Optimistic Posterior Sampling for Reinforcement Learning with Few Samples and Tight Guarantees

D. Tiapkin[1], D. Belomestny[2,1], D. Calandriello[3], É. Moulines[4], R. Munos[3], A. Naumov[1], M. Rowland[3], M. Valko[3], P. Ménard[5]

[1]HSE University, [2]Duisburg-Essen University, [3]DeepMind, [4]École Polytechnique, [5]ENS Lyon

Overview

- OPSRL algorithm with minimax optimal regret bound up to poly-log factors for large enough T;
- First theoretically and computationally efficient modification of posterior sampling;
- Resolves 2 open problems by [Agrawal and Jia, 2017];
- Novel tight anti-concentration inequality for weighted sums of Dirichlet random variables;

Setting

- Tabular MDP: H horizon, S states, A actions, $p_h(s'|s,a)$ unknown transitions, deterministic reward $r_h(s,a) \in [0,1]$.
- Regret: $\sum^{T} V_1^\star(s_1) - V_1^{\pi^t}(s_1)$

Optimistic Posterior Sampling for Reinforcement Learning

- UCBVI bonus-based exploration (theoretically near optimal, empirically bad)

$$\overline{Q}_h^t(s,a) = r_h(s,a) + \widehat{p}_h^t \overline{V}_{h+1}^t(s,a) + B_h^t(s,a), \qquad \overline{V}_h^t(s) = \max_a \overline{Q}_h^t(s,a),$$

where $\widehat{p}_h^t(s,a)$ is empirical transition probabilities, and $\widehat{p}_h^t f(s,a) \triangleq \sum_{s' \in \mathcal{S}} \widehat{p}_h^t(s'|s,a) f(s')$.

- PSRL exploration (no known regret guarantees, empirically good)

$$\widetilde{Q}_h^t(s,a) = r_h(s,a) + \widetilde{p}_h^t \widetilde{V}_{h+1}^t(s,a), \qquad \widetilde{V}_h^t(s) = \max_a \widetilde{Q}_h^t(s,a),$$

where $\widetilde{p}_h^t(s,a) \sim \rho_h^t(s,a)$ is sample from posterior distribution for transition probabilities.

- OPSRL exploration (theoretically near optimal, empirically good)

$$\overline{Q}_h^t(s,a) = r_h(s,a) + \max_{j \in [H]} \widetilde{p}_h^{t,j} \widetilde{V}_{h+1}^t(s,a), \qquad \overline{V}_h^t(s) = \max_a \overline{Q}_h^t(s,a),$$

where $\widetilde{p}_h^{t,j}(s,a) \sim \rho_h^t(s,a)$ are $J = \widetilde{\mathcal{O}}(1)$ samples from posterior distribution for transition probabilities.

