Tabular Analysis

State space S. $\rightarrow$ $|S| \times |A|$ < planning / learning.

< Method 1: Certainty Equivalence (Model-Based) P, R

Method 2: Valued-Function Approach (Model-Free)

e.g. Q-learning. Sarsa. ---.

$\hat{P}$, $\hat{R}$ (empirical estimate)

X : random variable

$\downarrow$

sampling $\hat{X}_n$: unbiased $\mathbb{E}[\hat{X}_n] = X$.

e.g. $\hat{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

by Law of Large Numbers (asymptotic) 渐近:

$n \to \infty$ $\hat{X}_n \xrightarrow{P} X$

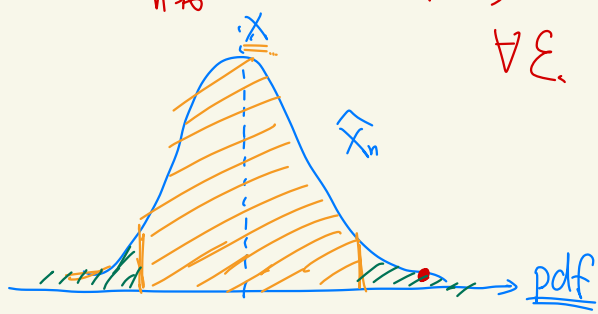
$$\lim_{n\to\infty} \Pr(|\hat{X}_n - X| > \varepsilon) = 0 \quad \forall \varepsilon.$$

Actually: non-asymptotic.

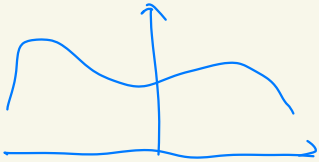
Concentration 集中.

$\downarrow$

Concentration inequality



n small



n large

probability mass (measure)

Generally: $Z_1, \cdots, Z_n$

deviation

$$\Pr\left(\left| \underbrace{f(Z_1, \cdots, Z_n)}_{\text{empirical estimate}} - \underbrace{\mu_n(f)}_{\text{true mean}} \right| > \varepsilon \right) < \delta_n$$

Markov's inequality:

non-negative r.v. Z: $\Pr(Z \geq \varepsilon) \leq \dfrac{\mathbb{E}[Z]}{\varepsilon}$.

$$\Pr(Z \geq t\,\mathbb{E}[Z]) \leq \frac{1}{t}$$

$\Downarrow$

Chebyshev's inequality:

$$\Pr(|Z - \mathbb{E}[Z]| \geq \varepsilon) = \Pr((Z - \mathbb{E}[Z])^2 \geq \varepsilon^2)$$

$$\leq \frac{\mathbb{E}[(Z - \mathbb{E}[Z])^2]}{\varepsilon^2} = \frac{\mathrm{Var}[Z]}{\varepsilon^2}.$$

$Z \sim (\mu, \sigma^2)$: $\Pr(|Z - \mu| \geq \varepsilon) \leq \dfrac{\sigma^2}{\varepsilon^2}$.

$\hat{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$ $\quad Z_i$ i.i.d from P_Z. $\Rightarrow$ $\mathbb{E}[\hat{Z}_n] = \mu$

$$\mathrm{Var}[\hat{Z}_n] = \frac{1}{n}\sigma^2 \Rightarrow \Pr\left(|\hat{Z}_n - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

$$= \mathrm{poly}\left(\frac{1}{n}, \frac{1}{\varepsilon}\right)$$

$$\downarrow$$

$$\exp\left(\frac{1}{n}, \frac{1}{\varepsilon}\right)$$

△ Chernoff's Method:

consider $\mathbb{E}[Z] = 0$

$$\Pr(Z \geq \varepsilon) = \Pr(e^Z \geq e^\varepsilon)$$

$$= \Pr(e^{tZ} \geq e^{t\varepsilon}) \leq \frac{\mathbb{E}[e^{tZ}]}{e^{t\varepsilon}}$$

$\forall t > 0$

MGF Moment Generating Function

$$\Rightarrow \Pr(Z \geq \varepsilon) \leq \inf_{t>0} \frac{\mathbb{E}[e^{tZ}]}{e^{t\varepsilon}}$$

We need to bound

$$\mathbb{E}[e^{tZ}] = \sum_{k=0}^{\infty} \frac{\mathbb{E}[Z^k]}{k!} \cdot t^k$$

Lemma (Hoeffding's Lemma) bounded r.v. Z: $a \leq Z \leq b$.

$$\mathbb{E}[e^{tZ}] \leq e^{t\mu + \frac{t^2}{8}(b-a)^2}$$

Intuition: bounded

$\Downarrow$

bound variance

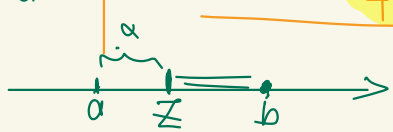
$$\mathrm{Var}[Z] = \mathbb{E}[(Z - \mathbb{E}[Z])^2] \leq \frac{(b-a)^2}{4}$$

Proof (Sketch): wlog. assume $\mu = 0$

$a \leq Z \leq b \rightarrow$ convex combination

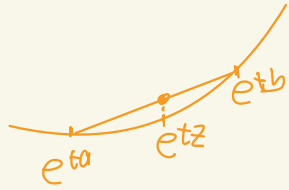
$$Z = \alpha \cdot b + (1-\alpha) \cdot a \qquad \alpha = \frac{Z-a}{b-a}$$

$Z \mapsto e^{tZ}$ is a convex function



By convexity:

$$e^{tZ} \le \frac{Z-a}{b-a}\cdot e^{tb} + \frac{b-Z}{b-a}\cdot e^{ta}.$$



$\mathbb{E}[Z]=0$

$$\Rightarrow \mathbb{E}[e^{tZ}] \le -\frac{a}{b-a}\cdot e^{tb} + \frac{b}{b-a}\cdot e^{ta} = e^{g(u)} \le ?$$

$$g(u) = \log\left(-\frac{a}{b-a}e^{tb} + \frac{b}{b-a}e^{ta}\right)$$

$$= -\gamma u + \log(1-\gamma+\gamma e^{u}) \qquad u = t(b-a).\ \gamma = -\frac{a}{b-a}.$$

Taylor:

$$g(u) = \underbrace{g(0)}_{0} + u\cdot \underbrace{g'(0)}_{0} + \frac{u^2}{2}\cdot g''(\xi). \qquad \xi \in (0,u)$$

$\forall u>0.\ g''(u) \le \frac{1}{4}$

$\left(l(1-l) \le \frac{1}{4} \right)$

$$\le \frac{u^2}{8} = \frac{t(b-a)^2}{8}.$$

$$\Rightarrow \mathbb{E}[e^{tZ}] \le e^{g(u)} \le e^{t^2(b-a)^2/8}$$

$\square$

Theorem 1 (Hoeffding's Inequality) $Z_1, Z_2, \cdots, Z_n$ independent,

$a \le Z_i \le b$. $\mathbb{E}[Z_i] = \mu$.

$$\Pr\left(|\bar{Z}_n - \mu| \ge \varepsilon\right) \le 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}} \nearrow \exp\left(\frac{1}{n}, \frac{1}{\varepsilon}\right)$$

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

Proof: Wlog. assume $\mu = 0$.

