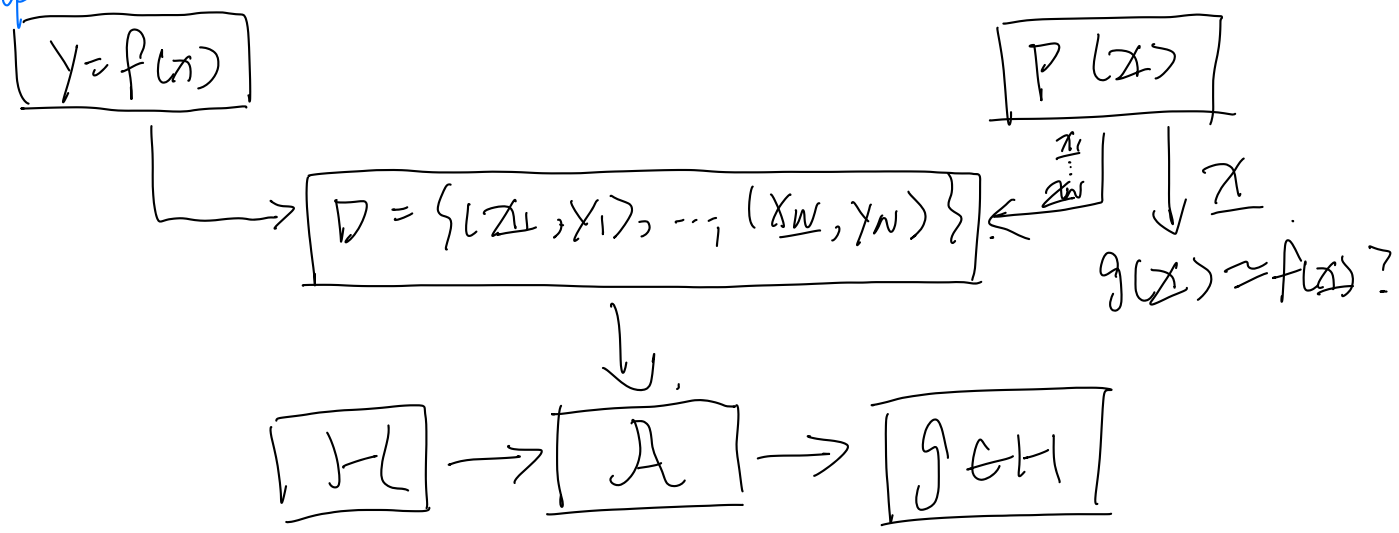
Lec 11 PAC Learning

① Recap:



For any given hypothesis $g \in \mathcal{H}$.

we define $e_n = e(y_n, g(x_n))$

$= 1\{y_n \neq g(x_n)\}$.

$E_{in}(g) = \frac{1}{N}\left(\sum_{n=1}^{N} 1\{y_n \neq g(x_n)\}\right)$ ← R.V.

└ avg of a seq of Bernoulli R.V.s

$E_{out}(g) = \mathbb{E}[1\{y \neq g(x)\}]$ ↕ same RV, same statistics, same distribution

└ expected val of Bernoulli R.V.s.

or expectation of the sequence.

Same distribution for in-sample & out-sample Bernoulli RVs.

$E_{in}(g) \simeq E_{out}(g)$.

↓ min. in training. ↓ care about in real-world.

Goal: $Pr\{E_{in}(g) \simeq E_{out}(g)\} = 1$ "PAC".

② Sequence of iid Bernoulli RVs.

e.g. Consider an urn containing infinitely many balls, each being either red or green.

Let μ be the fraction of red balls (aka %)

$$Z = \begin{cases} 1, & \text{if red ball is drawn} \\ 0, & \text{if green ball is drawn.} \end{cases}$$

$Z \sim$ Bernoulli(μ).

$$\Pr\{Z=1\} = E(Z) = \mu.$$

Suppose we want to estimate μ.

Randomly draw N balls from the urn (∵ infinitely many balls.)

