# March 18) Batch RL II.

## 1. Better error control for model misspec.

$M = (P, r)$ $\hat{M} = (\hat{P}, \hat{r})$

$\hat{\pi}$ ← $\hat{M}$ ← $M$

$V^{\hat{\pi}} \geq V^* - \sigma \mathbb{1}$

$\sigma \leq \frac{1}{(1-\gamma)^3} \left( \underbrace{\|P - \hat{P}\|}_{\sqrt{\frac{1}{n}}} + \|r - \hat{r}\| \right)$ $\forall s,a$

$H = \frac{1}{1-\gamma}$

$\Rightarrow n \simeq \frac{H^6 SA}{\varepsilon^2} \log(1/\delta)$ $\leq \varepsilon$

$\rightarrow \frac{H^3 SA}{\varepsilon^2} \log(1/\delta)$

(s,a) - sampling

## 2. Lower bound

## 3. Data collection by following policies

↳ Lower bound $\Omega(A^{H^*})$ trajectories to get $(\varepsilon, \frac{1}{4})$

$\forall \pi_b$

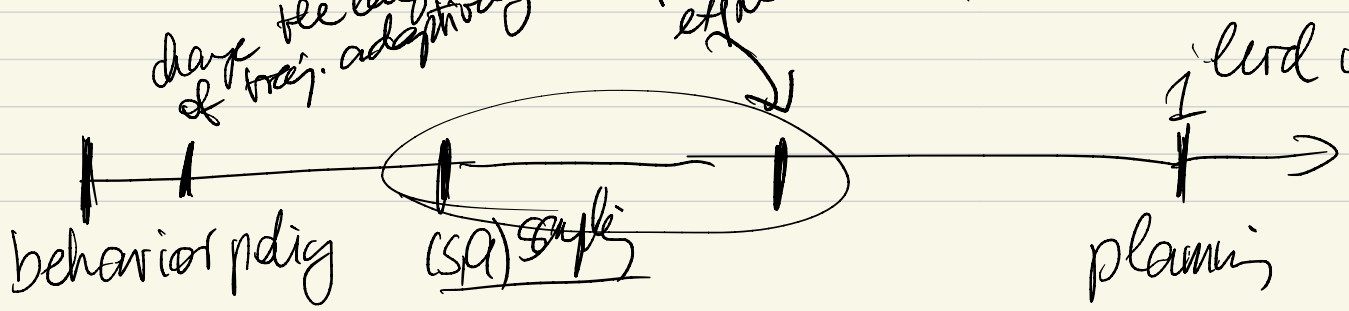
$S = H$

(s,a)-sampling:

$\rightarrow \frac{H^3 HA}{\varepsilon^2} \log(1/\delta)$

$H = H_{\gamma,\varepsilon}$

$\rightarrow H \cdot A^H$



1. $M = (P, r)$

$\hat{M} = (\hat{P}, \hat{r})$ : $\hat{\pi}$ $\qquad V^{\hat{\pi}} \approx V^*$

$$V^* - V^{\hat{\pi}} = V^* - \hat{V}^* + \hat{V}^* - \hat{V}^{\hat{\pi}} + \hat{V}^{\hat{\pi}} - V^{\hat{\pi}}$$

$$\leq V^* - \hat{V}^{\pi^*} + \hat{V}^* - \hat{V}^{\hat{\pi}} + \hat{V}^{\hat{\pi}} - V^{\hat{\pi}}$$

