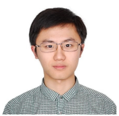
From Dirichlet to Rubin: Optimistic Exploration in RL without Bonuses

Daniil Tiapkin[1,2], Denis Belomestny[3,1], Éric Moulines[4,1], Alexey Naumov [1], Sergey Samsonov[1], Yunhao Tang [5], Michal Valko[5], Pierre Ménard[6]

[1]HSE [2]AIRI [3]Duisburg-Essen University [4]École Polytechnique [5]DeepMind [6]OvGU

What do we bring:

- New efficient Bayesian-inspired algorithm for tabular RL.

What do we bring:

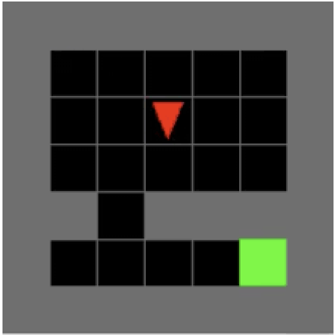
- New efficient Bayesian-inspired algorithm for tabular RL.
  - Good empirical performance.
  - Up till now no optimal Bayesian-inspired algorithm.

What do we bring:

- New efficient Bayesian-inspired algorithm for tabular RL.
  - Good empirical performance.
  - Up till now no optimal Bayesian-inspired algorithm.
- Algorithm that can be easily extended to the deep RL setting.
  - Link between our algorithm and Bayesian bootstrap.

Bridging tabular and deep RL

Tabular setting: $S \approx 100$;



Deep RL setting: $S \approx 10^{100}$;

Is there an algorithm that at the same time

- provably optimal in tabular setting?
- practically good in Deep RL setting?

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

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

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

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

Bonus-driven exploration

Basic idea: solve Bellman equation with upper approximations.

$$\overline{Q}_h^t(s,a) = r_h(s,a) + \underbrace{\overbrace{\widehat{p}_h^t}^{\text{empirical model}} \overline{V}_{h+1}^t(s,a) + \overbrace{B_h^t(s,a)}^{\text{exploration bonus}}}_{\text{upper approximation of } p_h V_{h+1}^\star(s,a)}$$

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

Bonus-driven exploration

Basic idea: solve Bellman equation with upper approximations.

$$\overline{Q}_h^t(s,a) = r_h(s,a) + \underbrace{\overbrace{\widehat{p}_h^t}^{\text{empirical model}} \overline{V}_{h+1}^t(s,a) + \overbrace{B_h^t(s,a)}^{\text{exploration bonus}}}_{\text{upper approximation of } p_h V_{h+1}^\star(s,a)}$$

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

- Near optimal in tabular setting: $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ regret.

Bonus-driven exploration

Basic idea: solve Bellman equation with upper approximations.

$$\overline{Q}_h^t(s,a) = r_h(s,a) + \underbrace{\overbrace{\widehat{p}_h^t}^{\text{empirical model}} \overline{V}_{h+1}^t(s,a) + \overbrace{B_h^t(s,a)}^{\text{exploration bonus}}}_{\text{upper approximation of } p_h V_{h+1}^\star(s,a)}$$

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

- Near optimal in tabular setting: $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ regret.
- Poor empirical performance.
- Difficult scale to deep RL.

Bayes-UCBVI: From Dirichlet...

Idea: use directly an upper quantile over posterior distribution (cf. Bayes-UCB [Kaufmann et al., 2012]).

$$\overline{Q}_h^t(s,a) = r_h(s,a) + \overbrace{\text{Quantile}_{\underbrace{p\sim\rho_h^t(s,a)}_{\text{Dirichlet distribution}}}}^{\text{quantile over posterior}}\left(p\overline{V}_{h+1}^t, \overbrace{\kappa}^{\text{quantile level}}\right)$$

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

where posterior $\rho_h^t(s,a) = \mathcal{D}\mathrm{ir}\left(n_h^t(s_1,s,a),\ldots,n_h^t(s_S,s,a), \underbrace{n_0}_{\text{pseudo transition}}\right)$

Bayes-UCBVI: From Dirichlet...

