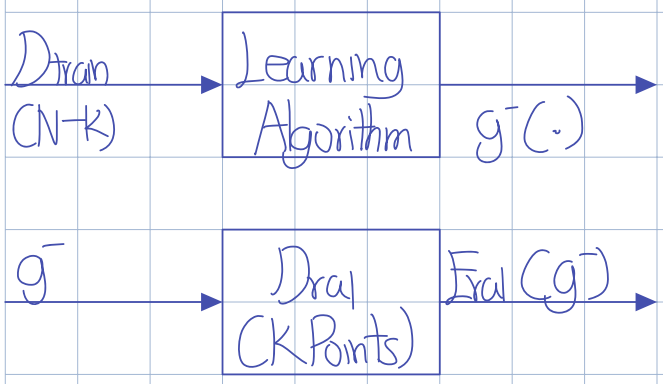
Note

$E_{val}(g^-)$ is random

Since D_{val} is random $\sim P_D(\cdot)$

1. $\mathbb{E}_{D_{val}}[E_{val}(g^-)] = \mathbb{E}_{out}(g^-)$
2. With probability $\geq 1-\delta$

$$E_{out}(\overline{g}) \leq E_{val}(g^-) + \sqrt{\frac{1}{2k} \log \frac{2}{\delta}}$$



$E_{out}(g^-) = \mathbb{E}_{D_{val}}[E_{val}(g^-)]$

Not g^- is fixed hypothesis wrt D_{val}

By Hoeffding Inequality,

$$P(|E_{val}(\bar{g}) - E_{out}(\bar{g})| > \varepsilon) \leq 2e^{-2k\varepsilon^2}$$

$$\delta = 2e^{-2k\varepsilon^2}$$

$$\varepsilon = \sqrt{\frac{1}{2k}\log\frac{2}{\delta}}$$

With probability $\geq 1-\delta$

$$|E_{out}(\bar{g}) - E_{val}(\bar{g})| < \sqrt{\frac{1}{2k}\log\frac{2}{\delta}}$$

$$E_{out}(\bar{g}) \leq E_{val}(\bar{g}) + \sqrt{\frac{1}{2k}\log\frac{2}{\delta}}$$

Proof of Part 1

$$\mathbb{E}_{D_{val}}[E_{val}(\bar{g})] = E_{out}(\bar{g})$$

$$= \mathbb{E}_{x_1,\dots,x_k}\left[\frac{1}{k}\sum_{(x_n,y_n)\in D_{val}} e(\bar{g}(x_n), y_n)\right]$$

$$= \frac{1}{k}\sum_{n=1}^{k} \mathbb{E}_{x_n}[e(\bar{g}(x_n), y_n)]$$

$$E_{out}(\bar{g})$$

$Var(E_{val})$ decrease as $\frac{1}{K}$

Final hypothesis $\underset{(N \text{ training points})}{D} \rightarrow$ [Learning Algo] $\rightarrow g(\cdot)$

$E_{out}(g) \underset{\substack{N-K \\ \text{large}}}{\approx} E_{out}(g^-) \underset{\substack{K \\ \text{large}}}{\approx} E_{val}(g^-)$

$K = \frac{N}{5}$ Good