Model based RL – Tabular

online setting

generative setting (last time)

(today)

Aarti Singh

Machine Learning 10-734
Oct 28, 2025

Slides courtesy: Yuejie Chi, Wen Sun

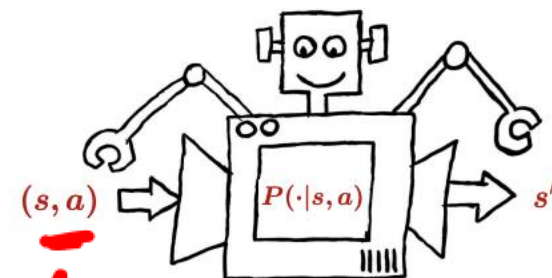
Data source in RL

offline RL

no control s, a, s', r

online RL

s_0 - control
a - control

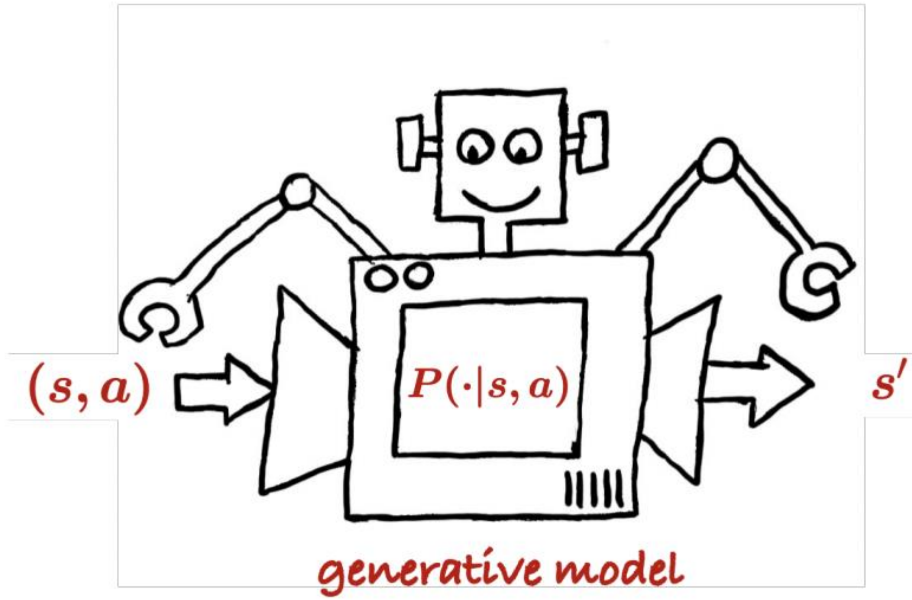


control

generative model

The capability of exploration increases from left to right.

RL with generative data



S, A finite (Tabular)

size of state space

size of action space

fixed samples per (s, a)

SAN total samples

For each (s, a), collect N independent samples $\{(s, a, s'_{(i)})\}_{1 \le i \le N}$

Empirical estimates: estimate $\widehat{P}(s'|s, a)$ by $\underbrace{\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{s'_{(i)} = s'\}}_{\text{empirical frequency}}$

Compute $\hat{\pi}$ given $(\hat{P}, r)$ using Value iteration or Policy iteration.

Simulation Lemma

Given policy π, does $P \approx \widehat{P}$ imply $V^\pi \approx V^\pi_{\widehat{P}}$?

Proposition

- Given any two transitions P and $\widehat{P}$, and any policy π, we have:

Infinite horizon setting

$$\forall s_0 : V^\pi_P(s_0) - V^\pi_{\widehat{P}}(s_0)$$

$$\leq \frac{\gamma}{1-\gamma}\mathbb{E}_{s,a\sim d^\pi_{s_0}} \left| \mathbb{E}_{s'\sim P(s,a)} \widehat{V}^\pi(s') - \mathbb{E}_{s'\sim \widehat{P}(s,a)} \widehat{V}^\pi(s') \right|$$

Where $V^\pi_{\hat{P}} = \hat{V}^\pi$ for simplicity.

