MVA: Reinforcement Learning (2021/2022) **Assignment 3**

Exploration in Reinforcement Learning (theory)

Lecturers: *M. Pirotta* *(December 16, 2021)*

Solution by Megi Dervishi

Instructions

- The deadline is **January 16, 2022. 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 $\mathbb{P}_{\mu_1,\dots,\mu_k}(\widehat{i} \neq i^\star) \leq \delta$ 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

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

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.

==========**Answer:**==========

First we want to find $U(t,\delta)$ such that:

$$\mathbb{P}\left(\bigcup_{t=1}^{+\infty} |\hat{\mu}_{i,t} - \mu_i| > U(t,\delta') \right) \leq \delta' \tag{1}$$

Note that:

$$\hat{\mu}_{i,t} = \frac{1}{t}\sum_{j=1}^{t} X_{i,j} \quad \text{and} \quad \mathbb{E}[\hat{\mu}_{i,t}] = \frac{1}{t}\sum_{j=1}^{t} \mathbb{E}[X_{i,j}] = \frac{1}{t}\sum_{j=1}^{t} \mu_i = \mu_i \tag{2}$$

Hence we can re-write inequality (1) as:

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{t=1}^{+\infty} |\hat{\mu}_{i,t} - \mu_i| > U(t,\delta') \right) &\leq \sum_{t=1}^{+\infty} \mathbb{P}\left(|\hat{\mu}_{i,t} - \mu_i| > U(t,\delta') \right) \\
&= \sum_{t=1}^{+\infty} \mathbb{P}\left(|\hat{\mu}_{i,t} - \mathbb{E}[\hat{\mu}_{i,t}]| > U(t,\delta') \right) \\
&= \sum_{t=1}^{+\infty} \mathbb{P}\left(\left| \sum_{j=1}^{t} X_{i,j} - \mathbb{E}\left[\sum_{j=1}^{t} X_{i,j} \right] \right| > t\, U(t,\delta') \right)
\end{aligned} \tag{3}$$

Using Hoefding's inequality we can express (3) as:

$$\sum_{t=1}^{+\infty} 2\exp\left(-2\frac{t^2 U(t,\delta)^2}{\sum_{j=1}^{t}(b_j - a_j)^2} \right) \quad \text{where} \quad a_j < X_{i,j} < b_j \tag{4}$$

We know $X_{i,j} \in [0,1]$ which gives $a_j = 0$ and $b_j = 1$, which results to:

$$\sum_{t=1}^{+\infty} 2\exp\left(-2tU(t,\delta')^2 \right) \leq \delta' \tag{5}$$

For the series to be convergent we can use the Riemann criterion and the dominated-convergence

theorem which constrain $U(t,\delta')$ as follows

$$2\exp\left(-2tU(t,\delta')^2\right) < \frac{\delta'}{t}$$
$$2tU(t,\delta')^2 > \ln\left(\frac{2t}{\delta'}\right)$$
$$U(t,\delta') > \sqrt{\frac{1}{2t}\ln\left(\frac{2t}{\delta'}\right)} \tag{6}$$
$$\tag{7}$$

Finally we have:

$$\mathbb{P}(\mathcal{E}) \leq \sum_{i=1}^{k} \mathbb{P}\left(\bigcup_{t=1}^{+\infty} |\hat{\mu}_{i,t} - \mu_i| > U(t,\delta')\right) \leq \sum_{i=1}^{k} \delta' \leq k\delta' = \delta \tag{8}$$

- 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:**==========

For arm $i^\star = \arg\max_i\{\mu_i\}$ to remain in the active set means that

$$\bigcap_{t=1}^{+\infty}\bigcap_{j\in S_t}\{\hat{\mu}_{j,t} - U(t,\delta') < \hat{\mu}_{i^\star,t} + U(t,\delta')\} \Leftrightarrow \bigcap_{t=1}^{+\infty}\bigcap_{j\in S_t}\{\hat{\mu}_{j,t} - \hat{\mu}_{i^\star,t} < 2U(t,\delta')\} \tag{9}$$

