Is Q-learning Provably Efficient?

Chi Jin*
University of California, Berkeley
chijin@cs.berkeley.edu

Zeyuan Allen-Zhu*
Microsoft Research, Redmond
zeyuan@csail.mit.edu

Sebastien Bubeck
Microsoft Research, Redmond
sebubeck@microsoft.com

Michael I. Jordan
University of California, Berkeley
jordan@cs.berkeley.edu

【强化学习 93】UCB+Q-learning

张楚珩
清华大学 交叉信息院博士在读

这篇文章把 UCB 和 Q-learning 相结合，得到两种算法 UCB-H 和 UCB-B，这里把它们统记为 UCB + Q-learning。

原文传送门

Jin, Chi, et al. "Is q-learning provably efficient?." Advances in Neural Information Processing Systems. 2018.

Arxiv version (It is also a best paper in ICML 2018 workshop "Exploration in RL")

特色

前面有两讲讲了 PG theory，里面提到要得到一个有效的强化学习算法，探索是一个必不可少的环节。不同于常用的 epsilon-greedy 的探索形式，这里使用 upper confidence bound（UCB） 来做探索。UCB 是 MAB 问题中一个有效的解法，本来以为 UCB+Q-learning 应该是很早就有人做了的，但是这篇还算比较近，在 NIPS 2018 。

Ps，看到作者才想到去年的这个时候 Michael Jordan 来清华讲课的时候，提到过这一篇；今天大佬又来了，但是今天懒了，没去。

过程

1. 背景

在 RL 方面，文章使用 finite horizon （每个 episode 都固定 H 步）+ cumulated reward （no discount）的设定。

	Algorithm	Regret	Time	Space
Model-based	UCRL2 [10] [1]	at least $\tilde{\mathcal{O}}(\sqrt{H^4S^2AT})$	$\Omega(TS^2A)$	$\mathcal{O}(S^2AH)$
	Agrawal and Jia [1] [1]	at least $\tilde{\mathcal{O}}(\sqrt{H^3S^2AT})$		
	UCBVI [5] [2]	$\tilde{\mathcal{O}}(\sqrt{H^2SAT})$	$\tilde{\mathcal{O}}(TS^2A)$	
	vUCQ [12] [2]	$\tilde{\mathcal{O}}(\sqrt{H^2SAT})$		
Model-free	Q-learning (ε-greedy) [14] (if 0 initialized)	$\Omega(\min\{T, A^{H/2}\})$	$\mathcal{O}(T)$	$\mathcal{O}(SAH)$
	Delayed Q-learning [25] [3]	$\tilde{\mathcal{O}}_{S,A,H}(T^{4/5})$		
	Q-learning (UCB-H)	$\tilde{\mathcal{O}}(\sqrt{H^4SAT})$		
	Q-learning (UCB-B)	$\tilde{\mathcal{O}}(\sqrt{H^3SAT})$		
	lower bound	$\Omega(\sqrt{H^2SAT})$	-	-

Table 1: Regret comparisons for RL algorithms on episodic MDP. $T = KH$ is totally number of steps, H is the number of steps per episode, S is the number of states, and A is the number of actions. For clarity, the table is presented for $T \geq \text{poly}(S, A, H)$, omitting low order terms.

上表中 T 表示 sample complexity；S 和 A 分别代表状态空间和动作空间的大小；H 表示一个回合有多少步。注意到，当 H=1 时就是一个普通的 contextual MAB 问题，因此这里可以看做是一个有 H『层』的 contextual MAP 问题。

从 regret 上来说，model-based 方法之前有 $O(\sqrt{T})$ 的结果，但是这类方法需要估计一个 model，这样需要一个比较大的存储空间；model-free 的方法就只需要 online 地更新，只需要存一个价值函数或者策略即可，空间上的复杂度会更低。

Ps，文章讲了两种算法，我暂时只看了正文里面讲的 UCB-H 方法，UCB-B 方法在附录中讲的，主要额外估计了 variance 并使用它来计算 upper confidence，从而得到更好的结果。没有仔细看，因此这里不讲了。

2. Q-learning combined with UCB

文章最核心的想法可以从如下公式看出：

$$Q_h(x,a) \leftarrow (1-\alpha_t)Q_h(x,a) + \alpha_t[r_h(x,a) + V_{h+1}(x') \boxed{+ b_t}] \ ,$$