Idea: use directly an upper quantile over posterior distribution (cf. Bayes-UCB [Kaufmann et al., 2012]).

$$\overline{Q}_h^t(s,a) = r_h(s,a) + \overbrace{\underbrace{\text{Quantile}_{p \sim \rho_h^t(s,a)}}_{\text{Dirichlet distribution}}}^{\text{quantile over posterior}}(p\overline{V}_{h+1}^t, \overbrace{\kappa}^{\text{quantile level}})$$

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

where posterior $\rho_h^t(s,a) = \mathcal{D}\mathrm{ir}\left(n_h^t(s_1,s,a), \ldots, n_h^t(s_S,s,a), \underbrace{n_0}_{\text{pseudo transition}}\right)$

- Near optimal in tabular setting: $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ regret.

`Bayes-UCBVI`: From Dirichlet...

Idea: use directly an upper quantile over posterior distribution (cf. `Bayes-UCB` [Kaufmann et al., 2012]).

$$\overline{Q}_h^t(s,a) = r_h(s,a) + \overbrace{\text{Quantile}_{\underbrace{p \sim \rho_h^t(s,a)}_{\text{Dirichlet distribution}}}}^{\text{quantile over posterior}}\Big(p\overline{V}_{h+1}^t, \overbrace{\kappa}^{\text{quantile level}}\Big)$$

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

where posterior $\rho_h^t(s,a) = \mathcal{D}\mathrm{ir}\Big(n_h^t(s_1,s,a),\ldots,n_h^t(s_S,s,a), \underbrace{n_0}_{\text{pseudo transition}}\Big)$

- Near optimal in tabular setting: $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ regret.
  - *Optimism:* novel anti-concentration inequality for Dirichlet weighted sum;
  - *Estimation error:* reduction to `UCBVI` [Azar et al., 2017].

Bayes-UCBVI: From Dirichlet...

Idea: use directly an upper quantile over posterior distribution (cf. Bayes-UCB [Kaufmann et al., 2012]).

$$\overline{Q}_h^t(s,a) = r_h(s,a) + \overbrace{\text{Quantile}_{\underbrace{p \sim \rho_h^t(s,a)}_{\text{Dirichlet distribution}}}}^{\text{quantile over posterior}}(p\overline{V}_{h+1}^t, \overbrace{\kappa}^{\text{quantile level}})$$

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

where posterior $\rho_h^t(s,a) = \mathcal{D}\mathrm{ir}\left(n_h^t(s_1,s,a),\ldots,n_h^t(s_S,s,a), \underbrace{n_0}_{\text{pseudo transition}}\right)$

- Near optimal in tabular setting: $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ regret.
  - *Optimism:* novel anti-concentration inequality for Dirichlet weighted sum;
  - *Estimation error:* reduction to UCBVI [Azar et al., 2017].
- Scalable with the magic of Bayesian bootstrap!

Where do we stand: Known guarantees

Algorithm	Upper bound (non-stationary)
UCBVI [Azar et al., 2017] UCB-Advantage [Zhang et al., 2020] RLSVI [Xiong et al., 2021]	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$
PSRL [Agrawal and Jia, 2017] BootNARL [Pacchiano et al., 2021]	$\widetilde{\mathcal{O}}(H^2S\sqrt{AT})$
Bayes-UCBVI (this paper)	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$
Lower bound [Jin et al., 2018, Domingues et al., 2021]	$\Omega(\sqrt{H^3SAT})$

Table: Regret upper bound for episodic, non-stationary, tabular MDPs.
Green: scalable, Yellow: scalable under simplifications, Red: not scalable.

Bayes-UCBVI: ...to Rubin - Scaling up!

Given: dataset $y^1, \dots, y^n \sim \mathcal{P}$.

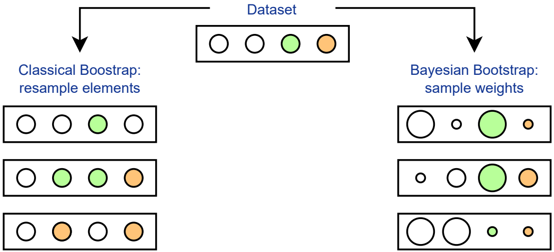
Goal: confidence interval for $\mathbb{E}_{y\sim\mathcal{P}}[y]$.

Classical Bootstrap [Efron, 1979]

- Resample $y^{1,b}, \dots, y^{n,b}$ B times;
- Compute mean estimates as $\bar{y}^b = \frac{1}{n}\sum_{i=1}^n y^{i,b}$ for all b;
- Compute quantile over $\bar{y}^b$.

Bayesian Bootstrap [Rubin, 1981]

- Sample $w^b \sim \mathcal{D}\mathrm{ir}(\underbrace{1, \dots, 1}_{n})$ B times;
- Compute mean estimates as $\bar{y}^b = \sum_{i=1}^n w^{b,i} y^i$ for all b;
- Compute quantile over $\bar{y}^b$.



