Posterior Sampling and Bayesian Bootstrap in Reinforcement Learning

Daniil Tiapkin[1], Alexey Naumov [1],

[1]HDI Lab, HSE University

Exploration in RL

November 1, 2022

Exploration in RL

Figure: Montezuma's Revenge from Atari. Solved by RL only in 2018

Offline RL as "anti-exploration"

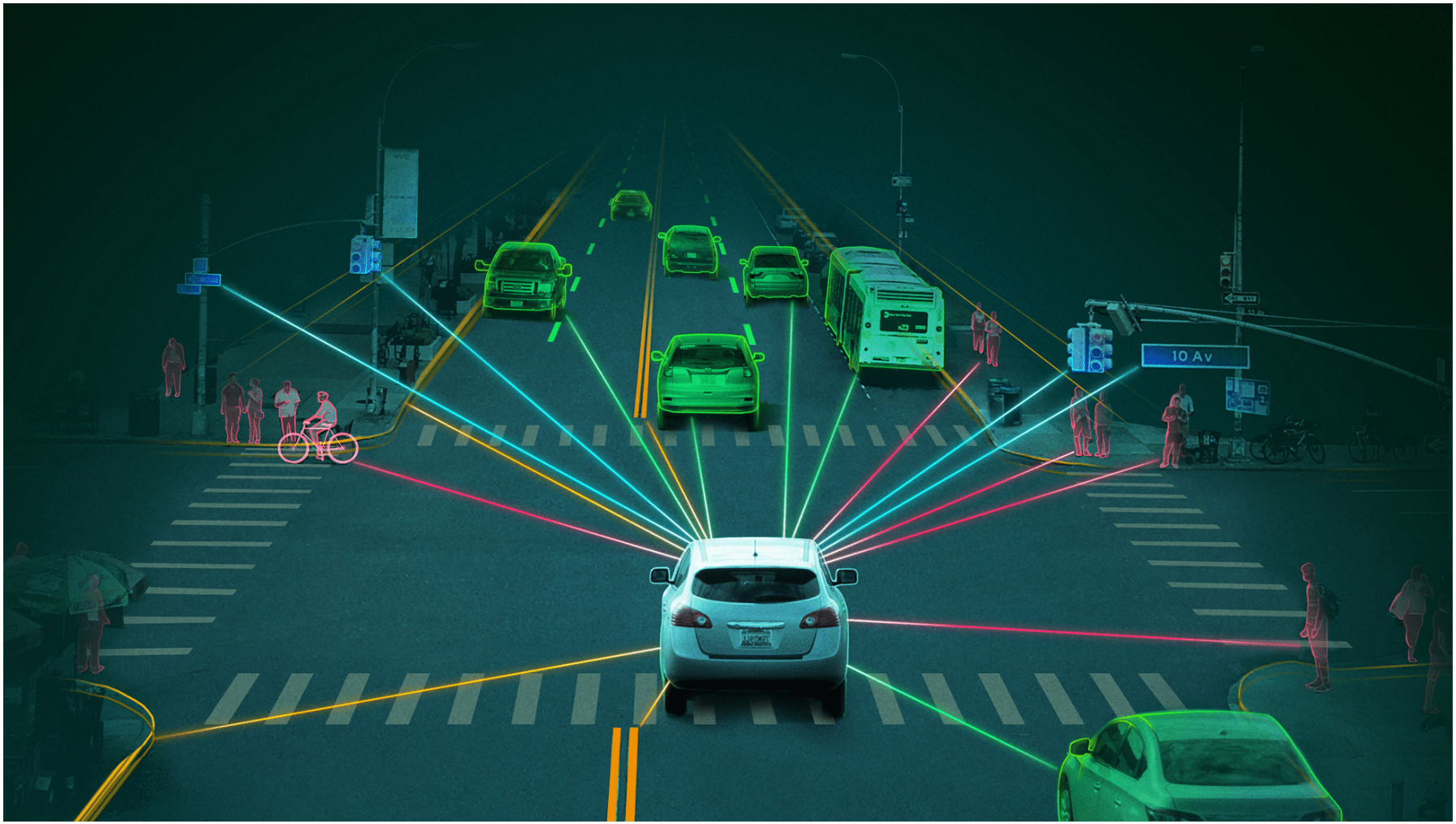


Figure: Autonomous driving potentially could be resolved as offline RL problem.

