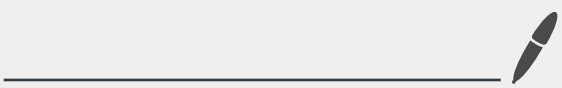
Model Based Reinforcement Learning

Yanjie Ze @CS19 SJTU

Mar 27, 2021



Model Based Reinforcement Learning

What's the difference between model-free and model-based reinforcement learning?

To answer this question, lets revisit the components of an MDP, the most typical decision making framework for RL.

An MDP is typically defined by a 4-tuple (S, A, R, T) where

S is the state/observation space of an environment

A is the set of actions the agent can choose between

$R(s, a)$ is a function that returns the reward received for taking action a in state s

$T(s'|s, a)$ is a transition probability function, specifying the probability that the environment will transition to state s' if the agent takes action a in state s.

Our goal is to find a policy π that maximizes the expected future (discounted) reward.

Now if we know what all those elements of an MDP are, we can just compute the solution before ever actually executing an action in the environment. In AI, we typically call computing the solution to a decision-making problem before executing an actual decision planning. Some classic planning algorithms for MDPs include Value Iteration, Policy Iteration, and whole lot more.

But the RL problem isn't so kind to us. What makes a problem an RL problem, rather than a planning problem, is the agent does *not* know all the elements of the MDP, precluding it from being able to plan a solution. Specifically, the agent does not know how the world will change in response to its actions (the transition function T), nor what immediate reward it will receive for doing so (the reward function R). The agent will simply have to try taking actions in the environment, observe what happens, and somehow, find a good policy from doing so.

So, if the agent does not know the transition function T nor the reward function R, preventing it from planning a solution out, how can it find a good policy? Well, it turns out there are lots of ways!

model based

One approach that might immediately strike you, after framing the problem like this, is for the agent to learn a model of how the environment works from its observations and then plan a solution using that model. That is, if the agent is currently in state $s1$, takes action $a1$, and then observes the environment transition to state $s2$ with reward $r2$, that information can be used to improve its estimate of $T(s2|s1,a1)$ and $R(s1,a1)$, which can be performed using supervised learning approaches. Once the agent has adequately modelled the environment, it can use a planning algorithm with its learned model to find a policy. RL solutions that follow this framework are model-based RL algorithms.

一旦 agent 完全掌握环境，就可以用规划算法

但是我们不一定非要这样

model free

As it turns out though, we don't have to learn a model of the environment to find a good policy. One of the most classic examples is Q-learning, which directly estimates the optimal Q-values of each action in each state (roughly, the utility of each action in each state), from which a policy may be derived by choosing the action with the highest Q-value in the current state. Actor-critic and policy search methods directly search over policy space to find policies that result in better reward from the environment. Because these approaches do not learn a model of the environment they are called model-free algorithms.

So if you want a way to check if an RL algorithm is model-based or model-free, ask yourself this question: after learning, can the agent make predictions about what the next state and reward will be before it takes each action? If it can, then it's a model-based RL algorithm. if it cannot, it's a model-free algorithm.

This same idea may also apply to decision-making processes other than MDPs.

Minimax Regret Bounds for RL

ICML 2017

We consider the problem of provably optimal exploration in reinforcement learning for finite horizon MDPs. We show that an optimistic modification to value iteration achieves a regret bound of $\widetilde{O}(\sqrt{HSAT} + H^2S^2A + H\sqrt{T})$ where H is the time horizon, S the number of states, A the number of actions and T the number of time-steps. This result improves over the best previous known bound $\widetilde{O}(HS\sqrt{AT})$ achieved by the UCRL2 algorithm of Jaksch et al. (2010). The key significance of our new results is that when $T \geq H^3S^3A$ and $SA \geq H$, it leads to a regret of $\widetilde{O}(\sqrt{HSAT})$ that matches the established lower bound of $\Omega(\sqrt{HSAT})$ up to a logarithmic factor. Our analysis contains two key insights. We use careful application of concentration inequalities to the optimal value function as a whole, rather than to the transitions probabilities (to improve scaling in S), and we define Bernstein-based "exploration bonuses" that use the empirical variance of the estimated values at the next states (to improve scaling in H).

