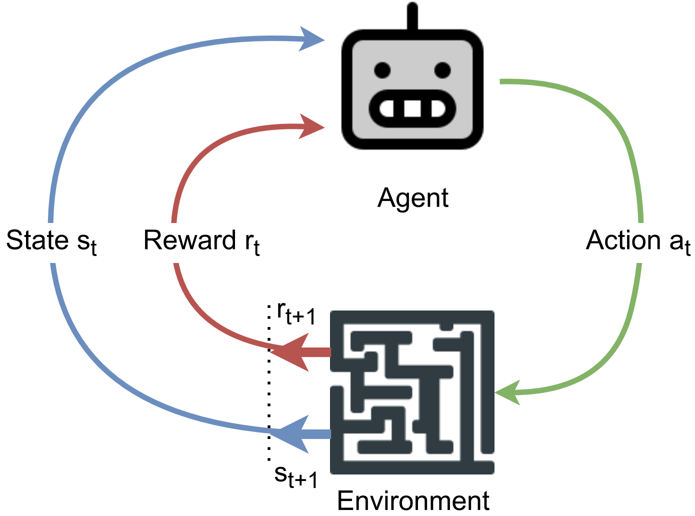
Math of Reinforcement Learning: Bayesian Approach

Daniil Tiapkin

HDI Lab, HSE University

Reinforcement Learning



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 $r_h(s_h^t, a_h^t)$ - known.

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

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

Examples

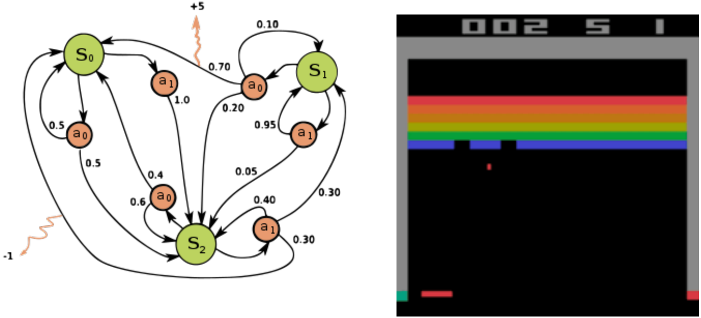


Figure: Left: MDP with $S = 3$, $A = 2$.
Right: Atari Breakout with $S = 256^{84\cdot 84} \approx 10^{17000}$, $A = 4$.

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

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 $p_h f(s,a) = \sum_{s'} p_h(s'|s,a) f(s')$. Then $\pi_h^\star(s) = \arg\max_a Q_h^\star(s,a)$.

Online Reinforcement Learning Algorithm

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

Goal: *regret* minimization

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

Good algorithm: sublinear regret $\mathfrak{R}^T = o(T)$.

Optimal algorithm: $\mathfrak{R}^T = \mathcal{O}(\sqrt{H^3SAT})$ (matches the lower bound).