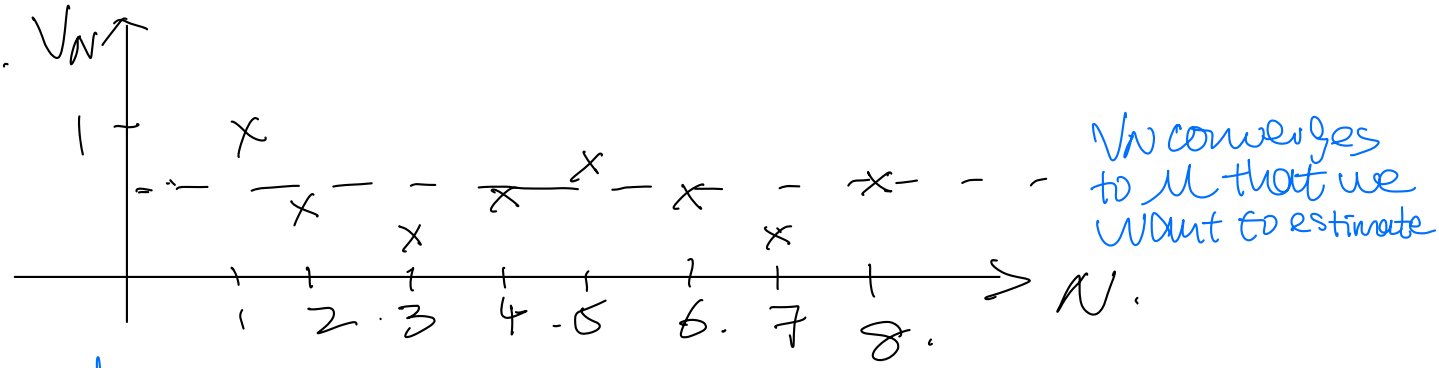
$Z_1, Z_2, \ldots, Z_N$ iid Bernoulli(μ).

e.g., Output sequence $(1, 0, 0, 1, 1, 1, 0, 0, 1)$

Define $V_N = \frac{1}{N}\sum_{n=1}^{N} Z_n$. (proportion of red balls).

IS $V_N \simeq \mu$? (same Q as $E_{in} \simeq E_{out}$?)

e.g. V_N



V_N converges to μ that we want to estimate

Note.. V_N is a R.V.

$$E[V_N] = E\left[\frac{1}{N}\sum_{n=1}^{N} Z_n\right] = \mu.$$ "unbiased estimator"

$$VAR[V_N] = VAR\left[\frac{1}{N}\sum_{n=1}^{N} Z_n\right]$$

$$\overset{\text{indep}}{=} \frac{1}{N^2}\sum_{n=1}^{N} VAR(Z_n)$$

$$= \frac{1}{N^2} \cdot N(\mu(1-\mu))$$

$$= \frac{1}{N}\mu(1-\mu)$$

∴ More balls ⇒ variance → 0.
Meaning: RV not longer random. "consistent estimator".
$N \to \infty \Rightarrow VAR \to 0$. (converged).

$\Rightarrow V_N \to \mu$ as $N \to \infty$. (in the mean squared sense)

(3) Chebyshev Inequality. (Not enough!)

- It tells how far away RV can be from mean val.

$$Pr(|V_N - \mu| > \varepsilon) \leq \frac{VAR(V_N)}{\varepsilon^2}$$

"upper bounded"

In Bernoulli'en case:

$$Pr = \frac{\mu(1-\mu)}{N\varepsilon^2}$$

Notice: Pr goes to 0 as $N \to \infty$ $\forall \varepsilon > 0$.

"We can get as close as we want if we have enough samples".

Can do better!