$(P-\hat{P})\cdot V^*$

$$\leq \frac{\gamma}{(1-\gamma)^2}\mathbb{E}_{s,a\sim d^\pi_{s_0}} \left\| \widehat{P}(\cdot|s,a) - P(\cdot|s,a) \right\|_1$$

Model accuracy

Sample complexity of RL using generative model

- With probability greater than 1-δ

$$V^* - V^{\hat{\pi}^*} = O\left(\frac{\gamma}{(1-\gamma)^2}\sqrt{\frac{S\ln(SA/\delta)}{N}}\right)$$

N - # sample per (s,a)

$\sim \varepsilon$

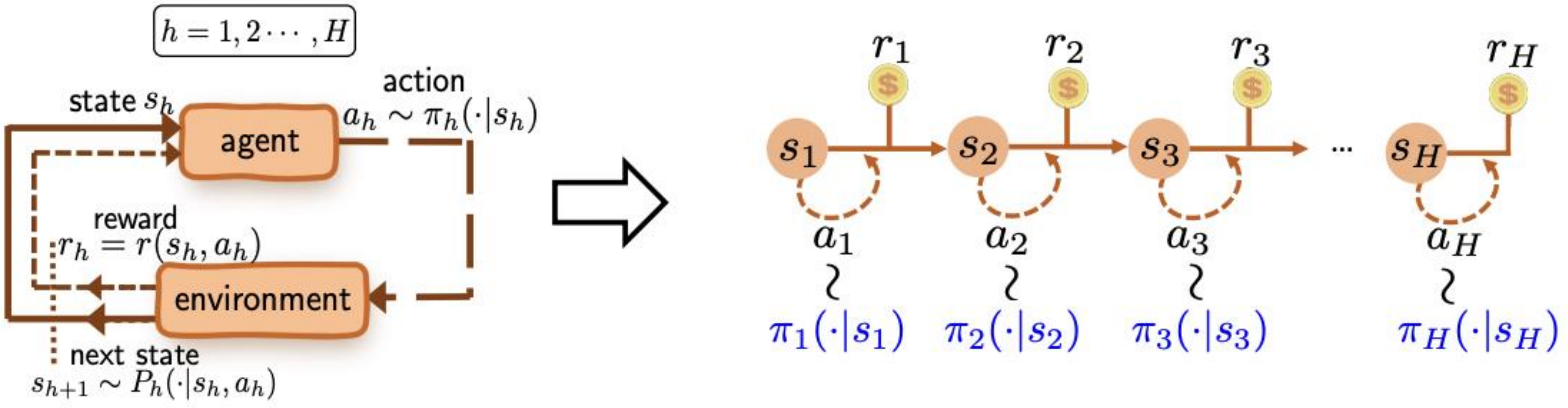
- Need $N \sim \frac{S}{\epsilon^2(1-\gamma)^4}\ln\left(\frac{SA}{\delta}\right)$ to get ϵ-accurate in policy value w.p. 1-δ

- Total samples *SAN* ~ $\frac{S^2A}{\epsilon^2(1-\gamma)^4}\ln\left(\frac{SA}{\delta}\right)$ matches parameter count argument

- Can improve scaling to SA (drop S term) if we only care about model error for high value state-action pairs - analyze model error projected on V*

RL with online data

- Tabular setting (finite S, A)
- Finite horizon $H \sim \frac{1}{1-\gamma}$
- Non-stationary $\mathscr{M} = \{\{r_h\}_{h=0}^{H-1}, \{P_h\}_{h=0}^{H}, H, \mu, S, A\}$
- Only reset to initial state $s_0 \sim \mu$
- For simplicity, μ is point mass at s_0



RL with online data

1. Learner initializes a policy π^1 $\qquad$ $\pi = \{\pi_0, \ldots, \pi_{H-1}\}$

2. At episode n, learner executes π^n and obtains trajectory

$$\{s_h^n, a_h^n, r_h^n\}_{h=0}^{H-1}$$

Can't guarantee fixed N samples from each state, action pair

with $a_h^n = \pi^n(s_h^n), r_h^n = r(s_h^n, a_h^n), s_{h+1}^n \sim P(\cdot \mid s_h^n, a_h^n)$