Introduction

Markov Decision Process (MDP)

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

Learning in MDP: at episode t, step h

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

Goal: find *a policy* $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes *a value function*

$$V_h^\pi(s) = \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \mid s_h = s \right].$$

Bellman Equations

Action-value function for policy π

$$Q_h^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'})) \mid s_h = s, a_h = a\right].$$

Bellman equations for policy π

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

where $p_h f(s,a) = \sum_{s'} p_h(s'|s,a) f(s') = \mathbb{E}_{s' \sim p_h(\cdot|s,a)}[f(s')]$.

Optimal Bellman Equations

Optimal policy $\pi^\star$ maximizes $V_h^\pi(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$.

Optimal value and action-value functions

$$V_h^\star(s) = V_h^{\pi^\star}(s), \quad Q_h^\star(s,a) = Q_h^{\pi^\star}(s,a).$$

Optimal Bellman equations

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

where $p_h f(s,a) = \sum_{s'} p_h(s'|s,a) f(s') = \mathbb{E}_{s' \sim p_h(\cdot|s,a)}[f(s')]$.

Choose $\pi_h^\star(s) = \arg\max_a Q_h^\star(s,a)$.

Least Squares Value Iteration (LSVI)

- Rewrite optimal Bellman equations as solution to linear regression ($V^\star_{h+1}$ is fixed)

$$Q^\star_h(s,a) = \arg\min_x \mathbb{E}_{s' \sim p_h(s,a)} \left[\left(x - [r_h(s,a) + V^\star_{h+1}(s')] \right)^2 \right]$$

Least Squares Value Iteration (LSVI)

- Rewrite optimal Bellman equations as solution to linear regression ($V^\star_{h+1}$ is fixed)

$$Q^\star_h(s,a) = \arg\min_x \mathbb{E}_{s' \sim p_h(s,a)} \left[\left(x - [r_h(s,a) + V^\star_{h+1}(s')] \right)^2 \right]$$

- Monte-Carlo approximation: for replay buffer $\mathcal{D} = \{(s^t_h, a^t_h, r^t_h)\}_{t \in [T], h \in [H]}$ define $y_{t,h} = r_h(s^t_h, a^t_h) + \widehat{V}_{h+1}(s^t_{h+1})$

$$\hat{Q}_h = \arg\min_{Q \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}} \sum_{t=1}^{T} \left(Q(s^t_h, a^t_h) - y_{t,h} \right)^2, \quad h = H, H-1, \ldots, 1$$

Least Squares Value Iteration (LSVI)

- Rewrite optimal Bellman equations as solution to linear regression ($V^\star_{h+1}$ is fixed)

$$Q^\star_h(s,a) = \arg\min_x \mathbb{E}_{s'\sim p_h(s,a)}\Big[\big(x - [r_h(s,a) + V^\star_{h+1}(s')]\big)^2\Big]$$

- Monte-Carlo approximation: for replay buffer $\mathcal{D} = \{(s_h^t, a_h^t, r_h^t)\}_{t\in[T],h\in[H]}$ define $y_{t,h} = r_h(s_h^t, a_h^t) + \widehat{V}_{h+1}(s_{h+1}^t)$

$$\hat{Q}_h = \arg\min_{Q\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}} \sum_{t=1}^{T}\big(Q(s_h^t, a_h^t) - y_{t,h}\big)^2, \quad h = H, H-1, \ldots, 1$$

- DQN: Q is neural network + discounted setting.

Online Reinforcement Learning Algorithm

Online algorithm: outputs a refined policy π^t after each episode $t = 1, \ldots, T$.

Goal: *regret* minimization

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

Lower bound: $\mathfrak{R}^T = \Omega(\sqrt{H^3SAT})$ [Domingues et al., 2021].

