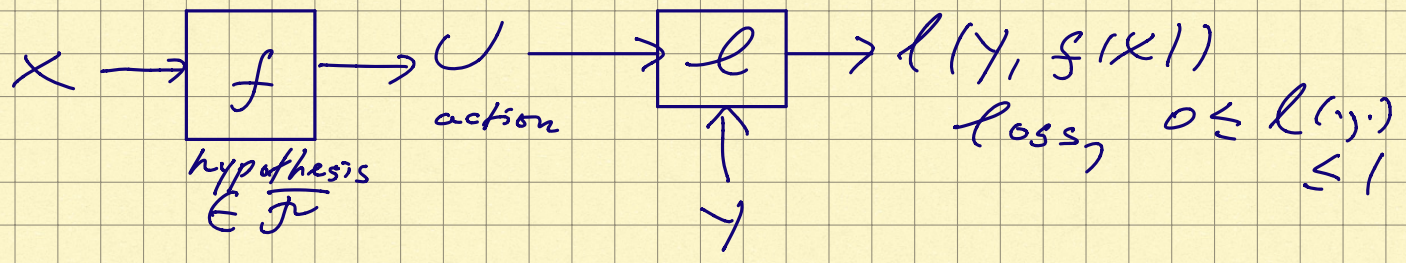
Review: Agnostic (Model-Free) Learning

$(X, Y) \sim P$ (unmodeled)

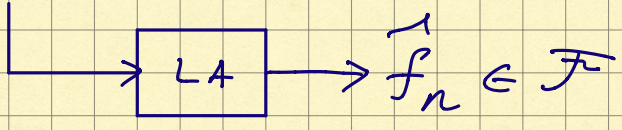
(feature), (label)



$L_P(f) := \mathbb{E}[\ell(Y, f(X))]$ — risk of f

$L_P^*(\mathcal{F}) := \inf_{f \in \mathcal{F}} L_P(f)$

Learning: $\underbrace{(X_1, Y_1), \ldots, (X_n, Y_n)}_{\text{data}} \overset{iid}{\sim} P$



Goal: design LA s.t., no matter what P is,

$L_P(\hat{f}_n) \approx L_P^*(\mathcal{F})$ w.h.p.

PAC learning: $\forall \varepsilon > 0$

$\lim_{n\to\infty} \sup_{P \in \mathcal{P}} P^n\left(L_P(\hat{f}_n) > L_P^*(\mathcal{F}) + \varepsilon\right) = 0.$

Do PAC learning algos exist?

$(X_1, Y_1), \dots, (X_n, Y_n)$ - data

Empirical risk of $f \in \mathcal{F}$: $\frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) =: P_n(f)$

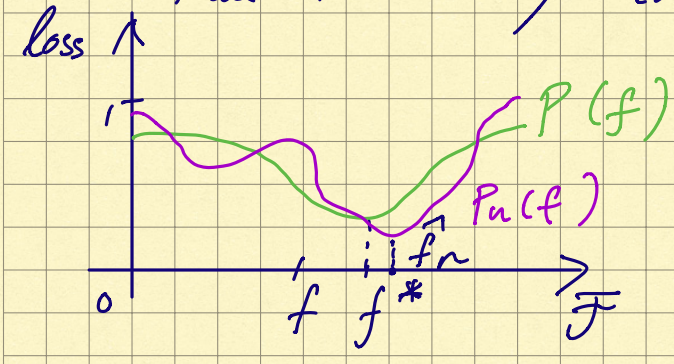
Population risk of $f \in \mathcal{F}$: $L_P(f) = P(f)$

1) $\mathbb{E} P_n(f) = P(f)$

2) $|P_n(f) - P(f)| \to 0$ a.s.

$$\mathbb{P}\{|P_n(f) - P(f)| > \varepsilon\} \le 2e^{-2n\varepsilon^2}$$

- true for every individual $f \in \mathcal{F}$



$$P(f^*) = \inf_{f \in \mathcal{F}} P(f)$$

$$P_n(\hat{f}_n) = \inf_{f \in \mathcal{F}} P_n(f)$$

Empirical Risk Minimization:

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$$

Mismatched Minimization Lemma

- $G, \hat{G} : \mathcal{U} \to \mathbb{R}$
- want to minimize G, can't do it directly
- can minimize $\hat{G}$
- G and $\hat{G}$ are "close": $\sup_{u \in \mathcal{U}} |G(u) - \hat{G}(u)| \le \varepsilon$

Then: 1) $G(u) \le \hat{G}(u) + \varepsilon$ for all $u \in \mathcal{U}$

2) let $\hat{u}^*$ be the minimizer of $\hat{G}$; then

$$G(\hat{u}^*) \le \inf_{u \in \mathcal{U}} G(u) + 2\varepsilon$$

Proof: 1) is obvious from hypotheses on $G, \hat{G}$.