3. Learner updates policy to π^{n+1} using all prior information

Performance measure: REGRET

$$\mathbb{E}\left[\sum_{n=1}^{N}\left(V^\star - V^{\pi^n}\right)\right] = \text{poly}(S, A, H)\sqrt{N}$$

gen $\frac{SA}{\sqrt{N}}$

RL with online data

- Need exploration (unlike generative model setting) to encourage visiting unexplored state-action pairs starting from s_0, while exploiting promising state-action pairs

"policy?" $\sqrt{KT}$ K - # policies ~ exp(S, H)

Attempt 1: Treat MDP as a Multi-armed bandit problem and run UCB

Doesn't work. Shouldn't treat policies as independent arms — they do share information

Attempt 2: The Upper Confidence Bound Value Iteration Algorithm (UCB-VI)

$P, r \rightarrow \hat{P}, r + b$ depends on # times visited

Attempt 2: UCB-VI

- Upper Confidence Bound Value Iteration (UCB-VI)

Optimistic Model-based Learning

At each iteration n

Use all previous data to estimate transitions $\widehat{P}^n_1, \ldots, \widehat{P}^n_{H-1}$

Design reward bonus $b^n_h(s, a), \forall s, a, h$

Optimistic planning with learned model: $\pi^n = \text{Value-Iter}\left(\{ \widehat{P}^n_h, r_h + b^n_h \}^{H-1}_{h=1} \right)$

Collect a new trajectory by executing π^n in the real world $\{P_h\}^{H-1}_{h=0}$ starting from s_0

UCB-VI: Model est. & reward bonus

Let us consider the **very beginning** of episode n:

$$\mathscr{D}_h^n = \{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}, \forall h$$

Estimate model $\widehat{P}_h^n(s' \mid s, a), \forall s, a, s', h$: $\quad \widehat{P}_h^n(s' \mid s, a) = \frac{N_h^n(s, a, s')}{N_h^n(s, a)}$

where

$$N_h^n(s, a, s') = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s, a, s')\}, \forall s, a, h$$

$$N_h^n(s, a) = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i) = (s, a)\}, \forall s, a, h$$

Not fixed N samples

Reward bonus

$$b_h^n(s, a) = cH\sqrt{\frac{\ln\left(SAHN/\delta\right)}{N_h^n(s, a)}}$$

Encourage to explore new state-actions

UCB-VI: Value iteration

Value iteration at episode n using $\{ \widehat{P}^n_h, r_h + b^n_h \}_{h=1}^{H-1}$

$$\widehat{V}^n_H(s) = 0, \forall s$$

For h = H-1, H-2, ..., 1

$$\widehat{Q}^n_h(s,a) = \min \left\{ r_h(s,a) + b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1}, \quad H \right\}, \forall s, a$$

$$\left\| \widehat{V}^n_h \right\|_\infty \leq H, \forall h, n$$

$$\widehat{V}^n_h(s) = \max_a \widehat{Q}^n_h(s,a)$$

$$\pi^n_h(s) = \arg\max_a \widehat{Q}^n_h(s,a), \forall s$$

UCB-VI

For $n = 1 \to N$:

1. Set $N_h^n(s, a) = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i) = (s, a)\}, \forall s, a, h$

2. Set $N_h^n(s, a, s') = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s, a, s')\}, \forall s, a, a', h$

3. Estimate $\widehat{P}^n$: $\widehat{P}_h^n(s' | s, a) = \frac{N_h^n(s, a, s')}{N_h^n(s, a)}, \forall s, a, s', h$

4. Plan: $\pi^n = VI\left(\{\widehat{P}_h^n, r_h + b_h^n\}_h\right)$, with $b_h^n(s, a) = cH\sqrt{\frac{\ln(SAHN/\delta)}{N_h^n(s, a)}}$

