

Лекция 11

Обучение с подкреплением

Варвара Руденко

Московский физико-технический институт
Физтех-школа прикладной математики и информатики

29 апреля 2024



Multi-Armed Bandit Problem

- Задача многорукого бандита: агент выбирает действие согласно политике $\pi(a)$ и получает награду за каждый эпизод. Формально задача сводится к MDP, зависящем только от выбора действия:

$$Q(a) = \mathbb{E}_{p(r|a)} r, \quad V^* = \max_a Q(a).$$

Multi-Armed Bandit Problem

- Стратегия меняется каждый эпизод: на k -ом эпизоде сэмплируется $a_k \sim \pi_k(a)$ и награда $r_k \sim p(r \mid a_k)$, политика обновляется в π_{k+1} .
- Для T итераций: оптимально было бы выбирать каждый раз действие, дающее максимальную награду $\Rightarrow TV^*$. Учитывая, что не всегда будет выбрано лучшее действие, вводится функция потерь (regret) R :

$$R_T = TV^* - \sum_{k=0}^T Q(a_k).$$

- Тогда целью задачи является минимизация функции R .

Exploration-exploitation dilemma

- На каждой итерации делается выбор: следует агенту проверить новую "руку" (exploration) или продолжать "играть с той же рукой" (exploitation).

Простое решение

- Можно оценить $Q(a)$ по Монте-Карло:

$$Q_k(a) = \frac{\sum_k r_k [a_k = a]}{\sum_k [a_k = a]},$$

$$\begin{aligned} Q_k(a_k) &= \left(1 - \frac{1}{n_k(a_k)}\right) Q_{k-1}(a_k) + \frac{1}{n_k(a_k)} r_k = \\ &= Q_{k-1}(a_k) + \frac{1}{n_k(a_k)} (r_k - Q_{k-1}(a_k)), \end{aligned} \tag{1}$$

где $n_k(a)$ - счетчик, сколько раз было выбрано действие a .

- Утверждение 74: При любом алгоритме скорость роста сожалений не более чем линейная: для некоторой константы C :

$$\mathbb{E}R_T \leq CT.$$

- Утверждение 75: При жадном выборе сожаления растут с линейной скоростью: для некоторой константы C :

$$\mathbb{E}R_T \geq CT.$$

Наивный бандит

- Наивное решение — решать проблему "исследования" при помощи ϵ -жадной стратегии:

Алгоритм 1: Наивный бандит

Гиперпараметры: $\epsilon(k)$ — стратегия исследования.

Инициализировать $Q_0(a)$ произвольно.

Обнулить счётчики $n_0(a) := 0$.

На очередном шаге k :

- 1 выбрать a_k случайно с вероятностью $\epsilon(k)$, иначе $a_k := \operatorname{argmax}_{a_k} Q_k(a_k)$;
- 2 увеличить счётчик $n_k(a) := n_{k-1}(a) + [a_k = a]$;
- 3 пронаблюдать r_k ;
- 4 обновить $Q_k(a) := Q_{k-1}(a) + [a_k = a] \frac{1}{n_k(a)} (r_k - Q_{k-1}(a))$.

- Утверждение 76: При ε -жадном выборе сожаления всё равно растут с линейной скоростью: для некоторой константы C :

$$\mathbb{E}R_T \geq CT.$$

Доказательство. Поскольку на каждом шаге с вероятностью $\frac{\varepsilon}{|\mathcal{A}|}$ мы выбираем некоторое неоптимальное действие a с ненулевым регретом $V^* - Q(a)$, в среднем регрете появится слагаемое $T \frac{\varepsilon}{|\mathcal{A}|} (V^* - Q(a))$, следовательно как минимум можно в качестве константы C выбрать $\frac{\varepsilon}{|\mathcal{A}|} (V^* - Q(a))$. ■

Нестационарный бандит

Нестационарных бандитов отличает возможность того, что распределения $p(r \mid a)$ тоже меняются со временем. Наивное решение:

- ε нельзя уменьшать к нулю, необходимо постоянно пробовать различные действия, поэтому можно выставить ε в константу.
- Вместо счётчиков обновлять информацию о Q -функции через экспоненциальное сглаживание:

$$Q_k(a_k) = Q_{k-1}(a_k) + \alpha (r_k - Q_{k-1}(a_k)),$$

где $\alpha < 1$ — константный гиперпараметр.