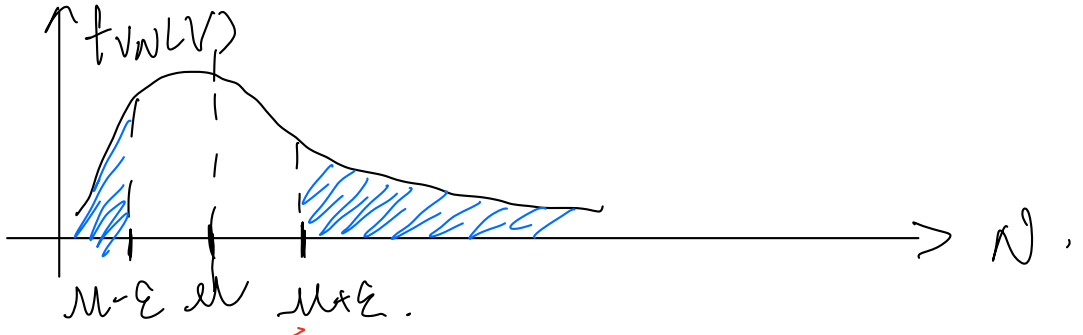
(4) Hoeffding's Inequality: (1963) ⇒ Stronger bound.

"tighter bound".

$$Pr(|V_N - \mu| > \varepsilon) \leq 2e^{-2N\varepsilon^2}.$$

This is a exponential decrease in N.

Confidence Interval



Confidence Interval (ε-conf Interval around μ val)

- $\Pr\{|\nu - \mu| > \varepsilon\}$ = area of shaded (tail) region.
  $\triangleq \delta$.
- $1 - \delta$ = area inside confidence interval
  └ "confidence level".

Consider upper bound:

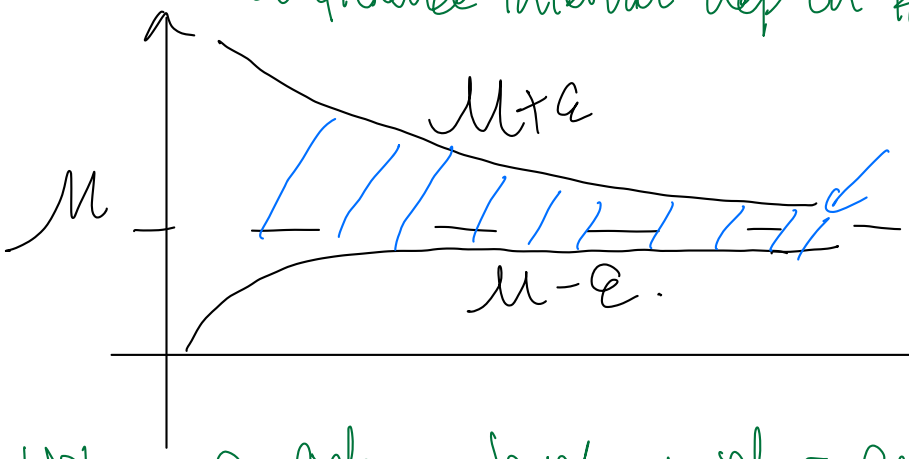
$$\delta = 2e^{-2N\varepsilon^2}$$

$$\Leftrightarrow \varepsilon = \sqrt{\frac{1}{2N} \log \frac{2}{\delta}}.$$

e.g., $\delta = 0.01$: what to be 99% sure / confident.

$$\varepsilon = \sqrt{\frac{1}{2N} \log 200} \simeq \frac{1.63}{\sqrt{N}}$$

└ confidence interval dep on # sample points.



Confidence Interval.
→ converges to μ as $N \uparrow$.
└ 99% confidents.

Higher confidence level, wider confidence interval.
⇒ dep on how much error to tolerate.
∴ Big data ≡ confidents.

