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

Exploration in Reinforcement Learning (theory)

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

Solution by EL OUAFI Moussa

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 $\mathbb{P}_{\mu_1,\ldots,\mu_k}(\widehat{i} \neq i^\star) \leq \delta$ 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
$S = \{1, \ldots, k\}$
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

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.

Solution: The goal of this question is to find a mapping U that satisfies the any-time confidence bound i.e:

$$\forall t > 0, \mathbb{P}(|\hat{\mu}_{i,t} - \mu_i| > U(t, \delta)) \leq \frac{\delta}{2t^2}.$$

The inequality above indicates that it's preferable to *assume* that $\forall t > 0, U(t, \delta) > 0$..

The reward observed $X_{i,t}$ are bounded in $[0, 1]$, therefore using Hoeffding's inequality we get :

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

Let's assume for the moment that for any t and δ, $2\exp(-2tU(t, \delta)) = \frac{\delta}{2t^2}$ i.e $U(t, \delta) = \sqrt{\frac{\log(\frac{4t^2}{\delta})}{2t}}$.

Now let's check if,for the U considered above, exists a particular choice of δ' such that $\mathbb{P}(\mathcal{E}) \leq \delta$. Using the Hoeffding's inequality we get:

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) = \mathbb{P}(\bigcup_{i=1}^{k} \bigcup_{t=1}^{\infty} \{|\mu_{i,t} - \mu_i| > U(t, \delta')\}) &\leq \sum_{i=1}^{k} \sum_{t=1}^{\infty} \mathbb{P}(|\hat{\mu}_{i,t} - \mu_i| > U(t, \delta')) \\
&\leq \sum_{i=1}^{k} \sum_{t=1}^{\infty} 2e^{-2tU(t,\delta')^2} \qquad \textit{(Hoeffding's inequality)} \\
&\leq k \sum_{t=1}^{\infty} \frac{\delta'}{2t^2} \qquad \textit{(Condition } 2\exp(-2tU(t,\delta)^2) = \frac{\delta}{2t^2}\textit{)} \\
&\leq \frac{k\delta'\pi^2}{12} \qquad (\sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6})
\end{aligned}$$

Then in order to have the desired inequality $\mathbb{P}(\mathcal{E}) \leq \delta$ for a particular choice of δ', we only need to choose δ' such that:

$$\frac{k\delta'\pi^2}{12} \leq \delta$$

Then for $\delta' = \dfrac{12\delta}{k\pi^2}$ and $U(t, \delta) = \sqrt{\dfrac{\log(\frac{4t^2}{\delta})}{2t}}$, we get $\mathbb{P}(\mathcal{E}) \leq \delta$.

- 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 $\neg\mathcal{E}$.

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

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

In the prvious solution we proved the existence of δ' such that $\mathbb{P}(\mathcal{E}) \le \delta$, then with probability at least $1-\delta$ we have:

$$\begin{aligned}
\forall i,t, |\hat{\mu}_{i,t}-\mu_i| \le U(t,\delta') &\iff \forall i,t, -U(t,\delta') \le \hat{\mu}_{i,t}-\mu_i \le U(t,\delta')\\
&\implies \forall i,t, \mu_i \le \hat{\mu}_{i,t} + U(t,\delta')\\
&\implies \forall t, \mu^* = \mu_i^* \le \hat{\mu}_{i^*,t} + U(t,\delta')
\end{aligned}$$

By the definition of a rejected arm, the optimal arm $i^* = argmax_i\mu_i$ is eliminated from the set S if $\exists t>0, \exists j \in S\setminus\{i^*\}$ such that:

$$\hat{\mu}_{j,t} - U(t,\delta') \ge \hat{\mu}_{i^*,t} + U(t,\delta')$$

$$\begin{aligned}
&\implies \hat{\mu}_{j,t} - U(t,\delta') \ge \hat{\mu}_{i^*,t} + U(t,\delta') \ge \hat{\mu}_{i^*,t} \ge \mu^* > \mu_j\\
&\implies \hat{\mu}_{j,t} - \mu_j > U(t,\delta')
\end{aligned}$$

Therefore using the results in red color we get $\hat{\mu}_{j,t}-\mu_j > U(t,\delta')$ and $|\hat{\mu}_{j,t}-\mu_j| \le U(t,\delta')$: **contradiction**. Hence:

With probability at least $1-\delta$, the optimal arm $i^\star = \arg\max_i\{\mu_i\}$ remains in the active set S.