(upper tail) $\Pr(\bar{Z}_n \geq \varepsilon) = \Pr\left(\frac{t}{n}\sum_{i=1}^{n} Z_i \geq t\varepsilon\right)$ ↓ Chernoff Method

$$= \Pr\left(e^{\frac{t}{n}\sum_{i=1}^{n} Z_i} \geq e^{t\varepsilon}\right)$$ ↓ Markov's ineq.

$$\leq e^{-t\varepsilon} \cdot \mathbb{E}\left[e^{\frac{t}{n}\sum_{i=1}^{n} Z_i}\right]$$ ↓ Z_i independent

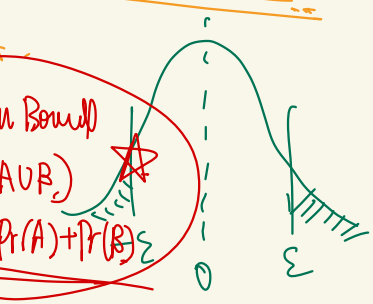
$$= e^{-t\varepsilon} \cdot \prod_{i=1}^{n} \mathbb{E}\left[e^{\frac{t}{n} Z_i}\right]$$ ↓ Hoeffding's Lemma

$$\leq e^{-t\varepsilon} \cdot \prod_{i=1}^{n} e^{\frac{t^2}{n^2}\frac{(b-a)^2}{8}}$$

$$= e^{\frac{t^2}{n}\cdot\frac{(b-a)^2}{8} - t\varepsilon} \quad \forall t > 0$$

$$\Rightarrow \min_t \frac{t^2}{n}\cdot\frac{(b-a)^2}{8} - t\varepsilon \quad \text{Let } t = \frac{4\varepsilon n}{(b-a)^2}$$

$$\Rightarrow \Pr(\bar{Z}_n \geq \varepsilon) \leq e^{-\frac{2n\varepsilon^2}{(b-a)^2}}$$



(lower tail) similarly

Union Bound
$\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$

$$\Rightarrow \Pr(|\bar{Z}_n| \geq \varepsilon) \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}} \quad \square$$

$$\Pr(|\bar{Z}_n - \mu| \geq \varepsilon) \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}$$

$$\Leftrightarrow \boxed{\text{With probability} \geq 1-\delta: \quad |\bar{Z}_n - \mu| \leq (b-a)\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}}$$

constant

Union Bound

Remark: 1) $Z_1, \cdots, Z_n$ independent . not i.i.d !

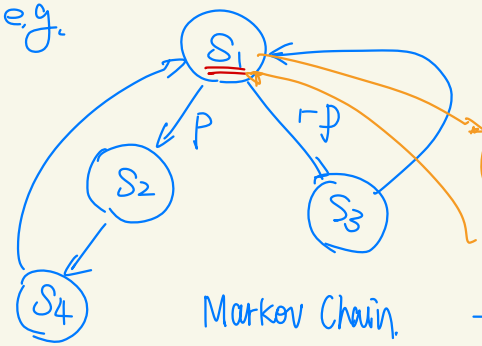
2) n constant

(even if n is a r.v.

$\Rightarrow$ n cannot depend on $Z_1, \cdots, Z_n$.

$\Rightarrow$ Hoeffding's inequality)

e.g.



Markov Chain

$s_1 \to s_2 \to s_4 \to s_1 \to s_3 \to s_1 \to s_2 \to \cdots$

n: # of times we visit s_1

(visitation)

$Z_i \sim \text{Bernoulli}(p)$

Case 1: treat n as # of coin tosses . ✗

Case 2: sample a path until we visit s_1 N times

$\to$ treat N as constant ✓

3) generalized: (McDiarmid's Inequality)

$\forall i$:

$$\sup_{z_1, \cdots, z_n, z_i'} \left| f(z_1, \cdots, z_{i-1}, z_i', z_{i+1}, \cdots, z_n) - f(z_1, \cdots, z_{i-1}, z_i, z_{i+1}, \cdots, z_n) \right| \le c_i$$

$$\Rightarrow \Pr\left(\left| f(Z_1, \cdots, Z_n) - \mathbb{E}[f(Z_1, \cdots, Z_n)] \right| \ge \varepsilon \right) \le 2e^{-2\varepsilon^2 / \sum_{i=1}^{n} c_i^2}$$

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

Low-Variance R.V $\rightarrow$ Tighter Bound

Lemma $|X| \le C$, $\mathbb{E}[X]=0$, $\forall t>0$:

$$\mathbb{E}[e^{tX}] \le \exp\left(t^2\sigma^2 \cdot \frac{e^{tC}-1-tC}{(tC)^2}\right)$$

Hoeffding: $t^2 \cdot \frac{(b-a)^2}{8}$

Proof: Let $F = \sum_{r=2}^{\infty} \frac{t^{r-2}\cdot \mathbb{E}[X^r]}{r!\cdot\sigma^2}$

$$\Rightarrow \mathbb{E}[e^{tX}] = \mathbb{E}\left[1+tX+\sum_{r=2}^{\infty}\frac{t^rX^r}{r!}\right] = 1+t^2\sigma^2 F \le e^{t^2\sigma^2 F}$$

It suffices to show: $F \le \frac{e^{tC}-1-tC}{(tC)^2}$

$$\mathbb{E}[X^r] = \mathbb{E}[X^{r-2}\cdot X^2] \le C^{r-2}\cdot\sigma^2$$

$$F = \sum_{r=2}^{\infty}\frac{t^{r-2}\cdot\mathbb{E}[X^r]}{r!\cdot\sigma^2} \le \sum_{r=2}^{\infty}\frac{t^{r-2}\cdot C^{r-2}\cdot\cancel{\sigma^2}}{r!\,\cancel{\sigma^2}} = \frac{1}{(tC)^2}\sum_{r=2}^{\infty}\frac{(tC)^r}{r!}$$

$$= \frac{1}{(tC)^2}\left(e^{tC}-1-tC\right) \quad \square$$

Theorem 2 (Bernstein's Inequality) $|X_i| \le C$, $\mathbb{E}[X_i]=\mu$.

$$\Pr(|\bar{X}_n-\mu| \ge \varepsilon) \le 2\cdot\exp\left(-\frac{n\varepsilon^2}{2\sigma^2+2C\varepsilon/3}\right) = \delta$$

$2\exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right)$

<u>Note</u>: When variance σ^2 is small

$\Rightarrow$ Bernstein better than Hoeffding. $\left(\exp\left(\frac{1}{\varepsilon}\right)\ \exp\left(\frac{1}{\varepsilon^2}\right)\right)$

$\Leftrightarrow$ with probability $\geq 1-\delta$:

$$|\bar{X}_n - \mu| \leq \sqrt{\frac{2\sigma^2 \cdot \log(1/\delta)}{n}} + \frac{2c \cdot \log(1/\delta)}{3n}$$

dominate term

In particular, if $\sigma^2 \leq \sqrt{(2c^2 \log 1/\delta)/(9n)}$

w.p. $\geq 1-\delta$: $|\bar{X}_n - \mu| \leq \frac{4c \cdot \log \frac{1}{\delta}}{3n}$.

w.p. $\geq 1-\delta$ $\begin{cases} \text{Hoeffding: } O\left(\sqrt{\frac{1}{n}}\right) \\ \text{Bernstein: } O\left(\frac{1}{n}\right) \end{cases}$ (with low variance)

Gaussian

$X_1, \cdots, X_n \sim N(\mu, \sigma^2)$

$$\Pr(|\bar{X}_n - \mu| \geq \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right)$$

"$1\sigma - 2\sigma - 3\sigma$"
79 95 99.7.

