Sample complexity of agnostic PAC learning

- Recall the notion of ε-approximation

Definition: Let D be a probability distribution over X. Let $H \subseteq \mathcal{P}(X)$. A finite set (multiset) $S \subseteq X$ is an **ε-approximation** for D, H if for all $h \in H$,

$$\left| D(h) - \frac{|S \cap h|}{|S|} \right| \leq \varepsilon$$

Lemma: Let D be a distribution over $X \times Y$. Let $H \subseteq Y^X$. Let $S \subseteq X \times Y$ be an **ε-approximation** for $D, G^c(H)$ then

1) $\forall h \in H: \quad |err_D(h) - \widehat{err_S}(h)| \leq \varepsilon$

2) $err_D(ERM(S)) \leq \min_{h \in H} err_D(h) + 2\varepsilon$

Proof:

- $err_D(h) = Pr[h(x) \neq y]$

$= D(\{(x,y) \in X \times Y : h(x) \neq y\})$

$= D(g^c(h))$

- $\widehat{err_S}(h) = \frac{|\{(x,y) \in S : h(x) \neq y\}|}{|S|}$

$|c \quad c(h)|$

$$= \frac{|S \cap g^c(h)|}{|S|}$$

- If S is ε-approximation for $D, G^c(H)$ then

$$\forall h \in H: \left| D(g^c(h)) - \frac{|S \cap g^c(h)|}{|S|} \right| \leq \varepsilon$$

- Therefore

$$\forall h \in H: \left| err_D(h) - \widehat{err}_S(h) \right| \leq \varepsilon$$

This proves part 1).

Consider ERM algorithm

Since ERM minimizes $\widehat{err}_S(\cdot)$,

- $\forall h \in H: \widehat{err}_S(ERM(S)) \leq \widehat{err}_S(h)$
- Use part 1) twice:

For all $h \in H$:

$$err_D(ERM(S)) \leq \widehat{err}_S(ERM(S)) + \varepsilon$$

$$\leq \widehat{err}_S(h) + \varepsilon$$

$$\leq err_D(h) + \varepsilon + \varepsilon$$

$$= err_D(h) + 2\varepsilon$$

Here we used part 1)

Here we use ...

- $err_D(ERM(S)) \leq err_D(h) + 2\varepsilon$
- Therefore

$$err_D(ERM(S)) \leq \min_{h \in H} err_D(h) + 2\varepsilon$$

- We want prove that if S is an i.i.d. from D then with high probability S is an ε-approximation.
- We will need two tools

from probability theory

1) McDiarmid's inequality

2) Massarts lemma

McDiarmid's inequality:

Let $X_1, X_2, \dots, X_n$ be independent random variables such that $X_i \in A_i$.

Let $f: A_1 \times A_2 \times \dots \times A_n \to \mathbb{R}$ be a function such that for any i, for any $x_1, x_2, \dots, x_n, x_i'$

$$|f(x_1, x_2, \dots, x_n) - f(x_1, x_2, \dots x_{i-1}, x_i', x_{i+1} \dots x_n)| \leq c_i$$

(bounded difference property).

Then

$$\Pr\left[\left|f(X_1,\dots X_n) - \mathbb{E}\left[f(X_1,\dots X_n)\right]\right| > \varepsilon\right] \leq 2e^{-\frac{\varepsilon^2}{\sum_{i=1}^{n} c_i^2}}$$

- We omit the proof

(The proof relies on martingales.)