Regret bounds

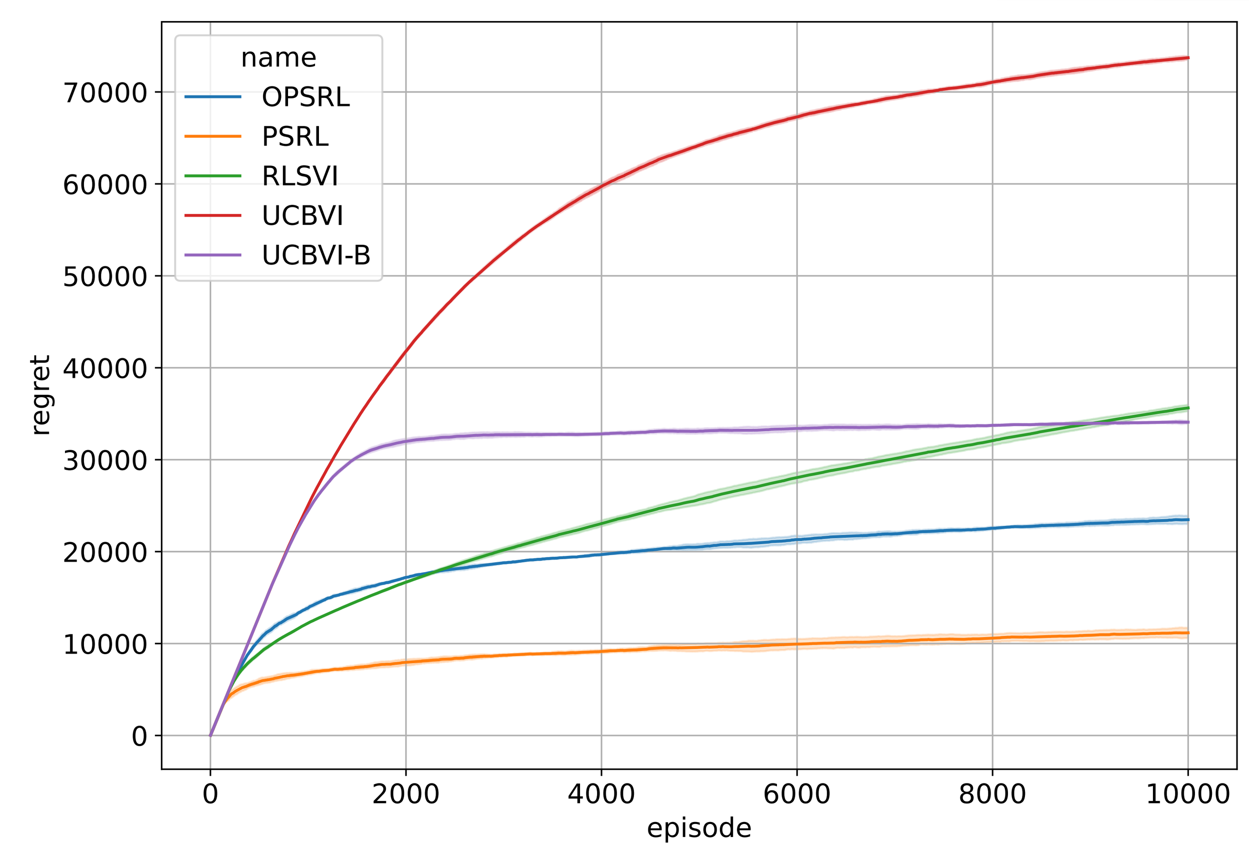
Algorithm	Upper bound (non-stationary)
UCBVI [Azar et al., 2017] UCB-Advantage [Zhang et al., 2020] RLSVI [Xiong et al., 2021]	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$
SOS-OPS-RL [Agrawal and Jia, 2017]	$\widetilde{\mathcal{O}}(\sqrt{H^4S^2AT})$
PSRL [Osband et al., 2013]	N/A
OPSRL (this paper)	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$
Lower bound [Jin et al., 2018, Domingues et al., 2021]	$\Omega(\sqrt{H^3SAT})$

Green: empirically efficient, Orange: empirically fair, Red: empirically poor.

Optimistic prior of OPSRL

- add an artificial isolated state s_0 with $r_h(s_0,a) > 1$;
- add n_0 pseudo-transitions from each state s to s_0 into the history of visits.
- use posterior inflation by $\kappa = \widetilde{\mathcal{O}}(1)$;