- Under event $\neg\mathcal{E}$, show that an arm $i \ne i^\star$ will be removed from the active set when $\Delta_i \ge C_1 U(t,\delta')$ 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]

Solution: In the previous solution we showed that under the event $\neg\mathcal{E}$, we get that at any time $t>0$, the optimal i^* remains in the set with probability at least $1-\delta$.

in order that $i \ne i^*$ be removed from the active set, by definition we must find $t>0, j\in S$ such that:

$$\hat{\mu}_{j,t} - U(t,\delta') \ge \hat{\mu}_{i,t} + U(t,\delta')$$

on the other hand, under the event $\neg\mathcal{E}$, $\hat{\mu}_{i,t}$ are in the confidence interval, therefore:

$$\forall t>0, j\in S, |\hat{\mu}_{j,t}-\mu_j| \le U(t,\delta')$$

[1]Note that $at \ge \log(bt)$ can be solved using Lambert W function. We thus have $t \ge \frac{-W_{-1}(-a/b)}{a}$ 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 $-1-\sqrt{2u}-u \le W_{-1}(-e^{-u-1}) \le -1-\sqrt{2u}-2u/3$ for $u>0$ [?]. Then $t \ge \frac{1+\sqrt{2u}+u}{a}$ with $u = \log(b/a)-1$.

Particulary, $\forall t > 0, |\hat{\mu}_{i,t} - \mu_i| \leq U(t, \delta')$ and since $i^* \in S$ we also get that $\forall t > 0, |\hat{\mu}_{i^*,t} - \mu^*| \leq U(t, \delta')$
Using the above inequalities, we get :

$$\begin{aligned}
\hat{\mu}_{i,t} - \mu_i - \hat{\mu}_{i^*,t} + \mu^* &\leq 2U(t, \delta') \\
\implies \hat{\mu}_{i,t} - \mu_i + \mu^* &\leq \hat{\mu}_{i^*,t} + 2U(t, \delta') \\
\implies \hat{\mu}_{i,t} - \mu_i + \mu^* - 3U(t, \delta') &\leq \hat{\mu}_{i^*,t} + 2U(t, \delta') - 3U(t, \delta') \\
\implies \hat{\mu}_{i,t} + U(t, \delta') - \mu_i + \mu^* - 4U(t, \delta') &\leq \hat{\mu}_{i^*,t} - U(t, \delta') \\
\implies \hat{\mu}_{i,t} + U(t, \delta') + (\Delta_i - 4U(t, \delta')) &\leq \hat{\mu}_{i^*,t} - U(t, \delta') \quad (\Delta_i = \mu^* - \mu_i)
\end{aligned}$$

Therefore :

$$\begin{aligned}
\Delta_i \geq 4U(t, \delta') &\implies \hat{\mu}_{i^*,t} - U(t, \delta') \geq \hat{\mu}_{i,t} + U(t, \delta') \\
&\implies \textbf{the arm } i \textbf{ will be removed from the set } S \textbf{ (with } C_1 = 4\textbf{).}
\end{aligned}$$

- Let's compute the time required to have such condition for each non-optimal arm:

Since $U(t, \delta') = \sqrt{\frac{\log(\frac{4t^2}{\delta'})}{2t}}$, then $\lim_{t\to\infty} U(t, \delta') = 0$. Hence there exists a time $t > 0$ such that a non-optimal arm $i \neq i^*$ will be removed from the set S.

Having,

$$\begin{aligned}
\Delta_i &\geq 4U(t, \delta') \\
\implies \frac{\Delta_i}{4} &\geq \sqrt{\frac{\log(\frac{4t^2}{\delta'})}{2t}} \\
\implies 2t(\frac{\Delta_i}{4})^2 &\geq log(\frac{4t^2}{\delta'})
\end{aligned}$$

Then the time required to have such condition for each non-optimal arm is the minimal solution t of the inequality $2t(\frac{\Delta_i}{4})^2 \geq log(\frac{4t^2}{\delta'}) \iff at \geq \log(bt)$ with $a = (\frac{\Delta_i}{4})^2$ and $b = \frac{2}{\sqrt{\delta'}}$.

using [1]Note we get for $-\frac{a}{b} \in (-\frac{1}{e}, 0)$ that $t \geq \frac{1+\sqrt{2u}+u}{a}$ with $u = log(\frac{b}{a}) - 1 = \log(\frac{32}{\sqrt{\delta'}\Delta_i^2}) - 1$.

then:

$$Time(i) = (\frac{4}{\Delta_i})^2 \times (1 + \sqrt{2(\log(\frac{32}{\sqrt{\delta'}\Delta_i^2}) - 1)} + (\log(\frac{32}{\sqrt{\delta'}\Delta_i^2}) - 1))$$

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

Solution: The algorithm stops after the all non-optimal arms $i \neq i^*$ are eliminated.

