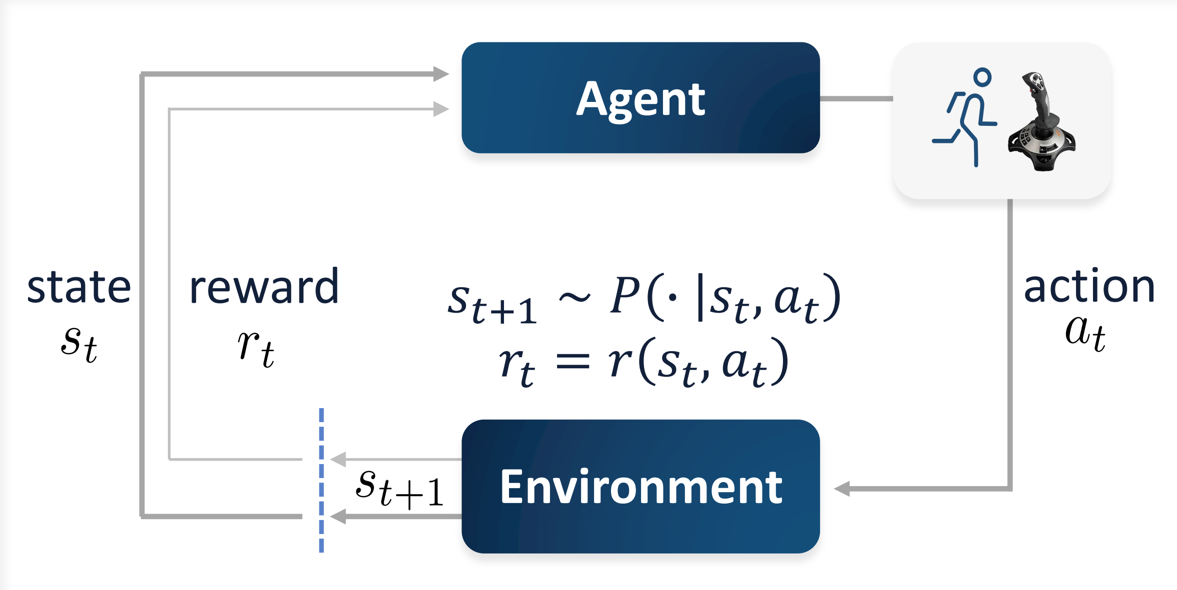
Nearly Optimal Reward-free Reinforcement Learning

Zihan Zhang Simon S. Du Xiangyang Ji

2021/7/21

Episodic Finite-Horizon MDP



Repeat **H** times
H: planning horizon / Episode length

Play **K** episodes in total

A policy π :
π: States(S) $\rightarrow$ Actions (A), $a = \pi(s)$

Goal: maximize value function

$$V^\pi(s_1) = \mathbb{E}\left[r_1 + r_2 + \cdots r_H \,\middle|\, s_1, \pi\right]$$

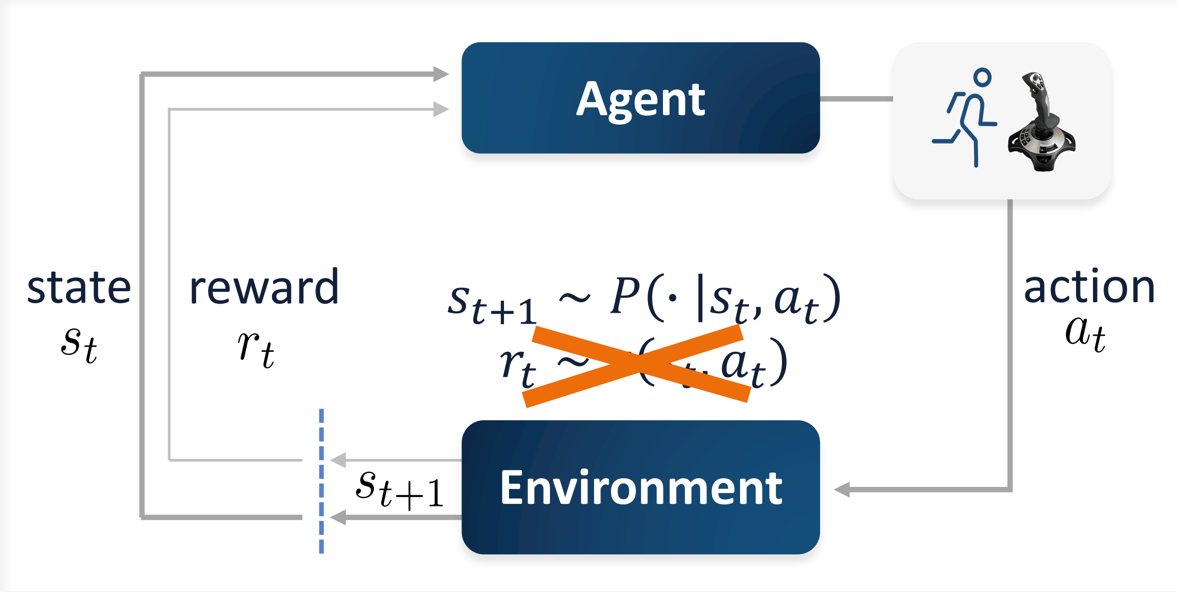
Goal: given $0 < \epsilon \leq 1$, find π such that