Since $i^\star = \arg\max_i\{\mu_i\}$ we know that $\forall j,\ \mu_j \leq \mu_{i^\star}$ hence:

$$\bigcap_{t=1}^{+\infty}\bigcap_{j\in S_t}\{\hat{\mu}_{j,t} - \mu_j + \mu_j - \hat{\mu}_{i^\star,t} < 2U(t,\delta')\} \supseteq \bigcap_{t=1}^{+\infty}\bigcap_{j\in S_t}\{\hat{\mu}_{j,t} - \mu_j + \mu_{i^\star} - \hat{\mu}_{i^\star,t} < 2U(t,\delta')\} \tag{10}$$

Using the triangular inequality and the fact that $|a| > a$ we can further simplify to obtain

$$\bigcap_{t=1}^{+\infty}\bigcap_{j\in S_t}\{\hat{\mu}_{j,t} - \mu_j + \mu_{i^\star} - \hat{\mu}_{i^\star,t} < 2U(t,\delta')\} \supseteq \bigcap_{t=1}^{+\infty}\bigcap_{j\in S_t}\{|\hat{\mu}_{j,t} - \mu_j| + |\mu_{i^\star} - \hat{\mu}_{i^\star,t}| < 2U(t,\delta')\} \tag{11}$$

Finally noting that $S_t \subseteq [1,k]\cap\mathbb{N}$ we can obtain the final expression

$$\bigcap_{t=1}^{+\infty}\bigcap_{j\in S_t}\{|\hat{\mu}_{j,t} - \mu_j| + |\mu_{i^\star} - \hat{\mu}_{i^\star,t}| < 2U(t,\delta')\} \supseteq \bigcap_{t=1}^{+\infty}\bigcap_{j=1}^{k}\{|\hat{\mu}_{j,t} - \mu_j| < U(t,\delta')\} = \neg\mathcal{E} \tag{12}$$

where we used that

$$\forall t \in [1,+\infty[\cap\mathbb{N}, \forall j \in [1,k[\cap\mathbb{N}, |\hat{\mu}_{j,t} - \mu_j| < U(t,\delta') \Rightarrow \begin{cases} |\hat{\mu}_{j,t} - \mu_j| < U(t,\delta') \\ \forall t, j, \quad |\mu_{i^\star} - \hat{\mu}_{i^\star,t}| < U(t,\delta') \end{cases} \tag{13}$$
$$\Rightarrow \forall t \in [1,+\infty[\cap\mathbb{N}, \forall j \in S_t\cap\mathbb{N},\ |\hat{\mu}_{j,t} - \mu_j| + |\mu_{i^\star} - \hat{\mu}_{i^\star,t}| < 2U(t,\delta') \tag{14}$$

Hence we have that

$$\mathbb{P}\left(\bigcap_{t=1}^{+\infty}\bigcap_{j\in S_t}\{\hat{\mu}_{j,t} - \hat{\mu}_{i^\star,t} < 2U(t,\delta')\}\right) \geq \mathbb{P}(\neg\mathcal{E}) = 1-\delta \tag{15}$$

- 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:**==========

Suppose we are in the event $\neg\mathcal{E}$. Hence

$$\forall t \in [1, +\infty[\cap\mathbb{N}, \forall j \in [1,k]\cap\mathbb{N}, \quad |\hat{\mu}_{j,t} - \mu_j| < U(t,\delta') \tag{16}$$

Hence

$$\Delta_i \geq C_1 U(t,\delta') \Leftrightarrow \mu^\star - \mu_i \geq C_1 U(t,\delta') \tag{17}$$

Now from the previous question we know that $\neg\mathcal{E} \subseteq \{\text{The optimal arm is never removed.}\}$. Hence a non-optimal arm j can be removed if for some t we have that

$$\hat{\mu}_{i^\star,t} - U(t,\delta') \geq \hat{\mu}_{j,t} + U(t,\delta') \Leftrightarrow \hat{\mu}_{i^\star,t} - \hat{\mu}_{j,t} \geq 2U(t,\delta') \tag{18}$$