Exploration

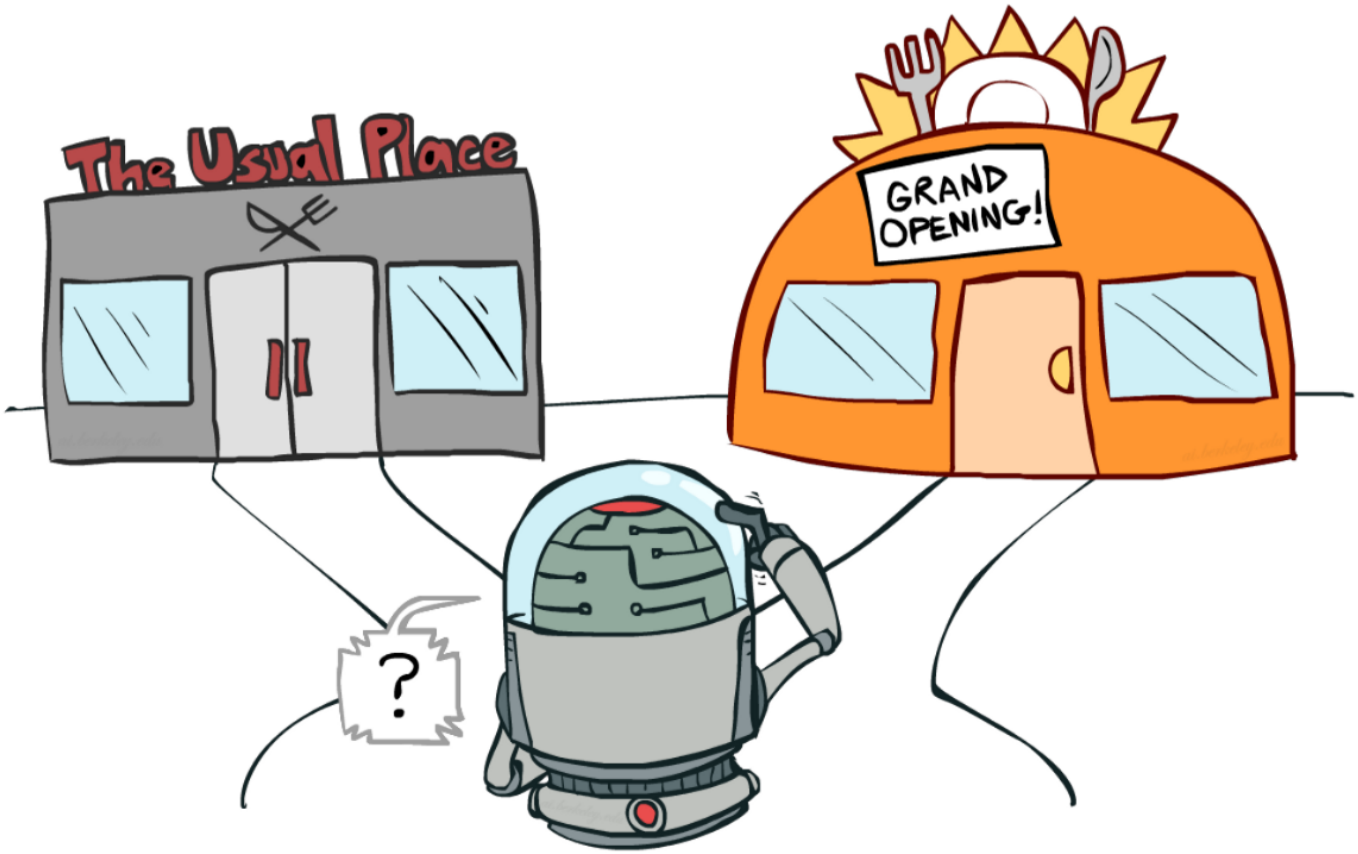


Figure: Image source: UC Berkeley Intro to AI course

ε-greedy exploration

- Classical way: ε-greedy exploration.

$$\pi_h(s) = \begin{cases} \text{uniform over } \mathcal{A} & \text{with probability } \varepsilon; \\ \max_{a \in \mathcal{A}} Q_h(s,a) & \text{otherwise.} \end{cases}$$

- Under the best choice of ε: $\mathfrak{R}^T = \mathcal{O}(T^{2/3})$;
- Moreover, there is a hard environment such that $\mathfrak{R}^T \geq 2^{S-1} - 1$

