Sample Complexity Reduction via Policy Difference Estimation in Tabular RL

Adhyyan Narang, Andrew Wagenmaker, Lillian Ratliff, Kevin Jamieson

University of Washington

Motivation: Exploration in RL

UCB: Plays estimated optimal actions with a bonus term for exploration.

Wagenmaker et.al, 2022: Plays "informative" actions to estimate the value of each policy individually.

$$\sum_{h=1}^{H} \inf_{\pi_{\exp}} \max_{\pi \in \Pi} \frac{\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}} + \|\phi_h^\star\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max(\Delta(\pi)^2, \epsilon^2)}$$

Li et.al, 2022: Obtains complexity in terms of estimating the value of *differences between policies*. This can be arbitrarily better when policies are similar (see right).

$$\rho_\Pi := \sum_{h=1}^{H} \inf_{\pi_{\exp}} \max_{\pi \in \Pi} \frac{\|\phi_h^\star - \phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2, \Delta(\pi)^2\}}$$

Present Work: Q1 Can we obtain this complexity for Tabular MDP?
Q2 If yes, what algorithmic insights does this provide?

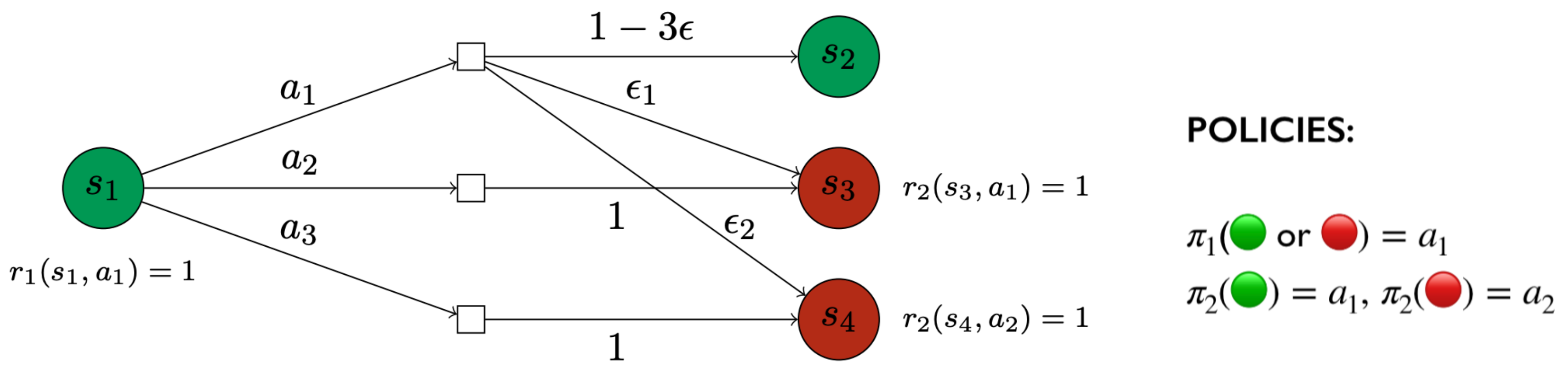
Preliminaries

- Episodic, finite-horizon, time inhomogeneous and tabular MDPs, denoted by $(\mathcal{S}, \mathcal{A}, H, \{P_h\}, \{r_h\})$.
- P_h denotes transition matrix and r_h the reward function at time h.
- Define $\phi_h^\pi(s,a)$ as the probability that policy π visits state s and plays action a at time h.
- Define $Q_h^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \middle| s_h = s, a_h = a\right]$.
- Define $V_h^\pi(s) = \mathbb{E}_{a\sim\pi}[Q_h^\pi(a,s)]$.

(ϵ, δ) **Best Policy Identification:** Given a set of policies Π, we want to find a policy $\hat{\pi}$ that is within ϵ of the best policy with probability $(1-\delta)$.

- Define $\Delta(\pi) = \max_{\mu \in \Pi} V_0^\mu - V_0^\pi$
- $\Lambda_h(\pi) = \sum_{s,a} \phi_h^\pi(s,a)\, e_{sa} e_{sa}^\top$

Lower Bound: Negative Answer to Q1



Proposition (Informal) For this example instance,
- ρ_Π = Constant,
- PEDEL from (Wagenmaker, 2022) = $1/\epsilon^2$,
- Lower Bound: Any (ϵ, δ)-PAC algorithm must consume at least $1/\epsilon$ samples.

Main Upper Bound: Semi-positive Answer to Q1

Theorem (Informal) PERP finds an ϵ-optimal policy with probability $(1-\delta)$ and consumes (upto lower order terms) at most
- For any MDP: $\left(\rho_\Pi + \frac{U(\pi,\pi^\star)}{\max\left(\epsilon^2, \Delta(\pi)^2\right)}\right) \log\left(\frac{|\Pi|}{\delta}\right)$ samples
- For contextual bandits: $\rho_\Pi \log\left(\frac{|\Pi|}{\delta}\right)$ samples.