Proof: Wlog assume $\mu = 0$.

(upper tail) $\Pr(\bar{X}_n \geq \varepsilon)$

$$= \Pr\left(e^{t\sum_{i=1}^n X_i} \geq e^{tn\varepsilon}\right)$$

$$\leq e^{-tn\varepsilon} \cdot \prod_{i=1}^n \mathbb{E}[e^{tX_i}]$$

$$\leq e^{-tn\varepsilon} \cdot \exp\left(nt^2\sigma^2 \cdot \frac{e^{tc} - 1 - tc}{(tc)^2}\right)$$

$\forall t > 0$

Let $t = \frac{1}{c} \cdot \log\left(1 + \frac{\varepsilon c}{\sigma^2}\right)$.

$$\Rightarrow \Pr(\bar{X}_n \geq \varepsilon) \leq \exp\left(-\frac{n\sigma^2}{c^2} \cdot h\left(\frac{c\varepsilon}{\sigma^2}\right)\right)$$

$h(u) = (1+u) \cdot \log(1+u) - u$

$$\boxed{h(u) \geq \frac{u^2}{2 + \frac{2}{3}u}}$$

$$\Rightarrow \Pr(\bar{X}_n \geq \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2 + \frac{2}{3}c\varepsilon}\right) \quad \square$$

Recall: (Hoeffding's inequality) independent r.v. $X_i \in [a,b]$, $\mathbb{E}[X_i] = \mu$.

$$\Rightarrow \Pr\left(|\bar{X}_n - \mu| \geq \varepsilon \right) \leq 2e^{-2n\varepsilon^2/(b-a)^2}$$

$$\Leftrightarrow \text{w.p.} \geq 1-\delta: \quad |\bar{X}_n - \mu| \leq (b-a)\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}.$$

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

Example 1. Supervised Learning

binary classification $<$ $\mathcal{X}$: feature space
$\mathcal{Y}$: label space. $\{0,1\}$

hypothesis

goal: learn $f: \mathcal{X} \to \mathcal{Y}$ with small loss (high accuracy)

$$(X_i, Y_i) \to \ell_f(X_i, Y_i) = \mathbb{1}\{f(X_i) \neq Y_i\}$$

i.i.d 独立同分布

$$\boxed{D = \{(X_i, Y_i)\}_{i=1}^{n}}$$

$(X_i, Y_i) \overset{i.i.d}{\sim} P_{X,Y}$ (unknown)

Hypothesis Class: $\mathcal{F}: \mathcal{X} \to \mathcal{Y}$. assume $|\mathcal{F}| < \infty$

Population Risk: $\mathcal{L}(f) = \mathbb{E}_{(X,Y)\sim P_{X,Y}}\left[\mathbb{1}\{f(X) \neq Y\}\right]$ ← unknown!

$$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \mathcal{L}(f)$$

Empirical Risk Minimization (ERM): $\boxed{\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \hat{\mathcal{L}}(f)}$ ☆ ← double

$$\hat{\mathcal{L}}(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{f(X_i) \neq Y_i\} \triangleq \hat{\mathbb{E}}[\mathbb{1}\{f(X) \neq Y\}]$$

Guarantee : compare $\hat{f}$ and f^*

$\mathcal{L}(f)$ small ×

restricted to function class $\mathcal{F}$

Excess Risk: $\mathcal{L}(\hat{f}) - \mathcal{L}(f^*)$

$$= \mathbb{E}[\mathbb{1}\{\hat{f}(X) \neq Y\}] - \mathbb{E}[\mathbb{1}\{f^*(X) \neq Y\}]$$

$$= \mathbb{E}[\mathbb{1}\{\hat{f}(X) \neq Y\}] - \hat{\mathbb{E}}[\mathbb{1}\{\hat{f}(X) \neq Y\}]$$

$$+ \underbrace{\hat{\mathbb{E}}[\mathbb{1}\{\hat{f}(X) \neq Y\}]}_{\leq \hat{\mathbb{E}}[\mathbb{1}\{f^*(X) \neq Y\}]} - \mathbb{E}[\mathbb{1}\{f^*(X) \neq Y\}]$$

☆ ($\hat{f}$ is ERM)

$$\leq 2 \max_{f \in \mathcal{F}} \left| \underbrace{\mathbb{E}[\mathbb{1}\{f(X) \neq Y\}] - \hat{\mathbb{E}}[\mathbb{1}\{f(X) \neq Y\}]}_{\text{generalization gap}} \right|$$

Uniform Convergence : bounding excess risk by generalization gap

$\mathbb{1}\{f(X) \neq Y\}$: i.i.d r.v. bounded in $[0,1]$

By Hoeffding: $\forall f \in \mathcal{F}$: w.p. $\geq 1-\delta$: (single-hypo)

$$\left| \mathbb{E}[\mathbb{1}\{f(X) \neq Y\}] - \hat{\mathbb{E}}[\mathbb{1}\{f(X) \neq Y\}] \right| \leq \sqrt{\frac{1}{2n} \cdot \log \frac{2}{\delta}}$$

$\Rightarrow$ w.p. $\geq 1-\delta$: $\forall f \in \mathcal{F}$. (all-hypo)

$$\left| \mathbb{E}[\mathbb{1}\{f(X) \neq Y\}] - \hat{\mathbb{E}}[\mathbb{1}\{f(X) \neq Y\}] \right| \leq \sqrt{\frac{1}{2n} \cdot \log \frac{2|\mathcal{F}|}{\delta}}$$

$$\Pr\left(|\mathcal{L}_f - \hat{\mathcal{L}}_f| \geq \varepsilon \right) \leq e^{-2n\varepsilon^2} = \delta \to \delta/|\mathcal{F}| \quad \forall f \in \mathcal{F}$$

$$\Pr\left(\exists f \in \mathcal{F} : |\mathcal{L}_f - \hat{\mathcal{L}}_f| \geq \varepsilon \right)$$

$$= \Pr\left(\bigcup_{f \in \mathcal{F}} |\mathcal{L}_f - \hat{\mathcal{L}}_f| \geq \varepsilon \right) \quad \text{Union Bound}$$

$$\leq \sum_{f \in \mathcal{F}} \Pr\left(|\mathcal{L}_f - \hat{\mathcal{L}}_f| \geq \varepsilon \right) = |\mathcal{F}| \cdot \delta \to \delta.$$