Efficient implementation

- targets for Q-function estimation $y^n = r_h(s,a) + \overline{V}^t_{h+1}(s^n_{h+1})$ for visits $n = 1, \ldots, n^t$.
- prior targets $y^n = r_h(s,a) + \overline{V}^t_h(s_0)$ for prior visits $n = -n^0 + 1, \ldots, 0$.

By aggregation property and sample quantile approximation

$$\begin{aligned}
\overline{Q}^t_h(s,a) &= r_h(s,a) + \mathrm{Quantile}_{p \sim \rho^t_h(s,a)}\left(p\overline{V}^t_{h+1}(s,a), \kappa\right) \\
&= \mathrm{Quantile}_{w \sim \mathcal{D}\mathrm{ir}(\underbrace{1,\ldots,1}_{n^t+n^0})}\left(\sum_{n=-n^0+1}^{n^t} w^n y^n, \kappa\right) \\
&\approx \underbrace{\mathrm{Quantile}_{b \sim \mathcal{U}\mathrm{nif}([1,B])}\left(\sum_{n=-n^0+1}^{n^t} \overbrace{w^{n,b}}^{\text{samples from Dirichlet}} y^n, \kappa\right)}_{\text{upper confidence bound by Bayesian bootstrap}}.
\end{aligned}$$

Deep RL extension: `Bayes-UCBDQN`

Recall $w^{n,b} \sim \mathcal{D}\text{ir}(\underbrace{1, \ldots, 1}_{n^t + n^0})$

$$\overline{Q}_h^t(s, a) \approx \text{Quantile}_{b \sim \mathcal{U}\text{nif}([1,B])}\left(\bar{y}^b, \kappa\right)$$

where Bayesian bootstrap sample $\bar{y}^b = \sum_{n=-n^0+1}^{n^t} w^{n,b} y^n$

Uniform Dirichlet distribution $=$ exponential with normalization

$$\bar{y}^b = \arg\min_x \sum_{n=-n^0+1}^{n^t} z^{n,b} (x - y^n)^2$$

where $z^{n,b} \sim \mathcal{E}(1)$ i.i.d..

$\rightarrow$ Weighted regression of the targets!

Experimental results

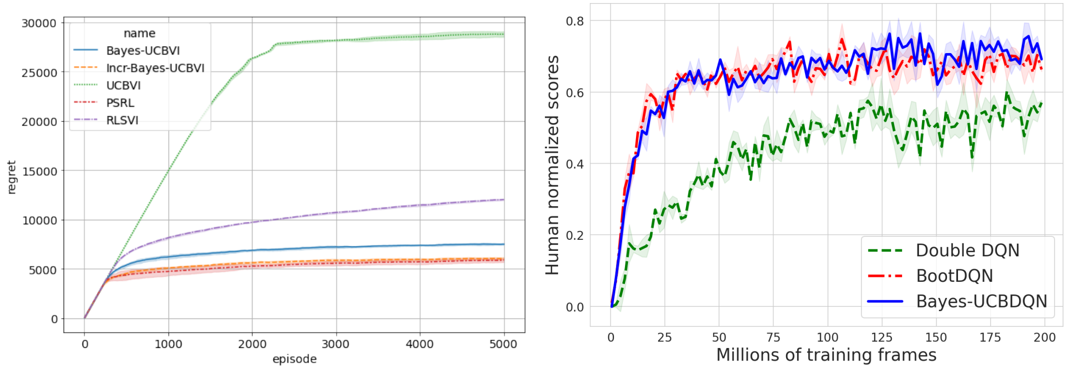


Figure: Left: Regret of `Bayes-UCBVI` and `Incr-Bayes-UCBVI` compared to baselines on grid-world with 5 rooms of size 5×5. Right: deep RL algorithms with median human normalized scores across Atari-57 games.

Takeaways

- `Bayes-UCBVI` $\rightsquigarrow$ near optimal optimistic algorithm **without bonuses**.
- New *anti-concentration* inequality for a Dirichlet weighted sum.
- `Bayes-UCBVI` scales to deep RL.

Bibliography

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.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. (2021).
Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited.
In Algorithmic Learning Theory.

Efron, B. (1979).
Bootstrap methods: Another look at the jackknife.
The Annals of Statistics, 7(1):1 – 26.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018).
Is Q-learning provably efficient?
In Neural Information Processing Systems.