其核心就是 Q-value 每次更新的 target 都会加上一个 exploration bonus b_t ，而这里的 $t = N_h(x,a)$ ，表示该 (h,x,a) 之前遇到过多少次。文章的关键就是给出了 α_t, b_t 的具体函数关系，并且导出了这样函数关系下相应的 regret。

UCB-H 算法如下：

Algorithm 1 Q-learning with UCB-Hoeffding

```
initialize Q_h(x,a) ← H and N_h(x,a) ← 0 for all (x,a,h) ∈ S × A × [H].
for episode k = 1, ..., K do
    receive x_1.
    for step h = 1, ..., H do
        Take action a_h ← argmax_{a'} Q_h(x_h, a'), and observe x_{h+1}.
        t = N_h(x_h, a_h) ← N_h(x_h, a_h) + 1;  b_t ← c√(H^3 ι/t).
        Q_h(x_h, a_h) ← (1 − α_t) Q_h(x_h, a_h) + α_t [r_h(x_h, a_h) + V_{h+1}(x_{h+1}) + b_t].
        V_h(x_h) ← min{H, max_{a'∈A} Q_h(x_h, a')}.
```


该算法中，选择 $b_t = c\sqrt{H^3\iota/t}$ ， $\alpha_t = \frac{H+1}{H+t}$ 。

该算法相应的 regret 有如下保证：

Theorem 1 (Hoeffding). *There exists an absolute constant $c > 0$ such that, for any $p \in (0,1)$, if we choose $b_t = c\sqrt{H^3\iota/t}$, then with probability $1-p$, the total regret of Q-learning with UCB-Hoeffding (see Algorithm 1) is at most $O(\sqrt{H^4 SAT\iota})$, where $\iota := \log(SAT/p)$.*

3. 证明

证明分为如下几个步骤：

- 第一步：给定的是 Q-learning 的更新公式，更新公式是 bootstrap 的形式，即新的 Q 值依赖于旧的 Q 值。因此，***第一步需要把这样的『递推公式』写成『通项公式』的形式***。这其中需要到一个 tradeoff：如果过度依赖于最新的 Q 值，估计会更为 unbiased，但是 variance 会比较大（因为参考的样本数目较少）；如果过度依赖以前的 Q 值，就会产生更大的 bias。 α_t 的选取方式就是在平衡这样一个 tradeoff。
- 第二步：虽然最后我们需要的 regret 是比较最优策略的性能和每一轮策略的性能，但是策略是依赖于所估计的 Q 函数值的，因此在***第二步中需要先 bound 所估计的 Q 函数值和最优 Q 函数值的差距，即*** $Q^k - Q^*$ 。这一步中最为核心的想法是，如果某一个 (h,x,a) 被访问的次数多（ t 比较大），那么对它的估计就更准确，相应地，可以匹配一个较小的 exploration bonus b_t ；反之，就需要一个更大的 exploration bonus 来鼓励探索。这一步依赖于前一步中写出的『通项公式』。
- 第三步：***把 regret 写出来，并且进行缩放改写，最后利用前述性质推导其上界***。其中核心的想法是，regret 关心的是 $V_1(x_1)$ 的差值，然而它的差值又可以写作后一个状态价值函数 $V_2(x_2)$ 的差值，以此类推，到最后（第 $H+1$ 步）各个价值函数又是相同的了（即，误差为零）。因此，误差是从 H 步往回累加起来的。因此，需要注意误差传递的途径，然后写出最后的 regret 形式。

预备：更新公式

$$Q_h^{k+1}(x,a) = \begin{cases} (1-\alpha_t)Q_h^k(x,a) + \alpha_t[r_h(x,a) + V_{h+1}^k(x_{h+1}^k) + b_t] & \text{if } (x,a) = (x_h^k, a_h^k) \\ Q_h^k(x,a) & \text{otherwise} \ . \end{cases} \tag{4.1}$$

$$V_h^k(x) \leftarrow \min\left\{H, \max_{a' \in \mathcal{A}} Q_h^k(x,a')\right\}, \quad \forall x \in \mathcal{S} \ .$$

其中上标 $k \in [K]$ 表示是第 k 个 episode，而下标 $h \in [H]$ 表示是 episode 中的第几步。注意到，每一步的奖励在 $[0,1]$ 之间，因此价值函数不超过 H，因此状态价值函数会使用 H 来截断。

预备：α_t 的性质

文章选择 $\alpha_t = \frac{H+1}{H+t}$ ，如果定义

$$\alpha_t^0 = \prod_{j=1}^{t}(1-\alpha_j), \qquad \alpha_t^i = \alpha_i \prod_{j=i+1}^{t}(1-\alpha_j) \ . \tag{4.2}$$

容易验证它具有如下性质。上述定义的作用后面将会看到。

Lemma 4.1. *The following properties hold for* α_t^i*:*

(a) $\frac{1}{\sqrt{t}} \le \sum_{i=1}^{t} \frac{\alpha_t^i}{\sqrt{i}} \le \frac{2}{\sqrt{t}}$ *for every* $t \ge 1$.