$\pi^*$ opt. in $M$ → $\qquad$ opt. error

$\pi \in \{\pi^*, \hat{\pi}\}$ $\qquad$ For simplicity: $r = \hat{r}$

$$V^{\pi} - \hat{V}^{\pi} = (I - \gamma P_\pi)^{-1} r_\pi - (I - \gamma \hat{P}_\pi)^{-1} r_\pi$$

$$= (I - \gamma P_\pi)^{-1} \left[ (I - \gamma \hat{P}_\pi) - (I - \gamma P_\pi) \right] (I - \gamma \hat{P}_\pi)^{-1} r_\pi$$

$$= \gamma (I - \gamma P_\pi)^{-1} (P_\pi - \hat{P}_\pi)(I - \gamma \hat{P}_\pi)^{-1} r_\pi$$

$$= \gamma (I - \gamma P_\pi)^{-1} M_\pi (P - \hat{P}) \hat{V}^{\pi}$$

$$\frac{1}{1-x} - \frac{1}{1-y} = \frac{1-y-(1-x)}{(1-x)(1-y)} = \frac{x-y}{(1-x)(1-y)}$$

$$\| V^{\pi} - \hat{V}^{\pi} \|_\infty \leq \frac{\gamma}{1-\gamma} \| (P - \hat{P}) \hat{V}^{\pi} \|_\infty$$

wp $1-\delta$:

$$\| (P - \hat{P}) \hat{V}^{\pi} \|_\infty \sim \sqrt{\frac{\log(SA/\delta)}{n}} \frac{1}{1-\gamma}$$

Hoeffding. $\qquad \frac{1}{1-\gamma}$ $\qquad$ Loose! $\qquad \frac{1}{1-\gamma}$ $\qquad$ cheat!

$$V^* - V^{\hat{\pi}} \le \underbrace{\frac{2\gamma}{(1-\gamma)^2}\sqrt{\frac{\log(\cdots)}{n}}}_{\le \varepsilon} + \varepsilon_{opt}$$

$\varepsilon_{opt} \approx 0$

$$n \ge \frac{H^4}{\varepsilon^2}\log(\cdots)$$

$H^6 \to H^4$

$\boxed{H^4 \to H^3 \ ?}$

---

Hoeffding : worst-case tight!

Bernstein's $\le$

$X_1, \dots, X_n \in [0, b]$ , $b > 0$

i.i.d.

$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$ , $\forall \zeta$ w.p. $1-\zeta$

$$|\bar{X}_n - \mathbb{E}[X_1]| \le \underbrace{\sigma\sqrt{\frac{2\log(2/\zeta)}{n}} + \frac{2b\log(2/\zeta)}{3n}}_{\le \varepsilon}$$

$\sigma^2 = \mathrm{Var}(X_1)$

$n \ge \frac{b^2}{\varepsilon^2}(\dots)$

$b\sqrt{\frac{\log(2/\zeta)}{2n}} \le \varepsilon$

Bernstein: $n \geq \frac{\sigma^2}{\varepsilon^2} \log(\dots)$ ← $\frac{H^3}{\varepsilon^2}$

$n \geq \frac{b}{\varepsilon}$ $\quad \frac{\sigma^2}{\varepsilon^2} \geq \frac{b}{\varepsilon}$

$\sigma^2 \leq b^2$ $\quad 1 = \frac{1/1-\gamma}{1/1-\gamma} =$ $\quad \boxed{\frac{\sigma^2}{b} \geq \varepsilon}$

---

$(P - \hat{P}) v^\pi$ $\quad s_1', \dots, s_n' \sim P_a(s)$

$(P_a(s) - \hat{P}_a(s)) v^\pi = P_a(s) v^\pi - \frac{1}{n} \sum_{i=1}^{n} v^\pi(s_i')$

$\hat{P}_a(s, s') = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(s_i' = s')$

$X_i = v^\pi(s_i')$ (!? ?!) $\quad \mathbb{E}[X_i] = P_a(s) v^\pi$

$\mathrm{Var}(X_i) = \sigma_\pi^2(s,a) \sim \frac{1}{(1-\gamma)^2}$ ✱

$\| (I - \gamma P_\pi)^{-1} M_\pi \sigma_\pi \| \leq \sqrt{\frac{2}{(1-\gamma)^3}} = \frac{1}{1-\gamma} \frac{1}{\sqrt{1-\gamma}}$

