MVA: Reinforcement Learning (2022/2023) **Assignment 3**

Exploration in Reinforcement Learning (theory)

Lecturers: *M. Pirotta* *(December 12, 2022)*

Solution by Denis Duval

Instructions

- The deadline is **January 20, 2023. 23h59**
- By doing this homework you agree to the *late day policy, collaboration and misconduct rules* reported on Piazza.
- **Mysterious or unsupported answers will not receive full credit**. A correct answer, unsupported by calculations, explanation, or algebraic work will receive no credit; an incorrect answer supported by substantially correct calculations and explanations might still receive partial credit.
- Answers should be provided in **English**.

1 Best Arm Identification

In best arm identification (BAI), the goal is to identify the best arm in as few samples as possible. We will focus on the fixed-confidence setting where the goal is to identify the best arm with high probability $1-\delta$ in as few samples as possible. A player is given k arms with expected reward μ_i. At each time step t, the player selects an arm to pull (I_t), and they observe some reward ($X_{I_t,t}$) for that sample. At any time step, once the player is confident that they have identified the best arm, they may decide to stop.

δ-correctness and fixed-confidence objective. Denote by τ_δ the stopping time associated to the stopping rule, by $i^\star$ the best arm and by $\widehat{i}$ an estimate of the best arm. An algorithm is δ-correct if it predicts the correct answer with probability at least $1-\delta$. Formally, if $\mathbb{P}_{\mu_1,\dots,\mu_k}(\widehat{i} \neq i^\star) \leq \delta$ and $\tau_\delta < \infty$ almost surely for any $\mu_1, \dots, \mu_k$. Our goal is to find a δ-correct algorithm that minimizes the sample complexity, that is, $\mathbb{E}[\tau_\delta]$ the expected number of sample needed to predict an answer. Assume that the best arm $i^\star$ is *unique* (i.e., there exists only one arm with maximum mean reward).

Notation

- I_t: the arm chosen at round t.
- $X_{i,t} \in [0,1]$: reward observed for arm i at round t.
- μ_i: the expected reward of arm i.
- $\mu^\star = \max_i \mu_i$.
- $\Delta_i = \mu^\star - \mu_i$: suboptimality gap.

Consider the following algorithm
The algorithm maintains an active set S and an estimate of the empirical reward of each arm $\widehat{\mu}_{i,t} = \frac{1}{t}\sum_{j=1}^{t} X_{i,j}$.

- Compute the function $U(t,\delta)$ that satisfy the any-time confidence bound. Let

$$\mathcal{E} = \bigcup_{i=1}^{k} \bigcup_{t=1}^{\infty} \left\{ |\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta') \right\}.$$

```
Input: $k$ arms, confidence $\delta$
$S = \{1, \dots, k\}$
for $t = 1, \dots$ do
    Pull all arms in $S$
    $S = S \setminus \left\{ i \in S \,:\, \exists j \in S,\ \widehat{\mu}_{j,t} - U(t, \delta') \geq \widehat{\mu}_{i,t} + U(t, \delta') \right\}$
    if $|S| = 1$ then
        STOP
        return $S$
    end
end
```

Using Hoeffding's inequality and union bounds, shows that $\mathbb{P}(\mathcal{E}) \leq \delta$ for a particular choice of δ'. This is called "bad event" since it means that the confidence intervals do not hold.

- Show that with probability at least $1-\delta$, the optimal arm $i^\star = \arg\max_i\{\mu_i\}$ remains in the active set S. Use your definition of δ' and start from the condition for arm elimination. From this, use the definition of $\neg\mathcal{E}$.
- Under event $\neg\mathcal{E}$, show that an arm $i \neq i^\star$ will be removed from the active set when $\Delta_i \geq C_1 U(t, \delta')$ for some constant $C_1 \in \mathbb{N}$. Compute the time required to have such condition for each non-optimal arm. Use the condition of arm elimination applied to arm $i^\star$.[1]
- Compute a bound on the sample complexity (after how many *pulls* the algorithm stops) for identifying the optimal arm w.p. $1-\delta$.
- We assumed that the optimal arm $i^\star$ is unique. Would the algorithm still work if there exist multiple best arms? Why?

Note that also a variations of UCB are effective in pure exploration.