(b) $\max_{i\in[t]} \alpha_t^i \le \frac{2H}{t}$ *and* $\sum_{i=1}^{t}(\alpha_t^i)^2 \le \frac{2H}{t}$ *for every* $t \ge 1$.

(c) $\sum_{t=i}^{\infty} \alpha_t^i = 1 + \frac{1}{H}$ *for every* $i \ge 1$.

第一步：价值函数的通项公式

$$Q_h^k(x,a) = \alpha_t^0 H + \sum_{i=1}^{t} \alpha_t^i \left[r_h(x,a) + V_{h+1}^{k_i}(x_{h+1}^{k_i}) + b_i \right] \ . \tag{4.3}$$

这里我们就能观察到刚刚所定义的 α_t^i 的作用了，它其实是在第 t 步时各项实际的权重。前面说到 α_t 的设置涉及到一个 bias-variance tradeoff：如果对于一个固定的 t， α_t^i 特别倾向于最新的 target Q value（上式方括号中的式子），那么就会产生较大的 variance；对于一个固定的 t， α_t^i 比较平均地对历史上遇到的 target Q value 加权，那么会导致较大的 bias。文章采取的这种加权方案能够使得不管 t 为多少，都几乎只考虑最后的 $1/H$ 份的样本，而最开始的 $1-1/H$ 份样本会被『遗忘』。从下图可以看出这一点。

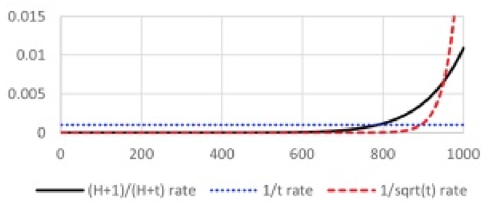

Figure 1: Illustration of $\{\alpha_{1000}^i\}_{i=1}^{1000}$ for learning rates $\alpha_t = \frac{H+1}{H+t}$, $\frac{1}{t}$ and $\frac{1}{\sqrt{t}}$ when $H = 10$.

注意到，第 h 层的价值函数依赖于第 h+1 层的价值函数，即价值函数是一层层传递过来的，因此总体上来看，所有的样本都能够得到有效地利用。

第二步：估计价值函数和最优价值函数的差距

首先可以得到，如下式子

Lemma 4.2 (recursion on Q). *For any $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and episode $k \in [K]$, let $t = N_h^k(x, a)$ and suppose (x, a) was previously taken at step h of episodes $k_1, \ldots, k_t < k$. Then:*

$$(Q_h^k - Q_h^\star)(x, a) = \alpha_t^0 (H - Q_h^\star(x, a)) + \sum_{i=1}^{t} \alpha_t^i \left[(V_{h+1}^{k_i} - V_{h+1}^\star)(x_{h+1}^{k_i}) + [(\hat{\mathbb{P}}_h^{k_i} - \mathbb{P}_h) V_{h+1}^\star](x, a) + b_i \right] .$$

注意到

$$[\mathbb{P}_h V_{h+1}](x, a) := \mathbb{E}_{x' \sim \mathbb{P}(\cdot|x,a)} V_{h+1}(x')$$

$$[\hat{\mathbb{P}}_h^{k_i} V_{h+1}](x, a) := V_{h+1}(x_{h+1}^{k_i})$$

接下来，考虑到 $\hat{\mathbb{P}} - \mathbb{P}$ 其实就是 sample mean - true mean，可以使用 concentration inequality 来 bound 这一项；直观来说，就是越多的样本估计的越准确。

由此，可以得到如下引理：

Lemma 4.3 (bound on $Q^k - Q^\star$). *There exists an absolute constant $c > 0$ such that, for any $p \in (0, 1)$, letting $b_t = c\sqrt{H^3 \iota / t}$, we have $\beta_t = 2 \sum_{i=1}^t \alpha_t^i b_i \le 4c\sqrt{H^3 \iota / t}$ and, with probability at least $1 - p$, the following holds simultaneously for all $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$:*

$$0 \le (Q_h^k - Q_h^\star)(x, a) \le \alpha_t^0 H + \sum_{i=1}^{t} \alpha_t^i (V_{h+1}^{k_i} - V_{h+1}^\star)(x_{h+1}^{k_i}) + \beta_t ,$$

where $t = N_h^k(x, a)$ and $k_1, \ldots, k_t < k$ are the episodes where (x, a) was taken at step h.

