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

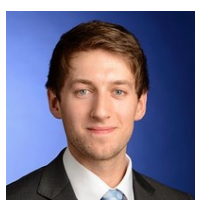
Daniil Tiapkin[1], Denis Belomestny[2,1], Daniele Calandriello[3], Éric Moulines[4], Remi Munos[3], Alexey Naumov[1], Mark Rowland[3], Michal Valko[3], Pierre Ménard[5]

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









OPSRL — NeurIPS 2022 **December 9, 2022**

What do we bring:

- Develop theoretically and computationally efficient version of `PSRL`:

What do we bring:

- Develop theoretically and computationally efficient version of PSRL:
  - Computationally and empirically efficient algorithm;
  - Up till now no near-optimal modifications of PSRL;

Markov Decision Process (MDP)

Tabular, episodic MDP: H horizon, S states, A actions.

Learning in MDP: at episode t, step h

- state s_h^t
- action a_h^t
- next state $s_{h+1}^t \sim p_h(\cdot|s_h^t, a_h^t)$
- reward $r_h(s_h^t, a_h^t)$ (known and bounded in $[0, 1]$)

Bellman equation policy π

$$
\begin{aligned}
Q_h^\pi(s,a) &= (r_h + p_h V_{h+1}^\pi)(s,a) \\
V_h^\pi(s) &= Q_h^\pi(s, \pi_h(s)) \\
V_{H+1}^\pi(s) &= 0
\end{aligned}
$$

where $p_h f(s,a) = \sum_{s'} p_h(s'|s,a) f(s')$

Markov Decision Process (MDP)

Tabular, episodic MDP: H horizon, S states, A actions.

Learning in MDP: at episode t, step h

- state s_h^t
- action a_h^t
- next state $s_{h+1}^t \sim p_h(\cdot|s_h^t, a_h^t)$
- reward $r_h(s_h^t, a_h^t)$ (known and bounded in $[0, 1]$)

Optimal Bellman equation

$$
\begin{aligned}
Q_h^\star(s,a) &= (r_h + p_h V_{h+1})(s,a) \\
V_h^\star(s) &= \max_a Q_h^\star(s,a) \\
V_{H+1}^\star(s) &= 0
\end{aligned}
$$

where $p_h f(s,a) = \sum_{s'} p_h(s'|s,a) f(s')$

Regret after T episodes: $R^T = \sum_{t=1}^T V_1^\star(s_1) - V_1^{\pi^t}(s_1)$

Posterior Sampling

Bonus-Based Exploration

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

- $\widehat{p}_h^t$ – empirical model, b_h^t – exploration bonus;

Posterior Sampling

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

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

Posterior Sampling

Bonus-Based Exploration

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

- $\widehat{p}_h^t$ – empirical model, b_h^t – exploration bonus;
- Nearly optimal regret: $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ [Azar et al., 2017];

Posterior Sampling

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

- $\widetilde{p}_h^t \sim \rho_h^t(s,a)$ is a sample from posterior distribution.
- Hard to analyze: only Bayesian regret [Osband and Van Roy, 2017];

Posterior Sampling

Bonus-Based Exploration

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

- $\widehat{p}_h^t$ – empirical model, b_h^t – exploration bonus;
- Nearly optimal regret: $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ [Azar et al., 2017];
- Bad empirical results.

Posterior Sampling

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

- $\widetilde{p}_h^t \sim \rho_h^t(s,a)$ is a sample from posterior distribution.
- Hard to analyze: only Bayesian regret [Osband and Van Roy, 2017];
- Good empirical results!

Optimistic Posterior Sampling

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

$$\overline{V}_h^t(s) = \max_a \overline{Q}_h^t(s,a)$$

- $\widetilde{p}_h^{j,t} \sim \rho_h^t(s,a)$ are J samples from posterior distribution.

