**MVA: Reinforcement Learning (2020/2021)** **Homework 3**

# Exploration in Reinforcement Learning (theory)

Lecturers: *A. Lazaric, M. Pirotta* *( December 10, 2020 )*

Solution by Imen AYADI

**Instructions**

- The deadline is **January 10, 2021. 23h00**
- 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 UCB

Denote by $S_{j,t} = \sum_{k=1}^{t} X_{i_k,k} \cdot \mathbb{1}(i_k = j)$ and by $N_{j,t} = \sum_{k=1}^{t} \mathbb{1}(i_k = j)$ the cumulative reward and number of pulls of arm $j$ at time $t$. Denote by $\widehat{\mu}_{j,t} = \frac{S_{j,t}}{N_{j,t}}$ the estimated mean. Recall that, at each timestep $t$, UCB plays the arm $i_t$ such that

$$i_t \in \arg\max_j \widehat{\mu}_{j,t} + U(N_{j,t}, \delta)$$

Is $\widehat{\mu}_{j,t}$ an unbiased estimator (i.e., $\mathbb{E}_{UCB}[\widehat{\mu}_{j,t}] = \mu_j$)? Justify your answer.
Solution: In general, $\widehat{\mu}_{j,t}$ a biased estimator for $\mu_j$.To show that, it is sufficient to construct a model where the bias is non null.
Assume that we have two arms $\{1, 2\}$ and (for initialisation) $i_1 = 1$ and $i_2 = 2$ as seen in the course. For $t \geq 2$, $i_{t+1} \in argmax_j \widehat{\mu}_{j,t} + U(N_{j,t})$ (actually we take the first minimizer) $U(n, t)$ decreases with respect to $n$.

- For $t = 2$:
  $N_{1,2} = 1$ ,$\widehat{\mu}_{1,2} = X_{1,1}$
  $N_{2,2} = 1$ ,$\widehat{\mu}_{2,2} = X_{2,2}$

- For $t = 3$:
  $i_3 \in argmax_{\{1,2\}}\{X_{1,1} + U(1, 2), X_{2,2} + U(1, 2)\}$
  Then,
  $$i_3 = \begin{cases} 1 & \text{if } X_{1,1} \geq X_{2,2} \\ 2 & \text{otherwise} \end{cases} \tag{1}$$
  $N_{1,3} = 1 + \mathbf{1}(X_{1,1} \geq X_{2,2})$ , $S_{1,3} = X_{1,1} + X_{1,3}\mathbf{1}(X_{1,1} \geq X_{2,2})$
  $N_{2,3} = 1 + \mathbf{1}(X_{1,1} < X_{2,2})$ , $S_{2,3} = X_{2,2} + X_{2,3}\mathbf{1}(X_{1,1} < X_{2,2})$

$$
\begin{aligned}
E(\widehat{\mu}_{1,3}) &= E(\mu_{1,3}\mathbf{1}(X_{1,1} \geq X_{2,2})\ ) + E(\mu_{2,3}\mathbf{1}(X_{1,1} < X_{2,2})\ ) \\
&= E(\frac{X_{1,1}+X_{1,3}}{2}\mathbf{1}(X_{1,1} \geq X_{2,2})\ ) + E(X_{1,1}\mathbf{1}(X_{1,1} < X_{2,2})\ ) \\
&= \frac{1}{2}E(X_{1,1}\mathbf{1}(X_{1,1} \geq X_{2,2})\ ) + \frac{\mu_1}{2}P(X_{1,1} \geq X_{2,2})\ ) + E(X_{1,1}\mathbf{1}(X_{1,1} < X_{2,2})\ ) \\
&= -\frac{1}{2}E(X_{1,1}\mathbf{1}(X_{1,1} \geq X_{2,2})\ ) + \mu_1\left(1+\frac{1}{2}P(X_{1,1} \geq X_{2,2})\right)
\end{aligned}
\tag{2}
$$

where we used independence of $(X_{j,t})_{j,t}$ to get the third equality.
We assume that $X_{j,t}$ $\mathcal{B}(\mu_j)$ i.e.$P(X_{j,t}=1)=\mu_j$ and $P(X_{j,t}=0)=1-\mu_j$.
Then,

$$
\begin{aligned}
E(X_{1,1}\mathbf{1}(X_{1,1} \geq X_{2,2}) &= E(X_{1,1}\mathbf{1}(X_{1,1}=1) \\
&= E(\mathbf{1}(X_{1,1}=1)) \\
&= P(X_{1,1}=1) \\
&= \mu_1
\end{aligned}
\tag{3}
$$