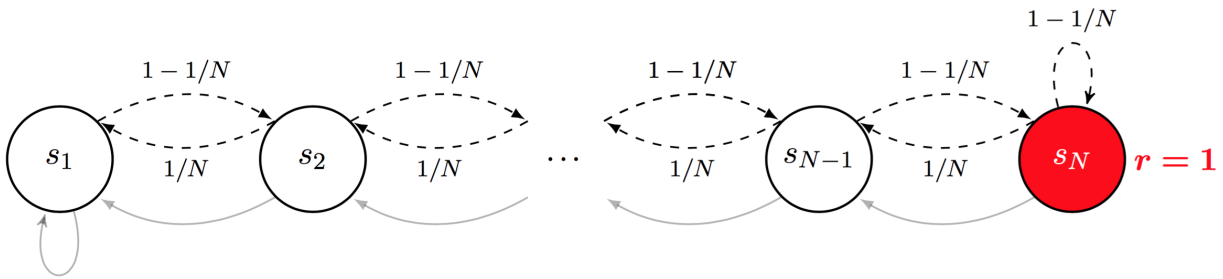


Figure: Hard instance for ε-greedy exploration

Result: sub-optimal exploration!

Bonus-driven exploration [Azar et al., 2017]

Basic idea: solve Bellman equation with upper approximations.

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

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

Bonus-driven exploration [Azar et al., 2017]

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 [Azar et al., 2017]

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.

Upper Confidence Bound Value Iteration UCBVI[Azar et al., 2017]

Recall the setup

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

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

Let $L = \log(5SAHT/\delta)$.

- UCBVI with Hoeffding bonuses

$$B_h^t(s,a) = \frac{7HL}{\sqrt{n_h^t(s,a)}}.$$

- UCBVI with Bernstein bonuses

$$B_h^t(s,a) = \sqrt{\frac{8L\mathrm{Var}_{s'\sim\widehat{p}_h^t(\cdot|s,a)}[\overline{V}_{h+1}^t(s')]}{n_h^t(s,a)}} + \frac{14HL}{3n_h^t(s,a)} + \text{correction}.$$

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

Randomized Exploration: RLSVI [Osband et al., 2016b]

Basic idea: solve *noisy* Bellman equation to increase robustness.

$$\widehat{Q}_h^t(s,a) = r_h(s,a) + \underbrace{\overbrace{\widehat{p}_h^t}^{\text{empirical model}} \widehat{V}_{h+1}^t(s,a) + \overbrace{g_h^t(s,a)}^{\text{Centered Gaussian noise}}}_{\text{noisy approximation of } p_h V_{h+1}^\star(s,a)}$$

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

Randomized Exploration: RLSVI [Osband et al., 2016b]

Basic idea: solve *noisy* Bellman equation to increase robustness.

$$\widehat{Q}_h^t(s,a) = r_h(s,a) + \underbrace{\overbrace{\widehat{p}_h^t}^{\text{empirical model}} \widehat{V}_{h+1}^t(s,a) + \overbrace{g_h^t(s,a)}^{\text{Centered Gaussian noise}}}_{\text{noisy approximation of } p_h V_{h+1}^\star(s,a)}$$

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

- Near optimal in tabular setting: $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ regret [Xiong et al., 2021].

Randomized Exploration: RLSVI [Osband et al., 2016b]

Basic idea: solve *noisy* Bellman equation to increase robustness.

$$\widehat{Q}_h^t(s,a) = r_h(s,a) + \underbrace{\overbrace{\widehat{p}_h^t}^{\text{empirical model}} \widehat{V}_{h+1}^t(s,a) + \overbrace{g_h^t(s,a)}^{\text{Centered Gaussian noise}}}_{\text{noisy approximation of } p_h V_{h+1}^\star(s,a)}$$

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

- Near optimal in tabular setting: $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ regret [Xiong et al., 2021].
- Fair empirical performance.
- Easy scale to deep RL with simplified noise shape.

General Randomized Least-Squares Value Iteration

Extension idea: move noise to rewards and use least-squares value iteration.

$$\widehat{Q}_h^t(s,a) = r_h(s,a) + g_h^t(s,a) + \widehat{p}_h^t \overline{V}_h^t(s,a) \Rightarrow$$

$$\widehat{Q}_h^T = \arg\min_Q \sum_{t=1}^{T} \left(Q_h(s_h^t, a_h^t) - \hat{r}_h^t - \widehat{V}_{h+1}^T(s_{h+1}^t) \right)^2,$$

where $\hat{r}_h^t \sim \mathcal{N}(r_h^t, \sigma^2)$ with σ^2 is a hyperparameter.

