PAC Learning

1) Unknown distribution $P(\underline{x})$

$\underline{x_1, \dots, x_N}$

2) Unknown function

$y = f(\underline{x}) \in \{-1, +1\}$

3) Dataset $D = \{(\underline{x_1, y_1}), \dots, (\underline{x_n, y_n})\}$

4) Hypothesis Class $\mathcal{H} = \{h_1, \dots, h_N\}$

5) Learning algorithm A

Output $g \in \mathcal{H}$

In Sample Training Error

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

Test Error

$$E_{out}(g) = Pr(y \neq g(\underline{x}))$$

Conditions under which

$E_{in}(g) \approx E_{out}(g)$

Example, Jar with many balls either red or green
Fraction of red balls = μ (unknown)
We want to draw N balls & estimate μ

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

$$\Pr(Z = 1) = \mu$$

$$\Pr(Z = 0) = 1 - \mu$$

$$Z = \text{Bernoulli}(\mu)$$

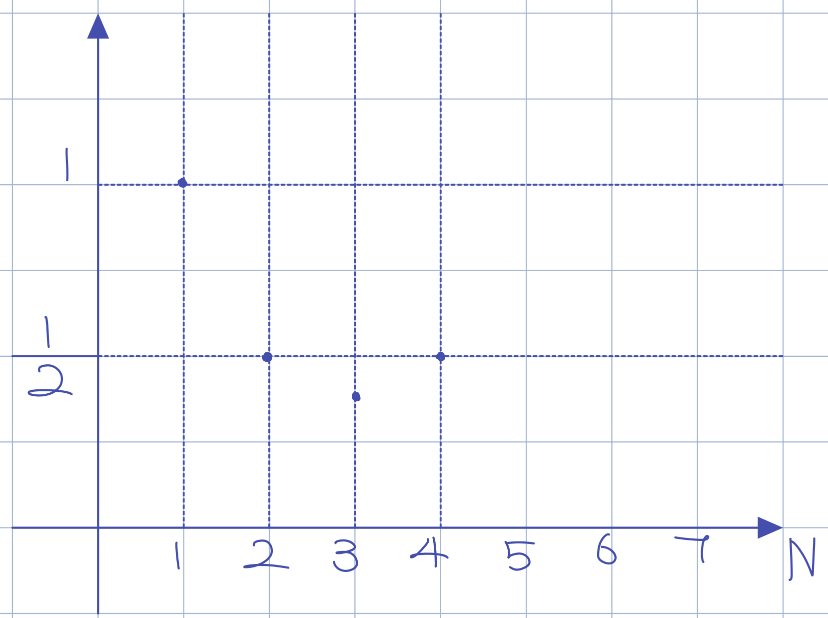
N drawns: $Z_1, Z_2, \ldots, Z_N \overset{iid}{\sim} \text{Bernoulli}(\mu)$
$N = 9$, $(1\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ 1)$

$$Y_n = \frac{1}{n} \sum_{i=1}^{n} Z_i$$

$$Y_1 = 1$$

$$Y_2 = \frac{1}{2}$$

$$Y_3 = \frac{1}{3}$$



As $N \longrightarrow \infty$

$Y_N \longrightarrow \mu$ (Law of large numbers)

Properties of Y_N

1) $\mathbb{E}[Y_N] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Z_i]$

Expectation

$$= \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

2) $Var[Y_N] = \mathbb{E}\left[(Y_N - \mathbb{E}[Y_N])^2\right]$

$$Var\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[Z_i] = \frac{\mu(1-\mu)}{n} \leq \frac{1}{n}$$