- The inequality is a generalization Hoeffding's inequality.
- We can recover Hoeffding's inequality if we take

$$f(x_1,\dots x_n) = x_1 + x_2 + \dots + x_n$$

Massart's lemma:

Let $\sigma_1, \dots \sigma_n$ be random variables

Let $X_1, X_2 \dots X_n$ be [illegible]

such that for all i, $s > 0$,

$$\mathbb{E}\left[e^{sX_i}\right] \le e^{s^2\sigma^2/2}$$

1) Then

$$\mathbb{E}\left[\max_{i=1..n} X_i\right] \le \sigma\sqrt{2\ln n}$$

2) If additionally for all i, $s > 0$,

$$\mathbb{E}\left[e^{-sX_i}\right] \le e^{s^2\sigma^2/2}$$

then

$$\mathbb{E}\left[\max_{i=1..n} |X_i|\right] \le \sigma\sqrt{2\ln(2n)}$$

Proof:

2) follows from by applying 1)

- 2) follows from [illegible]

to $\underbrace{X_1, \ldots X_n, -X_1, \ldots, -X_n}_{2n \text{ variables}}$

- It suffices to prove 1)

$$e^{s\mathbb{E}[\max_i X_i]} \leq \mathbb{E}\left[e^{s \max_i X_i}\right]$$

$$= \mathbb{E}\left[\max_i e^{sX_i}\right]$$

$$\leq \sum_{i=1}^{n} \mathbb{E}\left[e^{sX_i}\right]$$

$$\leq n e^{s^2\sigma^2/2}$$

$\leq n e$

Take logarithm and divide by $s > 0$:

$$\mathbb{E}\left[\max_i X_i\right] \leq \frac{\ln n}{s} + \frac{s\sigma^2}{2}$$

- Find $s > 0$ that minimizes right-hand side.
- Optimal $s = \frac{\sqrt{2 \ln n}}{\sigma}$
- $\mathbb{E}\left[\max_i X_i\right] \leq \sigma\sqrt{\frac{1}{2}\ln n} + \sigma\sqrt{\frac{1}{2}\ln n}$

$$= \sigma\sqrt{2 \ln n}$$

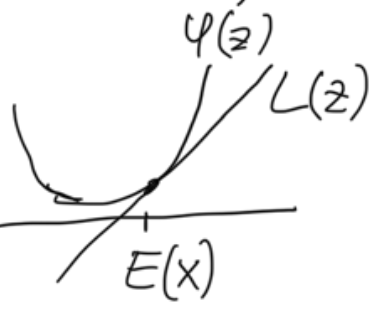
Lemma: Let X be a random variable. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function. Then,

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$$

Proof: $L(z) = \varphi(\mathbb{E}[X]) + (z - \mathbb{E}[X])\varphi'(\mathbb{E}[X])$

$\forall z \in \mathbb{R} \quad L(z) \leq \varphi(z)$

$$\mathbb{E}[L(X)] \leq \mathbb{E}[\varphi(X)]$$



Lemma: Let D be a distribution over X. Let $H \subseteq \mathcal{P}(X)$. Let S be a i.i.d. sample from D of size m. Then,

$$\mathbb{E}\left[\sup_{h \in H} \left| D(h) - \frac{|S \cap h|}{|S|} \right|\right] \leq \sqrt{\frac{2 \ln\left(2 \Pi_H(2m)\right)}{m}}.$$

Proof:

- Let $S = (x_1, x_2 \ldots x_m)$

- $$\frac{|S \cap h|}{|S|} = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}[x_i \in h]$$

- Let $Y_1, Y_2, \ldots, Y_m$ be an independent ghost sample.

$$D(h) = \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}[Y_i \in h]\right]$$

- Let

$$P_m(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}[X_i \in h]$$

$$Q_m(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}[Y_i \in h]$$

- We need to upper bound

$$\mathbb{E}\left[\sup_{h \in H} \left| P_m(h) - \mathbb{E}[Q_m(h)] \right| \right]$$

$$= \mathbb{E}\left[\sup_{h \in H} \left| \mathbb{E}[P_m(h) - Q_m(h) \mid X_1 \ldots X_m] \right| \right]$$

$$\leq \mathbb{E}\left[\sup_{h \in H} \mathbb{E}\left[\left| P_m(h) - Q_m(h) \right| \mid X_1 \ldots X_m \right] \right]$$

$$\leq \mathbb{E}\left[\mathbb{E}\left[\sup \left| P_m(h) - Q_m(h) \right| \mid X_1 \ldots X_m \right] \right]$$

$$\left[\sup_{h \in H} \right] \qquad \left] \right.$$

$$= \mathbb{E}\left[\sup_{h \in H} \left| P_m(h) - Q_m(h) \right| \right]$$

$$= \frac{1}{m} \mathbb{E}\left[\sup_{h \in H} \left| \sum_{i=1}^{m} \left(\mathbb{1}[X_i \in h] - \mathbb{1}[Y_i \in h] \right) \right| \right]$$

- Let us introduce Rademacher variables $\sigma_1, \sigma_2, \ldots, \sigma_m$.