Above, $U(\pi, \pi^\star) := \sum_{h=1}^{H} \mathbb{E}_{s\sim w_h^{\pi^\star}}\left[\left(Q_h^\pi(s, \pi_h(s)) - Q_h^\pi(s, \pi_h^\star(s))\right)^2\right]$,

- On example, the new term is $1/\epsilon$ and matches the lower bound.
- Best known complexity for Tabular MDPs.
- New Term → Estimating the value of a single reference policy $\bar{\pi}$, after which we pay ρ_Π to estimate the difference between $\bar{\pi}$ and any other π.

Algorithm

Algorithm 1 PERP: Policy Elimination with Reference Policy (shortened)

Require: tolerance ϵ, confidence δ, policies Π

$\Pi_1 \leftarrow \Pi$, $\epsilon_\ell \leftarrow 2^{-\ell}$
for $\ell = 1, 2, \ldots, \lceil \log \frac{1}{\epsilon} \rceil$ **do**
  Choose "centroid" policy $\bar{\pi}_\ell \in \Pi_\ell$
  Collect $\mathfrak{D}_{\bar{\pi}}$ by playing $\bar{\pi}_\ell$ with $\bar{n}_\ell \leftarrow O\left(\max_{\pi \in \Pi_\ell} \frac{\widehat{U}_{\ell-1}(\pi, \bar{\pi}_\ell)}{\epsilon_\ell^2} \cdot \log \frac{|\Pi_\ell|}{\delta}\right)$
  Estimate $\widehat{w}_h^{\bar{\pi}}$ from $\mathfrak{D}_{\bar{\pi}}$
  for $h = 1, 2, \ldots, H$ **do**
    Collect data $\mathfrak{D}_{\mathrm{FW}}$ using procedure from (Wagenmaker, 2022) satisfying:

$$\sup_{\pi \in \Pi_\ell} \|\widehat{\phi}_h^{\bar{\pi}_\ell} - \widehat{\phi}_h^\pi\|^2_{\Lambda_{\ell,h}^{-1}} \le \epsilon_\ell^2 \quad \text{for} \quad \Lambda_{\ell,h} = \sum_{(s,a)\in\mathfrak{D}_{\mathrm{FW}}} e_{sa} e_{sa}^\top$$

  end for
  Compute $\widehat{\Delta}_{\bar{\pi}_\ell}(\pi)$ and update:

$$\Pi_{\ell+1} \leftarrow \Pi_\ell \setminus \left\{\pi \in \Pi_\ell : \max_{\pi'} \widehat{\Delta}_{\bar{\pi}_\ell}(\pi') - \widehat{\Delta}_{\bar{\pi}_\ell}(\pi) > \epsilon_\ell\right\}$$

end for
return any $\pi \in \Pi_{\ell+1}$

- In the example, PERP would play a_2 because this gets us to the RED STATE that we care about.
- UCB, PEDEL would play a_1.

Keys to the Analysis: Answer to Q2

- Instead of estimating V_0^π directly, use estimator $\hat{\Delta}_{\bar{\pi}}(\pi)$ above for $\Delta_{\bar{\pi}}(\pi) = V_0^\pi - V_0^{\bar{\pi}}$.
- Actively collected data to cover states where policies disagree → $\hat{\Delta}_{\bar{\pi}}(\pi)$ is reduced-variance estimator → State of the art sample complexities.

Key insight: Playing informative actions to collect exploratory data where policies disagree can lead to large sample complexity savings!

Correspondence: adhyyan@uw.edu

Sample Complexity Reduction via Policy Difference Estimation in Tabular RL

Adhyyan Narang, Andrew Wagenmaker, Lillian Ratliff, Kevin Jamieson

University of Washington

Motivation

- In contextual bandits, (Li et.al, 2022) obtains complexity in terms of estimating the value of differences between policies.

$$\rho_\Pi := \sum_{h=1}^{H} \inf_{\pi_{\exp}} \max_{\pi \in \Pi} \frac{\|\phi_h^\star - \phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2, \Delta(\pi)^2\}}$$

- Best known complexity in Tabular MDP (Wagenmaker et.al, 2022) is terms of estimating the value of each policy individually.

$$\sum_{h=1}^{H} \inf_{\pi_{\exp}} \max_{\pi \in \Pi} \frac{\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}} + \|\phi_h^\star\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max(\Delta(\pi)^2, \epsilon^2)}$$

- This can be arbitrarily worse when policies are similar (see right).

Main Questions:

Q1 Can we obtain this complexity for Tabular MDP?

Q2 If yes, what algorithmic insights does this provide?

