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

Exploration in Reinforcement Learning (theory)

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

Solution by Matias Etcheverry

1 Best Arm Identification

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.

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.

  Answer:

  Let

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

  Then

$$\begin{aligned}\mathbb{P}(\mathcal{E}) &\le \sum_{i=1}^{k}\sum_{t=1}^{\infty}\mathbb{P}\left(|\hat{\mu}_{i,t} - \mu_i| > U(t,\delta')\right)\\ &\le 2\sum_{i=1}^{k}\sum_{t=1}^{\infty} e^{-2tU(t,\delta')^2} \quad \text{(Hoeffding's inequality, the reward being bounded)}\end{aligned}$$

  The goal is to find U verifying the above equation. We can set any form to U. We impose:

$$U(t,\delta') = \sqrt{\frac{\log f(t,\delta')}{2t^2}}$$

  with f to be determined. We have:

$$\mathbb{P}(\mathcal{E}) \le 2k\sum_{t=1}^{\infty}\frac{1}{f(t,\delta')}$$

Once again, we can set any form on f. Let

$$f(t,\delta') = t^2 u(\delta')$$

with u to be determined. This form of f assures that the sum is converging, and it equals the Riemann function $\zeta(2)$. We finally have:

$$\begin{aligned}\mathbb{P}(\mathcal{E}) &\leq \frac{2k\zeta(2)}{u(\delta')} \\ &\leq \frac{4k}{u(\delta')} \quad \text{as } \zeta(2) \leq 2\end{aligned}$$

Finally, we simply need $\frac{4k}{u(\delta')} = \delta$. We end up with:

$$U(t,\delta') = \sqrt{\frac{\log(\frac{4kt^2}{\delta})}{2t^2}}$$

We thus have $\mathbb{P}(\mathcal{E}) \leq \delta$ for $\delta' = \frac{\delta}{k}$, resulting in $\boxed{U(t,\delta) = \sqrt{\frac{\log(\frac{4t^2}{\delta})}{2t^2}}}$

- 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 $\neg\mathcal{E}$.

  Answer:

  The best arm $i^\star$ is dropped, at any time step t if there exists an arm j verifying:

  $$\begin{aligned}&\hat{\mu}_{j,t} - U(t,\delta') \geq \widehat{\mu}_{i^\star,t} + U(t,\delta') \\ \Rightarrow\quad &\hat{\mu}_{j,t} - U(t,\frac{\delta}{k}) \geq \widehat{\mu}_{i^\star,t} + U(t,\frac{\delta}{k}) \\ \Rightarrow\quad &\hat{\mu}_{j,t} - \mu_{j,t} - U(t,\frac{\delta}{k}) \geq \widehat{\mu}_{i^\star,t} - \mu_{i^\star,t} + U(t,\frac{\delta}{k}) \\ \text{ie}\quad &\mathbb{P}(\text{arm } i^\star \text{ is dropped}) \leq \mathbb{P}(\{\hat{\mu}_{j,t} - \mu_{j,t} - U(t,\frac{\delta}{k}) \geq \widehat{\mu}_{i^\star,t} - \mu_{i^\star,t} + U(t,\frac{\delta}{k})\})\end{aligned}$$

  Case 1: $\widehat{\mu}_{i^\star,t} - \mu_{i^\star,t} + U(t,\frac{\delta}{k}) > 0$. This means $\hat{\mu}_{j,t} - \mu_{j,t} > U(t,\frac{\delta}{k})$ and thus $\mathbb{P}(\text{arm } i^\star \text{ is dropped}) \leq \mathbb{P}(\mathcal{E}) \leq \delta$

  Case 2: $\widehat{\mu}_{i^\star,t} - \mu_{i^\star,t} + U(t,\frac{\delta}{k}) < 0$. This means $\mu_{i^\star,t} - \widehat{\mu}_{i^\star,t} > U(t,\frac{\delta}{k})$ and thus $\mathbb{P}(\text{arm } i^\star \text{ is dropped}) \leq \mathbb{P}(\mathcal{E}) \leq \delta$

  In either case, the best arm is dropped with probability δ.

