July, 13, '23

Recap $\quad M = (S, A, P, r, \gamma)$

Generative model (simulator): $\quad \forall (s,a) \in S \times A$, draw N samples $s' \sim P(\cdot \mid s,a)$

model-based estimator ("plug-in"): $\quad \hat{P}(s' \mid s,a) = \dfrac{\text{count}(s,a,s')}{N}$

$$\hat{M} = (S, A, \hat{P}, r, \gamma) \xrightarrow{\text{planning}} \pi^*_{\hat{M}} =: \hat{\pi}$$

Two metrics:

- (Value estimate) $\quad \| Q^* - \hat{Q}^* \|_\infty$
- (policy estimate) $\quad \| Q^* - Q^{\hat{\pi}} \|_\infty$

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

Result recaps & plan:

- lower bound (value & policy): $\#\text{samples} = N.|S|.|A| = \tilde{\Omega}\left(\frac{H^3 |S|.|A|}{\varepsilon^2}\right)$

  (Azar et al. '18)

- **upper bounds:**

  . (coarse analysis wf uniform convergence): $\#\text{samples}_{\text{policy/value}} = \tilde{O}\left(\frac{H^4 |S|^2 . |A|}{\varepsilon^2}\right)$

  . **Today:**

  (*) $\#\text{samples}_{\text{value}} = \tilde{O}\left(H^4 . |S| . |A| / \varepsilon^2\right)$

  (**) $\#\text{samples}_{\text{value}} = \tilde{O}\left(H^3 . \frac{|S|.|A|}{\varepsilon^2}\right)$ (Azar et al. '13)

  $\#\text{samples}_{\text{policy}} = \tilde{O}\left(H^3 \frac{|S|.|A|}{\varepsilon^2}\right)$ for $\varepsilon \in \left(0, \sqrt{\frac{H}{\gamma |S|}}\right)$ (Azar)

  (***) $\#\text{sample}_{\text{policy}} = \tilde{O}\left(H^3 \frac{|S|.|A|}{\varepsilon^2}\right)$ $\forall \varepsilon \in (0,1)$ (Agarwal '20)

(*)
Theorem (value estimation)

$$\text{wp } 1-\delta \quad \|Q^* - \hat{Q}^*\|_\infty \leq \gamma H^2 \sqrt{\frac{2\log(|S|\cdot|A|/\delta)}{N}}$$

Claim:

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

Proof:

$$\|Q^* - \hat{Q}^*\|_\infty = \|\gamma P^{\pi^*} Q^* - \gamma \hat{P}^{\hat{\pi}} \hat{Q}^*\|_\infty$$

$$= \gamma \| P^{\pi^*} Q^* - \hat{P}^{\pi^*} Q^* + \hat{P}^{\pi^*} Q^* - \hat{P}^{\hat{\pi}} \hat{Q}^* \|_\infty$$

$$\leq \gamma \| P^{\pi^*} Q^* - \hat{P}^{\pi^*} Q^* \|_\infty + \gamma \| \hat{P}^{\pi^*} Q^* - \hat{P}^{\hat{\pi}} \hat{Q}^* \|_\infty$$

$$= \gamma \| P V^* - \hat{P} V^* \|_\infty + \gamma \| \hat{P} V^* - \hat{P} \hat{V}^* \|_\infty$$

$$\leq \gamma \| (P-\hat{P}) V^* \|_\infty + \gamma \underbrace{\| V^* - \hat{V}^* \|_\infty}_{\leq \|Q^* - \hat{Q}^*\|_\infty}$$

$$\|(P-\hat{P})V^{*}\|_\infty = \max_{(s,a)} \left| \left(P(\cdot|s,a) - \hat{P}(\cdot|s,a)\right)^T V^{*} \right|$$

$$\lesssim \frac{1}{1-\gamma}\sqrt{\frac{\log(|S|\cdot|A|/\delta)}{N}} \quad \text{(Hoeffding's)}$$

Theorem : (**)

- (Value bound) $\| Q^* - \hat{Q}^* \|_\infty \leq \gamma \sqrt{H^3 \frac{\log(|S| \cdot |A| / \delta)}{N}} + \frac{\gamma}{(1-\gamma)^3} \frac{\log(|S| \cdot |A| / \delta)}{N}$