`RLSVI` **takeaway:** learn DQN with noisy rewards.
But: function approximation itself make many noise [Osband et al., 2019].

Posterior Sampling

Optimal Bellman Equations

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

- p_h - unknown!

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.

- Very hard to analyze: no regret bound for `PSRL` available, only for its optimistic extension.
- Very good in practice!
- Scalable to Deep RL: Bootstrap DQN [Osband et al., 2016a]

Bayes-UCBVI: From Dirichlet...

Idea: use directly an upper quantile over posterior distribution

(cf. Bayes-UCB [Kaufmann et al., 2012]).

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

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

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

Bayes-UCBVI: From Dirichlet...

Idea: use directly an upper quantile over posterior distribution

(cf. Bayes-UCB [Kaufmann et al., 2012]).

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

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

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

- 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]).

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

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

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

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

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

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

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

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

Known guarantees

Algorithm	**Upper bound** (non-stationary)
`UCBVI` [Azar et al., 2017] `UCB-Advantage` [Zhang et al., 2020] `RLSVI` [Xiong et al., 2021]	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$
`OPSRL` [Agrawal and Jia, 2017] `BootNARL` [Pacchiano et al., 2021]	$\widetilde{\mathcal{O}}(H^2S\sqrt{AT})$
`Bayes-UCBVI`	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$
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 $y^1, \dots, y^n \sim \mathcal{P}$.

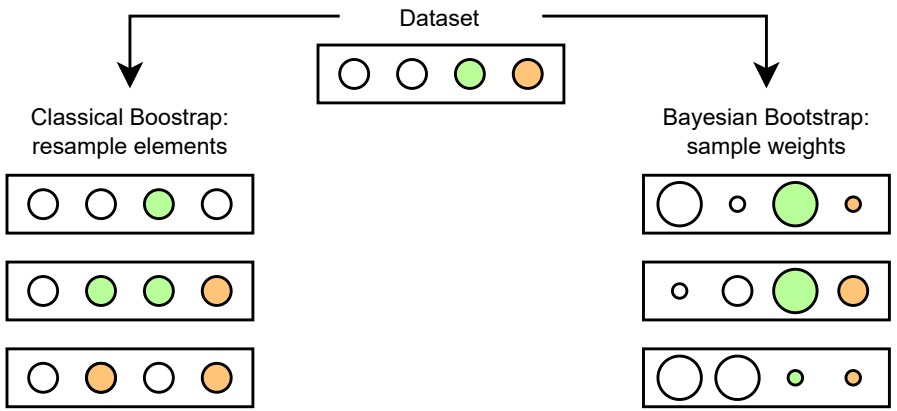
Goal: confidence interval for $\mathbb{E}_{y \sim \mathcal{P}}[y]$.

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 $w^b \sim \mathcal{D}\text{ir}(\underbrace{1, \dots, 1}_{n})$ 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 $n = 1, \dots, n^t$.
- prior targets $y^n = r_h(s,a) + \overline{V}^t_{h+1}(s_0)$ for prior visits $n = -n^0 + 1, \dots, 0$.