Bernoulli RV

$$Var[Z_i] = \mu(1-\mu)$$

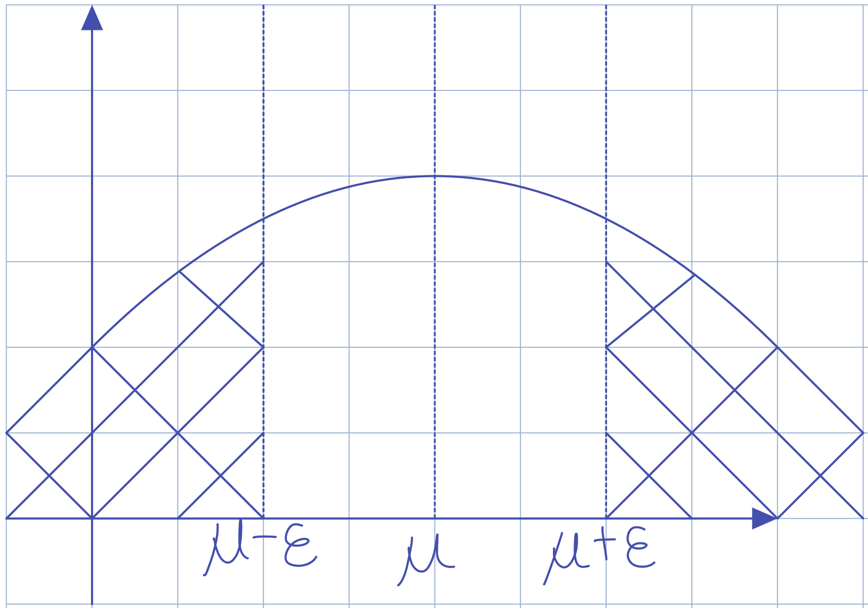
Chebeshev's Inequality

$$Pr(|Y_N - \mu| > \epsilon) \leq \frac{Var(Y_N)}{\varepsilon^2}$$

$$\leq \frac{1}{N\varepsilon^2}$$

$Pr(Y_N > \mu + \varepsilon \text{ or } Y_N < \mu + \varepsilon)$

$\leq \frac{1}{N\varepsilon^2}$



Confidence Interval

$$(\mu - \varepsilon, \mu + \varepsilon)$$

$$\delta = \Pr(V_N \notin \text{Confidence Interval})$$

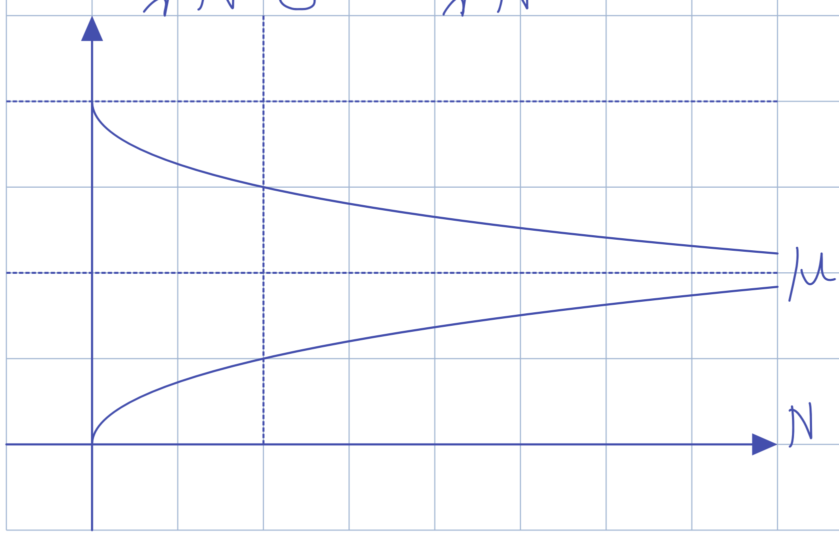
$$1 - \delta \equiv \text{Confidence Level}$$

$$\delta = 0.01$$

$$1 - \delta = 0.99$$

$$\delta \leq \frac{1}{N \varepsilon^2}$$

$$\varepsilon \leq \frac{1}{\sqrt{N \delta}} = \frac{10}{\sqrt{N}}$$



Not Tight Inequality

Hoeffding Inequality

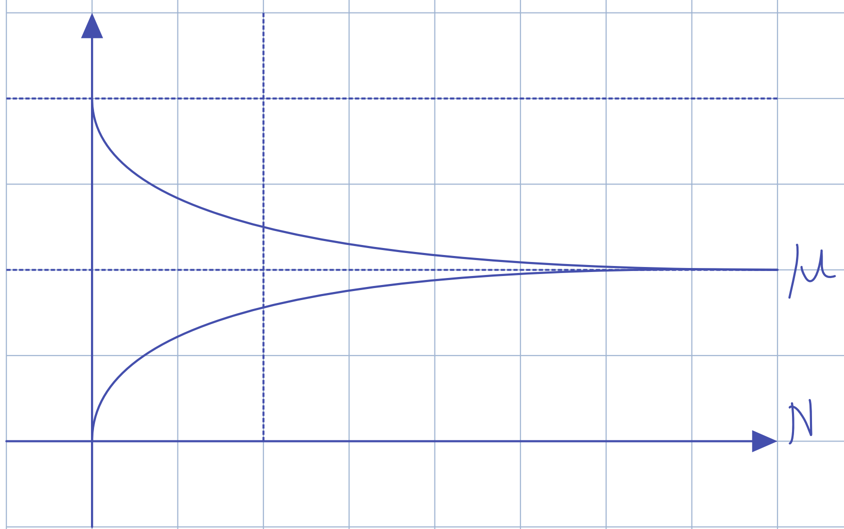
$$\Pr\left(|\nu_N - \mu| > \varepsilon\right) \leq 2e^{-2N\varepsilon^2}$$

Confidence Level $= 1 - \delta$

$$e^{2N\varepsilon^2} = \frac{2}{\delta} \longrightarrow \varepsilon = \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}$$

Confidence Interval

$$\left(\mu - \sqrt{\frac{1}{2N}\log\frac{2}{\delta}},\ \mu + \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}\right)$$