Теорема Лаи-Роббинса

- Можно переписать функцию регрета в виде:

$$R_T = \sum_a n_T(a)(V^* - Q(a)).$$

Theorem (Теорема Лаи-Роббинса (Lai-Robbins theorem))

$$\mathbb{E} R_T \geq \log T \sum_{a \neq a^*} \frac{V^* - Q(a)}{\text{KL}(p(r \mid a) \parallel p(r \mid a^*))}$$

- Алгоритм **асимптотически оптимален**, если он имеет логарифмическую скорость роста среднего регрета.

Upper Confidence Bound (UCB)

- Вводится семейство алгоритмов: на очередном шаге k выбирается действие:

$$a_k := \operatorname{argmax}_a [Q_k(a) + U_k(a)],$$

где $U_k(a)$ — некоторая положительная добавка, имеющая смысл **бонуса за исследования** (exploration bonus). Бонус положителен из принципа **оптимизма перед неопределённостью** (optimism in the face of uncertainty).

Upper Confidence Bound (UCB)

- Идея **upper confidence bounds** (UCB) алгоритмов следующая:

$$Q(a) \leq Q_k(a) + U_k(a).$$

- То есть: строится **доверительный интервал** (confidence interval) и берется его верхняя граница. $U_k(a)$ будет обратно пропорционален $n_k(a)$, ведь граница будет сжиматься к эмпирическому среднему.

Theorem (Неравенство Хёфдинга)

Пусть $X_1 \dots X_n$ — i.i.d выборка из распределения на домене^a $[0, 1]$ с истинным средним μ , $\hat{\mu} := \frac{1}{N} \sum_i^N X_i$ — выборочная оценка среднего. Тогда $\forall u > 0$:

$$P(\mu \geq \hat{\mu} + u) \leq e^{-2nu^2}.$$

^aмы всегда предполагаем ограниченность наград; для удобства записи будем считать, что диапазон награды — $[0, 1]$.

Upper Confidence Bound (UCB)

- Утверждение 78: Для любого δ с вероятностью $1 - \delta$ истинное значение $Q(a)$ не превосходит $Q_k(a) + U_k(a)$, где:

$$U_k(a) := \sqrt{\frac{-\ln \delta}{2n_k(a)}}.$$

Доказательство. В силу неравенства Хёфдинга:

$$P(Q(a) \geq Q_k(a) + U_k(a)) \leq \exp^{-2n_k(a)U_k(a)^2}.$$

Возьмём заданное δ и прогарантируем, что $\exp^{-2n_k(a)U_k(a)^2} = \delta$.
Разрешая это равенство относительно $U_k(a)$, получаем доказываемое. ■

- $$a_k = \operatorname{argmax}_a \left[Q_k(a) + c \sqrt{\frac{\log k}{n_k(a)}} \right].$$

Сэмплирование Томпсона

- Предположение: распределения наград $p(r \mid a)$ принадлежат некоторому параметрическому семейству $\theta_a \in \Theta$, награда генерируется из распределения $p(r \mid \theta_a)$.
- Предлагается **байесовский вывод**: Зададимся некоторым **априорным распределением** («прайором») $p(\theta_a)$ для каждого действия a и после очередного эпизода игры с автоматом a с исходом r_k будем обновлять распределение над θ_a по формуле Байеса на **апостериорное распределение**:

$$p(\theta_a) \leftarrow p(\theta_a \mid r_k) \propto p(r_k \mid \theta_a)p(\theta_a).$$

- Средний выигрыш из автомата a — $\mathbb{E}p(r \mid \theta_a)$, теперь будет являться случайной величиной, поскольку θ_a — случайное, с распределением $p(\theta_a)$.