$\Rightarrow$ Excess Risk: $\mathcal{L}(\hat{f}) - \mathcal{L}(f^*) \leq 2 \cdot \sqrt{\frac{1}{2n} \cdot \log \frac{2|\mathcal{F}|}{\delta}}$ finite

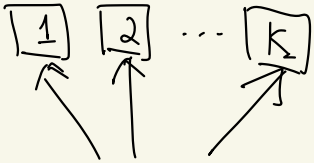
For $|\mathcal{F}| = \infty$ $\to$ complexity control

(covering number, Rademacher complexity)

Example 2. MAB (Multi-Armed Bandits)

stochastic bandits

K arms $\rightarrow$ K distribution $\{R_i\}_{i=1}^{K}$



R_i < bounded supported in $[0,1]$ / mean: μ_i

for round $t = 1, 2, \cdots, T$: ① choose arm $i_t \in [K]$

② receive reward $r_t \sim R_{i_t}$

Goal 1: minimize pseudo-regret.

$\mu^* = \max_{i \in [K]} \mu_i$

$$\text{Regret}_T = \sum_{t=1}^{T} (\mu^* - \mu_{i_t})$$

(online)

exploration-exploitation

Goal 2: BAI (Best Arm Identification)

$\{(i_t, r_t)\}_{t=1}^{T} \rightarrow \hat{i}$

pure-exploration

regret: $\mu^* - \mu_{\hat{i}}$.

BAI, whether $\hat{i} \in \arg\max_{i \in [K]} \mu_i$

1. Uniform Sampling: choose each arm T/K numbers

(assume $\frac{T}{K} \in \mathbb{N}$)

$\rightarrow \hat{i}$: arm with highest empirical mean

$$\hat{\mu}_i = \frac{1}{T/K} \cdot \sum_{t: i_t = i} r_t$$

$$\boxed{\hat{i} = \arg\max_{i \in [K]} \hat{\mu}_i}$$

"empirical risk minimizer"

$\mu^* - \mu_{\hat{i}}$: "excess risk"

$$= \mu^* - \hat{\mu}_{\hat{i}} + \hat{\mu}_{\hat{i}} - \mu_{\hat{i}}$$

Uniform Convergence

$$\leq \mu^* - \hat{\mu}_* + \hat{\mu}_{\hat{i}} - \mu_{\hat{i}} \leq 2 \cdot \max_{i \in [K]} |\mu_i - \hat{\mu}_i|$$

by Hoeffding

By Hoeffding: $\Pr(|\hat{\mu}_i - \mu_i| \geq \varepsilon) \leq 2e^{-2T\varepsilon^2/K}, \quad \forall i$

$$\Rightarrow \Pr(\exists i: |\hat{\mu}_i - \mu_i| \geq \varepsilon)$$

$$\leq \sum_{i=1}^{K} \Pr(|\hat{\mu}_i - \mu_i| \geq \varepsilon) \leq 2K \cdot e^{-2T\varepsilon^2/K} = \delta.$$

$$\Rightarrow \text{w.p.} \geq 1-\delta: \forall i \in [K], |\hat{\mu}_i - \mu_i| \leq \sqrt{\frac{K}{2T} \cdot \log \frac{2K}{\delta}}.$$

$$\Rightarrow \mu^* - \mu_{\hat{i}} \leq 2 \cdot \sqrt{\frac{K}{2T} \cdot \log \frac{2K}{\delta}}$$

(regret)

In order to achieve $\mu^* - \mu_{\hat{i}} \leq \varepsilon$,

(sample complexity)

w.p. $\geq 1-\delta$: we need $T = O\left(\frac{K}{\varepsilon^2} \cdot \log \frac{K}{\delta}\right)$

minimax optimal

Remark: dependency $\begin{cases} T: & o(1) \,/\, o(T) \text{ sublinear} \\ K & \text{tricky} \end{cases}$

Recall: Certainty Equivalence ← model-based

① estimate an MDP model from data

② perform policy evaluation / optimization in the estimated model

Batch learning

$M = (S, A, P, R, \gamma)$

Data $D = \{(s_1, a_1, r_1, s_2, a_2, r_2, \dots, s_H, a_H, r_H, s_{H+1})\}$

trajectory

$\downarrow$

$(s_1, a_1, r_1, s_2), (s_2, a_2, r_2, s_3), \dots, (s_H, a_H, r_H, s_{H+1})$

R P

(model-free)

Value-Based Q-Learning:

$$Q(s,a) \leftarrow (1-\alpha)\cdot Q(s,a) + \alpha\cdot\left(r + \gamma \cdot \max_{a'} Q(s',a')\right)$$

① online learning

② space complexity $O(|S|\cdot|A|)$

③ function approximation

Protocol: for each (s,a) pair:

draw n i.i.d samples of $\begin{cases} r \sim R(s,a) \\ s' \sim P(\cdot \mid s,a) \end{cases}$

$\Rightarrow$ dataset $D_{s,a}$

Goal: We want to learn a policy $\hat{\pi}$ from data $\{D_{s,a}\}_{(s,a)\in S\times A}$.

s.t. $\boxed{\| V_M^* - V_M^{\hat{\pi}} \|_\infty \le \varepsilon = f(n)}$

suboptimality gap

△ Given ε: how much data (n) do we need ?

Given n: how much accuracy (ε) can we get ?

A straightforward algorithm:

$\hat{P}$ and $\hat{R}$: MLE

$$\hat{R}(s,a) = \frac{1}{n} \sum_{r_i \in D_{s,a}} r_i$$

$$\hat{P}(s'|s,a) = \frac{1}{n} \sum_{s'_i \sim D_{s,a}} \mathbb{1}\{\underset{\text{r.v.}}{s'_i} = \underset{\text{const}}{s'}\}$$

$$\Updownarrow$$

$$\triangle \quad \hat{P}(s,a) = \frac{1}{n} \sum_{s'_i \sim D_{s,a}} \vec{e}_{s'_i}$$

unit vector $\in \mathbb{R}^{|S|}$.

$\Rightarrow$ Certainty equivalence model:

$$\hat{M} = (S, A, \hat{P}, \hat{R}, \gamma) \rightarrow \text{policy } \pi^*_{\hat{M}} = \hat{\pi}$$

Analysis Framework:

① concentration: $\hat{M} \approx M$

estimation error

w.p. $\geq 1-\delta$: $\forall (s,a)$

$$\begin{cases} |\hat{R}(s,a) - R(s,a)| \leq \varepsilon_R \\ \|\hat{P}(s,a) - P(s,a)\|_1 \leq \varepsilon_P \end{cases}$$

② error propagation analysis

☆ 误差传播分析

$$\|V^*_M - V^{\hat{\pi}}_M\|_\infty \leq f_n(\varepsilon_R, \varepsilon_P)$$

Concentration Analysis: (by Hoeffding)

$\forall (s,a) \in S \times A$: w.p. $\geq 1-\delta$: $|\hat{R}(s,a) - R(s,a)| \leq R_{max}\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$.