By aggregation property and sample quantile approximation

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

Deep RL extension: Bayes-UCBDQN

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

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

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

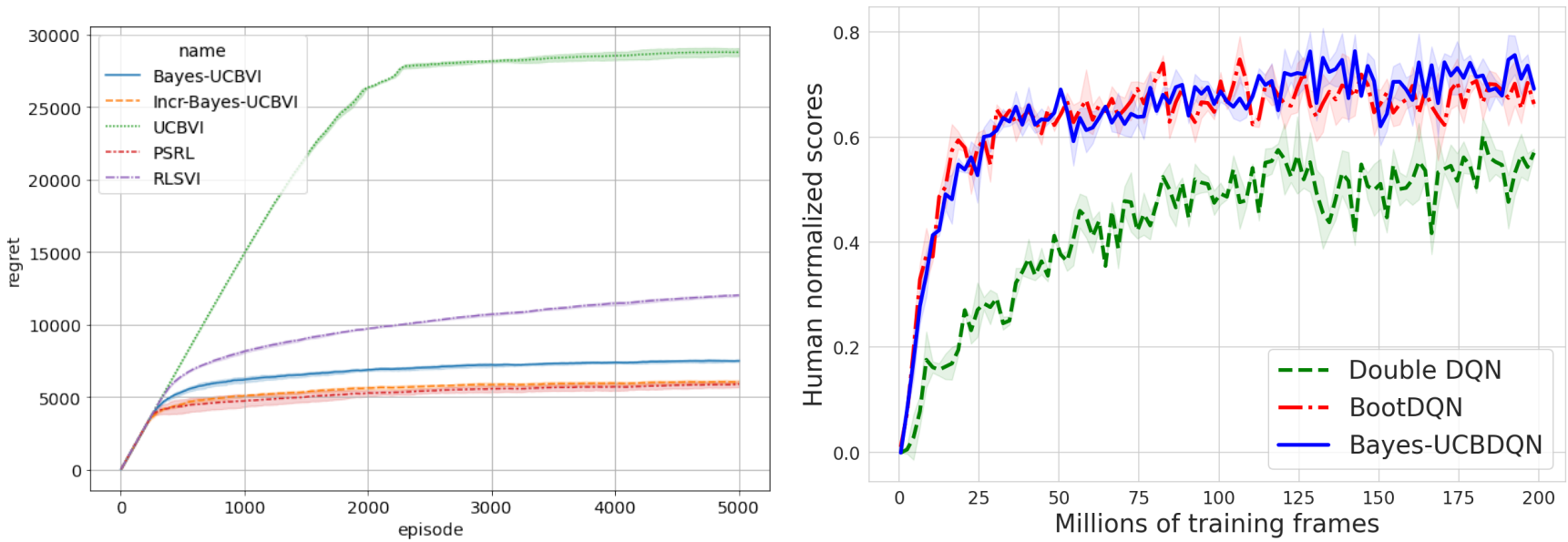


Figure: Left: Regret of `Bayes-UCBVI` and `Incr-Bayes-UCBVI` compared to baselines on grid-world with 5 rooms of size 5×5. Right: deep RL algorithms with median human normalized scores across Atari-57 games.

Takeaways

- *Optimism* (`UCBVI`) suggests non-scalable but theoretically optimal solution to exploration problem;
- *Randomization* (`RLSVI`, `PSRL`) suggests scalable but not always optimal solution;
- *Randomized optimism* (`Bayes-UCBVI`) takes the best from both worlds.

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.

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.

Kaufmann, E., Cappe, O., and Garivier, A. (2012).
On bayesian upper confidence bounds for bandit problems.
In Lawrence, N. D. and Girolami, M., editors, Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, volume 22 of Proceedings of Machine Learning Research, pages 592–600, La Palma, Canary Islands. PMLR.

Bibliography II


Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. (2016a).
Deep exploration via bootstrapped dqn.
In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 29. Curran Associates, Inc.


Osband, I., Roy, B. V., Russo, D. J., and Wen, Z. (2019).
Deep exploration via randomized value functions.
Journal of Machine Learning Research, 20(124):1–62.


Osband, I., Roy, B. V., and Wen, Z. (2016b).
Generalization and exploration via randomized value functions.
In Balcan, M. F. and Weinberger, K. Q., editors, Proceedings of The 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research, pages 2377–2386, New York, New York, USA. PMLR.


Pacchiano, A., Ball, P., Parker-Holder, J., Choromanski, K., and Roberts, S. (2021).
Towards tractable optimism in model-based reinforcement learning.
In de Campos, C. and Maathuis, M. H., editors, Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, volume 161 of Proceedings of Machine Learning Research, pages 1413–1423. PMLR.


Rubin, D. B. (1981).
The bayesian bootstrap.
The annals of statistics, pages 130–134.


Xiong, Z., Shen, R., Cui, Q., and Du, S. S. (2021).
Near-optimal randomized exploration for tabular mdp.

Bibliography III

Zhang, Z., Zhou, Y., and Ji, X. (2020).
Almost optimal model-free reinforcement learning via reference-advantage decomposition.
arXiv preprint arXiv:2004.10019.