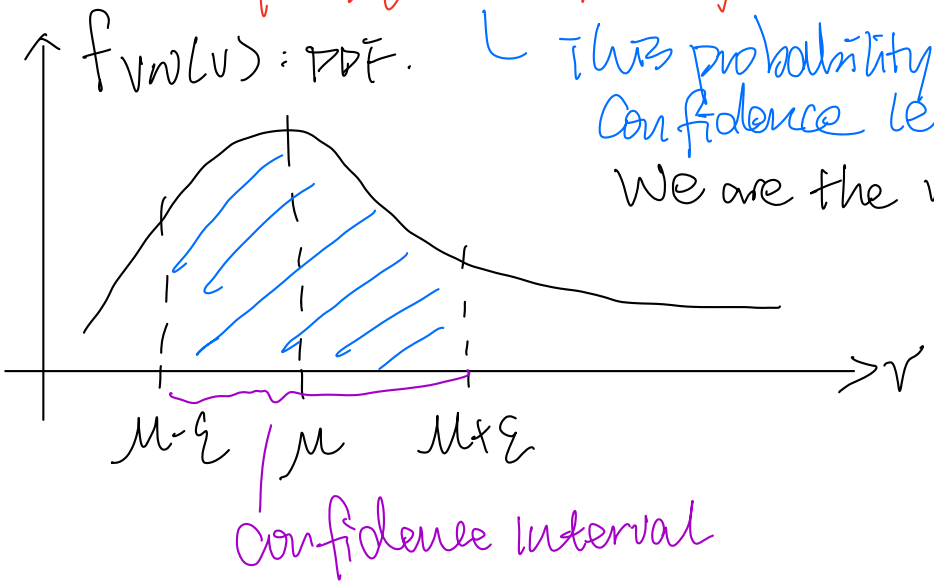
Lec32 PAC learning & Confidence Interval

① Recap:

For iid Bernoulli RVs $Z_1, Z_2, \dots, Z_N$ w/ $Pr\{Z\} = E[Z] = \mu$ & $V_N = \frac{1}{N}\sum_{n=1}^{N} Z_n$,

we have: $Pr\{|V_N - \mu| > \varepsilon\} \le 2e^{-2N\varepsilon^2}$

└ this probability is the confidence level: how sure we are the val is w/in the region?



② Back to PAC learning:

$E_{in}(h) = \frac{1}{N}\sum_{n=1}^{N} 1(y_n \neq h(x_n))$

$E_{out}(h) = E(1(y \neq h(x)))$.

Let $Z = 1(y \neq h(x))$.

$Z_n = 1(y_n \neq h_n(x_n))$, $n = 1, 2, \dots, N$.

For any given $h \in \mathcal{H}$ (indep of training dataset D)

$Z, Z_1, Z_2, \dots, Z_N$ are iid Bernoulli RVs.

$E_{in}(h) = V_N$

$E_{out}(h) = \mu$.

Hoeffding Inequality:

$\Rightarrow Pr\{|E_{in}(h) - E_{out}(h)| > \varepsilon\} \le 2e^{-2N\varepsilon^2}$ $\forall \varepsilon > 0$.

Let $\delta = 2e^{-2N\varepsilon^2}$

Then w/ probability $\geq 1-\delta$,

$$|E_{in}(h) - E_{out}(h)| \leq \varepsilon = \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}.$$

$$\Rightarrow \boxed{E_{out}(h) \leq E_{in}(h) + \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}.}$$

This creates an upper bound for test error

→ This is not a deterministic bound, but a probabilistic bound.
→ Not getting it for granted.
→ There's always a trade-off.

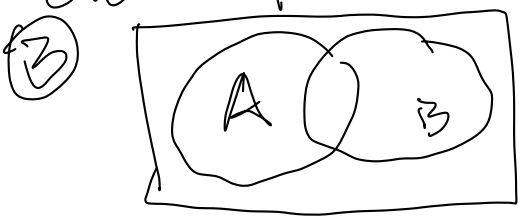
i.i.d Bernoulli RV only if hypothesis h is indep of dataset
if we choose h based on dataset, these properties & straightforward analysis no longer apply!

Trouble! This only works for given h **before** you observe $\mathcal{D}$ (indep of $\mathcal{D}$)

Our final hypothesis g is chosen based on $\mathcal{D}$ after observing $\mathcal{D}$.

$\Rightarrow g(x_1), \dots, g(x_N)$ are **not** indep.