Exploration-Exploitation Dilemma

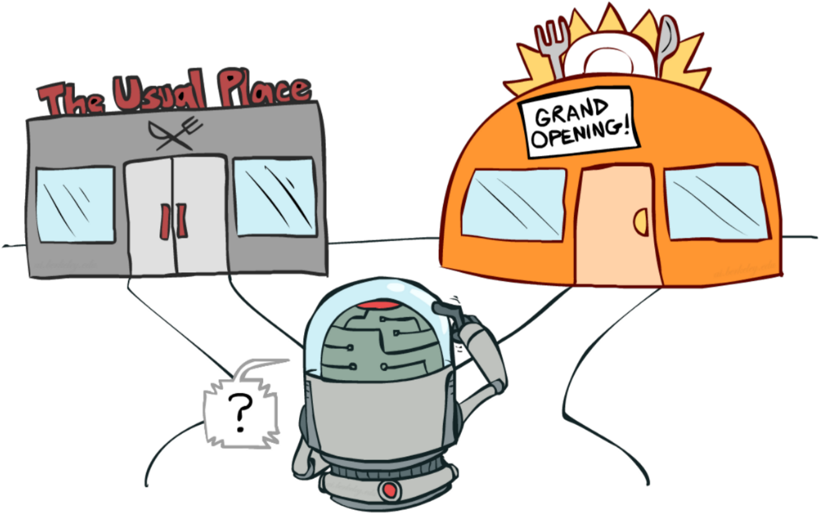


Figure: Image source: UC Berkeley Intro to AI course

Optimism in the Face of Uncertainty

Optimal Bellman Equations

$$Q^\star_h(s,a) = [r_h + p_h V^\star_{h+1}](s,a)$$
$$V^\star_h(s) = \max_a Q^\star_h(s,a)$$

- p_h - unknown!

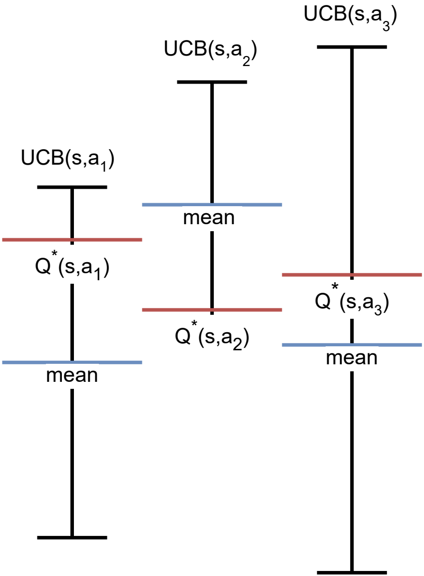
Upper confidence bound

$$\overline{Q}^t_h(s,a) = [r_h + \widehat{p}^t_h \overline{V}^t_{h+1} + B^t_h](s,a)$$
$$\overline{V}^t_h(s) = \max_a \overline{Q}^t_h(s,a)$$

- $\widehat{p}^t_h$ - empirical model (mean over transitions);
- B^t_h - exploration bonus.

The most important: $\overline{Q}^t_h(s,a) \geq Q^\star_h(s,a)$ with high probability.

Optimism in the Face of Uncertainty: visualization



How to Choose Bonuses: Hoeffding and Bernstein inequalities

Argument: bounded random variables concentrates near mean.
Given: $X_1, \dots, X_n$ i.i.d. random variables, $|X_i| < b$ a.s., $\mathbb{E}[X_i] = 0$.

Theorem (Hoeffding inequality)

With probability at least $1 - \delta$ *the following holds*

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i \right| \le \sqrt{\frac{2b^2 \log(2/\delta)}{n}}.$$

Theorem (Bernstein inequality)

With probability at least $1 - \delta$ *the following holds*

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i \right| \le \sqrt{\frac{2\mathrm{Var}[X_1] \log(2/\delta)}{n}} + \frac{2b \log(2/\delta)}{3n}.$$

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).

UCBVI with Hoeffding bonuses: optimism proof

Lemma

For all $s, a, h, t \in \mathcal{S} \times \mathcal{A} \times [H] \times [T]$ *it holds with high probability*

$$\overline{Q}_h^t(s,a) \geq Q_h^\star(s,a), \qquad \overline{V}_h^t(s) \geq V_h^\star(s).$$

UCBVI with Hoeffding bonuses: optimism proof

Lemma

For all $s, a, h, t \in \mathcal{S} \times \mathcal{A} \times [H] \times [T]$ *it holds with high probability*

$$\overline{Q}_h^t(s,a) \geq Q_h^\star(s,a), \qquad \overline{V}_h^t(s) \geq V_h^\star(s).$$

- First, by Hoeffding bound and union bound for all $s, a, h, t \in \mathcal{S} \times \mathcal{A} \times [H] \times [T]$

$$B_h^t(s,a) \geq \widehat{p}_h^t V_{h+1}^\star(s,a) - p_h V_{h+1}^\star(s,a) \geq -B_h^t(s,a)$$