These are i.i.d. variables independent of $X_1 \ldots X_m, Y_1 \ldots Y_m$ such that

$$\Pr[\sigma_i = +1] = \Pr[\sigma_i = -1] = 1/2$$

- Note that the distribution of $\left(\mathbb{1}[X_i \in h] - \mathbb{1}[Y_i \in h]\right)$ is the same as the distribution of $\sigma_i\left(\mathbb{1}[X_i \in h] - \mathbb{1}[Y_i \in h]\right)$.

- $$\mathbb{E}\left[\sup_{h \in H} \left| \sum_{i=1}^{m} \left(\mathbb{1}[X_i \in h] - \mathbb{1}[Y_i \in h]\right) \right|\right]$$

$$= \mathbb{E}\left[\sup_{h \in H} \left| \sum_{i=1}^{m} \sigma_i \left(\mathbb{1}[X_i \in h] - \mathbb{1}[Y_i \in h]\right) \right|\right]$$

- Let $U = (X_1 \dots X_m, Y_1 \dots Y_m)$ be sample of size $2m$.

- Condition on U
- There are at most $\Pi_H(2m)$ possible behaviors of H on U.

- After conditioning on U

$$\mathbb{E}\left[\sup_{h\in H}\left|\sum_{i=1}^{m}\sigma_i\left(\mathbb{1}[x_i\in h]-\mathbb{1}[y_i\in h]\right)\right|\right]$$

$$=\mathbb{E}\left[\max_{h\in \Pi_H(U)}\left|\sum_{i=1}^{m}\sigma_i\left(\mathbb{1}[x_i\in h]-\mathbb{1}[y_i\in h]\right)\right|\right]$$

- For any $h\in H$, conditioned on U,

$$Z_h=\sum_{i=1}^{m}\sigma_i\left(\mathbb{1}[x_i\in h]-\mathbb{1}[y_i\in h]\right)$$

is random variable such that

1) $\mathbb{E}[Z_n] = 0$

2) $Z_n = \sum_{i=1}^{m} Z_{n,i}$ where

$Z_{n,i} \in [-1, 1]$ and $\mathbb{E}[Z_{n,i}] = 0$

3) By Hoeffding's lemma, $s \in \mathbb{R}$,

$$\mathbb{E}[e^{sZ_{n,i}}] \leq e^{s^2/2}$$

$$\mathbb{E}[e^{sZ_n}] = \prod_{i=1}^{m} \mathbb{E}[e^{sZ_{n,i}}] \leq e^{ms^2/2}$$

- By Massart's lemma

$$\mathbb{E}\left[\max_{h \in \Pi_H(U)} \left| \sum_{i=1}^{m} \sigma_i \left(\mathbb{1}[x_i \in h] - \mathbb{1}[y_i \in h] \right) \right| \right]$$

$$= \mathbb{E}\left[\max_{h \in \Pi_H(U)} |Z_h| \right]$$

$$\leq \sqrt{2m \ln\left(2 \Pi_H(2m)\right)}$$

Theorem: Let D be a probability distribution over X. Let $\delta \in (0,1)$. Let $H \subseteq \mathcal{P}(X)$. Let S be an i.i.d. sample from D of size m. Then with probability at least $1-\delta$, S is an ε-approximation for

H, D where

$$\varepsilon = \sqrt{\frac{2\ln\left(2\Pi_H(2m)\right)}{m}} + \sqrt{\frac{\ln(2/\delta)}{2m}},$$

Proof:

- Let $Z = \sup_{h \in H} \left| D(h) - \frac{|S \cap h|}{|S|} \right|$

- Z is a random variable that is a function of $S = (X_1, X_2, \ldots, X_m)$

- We can apply McDiarmid's inequality to Z.

