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

Exploration in Reinforcement Learning (theory)

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

Solution by Thibault de SURREL

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 timestep t, the player selects an arm to pull (I_t), and they observe some reward ($X_{I_t,t}$) for that sample. At any timestep, 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 $\mathbb{P}_{\mu_1,\ldots,\mu_k}(\widehat{i} \neq i^\star) \leq \delta$ and $\tau_\delta < \infty$ almost surely for any $\mu_1,\ldots,\mu_k$. 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

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

  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.

Input: k arms, confidence δ
$S = \{1, \ldots, k\}$
for $t = 1, \ldots$ **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

Answer : We want to show the any-time confidence bound, that is, for $i \in \{1, ..., k\}$,

$$\mathbb{P}(|\hat{\mu}_{i,t} - \mu_i| > U(t, \delta)) \leq \delta$$

To show this bound, we can use Hoeffding's inequality, in fact, the random variables $X_{i,1}, ..., X_{i,t}$ are independents and bounded in $[0, 1]$, so we have

$$\forall u > 0, \ \mathbb{P} \left\{ |\sum_{j=1}^{t} X_{i,j} - \mathbb{E} \left[\sum_{j=1}^{t} X_{i,j} \right] | \geq u \right\} \leq 2 \exp \left(-2tu^2 \right)$$

So, we have

$$\mathbb{P} \left\{ |\hat{\mu}_{i,t} - \mu_i| > U(t, \delta) \right\} \leq 2 \exp(-2U(t, \delta)^2 t)$$

So, if one sets $U(t, \delta)$ to be equal to $\sqrt{\frac{\log(2/\delta)}{2t}}$, on has $\mathbb{P} \left\{ |\hat{\mu}_{i,t} - \mu_i| > U(t, \delta) \right\} \leq \delta$. So, the function $U(t, \delta)$ that satisfy the any-time confidence bound is

$$\boxed{U(t, \delta) = \sqrt{\frac{\log(2/\delta)}{2t}}} \tag{1}$$

Now, if one sets $\delta' = \frac{6}{\pi^2 k t^2} \delta$, one has :

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &= \mathbb{P} \left(\bigcup_{i=1}^{k} \bigcup_{t=1}^{\infty} \{ |\widehat{\mu}_{i,t} - \mu_i| > U(t, \delta') \} \right) \\
&\leq \sum_{i=1}^{k} \sum_{t=1}^{\infty} \mathbb{P}(|\widehat{\mu}_{i,t} - \mu_i| > U(t, \delta')) \quad \text{because we have a countable union} \\
&\leq \sum_{i=1}^{k} \sum_{t=1}^{\infty} \delta' \quad \text{thank to the first part of the question} \\
&\leq \sum_{i=1}^{k} \sum_{t=1}^{\infty} \frac{6}{\pi^2 k t^2} \delta \\
&\leq \delta \quad \text{because } \sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6}
\end{aligned}$$

So, for $\boxed{\delta' = \frac{6}{\pi^2 k t^2} \delta}$, we have $\mathbb{P}(\mathcal{E}) \leq \delta$.

- 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 : Let us suppose that $i^\star$ got removed from the active set. The condition for this to happen is :

$$\exists j \in S,\ \widehat{\mu}_{j,t} - U(t,\delta') \leq \widehat{\mu}^\star_t + U(t,\delta')$$

Let us suppose that $\mathbb{P}(\mathcal{E}) \leq \delta$, then we have that $\mathbb{P}(\neg\mathcal{E}) > 1-\delta$ where

$$\neg\mathcal{E} = \bigcap_{i=1}^{k}\bigcap_{t=1}^{\infty}\{|\widehat{\mu}_{i,t} - \mu_i| \leq U(t,\delta')\}$$

Then, for all $i \in \{1,...,k\}$, we have $\mathbb{P}(|\hat{\mu}_{i,t} - \mu_i| > U(t,\delta')) > 1-\delta$ (because $\neg\mathcal{E} \subset \{|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta')\}$) so, with probability $1-\delta$, we have

$$-U(t,\delta') \leq \widehat{\mu}_{i,t} - \mu_i \leq U(t,\delta').$$

So, still with with probability $1-\delta$, we have

$$\begin{cases} \mu_i + U(t,\delta') \geq \widehat{\mu}_{i,t} \\ \mu_i - U(t,\delta') \leq \widehat{\mu}_{i,t} \end{cases} \tag{2}$$