$\texttt{UCBVI}$ with Hoeffding bonuses: optimism proof

Lemma

For all $s, a, h, t \in \mathcal{S} \times \mathcal{A} \times [H] \times [T]$ it holds with high probability

$$\overline{Q}_h^t(s,a) \geq Q_h^\star(s,a), \qquad \overline{V}_h^t(s) \geq V_h^\star(s).$$

- First, by Hoeffding bound and union bound for all $s, a, h, t \in \mathcal{S} \times \mathcal{A} \times [H] \times [T]$

$$B_h^t(s,a) \geq \widehat{p}_h^t V_{h+1}^\star(s,a) - p_h V_{h+1}^\star(s,a) \geq -B_h^t(s,a)$$

- Next use backward induction over $h = H+1, \ldots, 1$

$$\begin{aligned}\overline{Q}_h^t(s,a) - Q_h^\star(s,a) &= \widehat{p}_h^t \overline{V}_{h+1}^t(s,a) + B_h^t(s,a) - p_h V_{h+1}^\star(s,a) \\ &\geq \widehat{p}_h^t V_{h+1}^\star(s,a) + B_h^t(s,a) - p_h V_{h+1}^\star(s,a) \geq 0.\end{aligned}$$

and

$$\overline{V}_h^t(s) \geq \overline{Q}_h^t(s, \pi^\star(s)) \geq Q_h^\star(s, \pi^\star(s)) = V_h^\star(s).$$

Scalability issues

Example: Go, $S \approx 10^{172}$ possible states.

Figure: Image source: Wikipedia

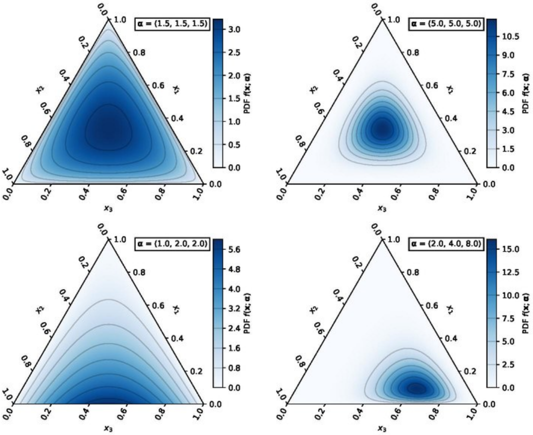
Bonus-based approach cannot be scaled: they required counters for all states.

Entering the Bayesian domain: posterior for transitions

- transitions $p_h(\cdot|s, a) \iff$ multinomial $\mathrm{Mult}(p_h(s'|s, a)_{s' \in \mathcal{S}})$;

Entering the Bayesian domain: posterior for transitions

- transitions $p_h(\cdot|s,a) \iff$ multinomial $\mathrm{Mult}(p_h(s'|s,a)_{s'\in\mathcal{S}})$;
- Conjugate prior for multinomial is Dirichlet distribution: if prior $\rho_h^0(s,a)$ is $\mathcal{D}\mathrm{ir}(\{\overline{n}_h^0(s'|s,a)\}_{s'\in\mathcal{S}})$, then posterior $\rho_h^t(s,a)$ is $\mathcal{D}\mathrm{ir}(\{\overline{n}_h^0(s'|s,a)+n_h^t(s'|s,a)\}_{s'\in\mathcal{S}})$.



Preliminaries: properties of Dirichlet distribution

The Dirichlet distribution $\mathcal{D}\mathrm{ir}(\alpha)$ for $\alpha = (\alpha_0, \ldots, \alpha_m) \in \mathbb{R}^m_{>0}$ is a distribution over m-dimensional simplex $\Delta_m = \{x \in \mathbb{R}^m \mid \sum_{i=1}^m x_i \leq 1\}$

$$p(x_1, \ldots, x_m) = \frac{1}{B(\alpha)}(1 - \sum_{i=1}^m x_i)^{\alpha_0 - 1} \prod_{i=1}^m x_i^{\alpha_i - 1},$$

where $B(\alpha)$ is a multivariate beta-function.