- We use careful application of Bernstein and Freedman inequalities (Bernstein, 1927; Freedman, 1975) to the concentration of the *optimal value function* directly, rather than building confidence sets for the transitions probabilities and rewards, like in UCRL2 (Jaksch et al., 2010) and UCFH (Dann & Brunskill, 2015).
- We use empirical-variance exploration bonuses based on Bernstein's inequality, which together with a recursive Bellman-type Law of Total Variance (LTV) provide tight bounds on the expected sum of the variances of the value estimates, in a similar spirit to the analysis from Azar et al. (2013); Lattimore & Hutter (2012).

At a high level, this work addresses the noted shortcomings of existing RL algorithms (Bartlett & Tewari, 2009; Jaksch et al., 2010; Osband & Van Roy, 2016b), in terms of dependency on S and H. We demonstrates that it is possible to design a simple and computationally efficient optimistic algorithm that simultaneously address both the loose scaling in S and H to obtain the first regret bounds that match the $\Omega(\sqrt{HSAT})$ lower bounds as T becomes large.

原文叫 minimax，是因为要求的是 max，
通过 min (max + bonus) 无法逼近到最优

Denote

时间范围 H , State Space S, Action Space A

策略 $\pi : S \times [H] \to A$

- $V_h^\pi(x)$: The value function starting from x in at step h with policy π.
- $V_h^*(x)$: The optimal value function starting from x in at step h. $V_h^*(x) = \max_\pi V_h^\pi(x)$
- $Q_h^\pi(x,a)$: The state-action value function starting from x in at step h with policy π.
- $Q_h^*(x,a)$: The optimal state-action value function starting from x with action a in at step h.
- $V_{k,h}$: Estimated V_h^* at step h in episode k.
- $Q_{k,h}$: Estimated Q_h^* at step h in episode k.

deterministic immediate reward $r_h^\pi(x) = R(x, \pi(x,h))$ $R(x,a) \in [0,1]$

Regret up to time K $R_k = \sum_{k=1}^{K} V_1^*(x_{k,1}) - V_1^{\pi_k}(x_{k,1})$

未知的 environment dynamics $P_h^\pi(y|x) = P(y|x, \pi(x,h))$

预估的 environment dynamics $\hat{P}_{k,h}(y|x,a)$

对未来价值的期望 $PV(x,a) = \sum_{y \in S} P(y|x,a) V(y) = \langle P(y|x,a), V(y) \rangle$

到第 k 轮第 h 步的 (x,a) 采样次数 $N_{k,h}(x,a)$

到第 k 轮第 h 步的 (x,a,y) 采样次数 $N_{k,h}(x,a,y)$

探索过 k 次及以上的 (x,a) pair 的集合 $K = \{(x,a) \in S \times A, N_{k,h}(x,a) > 0\}$

一个常数 C

$$T = H \cdot K = H \cdot N$$

Algorithm : Upper Confidence Bound Value Iteration

Algorithm 1: UCBVI

Initialize data buffer $H = \varnothing$;
for *epoch* $k = 1, 2, ..K$ **do**
  $Q_{k,h} =$ **UCBQ**(H) **for** *step* $h = 1,2,...,H$ **do**
    - Take action $a_{k,h} = \underset{a \in A}{argmax}\ Q_{k,h}(x_{k,h}, a)$
    - Sample $x_{k,h+1}$ from the dynamics of the Environment.
    - $H = H \cup (x_{k,h}, a_{k,h}, x_{k,h+1})$
  end
end

先探索一轮，探索完后进入UCBQ计算Q

Algorithm 2: UCBQ

Data: Data H
Result: Q-value $Q_{k,h}$
Initialize $V_{k,H+1}(x) = 0$ for all $x \in S$;
Estimate $\hat{P}_{k,h}(y|x,a) = \begin{cases} \frac{N_{k,h}(x,a,y)}{N_{k,h}(x,a)}, & \text{for } (x,a) \in K \\ 0, & \text{otherwise} \end{cases}$; 根据采样次数估算 dynamics
Calculate "Bonus" $b_{k,h}(x,a) = \begin{cases} CH\sqrt{\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}, & \text{for } (x,a) \in K \\ H, & \text{otherwise} \end{cases}$; explore 越少 bonus 越大，若没有explore过，bonus 为最大，H
Estimate $Q_{k,h}$ and $V_{k,h}$;
for $h = H, H-1, ..., 1$ **do**
  for $(x,a) \in S \times A$ **do**
    - $Q_{k,h} = \begin{cases} R(x,a) + (\hat{P}_{k,h}V_{k,h+1})(x.a) + b_{k,h}(x,a) & \text{for } (x,a) \in K \\ b_{k,h}(x,a), & \text{otherwise} \end{cases}$ bonus for exploration
    - $V_{k,h} = \max_{a \in A} Q_{k,h}(x,a)$
  end
end
return $Q_{k,h}$

