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 University, [2]Artificial Intelligence Research Institute, [3]Duisburg-Essen University, [4]École Polytechnique, [5]DeepMind, [6]Otto von Guericke University

Overview

- `Bayes-UCBVI` is Bayes-inspired algorithm with nearly minimax regret bound for large enough T;
- Novel anti-concentration inequality for weighted sums of Dirichlet random vector;
- Extension of exploration mechanism of `Bayes-UCBVI` in Deep RL setting;

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

`Bayes-UCBVI`

- Bonus-based exploration (theoretically near optimal, poor empirical performance, not scalable)

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

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

- `Bayes-UCBVI` exploration (theoretically near optimal, great empirical performance, scalable)

$$\overline{Q}_h^t(s,a) = r_h(s,a) + \mathrm{Quantile}_{p\sim\rho_h^t(s,a)}(p\overline{V}_{h+1}^t, \kappa), \qquad \overline{V}_h^t(s) = \max_a \overline{Q}_h^t(s,a),$$

where $\rho_h^t(s,a) = \mathcal{D}\mathrm{ir}\left(n_0, n_h^t(s_1,s,a), \ldots, n_h^t(s_S,s,a)\right)$ is a posterior distribution of transition probabilities.

Optimistic prior for `Bayes-UCBVI`:

- 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.

Regret bounds

Algorithm	Upper bound (non-stationary)
`UCBVI` [Azar et al., 2017] `UCB-Advantage` [Zhang et al., 2020] `RLSVI` [Xiong et al., 2021]	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$
`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]	$\Omega(\sqrt{H^3SAT})$

Green: scalable, Orange: scalable under simplifications, Red: not scalable.

Non-tabular extension: `Bayes-UCBDQN`

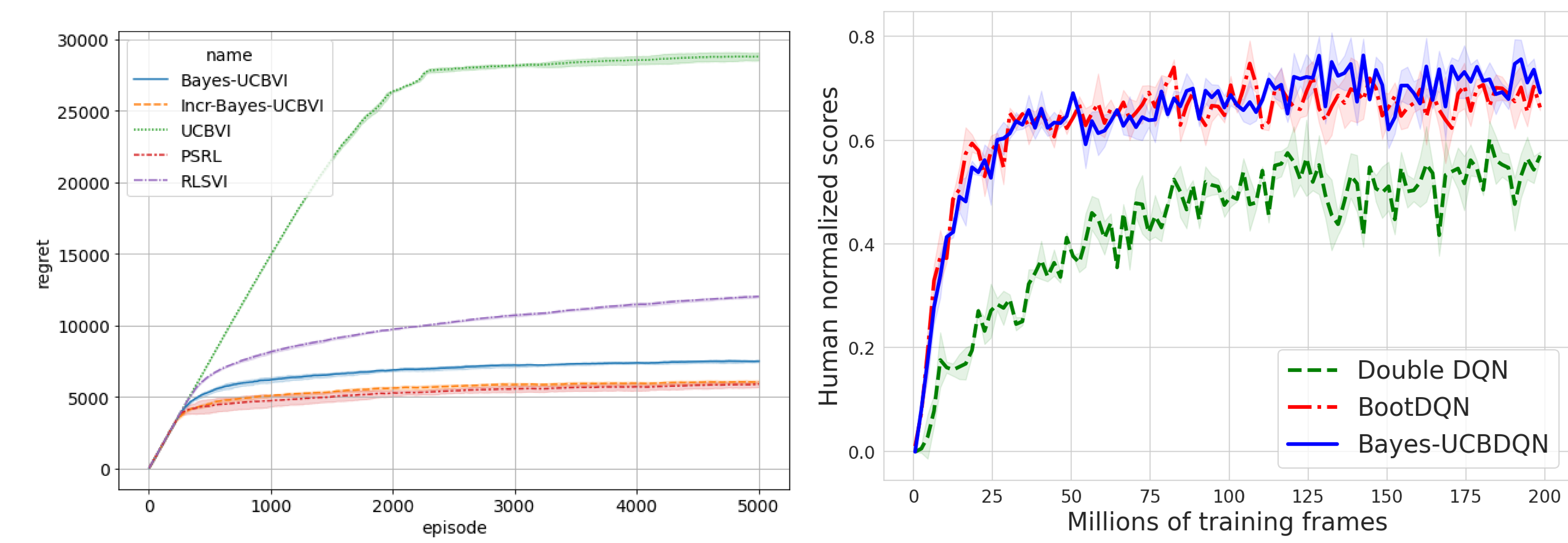

Figure 1:Left: Regret of `Bayes-UCBVI` and `Incr-Bayes-UCBVI` compared to baselines. Environment: grid-world with 5 rooms of size 5×5 ($S = 129, A = 4$).
Right: Evaluating deep RL algorithms with median human normalized scores across Atari-57 games.

- targets for Q-function estimation $y^n = r_h(s,a) + \overline{V}_{h+1}^t(s_{h+1}^n)$, $n = 1, \ldots, n_h^t(s,a)$;
- targets from prior transitions $y^n = r_h(s,a) + \overline{V}_h^t(s_0)$, $n = -n_0+1, \ldots, 0$.

Reformulate and approximate `Bayes-UCBVI` using B Bayesian bootstrap samples

$$\overline{Q}_h^t(s,a) = \mathrm{Quantile}_{w\sim\mathcal{D}\mathrm{ir}(\underbrace{1,\ldots,1}_{n_h^t(s,a)+n_0})}\left(\sum_{n=-n_0+1}^{n_h^t(s,a)} w^n y^n, \kappa\right)$$

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

where $\overline{Q}_h^{t,b}(s,a) = \sum_{n=-n_0+1}^{n_h^t(s,a)} w^{n,b} y^n$ Bayesian bootstrap sample with $w^{n,b} \sim \mathcal{D}\mathrm{ir}(\underbrace{1,\ldots,1}_{n_h^t(s,a)+n_0})$.

Since $w^{n,b}$ is a vector of normalized $z^{n,b} \sim \mathcal{E}(1)$ random variables

$$\overline{Q}_h^{t,b}(s,a) = \operatorname*{argmin}_x \sum_{n=-n_0+1}^{n_h^t(s,a)} z^{n,b}(x - y^n)^2 .$$

Anti-concentration inequality for Dirichlet distribution

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

$$\overline{\alpha}^{-3/2} \exp\left(-\overline{\alpha}\,\mathcal{K}_{\inf}(\overline{p}, \mu, f)\right) \le \mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \ge \mu] \le \exp\left(-\overline{\alpha}\,\mathcal{K}_{\inf}(\overline{p}, \mu, f)\right),$$

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

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

- Lower bound is an essential part for optimism;
- Upper bound is important for the reduction to `UCBVI`.