- 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]

  Answer:

  Arm i will ultimately be deleted by the best arm $i^\star$ when:

  $$\hat{\mu}_{i^\star,t} - U(t,\frac{\delta}{k}) \geq \hat{\mu}_{i,t} + U(t,\frac{\delta}{k})$$

[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 $u = \log(b/a) - 1$.

Under event $\mathcal{E}^c$, all the observed reward are close to the real reward. Especially, we have:

$$\hat{\mu}_{i^\star,t} \geq \mu_{i^\star} - U(t, \frac{\delta}{k}) \quad \text{and} \quad \hat{\mu}_{i,t} \leq \mu_i + U(t, \frac{\delta}{k})$$

So, if we meet the condition:

$$\begin{aligned} & \mu_{i^\star} - 2U(t, \frac{\delta}{k}) \geq \mu_i + 2U(t, \frac{\delta}{k}) \\ \text{ie} \quad & \boxed{\Delta_i \geq 4U(t, \delta')} \end{aligned}$$

the arm i will always be deleted under $\mathcal{E}^c$. This condition can be rewritten as:

$$\begin{aligned} & \Delta_i \geq 4U(t, \delta') \\ \Rightarrow \quad & \Delta_i^2 \geq 4\frac{\log(\frac{4kt^2}{\delta})}{t^2} \\ \Rightarrow \quad & \Delta_i^2 \geq 4\frac{\log(\frac{4kt^2}{\delta})}{t^2} \\ \Rightarrow \quad & at^2 \geq \log(bt^2) \end{aligned}$$

with $a = \frac{\Delta_i^2}{4}$ and $b = \frac{4k}{\delta}$. With small enough δ, we can insure $\frac{-a}{b} = \frac{-\Delta_i^2\delta}{16k} \in (-1/e, 0)$. We end up with $\boxed{t_i = \sqrt{\frac{1+\sqrt{2u}+u}{\frac{\Delta_i^2}{4}}} \quad \text{with} \quad u = \log(\frac{\Delta_i^2\delta}{16k}) - 1}$. t_i is the worst time from which the non-optimal arm i will be dropped. Hopefully, it can be dropped sooner.

- Compute a bound on the sample complexity (after how many *pulls* the algorithm stops) for identifying the optimal arm w.p. $1-\delta$.

  Answer:

  In the worst case, each arm needs to be dropped after t_i steps, of being chosen for exploration. An upper bound of the sample complexity with probability $1-\delta$ for being in $\mathcal{E}^c$ is then:

$$\boxed{\sum_{i, i\neq i^\star}^{k} \sqrt{\frac{1+\sqrt{2\log(\frac{\Delta_i^2\delta}{16k})-1} + \log(\frac{\Delta_i^2\delta}{16k}) - 1}{\frac{\Delta_i^2}{4}}}}$$

  Answer:

- We assumed that the optimal arm $i^\star$ is unique. Would the algorithm still work if there exist multiple best arms? Why?

  Answer:

  This algorithm wouldn't work if there exits multiple best arms. Indeed, the process is made to identify the best arm, without time condition. Without knowing 2 arms are optimal, a player would just play forever, until one of the arms is beaten by the other, which won't never occur.

  A simple solution is to allow a maximum number of iteration before choosing the best arm. One can also gives the player a gain in quickly choosing a (sub-)optimal arm.

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

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

<u>Answer:</u>

Let s, a, h, k. Let $N_{hk}(s,a)$ be the number of times we have seen the state s with action a, in the past.

- Using Hoeffding inequality on r:

$$\mathbb{P}(|\hat{r}_{hk}(s,a) - r_h(s,a)| \geq \beta^r_{hk}(s,a)) \leq 2e^{-2N_{hk}(s,a)\beta^r_{hk}(s,a)^2}$$

And we want:

$$\frac{\delta}{4SAHK} = 2e^{-2N_{hk}(s,a)\beta^r_{hk}(s,a)^2}$$
$$\Rightarrow \boxed{\beta^r_{hk}(s,a) = \sqrt{\frac{\log(\frac{8SAHK}{\delta})}{2N_{hk}(s,a)}}}$$

- Using Weissman inequality on p

$$\mathbb{P}(\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \geq \beta^p_{hk}(s,a)) \leq (2^S - 2)e^{\frac{-N_{hk}(s,a)\beta^p_{hk}(s,a)^2}{2}}$$