One useful thm: **Union bound**:



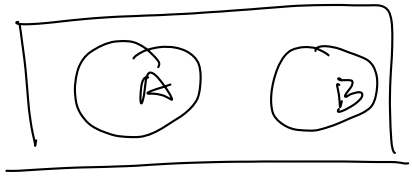
$$P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}.$$
$$\leq P\{A\} + P\{B\}.$$

- For arbitrary events $A_1, A_2, \dots, A_M$ (M: # hypothesis)

$$Pr\{A_1 \cup \dots \cup A_M\} \leq \sum_{i=1}^{M} P\{A_i\}.$$

w/ equality iff $A_1, \dots, A_M$ are mutually exclusive ("disjoint")

e.g.,



Since $g \in \mathcal{H} = \{h_1, h_2, \cdots, h_M\}$, the event that:

$$\{|E_{in}(g) - E_{out}(g)| > \varepsilon\} \leq \bigcup_{i=1}^{M} \{|E_{in}(h_i) - E_{out}(h_i)| > \varepsilon\}$$

g is one of the hypothesis
g is subset of all events.

event: set of outcomes in a probability space.
g is a subset of event.

probability of subset always $\leq$ probability of set.

$\Rightarrow$ the hypothesis we chose.

$$Pr\{|E_{in}(g) - E_{out}(g)| > \varepsilon\} \leq Pr\left(\bigcup_{i=1}^{M} \{|E_{in}(h_i) - E_{out}(h_i)| > \varepsilon\}\right)$$

union bound $\leq \sum_{i=1}^{M} Pr\{|E_{in}(h_i) - E_{out}(h_i)| > \varepsilon\}$.

$\leq \sum_{i=1}^{M} 2e^{-2N\varepsilon^2} = 2Me^{-2N\varepsilon^2}$.

h_i is indep of data prior to training.

$\therefore$ To get bound for final hypothesis we choose, we times M to Hoeffding's inequality

Note ① if $g \in \mathcal{H}$ is selected after observing D (data snooping), then Hoeffding bound is weakened by a factor of M.

② $\delta = 2Me^{-2N\varepsilon^2}$

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

④ Generalization Error:

Def: difference b/w E_{in} & E_{out}.