反过来求Q

Regret Bound:

$$O(H^{1.5}|S|\sqrt{|A|T})$$

Prof of the Upper Bound of Regret

Overview

首先定义 Regret：

$$R_K = \sum_{k=1}^{K} V_{k,1}^{*}(x) - V_{k,1}^{\pi_k}(x)$$

其中 $V_{k,1}^{\pi_k}(x)$ 是在第 k 轮的策略 π_k 下，h=1 时的 value

再定义 $\delta_{k,h} = (V_{k,h} - V_h^{\pi_k})(x)$

表示预估的 value 与 π_k 下的 value 之间的差。

总体证明思路是：

证明 6 个 Claim

Claim 1. *Define event* $\mathcal{G} = \left\{ \begin{cases} Q_{k,h}(x,a) \geq Q_h^*(x,a) \\ V_{k,h}(x) \geq V_h^*(x) \end{cases} \quad \forall k,h,x,a \right\}$

$$P(\mathcal{G}) \geq 1 - \delta$$

主要证明预估值比最优的高

Claim 2.

$$R_K = \mathbb{E}\sum_{k=1}^{K}(V_1^* - V_1^{\pi_k})(x) \leq \mathbb{E}\sum_{k=1}^{K}(V_{k,1} - V_1^{\pi_k})(x) + \delta T = \mathbb{E}\sum_{k=1}^{K}\delta_{k,1} + \delta T$$

where $\delta_{k,h} \triangleq (V_{k,h} - V_h^{\pi_k})(x)$

Claim 3. $\mathbb{E}[\delta_{k,h}] = \mathbb{E}\left[\left((\hat{P}_{k,h} - P)V_{k,h+1}\right)(x,a) + b_{k,h}(x,a)\right] + \mathbb{E}[\delta_{k,h+1}]$

Claim 4. $\mathbb{E}\left[\left((\hat{P}_{k,h} - P)V_{k,h+1}\right)(x,a)\right] \leq CH\sqrt{\frac{|S|}{N_{k,h}(x,a)}\ln(\frac{|S||A|T}{\delta})} + \frac{CH|S|}{N_{k,h}(x,a)}\ln(\frac{|S||A|T}{\delta}) + \delta H$

Claim 5.

$$\mathbb{E}[\delta_{k,1}] \leq \mathbb{E}\sum_{h=1}^{H}\left(CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}} + CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right) + \delta H^2$$

Claim 6.

$$R_K \leq \delta T^2 + C|S|^2|A|H^2\log^2(\frac{|S||A|T}{\delta}) + C|S|H^{1.5}\sqrt{|A|T\log(\frac{|S||A|T}{\delta})}$$

Claim 1: Define $G = \left\{ \begin{matrix} Q_{k,h}(x,a) \geq Q_h^*(x,a) \\ V_{k,h}(x) \geq V_h^*(x) \end{matrix} \quad \forall\, k,h,x,a \right\}$

$$P(G) \geq 1-\delta$$

Proof: 用 backward Induction 来证明.

注: backward Induction 后向归纳法

- Base case:

For $\hat{h} = H$,

- $Q_{k,h} = \begin{cases} R(x,a) + (\hat{P}_{k,h}V_{k,h+1})(x,a) + b_{k,h}(x,a) & \text{for } (x,a) \in K \\ b_{k,h}(x,a), & \text{otherwise} \end{cases}$ $= H$
- $V_{k,h} = \max_{a \in A} Q_{k,h}(x,a)$

$$b_{k,h}(x,a) = \begin{cases} CH\sqrt{\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}, & \text{for } (x,a) \in K \\ H, & \text{otherwise} \end{cases}$$

$$\therefore\ Q_{k,H}(x,a) = \begin{cases} R(x,a) + b_{k,H}(x,a) & \text{for } (x,a) \in K \\ H & \text{otherwise} \end{cases}$$

$$\text{又: } Q_H^*(x,a) = R(x,a) \quad \in [0,1]$$

$$\therefore\ Q_{k,H}(x,a) \geq Q_{k,H}^*(x,a)$$

$$V_{k,H}(x) = \max_a Q_{k,H}(x,a) \geq \max_a Q_{k,H}^*(x,a) = V_H^*(x)$$

Base Case 证毕.