文章中写 $\iota = \log(SAT/p)$ ，但是我推导了半天感觉是 $\iota = \log(SAH/p)$ 。。。反正，最重要的就是对于每一个 (h, x, a) 应用 Azuma-Hoeffding，然后再加 union bound 合起来，这样就能得到 $\hat{\mathbb{P}} - \mathbb{P}$ 项的界，接下来就需要设置 b_t 来匹配相应的界，使得所估计的 Q 值大概率是 Q^* 的上界，但是同时也比较紧（不会比 Q^* 大太多）。

第三步：计算 regret

首先注意到：由于估计的 Q 值是 upper confidence，因此第 k 轮估计的 Q 值 Q_h^k 、最优 Q 值 Q_h^* 和实际策略的 Q 值 $Q_h^{\pi_k}$ 的关系大致为： $Q_h^k > Q_h^* > Q_h^{\pi_k}$ 。定义

$$\delta_h^k := (V_h^k - V_h^{\pi_k})(x_h^k) \quad \text{and} \quad \phi_h^k := (V_h^k - V_h^\star)(x_h^k) .$$

因此，regret 可以被 bound 为

$$\mathrm{Regret}(K) = \boxed{\sum_{k=1}^K (V_1^\star - V_1^{\pi_k})(x_1^k) \le \sum_{k=1}^K (V_1^k - V_1^{\pi_k})(x_1^k)} = \sum_{k=1}^K \delta_1^k .$$

而 δ_h 和后续的 δ_{h+1} 又具有一定的联系

$$
\begin{aligned}
\delta_h^k &= (V_h^k - V_h^{\pi_k})(x_h^k) \overset{①}{\leq} (Q_h^k - Q_h^{\pi_k})(x_h^k, a_h^k) \\
&= (Q_h^k - Q_h^\star)(x_h^k, a_h^k) \boxed{+ (Q_h^\star - Q_h^{\pi_k})(x_h^k, a_h^k)} \\
&\overset{②}{\leq} \alpha_t^0 H + \textstyle\sum_{i=1}^t \alpha_t^i \phi_{h+1}^{k_i} + \beta_t + [\mathbb{P}_h(V_{h+1}^\star - V_{h+1}^{\pi_k})](x_h^k, a_h^k) \\
&\overset{③}{=} \alpha_t^0 H + \textstyle\sum_{i=1}^t \alpha_t^i \phi_{h+1}^{k_i} + \beta_t - \phi_{h+1}^k + \delta_{h+1}^k + \xi_{h+1}^k \ ,
\end{aligned}
\tag{4.6}
$$

① 中的小于等于号是由于 estimated V 定义中有一个 clip 操作，同时策略就是确定性地选择 a_h^k ，因此 $V_h^{\pi_k}(s_h^k) = Q^{\pi_k}(s_h^k, \pi_k) = Q^{\pi_k}(s_h^k, a_h^k)$ 。② 用到了前一步的结论，以及 $\mathbb{P}_h$ 算子的定义。③ 是恒等变化，需要注意到 $\hat{\mathbb{P}}_h$ 算子的定义，其中

$$\beta_t = 2\sum \alpha_t^i b_i \leq O(1)\sqrt{H^3\iota/t} \text{ and } \xi_{h+1}^k := [(\mathbb{P}_h - \hat{\mathbb{P}}_h^k)(V_{h+1}^\star - V_{h+1}^k)](x_h^k, a_h^k)$$

绿框部分：注意到 $Q_h^k > Q_h^* > Q_h^{\pi_k}$ ，这里其实是把 optimal V - policy V，放大成了 estimated V - policy V；但是看似矛盾的是，后面又拆成了 (estimated Q - optimal Q) + (optimal Q - policy Q)。但其实不矛盾的，一步过来的话，① 中的小于等于号是不成立的。

其中第一项

$$\boxed{\sum_{k=1}^K \alpha_{n_h^k}^0 H = \sum_{k=1}^K H \cdot \mathbb{I}[n_h^k = 0] \leq SAH}$$