1.1 Answers for BAI

- We want to find a function $U(t, \delta)$ s.t. $\mathbb{P}(|\widehat{\mu}_{i,t} - \mu_i| > U(t, \delta)) \leq \delta/2t^2$. We can use Hoeffding's inequality which stands because all $X_{i,t}/t$ are in $[0, 1/t]$:

$$\mathbb{P}(|\widehat{\mu}_{i,t} - \mu_i| > U(t, \delta)) \leq 2\exp(-2tU(t, \delta)^2).$$

Setting $2\exp(-2tU(t, \delta)^2) = \delta/2t^2$ we get $\boxed{U(t, \delta) = \sqrt{\frac{\log(4t^2/\delta)}{2t}}}$ Furthermore, we can show that for a certain choice of δ' we have $\mathbb{P}(\mathcal{E}) \leq \delta$.:

$$\mathbb{P}(\mathcal{E}) = \mathbb{P}(\bigcup_{i=1}^{k} \bigcup_{t=1}^{\infty} |\widehat{\mu}_{i,t} - \mu_i| > U(t, \delta'))$$

using union bound and the any-time confidence property of $U(t, \delta)$ we have derived above:

$$\mathbb{P}(\mathcal{E}) \leq k \sum_{t \geq 1} \delta'/2t^2 = k\delta'\pi^2/12 \leq k\delta'$$

By setting $\delta' = \delta/k$ we have the desired result.

[1] Note that $at \geq \log(bt)$ can be solved using Lambert W function. We thus have $t \geq \frac{-W_{-1}(-a/b)}{a}$ since, given $a = \Delta_i^2$ and $b = 2k/\delta$, $-a/b \in (-1/e, 0)$. We can make the bound more explicit by noticing that $-1 - \sqrt{2u} - u \leq W_{-1}(-e^{-u-1}) \leq -1 - \sqrt{2u} - 2u/3$ for $u > 0$ [Chatzigeorgiou, 2016]. Then $t \geq \frac{1+\sqrt{2u}+u}{a}$ with $u = \log(b/a) - 1$.

- If we use the same δ' as before we have $\mathbb{P}(\neg\mathcal{E}) \geq 1-\delta$. The event writes:

$$\bigcap_{t\geq 1}\bigcap_{i=1}^{k}(|\widehat{\mu}_{i,t}-\mu_i| < U(t,\delta'))$$

$$= \bigcap_{t\geq 1}\bigcap_{i=1}^{k}(|\widehat{\mu}_{i,t}-\mu_i| < U(t,\delta'))\bigcap(|\widehat{\mu}_{i^\star,t}-\mu_{i^\star}| < U(t,\delta'))$$

However $(|\widehat{\mu}_{i,t}-\mu_i| < U(t,\delta'))\bigcap(|\widehat{\mu}_{i^\star,t}-\mu_{i^\star}| < U(t,\delta'))$ is included in the event $(\widehat{\mu}_{i,t}-\mu_i+\mu_{i^\star}-\widehat{\mu}_{i^\star,t} \leq 2U(t,\delta'))$
Furthermore, at a given time t, $i^\star$ remains in S if and only if for all j (that are still in S at time t):

$$\widehat{\mu}_{j,t}-\widehat{\mu}_{i^\star,t} \leq 2U(t,\delta')$$

Using that $-\mu_i+\mu_{i^\star}$ is positive for all i (definition of the best arm) we can write:

$$\neg\mathcal{E} \subset \bigcap_{t\geq 1}\bigcap_{i=1}^{k}(\widehat{\mu}_{j,t}-\widehat{\mu}_{i^\star,t} \leq 2U(t,\delta')) \subset A$$

with A being the event we want to lower bound the probability. It follows that $\boxed{\mathbb{P}(A) \geq 1-\delta}$. (Actually, the event A is not necessarily equal to the event on its left because the intersection is on the arms that are still in the set at a given time t).