discounted stddev of value

$\frac{1}{1-\gamma}$

$\sigma_\pi^2 \sim \frac{1}{1-\gamma}$ ✱

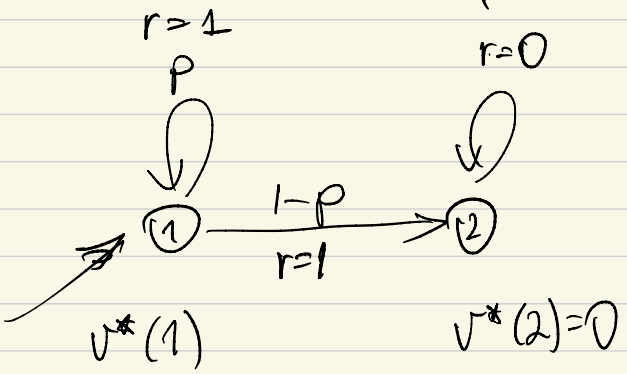
(2.) Can we improve upon $H^3 SA \log\left(\frac{SA}{\varepsilon}\right)$ / $\varepsilon^2$ ($\varepsilon$-subopt.) ?

Sketch: $X_1, \dots, X_n \sim \mathrm{Ber}(p)$ iid.

$$\Rightarrow |\bar{X}_n - p| \approx \sqrt{\frac{p(1-p)\log(2/\delta)}{2n}} + \dots$$

$\mathrm{Var}(X_1) = p(1-p)$

$$\sqrt{\frac{p(1-p)\log(2/\delta)}{2n}} \le \varepsilon \iff n \ge \frac{p(1-p)\log(2/\delta)}{\varepsilon^2}$$




$$v^*(1) = p(1 + \gamma v^*(1)) + (1-p)(1 + \gamma \underbrace{v^*(2)}_{=0})$$

$$v_p^*(1) = 1 + p\gamma v^*(1)$$

$$v_p^*(1) = \frac{1}{1-p\gamma}$$

$$\frac{d}{dp} v_p^*(1) = \frac{d}{dp} \frac{1}{1-p\gamma} = \frac{-(-\gamma)}{(1-\gamma p)^2} = \frac{\gamma}{(1-\gamma p)^2}$$

$$\varepsilon \geq |v^*_p(1) - v^*_{\bar{X}_n}(1)| \approx \frac{d}{dp} v^*_p(1) \, \underbrace{|p - \bar{X}_n|}$$

$$= \frac{\gamma}{(1-\gamma p)^2} \, \frac{\sqrt{p(1-p)\log(\dots)}}{2n}$$

$$\underbrace{(1-\gamma p)^2}_{\approx} \qquad 1-\gamma$$

$\gamma \to 1$

$$1 - \gamma p = 1 - \gamma\left(1 + \frac{\gamma-1}{2}\right) \qquad p = \frac{1+\gamma}{2} = \frac{1}{2} + \frac{\gamma}{2}$$

$$= 1 - \gamma + \frac{\gamma(\gamma-1)}{2} \qquad = 1 + \frac{\gamma-1}{2}$$


$$p(\gamma) = \frac{4\gamma - 1}{3\gamma} = \frac{4}{3} - \frac{1}{3\gamma}$$

$$= 1 + \frac{1}{3}\left(1 - \frac{1}{\gamma}\right) = 1 + \frac{1}{3}\frac{1-\gamma}{\gamma}$$