In particular, for $i = i^\star$ which is unique, we have

$$\mu_j + U(t,\delta') - U(t,\delta') > \mu_{i^\star} + U(t,\delta') - U(t,\delta')$$

so $\mu_j > \mu_{i^\star}$ This is not possible as $i^\star$ is the best arm. $\boxed{\text{So the optimal arm } i^\star \text{ remains in the active set.}}$

- 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 : Let $i \neq i^\star$. Then, if

$$\widehat{\mu}^\star_t - U(t,\delta') \geq \widehat{\mu}_{i,t} + U(t,\delta') \tag{3}$$

is verified, the arm i is removed from the active set. As $\neq \mathcal{E}$ holds, we still have the inequalities 2, in particular

$$\begin{cases} \mu_i + U(t,\delta') \geq \widehat{\mu}^\star_t \\ \mu_i - U(t,\delta') \leq \widehat{\mu}_{i,t} \end{cases} \tag{4}$$

Therefor, if

$$\mu^\star - 2U(t,\delta') \geq \mu_i + 2U(t,\delta')$$

is verified, we have that the condition 3 is also verified, therefor the arm i is removed from the active set. We can rewrite this condition as

$$\boxed{\Delta_i \geq 4U(t,\delta)}$$

Moreover, from the first question, we have the expression of $U(t,\delta')$. Then, we have

$$\begin{aligned} \Delta_i \geq 4U(t,\delta) &\iff \Delta_i \geq 4\sqrt{\frac{\log\left(\frac{\pi^2}{3\delta}t^2k\right)}{2t}} \\ &\iff \Delta_i^2 \geq 8\frac{\log\left(\frac{\pi^2}{3\delta}t^2k\right)}{t} \\ &\iff t\Delta_i^2 \geq 16\log\left(\pi\sqrt{\frac{k}{3\delta}}t\right) \\ &\iff at \geq \log(bt) \end{aligned}$$

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

Where $a = \frac{\Delta_i}{16}$ ($\neq 0$ because of the uniqueness of $i^\star$) and $b = \sqrt{\frac{k}{3\delta}}$. We can now use the footnote and get that the condition on the time t_i to have i removed from the active set is :

$$\boxed{t_i \geq \frac{\sqrt{2\log\left(\frac{16}{\Delta_i^2}\sqrt{\frac{k}{3\delta}}\right) - 2} + \log\left(\frac{16}{\Delta_i^2}\sqrt{\frac{k}{3\delta}}\right)}{\Delta_i^2/16}}$$

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

Answer : For each $i \neq i^\star$, that is for each non optimal arm i, it will be removed after t_i pulls with probability $1-\delta$. So, a bound on the sample complexity is the sum of the t_is :

$$\boxed{\mathcal{O}\left(\sum_{i\neq i^\star} \frac{\log\left(\frac{16}{\Delta_i^2}\sqrt{\frac{k}{3\delta}}\right)}{\Delta_i^2}\right)}$$

- 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.

Answer : If $i^\star$ is not unique, then there exists $i \in \{1, ..., k\} \setminus i^\star$ such that $i = i^\star$ so $\Delta_i = 0$. The algorithm would still remove all sub optimal with probability $1-\delta$. But, once all the sub optimal arms have been removed, the algorithm would keep iterating among the remaining one (which are all optimal) and the time t_i to remove i would be $+\infty$ as $\Delta_i = 0$. So the algorithm would not work.

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

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

Answer : We want to show that $\mathbb{P}(\neg\mathcal{E}) \leq \delta/2$. Let us start by looking at $\neg\mathcal{E}$. From the definition of $\neg\mathcal{E}$, we have

$$\begin{aligned}\neg\mathcal{E} &= \{\exists k\colon M^\star \notin \mathcal{M}_k\} \\ &= \{\exists k,s,a,h\colon |\widehat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a) \text{ or } \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 > \beta^p_{hk}(s,a)\} \\ &= \bigcup_{k,s,a,h} \left(\{|\widehat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)\} \cup \{\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 > \beta^p_{hk}(s,a)\}\right)\end{aligned}$$

Using this, we have, since the union is countable,

$$\begin{aligned}\mathbb{P}(\neg\mathcal{E}) &\leq \sum_{k,s,a,h} \mathbb{P}\left(\{|\widehat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)\} \cup \{\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 > \beta^p_{hk}(s,a)\}\right) \\ &\leq \sum_{k,s,a,h} \mathbb{P}\left\{|\widehat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)\right\} + \mathbb{P}\left\{\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 > \beta^p_{hk}(s,a)\right\}\end{aligned}$$

So, if we can find that the both terms of the sum are less than $\frac{\delta}{4KSAH}$, we will have the desired result.