$\forall (s,a,s') \in S \times A \times S$: w.p. $\geq 1-\delta$: $|\hat{P}(s'|s,a) - P(s'|s,a)| \leq \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$.

Union Bound.

total failure budget δ
- $\rightarrow \frac{\delta}{2}$ reward $\rightarrow$ evenly over $S \times A$. $\frac{\delta}{2} / |S| \times |A|$
- $\rightarrow \frac{\delta}{2}$ transition $\rightarrow$ evenly over $S \times A \times S$.

$\Rightarrow$ w.p. $\geq 1-\delta$: $\forall (s,a)$

$$|\hat{R}(s,a) - R(s,a)| \leq R_{max}\sqrt{\frac{1}{2n}\cdot\log\frac{4\cdot|S|\times|A|}{\delta}} = \varepsilon_R$$

$$|\hat{P}(s'|s,a) - P(s'|s,a)| \leq \sqrt{\frac{1}{2n}\log\frac{4\cdot|S|\times|A|\times|S|}{\delta}}$$

$$\Rightarrow \|\hat{P}(s,a) - P(s,a)\|_1 \leq |S|\cdot\sqrt{\frac{1}{2n}\cdot\log\frac{4\cdot|S|\times|A|\times|S|}{\delta}} = \varepsilon_P$$

$\|x\|_1 = \sum_i |x_i|$

Error Propagation Analysis

Lemma (Simulation lemma) $\forall$ deterministic policy $\pi: S \rightarrow A$.

$$\|V^{\pi}_{\hat{M}} - V^{\pi}_{M}\|_\infty \leq \frac{\varepsilon_R}{1-\gamma} + \frac{\gamma}{2(1-\gamma)}\cdot\varepsilon_P\cdot V_{max}$$

simulated model

$$\|V_M^* - V_M^{\pi^*_{\hat{M}}}\|_\infty = \|V_M^{\pi^*_M} - V_{\hat{M}}^{\pi^*_{\hat{M}}} + V_{\hat{M}}^{\pi^*_{\hat{M}}} - V_M^{\pi^*_{\hat{M}}}\|_\infty$$

optimal policy for $\hat{M}$.

local greedy in estimated model

$$\le \|V_M^{\pi^*_M} - V_{\hat{M}}^{\pi^*_M} + V_{\hat{M}}^{\pi^*_{\hat{M}}} - V_M^{\pi^*_{\hat{M}}}\|_\infty \qquad \left(V_{\hat{M}}^{\pi^*_{\hat{M}}} \ge V_{\hat{M}}^{\pi^*_M}\right)$$

$$\le \|V_M^{\pi^*_M} - V_{\hat{M}}^{\pi^*_M}\|_\infty + \|V_{\hat{M}}^{\pi^*_{\hat{M}}} - V_M^{\pi^*_{\hat{M}}}\|_\infty$$

uniform convergence

$$\le 2 \cdot \max_{\pi: S \to A} \|V_M^\pi - V_{\hat{M}}^\pi\|_\infty$$

Simulation Lemma

$$\le 2 \cdot \frac{\varepsilon_R}{1-\gamma} + \frac{\gamma}{1-\gamma} \cdot \varepsilon_P \cdot V_{max}$$

$$\Rightarrow \|V_M^* - V_M^{\hat{\pi}}\|_\infty \le \tilde{O}\left(|S| \cdot \frac{V_{max}}{\sqrt{n}\,(1-\gamma)}\right)$$

effective horizon

Suboptimality gap	Value error	estimation error
$\|V_M^* - V_M^\pi\|_\infty$	$\|V_M^\pi - V_{\hat{M}}^\pi\|_\infty$	$\varepsilon_R, \varepsilon_P$
by "ERM + Uniform Convergence"	by simulation lemma	by concentration

Proof of Simulation Lemma: $\forall$ deterministic policy $\pi: S \to A$

$$\|V_{\hat{M}}^\pi - V_M^\pi\|_\infty \le \frac{\varepsilon_R}{1-\gamma} + \frac{\gamma \cdot \varepsilon_P}{2(1-\gamma)} \cdot V_{max}$$

Proof: $\forall s \in S$: $V_{\hat{M}}^\pi(s) - V_M^\pi(s)$

$$= \left| \hat{R}(s,\pi) + \gamma \cdot \langle \hat{P}(s,\pi), V^{\pi}_{\hat{M}} \rangle - R(s,\pi) - \gamma \cdot \langle P(s,\pi), V^{\pi}_{M} \rangle \right|$$

$$\leq \varepsilon_R + \gamma \cdot \left| \langle \hat{P}(s,\pi), V^{\pi}_{\hat{M}} \rangle - \langle P(s,\pi), V^{\pi}_{M} \rangle \right|$$

$$= \varepsilon_R + \gamma \cdot \left| \langle \hat{P}(s,\pi) - P(s,\pi), V^{\pi}_{\hat{M}} \rangle + \langle P(s,\pi), V^{\pi}_{\hat{M}} - V^{\pi}_{M} \rangle \right|$$

$$\leq \varepsilon_R + \gamma \cdot \left| \langle \hat{P}(s,\pi) - P(s,\pi), V^{\pi}_{\hat{M}} \rangle \right| + \gamma \cdot \| V^{\pi}_{\hat{M}} - V^{\pi}_{M} \|_\infty$$

Hölder's inequality : $|\langle u, v \rangle| \leq \|u\|_p \cdot \|v\|_q$.

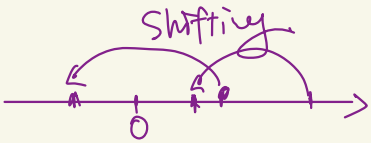
p, q conjugate: $\frac{1}{p} + \frac{1}{q} = 1$

$$\leq \varepsilon_R + \gamma \cdot \| \hat{P}(s,\pi) - P(s,\pi) \|_1 \, \| V^{\pi}_{\hat{M}} \|_\infty + \gamma \cdot \| V^{\pi}_{\hat{M}} - V^{\pi}_{M} \|_\infty$$

$$= \varepsilon_R + \gamma \cdot \varepsilon_p \cdot V_{max} + \gamma \cdot \| V^{\pi}_{\hat{M}} - V^{\pi}_{M} \|_\infty$$

$$\Rightarrow \| V^{\pi}_{\hat{M}} - V^{\pi}_{M} \|_\infty \leq \varepsilon_R + \gamma \cdot \varepsilon_p \cdot V_{max} + \gamma \cdot \| V^{\pi}_{\hat{M}} - V^{\pi}_{M} \|_\infty$$

$$\Rightarrow \| V^{\pi}_{\hat{M}} - V^{\pi}_{M} \|_\infty \leq \frac{\varepsilon_R}{1-\gamma} + \frac{\gamma}{1-\gamma} \cdot \varepsilon_p \cdot V_{max} \quad \square$$



$$\left\| V^{\pi}_{\hat{M}} - \frac{V_{max}}{2} \cdot \vec{1} \right\|_\infty$$

$$\leq \frac{V_{max}}{2} \rightarrow \frac{\gamma}{2(1-\gamma)} \cdot \varepsilon_p V_{max}$$

