Recap:

↳ **Growth function:**

$$m_H(N) = \max_{x_1, \cdots x_N \in \mathbb{R}^d} |H(x_1, \cdots x_N)|$$

worst case No. of dich. vectors $\equiv$ effective #. of hypothesis

↳ **VC - dimension:**

to bound $m_H(N)$:

$$d_{vc}(H) = N \quad \text{s.t.} \begin{cases} m_H(N) = 2^N \\ m_H(N+1) < 2^{N+1} \end{cases}$$

i.e. $N+1$ is the first breakpoint

e.g. Linear classifier $d = 2$

$m_H(3) = 8 = 2^3$, $m_H(4) = 14 < 2^4$

$\therefore d_{vc}(H) = 3$

↳ **Theorem:**

For any hypo. set H with $d_{vc}(H) < \infty$,

① $m_H(N) \leq \sum_{i=0}^{d_{vc}(H)} \binom{n}{i} \leq N^{d_{vc}(H)} + 1$

② $\Pr\{\Delta(g) > \varepsilon\} \leq 4\, m_H(2N)\, e^{-\frac{1}{8} N \varepsilon^2}$

Recall: $\underset{\uparrow}{\Delta(g)} = |E_{in}(g) - E_{out}(g)|$

generalization error

// 用处: previously we do $m_H(N) \leq 2^N$, exponential in N

Not the best way to bound

But above is poly. bound

↳ Note:

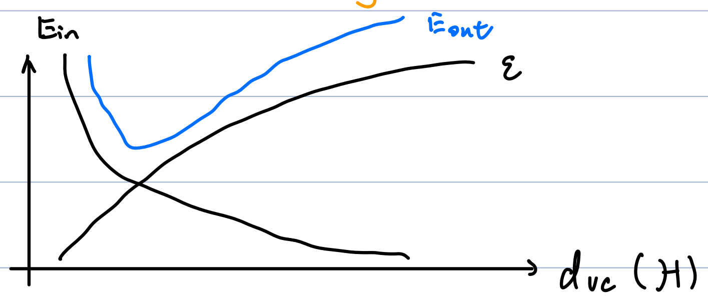
set $\delta = 4\, m_{\mathcal{H}}(2N)\, e^{-\frac{1}{8}N\varepsilon^2}$, with prob. $1-\delta$

$$E_{out}(g) \le E_{in}(g) + \varepsilon$$

$$\therefore\ \varepsilon = \sqrt{\frac{8}{N}\log\frac{4\,m_{\mathcal{H}}(2N)}{\delta}} \le \sqrt{\frac{8}{N}\log\frac{4\left((2N)^{d_{vc}(\mathcal{H})}+1\right)}{\delta}}$$

$$= \theta\left(\sqrt{\frac{d_{vc}(\mathcal{H})}{N}\log\frac{N}{\delta}}\right)$$

scaling



e.g.1. Linear classifier in $\mathbb{R}^d$

$d_{vc}(\mathcal{H}) = d+1$

e.g.2. 2^{nd} order model

$\underline{x} = [1, x_1, x_2]$ $\quad d=2$

Transform to new space:

$\underline{z} = [1, x_1, x_2, x_1x_2, x_1^2, x_2^2]$ $\quad d=5$

lin. classification in $\underline{z}$ space $\quad d_{vc}(\mathcal{H}) = 6$

↳ In general, $\underline{x} \in \mathbb{R}^d$, k^{th} order polynomial

$$d_{vc}(\mathcal{H}) = \binom{k+d}{d} = O(k^d)$$

↳ In practice, a rule of thumb is to choose $\mathcal{H}$ s.t.

$d_{vc}(\mathcal{H}) \approx \frac{N}{10}$

Generalization Bound in Regression

1. Squared error:

↳ $E_{in}(g) = \frac{1}{N} \sum_{n=1}^{N} \left(y_n - g(\underline{x}_n) \right)^2$

$E_{out}(g) = \mathbb{E}\left[(y - g(\underline{x}_n)^2 \right]$

↳ We use the bias-variance trade off.

2. Bias-variance trade off (ch. 2.3)

Today: setup for learning model　　　　明: trade off.

↳ training set:

$\mathcal{D} = \{ (\underline{x}_1, y_1) \cdots (\underline{x}_N, y_N) \}$

$\sim P(\mathcal{D}) = P(\underline{x}_1)\, P(\underline{x}_2) \cdots P(\underline{x}_N)$

↳ Output hypothesis: g^D

放一个 D to tell that we have train the hypo. w/ data.

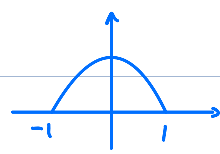
$\hat{y} = g^D(\underline{x})$

output label

↳ Average hypothesis:

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

e.g. $d = 1$, $x \in [-1, 1]$



$y = \sin(\pi x)$ // unknown target func.

$\mathcal{D} = \{ (u, v) \}$, $u \sim \mathcal{U}(-1, 1)$ ← $p(x)$

$\therefore v = \sin(\pi u)$

Assume $\mathcal{H} \equiv$ constant hypothesis

(meaning that whatever the model is, the output is the same)

$\therefore g^D(x) = v$, for any x.

$$\therefore \bar{g} = \mathbb{E}_u[g^D(x)]$$

Avg. Hypo.

$$= \mathbb{E}_u[v]$$

$$= \mathbb{E}_u[\sin(\pi u)]$$

$$= 0$$

↳ "best" approx. to $f(\underline{x})$ given infinite amount of data

But cannot find $\bar{g}$ in practice.

↳ <u>Def</u>:

<u>Bias of learning model</u> for any x:

$$\text{bias}(\underline{x}) = (\bar{g}(\underline{x}) - f(\underline{x}))^2$$

<u>Variance</u> of 〃 〃 〃 〃

$$\text{var}(\underline{x}) = \mathbb{E}_D[(g^D(\underline{x}) - \bar{g}(\underline{x}))^2]$$

e.g. In the prev example.

$$\text{bias}(\underline{x}) = (0 - \sin(\pi x))^2 = \sin^2(\pi x)$$

$$\text{var}(\underline{x}) = \mathbb{E}_u[(v - 0)^2]$$

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

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

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