Lemma (point-wise bounds)

$$Q^* - \hat{Q}^* \leq \gamma (I - \gamma \hat{P}^{\pi^*})^{-1} (P - \hat{P}) V^*$$
$$Q^* - \hat{Q}^* \geq \gamma (I - \gamma \hat{P}^{\hat{\pi}})^{-1} (P - \hat{P}) V^*$$

<u>Proof:</u>

- $Q^* - \hat{Q}^* \leq Q^* - \hat{Q}^{\pi^*} = \gamma (I - \gamma \hat{P}^{\pi^*})^{-1} (P - \hat{P}) V^*$

- Recall: $Q^* = r + \gamma P^{\pi^*} Q^*$, $\hat{Q}^* = r + \gamma \hat{P}^{\hat{\pi}} \hat{Q}^*$

Thus, $Q^* - \hat{Q}^* = Q^* - (I - \gamma \hat{P}^{\hat{\pi}})^{-1} r$

$$= Q^* - (I - \gamma \hat{P}^{\hat{\pi}})^{-1} (I - \gamma P^{\pi^*}) Q^*$$
$$= (I - \gamma \hat{P}^{\hat{\pi}})^{-1} \left[(I - \gamma \hat{P}^{\hat{\pi}}) - (I - \gamma P^{\pi^*}) \right] Q^*$$
$$= \gamma (I - \gamma \hat{P}^{\hat{\pi}})^{-1} (P^{\pi^*} - \hat{P}^{\hat{\pi}}) Q^*$$

Note: $P^{\pi^*} Q^* = P V^*$; $\hat{P}^{\hat{\pi}} Q^* \leq \hat{P}^{\pi^*} Q^* = \hat{P} V^*$

(point_wise bounds)

Claim

1 $$\boxed{\begin{aligned} Q^* - \hat{Q}^* &\leq \gamma (I - \gamma \hat{P}^{\pi^*})^{-1} (P - \hat{P}) V^* \\ Q^* - \hat{Q}^* &\geq \gamma (I - \gamma \hat{P}^{\hat{\pi}})^{-1} (P - \hat{P}) V^* \end{aligned}}$$

(todo: show this is time permitted)

Bernstein's inequality:

2 $$\left| (P - \hat{P}) V^* \right| \leq \sqrt{\mathrm{Var}_P(V^*) \frac{\log(|S| \cdot |A| / \delta)}{N}} + H \frac{\log(|S| \cdot |A| / \delta)}{N}$$

here $$[\mathrm{Var}_P(V^*)](s,a) = \mathrm{Var}_{s' \sim P(\cdot|s,a)} [V^*(s')]$$

Plug in 2 into 1:

- $$Q^* - \hat{Q}^* \leq \gamma (I - \gamma \hat{P}^{\pi^*})^{-1} \sqrt{\mathrm{Var}_P(V^*)} \cdot \tilde{O}\left(\frac{1}{\sqrt{N}}\right) + \tilde{O}\left(\left(\frac{1}{1-\gamma}\right)^2 \frac{1}{N}\right)$$
- $$Q^* - \hat{Q}^* \geq -\gamma (I - \gamma \hat{P}^{\hat{\pi}})^{-1} \sqrt{\mathrm{Var}_P(V^*)} \cdot \tilde{O}\left(\frac{1}{\sqrt{N}}\right) +$$

It remains to upper bound:

$$\left\| (I - \gamma \hat{P}^{\pi})^{-1} \sqrt{\mathrm{Var}_P(V^*)} \right\|_\infty$$

where $\pi \in \{\pi^*, \hat{\pi}\}$

- A trivial upper bound: $\left(\frac{1}{1-\gamma}\right)^2$

- A more intricate analysis gives: $\left(\frac{1}{1-\gamma}\right)^{3/2} + \tilde{O}\left(\frac{1}{1-\gamma} \cdot \frac{1}{N^{1/4}} + \left(\frac{1}{1-\gamma}\right)^2 \frac{1}{\sqrt{N}}\right)$

③ <u>Claim</u> $\boxed{\left\| (I - \gamma P^\pi)^{-1} \sqrt{Var_P(V^\pi)} \right\|_\infty \leq \sqrt{H^3}}$