Alternative "Modern" Proof of Simulation Lemma

Claim: $\forall f \in \mathbb{R}^S$, $s_0 \in S$

Bellman Error

$$\star \quad \underbrace{f(s_0) - V_M^\pi(s_0)}_{J_f(\pi) - J(\pi)} = \frac{1}{1-\gamma} \cdot \boxed{\mathbb{E}\left[f(s) - r - \gamma f(s') \right]}$$

$s \sim d_M^{\pi, s_0}$, $a \sim \pi(\cdot|s)$, $r \sim R(s,a)$, $s' \sim P(\cdot|s,a)$

TD Error

$$f(s) - (\mathcal{T}_M^\pi f)(s)$$

$$f(s_0) - V_M^\pi(s_0) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{s \sim d_M^{\pi, s_0}}\left[(f - \mathcal{T}_M^\pi f)(s) \right]$$

discounted occupancy from s_0.

$$d^{\pi, s_0} = (1-\gamma) \cdot \sum_{t=1}^{\infty} d_t^{\pi, s_0}$$

$$d_t^{\pi, s_0} = \Pr(s_t = \cdot \mid s_0)$$

Corollary

$$\boxed{\|f - V_M^\pi\|_\infty \le \frac{1}{1-\gamma} \|f - \mathcal{T}_M^\pi f\|_\infty} \quad \star$$

$$\|f - V_M^*\|_\infty \le \frac{1}{1-\gamma} \cdot \|f - \mathcal{T}_M f\|_\infty$$

$f \in \mathbb{R}^{S \times A}$:

$$\|f - Q_M^\pi\|_\infty \le \frac{1}{1-\gamma} \cdot \|f - \mathcal{T}_M^\pi f\|_\infty$$

$$\star\star\star \quad \|f - Q_M^*\|_\infty \le \frac{1}{1-\gamma} \|f - Q_M^* f\|_\infty$$

bounding value error by Bellman error.

(paying an effective horizon)

Proof of Claim:

$$V_M^\pi(s_0) = \mathbb{E}_{M,\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_0 \right]$$

$$= \sum_{t=1}^{\infty} \gamma^{t-1} \cdot \mathbb{E}_{M,\pi}[r_t \mid s_0]$$

$\mathbb{E}[r]$

discounted average via t

$s \sim d_t^{\pi, s_0}$, $a \sim \pi(\cdot|s)$, $r \sim R(s,a)$

$$\Rightarrow V_M^\pi(s_0) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{\substack{s\sim d^{\pi,s_0} \\ a\sim\pi(\cdot|s) \\ r\sim R(s,a)}}[r]$$

$$s\sim d^{\pi,s_0}, \quad a\sim\pi(\cdot|s), \quad s'\sim p(\cdot|s,a)$$

It suffices to show:

$$f(s_0) = \frac{1}{1-\gamma} \cdot \mathbb{E}[f(s) - \gamma f(s')]$$

$$\text{RHS} = \sum_{t=1}^{\infty} \gamma^{t-1} \cdot \mathbb{E}_{s\sim d_t^{\pi,s_0}}[f(s) - \gamma \cdot f(s')]$$

$$s\sim d_t^{\pi,s_0}, \quad a\sim\pi(\cdot|s), \quad s'\sim p(\cdot|s,a) \to s'\sim d_{t+1}^{\pi,s_0}$$

$$= \mathbb{E}_{\substack{s\sim d_1^{\pi,s_0} \\ s'\sim d_2^{\pi,s_0}}}[f(s) - \cancel{\gamma f(s')}]$$

$$+ \mathbb{E}_{\substack{s\sim d_2^{\pi,s_0} \\ s'\sim d_3^{\pi,s_0}}}[\cancel{\gamma \cdot f(s)} - \cancel{\gamma^2 f(s')}]$$

$$+ \mathbb{E}_{\substack{s\sim d_3^{\pi,s_0} \\ s'\sim d_4^{\pi,s_0}}}[\cancel{\gamma^2 f(s)} - \cancel{\gamma^3 f(s')}]$$

$$+ \cdots \qquad = \mathbb{E}_{s\sim d_1^{\pi,s_0}}[f(s)]$$

$$= f(s_0) = \text{LHS} \quad \square$$

Similar Analysis: Bellman Flow Equation

Let $f = V_{\hat{M}}^\pi \Rightarrow$

$$\|V_{\hat{M}}^\pi - V_M^\pi\|_\infty \le \frac{1}{1-\gamma} \cdot \|V_{\hat{M}}^\pi - \mathcal{T}_M^\pi V_{\hat{M}}^\pi\|_\infty$$

$$= \frac{1}{1-\gamma} \cdot \|\mathcal{T}_{\hat{M}}^\pi V_{\hat{M}}^\pi - \mathcal{T}_M^\pi V_{\hat{M}}^\pi\|_\infty$$

$$= \frac{1}{1-\gamma} \cdot \|\hat{R}(s,\pi) + \gamma \cdot \langle \hat{P}(s,\pi), V_{\hat{M}}^\pi \rangle - R(s,\pi) - \gamma \cdot \langle P(s,\pi), V_{\hat{M}}^\pi \rangle\|_\infty$$

$$\leq \frac{1}{1-\gamma}\cdot\left(\varepsilon_R + \gamma\cdot \underline{\langle \hat{P}(s,\pi) - P(s,\pi), V^{\pi}_{\hat{M}} \rangle}\right)$$

$$\leq \frac{1}{1-\gamma}\left(\varepsilon_R + \gamma\cdot\varepsilon_P\cdot V_{max}\right) \rightarrow \text{simulation lemma.}$$

<u>Recall</u>: tabular CE $\hat{M} = (S, A, \underline{\hat{P}}, \underline{\hat{R}}, \gamma)$

$$\Rightarrow \pi^*_{\hat{M}} = \hat{\pi}.$$

<u>error propagation</u>: $\| V^*_M - V^{\pi^*_{\hat{M}}}_M \|_\infty$

$$\leftarrow \| V^{\pi}_M - V^{\pi}_{\hat{M}} \|_\infty$$

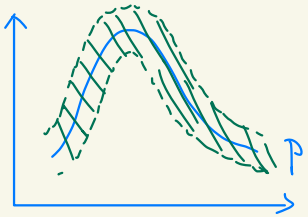
$$\leftarrow \begin{cases} |\hat{R}(s,a) - R(s,a)| \leq \varepsilon_R \\ \underline{\|\hat{P}(s,a) - P(s,a)\|_1} \leq \varepsilon_P \end{cases}$$

$$\| V^*_M - V^{\pi^*_{\hat{M}}}_M \|_\infty \leq \tilde{O}\left(\frac{|S|\cdot V_{max}}{\sqrt{n}\cdot(1-\gamma)}\right)$$

loose

△ Improving $|S|$ to $\sqrt{|S|}$

$$\max_{s,a,s'} \left| \hat{P}(s'|s,a) - P(s'|s,a) \right| \leq \frac{\varepsilon_P}{|S|}$$



$\hat{P}$: any curve in band ?