- Assume for $\hat{h} = H, H-1, \cdots, h+1$ hold

$$P(G) \geq 1-\delta \quad \text{, where } G: \begin{matrix} Q \geq Q^* \\ V \geq V^* \end{matrix}$$

- Now to prove $\hat{h} = h$ holds:

$$Q_{k,h}(x,a) - Q_h^*(x,a) = \left[(\hat{P}_{k,h}V_{k,h+1})(x,a) + R(x,a) + b_{k,h}(x,a) \right] - \left[(PV_{h+1}^*)(x,a) + R(x,a) \right]$$

减去和加上一项

$$= b_{k,h}(x,a) + \left[(\hat{P}_{k,h} - P)V_{h+1}^* \right](x,a) + \underbrace{\left[\hat{P}_{k,h}(V_{k,h+1} - V_{h+1}^*) \right](x,a)}$$

By induction, $V_{k,h+1} - V_{h+1}^* \geq 0$

So this part ≥ 0

Now let's prove $b_{k,h}(x,a) + \left[(\hat{P}_{k,h} - P)V^*_{h+1}\right](x,a) \geq 0$ with high probability

Define events: $\mathcal{E} = \left\{ \left| \left[(\hat{P}_{k,h} - P)V^*_{h+1}\right](x,a) \right| \leq b_{k,h}(x,a), \forall k,h,x,a \right\}$

$$\mathcal{E}_{k,h,x,a} = \left\{ \left| \left[(\hat{P}_{k,h} - P)V^*_{h+1}\right](x,a) \right| \leq b_{k,h}(x,a) \right\}$$

$\mathcal{E}^c$, $\mathcal{E}^c_{k,h,x,a}$ 代表这两个事件的补集。

引入 Theorem 1: Union Bound

For $A_1, A_2, \cdots, A_n$,

$$P\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n P(A_i)$$

注：容斥原理：

$$\left|\bigcup_{i=1}^n A_i\right| = \sum_{\phi \neq I \subseteq \{1,2,\cdots,n\}} (-1)^{|I|-1} \left|\bigcap_{i \in I} A_i\right|$$

所以：

$$\mathcal{E} = \bigcap_{k,h,x,a} \mathcal{E}_{k,h,x,a}$$

$$\Rightarrow \mathcal{E}^c = \bigcup_{k,h,x,a} \mathcal{E}^c_{k,h,x,a}$$

使用 Theorem 1：

$$P(\mathcal{E}^c) \leq \sum_{k,h,x,a} P(\mathcal{E}^c_{k,h,x,a}) \approx |S||A| \cdot H \cdot N \cdot P(\mathcal{E}^c_{k,h,x,a})$$

再算下这个就行了

引入 Theorem 2: Hoeffding Inequality.

$Z_1, Z_2, \cdots Z_n \in [a,b]$.

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n (Z_i - E(Z_i))\right| \geq t\right) \leq 2e^{-\frac{2nt^2}{(b-a)^2}}$$

$$P(\mathcal{E}_{k,h}^c(x,a)) = P\left(\left|\left[(\hat{P}_{k,h}-P)V_{h+1}^*\right](x,a)\right| \geq b_{k,h}(x,a)\right)$$

PV的定义

$$= P\left(\left|\sum_{y\in S}\hat{P}_{k,h}(y|x,a)V_{h+1}^*(y) - \sum_{y\in S}P(y|x,a)V_{h+1}^*(y)\right| \geq b_{k,h}(x,a)\right)$$

令 $Z = \sum_{y\in S}\hat{P}_{k,h}(y|x,a)V_{h+1}^*(y)$，那么 $E(Z) = \sum_{y\in S}P(y|x,a)V_{h+1}^*(y)$

