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

Exploration in Reinforcement Learning (theory)

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

Solution by DI PIAZZA Théo

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

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

Input: k arms, confidence δ
$S = \{1, \dots, k\}$
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]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 $u = \log(b/a) - 1$.

1.1 Answer - Best Arm Identification

- To start, let's define the any-time confidence bound such that:

$$\mathbb{P}\left(\bigcup_{t=1}^{\infty}\{|\widehat{\mu}_{i,t}-\mu_i|>U(t,\delta)\}\right)\leq\delta$$

It comes that:

$$\mathbb{P}\left(\bigcup_{t=1}^{\infty}\{|\widehat{\mu}_{i,t}-\mu_i|>U(t,\delta)\}\right)\leq\sum_{t=1}^{\infty}\mathbb{P}\left\{|\widehat{\mu}_{i,t}-\mu_i|>U(t,\delta)\right\}$$

Therefore, it's to find $U(t,\delta)$ that satisfies:

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

Using Hoeffding's inequality, it comes that:

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

Hence,

$$\frac{\delta}{2t^2}=2exp(-2t(U(t,\delta))^2)\iff\boxed{U(t,\delta)=\sqrt{\frac{log(\frac{4t^2}{\delta})}{2t}}}$$

Now, let's find a particular δ' such that $\mathbb{P}(\mathcal{E})\leq\delta$.

$$\mathbb{P}(\mathcal{E})=\mathbb{P}(\bigcup_{i=1}^{k}\bigcup_{t=1}^{\infty}\{|\widehat{\mu}_{i,t}-\mu_i|>U(t,\delta')\})\leq\sum_{i=1}^{k}\sum_{t=1}^{\infty}\mathbb{P}\left\{|\widehat{\mu}_{i,t}-\mu_i|>U(t,\delta')\right\}$$
$$\leq\sum_{i=1}^{k}\sum_{t=1}^{\infty}\frac{\delta'}{2t^2}=\sum_{i=1}^{k}\frac{\delta'}{2}\sum_{t=1}^{\infty}\frac{1}{t^2}=\frac{k\delta'\pi}{12}\leq k\delta'\implies\boxed{\delta'=\frac{\delta}{k}}$$

.

- Now, let's find that with probability at least $1-\delta$, the optimal arm $i^\star=\arg\max_i\{\mu_i\}$ remains in the active set S.
Recall: the optimal arm $i^\star$ doesn't remain in S if $\exists j\in S,\ \widehat{\mu}_{j,t}-U(t,\delta')\geq\widehat{\mu}_{i,t}+U(t,\delta')$ (1)
We suppose that $\mathcal{E}$ (resp. $\mathcal{E}^c$) holds with probability δ (resp. $1-\delta$), where:

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

It comes that:

$$-U(t,\delta')\leq\widehat{\mu}_{i^\star,t}-\mu_{i^\star}\leq U(t,\delta')\implies\widehat{\mu}_{i^\star,t}+U(t,\delta')\geq\mu_{i^\star}$$
$$-U(t,\delta')\leq\widehat{\mu}_{j,t}-\mu_j\leq U(t,\delta')\implies\widehat{\mu}_{j,t}-U(t,\delta')\leq\mu_j$$

Hence, with (1) inequality:

$$\mu_j\geq\widehat{\mu}_{j,t}-U(t,\delta')\geq\widehat{\mu}_{i^\star,t}+U(t,\delta')\geq\mu_{i^\star}$$

Which cannot hold since $i^\star$ is the optimal arm.
$\boxed{\text{Hence with probability } 1-\delta\text{, the optimal arm } i^\star \text{ remains in the active set } S.}$

- Now, let's show that under $\mathcal{E}^c$, an arm $i \neq i^\star$ will be removed from the active set when:

$$\Delta_i \geq C_1 U(t, \delta')$$

Recall: the arm i doesn't remain in S if $\widehat{\mu}_t^* - U(t,\delta') \geq \widehat{\mu}_{i,t} + U(t,\delta')$ (1). Where $\widehat{\mu}_t^*$ is the estimated reward of the arm with the largest expected reward μ^*.
As it has been seen previously, if $\mathcal{E}^c$ holds, it comes that:

$$-U(t,\delta') \leq \widehat{\mu}_t^* - \mu^* \leq U(t,\delta') \Longrightarrow \widehat{\mu}_t^* \geq \mu^* - U(t,\delta')$$