And we want:

$$\frac{\delta}{4SAHK} = (2^S - 2)e^{\frac{-N_{hk}(s,a)\beta^p_{hk}(s,a)^2}{2}}$$
$$\Rightarrow \boxed{\beta^p_{hk}(s,a) = \sqrt{2\frac{\log(2^S - 2) + \log(\frac{4SAHK}{\delta})}{2N_{hk}(s,a)}}}$$

Indeed, with such values for $\beta^r_{hk}(s,a)$ and $\beta^p_{hk}(s,a)$, we have:

$$\begin{aligned}
&\mathbb{P}\Big(\forall k,h,s,a : \widehat{r}_{hk}(s,a) - r_h(s,a)| \leq \beta^r_{hk}(s,a) \wedge \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \leq \beta^p_{hk}(s,a)\Big) \\
&\qquad = 1 - \mathbb{P}\Big(\exists k,h,s,a : \widehat{r}_{hk}(s,a) - r_h(s,a)| \geq \beta^r_{hk}(s,a) \wedge \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \geq \beta^p_{hk}(s,a)\Big) \\
&\qquad \geq 1 - \left(\sum_{k,h,s,a} \mathbb{P}(|\hat{r}_{hk}(s,a) - r_h(s,a)| \geq \beta^r_{hk}(s,a)) + \mathbb{P}(\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \geq \beta^p_{hk}(s,a))\right) \\
&\qquad \geq 1 - \left(\sum_{k,h,s,a} \frac{\delta}{4SAHK} + \frac{\delta}{4SAHK}\right) \\
&\qquad \geq 1 - \frac{\delta}{2}
\end{aligned}$$

- 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.,

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

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

Answer:

The goal of this question is to find a bonus function so that Q_k is optimistic under $\mathcal{E}$. Let's prove that Q_k is optimistic by induction on h, under $\mathcal{E}$. This induction method will give us some conditions on the bonus function $b_{h,k}(s,a)$.

  - for $h = H$, having $Q_{h,k}(s,a) \geq Q^\star_h(s,a)$ implies $\widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) \geq r_h(s,a)$. Under $\mathcal{E}$, this latter condition is met if $b_{h,k}(s,a) \geq \beta^r_{hk}(s,a)$
  - for $h < H$, we suppose that $Q_{h+1,k}(s,a) \geq Q^\star_{h+1}(s,a)$. Thus, $V_{h+1,k}(s) = \min\{H, \max_a Q_{h+1,k}(s,a)\} \geq V^\star_{h+1}(s)$. Then, we also have:

$$\begin{aligned}
Q_{h,k}(s,a) &= \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \sum_{s'} \widehat{p}_{h,k}(s'|s,a) V_{h+1,k}(s') \\
&= \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \sum_{s'} \left(\widehat{p}_{h,k}(s'|s,a) - p_h(s'|s,a) + p_h(s'|s,a)\right) V_{h+1,k}(s') \\
&\geq \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \sum_{s'} \left(\widehat{p}_{h,k}(s'|s,a) - p_h(s'|s,a)\right) V_{h+1,k}(s') + p_h(s'|s,a) V^\star_{h+1}(s')
\end{aligned}$$

And we want, $Q_{h,k}(s,a) \geq Q^\star_h(s,a)$. This condition is met when having

$$\widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \sum_{s'} \left(\widehat{p}_{h,k}(s'|s,a) - p_h(s'|s,a)\right) V_{h+1,k}(s) \geq r_h(s,a)$$

To insure the above condition is met, we simply choose $b_{h,k}(s,a) \geq \beta^r_{hk}(s,a) + \beta^p_{hk}(s,a)\|V_{h+1,k}(s)\|_\infty$, ie $\boxed{b_{h,k}(s,a) = \beta^r_{hk}(s,a) + H\beta^p_{hk}(s,a)}$.

Thus, if we define the bonus function as:

$$\boxed{b_{h,k}(s,a) = \sqrt{\frac{\log(\frac{8SAHK}{\delta})}{2N_{hk}(s,a)}} + H\sqrt{2\frac{\log(2^S - 2) + \log(\frac{4SAHK}{\delta})}{2N_{hk}(s,a)}}}$$

then Q_k is optimistic under $\mathcal{E}$.