Then with probability $1-\delta$ for identifying the optimal arm i^*, the number of pulls or the excution time is given by:

$$Pulls \leq \mathcal{O}(\sum_{i \neq i^*} Time(i))$$

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

  Solution: We assumed that optimal arm i^* is unique in order to the algorithm stops when the all other non optimal arms are eliminated (i.e $|S| = 1$ id the only condition for our algorithm to stop).

  Having multiple optimal arms in the set S the algorithm won't workn it will run forever. In order for the algorithm to stop, we can modifty it to filter the sub-optimal arms like it's shown in this paper `https://arxiv.org/pdf/2006.06792.pdf`

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

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

  Solution: We want to prove that $\mathbb{P}(\neg\mathcal{E}) \le \delta/2$.

  By definition,

$$\neg\mathcal{E} = \{(S, A, p_h, r_h), |\hat{r}_{h,k}(s, a) - r_h(s, a)| > \beta^r_{h,k}(s, a) \vee \| \|\widehat{p}_{hk}(\cdot|s, a) - p_h(\cdot|s, a)\|_1 > \beta^p_{hk}(s, a), \forall k, \forall (s, a) \in S \times A\}$$

  Using Hoeffding inequality we get:

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

  Considering $2e^{(-2N_{h,k}(s,a)\beta^r_{h,k}(s,a)^2)} \le \frac{\delta}{4SAHK}$ we get:

$$-2N_{h,k}(s, a)\beta^r_{h,k}(s, a)^2 \le \log(\frac{\delta}{4SAHK})$$
$$\implies \beta^r_{h,k}(s, a) \le \sqrt{\frac{\log(\frac{8SAHK}{\delta})}{2N_{h,k}(s, a)}}$$

Then choosing $\beta^r_{h,k}(s,a) = \sqrt{\frac{\log(\frac{8SAHK}{\delta})}{2N_{h,k}(s,a)}}$ leads to:

$$\begin{aligned}
&\mathbb{P}(|\hat{r}_{h,k}(s,a) - r_h(s,a)| > \beta^r_{h,k}(s,a)) \leq 2e^{(-2N_{h,k}(s,a)\beta^r_{h,k}(s,a)^2)} \\
\implies &\mathbb{P}(|\hat{r}_{h,k}(s,a) - r_h(s,a)| > \beta^r_{h,k}(s,a)) \leq \frac{\delta}{4SAHK} \\
\implies &\mathbb{P}(|\hat{r}_{h,k}(s,a) - r_h(s,a)| > \beta^r_{h,k}(s,a)) \leq \frac{\delta}{4}
\end{aligned}$$

- On the other hand by Weissmain inequality we get:

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

Therefore if we fixed the condition, $(2^s - 2)\exp\big(-\frac{N_{h,k}(s,a)\beta^p_{h,k}(s,a)^2}{}\big) \leq \frac{\delta}{4SAHK}$ we get:

$$\begin{aligned}
&(2^s - 2)\exp\big(-\frac{N_{h,k}(s,a)\beta^p_{h,k}(s,a)^2}{}\big) \leq \log(\frac{\delta}{4SAHK}) \\
\implies &\beta^p_{h,k}(s,a) \leq \sqrt{\frac{2\log(\frac{(2^s-2)4SAHK}{\delta})}{N_{h,k}(s,a)}}
\end{aligned}$$

Then to get the inequality desired it's enough to choose

$$\beta^p_{h,k}(s,a) = \sqrt{\frac{2\log(\frac{(2^s-2)4SAHK}{\delta})}{N_{h,k}(s,a)}}$$

Therfore,

$$\mathbb{P}\Big(\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \geq \beta^p_{hk}(s,a)\Big) \leq \frac{\delta}{4SAHK} \leq \frac{\delta}{4}.$$

Combining the both proven inequalities above we get :

$$\begin{aligned}
\mathbb{P}\Big(\neg\mathcal{E}\Big) &\leq \frac{\delta}{4} + \frac{\delta}{4} \\
&\leq \frac{\delta}{2}
\end{aligned}$$

- 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.