使用 Theorem 2：

$$\text{上式} = P(|Z-E(Z)| \geq b_{k,h}(x,a))$$

$$\leq 2e^{\frac{-2\cdot b_{k,h}^2(x,a)}{H^2}\cdot N_{k,h}(x,a)}$$

（$N_{k,h}(x,a)$：采样次数）

$$\therefore P(\mathcal{E}) = 1-P(\mathcal{E}^c) \geq 1-|S||A|HN\cdot 2e^{\frac{-2b_{k,h}^2(x,a)}{H^2}\cdot N_{k,h}(x,a)}$$

得证 Claim 1

$\Rightarrow \delta$

这与：$P(G) \geq 1-\delta$ 等价

此外：

由

$$\delta = |S||A|HN\cdot 2e^{\frac{-2b_{k,h}^2(x,a)}{H^2}\cdot N_{k,h}(x,a)}$$

解得

$$b_{k,h} = \underbrace{CH}_{\text{常数项}}\cdot\sqrt{\frac{\ln\left(\frac{|S||A|HN}{\delta}\right)}{N_{k,h}(x,a)}}$$

即 bonus 项

Claim 2.

$$E[R_K] = E\sum_{k=1}^{K}(V_1^* - V_{k,1}^{\pi_k})(x) \le E\sum_{k=1}^{K}(V_{k,1} - V_{k,1}^{\pi_k})(x) + \delta T$$

$$= E\sum_{k=1}^{K}\delta_{k,1} + \delta T$$

其中，$\delta_{k,h} \triangleq (V_{k,h} - V_h^{\pi_k})(x)$

Proof:

由 Claim1，我们知道：$P(G) \ge 1-\delta$，也可以说是 $P(\mathcal{E}) \ge 1-\delta$

因此

$$E[R_K] = E[R_K|\mathcal{E}] \times (1-\delta) + E[R_K|\mathcal{E}^c] \times \delta$$

$$= E[R_K|\mathcal{E}] + \delta(E[R_K|\mathcal{E}^c] - E[R_K|\mathcal{E}])$$

而

① $E[R_K|\mathcal{E}] = E[R_K|\{V_{k,h} \ge V^*\}]$

$$\le E\sum_{k=1}^{K}(V_{k,1} - V_{k,1}^{\pi_k})(x) = E\sum_{k=1}^{K}\delta_{k,1}$$

② $E[R_K|\mathcal{E}^c] - E[R_K|\mathcal{E}] \le H \times K = T$

所以

$$E[R_K] \le E\sum_{k=1}^{K}\delta_{k,1} + \delta T$$

Claim 3:

$$E[\delta_{k,h}] = E[((\hat{P}_{k,h} - P)V_{k,h+1})(x,a) + b_{k,h}(x,a)] + E[\delta_{k,h+1}]$$

Proof:

根据定义，有：

$$\delta_{k,h} \triangleq (V_{k,h} - V_h^{\pi_k})(x) \overset{\text{取action } a}{=} (Q_{k,h} - Q^{\pi_k})(x,a)$$

现

$$\delta_{k,h} = (Q_{k,h} - Q^{\pi_k})(x,a)$$

$$= Q_{k,h}(x,a) - Q^{\pi_k}(x,a)$$

根据

- $Q_{k,h} = \begin{cases} R(x,a) + (\hat{P}_{k,h}V_{k,h+1})(x.a) + b_{k,h}(x,a) & \text{for } (x,a) \in K \\ b_{k,h}(x,a), & \text{otherwise} \end{cases}$
- $V_{k,h} = \max_{a \in A} Q_{k,h}(x,a)$

$$= \hat{P}_{k,h}V_{k,h+1}(x,a) + b_{k,h}(x,a) - PV_{k,h+1}^{\pi_k}(x)$$

加一项减一项

$$= \hat{P}_{k,h}V_{k,h+1}(x,a) + b_{k,h}(x,a) - PV_{k,h+1}^{\pi_k}(x) - PV_{k,h+1}(x,a) + PV_{k,h+1}(x,a)$$

$$= ((\hat{P}_{k,h} - P)V_{k,h+1})(x,a) + b_{k,h}(x,a) + \underbrace{PV_{k,h+1}(x,a) - PV_{k,h+1}^{\pi_k}(x)}_{\delta_{k,h+1}}$$