Optimistic Posterior Sampling

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

$$\overline{V}_h^t(s) = \max_a \overline{Q}_h^t(s,a)$$

- $\widetilde{p}_h^{j,t} \sim \rho_h^t(s,a)$ are J samples from posterior distribution.
- Near optimal regret: $\widetilde{\mathcal{O}}\left(\sqrt{H^3SAT}\right)$
  - *Optimism:* novel tight Gaussian anti-concentration inequality for Dirichlet weighted sum;
  - *Overestimation error*: reduction to `UCBVI`[Azar et al., 2017];

Optimistic Posterior Sampling

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

$$\overline{V}_h^t(s) = \max_a \overline{Q}_h^t(s,a)$$

- $\widetilde{p}_h^{j,t} \sim \rho_h^t(s,a)$ are J samples from posterior distribution.
- Near optimal regret: $\widetilde{\mathcal{O}}\left(\sqrt{H^3SAT}\right)$
  - *Optimism:* novel tight Gaussian anti-concentration inequality for Dirichlet weighted sum;
  - *Overestimation error*: reduction to `UCBVI`[Azar et al., 2017];
- Logarithmic number of posterior samples: $J = \mathcal{O}(\log(SATH/\delta))$;

Optimistic Posterior Sampling

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

- $\widetilde{p}_h^{j,t} \sim \rho_h^t(s,a)$ are J samples from posterior distribution.
- Near optimal regret: $\widetilde{\mathcal{O}}\left(\sqrt{H^3SAT}\right)$
  - *Optimism:* novel tight Gaussian anti-concentration inequality for Dirichlet weighted sum;
  - *Overestimation error*: reduction to `UCBVI`[Azar et al., 2017];
- Logarithmic number of posterior samples: $J = \mathcal{O}(\log(SATH/\delta))$;

All of them: open problems raised by [Agrawal and Jia, 2017].

Gaussian anti-concentration for Dirichlet weighted sums

Theorem

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* $\overline{\alpha} = \sum_{j=0}^{m} \alpha_j$. *For any* $\varepsilon > 0$, *under technical assumptions on* $\alpha_0, \overline{\alpha}$, *for "good"* $f : \{0, \ldots, m\} \to [0, b_0]$ *and* $\mu \in (\overline{p}f, b_0)$

$$\mathbb{P}_{w \sim \mathcal{D}\mathrm{ir}(\alpha)}[wf \geq \mu] \geq (1 - \varepsilon)\mathbb{P}_{g \sim \mathcal{N}(0,1)}\left[g \geq \sqrt{2\overline{\alpha}\, \mathcal{K}_{inf}(\overline{p}, \mu, f)}\right]$$

where $\mathcal{K}_{inf}(p, u, f)$ *is given by*

$$\mathcal{K}_{inf}(p, u, f) \triangleq \max_{\lambda \in [0,1]} \mathbb{E}_{X \sim p}\left[\log\left(1 - \lambda \frac{f(X) - u}{b_0 - u}\right)\right].$$

- Essential part for providing optimism for `OPSRL`;
- The rest of proof: "`OPSRL` is no worse than `UCBVI` with Bernstein bonuses" [Azar et al., 2017];