- One event that can eliminate i is $(\widehat{\mu}_{i^\star}-U(t,\delta') \geq \widehat{\mu}_{i,t}+U(t,\delta'))$. Using that event $\neg\mathcal{E}$ is happening, we have $\widehat{\mu}_{i^\star,t} \geq \mu_{i^\star}-U(t,\delta')$ and $\widehat{\mu}_{i,t} \leq \mu_i+U(t,\delta')$. So we can see that if we have

$$\mu_{i^\star}-2U(t,\delta') \geq \mu_i+2U(t,\delta')$$

the elimination inequality would immediately follow. This writes :

$$\boxed{\Delta_i \geq 4U(t,\delta')}$$

so $C_1=4$. We can then upper bound the time T_i when i (a non-optimal arm) will be eliminated by solving:

$$\begin{aligned}\Delta_i &\geq 4U(t,\delta')\\ \Leftrightarrow 2(\Delta_i/4)^2 t &\geq \log(4t^2k/\delta)\\ \Leftrightarrow (\Delta_i/4)^2 t &\geq \log(t2\sqrt{k/\delta})\\ \Leftrightarrow at &\geq \log(bt)\end{aligned}$$

with $a=(\Delta_i/4)^2$ and $b=2\sqrt{k/\delta}$. Using the hint in the footnote we have:

$$t \geq \frac{1+\sqrt{2u}+u}{a}$$

with $u=\log(b/a)-1$.. T_i can be upper bounded by the minimum t that satisfies the above inequality which means:

$$\boxed{T_i \leq \frac{1+\sqrt{2u}+u}{a}}$$

.

- We can then find a bound on the sample complexity w.p. $1-\delta$. Assuming again that $\neg\mathcal{E}$ is happening we can say that the sample complexity C verifies:

$$\boxed{C=\mathbb{E}[\max_{i\neq i^\star} T_i] = \mathcal{O}(\sum_{i\neq i^\star}\frac{\log(\frac{\sqrt{k/\delta}}{\Delta_i^2})}{\Delta_i^2})}$$

- First if there are multiple optimal arms, they would all remain in the active set (the setting is unchanged) w.p. $1-\delta$. In fact the proof that we did still applies because the only difference is that $-\mu_i + \mu_{i^\star}$ could be zero but it doesn't change the validity of the proof. Also, the non-optimal arms would be eliminated, the proof of this is still valid as well. Now the problem is when to stop. In fact, if we don't know that there are multiple optimal arms then we may have an active set that will remain the same at some point (with a cardinality greater than one, containing all the optimal arms). If we know the number of optimal arms, then we could adapt the stopping condition of the algorithm, and the bounds on the sample complexity would be still valid. If we don't have the exact number, we may have a knowledge of how many arms at most are optimal so we could again adapt the stopping condition. However if we don't have any knowledge then we may for example set a stopping condition "if the set hasn't changed for n iterations" where n is to choose.

2 Regret Minimization in RL

Consider a finite-horizon MDP $M^\star = (S, A, p_h, r_h)$ with stage-dependent transitions and rewards. Assume rewards are bounded in $[0,1]$. We want to prove a regret upper-bound for UCBVI. We will aim for the suboptimal regret bound ($T = KH$)

$$R(T) = \sum_{k=1}^{K} V_1^\star(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) = \widetilde{O}(H^2 S\sqrt{AK})$$

Define the set of plausible MDPs as

$$\mathcal{M}_k = \{M = (S, A, p_{h,k}, r_{h,k}) \ : \ r_{h,k}(s,a) \in \beta^r_{h,k}(s,a), p_{h,k}(\cdot|s,a) \in \beta^p_{h,k}(s,a)\}$$

Confidence intervals can be anytime or not.

- Define the event $\mathcal{E} = \{\forall k, M^\star \in \mathcal{M}_k\}$. Prove that $\mathbb{P}(\neg\mathcal{E}) \le \delta/2$. First step, construct a confidence interval for rewards and transitions for each (s,a) using Hoeffding and Weissmain inequality (see appendix), respectively. So, we want that

$$\mathbb{P}\Big(\forall k, h, s, a : \widehat{r}_{hk}(s,a) - r_h(s,a)| \le \beta^r_{hk}(s,a) \wedge \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \le \beta^p_{hk}(s,a)\Big) \ge 1 - \delta/2$$

- Define the bonus function and consider the Q-function computed at episode k

$$Q_{h,k}(s,a) = \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \sum_{s'} \widehat{p}_{h,k}(s'|s,a) V_{h+1,k}(s')$$

  with $V_{h,k}(s) = \min\{H, \max_a Q_{h,k}(s,a)\}$. Recall that $V_{H+1,k}(s) = V^\star_{H+1}(s) = 0$. Prove that under event $\mathcal{E}$, Q_k is optimistic, i.e.,

$$Q_{h,k}(s,a) \ge Q^\star_h(s,a), \forall s, a$$

  where $Q^\star$ is the optimal Q-function of the unknown MDP $M^\star$. Note that $\widehat{r}_{H,k}(s,a) + b_{H,k}(s,a) \ge r_{H,k}(s,a)$ and thus $Q_{H,k}(s,a) \ge Q^\star_H(s,a)$ (for a properly defined bonus). Then use induction to prove that this holds for all the stages h.