Now starting from the separation hypothesis

$$\begin{aligned}
\Delta_j \geq C_1 U(t,\delta') &\Leftrightarrow (\mu_{i^\star} - \hat{\mu}_{i^\star,t}) + (\hat{\mu}_{j,t} - \mu_j) + (\hat{\mu}_{i^\star,t} - \hat{\mu}_{j,t}) \geq C_1 U(t,\delta') && (19)\\
&\Leftrightarrow \hat{\mu}_{i^\star,t} - \hat{\mu}_{j,t} \geq C_1 U(t,\delta') - (\mu_{i^\star} - \hat{\mu}_{i^\star,t}) - (\hat{\mu}_{j,t} - \mu_j) && (20)\\
&\Rightarrow \hat{\mu}_{i^\star,t} - \hat{\mu}_{j,t} \geq C_1 U(t,\delta') - |\mu_{i^\star} - \hat{\mu}_{i^\star,t}| - |\hat{\mu}_{j,t} - \mu_j| && (21)\\
&\Rightarrow \hat{\mu}_{i^\star,t} - \hat{\mu}_{j,t} \geq C_1 U(t,\delta') - 2U(t,\delta') = (C_1 - 2)U(t,\delta') && (22)
\end{aligned}$$

To pass from 20 to 21 we simply use that $a < |a|$ and to pass from 21 to 22 we use 16. Hence we see that for $C_1 \geq 4$ then if $\Delta_j \geq C_1 U(t,\delta')$ the arm j will be removed. In order to have the tightest bound we take $C_1 = 4$. Now we take the previously computed explicit form of $U(t,\delta')$

$$U(t,\delta') > \sqrt{\frac{1}{2t}\ln\left(\frac{2t}{\delta'}\right)} \tag{23}$$

where $\delta' = \delta/k$, then the condition for arm j is verified if

$$\sqrt{\frac{1}{2t}\ln\left(\frac{2kt}{\delta}\right)} \leq \frac{\Delta_j}{4} \Leftrightarrow \ln\left(\frac{2kt}{\delta}\right) \leq \frac{\Delta_j^2}{8}t \tag{24}$$

Note that this is the inequality detailed in the footnote where

$$a = \frac{\Delta_j^2}{8} = \tilde{\Delta}_j^2 \quad \text{and} \quad b = \frac{2k}{\delta} \tag{25}$$

Hence we know that it can be written as

$$t \geq \frac{-W_{-1}(a/b)}{a} \tag{26}$$

and we can get a more explicit bound by setting

$$u = \log(b/a) - 1 = \log\left(\frac{2k}{\delta\tilde{\Delta}_j^2}\right) - 1 \tag{27}$$

Then

$$t \geq \frac{1 + \sqrt{2u} + u}{a} = \left(\tilde{\Delta}_j^{-2}\log\left(\frac{2k}{\delta\tilde{\Delta}_j^2}\right)\right) + \frac{\sqrt{2u}}{a} \gtrsim \left(\tilde{\Delta}_j^{-2}\log\left(\frac{2k}{\delta\tilde{\Delta}_j^2}\right)\right) \tag{28}$$

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

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

  ==========Answer:==========

  Since to eliminate arm j we need

$$t \geq \frac{1}{a_j}\left(1+\sqrt{2u_j}+u_j\right) \sim \frac{1}{\tilde{\Delta}_j^2}\left(1+\sqrt{2\log\left(\frac{1}{\tilde{\Delta}_j^2}\right)}+\log\left(\frac{1}{\tilde{\Delta}_j^2}\right)\right) \tag{29}$$

  which is a strictly decreasing function of Δ_j hence for all non-optimal arms to be supressed and to identify the optimal arm with probability $1-\delta$ we need at least

$$t \geq \frac{1}{a^\star}(1+\sqrt{2u^\star}+u^\star) \quad \text{where} \quad a^\star, u^\star \quad \text{are defined as above with} \quad \min_j \Delta_j \tag{30}$$

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

  ==========Answer:==========

  Yes it would also work however instead of being left with one optimal arm we might be left with a set of the multiple best arms even at long times.