$$\Delta(g) = |E_{in}(g) - E_{out}(g)|$$

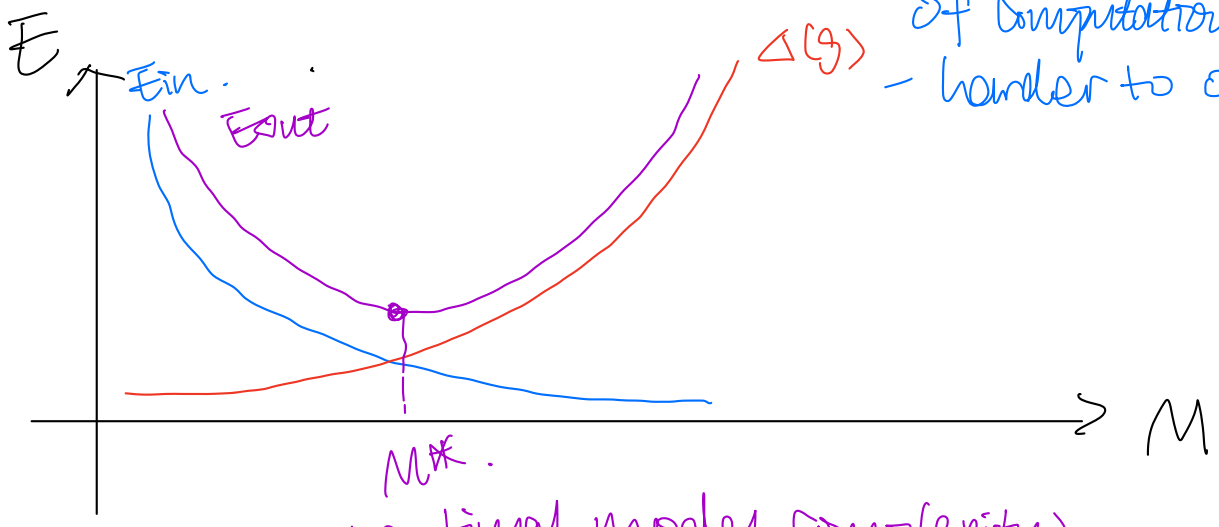
w/ probability $1 - \delta$

$$\Delta(g) \leq \sqrt{\frac{1}{2N} \log \frac{2M}{\delta}}$$

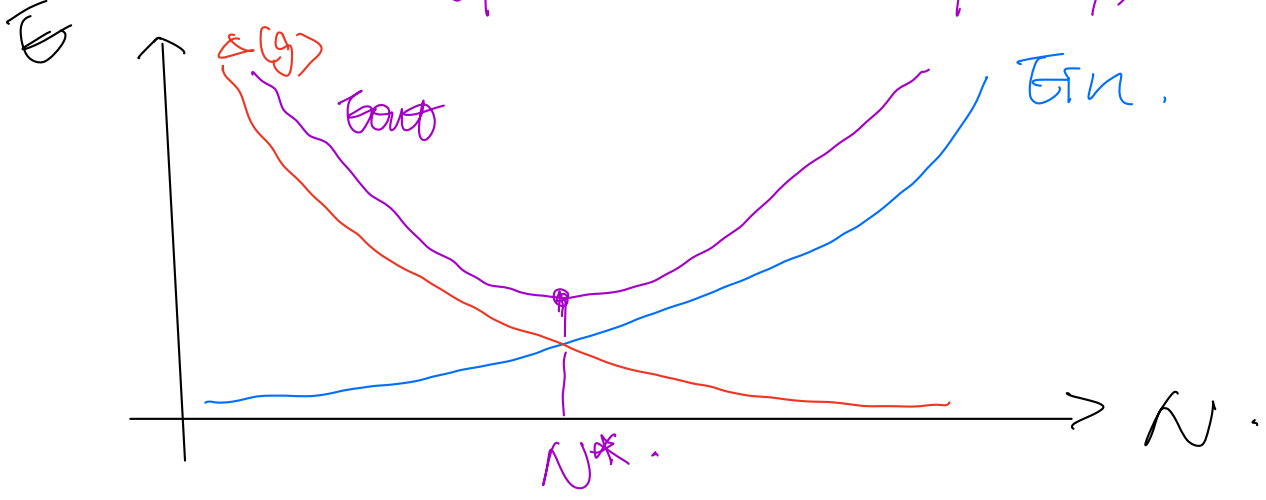
related to model complexity.
M: # hypothesis.
N: # dataset.
(# samples.

	$\Delta(g)$	$E_{in}(g)$
Increase M	increases	decreases
Increase N	decreases	increases.

- from same amount of computation, E_{in} larger
- harder to optimize.



(optimal model complexity)



Note: test error (E_{out}) is the only thing matters in the real world.

Note: Still trouble! Most useful/practical ML models in the world have ∞ M.

e.g., $y = \text{sign}(w^T x)$, $x \in \mathbb{R}^{d+1}$.

choices of w: infinite / uncountable

$\therefore$ Union bound NOT Enough! (Loose bound).

we need effective number of hypothesis

Lec 33 PAC Learning for final hypothesis

① Recap: PAC learning for final hypothesis $g \in \mathcal{H}$

$\mathcal{H} = \{h_1, h_2, \dots, h_M\}$.

Generalization error:

$$\Delta(g) = |E_{in}(g) - E_{out}(g)|$$

$$Pr\{\Delta(g) > \varepsilon\} \le Pr\{\bigcup_{i=1}^{M} \Delta(h_i) > \varepsilon\}$$

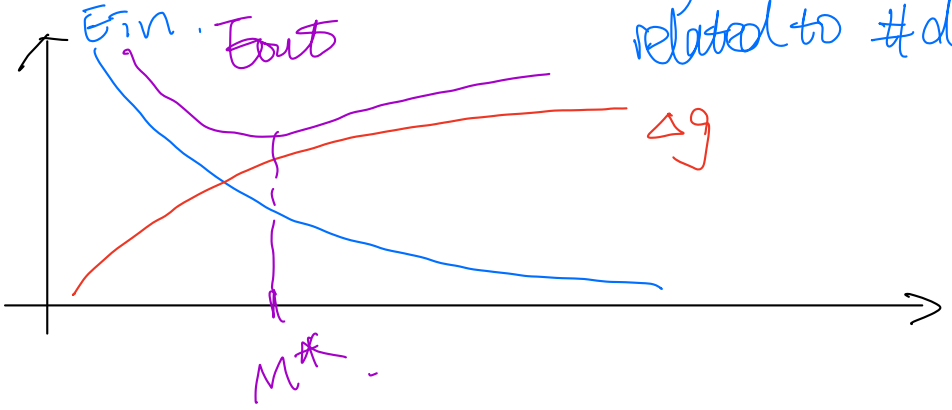
$$\overset{\text{union bound}}{\le} 2Me^{-2N\varepsilon^2}$$