2) For any $u \in \mathcal{U}$,

$$\begin{aligned} G(u) &\ge \hat{G}(u) - \varepsilon && |G - \hat{G}| \le \varepsilon \\ &\ge \hat{G}(\hat{u}^*) - \varepsilon && \hat{G}(u) \ge \hat{G}(\hat{u}^*) \\ &\ge G(\hat{u}^*) - 2\varepsilon && |G - \hat{G}| \le \varepsilon \end{aligned}$$

$$\Rightarrow \inf_{u \in \mathcal{U}} G(u) \ge G(\hat{u}^*) - 2\varepsilon. \qquad \square$$

Our setting: $P, P_n : \mathcal{F} \to [0,1]$

if $\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \le \varepsilon$, then:

1) for any LA that generates $\hat{f}_n$,

$$P(\hat{f}_n) \le P_n(\hat{f}_n) + \varepsilon$$

2) for ERM, $\hat{f}_n = \operatorname{argmin}_{f \in \mathcal{F}} P_n(f)$,

$$P(\hat{f}_n) \le \inf_{f \in \mathcal{F}} P(f) + 2\varepsilon.$$

A sufficient condition for ERM to be PAC:

$$\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \xrightarrow{a} 0 \quad \text{w.h.p.}$$

Formalize: $\ell, \mathcal{F}, \mathcal{P}$ given

$$q(n,\varepsilon) := \sup_{P\in\mathcal{P}} P^n\left(\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| \geq \varepsilon \right)$$

ERM is PAC if $q(n,\varepsilon) \to 0$ as $n\to\infty$ for any $\varepsilon > 0$.

Example: any finite $\mathcal{F} = \{f_1, \dots, f_M\}$

$$P^n\left(\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| \geq \varepsilon \right)$$

$$= P^n\left(\bigcup_{m=1}^{M} \{ |P_n(f_m) - P(f_m)| \geq \varepsilon \} \right)$$

$$\leq \sum_{m=1}^{M} P^n\left(|P_n(f_m) - P(f_m)| \geq \varepsilon \right) \quad \text{(union bound)}$$

$$\leq 2Me^{-2n\varepsilon^2} \quad \text{(Hoeffding)}$$

$$q(n,\varepsilon) = \sup_P P^n\left(\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| \geq \varepsilon \right)$$

$$\leq 2Me^{-2n\varepsilon^2}$$

$$\to 0 \quad \text{as } n\to\infty.$$

Generally: for a broad class of $(\ell, \mathcal{F})$,

$$\sup_P \mathbb{E}\left[\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| \right] \leq \frac{C}{\sqrt{n}}$$

where $C \geq 0$ depends on "complexity" of $\ell, \mathcal{F}$ ($\mathcal{F}$ may be infinite!)

Then ERM is PAC:

$$P^n\left\{ \underbrace{\sup_{f\in\mathcal{F}} |P_n(f) - P(f)|}_{:= g(z^n)} \geq \varepsilon \right\}$$

$z_i := (x_i, y_i)$

$z^n = (z_1, \dots, z_n)$

Know $\mathbb{E}\, g(z^n) \leq \frac{C}{\sqrt{n}}$;

$$g(z^n) = \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i)) - \mathbb{E}_P \ell(Y, f(X)) \right|$$

has bdd diff. $c_1 = \dots = c_n = \frac{1}{n}$

$$\Rightarrow P\{ g(z^n) \geq \mathbb{E} g(z^n) + t \} \leq e^{-2nt^2}$$

$\forall t > 0$ (by McDiarmid)

$$C/\sqrt{n} \leq \varepsilon/2 \iff n \geq \frac{4C^2}{\varepsilon^2}$$

$$P\{ g(z^n) \geq \varepsilon \} = P\{ g(z^n) - \mathbb{E} g(z^n) \geq \varepsilon - \mathbb{E} g(z^n) \}$$

$$\leq P\left\{ g(z^n) - \mathbb{E} g(z^n) \geq \varepsilon - \frac{C}{\sqrt{n}} \right\}$$

$$\leq P\{ g(z^n) - \mathbb{E} g(z^n) \geq \varepsilon/2 \}$$

$$\leq e^{-2n(\varepsilon/2)^2}$$

$$= e^{-n\varepsilon^2/2}$$

$$\Rightarrow q(n, \varepsilon) \leq e^{-2n\varepsilon^2/2} \quad \text{for} \quad n \geq \frac{4C^2}{\varepsilon^2}$$