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

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:==========

  Similarly to what was done before be start by computing the any-time confidence interval for transitions. Indeed from the Weissmain inequality

$$\mathbb{P}(||\hat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)||_1 \geq \beta^p_{hk}(s,a)) \leq (2^S - 2)\exp\left(-\frac{N_{hk}(s,a)\beta^p_{hk}(s,a)^2}{2}\right) \tag{31}$$

and similarly as before

$$\mathbb{P}(\exists k,h,s,a \,:\, ||\hat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)||_1 \geq \beta^p_{hk}(s,a)) \leq \sum_{k,h,s,a} (2^S - 2)\exp\left(-\frac{N_{hk}(s,a)\beta^p_{hk}(s,a)^2}{2}\right) \tag{32}$$

hence choosing $\beta^p_{hk}(s,a)$ such that

$$(2^S - 2)\exp\left(-\frac{N_{hk}(s,a)\beta^p_{hk}(s,a)^2}{2}\right) = \frac{\delta'_p}{HSAK} \tag{33}$$

which is readily satisfied by

$$\beta^p_{hk}(s,a) = \sqrt{\frac{2\log\left(\frac{(2^S-2)AHKS}{\delta'_p}\right)}{N_{hk}(s,a)}} \tag{34}$$

leads to

$$\mathbb{P}(\exists k,h,s,a \,:\, ||\hat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)||_1 \geq \beta^p_{hk}(s,a)) \leq \delta'_p \tag{35}$$

The exact same reasoning can be repeated for the rewards using Hoeffding's inequality instead of Weissmain analogously to what was done in the previous section (note that rewards are bounded in $[0,1]$ so the denominator simplifies same as before).

$$\mathbb{P}(|\hat{r}_{hk}(s,a) - r_h(s,a)| \geq \beta^r_{hk}(s,a)) \leq 2\exp\left(-\frac{2}{h}\beta^r_{hk}(s,a)^2\right) \tag{36}$$

and the equation

$$2\exp\left(-\frac{2}{H}\beta^r_{hk}(s,a)^2\right) = \frac{\delta'_r}{HSAK} \tag{37}$$

is solved by

$$\beta^r_{hk}(s,a) = \sqrt{\frac{H}{2}\log\left(\frac{2AHKS}{\delta'_r}\right)} \tag{38}$$

which leads to

$$\mathbb{P}(\exists k,h,s,a : |\hat{r}_{hk}(s,a) - r_h(s,a)| \geq \beta^r_{hk}(s,a)) \leq \delta'_r \tag{39}$$

Then we have that

$$\mathbb{P}(\exists k,h,s,a \,:\, ||\hat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)||_1 \geq \beta^p_{hk}(s,a) \vee |\hat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)) \leq \delta'_p + \delta'_r = \delta/2 \tag{40}$$

where we simply took $\delta'_p = \delta'_r = \delta/4$. Hence

$$\mathbb{P}(\neg\mathcal{E}) \leq \delta/2 \tag{41}$$

- 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:**==========

Since we are in event $\mathcal{E}$ we know that

$$\forall k, h, s, a, \quad |\hat{r}_{hk}(s,a) - r_h(s,a)| \leq \beta^r_{hk}(s,a) \tag{42}$$

Hence if we take

$$b_{h,k}(s,a) = \beta^r_{hk}(s,a) + f(h,k,\beta^p_{hk}(s,a)) \tag{43}$$

with only constraint that $f(H,k,\beta^p_{hk}(s,a)) \geq 0$ then

$$Q_{Hk}(s,a) = \hat{r}_{Hk}(s,a) + b_{Hk}(s,a) \geq r_H(s,a) = Q^\star_H(s,a) \tag{44}$$