Preliminaries

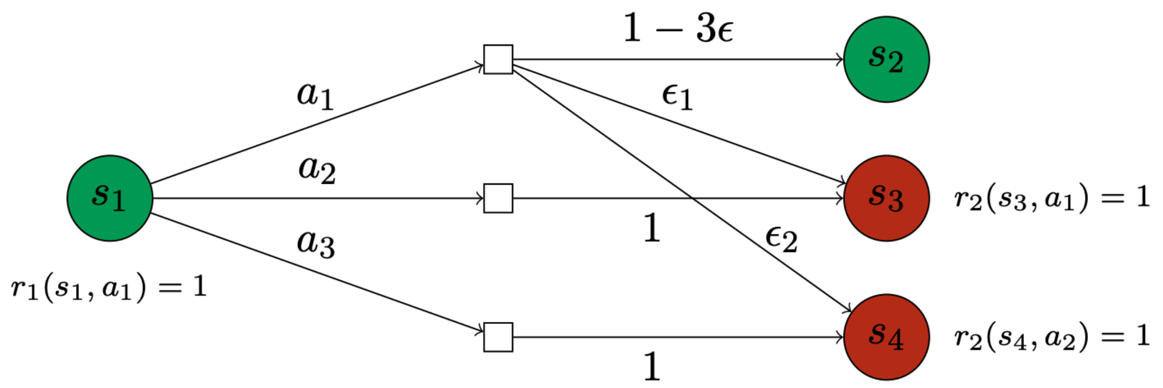
- Episodic, finite-horizon, time inhomogeneous and tabular MDPs, denoted by $(\mathcal{S}, \mathcal{A}, H, \{P_h\}, \{r_h\})$.
- P_h denotes transition matrix and r_h the reward function at time h.
- Define $\phi_h^\pi(s, a)$ as the probability that policy π visits state s and plays action a at time h.
- Define $Q_h^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \middle| s_h = s, a_h = a \right]$.
- Define $V_h^\pi(s) = \mathbb{E}_{a \sim \pi}[Q_h^\pi(a, s)]$.

(ϵ, δ) Best Policy Identification

Given a set of policies Π, we want to find a policy $\hat{\pi}$ that is within ϵ of the best policy with probability $(1 - \delta)$.

- Define $\Delta(\pi) = \max_{\mu \in \Pi} V_0^\mu - V_0^\pi$

Lower Bound: Negative Answer to Q1



POLICIES:

π_1(green or red) $= a_1$

π_2(green) $= a_1$, π_2(red) $= a_2$

Proposition (Informal) For this example instance,
- ρ_Π = Constant,
- PEDEL from (Wagenmaker, 2022) = $1/\epsilon^2$,
- Lower Bound: Any (ϵ, δ)-PAC algorithm must consume at least $1/\epsilon$ samples.

Main Upper Bound: Semi-positive Answer to Q1

Theorem (Informal) PERP finds an ϵ-optimal policy with probability $(1-\delta)$ and consumes (upto lower order terms) at most
- For any MDP: $\left(\rho_\Pi + \frac{U(\pi, \pi^\star)}{\max\left(\epsilon^2, \Delta(\pi)^2\right)}\right) \log\left(\frac{|\Pi|}{\delta}\right)$ samples
- For contextual bandits: $\rho_\Pi \log\left(\frac{|\Pi|}{\delta}\right)$ samples.

Above, $U(\pi, \pi^\star) := \sum_{h=1}^{H} \mathbb{E}_{s \sim w_h^{\pi^\star}} \left[\left(Q_h^\pi(s, \pi_h(s)) - Q_h^\pi(s, \pi_h^\star(s)) \right)^2 \right]$,

- On example, the new term is $1/\epsilon$ and matches the lower bound.
- Best known complexity for Tabular MDPs.

Algorithm

Algorithm 1 PERP: Policy Elimination with Reference Policy (shortened)

Require: tolerance ϵ, confidence δ, policies Π

1: $\Pi_1 \leftarrow \Pi$, $\epsilon_\ell \leftarrow 2^{-\ell}$
2: **for** $\ell = 1, 2, \ldots, \lceil \log \frac{1}{\epsilon} \rceil$ **do**
3: Choose "centroid" policy $\bar{\pi}_\ell \in \Pi_\ell$
4: Collect $\mathfrak{D}_{\bar{\pi}}$ by playing $\bar{\pi}_\ell$ with $\bar{n}_\ell \leftarrow O\left(\max_{\pi \in \Pi_\ell} \frac{\widehat{U}_{\ell-1}(\pi, \bar{\pi}_\ell)}{\epsilon_\ell^2} \cdot \log \frac{|\Pi_\ell|}{\delta}\right)$
5: Estimate $\widehat{w}_h^{\bar{\pi}}$ from $\mathfrak{D}_{\bar{\pi}}$
6: **for** $h = 1, 2, \ldots, H$ **do**
7: Collect data $\mathfrak{D}_{\mathrm{FW}}$ satisfying:

$$\sup_{\pi \in \Pi_\ell} \|\widehat{\phi}_h^{\bar{\pi}_\ell} - \widehat{\phi}_h^\pi\|^2_{\Lambda_{\ell,h}^{-1}} \le \epsilon_\ell^2 \quad \text{for} \quad \Lambda_{\ell,h} = \sum_{(s,a) \in \mathfrak{D}_{\mathrm{FW}}} e_{sa} e_{sa}^\top$$