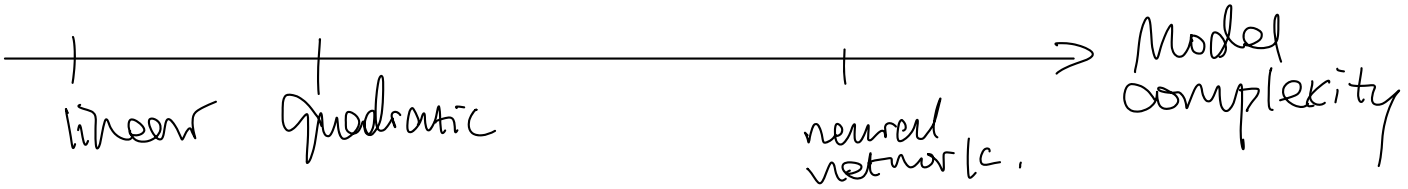
$\Rightarrow$ w/ probability $\ge 1-\delta$

$$E_{out}(g) \le E_{in}(g) + \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}}$$

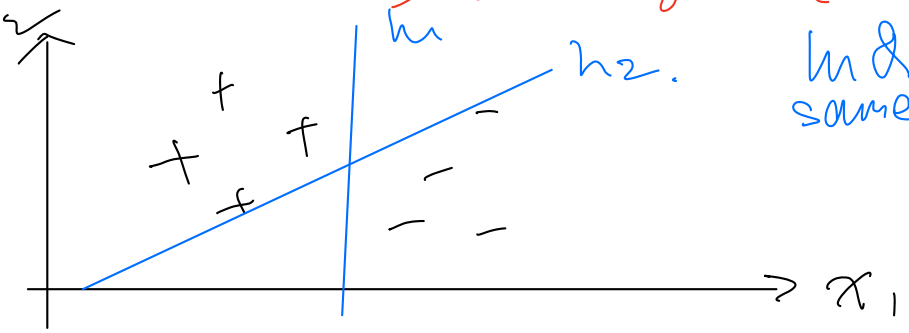
related to model complexity

related to # data points,





Trouble! Union bound is loose in general.



h_1 & h_2 has same effect on D.

② Idea: (drop 2.1, 2.2)

Consider $E_{in}(h) = \frac{1}{N}\sum_{n=1}^{N} 1\{y_n \neq h(x_n)\}$ dep on h only through $\{h(x_1), h(x_2), \dots, h(x_N)\}$.

⇒ Replace M w/ an effective # hypothesis $m_{\mathcal{H}}(N)$.

if 2 hs are the same ⇒ essentially 1 line.

e.g. $h_1 \equiv h_2$.

Dichotomy:

Let $x_1, x_2, \dots, x_N \in \mathbb{R}^d$ be fixed points.

Let $h \in \mathcal{H}$ be some hypothesis, $h: \mathbb{R}^d \to \{+1, -1\}$.

Def: Dichotomy vector

$(h(x_1), h(x_2), \dots, h(x_N)) \in \{+1, -1\}^N$.