Hence the property holds for $h = H$, now assume it holds for $h+1 \in [0,H] \cap \mathbb{N}$, we want to show that it holds for h. We have that

$$Q^\star_h(s,a) = r_h(s,a) + \sum_{s'} p_h(s'|s,a) V_{h+1}(s') \tag{45}$$

now we use our induction hypothesis

$$V_{h+1}(s) = \max_a Q_{h+1}(s,a) \leq \max_a Q_{h+1,k}(s,a) = V_{h+1,k}(s,a) \tag{46}$$

hence we have that

$$Q_{h,k}(s,a) - Q^\star_h(s,a) \geq (\hat{r}_{h,k}(s,a) - r_h(s,a)) + \sum_{s'} (\hat{p}_{h,k}(s'|s,a) - p_h(s'|s,a)) V_{h+1,k}(s') + b_{h,k}(s,a) \tag{47}$$

Now notice that

$$\sum_{s'} (\hat{p}_{hk}(s'|s,a) - p_h(s'|s,a)) V_{h+1,k}(s') \leq ||\hat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)||_1 ||V_{h+1,k}||_\infty \tag{48}$$

$$\leq H ||\hat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)||_1 \tag{49}$$

and since we are in event $\mathcal{E}$ we further have that

$$\sum_{s'} (\hat{p}_{hk}(s'|s,a) - p_h(s'|s,a)) V_{h+1,k}(s') \leq H \beta^p_{hk}(s,a) \tag{50}$$

Hence taking

$$b_{h,k}(s,a) = \beta^r_{h,k}(s,a) + H\beta^p_{hk}(s,a) \tag{51}$$

ensures that if we are in $\mathcal{E}$ then 47 reduces to

$$Q_{h,k}(s,a) - Q^\star_h(s,a) \geq 0 \tag{52}$$

Hence we have proved by induction that

$$\forall h, k, s, a, \quad Q_{h,k}(s,a) \geq Q^\star_h(s,a) \tag{53}$$

if we are in event $\mathcal{E}$.

- In class we have seen that

$$\delta_{1k}(s_{1,k}) \leq \sum_{h=1}^{H} Q_{hk}(s_{hk},a_{hk}) - r(s_{hk},a_{hk}) - \mathbb{E}_{Y \sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + m_{hk} \tag{54}$$

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 $V_h^{\pi_k}(s_{hk}) = r(s_{hk}, a_{hk}) + \mathbb{E}_p[V_{h+1,k}(s')] - \delta_{h+1,k}(s_{h+1,k}) - m_{h,k}$
2. Show that $V_{h,k}(s_{hk}) \leq Q_{h,k}(s_{hk}, a_{hk})$.
3. Putting everything together prove Eq. 54.

==========**Answer:**==========

1. We have

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

2. From the algorithm we see that

$$V_{h,k}(s_{hk}) = \min\{H, \max_a Q_{h,k}(s_{hk}, a)\} \leq \max_a Q_{h,k}(s_{hk}, a) = Q_{h,k}(s_{hk}, \arg\max_a Q_{h,k}(s_{hk}, a))\} \tag{55}$$

From the definition of a_{hk} we then have

$$V_{h,k}(s_{hk}) \leq Q_{h,k}(s_{hk}, a_{hk}) \tag{56}$$

3. We have

$$\delta_{1k}(s_{1k}) = V_{1k}(s_{1k}) - V_1^{\pi_k}(s_{1k}) \tag{57}$$

Now using our answer to question 1 we write

$$\delta_{1k}(s_{1k}) = V_{1k}(s_{1k}) - r(s_{1k}, a_{1k}) - \mathbb{E}_p[V_{2,k}(s')] + \delta_{2,k}(s_{2,k}) + m_{1,k} \tag{58}$$

now using Question 1 again to repeatedly expand $\delta_{2,k}(s_{2,k}), \delta_{3,k}(s_{3,k}), \cdots$ we get

$$\delta_{1,k}(s_{1,k}) = \delta_{H+1,k}(s_{H+1,k}) + \sum_{h=1}^{H} V_{hk}(s_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_p[V_{h+1,k}(s')] + m_{hk} \tag{59}$$

Now by definition we have set

$$\delta_{H+1,k}(s_{H+1,k}) = 0 \tag{60}$$

and using our answer to Question 2 we can re-write the previous equation as

$$\delta_{1,k}(s_{1,k}) \leq \sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_p[V_{h+1,k}(s')] + m_{hk} \tag{61}$$

which is the desired form.