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

Hence, with inequality (1):

$$\mu^* - 2U(t,\delta') \geq \mu_i + 2U(t,\delta') \Longleftrightarrow \mu^* - \mu i = \Delta_i \geq 4U(t,\delta')$$

Now, let's compute the time required to have such condition for each non-optimal arm.
Since $U(t,\delta') = \sqrt{\frac{log(\frac{4kt^2}{\delta})}{2t}}$, the condition comes that:

$$\Delta_i \geq 4\sqrt{\frac{log(\frac{4kt^2}{\delta})}{2t}}$$

Then, it's needed to find the minimum value of t which holds this inequality, for each arm (T_i).

$$\Delta_i \geq 4\sqrt{\frac{log(\frac{4kt^2}{\delta})}{2t}} \Longleftrightarrow \frac{\Delta_i t}{8} \geq log(\frac{4t^2}{\delta})$$

$$\Longrightarrow \forall i\ (\neq i^\star) \in S,\ T_i \leq C\ \frac{log(\frac{k\ log(\Delta_i^{-2})}{\delta})}{\Delta_i^2}$$

- Now, let's compute a bound on the sample complexity. As it has been seen previously, each arm $i \neq i^\star$ will be removed from the active set S after it has been sampled at most T_i times. Therefore to find upper bound on the sample complexity, it's need to sum the upper bounds on the sample on each non-optimal arm. The bound is:

$$\mathcal{O}\left(\sum_{i=1, i\neq i^\star}^{k} \frac{log(k\ log(\Delta_i^{-2}))}{\Delta_i^2}\right)$$

- Since the optimal arm $i^\star$ doesn't remain in S if $\exists j \in S,\ \widehat{\mu}_{j,t} - U(t,\delta') \geq \widehat{\mu}_{i,t} + U(t,\delta')$, it should not work since it won't converge. Indeed, once all non-optimal arms will be removed, the algorithm will iterate to infinity.

End of Answer - Best Arm Idenfitication.

Let's move on - 2 Regret Minimization in RL

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

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

- 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.

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

- As explained in the question, it's needed to show that:

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

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

$$\mathbb{P}\Big(|\widehat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)\Big) + \mathbb{P}\Big(\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \ > \beta^p_{hk}(s,a)\Big) \le \delta/2$$

<u>For rewards</u> Using Hoeffding's inequality, it comes that:

$$\mathbb{P}\Big(|\widehat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)\Big) \le 2exp(-2N_{h,k}(\beta^r_{hk}(s,a))^2)$$

We want that :

$$2exp(-2N_{h,k}(\beta^r_{hk}(s,a))^2) = \frac{\delta}{4|S||A|HK} \iff \beta^r_{hk}(s,a) = \sqrt{\frac{log(\frac{8|S||A|HK}{\delta})}{2N_{h,k}}}$$

It comes that:

$$\mathbb{P}\Big(|\widehat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)\Big) \le 2exp(\frac{-2N_{h,k}log(\frac{\delta}{8|S||A|HK})}{-2N_{h,k}}) = \frac{\delta}{4|S||A|HK} \le \frac{\delta}{4}$$

<u>For transitions</u> Using Weissmain inequality, it comes that:

$$\mathbb{P}\Big(\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \ > \beta^p_{hk}(s,a)\Big) \le (2^{|S|}-2)exp(-\frac{N_{h,k}(\beta^p_{hk}(s,a))^2}{2})$$

We want that:

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

It comes that:

$$\mathbb{P}\Big(\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \ > \beta^p_{hk}(s,a)\Big) \le (2^{|S|}-2)exp(-\frac{N_{h,k}(\sqrt{\frac{2log(\frac{(2^{|S|}-2)4|S||A|HK}{\delta})}{N_{h,k}}})^2}{2}) = \frac{\delta}{4|S||A|HK} \le \frac{\delta}{4}$$

<u>So finally, it comes that:</u>

$$\mathbb{P}\Big(\exists k,h,s,a : |\widehat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a) \vee \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \ > \beta^p_{hk}(s,a)\Big) \le \frac{\delta}{4} + \frac{\delta}{4} = \frac{\delta}{2}$$