Preliminaries: properties of Dirichlet distribution

The Dirichlet distribution $\mathcal{D}\text{ir}(\alpha)$ for $\alpha = (\alpha_0, \ldots, \alpha_m) \in \mathbb{R}^m_{>0}$ is a distribution over m-dimensional simplex $\Delta_m = \{x \in \mathbb{R}^m \mid \sum_{i=1}^m x_i \leq 1\}$

$$p(x_1, \ldots, x_m) = \frac{1}{B(\alpha)}(1 - \sum_{i=1}^{m} x_i)^{\alpha_0 - 1} \prod_{i=1}^{m} x_i^{\alpha_i - 1},$$

where $B(\alpha)$ is a multivariate beta-function.

- Representation using gamma distribution

$$(w_0, \ldots, w_m) \sim \mathcal{D}\text{ir}(\alpha) \iff w_i = \frac{Y_i}{\sum_{i=0}^m Y_i}, \quad Y_i \overset{\text{i.i.d}}{\sim} \Gamma(\alpha_i, 1).$$

Preliminaries: properties of Dirichlet distribution

The Dirichlet distribution $\mathcal{D}\mathrm{ir}(\alpha)$ for $\alpha = (\alpha_0, \ldots, \alpha_m) \in \mathbb{R}^m_{>0}$ is a distribution over m-dimensional simplex $\Delta_m = \{x \in \mathbb{R}^m \mid \sum_{i=1}^m x_i \leq 1\}$

$$p(x_1, \ldots, x_m) = \frac{1}{B(\alpha)}(1 - \sum_{i=1}^m x_i)^{\alpha_0 - 1} \prod_{i=1}^m x_i^{\alpha_i - 1},$$

where $B(\alpha)$ is a multivariate beta-function.

- Representation using gamma distribution

$$(w_0, \ldots, w_m) \sim \mathcal{D}\mathrm{ir}(\alpha) \iff w_i = \frac{Y_i}{\sum_{i=0}^m Y_i}, \quad Y_i \overset{\text{i.i.d}}{\sim} \Gamma(\alpha_i, 1).$$

- Aggregation property: if $\alpha \in \mathbb{N}^m$ and $\overline{\alpha} = \sum_{i=0}^m \alpha_i$

$$\sum_{i=0}^m w_i x_i = \sum_{j=1}^{\overline{\alpha}} \hat{w}_j y_j,$$

where $w \sim \mathcal{D}\mathrm{ir}(\alpha)$, $\hat{w} \sim \mathcal{D}\mathrm{ir}(\mathbf{1}^{\overline{\alpha}})$, y_j are copies of x_i repeated α_i times.

Bayes-UCBVI: From Dirichlet...

Based on joint work with D.Belomenstny, E.Moulines, A.Naumov, S.Samsonov, Y.Tang, M.Valko, P.Menard. "From Dirichlet to Rubin: Optimistic Exploiration in RL without Bonuses", Oral at ICML-2022.

Idea: use directly an upper quantile over posterior distribution.

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

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

Bayes-UCBVI: From Dirichlet...

Based on joint work with D.Belomenstny, E.Moulines, A.Naumov, S.Samsonov, Y.Tang, M.Valko, P.Menard. "From Dirichlet to Rubin: Optimistic Exploiration in RL without Bonuses", Oral at ICML-2022.

Idea: use directly an upper quantile over posterior distribution.

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

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

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

...to Rubin: Bayesian bootstrap

Given: sample $y^1, \ldots, y^n \sim \mathcal{P}$.
Goal: confidence interval for $\mathbb{E}_{y \sim \mathcal{P}}[y]$.