Better Constant

Amelie's Hand

Recall: PAC Learning

2 Types of error

Training Error

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

Test Error

$$E_{out}(g) = \Pr(y \neq g(\underline{x}))$$

Example

$Z \sim \text{Bernoulli}(\mu)$

$\Pr(Z = 0) = 1 - \mu$

$\Pr(Z = 1) = \mu$

$$\Pr\left(\left| \frac{1}{N}\sum_{n=1}^{N} Z_i - \mu \right| > \epsilon \right) \leq 2e^{-2N\epsilon^2}$$

$E_{in}(h)$ $E_{out}(h)$

$$Z = \begin{cases} 0 & \text{if green ball drawn, } y_i \neq h(x_i) \\ 1 & \text{if red ball drawn, } y_i = h(x_i) \end{cases}$$

$$E_{out}(h) = \Pr(Z_i = 1) = \mu$$

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

$$\Pr\left(|E_{in}(h) - E_{out}(h)| > \varepsilon\right) \leq 2e^{-2N\varepsilon^2}$$

$$= \Pr\left(\left|\frac{1}{N}\sum_{n=1}^{N} Z_i - \mu\right| > \varepsilon\right)$$

Given $\underline{X_1, \dots, X_N} \xrightarrow{\text{iid}} P(\underline{X})$

$y_i = f(\underline{x_i})$

$Z_1 = \mathbb{1}\{Y_1 \neq h(\underline{X_1})\}$

$Z_2 = \mathbb{1}\{Y_2 \neq h(\underline{X_2})\}$

Suppose $h(\cdot)$ is a fixed hypothesis selected before traning samples are observed

$\downarrow$

$Z_1, \dots, Z_n$ is iid

$X_1, \dots, X_N$ iid $P(X)$

$h(\cdot)$ is not function of $X_1, \dots, X_n$

$h(X_1), \dots, h(X_n)$ is iid

$y_i = f(\underline{x_i})$, where $f(\cdot)$ is some fixed function

$\underbrace{\mathbb{1}\{h(\underline{x_1}) \neq f(\underline{x_1})\}}_{Z_1}, \dots$ also iid

Hoeffding Applies

$$\Pr\left(|E_{in}(h) - E_{out}(h)| > \varepsilon\right) \leq 2e^{-2N\varepsilon^2}$$

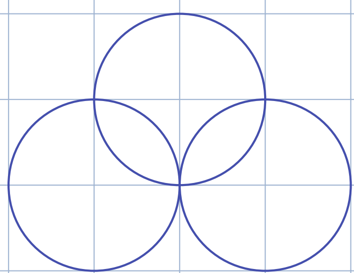
Not Useful

We want to extend this analysis when $h(\cdot)$ is not fixed but selected after observing dataset

Union Bound

$A_1, \ldots, A_N$ be arbitrary events

$$Pr(A_1 \cup \ldots \cup A_N) \leq \sum_{i=1}^{N} Pr(A_i)$$



Equality holds when disjoint events

Task, output hypothesis

$g(\cdot)$ is selected after observing training data

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

Hypothesis Class

$$H = \{h_1, h_2, \ldots, h_M\}$$

$$g \in \{h_1, h_2, \ldots, h_M\}$$

$Pr(|E_{in}(h) - E_{out}(h)| > \varepsilon)$ Bound

Claim,

Given dataset

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

$|E_{in}(h) - E_{out}(h)| > \epsilon$

Then the following must hold

$\downarrow$

At Least One Of

1) $|E_{in}(h_1) - E_{out}(h_1)| > \epsilon$

$\vdots$

m) $|E_{in}(h_m) - E_{out}(h_m)| > \epsilon$

Suppose

$|E_{in}(h_i) - E_{out}(h_i)| < \epsilon$

for all $i = 1, 2, \ldots, M$

Since $g \in \{h_1, \ldots, h_M\}$

$|E_{in}(g) - E_{out}(g)| \leq \epsilon$

Cannot be true since we assumed that

$|E_{in}(g) - E_{out}(g)| > \epsilon$

for at least one of $h_1, \ldots, h_M$

Event:

$|E_{in}(g) - E_{out}(g)| > \epsilon$

$$\subseteq \bigcup_{i=1}^{M} \{|E_{in}(h_i) - E_{out}(h_i)| > \epsilon\}$$

$$\Pr\left(|E_{in}(g) - E_{out}(g)| > \varepsilon\right) \leq$$

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

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

$$\sum_{i=1}^{M} \Pr\{|E_{in}(\underbrace{h_i}_{\text{fixed hypothesis before observing dataset } D}) - E_{out}(h_i)| > \varepsilon\} \leq 2Me^{-2N\varepsilon^2}$$

Summary

$g \in \{h_1, h_2, \ldots, h_M\}$

after observing D

$$\sum_{i=1}^{M} \Pr\{|E_{in}(h_i) - E_{out}(h_i)| > \varepsilon\} \leq 2Me^{-2N\varepsilon^2} = \delta$$

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

$$\left(|E_{in}(h_i) - E_{out}(h_i)| > \varepsilon\right) \leq \sqrt{\frac{1}{2N} \log \frac{2M}{\delta}}$$

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

M (Candidate Hypothesis)

N (Training Samples)