Def: Dichotomy set

$$\mathcal{H}(x_1, x_2, \dots, x_N) = \{(h(x_1), h(x_2), \dots, h(x_N)) : h \in \mathcal{H}\}$$

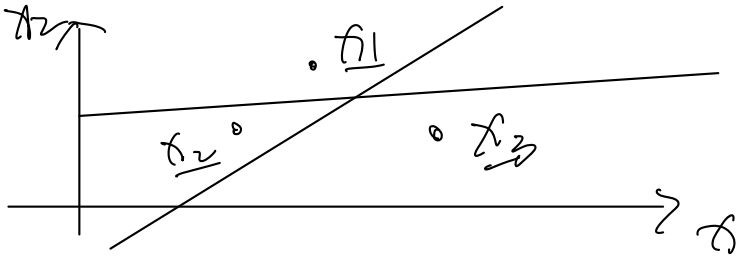
i.e., the collection of all binary vectors generated by $\mathcal{H}$ on $x_1, x_2, \dots, x_N$.

Note: there are no repeated elements in a set.

$|\mathcal{H}(x_1, x_2, \dots, x_N)| \leq 2^N$.

(No matter how large $M = |\mathcal{H}|$ is !).

e.g.: $d=2$



perceptron learning.

$|\mathcal{H}| = \infty$.

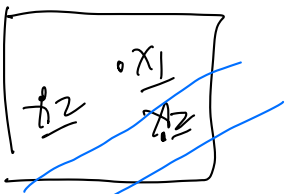
$|\mathcal{H}(x_1, x_2, x_3)| \leq 8$

Def the hypothesis set $\mathcal{H}$ shatters $(x_1, x_2, \dots, x_N)$ if $|\mathcal{H}(x_1, x_2, \dots, x_N)| = 2^N$.

e.g.: linear classification in $d=2$.

$\mathcal{H}$ = all $w \in \mathbb{R}^3$.

$y = \text{sign}(w_0 + w_1 x_1 + w_2 x_2)$



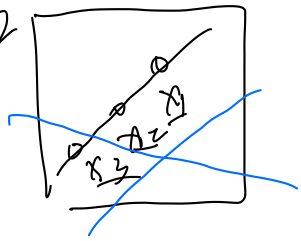
8/8

$h(x_1)$	$h(x_2)$	$h(x_3)$	
+	+	+	✓
+	+	−	✓
+	−	+	✓
+	−	−	✓
−	+	+	✓
−	+	−	✓
−	−	+	✓
−	−	−	✓

by symmetry.

Shattering

#2



6/8

$h(x_1)$	$h(x_2)$	$h(x_3)$	
+	+	+	✓
+	+	−	✓
+	−	+	✗
+	−	−	✓

not linearly separable

⇒ has to go to quadratic classifier.

$|H((x_1, x_2, x_3))| = 6 < 2^3$

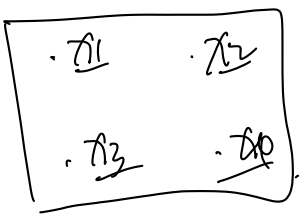
Not shattering

By symmetry in the opposite side ✓

✓ ✓ ✗

-, Linear classifier can shatter points not in one linear line (aka co-linear).

#3



-- A linear classifier (hypotheses) cannot shatter 4 points either.

$|H(x_1, x_2, x_3, x_4)| = 14 < 2^4$.

even though they are not co-linear)

Def Growth Function

$$m_H(N) = \max_{x_1, x_2, \ldots x_N \in \mathbb{R}^d} |H(x_1, x_2, \ldots, x_N)|.$$

↑ (all possible location of points)

≡ Effective # of hypotheses

e.g., For linear classification in $d=2$

$m_H(3) = 8$ $\quad$ $m_H(4) = 14$ (worst case) (at most have 2^4 hypothesis small comparing to ∞)

(Recall: hypothesis represents model complexity wrt dataset).

Note:
1) $m_H(N) \leq 2^N$
2) A tighter bound can be found, using the VC dimension.

Def: Let k be an integer s.t. $m_H(k) < 2^k$, then k is a breakpoint of H.

e.g., for linear classifier, the (first) break point is 4.

Def Vapnik-Chervonenkis (VC) Dimension

Let N be an integer s.t. $m_H(N) = 2^N$ & $m_H(N+1) < 2^{N+1}$

i.e., $N+1$ is the first (smallest) break point.

then VC dimension of H:

$d_{VC}(H) = N$.

Next week, VC dimension to bound growth func.