Сэмплирование Томпсона

- **Probability matching**: Согласно текущим $p(\theta_a)$ жадный выбор действия имеет вид:

$$a := \operatorname{argmax}_a \mathbb{E}_{\theta_a \sim p(\theta_a)} \mathbb{E} p(r \mid \theta_a).$$

- Появляется вероятность неоптимального действия, то есть с учётом распределений θ_a посчитать вероятность, что хотя бы для одного другого автомата $\hat{a} \neq a$:

$$\mathbb{E} p(r \mid \theta_{\hat{a}}) > \mathbb{E} p(r \mid \theta_a).$$

- **Сэмплирование Томпсона** (Thompson Sampling) – процедура, при которой на k -ом шаге при решении задачи многоруких бандитов действие a выбирается с вероятностью того, что оно оптимально в рамках выученных моделей $p(\theta_a)$:

$$\pi(a) := P \left(\mathbb{E} p(r \mid \theta_a) = \max_{\hat{a}} \mathbb{E} p(r \mid \theta_{\hat{a}}) \right).$$

Bernoulli-бандиты

Положим, что автоматы выдают только награду 0 или 1 – распределение Бернулли с вероятностью θ_a :

$$p(r \mid a) := \text{Bernoulli}(r \mid \theta_a) = \theta_a^{\mathbb{I}[r=1]}(1 - \theta_a)^{\mathbb{I}[r=0]} = \theta_a^r(1 - \theta_a)^{1-r}.$$

Априорное распределение зададим при помощи Бета-распределения¹:

$$p(\theta_a) := \text{Beta}(\theta_a \mid \alpha, \beta) \propto \theta_a^{\alpha-1}(1 - \theta_a)^{\beta-1},$$

где α и β — некоторые параметры (свои для каждого автомата a). Мы можем обновлять эти знания при помощи новых сэмплов $r_k \sim p(r \mid a)$, проводя байесовский вывод. Применяем формулу Байеса:

$$p(\theta_a \mid r) \propto \underbrace{\theta_a^{r_k}(1 - \theta_a)^{1-r_k}}_{p(r_k \mid \theta_a)} \underbrace{\theta_a^{\alpha-1}(1 - \theta_a)^{\beta-1}}_{p(\theta_a)} = \text{Beta}(\theta_a \mid \alpha + r_k, \beta + 1 - r_k).$$

¹это распределение является **сопряжённым** (conjugate) к Бернулли, что означает, что после применения формулы Байеса наше распределение останется в этом же семействе распределений.

Bernoulli-бандиты

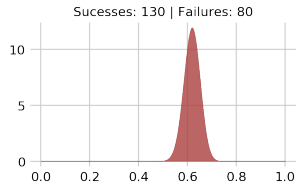
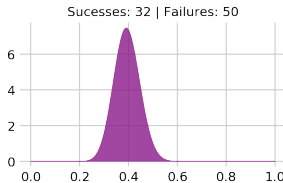
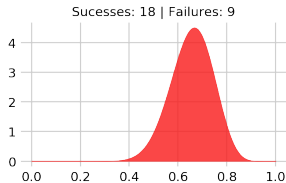
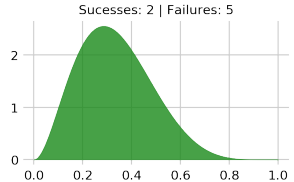
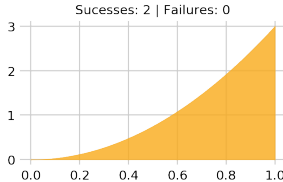
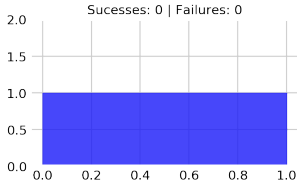
Для обновления параметров α, β достаточно увеличить α на r_k , а β — на $1 - r_k$. При заданном θ_a мы можем посчитать среднее значение выигрыша, ценность автомата a :

$$\hat{Q}(a) = \mathbb{E}p(r \mid \theta_a) = \theta_a.$$

Мат.ожидание выигрыша в нашей модели:

$$\mathbb{E}_{\theta_a} \hat{Q}(a) = \mathbb{E}_{\text{Beta}(\theta_a | \alpha, \beta)} \theta_a = \frac{\alpha}{\alpha + \beta}.$$