第二项

$$\sum_{k=1}^K \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \phi_{h+1}^{k_i(x_h^k, a_h^k)} \leq \sum_{k'=1}^K \phi_{h+1}^{k'} \sum_{t=n_h^{k'}+1}^{\infty} \alpha_t^{n_h^{k'}} \leq \left(1 + \frac{1}{H}\right) \sum_{k=1}^K \phi_{h+1}^k,$$

根据 $\delta_h^k \geq \phi_h^k$ ，有

$$
\begin{aligned}
\sum_{k=1}^K \delta_h^k &\leq SAH + \left(1 + \frac{1}{H}\right) \sum_{k=1}^K \phi_{h+1}^k - \sum_{k=1}^K \phi_{h+1}^k + \sum_{k=1}^K \delta_{h+1}^k + \sum_{k=1}^K (\beta_{n_h^k} + \xi_{h+1}^k) \\
&\leq SAH + \left(1 + \frac{1}{H}\right) \sum_{k=1}^K \delta_{h+1}^k + \sum_{k=1}^K (\beta_{n_h^k} + \xi_{h+1}^k) \ ,
\end{aligned}
\tag{4.7}
$$

根据递推关系，能够得到

$$\sum_{k=1}^K \delta_1^k \leq O\Big(H^2 SA + \sum_{h=1}^H \sum_{k=1}^K (\beta_{n_h^k} + \xi_{h+1}^k)\Big).$$

最后 bound β 项和 ξ 项即可：

$$\sum_{k=1}^{K}\beta_{n_h^k} \le O(1)\cdot\sum_{k=1}^{K}\sqrt{\frac{H^3\iota}{n_h^k}} = O(1)\cdot\sum_{x,a}\sum_{n=1}^{N_h^K(x,a)}\sqrt{\frac{H^3\iota}{n}} \overset{①}{\le} O\big(\sqrt{H^3SAK\iota}\big) = O\big(\sqrt{H^2SAT\iota}\big) \quad (4.8)$$

其中，① 需要注意到 $\sum_{x,a} N_h^k(x,a) = K$ ，因此取 $N_h^k(x,a) = K/SA, \forall h,x,a$ 时能得到 upper bound。同时用积分来 bound 级数，有 $\sum_{i=1}^{n}\frac{1}{\sqrt{i}} \le 1 + \int_1^n \frac{1}{\sqrt{x}}dx = 2\sqrt{n}$ 。

$$\Big|\sum_{h=1}^{H}\sum_{k=1}^{K}\xi_{h+1}^k\Big| = \Big|\sum_{h=1}^{H}\sum_{k=1}^{K}[(\mathbb{P}_h - \hat{\mathbb{P}}_h^k)(V_{h+1}^\star - V_{h+1}^k)](x_h^k, a_h^k)\Big| \le cH\sqrt{T\iota}.$$

注意到 ξ 是 martingale difference sequence，使用 Azuma-Hoeffding 就可以得到上式。

最后得到：

In sum, we have $\sum_{k=1}^{K}\delta_1^k \le O\big(H^2SA + \sqrt{H^4SAT\iota}\big)$, with probability at least $1-2p$.

Azuma-Hoeffding Inequality

Theorem 1.1. *(Azuma-Hoeffding) Let S_n be a martingale (relative to some sequence $Y_0, Y_1, \ldots$) satisfying $S_0 = 0$ whose increments $\xi_n = S_n - S_{n-1}$ are bounded in absolute value by 1. Then for any $\alpha > 0$ and $n \ge 1$,*

(1) $$P\{S_n \ge \alpha\} \le \exp\{-\alpha^2/2n\}.$$

More generally, assume that the martingale differences ξ_k satisfy $|\xi_k| \le \sigma_k$. Then

(2) $$P\{S_n \ge \alpha\} \le \exp\left\{-\alpha^2/2\sum_{j=1}^{n}\sigma_j^2\right\}.$$

各种 paper 里面比较喜欢写成如下形式：

With probability at least $1-p$ ，$|S_n| \le \sqrt{2\log(\frac{2}{p})\sum_{j=1}^{n}\sigma_j^2}$ 。