- In class we have seen that

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

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

$$\begin{aligned} m_{h,k} &= \mathbb{E}_{s'\sim p(\cdot|s_{h,k},a_{h,k})}[\delta_{h+1,k}(s')] - \delta_{h+1,k}(s_{h+1,k}) \\ &= \mathbb{E}_{s'\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(s') - V_{h+1}^{\pi_k}(s')] - \delta_{h+1,k}(s_{h+1,k}) \end{aligned}$$

   ie $\boxed{m_{h,k} = \mathbb{E}_{s'\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(s')] - V_h^{\pi_k}(s_{h,k}) + r(s_{h,k}, a_{h,k}) - \delta_{h+1,k}(s_{h+1,k})}$ with greedy policy π_k

2. Show that $V_{h,k}(s_{hk}) \leq Q_{h,k}(s_{hk}, a_{hk})$.
   Answer:
   The greedy policy consists in choosing $a_{h,k} = \arg\max_a Q_{h,k}(s,a)$. And we define $V_{h,k}(s,a) = \min(H, \max_a Q_{h,k}(s,a))$. Thus, we have $\boxed{V_{h,k}(s_{hk}) \leq Q_{h,k}(s_{hk}, a_{hk})}$

3. Putting everything together prove Eq. 1.
   Answer:
   First, we prove:

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

   Eq. 1 is simply the above equation with $h = 1$. We prove the above equation with induction on h.

   - for $h = H$, $m_{H,k} = 0$ and $\mathbb{E}_{s'\sim p(\cdot|s_{H,k},a_{H,k})}[V_{H+1,k}(s')]) = 0$. We indeed have $\delta_{H,k}(s_{H,k}) = V_{H,k}(s_{H,k}) - V_H^{\pi_k}(s_{H,k}) \leq Q_{H,k}(s_{H,k}, a_{H,k}) - r(s_{H,k}, a_{H,k})$
   - for $h < H$. We suppose:

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

   And we have:

$$\begin{aligned} \delta_{h,k}(s_{h,k}) &= V_{h,k}(s_{h,k}) - V_h^{\pi_k}(s_{h,k}) \\ &= V_{h,k}(s_{h,k}) + m_{h,k} - \mathbb{E}_{s'\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(s')] - r(s_{h,k}, a_{h,k}) + \delta_{h+1,k}(s_{h+1,k}) \\ &\leq Q_{h,k}(s_{hk}, a_{hk}) - \mathbb{E}_{s'\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(s')] - r(s_{h,k}, a_{h,k}) + \delta_{h+1,k}(s_{h+1,k}) \end{aligned}$$

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

   thanks to q1 and q2 and with the equation true at level $h+1$

- 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

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

Answer:

Under $\mathcal{E}$, ie with probability $1-\delta/2$, and supposing $\sum_{k,h} m_{hk}$ is bounded with probability $1-\delta/2$, we have with probability $1-\delta$ (implicit union bound):

$$\begin{aligned}
R(T) &= \sum_{k=1}^{K} V_1^\star(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) \\
&= \sum_{k=1}^{K} V_{1,k}(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) \\
&= \sum_{k=1}^{K} \delta_{1,k}(s_{1,k}) \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} Q_{h,k}(s_{h,k}, a_{h,k}) - r(s_{h,k}, a_{h,k}) - \mathbb{E}_{s'\sim p(\cdot|s_{h,k},a_{h,k})}[V_{h+1,k}(s')]) + m_{hk} \\
&\leq 2H\sqrt{KH\log(2/\delta)} + \sum_{k,h}^{K,H} Q_{h,k}(s_{h,k}, a_{h,k}) - r(s_{h,k}, a_{h,k}) - \mathbb{E}_{s'\sim p(\cdot|s_{h,k},a_{h,k})}[V_{h+1,k}(s')]) \\
&\leq 2H\sqrt{KH\log(2/\delta)} + \sum_{k,h}^{K,H} \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \mathbb{E}_{s'\sim\widehat{p}_{h,k}}[V_{h+1,k}(s')] - r(s_{h,k}, a_{h,k}) - \mathbb{E}_{s'\sim p}[V_{h+1,k}(s')]
\end{aligned}$$

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

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

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

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

Answer not found.

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.