- Suppose $S = (x_1, \ldots, x_m)$,

$S' = (x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_m)$

and $$Z' = \sup_{h \in H} \left| D(h) - \frac{|S' \cap h|}{|S'|} \right|$$

- For all $h \in H$:

$$\left| \left| D(h) - \frac{|S' \cap h|}{|S'|} \right| - \left| D(h) - \frac{|S \cap h|}{|S|} \right| \right|$$

$$\leq \left| \frac{|S' \cap h|}{|S'|} - \frac{|S \cap h|}{|S|} \right|$$

$$= \frac{1}{m} \underbrace{\left| |S' \cap h| - |S \cap h| \right|}_{\leq 1}$$

$$\leq \frac{1}{m}$$

- Therefore

$$\sup_{h \in H} \left| D(h) - \frac{|S \cap h|}{|S|} \right| - \sup_{h \in H} \left| D(h) - \frac{|S' \cap h|}{|S'|} \right| \leq \frac{1}{m}$$

- So Z satisfies McDiarmid's theorem with $c_i = \frac{1}{m}$.

- Therefore, with probability at least $1-\delta$,

$$|Z - \mathbb{E}[Z]| \leq \sqrt{\frac{\ln(2/\delta)}{2m}}$$

- By previous lemma

$$\mathbb{E}[Z] \leq \sqrt{\frac{2 \ln(2 \Pi_H(2m))}{m}}$$

- Thus, w.p. at least $1-\delta$,

$$\leq \mathbb{E}[Z] + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

$$\leq \sqrt{\frac{2\ln(2\Pi_H(2m))}{m}} + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

∎

Corrollary: Let D be a distribution over X. Let $H \subseteq \mathcal{P}(X)$ and let $d = VC(H)$. Let $\varepsilon, \delta \in (0,1)$. Let S be an i.i.d. sample from D of size m. With probability at least $1-\delta$, S is an ε-approximation provided

$$m \geq \max\left(\frac{2\ln(2/\delta)}{\varepsilon^2}, \frac{4\ln 2}{\varepsilon^2}, \frac{16d\ln\left(\left(\frac{32e}{\varepsilon^2}\right)^2\right)}{\varepsilon^2}\right)$$

Proof:

- It suffices to show that

$$\sqrt{\frac{2\ln\left(2\Pi_H(2m)\right)}{m}} + \sqrt{\frac{\ln(2/\delta)}{2m}} \leq \varepsilon.$$

- Clearly

$$\sqrt{\frac{\ln(2/\delta)}{2m}} \leq \varepsilon/2$$

- It remains to show that

$$\sqrt{\frac{2\ln\left(2\Pi_H(2m)\right)}{m}} \leq \varepsilon/2$$

- Since $2m \geq d$, $\Pi_H(2m) \leq \left(\frac{2me}{d}\right)^d$

and

$$\sqrt{\frac{2\ln\left(2\Pi_H(2m)\right)}{m}}$$

$$\leq \sqrt{\frac{2\ln\left(2\left(\frac{2me}{d}\right)^d\right)}{m}}$$

$$= \sqrt{\frac{2\ln 2 + 2d\ln\left(\frac{2me}{d}\right)}{m}}$$

- Squaring both sides

$$\frac{2\ln 2 + 2d\ln\left(\frac{2me}{d}\right)}{m} \leq \frac{\varepsilon^2}{4}$$

- $\frac{2\ln 2}{m} \leq \frac{\varepsilon^2}{\delta}$. It suffices to prove

$$\frac{2d \ln\left(\frac{2me}{d}\right)}{m} \leq \frac{\varepsilon^2}{\delta}$$

- Left-hand side is decreasing in m. It suffices to verify if for

$$m = \frac{16d \ln\left(\left(\frac{32e}{\varepsilon^2}\right)^2\right)}{\varepsilon^2}$$