UCB

5. Execute π^n : $\{s_0^n, a_0^n, r_0^n, \ldots, s_{H-1}^n, a_{H-1}^n, r_{H-1}^n, s_H^n\}$

UCB-VI regret bound

Regret

$$\mathbb{E}\left[\sum_{n=1}^{N}\left(V^\star - V^{\pi^n}\right)\right] \leq \widetilde{O}\left(H^2\sqrt{S^2AN}\right)$$

P S^2A parameter

Dependency on H and S are suboptimal; but the **same** algorithm can achieve $H^2\sqrt{SAN}$ in the leading term
[Azar et.al 17 ICML, and AJKS book Ch 7]

Proof sketch

Bonus $b_h^n(s,a)$ is related to $\left(\left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V_{h+1}^\star \right)$

VI with bonus inside the learned model gives optimism, i.e.,

$$\widehat{V}_h^n(s) \geq V_h^\star(s), \forall h, n, s, a$$

$r \in [\hat{r} \pm UCB]$

Upper bound per-episode regret:

$$V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$$

Apply simulation lemma: $\widehat{V}_0^n(s_0) - V^{\pi^n}(s_0)$

Model error projected on V*

Given a fixed function $f : S \mapsto [0,H]$, w/ prob $1-\delta$:

$$\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right)^\top f \right| \leq O(H\sqrt{\ln(SAHN/\delta)/N_h^n(s,a)}), \forall s,a,h,N$$

Bonus $b_h^n(s,a)$

Intuition:

1. Assume for some i, $s_h^i = s, a_h^i = a,$

then $f(s_{h+1}^i)$ is an unbiased estimate of $\mathbb{E}_{s' \sim P_h(\cdot|s,a)} f(s')$

2. Note $\widehat{P}_h^n(\cdot \mid s,a) \cdot f = \frac{1}{N_h^n(s,a)} \sum_{i=1}^{n-1} \mathbf{1}[(s_h^i, a_h^i) = (s,a)] f(s_{h+1}^i)$

Optimism via induction

Lemma [Optimism]: $\widehat{V}_h^n(s) \geq V_h^\star(s), \forall n, h, s$

Recall Bonus-enhanced Value Iteration at episode n:

$$\widehat{V}_H^n(s) = 0, \quad \widehat{Q}_h^n(s,a) = \min\left\{ r_h(s,a) + b_h^n(s,a) + \widehat{P}_h^n(\cdot\,|\,s,a)\cdot \widehat{V}_{h+1}^n, H \right\}$$

$$\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \quad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a), \forall s$$

Inductive hypothesis: $\widehat{V}_{h+1}^n(s) \geq V_{h+1}^\star(s), \quad \forall s$

$$\begin{aligned}
\widehat{Q}_h^n(s,a) - Q_h^\star(s,a) &= r_h(s,a) + b_h^n(s,a) + \widehat{P}_h^n(\cdot\,|\,s,a)\cdot \widehat{V}_{h+1}^n - r_h(s,a) - P_h(\cdot\,|\,s,a)\cdot V_{h+1}^\star \\
&\geq b_h^n(s,a) + \widehat{P}_h^n(\cdot\,|\,s,a)\cdot V_{h+1}^\star - P_h(\cdot\,|\,s,a)\cdot V_{h+1}^\star \\
&= b_h^n(s,a) + \left(\widehat{P}_h^n(\cdot\,|\,s,a) - P_h(\cdot\,|\,s,a)\right)\cdot V_{h+1}^\star \\
&\geq b_h^n(s,a) - b_h^n(s,a) = 0, \quad \forall s, a
\end{aligned}$$

w.p. > 1-δ

Bounding regret using optimism

per-episode regret $:= V_0^{\star}(s_0) - V_0^{\pi_n}(s_0)$

$$\leq \widehat{V}_0^n(s_0) - V_0^{\pi_n}(s_0)$$

This is something
we can control!
And this is related
to our policy π^n

Bounding regret using Simulation lemma

$$\widehat{V}^n_H(s) = 0, \quad \widehat{Q}^n_h(s,a) = \min\left\{ r_h(s,a) + b^n_h(s,a) + \widehat{P}^n_h(\cdot\,|\,s,a)\cdot \widehat{V}^n_{h+1}, H \right\}$$

$$\widehat{V}^n_h(s) = \max_a \widehat{Q}^n_h(s,a), \quad \pi^n_h(s) = \arg\max_a \widehat{Q}^n_h(s,a), \forall s$$