因此，

$$E[\delta_{k,h}] = E[((\hat{P}_{k,h} - P)V_{k,h+1})(x,a) + b_{k,h}(x,a)] + E[\delta_{k,h+1}]$$

Claim 4

$$E\left[\left((\hat{P}_{k,h}-P)V_{k,h+1}\right)(x,a)\right] \le CH\sqrt{\frac{|S|}{N_{k,h}(x,a)}\ln\left(\frac{|S||A|T}{\delta}\right)} + E[\delta_{k,h+1}]$$

Proof:

① 首先有：

$$E\left[\left((\hat{P}_{k,h}-P)V_{k,h+1}\right)(x,a)\right] \le E\left[\left|\left((\hat{P}_{k,h}-P)V_{k,h+1}\right)(x,a)\right|\right]$$

然后：

$$\left|\left((\hat{P}_{k,h}-P)V_{k,h+1}\right)(x,a)\right| = \left|\sum_{y\in S}\hat{P}_{k,h}(y|x,a)V_{k,h+1}(y) - P(y|x,a)V_{k,h+1}(y)\right|$$

求和号提出来

$$\le \sum_{y\in S}\left|\hat{P}_{k,h}(y|x,a)V_{k,h+1}(y) - P(y|x,a)V_{k,h+1}(y)\right|$$

$V \le H$

$$\le H\sum_{y\in S}\left|\hat{P}_{k,h}(y|x_{k,h},a) - P(y|x_{k,h},a)\right|$$

② 引入 Theorem 3: Bernstein inequality

Theorem 3: Bernstein inequality

Let $Z_1, Z_2, \ldots, Z_n$ be independent bounded random variables with $|Z_i| \le M$ for all i. Then we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - E[Z_i])\right| \ge t\right) \le 2e^{-\frac{\frac{1}{2}(nt)^2}{\sum_{i=1}^{n}E[(Z_i-E[Z_i])^2]+\frac{1}{3}Mnt}}$$

When Z_i are i.i.d. random variable, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - E[Z_i])\right| \ge t\right) \le 2e^{-\frac{\frac{1}{2}nt^2}{Var(Z_i)+\frac{1}{3}Mt}}$$

When $Z_i \in [0,1]$, we have $E[Z_i] \ge E[Z_i^2] \ge Var(Z_i)$.Thus, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - E[Z_i])\right| \ge t\right) \le 2e^{-\frac{\frac{1}{2}nt^2}{E(Z_i)+\frac{1}{3}t}}$$

（用的是这个）

因此，根据 Bernstein Inequality：

来 bound $|\hat{P}_{k,h}(y|x_{k,h},a) - P(y|x_{k,h},a)|$：

$$P\left(|\hat{P}_{k,h}(y|x_{k,h},a) - P(y|x_{k,h},a)| \geq t\right) \leq 2e^{-\frac{\frac{1}{2}N_{k,h}(x,a)t^2}{u+\frac{1}{3}t}}$$

其中，$u = E[\hat{P}_{k,h}(y|x_{k,h},a)] = P(y|x_{k,h},a)$

这里的 δ 当作 claim 中的 δ 吧？

再让 $\frac{\delta}{|S||A|T} = e^{-\frac{\frac{1}{2}N_{k,h}(x,a)t^2}{u+\frac{1}{3}t}}$，可以求得 t：

$$t = C_1\frac{\ln(\frac{\|S||A|T}{\delta})}{N_{k,h}(x,a)} + \sqrt{C_2\frac{\ln(\frac{|S||A|T}{\delta})\mu}{N_{k,h}(x,a)} + C_3\left(\frac{\ln(\frac{|S||A|T}{\delta})\mu}{N_{k,h}(x,a)}\right)^2} \approx C\left(\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)} + \sqrt{\frac{\ln(\frac{|S||A|T}{\delta})\mu}{N_{k,h}(x,a)}}\right)$$