× valid <u>probability distribution</u>

$\Rightarrow$ <u>$\|\cdot\|_\infty$</u> doesn't catch such dependency

Claim: $\forall v \in \mathbb{R}^{|S|}$: $\|v\|_1 = \max_{u \in \{-1,1\}^{|S|}} u^T v$

$\|v\|_1 = |v_1| + |v_2| + \cdots + |v_{|S|}|$

$\rightarrow 2^{|S|}$ 种不同的 u

$$\Rightarrow \|\hat{P}(s,a) - P(s,a)\|_1 = \max_{u \in \{-1,1\}^{|S|}} u^T(\hat{P}(s,a) - P(s,a))$$

$$= u^T \frac{1}{n} \sum_{s' \in D_{s,a}} \vec{e}_{s'} - u^T P(s,a)$$

$$= \frac{1}{n} \sum_{s' \in D_{s,a}} u(s') - \mathbb{E}_{s' \sim P(s,a)}[u(s')]$$

Hoeffding

$$\leq 2 \cdot \sqrt{\frac{1}{2n} \log \frac{2}{\delta}} \qquad \forall (s,a) \in S \times A.$$

w.p. $\geq 1-\delta$

By Union Bound: $\|\hat{P}(s,a) - P(s,a)\|_1 \leq 2 \cdot \sqrt{\frac{1}{2n} \log \frac{2|S| \times |A| \times 2^{|S|}}{\delta}}$

error bound $= \tilde{O}\left(\sqrt{\frac{|S|}{n}} \cdot \frac{V_{max}}{1-\gamma}\right)$

△ No dependency on $|S|$

↳ $O(1)$ sample per (s,a) pair to get a near-optimal error rate

$P(s'|s,a)$

s' 有 $|S|$ 种可能性

⇓

Actually: pay another factor of effective horizon

$$\Rightarrow \text{error bound} = \tilde{O}\left(\frac{V_{max}}{\sqrt{n}(1-\gamma)^2}\right)$$

$$\|V_M^* - V_M^{\pi^*_{\hat{M}}}\|_\infty \le \frac{2}{1-\gamma} \cdot \|Q_M^* - Q_{\hat{M}}^*\|_\infty$$

baseline

$$\le \frac{2}{(1-\gamma)^2} \cdot \|Q_M^* - \mathcal{T}_{\hat{M}} Q_M^*\|_\infty$$

Bellman Error.

$$\|V^* - V^{\pi_f}\|_\infty \le \frac{2\|Q^* - f\|_\infty}{1-\gamma}$$

$$\|Q_M^* - Q_{\hat{M}}^*\|_\infty = \|Q_M^* - \mathcal{T}_{\hat{M}} Q_M^* + \mathcal{T}_{\hat{M}} Q_M^* - \mathcal{T}_{\hat{M}} Q_{\hat{M}}^*\|_\infty$$

$$\le \|Q_M^* - \mathcal{T}_{\hat{M}} Q_M^*\|_\infty + \|\mathcal{T}_{\hat{M}} Q_M^* - \mathcal{T}_{\hat{M}} Q_{\hat{M}}^*\|_\infty$$

$$\le \|Q_M^* - \mathcal{T}_{\hat{M}} Q_M^*\|_\infty + \gamma \cdot \|Q_M^* - Q_{\hat{M}}^*\|_\infty$$

$$\Rightarrow \|Q_M^* - Q_{\hat{M}}^*\|_\infty \le \frac{1}{1-\gamma} \cdot \|Q_M^* - \mathcal{T}_{\hat{M}} Q_M^*\|_\infty$$

$$\forall (s,a): \left|(\mathcal{T}_M Q_M^* - \mathcal{T}_{\hat{M}} Q_M^*)(s,a)\right|$$

$$= \left| Q_M^*(s,a) - \left(\hat{R}(s,a) - \gamma \langle \hat{P}(s,a), V_M^* \rangle\right) \right|$$

$$= \left| \left(R(s,a) + \gamma \cdot \langle P(s,a), V_M^* \rangle\right) - \left(\hat{R}(s,a) + \gamma \cdot \langle \hat{P}(s,a), V_M^* \rangle\right) \right|$$

$\rightarrow$ true mean

$\rightarrow$ estimated mean

Hoeffding $\le V_{max} \cdot \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}$.

↳ 直接对Q作 concentration，无需引入 $\varepsilon_R, \varepsilon_P$

$$\Rightarrow \| V_M^* - V_M^{\pi_{\hat{M}}^*} \|_\infty \le \tilde{O}\left(\frac{V_{max}}{\sqrt{n}(1-\gamma)^2} \right)$$

$$\| V_M^* - V_M^{\hat{\pi}} \|_\infty \xrightarrow{\text{Way 1}} \| V_M^{\pi} - V_{\hat{M}}^{\pi} \|_\infty \longleftarrow \varepsilon_R, \varepsilon_P \, (|S|)$$

$$\| V_M^* - V_M^{\hat{\pi}} \|_\infty \xrightarrow{\text{Way 2}} \| Q_M^* - Q_{\hat{M}}^* \|_\infty \longleftarrow \underbrace{\| Q_M^* - \mathcal{T}_{\hat{M}} Q_M^* \|_\infty}_{\Delta}$$

<u>Remark</u>: we can also get

$$\| Q_M^* - Q_{\hat{M}}^* \|_\infty \le \frac{1}{1-\gamma} \cdot \| Q_{\hat{M}}^* - \mathcal{T}_M Q_{\hat{M}}^* \|_\infty$$

$$= \frac{1}{1-\gamma} \cdot \| \mathcal{T}_{\hat{M}} Q_{\hat{M}}^* - \mathcal{T}_M Q_{\hat{M}}^* \|_\infty$$

$$= \left| \left(\hat{R}(s,a) + \gamma \cdot \langle \hat{P}(s,a), V_{\hat{M}}^* \rangle \right) - \left(R(s,a) + \gamma \cdot \langle P(s,a), V_{\hat{M}}^* \rangle \right) \right|$$

$V_{\hat{M}}^*$ still depends on data!

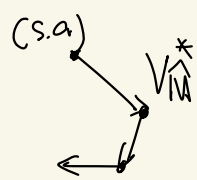
from data

→ CANNOT APPLY HOEFFDING!