- In class we have seen that

$$\delta_{1k}(s_{1,k}) \le \sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + m_{hk} \tag{1}$$

  where $\delta_{hk}(s) = V_{hk}(s) - V_h^{\pi_k}(s)$ and $m_{hk} = \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[\delta_{h+1,k}(Y)] - \delta_{h+1,k}(s_{h+1,k})$. We now want to prove this result. Denote by a_{hk} the action played by the algorithm (you will have to use the greedy property).

  1. Show that $V_h^{\pi_k}(s_{hk}) = r(s_{hk}, a_{hk}) + \mathbb{E}_p[V_{h+1,k}(s')] - \delta_{h+1,k}(s_{h+1,k}) - m_{h,k}$

```
Initialize $Q_{h1}(s,a) = 0$ for all $(s,a) \in S \times A$ and $h = 1, \dots, H$

for $k = 1, \dots, K$ do
    Observe initial state $s_{1k}$ (arbitrary)
    Estimate empirical MDP $\widehat{M}_k = (S, A, \widehat{p}_{hk}, \widehat{r}_{hk}, H)$ from $\mathcal{D}_k$

        $\widehat{p}_{hk}(s'|s,a) = \frac{\sum_{i=1}^{k-1} \mathbb{1}\{(s_{hi}, a_{hi}, s_{h+1,i}) = (s,a,s')\}}{N_{hk}(s,a)}, \quad \widehat{r}_{hk}(s,a) = \frac{\sum_{i=1}^{k-1} r_{hi} \cdot \mathbb{1}\{(s_{hi}, a_{hi}) = (s,a)\}}{N_{hk}(s,a)}$

    Planning (by backward induction) for $\pi_{hk}$ using $\widehat{M}_k$
    for $h = H, \dots, 1$ do
        $Q_{h,k}(s,a) = \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \sum_{s'} \widehat{p}_{h,k}(s'|s,a) V_{h+1,k}(s')$
        $V_{h,k}(s) = \min\{H, \max_a Q_{h,k}(s,a)\}$
    end
    Define $\pi_{h,k}(s) = \arg\max_a Q_{h,k}(s,a), \forall s, h$
    for $h = 1, \dots, H$ do
        Execute $a_{hk} = \pi_{hk}(s_{hk})$
        Observe $r_{hk}$ and $s_{h+1,k}$
        $N_{h,k+1}(s_{hk}, a_{hk}) = N_{h,k}(s_{hk}, a_{hk}) + 1$
    end
end
```

Algorithm 1: UCBVI

2. Show that $V_{h,k}(s_{hk}) \leq Q_{h,k}(s_{hk}, a_{hk})$.
3. Putting everything together prove Eq. 1.

- Since $(m_{hk})_{hk}$ is an MDS, using Azuma-Hoeffding we show that with probability at least $1 - \delta/2$

$$\sum_{k,h} m_{hk} \leq 2H\sqrt{KH \log(2/\delta)}$$

Show that the regret is upper bounded with probability $1 - \delta$ by

$$R(T) \leq 2 \sum_{kh} b_{hk}(s_{hk}, a_{hk}) + 2H\sqrt{KH \log(2/\delta)}$$

- Finally, we have that [Domingues et al., 2021]

$$\sum_{h,k} \frac{1}{\sqrt{N_{hk}(s_{hk}, a_{hk})}} \lesssim H^2 S^2 A + 2 \sum_{h=1}^{H} \sum_{s,a} \sqrt{N_{hK}(s,a)}$$

Complete this by showing an upper-bound of $H\sqrt{SAK}$, which leads to $R(T) \lesssim H^2 S\sqrt{AK}$