因为 $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$，

我们可以让 C 足够大，把 $\approx$ 变成 $\leq$

③ 现在我们定义新的事件：

$$\mathcal{F}_{k,h,s,a} \triangleq \{\left|\hat{P}_{k,h}(y|x,a) - P(y|x,a)\right| \leq C\left(\sqrt{\frac{\ln(\frac{|S||A|T}{\delta})P(y|s,a)}{N_{k,h}(x,a)}} + \frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right)\}$$

$$\mathcal{F} \triangleq \{\forall k,h,s,a\left|\hat{P}_{k,h}(y|x,a) - P(y|x,a)\right| \leq C\left(\sqrt{\frac{\ln(\frac{|S||A|T}{\delta})P(y|s,a)}{N_{k,h}(x,a)}} + \frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right)\}$$

由于上面已经让 $\frac{\delta}{|S||A|T} =$ ，所以：

$$P(F^c_{k,h,s,a}) \leq \frac{\delta}{|S||A|T}$$

$\because F = \bigcap F_{k,h,s,a}$

$F^c = \bigcup F^c_{k,h,s,a}$

$\therefore P(F^c) = P(\bigcup F^c_{k,h,s,a}) \le |S||A|T \cdot P(F^c_{k,h,s,a})$

$\le |S||A|T \cdot \frac{\delta}{|S||A|T}$

$= \delta$

$\therefore P(F) \ge 1-\delta$

④ 利用 ① 和 ③

$$H \sum_{y \in S} \left| \hat{P}_{k,h}(y|x_{k,h},a) - P(y|x_{k,h},a) \right|$$

$$\le H \sum_{y \in S} C\left(\sqrt{\frac{\ln(\frac{|S||A|T}{\delta}) P(y|s,a)}{N_{k,h}(x,a)}} + \frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)} \right) + H\delta$$

因为 F 不满足时 差 ≤ 1

⑤

引入 Theorem 4:

Theorem 4: Cauchy–Schwarz inequality

For vectors u and v, we have

$$\left|\sum_i^n u_i v_i\right| \le ||u|| \cdot ||v||$$

根据 Cauchy, 我们有

$$\sum_{y \in S} \sqrt{P(y|x,a)} \le \sqrt{\sum_{y \in S} P(y|x,a) \sum_{y \in S} 1} = \sqrt{|S|}$$

Then, for inequality (4), we have

$$H\sum_{y\in S}\left|\hat{P}_{k,h}(y|x,a)-P(y|x,a)\right| \le H\sum_{y\in S}C\Big(\sqrt{\frac{\ln(\frac{|S||A|T}{\delta})P(y|s,a)}{N_{k,h}(x,a)}}+\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\Big)+H\delta \tag{5}$$

$$\le CH\sqrt{\frac{\ln(\frac{|S||A|T}{\delta})|S|}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}+H\delta \tag{6}$$

Thus, we have

$$\mathbb{E}\left[\Big((\hat{P}_{k,h}-P)V_{k,h+1}\Big)(x,a)\right] \le \mathbb{E}\left[CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right]+\delta H$$

□

Claim 5

根据 Claim. 3. 4 . 可以得到 Claim 5

Claim 5.

$$\mathbb{E}[\delta_{k,1}] \leq \mathbb{E}\sum_{h=1}^{H}\left(CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right)+\delta H^2$$

Proof. Given by Claim 3, we have

$$\mathbb{E}[\delta_{k,h}]-\mathbb{E}[\delta_{k,h+1}]=\mathbb{E}\left[\left((\hat{P}_{k,h}-P)V_{k,h+1}\right)(x,a)+b_{k,h}(x,a)\right]$$

We know

$$\begin{aligned}
&\mathbb{E}\left[\left((\hat{P}_{k,h}-P)V_{k,h+1}\right)(x,a)+b_{k,h}(x,a)\right]\\
\leq&\mathbb{E}\left[CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}+b_{k,h}(x,a)\right]+\delta H\\
\leq&\mathbb{E}\left[CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}+CH\sqrt{\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}\right]+\delta H\\
\approx&\mathbb{E}\left[CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right]+\delta H
\end{aligned}$$

Note that the last approximate equality comes from the fact the bonus function has the same order as the first term. With appropriate select C, the equality can hold.