Solution: We want to prove by induction that under event $\mathcal{E}$, Q_k is optimistic, i.e.,

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

We define the bonus function by $b_{h,k}(s,a) = \beta^r_{h,k}(s,a) = \sqrt{\frac{\log(\frac{8SAHK}{\delta})}{2N_{h,k}(s,a)}}$

- if $h = H$, then $\widehat{r}_{h,k}(s,a)+b_{h,k}(s,a) \geq r_{h,k}$ which implies that $Q_{h,k}(s,a) \geq Q^\star_h(s,a), \forall (s,a) \in S\times A$.
- Suppose that for $h < H$, $Q_{h,k}(s,a) \geq Q^\star_h(s,a) \forall (s,a) \in S \times A$.

We have $V_{h,k}(s) = \min\{H, \max_a Q_{h,k}(s,a)\}$ and $Q_{h,k}(s,a) \geq Q^\star_h(s,a)\forall (s,a) \in S \times A$.. and since $\forall k, V_{h,k}(s) \geq V^*_h(s)$. We get :

$$\begin{aligned}
Q_{h-1,k}(s,a) &= \widehat{r}_{h-1,k}(s,a) + b_{h-1,k}(s,a) \sum_{s'} \widehat{p}_{h-1,k}(s'|s,a) V_{h,k}(s') \\
&\geq \widehat{r}_{h-1,k}(s,a) + b_{h-1,k}(s,a) \sum_{s'} \widehat{p}_{h-1,k}(s'|s,a) V^*_h(s') \qquad (V_{h,k}(s) \geq V^*_h(s)) \\
&\geq r_{h-1}(s,a) + \sum_{s'} p_{h-1}(s'|s,a) V^*_h(s') \qquad (\text{under the event} \mathcal{E} = \{\forall k, M^\star \in \mathcal{M}_k\}) \\
&\geq Q^\star_{h-1}(s,a), \forall (s,a) \in S \times A \qquad (\text{ since } r_{h-1}(s,a) + \sum_{s'} p_{h-1}(s'|s,a) V^*_h(s') = Q^\star_{h-1}(s,a))
\end{aligned}$$

Hence we've proved by induction that under event $\mathcal{E}$, Q_k is optimistic, i.e.,

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

- In class we have seen that

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

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. 2.

Solution:

1. We have

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

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

   Therefore:

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

   Which proves 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. We want to show that $V_{h,k}(s_{hk}) \leq Q_{h,k}(s_{hk}, a_{hk})$.

   Recall that $V_{h,k}(s) = \min\{H, \max_a Q_{h,k}(s,a)\}$.
   Since the greedy action is $a_{h,k}$ we get :

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

3. Using the results proven in 1 and 2 we get :

$$\begin{aligned}
\delta_{1k}(s_{1,k}) &= V_{1,k}(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) \\
&\leq Q_{1,k}(s_{1,k}, a_{1,k}) - \big(r(s_{1,k}, a_{1,k}) + \mathbb{E}_p[V_{2,k}(s')] - \delta_{2,k}(s_{2,k}) - m_{1,k}\big) \\
&= Q_{1,k}(s_{1,k}, a_{1,k}) - r(s_{1,k}, a_{1,k}) - \mathbb{E}_p[V_{2,k}(s')] + \delta_{2,k}(s_{2,k}) + m_{1,k}
\end{aligned}$$

let's proove equation 1 by induction:
and $V_{H+1,k}(s') = V_1^{\pi_k}(s') = 0$, we have

$$\delta_{1k}(s_{1,k}) \leq \sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + m_{hk} \tag{2}$$

then it's easy to prove that

- 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

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

Solution: under the event $\mathcal{E}$ we have:

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

Because

$$\hat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \mathbb{E}_{\hat{p}}[V_{h+1,k}(s')]) - r(s_{h,k}, a_{h,k}) - \mathbb{E}_p[V_{h+1,k}(s')]) \leq 2b_{hk}(s_{hk}, a_{hk})$$

- Finally, we have that [?]

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

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

Solution:
In the previous solution we showed that the regret is upper bounded with probability $1 - \delta$ by

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

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

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