Simulation lemma for finite horizon: Value of policy π^n under $\hat{P}$ vs. P at step h

$$\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) \leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d^{\pi^n}_h}\left[b^n_h(s,a) + (\widehat{P}^n_h(\cdot\,|\,s,a) - P_h(\cdot\,|\,s,a))\cdot \widehat{V}^n_{h+1} \right]$$

Proof: $\widehat{V}^n_0(s_0) - V^{\pi^n}_0(s_0) = \widehat{Q}^n_0(s_0, \pi^n(s_0)) - Q^{\pi^n}_0(s_0, \pi^n(s_0))$

$$\leq r_0(s_0,\pi^n(s_0)) + b^n_h(s_0,\pi^n(s_0)) + \widehat{P}^n_0(\cdot\,|\,s_0,\pi^n(s_0))\cdot \widehat{V}^n_1 - r_0(s_0,\pi^n(s_0)) - P_0(\cdot\,|\,s_0,\pi^n(s_0))\cdot V^{\pi^n}_1$$

$$= b^n_h(s_0,\pi^n(s_0)) + \widehat{P}^n_0(\cdot\,|\,s_0,\pi^n(s_0))\cdot \widehat{V}^n_1 - P_0(\cdot\,|\,s_0,\pi^n(s_0))\cdot V^{\pi^n}_1$$

$$= b^n_h(s_0,\pi^n(s_0)) + \left(\widehat{P}^n_0(\cdot\,|\,s_0,\pi^n(s_0)) - P_0(\cdot\,|\,s_0,\pi^n(s_0)) \right)\cdot \widehat{V}^n_1 + P_0(\cdot\,|\,s_0,\pi^n(s_0))\cdot \left(\widehat{V}^n_1 - V^{\pi^n}_1 \right)$$

$$= \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d^{\pi^n}_h}\left[b^n_h(s,a) + (\widehat{P}^n_h(\cdot\,|\,s,a) - P_h(\cdot\,|\,s,a))\cdot \widehat{V}^n_{h+1} \right]$$

Bounding regret using Simulation lemma

$$\text{per-episode regret} := V_0^\star(s_0) - V_0^{\pi_n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi_n}(s_0)$$

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + \left(\widehat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a) \right) \cdot \widehat{V}_{h+1}^n \right]$$

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}} \right]$$ w.p. > 1-δ

$$\leq 2\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}} \right]$$

$$\left(\widehat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a) \right) \cdot \widehat{V}_{h+1}^n \leq \|P_h(\cdot|s,a) - \widehat{P}_h^n(\cdot|s,a)\|_1 \|\widehat{V}_{h+1}^n\|_\infty$$

$$\leq H\|P_h(\cdot|s,a) - \widehat{P}_h^n(\cdot|s,a)\|_1 \leq H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}}, \forall s,a,h,n, \text{with prob} 1-\delta$$

Bounding regret using Simulation lemma

$$\text{per-episode regret} := V_0^\star(s_0) - V_0^{\pi_n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi_n}(s_0)$$

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + (\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a)) \cdot \widehat{V}_{h+1}^n \right]$$

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[b_h^n(s,a) + H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}} \right]$$

w.p. > 1-δ

$$\leq 2\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[H\sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}} \right]$$

$$= 2H\sqrt{S\ln(SAHN/\delta)} \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}} \left[\sqrt{\frac{1}{N_h^n(s,a)}} \right]$$

Regret bound UCB-VI

per-episode regret $:= V_0^\star(s_0) - V_0^{\pi_n}(s_0)$

$$\leq 2H\sqrt{S\ln(SAHN/\delta)}\sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim d_h^{\pi^n}}\left[\sqrt{\frac{1}{N_h^n(s,a)}}\right]$$