We Telescoping Sum, we have

$$\begin{aligned}
\mathbb{E}[\delta_{k,1}]-\mathbb{E}[\delta_{k,2}]&\leq\mathbb{E}\left[CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right]+\delta H\\
\mathbb{E}[\delta_{k,2}]-\mathbb{E}[\delta_{k,3}]&\leq\mathbb{E}\left[CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right]+\delta H\\
&\vdots\\
\mathbb{E}[\delta_{k,H}]-\mathbb{E}[\delta_{k,H+1}]&\leq\mathbb{E}\left[CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right]+\delta H
\end{aligned}$$

where $\mathbb{E}[\delta_{k,H+1}]=0$. We have

$$\mathbb{E}[\delta_{k,1}] \leq \mathbb{E}\sum_{h=1}^{H}\left(CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right)+\delta H^2$$

□

Claim 6

$$R_K \le \delta T^2 + C|S|^2|A|H^2 \log^2(\frac{|S||A|T}{\delta}) + C|S|H^{1.5}\sqrt{|A|T\log(\frac{|S||A|T}{\delta})}$$

Proof. We know

$$\begin{aligned}
R_K &\le \mathbb{E}\sum_{k=1}^{K}\delta_{k,1} + \delta T \\
&\le \mathbb{E}\sum_{k=1}^{K}\left[\sum_{h=1}^{H}\left(CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}} + CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right) + \delta H^2\right] + \delta T \\
&\le \mathbb{E}\sum_{k=1}^{K}\sum_{h=1}^{H}\left(CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}} + CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right) + \delta T^2 \\
&= \mathbb{E}\sum_{x,a,h}\sum_{n=1}^{N_{k,h}(x,a)}\left(CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{n}} + CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{n}\right) + \delta T^2
\end{aligned}$$

Using the following three facts, we can bound the term above.

Fact 1.

$$\sum_{k=1}^{n}\sqrt{\frac{1}{k}} < 2\sqrt{n}$$

Proof.

$$\frac{1}{\sqrt{k}} - 2(\sqrt{k} - \sqrt{k-1})) = \frac{1}{\sqrt{k}} - \frac{2}{(\sqrt{k}+\sqrt{k-1})} < \frac{1}{\sqrt{k}} - \frac{2}{(\sqrt{k}+\sqrt{k})} = 0$$

Thus, we have

$$\frac{1}{\sqrt{k}} < 2(\sqrt{k} - \sqrt{k-1})$$

Therefore, we have

$$\sum_{k=1}^{n}\sqrt{\frac{1}{k}} < 2\sqrt{n}$$

□

Fact 2. (It is trial that summation is smaller than integral.)

$$\sum_{k=1}^{n}\frac{1}{k} \approx \ln(n)$$

Fact 3.(Cauchy-Schwarz inequality)

$$\sum_{s,a,h}\sqrt{N_{k,h}} \le \sqrt{\sum_{s,a,h}N_{k,h}}\sqrt{\sum_{s,a,h}1} = \sqrt{T}\sqrt{|S||A|H}$$

Then

$$\begin{aligned}
&\mathbb{E}\sum_{x,a,h}\sum_{n=1}^{N_{k,h}(x,a)}\left(CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{n}} + CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{n}\right) + \delta T^2 \\
\le &\mathbb{E}\Big[\sum_{x,a,h}\left(CH\sqrt{|S|N_{k,h}(x,a)\ln(\frac{|S||A|T}{\delta})} + CH|S|\ln^2(\frac{|S||A|T}{\delta})\right)\Big] + \delta T^2 \\
= &\mathbb{E}\Big[\sum_{x,a,h}CH\sqrt{|S|N_{k,h}(x,a)\ln(\frac{|S||A|T}{\delta})}\Big] + CH^2|A||S|^2\ln^2(\frac{|S||A|T}{\delta}) + \delta T^2 \\
\le &CH\sqrt{T|S|^2|A|H\ln(\frac{|S||A|T}{\delta})} + CH^2|A||S|^2\ln^2(\frac{|S||A|T}{\delta}) + \delta T^2 \\
= &CH^{1.5}|S|\sqrt{T|A|\ln(\frac{|S||A|T}{\delta})} + CH^2|A||S|^2\ln^2(\frac{|S||A|T}{\delta}) + \delta T^2
\end{aligned}$$

□

If we choose $\delta = \frac{1}{T^2}$, then the regret R_k is bounded by $\mathcal{O}(H^{1.5}|S|\sqrt{|A|T})$.