$$\mathbb{E}_{s_1 \sim \mu}[V^*(s_1) - V^\pi(s_1)] \leq \epsilon$$

$V^* = V^{\pi^*}$: value function of opt policy

V^π: value function of policy π

Reward-Free RL [Jin et al. 2020]



Reward is unknown during interactions

Reward is defined by the user afterward (depends on the collected data)

Exploration Phase:

Interacts with the environment and collects a dataset:

$$\mathcal{D} = \{(s_h^k, a_h^k)\}_{(h,k)=(1,1)}^{(H,K)}$$

Planning Phase:

Given an arbitrary reward $r(\cdot,\cdot)$, compute a policy π:

$$\mathbb{E}_{s_1\sim\mu}[V^*(s_1) - V^\pi(s_1)] \le \epsilon$$

Sample Complexity:

How many episodes (K) needed?

Motivations

Batch Reinforcement Learning

- Existing results: if the collected dataset has a **good coverage**, we can compute a near-optimal policy.
- Reward free RL: how to collect a dataset with **good coverage?**

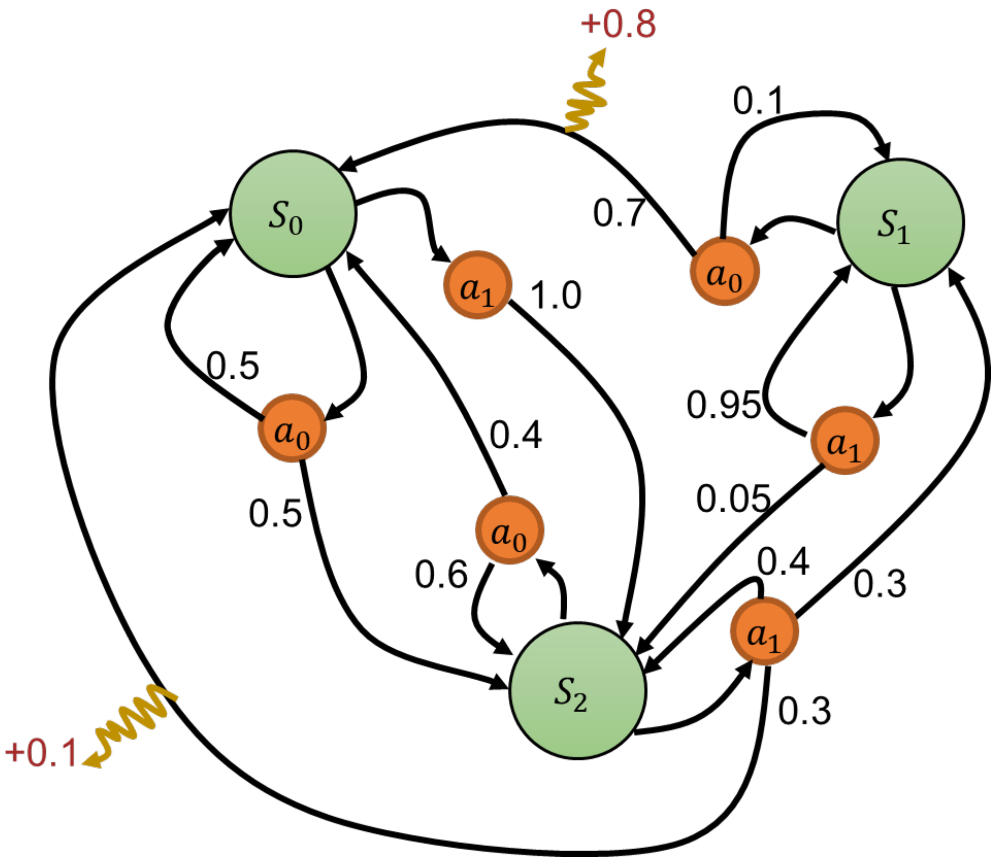
Constrained Reinforcement Learning

- Reward functions are iteratively engineered to encourage desired behavior via trial and error
- Don’t want to repeatedly interact with the environment.

Tabular Markov Decision Process

Assumptions:

1. \# of States $\boldsymbol{S} < \infty$
2. \# of actions $\mathbf{A} < \infty$
3. Homogenous transition: $P(\cdot \mid \cdot,\cdot)$ is independent of h
4. Bounded total rewards:
   $r_h \geq 0, h = 1, \ldots, H$
   $r_1 + r_2 + \cdots + r_H \leq \mathbf{1}$



Reward Scaling Assumptions

$\epsilon \in (0,1)$: measures the performance relative to the **total reward**

Uniformly Bounded Reward