And,

$$
P(X_{1,1} \geq X_{2,2}) = P(X_{1,1}=1) = \mu_1 \tag{4}
$$

Finally, we get $E(\widehat{\mu}_{1,3}) = \frac{\mu_1(\mu_1+1)}{2}$.
If we assume that $\mu_1 \neq 0$ and $\mu_1 \neq 1$, then, $\frac{\mu_1(\mu_1+1)}{2} \neq \mu_1$
Hence, $\widehat{\mu}_{1,3}$ is a biased estimator of $\mu_1$.

## 2 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 $\mu_i$. 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.

**$\delta$-correctness and fixed-confidence objective.** Denote by $\tau_\delta$ 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 $\delta$-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 $\delta$-correct algorithm that minimizes the sample complexity, that is, $\mathbb{E}[\tau_\delta]$ the expected number of sample needed to predict an answer.

Notation

- $I_t$: the arm chosen at round $t$.
- $X_{i,t} \in [0,1]$: reward observed for arm $i$ at round $t$.
- $\mu_i$: 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 $\widehat{\mu}_{i,t} = \frac{1}{t}\sum_{j=1}^{t} X_{i,j}$.

**Input:** $k$ arms, confidence $\delta$
$S = \{1, \dots, k\}$
**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**

- Compute the function $U(t,\delta)$ that satisfy the any-time confidence bound. For any arm $i \in [k]$

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

Use Hoeffding's inequality.

- Solution: Fist, let us recall the Hoeffding's inequality:
Let $(Y_i)_{1 \leq i \leq n}$ be $n$ independant rv. with means $E(Y_i)$ and such that $Y_i \in [a_i, b_i]$, then $\forall \epsilon > 0$:

$$P(|\sum_{i=1}^{n} (Y_i - E(Y_i))| \geq \epsilon) \leq 2 \ exp\left( -\frac{2\epsilon^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right)$$

We apply this inequality and we get $\forall t \in \mathbb{N}^*$ and $\forall \epsilon > 0$:

$$\begin{aligned} P(|\widehat{\mu}_{i,t} - \mu_i| \geq U(t,\delta)) &= P(|\frac{1}{t} \sum_{j=1}^{t} (X_{i,j} - E(X_{i,j})| \geq U(t,\delta)) \\ &= P(|\sum_{j=1}^{t} (X_{i,j} - E(X_{i,j})| \geq tU(t,\delta)) \\ &\leq 2 \ exp\left( -\frac{2t^2\epsilon^2}{\sum_{j=1}^{t} (1-0)^2} \right) \\ &= 2 \ exp(-2t \ U(t,\delta)^2) \end{aligned} \tag{5}$$

Therefore,

$$\begin{aligned} P\left( \bigcup_{t=1}^{\infty} \{ |\widehat{\mu}_{i,t} - \mu_i| \geq U(t,\delta) \} \right) &\leq \sum_{t=1}^{\infty} P(|\widehat{\mu}_{i,t} - \mu_i| \geq U(t,\delta)) \\ &\leq 2 \sum_{t=1}^{\infty} exp(-2t \ U(t,\delta)^2) \end{aligned} \tag{6}$$