Hence,

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

- The bonus function is defined as:

$$b_{h,k}(s,a) = H\sqrt{\frac{2log(\frac{(2^{|S|}-2)4|S||A|HK}{\delta})}{N_{h,k}}}$$

Induction will be used to prove that for all the stages:

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

For $h = H$: 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 $h < H$: We suppose that $Q_{h,k}(s,a) \geq Q^\star_h(s,a), \forall s,a$. It's known that $V_{h,k}(s) = \min\{H, \max_a Q_{h,k}(s,a)\}$, $Q_{h,k}(s,a) \geq Q^\star_h(s,a)$ and then $V_{h,k}(s) \geq V^*_h(s)$.
It comes that:

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

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

$$Q_{h-1,k}(s,a) \geq r_{h-1,k}(s,a) + \sum_{s'} p_{h-1,k}(s'|s,a)V^*_h(s') \quad \text{under } \mathcal{E}$$

Hence,

$$Q_{h-1,k}(s,a) \geq Q^*_{h-1}(s,a) \ \forall(s,a)$$

$\boxed{\text{So finally, } Q_{h,k}(s,a) \geq Q^*_h(s,a) \ \forall(s,a)}$

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

$$m_{h,k} = \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[\delta_{h+1,k}(Y)] - \delta_{h+1,k}(s_{h+1,k})$$

$$\implies \delta_{h+1,k}(s_{h+1,k}) = \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[\delta_{h+1,k}(Y)] - mh,k$$

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

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

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

  2. Then, let's show that $V_{h,k}(s_{hk}) \leq Q_{h,k}(s_{hk}, a_{hk})$.
  It's known $V_{h,k}(s) = \min\{H, \max_a Q_{h,k}(s,a)\}$ and $\max_a Q_{h,k}(s_{hk}, a) = Q_{h,k}(s_{hk}, a_{hk})\}$.
  Hence,

$$V_{h,k}(s_{hk}) = \min\{H, \max_a Q_{h,k}(s_{hk}, a)\} \leq Q_{h,k}(s_{hk}, a_{hk})$$

  3. Finally, let's show that

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

  Induction will be used to prove that.
  For h=H:

$$\begin{aligned} \delta_{Hk}(s_{H,k}) &= V_{H,k}(s_{Hk}) - V^{\pi_k}_H(s_{Hk}) \\ &\leq Q_{H,k}(s_{Hk}, a_{Hk})) - V^{\pi_k}_H(s_{Hk}) \\ &= Q_{H,k}(s_{Hk}, a_{Hk}) - \mathbb{E}_p[V_{H+1,k}(s')] - r(s_{H,k}, a_{H,k}) + m_{H,k} + \delta_{H+1,k}(s_{H+1,k}) \\ &= \sum_{h=H}^{H} Q_{H,k}(s_{Hk}, a_{Hk}) - \mathbb{E}_p[V_{h+1,k}(s')] - r(s_{h,k}, a_{h,k}) + m_{h,k} \end{aligned}$$

  For $h < H$: We suppose that:

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

  Then, it comes that:

$$\delta_{h-1,k}(s_{h-1,k}) = V_{h-1,k}(s_{h-1,k}) - V^{\pi_k}_{h-1}(s_{h-1,k})$$

$$\leq Q_{h-1,k}(s_{h-1,k}, a_{h-1,k})) - V_{h-1}^{\pi_k}(s_{h-1,k})$$
$$= Q_{h-1,k}(s_{h-1,k}, a_{h-1,k}) - \mathbb{E}_p[V_{h,k}(s')] - r(s_{h-1,k}, a_{h-1,k}) + m_{h-1,k} + \delta_{h,k}(s_{h,k})$$
$$= Q_{h-1,k}(s_{h-1,k}, a_{h-1,k}) - \mathbb{E}_p[V_{h,k}(s')] - r(s_{h-1,k}, a_{h-1,k}) + m_{h-1,k} + \delta_{h,k}(s_{h,k})$$
$$\leq Q_{h-1,k}(s_{h-1,k}, a_{h-1,k}) - \mathbb{E}_p[V_{h,k}(s')] - r(s_{h-1,k}, a_{h-1,k}) + m_{h-1,k}$$
$$+ Q_{h,k}(s_{h,k}, a_{h,k}) - \mathbb{E}_p[V_{h+1,k}(s')] - r(s_{h,k}, a_{h,k}) + m_{h,k}$$
$$= \sum_{j=h-1}^{H} Q_{j,k}(s_{jk}, a_{jk}) - \mathbb{E}_p[V_{j+1,k}(s')] - r(s_{j,k}, a_{j,k}) + m_{j,k}$$