Experiments



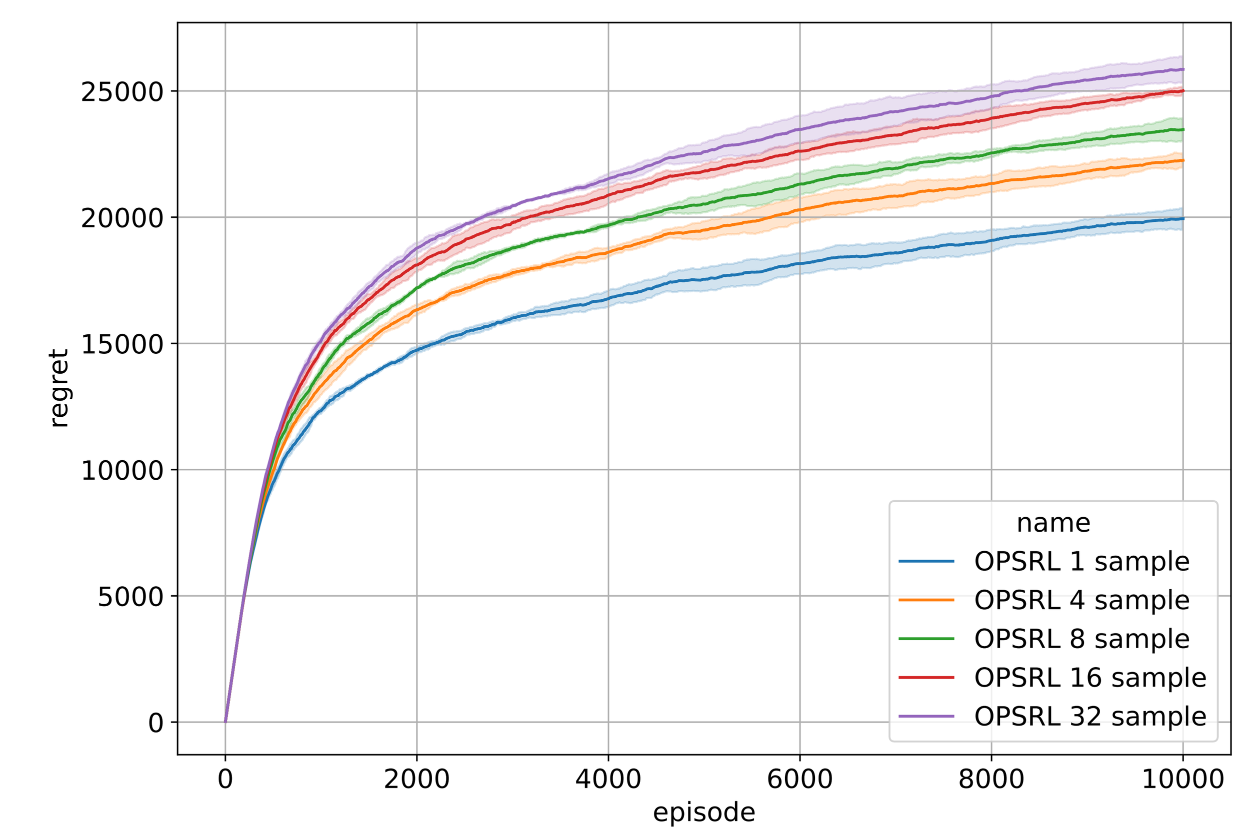


Figure 1: **Left:** Regret of OPSRL and baselines on grid-world environment with 100 states and 4 action for $H = 50$ an transitions noise 0.2, average over 4 seeds. **Right:** regret of OPSRL for $J \in \{1,4,8,16,32\}$ on the same environment.

Upper and lower bounds on tails for Dirichlet weighted sum

For any $\alpha = (\alpha_0 + 1, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}_+^{m+1}$ define $\overline{p} \in \Delta_m$ with $\overline{p}(\ell) = \alpha_l/\overline{\alpha}, \ell = 0, \ldots, m$, where $\overline{\alpha} = \sum_{j=0}^m \alpha_j$ and $\overline{p}'(\ell) = (\alpha_\ell + \mathbb{1}\{\ell = 0\})/(\overline{\alpha}+1)$. Under technical assumptions, for $f\colon \{0,\ldots,m\} \to [0,b_0]$ and $\mu \in (\overline{p}f, b_0)$

$$(1-\varepsilon)\left(1 - \Phi\left(\sqrt{2\overline{\alpha}\,\mathcal{K}_{\inf}(\overline{p},\mu,f)}\right)\right) \le \mathbb{P}_{w \sim \mathcal{D}\mathrm{ir}(\alpha)}[wf \ge \mu] \le \exp\left(-(\overline{\alpha}+1)\,\mathcal{K}_{\inf}(\overline{p}',\mu,f)\right),$$

where $\Phi(\cdot)$ is CDF of standard normal law and $\mathcal{K}_{\inf}(p,u,f)$ is given by

$$\mathcal{K}_{\inf}(p,u,f) \triangleq \max_{\lambda \in [0,1]} \mathbb{E}_{X \sim p}\left[\log\left(1 - \lambda \frac{f(X)-u}{b_0 - u}\right)\right] = \inf\left\{\mathrm{KL}(p,q) : q \in \Delta_m, qf \ge u\right\}$$

- Lower bound is an essential part for optimism and small number of samples J;
- Upper bound is important for the reduction to UCBVI.
- Main application: bounding linear forms of Dirichelet r.v. (e.g., $\widetilde{p}_h^{t,j}\widetilde{V}_{h+1}^t(s,a)$)

$$\widetilde{p}_h^{t,j}(s,a) \sim \mathcal{D}\mathrm{ir}(\alpha_0 + 1, \alpha_1, \ldots, \alpha_S) \text{ where } \begin{cases} \alpha_0 = n_0/\kappa - 1 \\ \alpha_i = n_h^t(s_i|s,a)/\kappa \end{cases}$$

and $\overline{\alpha} = (\overline{n}_h^t(s,a) - \kappa)/\kappa$.