- 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:**==========

From the definition we have

$$R(T) = \sum_{k=1}^{K} V_1^\star(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) = \sum_{k=1}^{K} \delta_{1k}(s_{1,k}) \tag{62}$$

Using our previous result we get

$$R(T) \leq \sum_{k,h} m_{h,k} + \sum_{k,h} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_p[V_{h+1,k}(s')] \tag{63}$$

$$\leq \sum_{k,h} m_{h,k} + \sum_{k,h} (\hat{r}_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk})) + \mathbb{E}_{\hat{p}-p}[V_{h+1,k}(s')] + b_{hk}(s_{hk}, a_{hk}) \tag{64}$$

from Question 2 we know that with probability $\mathbb{P}(\mathcal{E}) = 1 - \delta/2$ then

$$(\hat{r}_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk})) + \mathbb{E}_{\hat{p}-p}[V_{h+1,k}(s')] \leq b_{hk}(s_{hk}, a_{hk}) \tag{65}$$

then from Azuma-Hoeffding with probability at least $1 - \delta/2$ we get

$$\sum_{k,h} m_{hk} \leq 2H\sqrt{KH\log(2/\delta)} \tag{66}$$

hence putting the two inequalities together we have that with probability at least $1 - \delta$ we have that

$$R(T) \leq 2H\sqrt{KH\log(2/\delta)} + 2\sum_{kh} b_{hk}(s_{hk}, a_{hk}) \tag{67}$$

- 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 have that

$$\sum_{h=1}^{H}\sum_{s,a}\sqrt{N_{hK}(s,a)} \leq \sum_{h=1}^{H}\sqrt{SAN_{hK}(s,a)} \leq \sum_{h=1}^{H}\sqrt{SAK} = H\sqrt{SAK} \tag{68}$$

Now we want to bound the bonus. We have that

$$H\beta^p_{hk}(s,a) = H\sqrt{2\log\left(\frac{4(2^S-2)AHKS}{\delta}\right)}\sqrt{\frac{1}{N_{hk}(s,a)}} \quad \text{and} \quad \beta^r_{hk} = \sqrt{\frac{H}{2}\log\left(\frac{8AHKS}{\delta}\right)} \tag{69}$$

hence aplpying the inequalities given above we get

$$\sum_{kh} b_{kh}(s_{hk}, a_{hk}) \leq \sqrt{\frac{1}{2}\log\left(\frac{8AHKS}{\delta}\right)}HK\sqrt{H} + H\sqrt{2\log\left(\frac{4(2^S-2)AHKS}{\delta}\right)}(H^2S^2A + 2H\sqrt{SAK}) \tag{70}$$

$$\leq HK\sqrt{H}\sqrt{\log\frac{AHKS}{\delta} + \mathcal{O}(1)} + H\sqrt{S}\sqrt{\log\frac{AHKS}{\delta} + \mathcal{O}(1)}(H^2S^2A + 2H\sqrt{SAK}) \tag{71}$$

$$= H\sqrt{\log\frac{AHKS}{\delta} + \mathcal{O}(1)}\left(K\sqrt{H} + H^2S^{5/2}A + 2HS\sqrt{AK}\right) \tag{72}$$

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

The leading order of $R(T)$ then seems to be $H^3 S^{5/2} A$ which is not the desired behavior, however if there was no $H^2 S^2 A$ term in the originally stated inequality we would get

$$R(T) \lesssim H^2 S \sqrt{AK} \tag{73}$$

which is the desired behavior.

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.