- We need to prove

$$\frac{2d \ln\left(\frac{2me}{d}\right)}{m} \leq \frac{\varepsilon^2}{\delta}$$

$$\Updownarrow$$

$$\frac{2d \ln\left(\frac{32ed}{\varepsilon^2 d} \ln\left(\left(\frac{32e}{\varepsilon^2}\right)^2\right)\right)}{\frac{16d}{\varepsilon^2} \ln\left(\left(\frac{32e}{\varepsilon^2}\right)^2\right)} \leq \frac{\varepsilon^2}{8}$$

$$\Updownarrow$$

$$\frac{\ln\left(\frac{32e}{\varepsilon^2} \ln\left(\left(\frac{32e}{\varepsilon^2}\right)^2\right)\right)}{\ln\left(\left(\frac{32e}{\varepsilon^2}\right)^2\right)} \leq 1$$

$$\Updownarrow$$

$$\frac{32e}{\varepsilon^2} \ln\left(\left(\frac{32e}{\varepsilon^2}\right)^2\right) \leq \left(\frac{32e}{\varepsilon^2}\right)^2$$

ε^2 $\left(\varepsilon^2/2\right)$ $\left(\varepsilon^2\right)$

$\Updownarrow$

$$2\ln\frac{32e}{\varepsilon^2} \leq \frac{32e}{\varepsilon^2}$$

$\Updownarrow$

$$2\ln x \leq x$$

which is true for all $x > 0$.

∎

Corollary: Let D be a distribution over $X \times Y$. Let $H \subseteq Y^X$. $d = VC(H)$. Let $\varepsilon, \delta \in (0,1)$. Let S be an i.i.d. sample from D

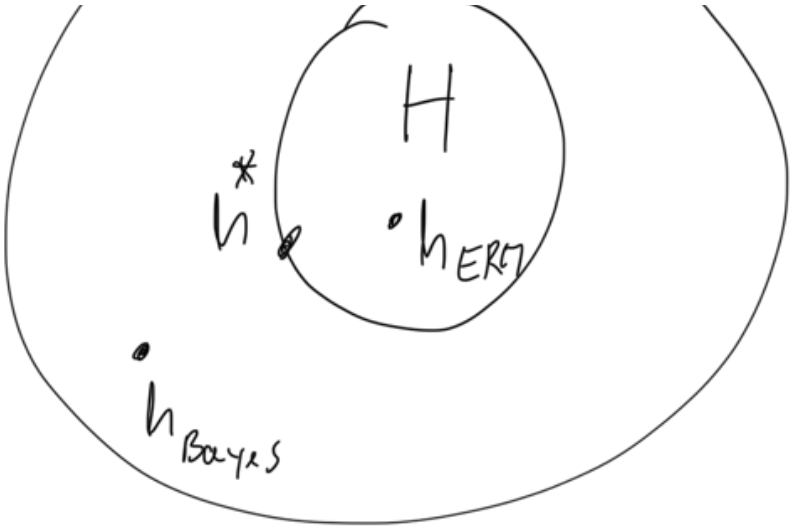
of size

$$m \geq \max\left(\frac{8\ln(2/\delta)}{\varepsilon^2}, \frac{16\ln 2}{\varepsilon^2}, \frac{64d\ln\left(\left(\frac{12e}{\varepsilon^2}\right)^2\right)}{\varepsilon^2} \right)$$

Then, with probability at least $1-\delta$,

$$\text{err}_D(\text{ERM}(S)) \leq \inf_{h \in H} \text{err}_D(h) + \varepsilon.$$

Proof: Follows directly from previous corollary with $\varepsilon \leftarrow \varepsilon/2$ and lemmas above.



$$h^* = \underset{h \in H}{\operatorname{argmin}} \; err_D(h)$$

$$err_D(h_{Bayes}) - err_D(h^*) = \text{approximation error}$$

$$err_D(h_{ERM}) - err_D(h^*) = \text{estimation error}$$