Let us start with the first term. As in part 1, we can use Hoeffding's inequality on the rewards that are independents and bounded in $[0,1]$. From this we get :

$$\mathbb{P}\left\{|\widehat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)\right\} \leq 2e^{-2N_{h,k}(s,a)\beta^r_{hk}(s,a)^2}$$

Moreover, we have

$$2e^{-N_{h,k}(s,a)\beta^r_{hk}(s,a)^2} = \frac{\delta}{4KSAH} \iff \beta^r_{hk}(s,a) = \sqrt{\frac{\log\left(\frac{8KSAH}{\delta}\right)}{2N_{h,k}(s,a)}}$$

For the second term, we can use the Weissmain inequality of section A :

$$\mathbb{P}(\|\widehat{p}_h(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \geq \beta^p_{hk}(s,a)) \leq (2^S - 2)\exp\left(-\frac{N_{h,k}(s,a)\beta^p_{hk}(s,a)^2}{2}\right)$$

Once again, we have

$$(2^S - 2)\exp\left(-\frac{N_{h,k}(s,a)\beta^p_{hk}(s,a)^2}{2}\right) = \frac{\delta}{4KSAH} \iff \beta^p_{hk}(s,a) = \sqrt{\log\left(\frac{4KSAH(2^S-2)}{\delta}\right)\frac{2}{N_{h,k}(s,a)}}$$

So, by choosing

$$\boxed{\beta^r_{hk}(s,a) = \sqrt{\frac{\log\left(\frac{8KSAH}{\delta}\right)}{2N_{h,k}(s,a)}} \text{ and } \beta^p_{hk}(s,a) = \sqrt{\frac{2}{N_{h,k}(s,a)}\log\left(\frac{4KSAH(2^S-2)}{\delta}\right)}}$$

we have that $\mathbb{P}(\neg\mathcal{E}) \leq \delta/2$

- 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.,

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

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 : Let us show the result by backward induction on $h \in \{1, ..., H\}$.

- For $h = H$. Under $\mathcal{E}$, we have that, for all a and s, $|\widehat{r}_{H,k}(s,a) - r_{H,k}(s,a)| \leq \beta^r_{Hk}(s,a)$, thus, $r_{H,k}(s,a) - \widehat{r}_{H,k}(s,a) \leq \beta^r_{Hk}(s,a)$ and so $r_{H,k}(s,a) \leq \beta^r_{Hk}(s,a) + \widehat{r}_{H,k}(s,a)$. So, if we choose the bonus function $b_{H,k}$ such that, for all a and s we have $b_{H,k}(s,a) \geq \beta^r_{Hk}(s,a)$, we then have $\widehat{r}_{H,k}(s,a) + b_{H,k}(s,a) \geq r_{H,k}(s,a)$. Moreover, we have that $V_{H+1,k}(s) = V^\star_{H+1}(s) = 0$. Putting everything together we get :

$$Q_{H,k}(s,a) = \widehat{r}_{H,k}(s,a) + b_{H,k}(s,a) \geq r_{H,k}(s,a) = Q^\star_H(s,a) \quad \forall a, s$$

  We then have the initialization of the induction.

- Let $h \in \{1, ..., H-1\}$. Let us suppose that, for all s, a, we have $Q_{h+1,k}(s,a) \geq Q^\star_{h+1}(s,a)$. We want to show the result for the rank h. A first remark is that we have, using the induction, that

$$V_{h+1,k}(s) = \min\{H, \max_a Q_{h,k}(s,a)\} \geq \max_a Q^\star_h(s,a) = V^\star_{h+1}(s)$$

  Using the Bellman equation and the definition of $Q_{h,k}$, we have

$$\begin{cases} Q_{h,k}(s,a) = \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \sum_{s'} \widehat{p}_{h,k}(s'|s,a)V_{h+1}(s') \\ Q^\star_h(s,a) = r_{h,k}(s,a) + \sum_{s'} p_{h,k}(s'|s,a)V^\star_{h+1}(s') \end{cases}$$