Experimental results

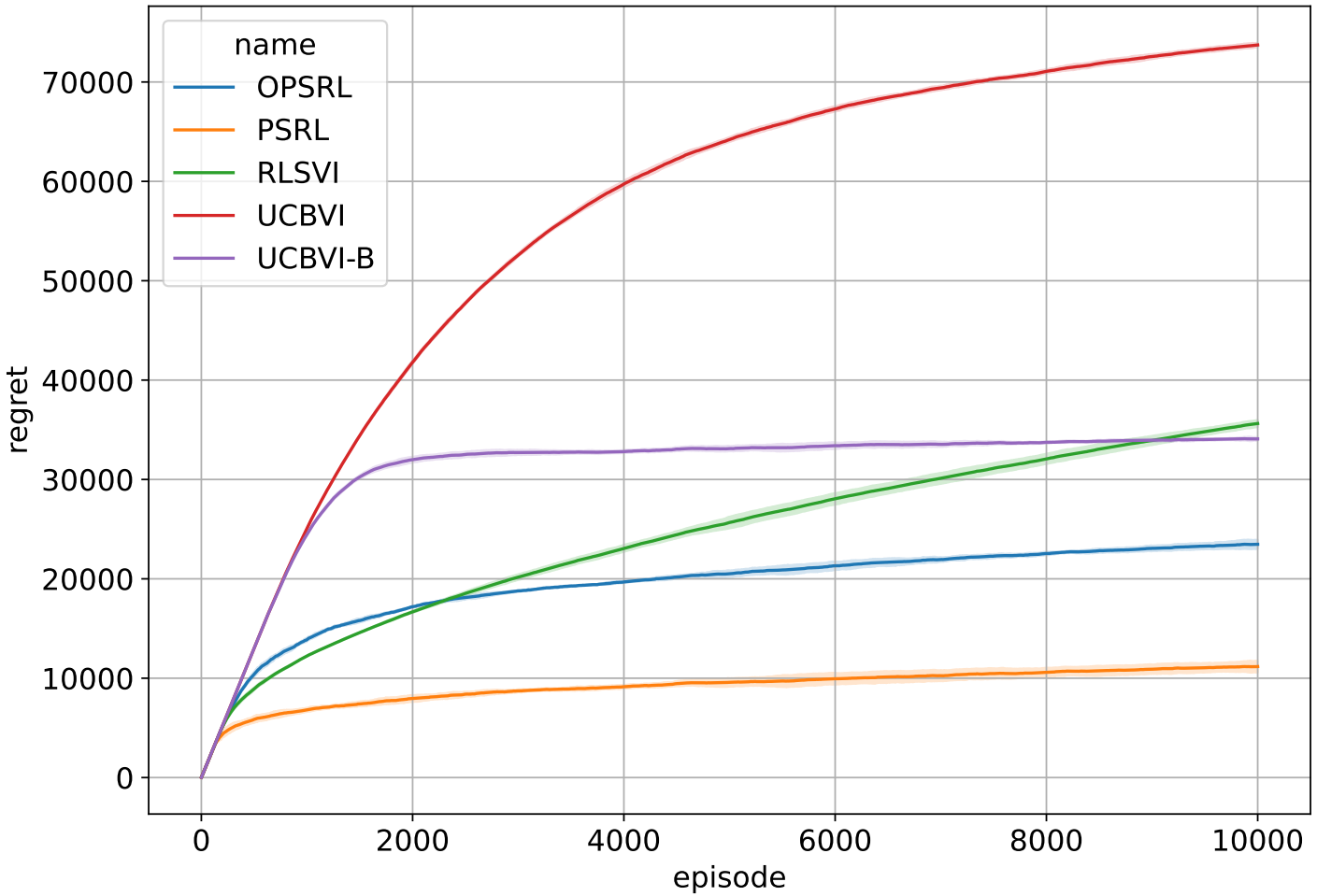


Figure: Regret of `OPSRL` and baselines on grid-world environment with 100 states and 4 action for $H = 50$ an transitions noise 0.2. We show average over 4 seeds.

Lazy version

- Computational complexity of `OPSRL`: $\widetilde{\mathcal{O}}(HS^2A)$. Can we do better?
- Lazy updates only on visited state-action pairs: $\widetilde{\mathcal{O}}(HSA)$ time.

$$\overline{Q}_h^t(s,a) \triangleq \mathbb{1}\{s = s_h^{t+1}\}\left(r_h(s,a) + \max_{j\in[J]}\{\widetilde{p}_h^{t,j}\overline{V}_{h+1}^{t-1}(s,a)\}\right) + (1 - \mathbb{1}\{s = s_h^{t+1}\})\overline{Q}_h^{t-1}(s,a)\,,$$

$$\overline{V}_h^t(s) \triangleq \min\left\{\max_{a\in\mathcal{A}} \overline{Q}_h^t(s,a), \overline{V}_h^{t-1}(s)\right\},$$

$$\pi_h^{t+1}(s) \in \arg\max_{a\in\mathcal{A}} \overline{Q}_h^t(s,a)\,,$$

- Also nearly optimal regret guarantees!

Takeaways

- `OPSRL` is an optimistic extension of posterior sampling algorithm;
- Computationally efficient and theoretically (near) optimal;
- Resolve two open questions stated by [Agrawal and Jia, 2017].

Thank you!

Bibliography I

Agrawal, S. and Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: worst-case regret bounds.
In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Azar, M. G., Osband, I., and Munos, R. (2017).
Minimax regret bounds for reinforcement learning.
In International Conference on Machine Learning.

Osband, I. and Van Roy, B. (2017).
Why is posterior sampling better than optimism for reinforcement learning?
In Precup, D. and Teh, Y. W., editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 2701–2710. PMLR.