Total regret

$$\mathbb{E}\left[\text{Regret}_N\right] \leq \mathbb{E}\left[\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right] + 2\delta NH$$

$$\leq H\sqrt{S\ln(SANH/\delta)}\mathbb{E}\left[\sum_{n=1}^{N}\sum_{h=0}^{H-1}\frac{1}{\sqrt{N_h^n(s_h^n, a_h^h)}}\right] + 2\delta NH$$

Regret bound UCB-VI

$$\sum_{n=1}^{N}\sum_{h=0}^{H-1}\frac{1}{\sqrt{N_h^n(s_h^n,a_h^n)}} = \sum_{h=0}^{H-1}\sum_{s,a}\sum_{i=1}^{N_h^N(s,a)}\frac{1}{\sqrt{i}} \leq \sum_{h=0}^{H-1}\sum_{s,a}\sqrt{N_h^N(s,a)}$$

$$\leq \sum_{h=0}^{H-1}\sqrt{SA\sum_{s,a}N_h^N(s,a)}$$

$$\leq \sum_{h=0}^{H-1}\sqrt{SAN} = H\sqrt{SAN}$$

$$\mathbb{E}\left[\text{Regret}_N\right] \leq 2H^2S\sqrt{AN\ln(SAHN/\delta)} + 2\delta NH$$

Set $\delta = 1/(HN)$

$$\leq 2H^2S\sqrt{AN\cdot\ln(SAH^2N^2)} = \widetilde{O}\left(H^2S\sqrt{AN}\right)$$

High-level idea: Exploration-Exploitation tradeoff

Upper bound per-episode regret: $V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$

1. What if $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \leq \epsilon$?

Then π^n is close to $\pi^\star$, i.e., we are doing exploitation

2. What if $\widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \geq \epsilon$?

$$\epsilon \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^n}}\left[b_h^n(s,a) + (\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a)) \cdot \widehat{V}_{h+1}^n \right]$$

We collect data at steps where bonus is large or model is wrong, i.e., exploration

RL in generative vs. online setting

Generative

reset to any state

Online

reset to initial state only

RL in generative vs. online setting

Generative

reset to any state

obtain fixed amount of data for each state-action pair

Online

reset to initial state only

online roll-out don't guarantee fixed amount of data per (s,a)

RL in generative vs. online setting

Generative	Online
reset to any state	reset to initial state only
obtain fixed amount of data for each state-action pair	online roll-out don't guarantee fixed amount of data per (s,a)
plug-in and exploit	explore-exploit using confidence

RL in generative vs. online setting

Generative

reset to any state

obtain fixed amount of data for each state-action pair

plug-in and exploit

Regret, $E[V^* - V^{\hat{\pi}^*}] \leq \epsilon$

if scalar samples SAN = $\tilde{O}\left(\frac{S^2 A}{\epsilon^2 (1-\gamma)^4}\right)$

Online

reset to initial state only

online roll-out don't guarantee fixed amount of data per (s,a)

explore-exploit using confidence

Regret, $E[\sum_{n=1}^{N}(V^* - V^{\pi_n})] \leq N\epsilon$

if scalar samples NH = $\tilde{O}\left(\frac{H^5 S^2 A}{\epsilon^2}\right)$

RL in generative vs. online setting

Generative	**Online**
reset to any state	reset to initial state only
obtain fixed amount of data for each state-action pair	online roll-out don't guarantee fixed amount of data per (s,a)
plug-in and exploit	explore-exploit using confidence
Regret, $E[V^* - V^{\hat{\pi}^*}] \leq \epsilon$	Regret, $E[\sum_{n=1}^{N}(V^* - V^{\pi_n})] \leq N\epsilon$
if scalar samples SAN = $\tilde{O}\left(\frac{S^2 A}{\epsilon^2(1-\gamma)^4}\right)$	if scalar samples NH = $\tilde{O}\left(\frac{H^5 S^2 A}{\epsilon^2}\right)$
improve to remove S, 1/(1-γ)	improve to remove S, H (tighter bonus via Bernstein concentration)

Online worse by H since assume non-stationary at each of the H steps!

Next Questions

- How to handle unknown state transition and reward functions?

Done!

- How to handle continuous states and actions?

Next