To have convergence of the sum, we can take $U(t,\delta)$ independant from $t$. However, we are interested to functions $U(t,\delta)$ that decreases with $t$ .
I choose $exp(-2tU(t,\delta)^2) = \frac{3\delta}{\pi^2} \frac{1}{t^2}$ (in fact, $\frac{\pi^2}{6} = \sum_{n=1}^{\infty} \frac{1}{n^2}$) i.e. $-2tU(t,\delta)^2 = log(\frac{3\delta}{\pi^2}) - 2log(t)$.
Notice that $\forall t \in \mathbb{N}^*, \ t^2 \geq 1 \geq \frac{3}{\pi^2} \geq \frac{3\delta}{\pi^2}$. Therefore, our expression is well defined.
We get finally, $U(t,\delta) = \sqrt{\frac{2log(t) + log(\frac{\pi^2}{3\delta})}{2t}}$ wih is decreasing with respect to $t \in \mathbb{N}$
0, 1.
Since $\forall t \geq 1, \ log(t) \leq t$, then, $\sqrt{\frac{2log(t) + log(\frac{\pi^2}{3\delta})}{2t}} \leq \sqrt{1 + \frac{log(\frac{\pi^2}{3\delta})}{2t}} \leq \sqrt{1 + \frac{log(\frac{4}{\delta})}{2t}}$.

Notice that $t \mapsto \sqrt{1+\frac{log(\frac{4}{\delta})}{2t}}$ is decreasing over $[1,+\infty[$.
Therefore,

$$\begin{aligned}
P\left(\bigcup_{t=1}^{\infty}\left\{|\widehat{\mu}_{i,t}-\mu_i| > \sqrt{1+\frac{log(\frac{4}{\delta})}{2t}}\right\}\right) &\leq P\left(\bigcup_{t=1}^{\infty}\left\{|\widehat{\mu}_{i,t}-\mu_i| \geq \sqrt{\frac{2log(t)+log(\frac{\pi^2}{3\delta})}{2t}}\right\}\right) \\
&\leq 2\sum_{t=1}^{\infty}\frac{3\delta}{\pi^2}\frac{1}{t^2} \\
&= \delta
\end{aligned} \tag{7}$$

Hence, we can choose $\boxed{U(t,\delta)=\sqrt{1+\frac{log(\frac{4}{\delta})}{2t}}}$

It is evident that there are several possible choices for $U(t,\delta)$. We could optimize the choice bu picking up the "most" decreasing functions ensuring the convergence of the sum.

- Let $\mathcal{E}=\bigcup_{i=1}^{k}\bigcup_{t=1}^{\infty}\{|\widehat{\mu}_{i,t}-\mu_i|>U(t,\delta')\}$. Using previous result shows that $\mathbb{P}(\mathcal{E})\leq\delta$ for a particular choice of $\delta'$. This is called "bad event" since it means that the confidence intervals do not hold.

- Solution:

,

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

Hence, we can choose $\boxed{\delta'=\frac{\delta}{k}}$

- 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 $\delta'$ and start from the condition for arm elimination. From this, use the definition of $\neg\mathcal{E}$.

- Solution: The update of the active set at iteration $t+1$ is given by:

$$S_{t+1}=S_t\setminus\left\{i\in S_t\ :\ \exists j\in S_t,\ \widehat{\mu}_{j,t}-U(t,\delta)>\widehat{\mu}_{i,t}+U(t,\delta)\right\}$$

Then,

$$\begin{aligned}
P(i^*\text{ remains in } S) &= P(\forall t\in\mathbb{N}^*,\forall i\in S_t,\ \widehat{\mu}_{i,t}-U(t,\delta')\leq\widehat{\mu}_{i^*,t}+U(t,\delta')) \\
&= 1-P(\exists t\in\mathbb{N}^*,\ \exists i\in S_t\ ,\widehat{\mu}_{i,t}-\widehat{\mu}_{i^*,t}>2U(t,\delta')) \\
&= 1-P(\bigcup_{t=1}^{\infty}\bigcup_{i\in S_t}\widehat{\mu}_{i,t}-\widehat{\mu}_{i^*,t}>2U(t,\delta'))
\end{aligned} \tag{9}$$