2.1 Regret minimization answers

- We have :

$$\neg\mathcal{E} = \bigcup_{k,h,s,a} (|\widehat{r}_{h,k}(s,a) - r_h(s,a)| > \beta^r_{h,k}(s,a)) \bigcup (||\widehat{p}_{h,k}(\cdot|s,a) - p_h(\cdot|s,a)||_1 > \beta^p_{h,k}(s,a))$$

There are $2KHSA$ terms in this union and we want the probability of this event to be less or equal than $\delta/2$. Thus we should set the confidence sets so every term is less or equal than $\delta/4KHSA$. One half of the terms can be bounded in probability using Hoeffding's inequality and the other half using Weissmain's inequality. In Hoeffding's inequality this leads to:

$$2\exp(-2N_{h,k}(s,a)\beta^r_{h,k}(s,a)^2) = \delta/4KHSA.$$

So :

$$\Leftrightarrow \boxed{\beta^r_{h,k}(s,a) = \sqrt{\frac{\log(8KHSA/\delta)}{2N_{h,k}(s,a)}}}$$

In Weissmain's inequality:

$$(2^S - 2)\exp(-N_{h,k}(s,a)\beta^p_{h,k}(s,a)^2) = \delta/4KHSA.$$

So :

$$\Leftrightarrow \boxed{\beta^p_{h,k}(s,a) = \sqrt{\frac{2\log((2^S-2)4KHSA/\delta)}{N_{h,k}(s,a)}}}$$

It follows by union bound that

$$\boxed{\mathbb{P}(\neg\mathcal{E}) \le \delta/2}.$$

- Let's prove by induction on h, that : $Q_{h,k}(s,a) \ge Q^\star_h(s,a), \forall s,a$.
  - for $h = H$, we want to have $\widehat{r}_{H,k}(s,a) + b_{H,k}(s,a) \ge r_H(s,a)$. With the assumption that we are under $\mathcal{E}$, we should take $b_{h,k}(s,a) \ge \beta^r_{h,k}(s,a)$ for all k, h, s, a and this inequality will hold because of $\mathcal{E}$.
  - suppose the property is true for some h. Let's prove it for $h-1$:

    $$Q_{h-1,k}(s,a) = \widehat{r}_{h-1,k}(s,a) + b_{h-1,k}(s,a) + \sum_{s'} \widehat{p}_{h-1,k}(s'|s,a)V_{h,k}(s')$$

    By induction, we can lower bound $V_{h,k}(s')$ by $V^\star_h(s')$:

    $$Q_{h-1,k}(s,a) \ge \widehat{r}_{h-1,k}(s,a) + b_{h-1,k}(s,a) + \sum_{s'} \widehat{p}_{h-1,k}(s'|s,a)V^\star_h(s')$$

    Now we want to lower bound $\sum_{s'} \widehat{p}_{h-1,k}(s'|s,a)V^\star_h(s')$. By adding the term $H\beta^p_{h,k}(s,a)$ to the bonus we can lower bound this term. In fact we have:

    $$H\beta^p_{h,k}(s,a) \ge \sum_{s'} |p_h(s'|s,a) - \widehat{p}_{h-1,k}(s'|s,a)|V^\star_h(s')$$

    which immediately follows from $\mathcal{E}$ and the fact that the value function is upper bounded by H (another way to see it is a Holder inequality with $||\cdot||_1$ and $||\cdot||_\infty$. We can further drop the absolute values, so that if we take $\boxed{b_{h,k}(s,a) = \beta^r_{h,k}(s,a) + H\beta^p_{h,k}(s,a)}$ we have:

    $$Q_{h-1,k}(s,a) \ge r_{h-1}(s,a) + \sum_{s'} p_{h-1}(s'|s,a)V^\star_{h-1}(s') = Q^\star_{h-1}(s,a).$$

    The property is then true for every h.

- Let's prove (1).