We need <u>deterministic function class</u> (V_M^*)

* Sometimes can work.

e.g. $M(S)$ is a directed acyclic graph (DAG)



* Sparse Sampling $\rightarrow$ MCTS (Monte Carlo Tree Search)

Best-of-Both Worlds?

Way 1 $\rightarrow$ extra factor of $|S|$

Way 2 $\rightarrow$ extra factor of $1/(1-\gamma)$

Answer: Yes! But in large-sample regime.

$\|Q^*_M - Q^*_{\hat{M}}\|_\infty$ $\rightarrow$ $\frac{1}{1-\gamma} \cdot \|Q^*_M - \mathcal{T}_{\hat{M}} Q^*_M\|_\infty$ ✓

$\rightarrow$ $\frac{1}{1-\gamma} \cdot \|Q^*_{\hat{M}} - \mathcal{T}_M Q^*_{\hat{M}}\|_\infty$ ✗

Lemma For any $f \in \mathbb{R}^{S \times A}$, initial state s_0, policy π:

$$V^\pi(s_0) - V^{\pi_f}(s_0) \le \frac{1}{1-\gamma} \cdot \left(\mathbb{E}_{d^{\pi, s_0}}[\mathcal{T} f - f] + \mathbb{E}_{d^{\pi_f, s_0}}[f - \mathcal{T} f] \right)$$

Proof: (Q-variant of Claim) $\mathbb{E}_\mu[f] = \mathbb{E}_{(s,a)\sim\mu}[f(s,a)]$

$$f(s_0, \pi) - V^\pi(s_0) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{d^{\pi, s_0}}[f - \mathcal{T}^\pi f]$$

$$\Rightarrow V^\pi(s_0) - V^{\pi_f}(s_0) = V^\pi(s_0) - f(s_0, \pi) + f(s_0, \pi) + V^{\pi_f}(s_0)$$

$$\le V^\pi(s_0) \underbrace{- f(s_0, \pi) + f(s_0, \pi_f)}_{\text{advantage}} + V^{\pi_f}(s_0)$$

$$= \frac{1}{1-\gamma}\left(\mathbb{E}_{d^{\pi_f, s_0}}[\mathcal{T}^{\pi} f - f] + \mathbb{E}_{d^{\pi_f, s_0}}[f - \mathcal{T}^{\pi_f} f]\right)$$

$$\leq \frac{1}{1-\gamma}\left(\mathbb{E}_{d^{\pi, s_0}}[\mathcal{T} f - f] + \mathbb{E}_{d^{\pi_f, s_0}}[f - \mathcal{T} f]\right)$$

$\pi = \pi^*_M$

$\pi_f = \pi^*_{\hat{M}} \quad (f = Q^*_{\hat{M}})$

$$\left(\mathcal{T}^{\pi} f \leq \mathcal{T} f. \quad \underline{\mathcal{T}^{\pi_f} f = \mathcal{T} f}\right) \quad \square$$

$$\Rightarrow \underline{\| V^*_M - V^{\pi^*_{\hat{M}}}_M \|_\infty} \leq \frac{2}{1-\gamma} \underline{\| \mathcal{T}_M Q^*_{\hat{M}} - Q^*_{\hat{M}} \|_\infty}$$

drop an effective horizon !

<u>Goal</u>: bound $\| \mathcal{T}_M Q^*_{\hat{M}} - Q^*_{\hat{M}} \|_\infty$

$$\forall (s,a) \quad \left| (\mathcal{T}_M Q^*_{\hat{M}} - Q^*_{\hat{M}})(s,a) \right|$$

$$= \left| (\mathcal{T}_M Q^*_{\hat{M}})(s,a) - (\mathcal{T}_{\hat{M}} Q^*_{\hat{M}})(s,a) \right|$$

$$\leq \boxed{\left| (\mathcal{T}_M Q^*_M)(s,a) - (\mathcal{T}_{\hat{M}} Q^*_M)(s,a) \right|} \quad ⓪$$

$$① + \left| (\mathcal{T}_M Q^*_{\hat{M}})(s,a) - (\mathcal{T}_M Q^*_M)(s,a) \right|$$

$$② + \left| (\mathcal{T}_{\hat{M}} Q^*_M)(s,a) - (\mathcal{T}_{\hat{M}} Q^*_{\hat{M}})(s,a) \right|$$

$$V_{max} \sqrt{\frac{1}{2n} \cdot \log \frac{2}{\delta}}$$

already bounded via Hoeffding's ineq. (Q^*_M deterministic)

same Bellman operator $\mathcal{T}_M$ (1), $\mathcal{T}_{\hat{M}}$ (2)

$$\Rightarrow \left| (\mathcal{T}_M Q^*_{\hat{M}})(s,a) - (\mathcal{T}_M Q^*_M)(s,a) + (\mathcal{T}_{\hat{M}} Q^*_M)(s,a) - (\mathcal{T}_{\hat{M}} Q^*_{\hat{M}})(s,a) \right|$$

$$= \Big| \cancel{R(s,a)} + \gamma \cdot \langle P(s,a), V^*_{\hat{M}} \rangle - \cancel{R(s,a)} - \gamma \cdot \langle P(s,a), V^*_M \rangle$$
$$+ \cancel{\hat{R}(s,a)} + \gamma \cdot \langle \hat{P}(s,a), V^*_M \rangle - \cancel{\hat{R}(s,a)} - \gamma \cdot \langle \hat{P}(s,a), V^*_{\hat{M}} \rangle \Big|$$

$$= \gamma \cdot \left| \langle P(s,a) - \hat{P}(s,a), V^*_{\hat{M}} - V^*_M \rangle \right|$$

inevitable

$$\leq \gamma \cdot \underbrace{\| P(s,a) - \hat{P}(s,a) \|_1}_{O\left(\sqrt{\frac{|S|}{n}}\right)} \cdot \underbrace{\| V^*_{\hat{M}} - V^*_M \|_\infty}_{\frac{1}{1-\gamma} \cdot \sqrt{\frac{1}{2n} \log \frac{2}{\delta}} \cdot V_{max}}$$

$$\downarrow$$

$$O\left(\frac{1}{1-\gamma} \cdot \frac{1}{n} \cdot \sqrt{|S|} \right) \quad \star$$

fast

$$\Rightarrow \| \mathcal{T}_M Q^*_{\hat{M}} - Q^*_{\hat{M}} \|_\infty \leq O\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{|S|}}{1-\gamma} \cdot \frac{1}{n} \right) \cdot V_{max}$$

In large-sample regime: $n \gtrsim \frac{\sqrt{|S|}}{1-\gamma}$.

$$\Rightarrow \| \mathcal{T}_M Q^*_{\hat{M}} - Q^*_{\hat{M}} \|_\infty \leq O\left(\frac{1}{\sqrt{n}} \right) \cdot V_{max}$$

dominating term

$$\Rightarrow \| V^*_M - V^{\pi^*_{\hat{M}}}_M \|_\infty \leq O\left(\frac{1}{\sqrt{n}} \cdot \frac{V_{max}}{1-\gamma} \right)$$