Bernoulli-бандиты



Планирование

- Для данного состояния s_0 набор действий a_0, a_1, a_2, \dots – **план** (plan).
- Допустим агент находится в некотором s_0 и хочет найти хорошее действие a_0 , после чего сэмплируется s_1 , выбирается $a_1 \Rightarrow$ строится план. Но подобные алгоритмы могут «спланировать» свои будущие действия, а не выучить непосредственно зависимость $\pi(a \mid s)$ («обучить стратегию»).
- При доступной функции переходов $p(s' \mid s, a)$ и функции награды такая задача называется **планированием** (planning).

$$\operatorname{argmax}_{a_0, a_1, a_2 \dots} \mathbb{E}_{\mathcal{T} \mid s_0, a_0, a_1, a_2 \dots} R(\mathcal{T}).$$

World Models

- **Моделью мира** (world model) называется любая модель, явно или неявно обучающаяся модели динамики среды.

Алгоритм 2: Общая схема Model-based подхода

Гиперпараметры: Планировщик, модель мира.

Проинициализировать стратегию $\pi_0(a \mid s)$ случайно.

Проинициализировать модель мира случайно.

Проинициализировать датасет пустым множеством.

На k -ом шаге:

- 1 Повзаимодействовать со средой при помощи π_k , добавив встреченные переходы $\{s, a, r, s'\}$ в датасет.
- 2 Провести дообучение модели мира на собранном датасете.
- 3 Получить π_k из планировщика, используя текущую модель мира.

Модель прямой динамики

- Модель функции переходов $p_\theta(s' \mid s, a)$ называется **моделью прямой динамики** (forward dynamics model).
- Учить генеративную модель может быть дорого, и простым удешевлением является обучение детерминированного приближения $s' \approx f_\theta(s, a)$. Это можно делать по любым доступным траекториям, собранным любым способом:

$$\sum_{s,a,s'} \|f_\theta(s, a) - s'\|^2 \rightarrow \min_{\theta};$$

$$\sum_{s,a,r} (r_\psi(s, a) - r)^2 \rightarrow \min_{\psi}.$$

Сновидения

- Наличие модели прямой динамики $p_{\theta}(s', r \mid s, a)$ позволяет при помощи текущей стратегии π генерировать траектории, используя полностью «внутренние модели» и никак не используя реальную внешнюю среду.
- **Сновидениями** (dreaming) называется обучение агента на опыте, собранном при помощи приближения динамики среды $p_{\theta}(s', r \mid s, a)$.

Интересные результаты

- Optimistic Posterior Sampling for Reinforcement Learning with Few Samples and Tight Guarantees by Daniil Tiapkin, Denis Belomestny et al. (2022)
- From Dirichlet to Rubin: Optimistic Exploration in RL without Bonuses by Daniil Tiapkin, Denis Belomestny et al. (2022)

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

- Reinforcement learning in an environment modeled by an episodic, finite, stage-dependent Markov decision process of horizon H with S states, and A actions.
- The performance of an agent is measured by the regret after interacting with the environment for T episodes.
- Proposed an optimistic posterior sampling algorithm for reinforcement learning OPSRL, a simple variant of posterior sampling that only needs a number of posterior samples logarithmic in H , S , A , and T per state-action pair.

- For OPSRL guaranteed a high-probability regret bound of order at most $\tilde{O}(\sqrt{H^3 SAT})$ ignoring $\text{poly} \log(HSAT)$ terms.
- The key novel technical ingredient is a new sharp anti-concentration inequality for linear forms which may be of independent interest. Specifically, extend the normal approximation-based lower bound for Beta distributions by Alfers and Dinges¹ to Dirichlet distributions. Bound matches the lower bound of order $\Omega(\sqrt{H^3 SAT})$, thereby answering the open problems raised by Agrawal and Jia² for the episodic setting.

¹A normal approximation for beta and gamma tail probabilities. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 65(3):399–420, Feb 1984. ISSN 1432-2064. doi: 10.1007/BF00533744.

²Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017b.