8: **end for**
9: Compute $\widehat{\Delta}_{\bar{\pi}_\ell}(\pi)$ and update:

$$\Pi_{\ell+1} \leftarrow \Pi_\ell \setminus \left\{ \pi \in \Pi_\ell : \max_{\pi'} \widehat{\Delta}_{\bar{\pi}_\ell}(\pi') - \widehat{\Delta}_{\bar{\pi}_\ell}(\pi) > \epsilon_\ell \right\}$$

10: **end for**
11: **return** any $\pi \in \Pi_{\ell+1}$

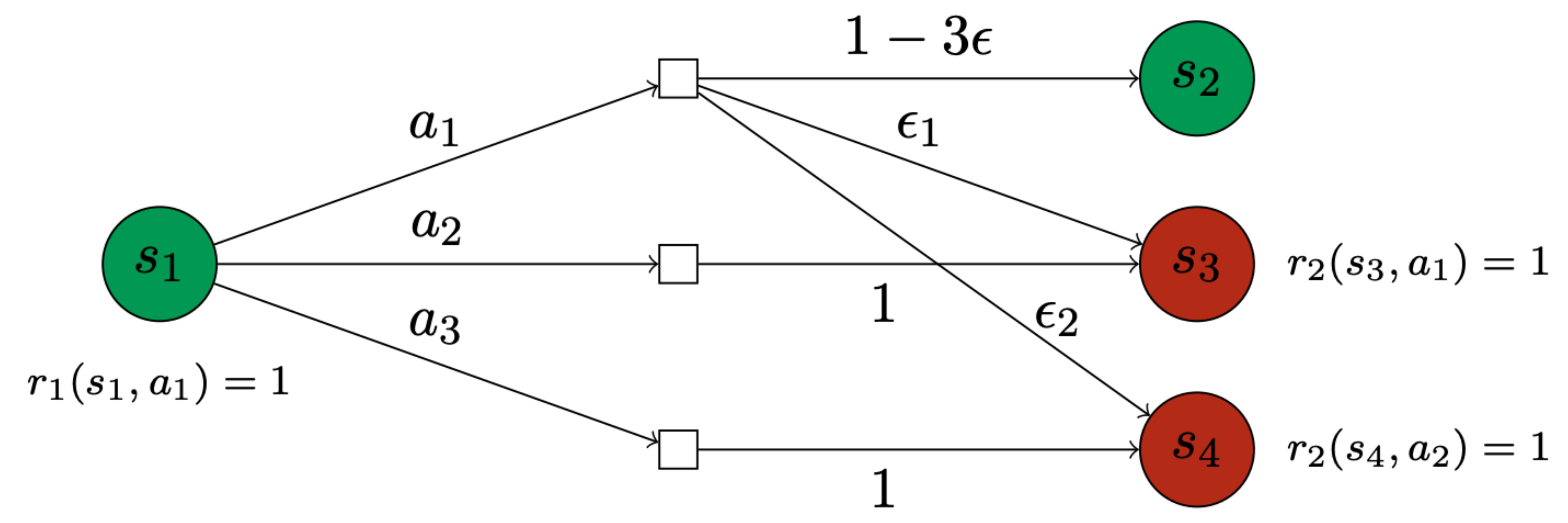
- In the example, PERP would play a_2 because this gets us to the RED STATE that we care about.
- UCB, PEDEL would play a_1.

Keys to the Analysis: Answer to Q2

- Instead of estimating V_0^π directly, use estimator $\hat{\Delta}_{\bar{\pi}}(\pi)$ above for $\Delta_{\bar{\pi}}(\pi) = V_0^\pi - V_0^{\bar{\pi}}$.
- Actively collected data to cover states where policies disagree → $\hat{\Delta}_{\bar{\pi}}(\pi)$ is reduced-variance estimator → State of the art sample complexities.

Key insight: Playing informative actions to collect exploratory data where policies disagree can lead to large sample complexity savings!

Correspondence: adhyyan@uw.edu



POLICIES:

π_1(🟢 or 🔴) $= a_1$

π_2(🟢) $= a_1$, π_2(🔴) $= a_2$