Regret 和 PAC Guarantee 的关系

前面说到分析 regret 和分析 PAC（或者应该也可以叫做 finite sample analysis）是等价的。这里给出具体的等价关系。

PAC 分析的是需要多少样本能够找到一个 ϵ -optimal 的策略，即 $V_1^*(x_1) - V_1^\pi(x_1) \le \epsilon$ ； regret 分析的结论通常可以表示为 $\sum_{k=1}^{K}[V_1^*(x_1) - V_1^{\pi_k}(x_1)] \le C\cdot T^{1-\alpha} = CHKT^{-\alpha}$ 。

- 从 regret 到 PAC：随机取一个 $\pi = \pi_k, k \in [K]$ ，大概率地有 $V_1^*(x_1) - V_1^\pi(x_1) \le nCHT^{-\alpha}, n > 1$ ，因此找到一个 ϵ -optimal 的策略需要的样本大概是 $T = O\left(\left(\frac{CH}{\epsilon}\right)^{1/\alpha}\right)$ 。比如文章的 UCB-H、UCB-B 的 sample complexity 分别是 $\tilde{O}(H^5SA/\epsilon^2)$ 和 $\tilde{O}(H^4SA/\epsilon^2)$ 。
- 从 PAC 到 regret：如果用 $T_1 = C\epsilon^{-\beta}$ 的样本找到了一个 ϵ -optimal 的策略，那么可以用剩下的 $T - T_1$ 步继续这个策略，从而得到总 regret = $O(T_1 + \epsilon(T - T_1)/H)$ ，选择合适的 T_1 使得 regret 最小，可以得到最小的 regret 为 $O(C^{1+\beta}(T/H)^{\beta/1+\beta})$ 。比如，如果 sample complexity $\propto 1/\epsilon^4$ ，则 regret $\propto T^{4/5}$ 。

疑问

这篇文章的推导总感觉有点问题（也可能是我自己搞错了），在这里记录一下，如果之后需要用，需要核实一下。同时有些地方也怪怪的，也记录一下，有空再想想。

- 关于 ι ：产生在 Lemma 4.3 的 Azuma-Hoeffding + union bound 中，我得到的 $\iota = \log(SAH/p)$ 不含 T，文中是 $\iota = \log(SAT/p)$ ，含有 T 的，并且文章后面多次出现，应该不是打印错了。
- Theorem 1 证明的最后面：得到 $O(H^2SA + \sqrt{H^4SAT\iota})$ 之后，文章里面下面一段话完全没理解。要是令 $H^2SA = \sqrt{H^4SAT\iota}$ 得到的结果也是 $T = SA/\iota$ 啊。

> This establishes $\sum_{k=1}^K \delta_1^k \le O\big(H^2SA + \sqrt{H^4SAT\iota}\big)$. We note that when $T \ge \sqrt{H^4SAT\iota}$, we have $\sqrt{H^4SAT\iota} \ge H^2SA$, and when $T \le \sqrt{H^4SAT\iota}$, we have $\sum_{k=1}^K \delta_1^k \le HK = T \le \sqrt{H^4SAT\iota}$. Therefore, we can remove the H^2SA term in the regret upper bound.

- 关于最后的结论：文章假定了 $R_{\max} = 1$ ，其实最后的 regret 中至少应该有一个 $H \to HR_{\max}$ 。
- 文章的设定也比较奇怪：
  - 首先，使用 finite horizon + undiscounted cumulative reward，这一点也不算太奇怪，但是比较奇怪的是默认当前的步数 h 也是状态的一部分；这相当于每回合 H 步都不可能有重复的状态，并且每『层』的状态空间都是分离的。这一点对于分析有什么特别的影响呢？文章这样的设定看起来更强，因为对于不同的 h 可以有不同的 $\mathbb{P}_h$ 。
  - 其次，一般认为初始状态是从一个分布中采集的，但是这里认为是对手选定的。看起来文章的这种假设更强，是不是这样呢？
- MDP 本身的探索难问题反映在分析的哪个地方：有些 MDP 本身就探索难（某些状态不容易被访问到，比如 Kakede&Langford02 中的第一个例子），如果有些状态学习中一直没有探索到，那么突然遇到时肯定『不知所措』，这时文章假定会遭受一个最大的损失 H （蓝框部分），这产生了 regret 中的常数项（与 T 无关项）。但是下一次再遇到该状态的时候，就不会有这么大的损失了。
- 有些时候 $\mathbb{P}$ 随机性大小不定，如果除了估计均值之外还估计方差，应该可以得到更多信息。这大概是附录里面讲的 Bernstein 探索方法。再往后想，大概是 distributional RL + UCB？

发布于 2019-09-18