  So, we have that, for all s, a,

$$\begin{aligned} Q^\star_h(s,a) - Q_{h,k}(s,a) &= r_{h,k}(s,a) - \widehat{r}_{h,k}(s,a) - b_{h,k}(s,a) + \sum_{s'} \left(p_{h,k}(s'|s,a)V^\star_{h+1}(s') - \widehat{p}_{h,k}(s'|s,a)V_{h+1}(s')\right) \\ &\leq |r_{h,k}(s,a) - \widehat{r}_{h,k}(s,a)| - b_{h,k}(s,a) + \sum_{s'} \left(p_{h,k}(s'|s,a) - \widehat{p}_{h,k}(s'|s,a)\right) V_{h+1,k}(s') \\ &\leq |r_{h,k}(s,a) - \widehat{r}_{h,k}(s,a)| - b_{h,k}(s,a) + \sum_{s'} |p_{h,k}(s'|s,a) - \widehat{p}_{h,k}(s'|s,a)| H \\ &\leq |r_{h,k}(s,a) - \widehat{r}_{h,k}(s,a)| - b_{h,k}(s,a) + H\|p_{h,k}(s'|s,a) - \widehat{p}_{h,k}(s'|s,a)\|_1 \\ &\leq \beta^r_{hk}(s,a) - b_{h,k}(s,a) + H\beta^p_{hk}(s,a) \text{ since we are under } \mathcal{E} \end{aligned}$$

  So, if we choose $b_{h,k}(s,a) \geq \beta^r_{hk}(s,a) + H\beta^p_{hk}(s,a)$, we have the results that we want.

To conclude the induction, for $b_{h,k}(s,a) \geq \beta^r_{hk}(s,a) + H\beta^p_{hk}(s,a)$, we have, for all stages h that

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

- In class we have seen that

$$\delta_{1k}(s_{1,k}) \leq \sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y \sim p(\cdot|s_{hk}, a_{hk})}[V_{h+1,k}(Y)]) + m_{hk} \tag{5}$$

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

Answer : We start by using the Bellman equation :

$$\begin{aligned}
V_h^{\pi_k}(s_{h,k}) &= r(s_{h,k}, \pi_k(s_{h,k})) + \sum_{s'} p_{h+1,k}(s' \mid s_{h,k}, \pi_k(s_{h,k}))V_{h+1}^{\pi_k}(s') \\
&= r(s_{h,k}, a_{h,k}) + \sum_{s'} p_{h+1,k}(s' \mid s_{h,k}, a_{h,k})V_{h+1}^{\pi_k}(s') \quad \text{as } a_{h,k} \text{ is the action played} \\
&= r(s_{h,k}, a_{h,k}) + \sum_{s'} p_{h+1,k}(s' \mid s_{h,k}, a_{h,k})(V_{h+1,k}(s') - \delta_{h+1,k}(s')) \text{ by defintion of } \delta_{h+1,k}(s') \\
&= r(s_{h,k}, a_{h,k}) + \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)] - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[\delta_{h+1,k}(Y)]
\end{aligned}$$

Finally, using the definition of $m_{h,k}$, we have

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

2. Show that $V_{h,k}(s_{hk}) \leq Q_{h,k}(s_{hk}, a_{hk})$.

Answer : In order to show this, we use the definition of $V_{h,k}(s)$ and the fact that the action $a_{h,k}$ played by the algorithm is the greedy action i.e. $a_{h,k} \in \arg\max_a Q_{h,k}(s_{h,k}, a)$:

$$\begin{aligned}
V_{h,k}(s_{h,k}) &= \min\{H, \max_a Q_{h,k}(s_{h,k}, a)\} \\
&\leq \min\{H, Q_{h,k}(s_{h,k}, a_{h,k})\} \\
&\leq Q_{h,k}(s_{h,k}, a_{h,k})
\end{aligned}$$

3. Putting everything together prove Eq. 5.

Answer : First, using the two last questions, we have that :

$$\delta_{hk}(s) = V_{hk}(s) - V_h^{\pi_k}(s) \leq Q_{h,k}(s_{h,k}, a_{h,k}) - r(s_{hk}, a_{hk}) - \mathbb{E}_p[V_{h+1,k}(s')] + \delta_{h+1,k}(s_{h+1,k}) + m_{h,k} \tag{6}$$

Let us now do a backward induction on $h \in \{1, ..., H\}$ in order to prove the following property

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

- For $h = H$. We just have to use the equation 6 for $h = H$ knowing that $\delta_{H+1,k}(s_{H+1,k}) = 0$. We then have :

$$\delta_{Hk}(s) \leq Q_{Hk}(s_{Hk}, a_{Hk}) - r(s_{Hk}, a_{Hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{Hk},a_{Hk})}[V_{H+1,k}(Y)] + m_{Hk}$$

- Let $h \in \{1, ..., H-1\}$. Let us suppose that the property we want to show is true at rank $h+1$. Let us show it for the rank h. Using equation 6 and the induction hypothesis, we have that

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

Thus, for $h = 1$, we have the desired result :