Hence:

$$\delta_{h,k}(s_{h,k}) \leq \sum_{j=h}^{H} Q_{j,k}(s_{jk}, a_{jk}) - \mathbb{E}_p[V_{j+1,k}(s')] - r(s_{j,k}, a_{j,k}) + m_{j,k}$$

So finally:

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

- Now, let's show that the regret is upper bounded with probability $1 - \delta$ by

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

To start with the definition of $R(T)$:

$$R(T) = \sum_{k=1}^{K} V_1^*(s_1, k) - V_1^{\pi_k}(s_1, k) \leq \sum_{k=1}^{K} V_{1,k}(s_1, k) - V_1^{\pi_k}(s_1, k) \text{ when } \mathcal{E} \text{ holds}$$

$$\implies R(T) \leq \sum_{k=1}^{K} \delta_{1,k}(s_1, k)$$

$$\leq \sum_{k=1}^{K}\sum_{h=1}^{H} Q_{h,k}(s_{hk}, a_{hk}) - \mathbb{E}_p[V_{h+1,k}(s')] - r(s_{h,k}, a_{h,k}) + m_{h,k}$$

$$\leq \sum_{k=1}^{K}\sum_{h=1}^{H} Q_{h,k}(s_{hk}, a_{hk}) - \mathbb{E}_p[V_{h+1,k}(s')] - r(s_{h,k}, a_{h,k}) + 2H\sqrt{KHlog(\frac{2}{\delta})}$$

$$= \sum_{k=1}^{K}\sum_{h=1}^{H} \hat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \mathbb{E}_{\hat{p}}[V_{h+1,k}(s')] - \mathbb{E}_p[V_{h+1,k}(s')] - r(s_{h,k}, a_{h,k}) + 2H\sqrt{KHlog(\frac{2}{\delta})}$$

$$\leq \sum_{k=1}^{K}\sum_{h=1}^{H} 2b_{h,k}(s,a) + 2H\sqrt{KHlog(\frac{2}{\delta})}$$

So finally:

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

- Previously, it has been shown that:

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

Then, using bonus function previously defined, it comes that:

$$R(T) \leq 2H\sqrt{2log(\frac{(2^{|S|}-2)4|S||A|HK}{\delta})}\sum_{kh}\frac{1}{\sqrt{N_{h,k}(s_{hk}, a_{hk})}} + 2H\sqrt{KH\log(2/\delta)}$$

$$R(T) \leq 2H\sqrt{2log(\frac{(2^{|S|}-2)4|S||A|HK}{\delta})}\sum_{kh}\frac{1}{\sqrt{N_{h,k}(s_{hk},a_{hk})}} + 2H\sqrt{KH\log(2/\delta)}$$

It's also known that:

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

Hence,

$$R(T) \leq 2H\sqrt{2log(\frac{(2^{|S|}-2)4|S||A|HK}{\delta})}(H^2S^2A + 2\sum_{h=1}^{H}\sum_{s,a}\sqrt{N_{hK}(s,a)}) + 2H\sqrt{KH\log(2/\delta)}$$

$$R(T) \leq 2H\sqrt{2|S|log(\frac{(8|S||A|HK}{\delta})}(H^2S^2A + 2\sum_{h=1}^{H}\sum_{s,a}\sqrt{N_{hK}(s,a)}) + 2H\sqrt{KH\log(2/\delta)}$$

Finally,

$$\boxed{\text{R(T)} \leq H^2|S|\sqrt{AK}}$$

End of Answer - Regret Minimization in RL.

End of HW3 RL - Thank you for reading!

Théo Di Piazza - theo.dipiazza@gmail.com

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

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.