  1. We have :

     $$m_{h,k} = \mathbb{E}_p[\delta_{h+1,k}(s')] - \delta_{h+1,k}(s_{h+1,k})$$

     And $\delta_{h+1,k}(s') = V_{h+1,k}(s') - V^{\pi_k}_{h+1}(s')$. By taking the expectation :

     $$m_{h,k} = \mathbb{E}_p[V_{h+1,k}(s')] - V^{\pi_k}_h(s_{h,k}) + r(s_{h,k},a_{h,k}) - \delta_{h+1,k}(s_{h+1,k})$$

     So

     $$\boxed{V^{\pi_k}_h(s_{h,k}) = r(s_{h,k},a_{h,k}) + \mathbb{E}_p[V_{h+1,k}(s')] - \delta_{h+1,k}(s_{h+1,k}) - m_{h,k}}$$

2. By definition we have $V_{h,k}(s) \leq \max_a Q_{h,k}(s,a)$. By using that $a_{h,k}$ is the greedy action in state $s_{h,k}$ we have:

$$\boxed{V_{h,k}(s_{h,k}) \leq Q_{h,k}(s_{h,k}, a_{h,k})}$$

3. The property follows from an easy induction (backward from H to 1), combining each time the last two results. It just has to be said that for the initialization we must use that $\delta_{H+1,k}(s_{H+1,k}) = 0 - 0 = 0$. So we have :

$$\boxed{\delta_{1k}(s_{1,k}) \leq \sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y \sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + m_{hk}}$$

- The regret is defined as :

$$R(T) = \sum_{k=1}^{K} V_1^\star(s_{1,k}) - V_1^{\pi_k}(s_{1,k})$$

Under $\mathcal{E}$, the value function is optimistic:

$$R(T) \leq \sum_{k=1}^{K} V_{1,k}(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) = \sum_{k=1}^{K} \delta_{1k}(s_{1,k})$$

By using the previous question:

$$R(T) \leq \sum_{k,h} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y \sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)] + m_{hk}$$

We can plug in the value of Q:

$$R(T) \leq \sum_{k,h} \widehat{r}_{h,k}(s,a) - r(s_{hk}, a_{hk}) + \mathbb{E}_{Y \sim \widehat{p}}[V_{h+1,k}(Y)] - \mathbb{E}_{Y \sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)] + b_{h,k}(s_{h,k}, a_{h,k}) + m_{hk}$$

The sum of the first four terms in the sum can be upper bounded by $b_{h,k}(s_{h,k}, a_{h,k})$, in fact we have precisely chosen $b_{h,k}(s_{h,k}, a_{h,k})$ to upper bound this quantity with absolute values around the differences (under $\mathcal{E}$). Using also the inequality in the handout:

$$\boxed{R(T) \leq 2\sum_{h,k} b_{h,k}(s_{h,k}, a_{h,k}) + 2H\sqrt{KH\log(2/\delta)}}$$

- We have that :

$$\sum_{h,k} \frac{1}{\sqrt{N_{h,k}(s_{h,k}, a_{h,k})}} = \sum_h \sum_{s,a} \sum_{i=1}^{N_{h,K}(s,a)} \frac{1}{\sqrt{i}}$$

This can be upper bounded comparing series and integral:

$$\sum_{h,k} \frac{1}{\sqrt{N_{h,k}(s_{h,k}, a_{h,k})}} \leq 2\sum_h \sum_{s,a} \sqrt{N_{h,K}(s,a)}$$

By using C-S inequality :

$$\sum_{h,k} \frac{1}{\sqrt{N_{h,k}(s_{h,k}, a_{h,k})}} \leq 2\sum_h \sqrt{SA}\sqrt{\sum_{s,a} N_{h,K}(s,a)} = 2H\sqrt{SAK}$$

Furthermore, we have $b_{h,k}(s_{h,k}, a_{h,k}) \lesssim H\sqrt{S}$. Combining the previous question and this bound of the bonus we get :

$$\boxed{R(T) \lesssim H^2 S\sqrt{AK}}.$$

A Weissmain inequality

Denote by $\widehat{p}(\cdot|s,a)$ the estimated transition probability build using n samples drawn from $p(\cdot|s,a)$. Then we have that

$$\mathbb{P}(\|\widehat{p}_h(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \geq \epsilon) \leq (2^S - 2)\exp\Big(-\frac{n\epsilon^2}{2}\Big)$$

References

Ioannis Chatzigeorgiou. Bounds on the lambert function and their application to the outage analysis of user cooperation. *CoRR*, abs/1601.04895, 2016.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. Kernel-based reinforcement learning: A finite-time analysis. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 2783–2792. PMLR, 2021.