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

- 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 : We have shown in the previous questions, that under $\mathcal{E}$ (that occurs with probability $1-\delta$), that Q_k is optimistic, so V_k too and we have $V_1^\star(s_1,k) \leq V_{1,k}(s_1,k)$. So, we have the following derivation :

$$\begin{aligned}
R(T) &= \sum_{k=1}^{K} V_1^\star(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) \\
&\leq \sum_{k=1}^{K} V_{1,k}(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) \\
&\leq \sum_{k=1}^{K} \delta_{1k}(s_{1,k}) \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + m_{hk} \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} \big(Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]\big)\big) + 2H\sqrt{KH\log(2/\delta)} \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} \widehat{r}_{h,k}(s_{hk}, a_{hk}) + b_{h,k}(s,a) + \mathbb{E}_{Y\sim \widehat{p}(\cdot|s_{hk},a_{hk})}[V_{h+1}(Y)] - r(s_{hk}, a_{hk}) \\
&- \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + 2H\sqrt{KH\log(2/\delta)}
\end{aligned}$$

We now have an upper bound of $R(T)$ composed of the sum of multiples differences, that we have all already bounded in the previous questions. In fact, we have that, with probability higher than $1-\delta/2$,

$$\begin{aligned}
&\widehat{r}_{h,k}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) + \mathbb{E}_{Y\sim \widehat{p}(\cdot|s_{hk},a_{hk})}[V_{h+1}(Y)] - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) \\
&\leq |\widehat{r}_{h,k}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk})| + \sum_{s'} |p_{h,k}(s'|s_{hk}, a_{hk}) - \widehat{p}_{h,k}(s'|s,a)| V_{h+1}(s') \\
&\leq |\widehat{r}_{h,k}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk})| + H\sum_{s'} |p_{h,k}(s'|s_{hk}, a_{hk}) - \widehat{p}_{h,k}(s'|s_{hk}, a_{hk})| \\
&\leq \beta^r_{hk}(s_{hk}, a_{hk}) + H\beta^p_{hk}(s,a) \\
&\leq b_{h,k}(s_{hk}, a_{hk})
\end{aligned}$$

So, putting everything together, we have, with probability $1-\delta$:

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

- Finally, we have that [Domingues et al., 2021]
$$\sum_{h,k} \frac{1}{\sqrt{N_{hk}(s_{hk}, a_{hk})}} \lesssim H^2S^2A + 2\sum_{h=1}^{H}\sum_{s,a}\sqrt{N_{hK}(s,a)}$$
Complete this by showing an upper-bound of $H\sqrt{SAK}$, which leads to $R(T) \lesssim H^2S\sqrt{AK}$

Answer : We start by using Hölder's inequality :

$$\sum_{s,a}\sqrt{N_{hK}(s,a)} \leq \sqrt{\sum_{s,a}1}\sqrt{\sum_{s,a}\sqrt{N_{hK}(s,a)}^2} = \sqrt{SA\sum_{s,a}N_{hK}(s,a)}$$

So, we have, since $\sum_{s,a} N_{hK}(s,a) \leq K$,

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

According to the deviations lead previously, we have, by choosing $b_{h,k}(s,a) = \beta^r_{hk}(s,a) + H\beta^p_{hk}(s,a)$

$$\begin{aligned}
R(T) &\leq 2\sum_{kh} b_{hk}(s_{hk},a_{hk}) + 2H\sqrt{KH\log(2/\delta)} \\
&\leq 2\sum_{kh} \beta^r_{hk}(s_{hk},a_{hk}) + H\beta^p_{hk}(s_{hk},a_{hk}) + 2H\sqrt{KH\log(2/\delta)} \\
&\leq 2\sum_{kh}\sqrt{\frac{\log\left(\frac{8KSAH}{\delta}\right)}{2N_{h,k}(s,a)}} + H\sqrt{\frac{2}{N_{h,k}(s,a)}\log\left(\frac{4KSAH(2^S-2)}{\delta}\right)} + 2H\sqrt{KH\log(2/\delta)} \\
&\leq \left(\sqrt{\frac{\log\left(\frac{8KSAH}{\delta}\right)}{2}} + 2H\sqrt{2\log\left(\frac{4KSAH(2^S-2)}{\delta}\right)}\right)\sum_{kh}\sqrt{\frac{1}{N_{h,k}(s,a)}} + 2H\sqrt{KH\log(2/\delta)} \\
&\lesssim \left(\sqrt{\frac{\log\left(\frac{8KSAH}{\delta}\right)}{2}} + 2H\sqrt{2\log\left(\frac{4KSAH(2^S-2)}{\delta}\right)}\right)(H^2S^2A + 2H\sqrt{SAK}) + 2H\sqrt{KH\log(2/\delta)}
\end{aligned}$$

So, after a lot of approximations and computations, we have that :

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

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

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

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.

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 π_{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