- <u>Bellman equation for variance</u>

$$\boxed{\Sigma_M^\pi = \gamma^2 (I - \gamma^2 P^\pi)^{-1} Var_P(V_M^\pi)}$$

where: $\Sigma_N^\pi = \mathbb{E}\left[\left(\sum_{t=0}^{\infty} \gamma^t r_t - V_M^\pi(s,a) \right)^2 \middle| (s_0,a_0) = (s,a) \right]$

- $\| \Sigma_M^\pi \|_\infty \leq H^2$
- $\| (I - \gamma P^\pi)^{-1} \sqrt{v} \|_\infty = \frac{1}{1-\gamma} \| (1-\gamma)(1-\gamma P^\pi)^{-1} \sqrt{v} \|_\infty$

$$\leq \frac{1}{1-\gamma} \left\| \sqrt{(1-\gamma)(1-\gamma P^\pi)^{-1} v} \right\|_\infty$$

$$\leq \frac{1}{1-\gamma} \left\| \sqrt{2(1-\gamma)(1-\gamma^2 P^\pi)^{-1} v} \right\|_\infty$$

$$= \frac{2}{\sqrt{1-\gamma}} \left\| \sqrt{(1-\gamma^2 P^\pi)^{-1} v} \right\|_\infty$$

$\because \| (1-\gamma P^\pi)^{-1} v \|_\infty \leq \| 2(1-\gamma^2 P^\pi)^{-1} v \|_\infty$

$$\begin{cases} \mathrm{Var}_P(V^*) \longrightarrow \mathrm{Var}_{\hat{P}}(\hat{V}^*) \ ? \\ \mathrm{Var}_P(V^*) \longrightarrow \mathrm{Var}_{\hat{P}}(\hat{V}^{\hat{\pi}}) \ ? \end{cases}$$

Crude bounds are fine!

(e.g. Hoeffding's ineq)

- $$\mathrm{Var}_P(V^*) = \mathrm{Var}_P(V^*) - \mathrm{Var}_{\hat{P}}(V^*) + \mathrm{Var}_{\hat{P}}(V^*)$$

$$= \left[P(V^*)^2 - (PV^*)^2\right] - \left[\hat{P}(V^*)^2 - (\hat{P}V^*)^2\right] + \mathrm{Var}_{\hat{P}}(V^*)$$

$$= \underbrace{(P-\hat{P})(V^*)^2}_{} + \underbrace{(\hat{P}-P)V^*\left[(\hat{P}+P)V^*\right]}_{} + \mathrm{Var}_{\hat{P}}(V^*)$$

$$= \tilde{O}\left(\left(\frac{1}{1-\gamma}\right)^2 \frac{1}{\sqrt{N}}\right) + \tilde{O}\left(\left(\frac{1}{1-\gamma}\right)^2 \frac{1}{\sqrt{N}}\right) + \mathrm{Var}_{\hat{P}}(V^*)$$

- $$\mathrm{Var}_{\hat{P}}(V^*) = \mathrm{Var}_{\hat{P}}(\hat{V}^* + V^* - \hat{V}^*)$$

$$\leq \left(\sqrt{\mathrm{Var}_{\hat{P}}(\hat{V}^*)} + \sqrt{\mathrm{Var}_{\hat{P}}(V^* - \hat{V}^*)}\right)^2$$

$$\leq 2\left(\mathrm{Var}_{\hat{P}}(\hat{V}^*) + \underbrace{\mathrm{Var}_{\hat{P}}(V^* - \hat{V}^*)}_{\leq \|V^* - \hat{V}^*\|_\infty^2 = \tilde{O}\left(\left(\frac{1}{1-\gamma}\right)^4 \frac{1}{N}\right)}\right)$$

(new)

Theorem (policy bound) (***)

- (policy bound) $$\|Q^* - Q^{\hat{\pi}}\|_\infty \leq \gamma \sqrt{H^3 \frac{\log(|S|.|A|/\delta)}{N}}, \text{ if } N \geq H^2$$

more subtle

Ref: Agarwal & Kakade & Yang : "Model-based RL w/ a generative model is minimax optimal;" 2020