Classical (Efron) Bootstrap

- Resample $y^{1,b}, \ldots, y^{n,b}$.
- Compute mean estimate as $\frac{1}{n} \sum_{i=1}^{n} y^{i,b}$.
- Repeat B times.

Bayesian Bootstrap

- Sample $w^b \sim \mathcal{D}\mathrm{ir}(\mathbf{1}^n)$;
- Compute mean estimate as $\sum_{i=1}^{n} w^{b,i} y^i$;
- Repeat B times.

Then use quantiles of B mean estimates to construct a confidence interval.

Scalable implementation

- targets for Q-function estimation $y_h^n(s,a) \triangleq r_h(s,a) + \overline{V}^t_{h+1}(s^n_{h+1})$ for $n = 1, \ldots, n_h^t(s,a)$.
- prior targets $y_h^n(s,a) \triangleq r_h(s,a) + \overline{V}^t_h(s_0)$ for $n = -n_0 + 1, \ldots, 0$.

By aggregation property and sample quantile approximation

$$
\begin{aligned}
\overline{Q}^t_h(s,a) &\triangleq r_h(s,a) + \mathbb{Q}_{p \sim \rho_h^t(s,a)}\left(p\overline{V}^t_{h+1}(s,a), \kappa_h^t(s,a)\right) \\
&= \mathbb{Q}_{w \sim \mathcal{D}\mathrm{ir}(\mathbf{1}^{\bar{n}_h^t(s,a)})}\left(\sum_{n=-n_0+1}^{n_h^t(s,a)} w_n y_h^n(s,a), \kappa_h^t(s,a)\right) \\
&\approx \underbrace{\mathbb{Q}_{b \sim \mathcal{U}\mathrm{nif}([B])}\left(\sum_{n=-n_0+1}^{n_h^t(s,a)} w_h^{n,b}(s,a) y_h^n(s,a), \kappa_h^t(s,a)\right)}_{\text{upper confidence bound by Bayesian bootstrap}} .
\end{aligned}
$$

Deep RL extension: Bayes-UCBDQN

Recall

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

$$\text{where } \overline{Q}_h^{t,b}(s,a) \triangleq \sum_{n=-n_0+1}^{n_h^t(s,a)} w_h^{n,b}(s,a) y_h^n(s,a)\,.$$

Uniform Dirichlet distribution = exponential ($\Gamma(1,1)$) with normalization

$$\overline{Q}_h^{t,b}(s,a) = \arg\min_x \sum_{n=-n_0+1}^{n_h^t(s,a)} z_h^{n,b}(s,a)(x - y_h^n(s,a))^2$$

$$\text{where } z_h^{n,b}(s,a) \sim \mathcal{E}(1) \text{ i.i.d.}\,.$$

Deep RL:

- sample minibatch of targets;
- update parameters by the gradient of weighted linear regression.

Experimental results

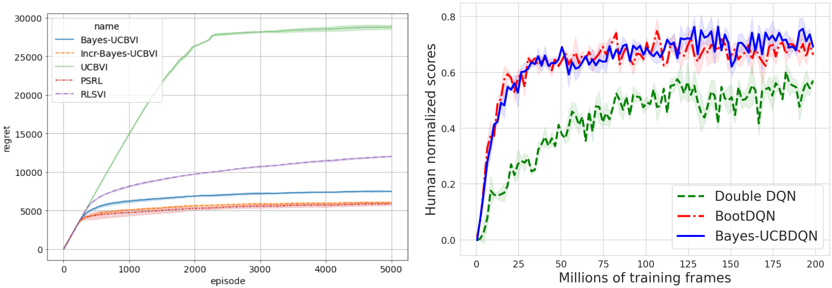


Figure: Left: Regret of `Bayes-UCBVI` and `Incr-Bayes-UCBVI` compared to baselines on grid-world with 5 rooms of size 5×5. Right: deep RL algorithms with median human normalized scores across Atari-57 games.