$$1 - \gamma p = 1 - \gamma\left(1 + \frac{1}{3}\frac{1-\gamma}{\gamma}\right)$$

$$= 1 - \gamma - \frac{1}{3}(1-\gamma) = (1-\gamma)\cdot\frac{2}{3}$$

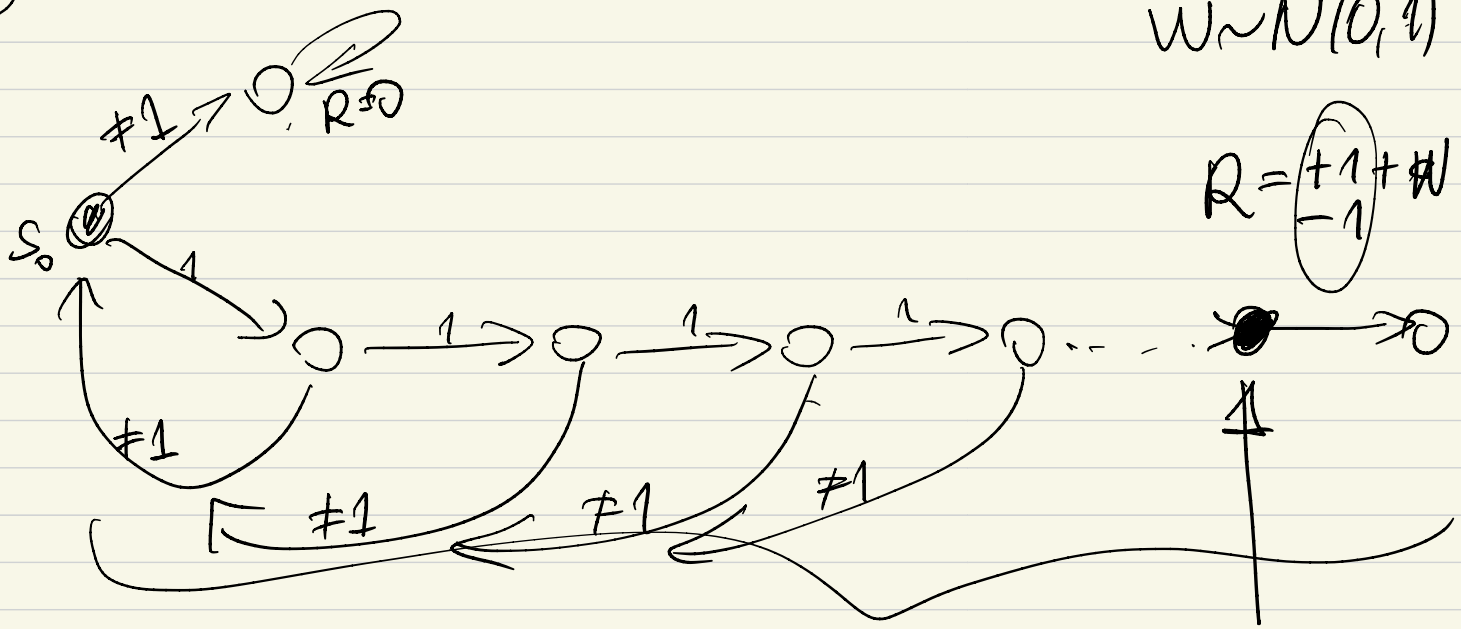
$$(1 - p(\gamma))\, p(\gamma) = \underbrace{\frac{1}{3}\frac{1-\gamma}{\gamma}}_{\approx} \cdot \underbrace{\frac{4\gamma-1}{3\gamma}}_{\approx 1} \approx \frac{1}{3}(1-\gamma) \qquad \gamma \to 1$$

$$\frac{\gamma}{(1-\gamma p)^2}\sqrt{\frac{p(1-p)\log(\cdots)}{2n}} \approx \frac{1}{(1-\gamma)^2\left(\frac{2}{3}\right)^2}\sqrt{\frac{\frac{1}{2}(1-\gamma)\log(\cdots)}{2n}}$$

$$\leq \varepsilon$$

$$\boxed{n \geq H^3/\varepsilon^2 \log(\cdots)} \qquad H = \frac{1}{1-\gamma}$$

(3.)




$H = H_{\gamma,\varepsilon}$ $\quad p_0 \geq \left(\frac{1}{A}\right)^H$

$\pi_b(1|s)$

$\leq \pi_b(a|s) \quad \forall a$

$$\boxed{\pi_b(1|s) \leq \frac{1}{A}}$$

$$\boxed{n \geq p_0^{-1}} \qquad \boxed{= A^H}$$