$0 \leq r_h \leq 1, h = 1, \ldots, H$

Total Reward: H => rescale $\epsilon = \epsilon \times H$

Uniformly Bounded Reward (Rescaled)
$0 \leq r_h \leq 1/H, h = 1, \ldots, H$

Total Reward: 1 => right scaling
but per-step reward is tiny

Bounded Total Reward
$r_h \geq 0, h = 1, \ldots, H$ $r_1 + r_2 + \cdots + r_H \leq \mathbf{1}$

Benefit: can model sparse spiky reward
[Kakade 03, Jiang and Agarwal 18]

Existing Results

Paper	Algorithm	Sample Complexity
Brafman and Tenenholtz 2002	Rmax / ZeroRMax	$\frac{\boldsymbol{H}^{\mathbf{8}} S^2 A}{\epsilon^3}$
Jin Krishnamurthy Simchowitz Yu 2/2020	RF-RL-Explore	$\frac{\boldsymbol{H}^{\mathbf{3}} S^2 A}{\epsilon^2}$
Kaufmann Ménard Domingues Jonsson Laurent Valko 6/2020	RF-UCRL	$\frac{\boldsymbol{H}^{\mathbf{2}} S^2 A}{\epsilon^2}$
Ménard Domingues Jonsson Kaufmann Laurent Valko 7/2020	RF-Express	$\frac{\boldsymbol{H} S^2 A}{\epsilon^2}$
This work	**SSTP**	$\boldsymbol{\frac{S^2 A}{\epsilon^2}}$
Jin Krishnamurthy Simchowitz Yu 2/2020	Lower Bound	$\boldsymbol{\frac{S^2 A}{\epsilon^2}}$

All bounds are in Big-O / Big-Omega and ignore logarithmic factors.

Main Result

Under the bounded total reward assumption: $r_h \geq 0$, for $h = 1, \ldots, H$, and $r_1 + r_2 + \cdots + r_H \leq 1$, Staged Sampling + Truncated Planning (**SSTP**) solves reward-free RL using at most $\tilde{\boldsymbol{O}}(\frac{\boldsymbol{S^2 A}}{\boldsymbol{\epsilon^2}})$ episodes in the exploration phase.

- Matches $\Omega(\frac{\boldsymbol{S^2 A}}{\boldsymbol{\epsilon^2}})$ lower bound up to logarithmic factors.