Theorem (Regret bound for OPSRL)

Consider a parameter $\delta \in (0, 1)$. Let

$$\kappa \triangleq 2(\log(12SAH/\delta) + 3\log(\exp \pi(2T + 1))),$$

$n_0 \triangleq \lceil \kappa(c_0 + \log_{17/16}(T)) \rceil$, $r_0 \triangleq 2$, where c_0 is an absolute constant defined as:

$$c_0 \triangleq \left(\frac{4}{\sqrt{\log(17/16)}} + 8 + \frac{49 \cdot 4\sqrt{6}}{9} \right)^2 \frac{8}{\pi} + \log_{17/16} \left(\frac{20}{32} \right) + 1. \quad (2)$$

Then for OPSRL, with probability at least $1 - \delta$,

$$\mathfrak{R}^T = \tilde{O} \left(\sqrt{H^3 SATL^3} + H^3 S^2 AL^3 \right),$$

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

Алгоритм 3: OPSRL

Input: Family of probability distributions $\rho : \mathcal{N}_+^{S+1} \rightarrow \Delta_{\mathcal{S}'}$ over transitions, initial pseudo-count \bar{n}_h^0 , number of posterior samples J .

- **for** $t \in [T]$ **do**
- For all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, sample J independent transitions

$$\tilde{p}_h^{t-1,j}(s, a) \sim \rho(\bar{n}_h^{t-1}(s'|s, a)_{s' \in \mathcal{S}'}), \quad j \in [J].$$

Алгоритм 3: OPSRL (продолжение)

- Optimistic backward induction: set $\bar{V}_{H+1}^{t-1}(s) = 0$ and recursively for $h \in [H]$, compute

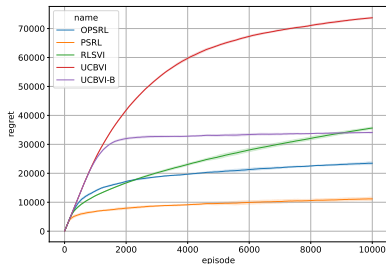
$$\bar{Q}_h^{t-1}(s, a) = r_h(s, a) + \max_{j \in [J]} \{ \tilde{p}_h^{t-1,j} \bar{V}_{h+1}^{t-1}(s, a) \},$$

$$\bar{V}_h^{t-1}(s) = \max_{a \in \mathcal{A}} \bar{Q}_h^{t-1}(s, a),$$

$$\pi_h^t(s) \in \operatorname{argmax}_{a \in \mathcal{A}} \bar{Q}_h^{t-1}(s, a).$$

- **for** $h \in [H]$ **do**
- Play $a_h^t = \pi_h^t(s_h^t)$.
- Observe $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$.
- Increment the pseudo-count $\bar{n}_h^t(s_{h+1}^t | s_h^t, a_h^t)$.
- **end for**
- **end for**

Experimental results



Regret of OPSRL and baselines on grid-world environment with 100 states and 4 action for $H = 50$ and transitions noise 0.2, average over 4 seeds.


From Dirichlet to Rubin: Optimistic Exploration in RL without Bonuses

- Proposed the BayesUCBVI algorithm for reinforcement learning in tabular, stage-dependent, episodic Markov decision process: a natural extension of the BayesUCB algorithm by Kaufmann et al.¹ for multiarmed bandits.
- For BayesUCBVI, we prove a regret bound of order $\tilde{O}(\sqrt{H^3 SAT})$ where H is the length of one episode, S is the number of states, A the number of actions, T the number of episodes, that matches the lower-bound of $\Omega(\sqrt{H^3 SAT})$ up to poly-log terms in H, S, A, T for a large enough T .

¹On bayesian upper confidence bounds for bandit problems. In Neil D. Lawrence and Mark Girolami, 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, 2012. PMLR.

From Dirichlet to Rubin: Optimistic Exploration in RL without Bonuses

- This method uses the quantile of a Q-value function posterior as upper confidence bound on the optimal Q-value function.
- To the best of our knowledge, this is the first algorithm that obtains an optimal dependence on the horizon H (and S) *without the need of an involved Bernstein-like bonus or noise*. Crucial to our analysis is a new fine-grained anticoncentration bound for a weighted Dirichlet sum that can be of independent interest.
- And then explain how BayesUCBVI can be easily extended beyond the tabular setting, exhibiting a strong link between our algorithm and Bayesian bootstrap Rubin¹.

¹The bayesian bootstrap. The annals of statistics, pages 130–134, 1981. 