Error decomposition:

$$- \underbrace{Q^* - Q^{\hat\pi}}_{\geq 0} = Q^* - \hat{Q}^{\pi^*} + \hat{Q}^{\pi^*} - Q^{\hat\pi}$$

$$\leq \underbrace{Q^* - \hat{Q}^{\pi^*} + \hat{Q}^{\hat\pi} - Q^{\hat\pi}}$$

$$\Rightarrow \| Q^* - Q^{\hat\pi} \|_\infty \leq \| Q^* - \hat{Q}^{\pi^*} \|_\infty + \| \hat{Q}^{\hat\pi} - Q^{\hat\pi} \|_\infty$$

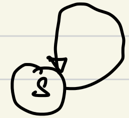
Simulation Lemma:

$$Q^* - \hat{Q}^{\pi^*} = \gamma (I - \gamma P^{\pi^*})(P - \hat{P}) \hat{V}^{\pi^*}$$

$$\hat{Q}^{*} - Q^{\hat\pi} = \gamma (I - \gamma P^{\hat\pi})(P - \hat{P}) \hat{V}^{\hat\pi}$$

<u>idea:</u> s-absorbing MDP:

- For any $s \in S$, $u \in \mathbb{R}$, $M_{s,u}$ is identical to M except that:



$$\begin{cases} P_{M_{s,u}}(s \mid s,a) = 1 & \forall a \\ r_{M_{s,u}}(s,a) = u & \forall a \end{cases}$$

- $$(P_{s,a} - \hat{P}_{s,a})\hat{V}^* = (P_{s,a} - \hat{P}_{s,a}) V^*_{\hat{M}_{s,u}} + \underbrace{(P_{s,a} - \hat{P}_{s,a})\left(\hat{V}^* - V^*_{\hat{M}_{s,u}}\right)}_{\leq \|\hat{V}^* - V^*_{\hat{M}_{s,u}}\|_\infty =: \Delta}$$

Note: $\hat{P}_{s,a} \perp V^*_{\hat{M}_{s,u}}$

<u>Bernstein's:</u> Fix (s,a), $\forall u \in \mathcal{U}$:

$$(P_{s,a} - \hat{P}_{s,a})\hat{V}^* \leq \sqrt{\frac{\mathrm{Var}_{P_{s,a}}\left(V^*_{\hat{M}_{s,u}}\right)\log(|\mathcal{U}|)}{N}} + \frac{\log(|\mathcal{U}|)}{(1-\gamma)N} + \Delta$$

$$\leq \sqrt{\frac{\mathrm{Var}_{P_{s,a}}\left(\hat{V}^* + V^*_{\hat{M}_{s,a}} - \hat{V}^*\right) \cdot \log U}{N}} + \frac{\log U}{(1-\gamma)N} + \Delta$$

$$= \sqrt{\frac{\mathrm{Var}_{P_{s,a}}\left(\hat{V}^*\right)}{N} \log(|U|)} + \inf_{u \in U} \left\| \hat{V}^* - V^*_{\hat{M}_{s,u}} \right\| \left(1 + \cancel{\frac{\log U}{\sqrt{N}}}\right)$$

$$+ \cancel{\frac{\log U}{(1-\gamma) N}}$$

<u>Controlling:</u> $\| V^*_{\hat{M}_{s,u}} - V^*_{\hat{M}} \|_\infty$ (Goal: $\tilde{O}\left(\frac{1}{\sqrt{N}}\right)$)

$$= |u - \hat{V}^*(s)|$$

choose $\mathcal{U}$: even interval of $\left[V^*(s) \pm \tilde{O}\left(\left(\frac{1}{1-\gamma}\right)^2 \frac{1}{\sqrt{N}} \right) \right]$

$$\min_{u \in \mathcal{U}} |u - \hat{V}^*(s)| \leq \frac{\tilde{O}\left(\frac{1}{(1-\gamma)^2}\frac{1}{\sqrt{N}}\right)}{|\mathcal{U}| - 1} = \tilde{O}\left(\frac{1}{\sqrt{N}}\right)$$

if choose $|\mathcal{U}| - 1 = \frac{1}{(1-\gamma)^2}$