- Reward-free Tabular RL is almost independent of the planning horizon:
  - $\mathbf{log\ H}$ bounds have been obtained in tabular RL: [Wang D. Yang Kakade 2020], [Zhang Ji D. 2020]: $\tilde{O}(\sqrt{SAK} + S^2 A)$ regret / $\tilde{O}(\frac{\boldsymbol{SA}}{\boldsymbol{\epsilon^2}} + \frac{\boldsymbol{S^2 A}}{\boldsymbol{\epsilon}})$ sample complexity.

A Sufficient Condition

Observation [Jin et al. 2020]:
Maximal expected visitation count:

$$\lambda(s,a) = \max_{\pi} \mathbb{E}[\, N(s,a) \mid \pi \,], \quad 0 \le \lambda(s,a) \le H$$

$N(s,a)$: number of visitation in an episode.

- If in the dataset, for every (s,a) pair, we have $\tilde{N}\lambda(s,a)$ data with $\tilde{N} \sim \mathrm{poly}(S, H, 1/\epsilon)$, then we can compute an ϵ-optimal policy with any (approximate) MDP solver.

Algorithm for exploration:

- For every state-action pair (s,a):
  - Set reward $r(s,a) = 1$ and all other pair $(s',a') \neq (s,a)$, $r(s',a') = 0$
  - Run a SOTA algorithm for tabular RL (as blackbox) to collect as many (s,a) as possible.

A Tighter Sufficient Condition

Observation [This work]:
If in the dataset, for every (s, a) pair, we have $\tilde{N}\lambda(s, a)$ data with $\boldsymbol{\tilde{N} \sim S/\epsilon^2}$ then we can compute an ϵ-optimal policy with an **optimistic planning algorithm**.

- Planning algorithm: adding **Bernstein bonus** in dynamic programming.

Q1: $\lambda(s, a)$ is unknown

A Tighter Sufficient Condition (Cont'd)

Discretization by doubling:

- $\mathcal{S} \times \mathcal{A} = \mathcal{X}_1 \cup \mathcal{X}_2 \cdots \mathcal{X}_M$: $(s, a) \in \mathcal{X}_i \Rightarrow \lambda(s, a) \sim H/2^i$. $M = \log_2(H/\epsilon)$
- Sufficient condition:
  - For every $(s, a) \in \mathcal{X}_i$, we have $N_{s,a}(\mathcal{D}) = \Omega(\frac{SH}{2^i \epsilon^2})$ where $N_{s,a}(\mathcal{D})$ is the visitation counts of (s, a) in the collected dataset $\mathcal{D}$.

Q2: How to collect a dataset that satisfies this condition?

Staged Sampling

Initialize $\mathcal{Y}_1 = \mathcal{S} \times \mathcal{A}$
For $i = 1, \ldots, M$
- Set $r(s,a) = 1$ for **all** $(s,a) \in \mathcal{Y}_i$:
- Run a regret-minimization tabular algorithm with n episodes. For each episode:
  - For $(s,a) \in \mathcal{Y}_i$, if $N(s,a) \geq \frac{SH}{2^i \epsilon^2}$ where $N(s,a)$ is the visitation count of (s,a) up to this episode: set $r(s,a) = 0$.
- Denote $\mathcal{Y}_{i+1} = \{(s,a) \in \mathcal{Y}_i, N(s,a) < O(\frac{SH}{2^i \epsilon^2})\}$,

Proof:

- Need: n is large enough such that the algorithm guarantees to collect $\Omega(\frac{SH}{2^i \epsilon^2})$ samples for $\lambda(s,a) \geq \frac{H}{2^i}$.
- Choose **MVP** in [Zhang Ji D. 2020] ($\tilde{O}(\sqrt{SAK} + S^2A)$ regret for standard RL setting) as the algorithm.
- Use a reward-varying regret analysis.

=> $n = \tilde{O}\left(\frac{S^2A}{\epsilon^2}\right)$.

Conclusion

New Algorithm: SSTP for Reward-free Reinforcement Learning

- Sample complexity: $\tilde{O}(S^2 A/\epsilon^2)$.
- The sample complexity can be almost **independent** of planning horizon $\boldsymbol{H}$.
- Matches $\Omega(S^2 A/\epsilon^2)$ lower bound up to logarithmic factors.

Thank You