Regret upper bound for episodic, non-stationary, tabular MDPs

Algorithm	Upper bound (non-stationary)
UCBVI (Azar et al.) ¹	$\tilde{O}(\sqrt{H^3 SAT})$
UCBAdventage (Zhang et al.) ²	
RLSVI (Xiong et al.) ³	
PSRL Agrawal and Jia ⁴	$\tilde{O}(H^2 S \sqrt{AT})$
BootNARL(Pacchiano et al.) ⁵	
BayesUCBVI (this paper)	$\tilde{O}(\sqrt{H^3 SAT})$
Lower bound (Jin et al., Domingues et al.) ⁶	$\Omega(\sqrt{H^3 SAT})$

Regret upper bound for episodic, non-stationary, tabular MDPs

- 1) Minimax regret bounds for reinforcement learning. In International Conference on Machine Learning, 2017.
- 2) Almost optimal model-free reinforcement learning via reference-advantage decomposition. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.
- 3) Near-optimal randomized exploration for tabular mdp, 2021.
- 4) Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017b.

Regret upper bound for episodic, non-stationary, tabular MDPs

- 5) Towards tractable optimism in model-based reinforcement learning. In Cassio de Campos and Marloes H. Maathuis, editors, Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, volume 161 of Proceedings of Machine Learning Research, pages 1413–1423. PMLR, 2021.
- 6) Is Q-learning provably efficient? In Neural Information Processing Systems, 2018.

From Dirichlet to Rubin: Optimistic Exploration in RL without Bonuses

Theorem

Consider a parameter $\delta > 0$. Let $n_0 \triangleq \lceil c_{n_0} + \log_{17/16}(T) \rceil$, $r_0 \triangleq 2$, where c_{n_0} is an absolute constant defined as:

$$c_{n_0} = \frac{1}{(\sqrt{2\pi} - 1)^2} \cdot \left(\frac{2\sqrt{2}}{\sqrt{\log(17/16)}} + \frac{98\sqrt{6}}{9} \right)^2 + \frac{\log(10\pi)}{\log(17/16)}. \quad (3)$$

Then for BayesUCBVI, with probability at least $1 - \delta$,

$$\mathfrak{R}^T = \tilde{O} \left(\sqrt{H^3 S A T L} + H^3 S^2 A L^2 \right),$$

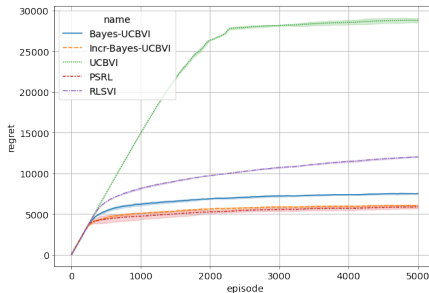
where $L \triangleq \tilde{O}(\log(HSAT/\delta))$.

Алгоритм 4: BayesUCBVI

Input: quantile functions $(\kappa^t)_{t \in [T]}$, prior dist. ρ^0

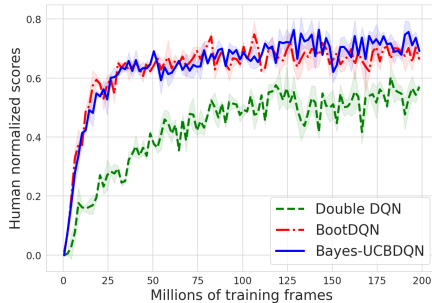
- **for** $t \in [T]$ **do**
- Optimistic planning, see
- **for** $h \in [H]$
- Play $a_h^t \in \operatorname{argmax}_{a \in \mathcal{A}} \overline{Q}_h^{t-1}(s_h^t, a)$
- Observe $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$
- Update the posterior distributions $\rho_h^t(s_h^t, a_h^t)$ with $(s_h^t, a_h^t, s_{h+1}^t)$
- **end for**
- **end for**

Experimental results



Regret of BayesUCBVI and IncrBayesUCBVI compared to baselines for $H = 30$ and transitions noise 0.1, average over 4 seeds.

Experimental results



Evaluating deep RL algorithms with median human normalized scores across Atari-57 games. We compare DoubleDQN, BootDQN and BayesUCBDQN. The training curves show the average \pm std over 3 seeds.