Back to theory: optimistic prior

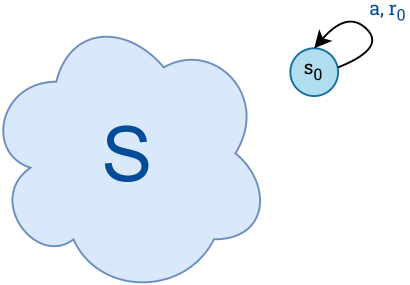


Figure: Extended state space by a fake state s_0, $r_0 > 1$.

Goal: encourage initial exploration.

- *Tabular:* prior $\rho_h^0(s'|s,a) = \mathcal{D}\mathrm{ir}(\{n_0\}_{s'=s_0} \cup \{0\}_{s' \in \mathcal{S}})$.
- *Deep RL:* Add n_0 prior transitions to s_0;

Theoretical analysis

Let us fix $\delta \in (0, 1)$, $r_0 \triangleq 2$, $n_0 \triangleq \mathcal{O}(\log(T))$, and the quantile function

$$\kappa_h^t(s, a) \triangleq 1 - \underbrace{\frac{C_\kappa \delta}{SAH[2n_h^t(s, a) + 1]^3 [\overline{n}_h^t(s, a)]^{3/2}}}_{\text{polynomial in parameters}} .$$

Theoretical analysis

Let us fix $\delta \in (0,1)$, $r_0 \triangleq 2$, $n_0 \triangleq \mathcal{O}(\log(T))$, and the quantile function

$$\kappa_h^t(s,a) \triangleq 1 - \underbrace{\frac{C_\kappa \delta}{SAH[2n_h^t(s,a)+1]^3[\overline{n}_h^t(s,a)]^{3/2}}}_{\text{polynomial in parameters}}.$$

Theorem (Regret bound)

For `Bayes-UCBVI`, with probability at least $1-\delta$,

$$\mathfrak{R}^T = \mathcal{O}\Big(\sqrt{H^3SAT}L + H^3S^2AL^2\Big),$$

where $L \triangleq \mathcal{O}(\log(HSAT/\delta))$.

Theoretical analysis

Let us fix $\delta \in (0,1)$, $r_0 \triangleq 2$, $n_0 \triangleq \mathcal{O}(\log(T))$, and the quantile function

$$\kappa_h^t(s,a) \triangleq 1 - \underbrace{\frac{C_\kappa \delta}{SAH[2n_h^t(s,a)+1]^3[\overline{n}_h^t(s,a)]^{3/2}}}_{\text{polynomial in parameters}}.$$

Theorem (Regret bound)

For Bayes-UCBVI, with probability at least $1-\delta$,

$$\mathfrak{R}^T = \mathcal{O}\Big(\sqrt{H^3SAT}L + H^3S^2AL^2\Big),$$

where $L \triangleq \mathcal{O}(\log(HSAT/\delta))$.

Matches the lower bound $\Omega(\sqrt{H^3SAT})$ up to poly-log terms.

Sketch of proof

The heart of the analysis is a novel anti-concentration inequality.

Sketch of proof

The heart of the analysis is a novel anti-concentration inequality.

Theorem (Dirichlet boundary crossing, Informal)

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* $f : \{0, \ldots, m\} \to [0, b_0]$ *and* $\mu \in (\overline{p}f, b_0)$

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

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

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

Sketch of proof

The heart of the analysis is a novel anti-concentration inequality.

Theorem (Dirichlet boundary crossing, Informal)

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* $f : \{0, \ldots, m\} \to [0, b_0]$ *and* $\mu \in (\overline{p}f, b_0)$

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

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`.

Takeaways

- Optimism in the face of uncertainty principle as a solution to exploration-exploitation dilemma;
- Bayesian perspective gives more possibility to scale up algorithms;
- Reinforcement learning is full of mathematical questions and fun!

Thank you!

Bibliography I

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