We notice that for $1\leq i\leq n$:

$$\widehat{\mu}_{i,t}-\mu_{i^*,t}=(\widehat{\mu}_{i,t}-\mu_i)-(\widehat{\mu}_{i^*,t}-\mu_{i^*})+(\mu_i-\mu_{i^*})\leq(\widehat{\mu}_{i,t}-\mu_i)-(\widehat{\mu}_{i^*,t}-\mu_{i^*})\leq|\widehat{\mu}_{i,t}-\mu_i|+|\widehat{\mu}_{i^*,t}-\mu_{i^*}| \tag{10}$$

where we used the fact that $i^*=argmax_i\mu_i$. Therefore,

$$\begin{aligned}
\{\widehat{\mu}_{i,t}-\widehat{\mu}_{i^*,t}>2U(t,\delta')\} &\subseteq \{|\widehat{\mu}_{i,t}-\mu_i|+|\widehat{\mu}_{i^*,t}-\mu_{i^*}|>2U(t,\delta')\} \\
&\subseteq \{|\widehat{\mu}_{i,t}-\mu_i|>U(t,\delta')\}\cup\{|\widehat{\mu}_{i^*,t}-\mu_{i^*}|>U(t,\delta')\}
\end{aligned} \tag{11}$$

Hence,

$$\begin{aligned}
\bigcup_{i\in S}\bigcup_{t=1}^{\infty}\{\widehat{\mu}_{i,t}-\widehat{\mu}_{i^*,t}>2U(t,\delta')\} &\subseteq \bigcup_{i=1}^{k}\bigcup_{t=1}^{\infty}\{\widehat{\mu}_{i,t}-\widehat{\mu}_{i^*,t}>2U(t,\delta')\}\\
&\subseteq \bigcup_{i=1}^{k}\bigcup_{t=1}^{\infty}\{|\widehat{\mu}_{i,t}-\mu_i|>U(t,\delta')\}\cup\{|\widehat{\mu}_{i^*,t}-\mu_i|>U(t,\delta')\}\\
&= \bigcup_{t=1}^{\infty}(\bigcup_{i=1}^{k}\{|\widehat{\mu}_{i,t}-\mu_i|>U(t,\delta')\}\cup\{|\mu_{i^*,t}-\mu_i|>U(t,\delta')\})\\
&= \bigcup_{t=1}^{\infty}\bigcup_{i=1}^{k}\{|\widehat{\mu}_{i,t}-\mu_i|>U(t,\delta')\} \quad \text{(because there is } i \text{ equal to } i^*)\\
&= \mathcal{E}
\end{aligned} \tag{12}$$

Then, $P(\bigcup_{i=1}^{k}\bigcup_{t=1}^{\infty}\{\widehat{\mu}_{i,t}-\widehat{\mu}_{i^*,t}>2U(t,\delta')\}) \leq P(\mathcal{E}) \leq \delta$. Finally, we get: $\boxed{P(i^* \text{ remains in } S) \geq 1-\delta}$

- Under event $\neg\mathcal{E}$, show that an arm $i \neq i^\star$ will be removed from the active set when $\Delta_i \geq C_1 U(t,\delta')$ where $C_1 > 1$ is a constant. Compute the time required to have such condition for each non-optimal arm. Use the condition of arm elimination applied to arm $i^\star$.

- Solution : Recall that

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

We proved in the previous question that under the event $\neg\mathcal{E}$, $i^*$ remains i.e

$$\forall t \in \mathbb{N}^2*,\ \forall j \in S_t,\ \widehat{\mu}_{j,t}-\widehat{\mu}_{i^*,t} \leq 2U(t,\delta')$$

Let an arm $i \neq i^*$.Suppose that we are under $\neg\mathcal{E}$ and there exists $C_1 > 4$ and an iteration $s \in \mathbb{N}^*$ such that $\Delta_i \geq C_1 U(s,\delta')$.
We want to prove that $i$ is removed from the active set at the iteration $t$ .
Then, at iteration $t$ :

$$\begin{aligned}
\widehat{\mu}_{i,t}-\widehat{\mu}_{i,t} &= (\widehat{\mu}_{i^*,t}-\mu^*)+(\mu^*-\mu_i)-(\widehat{\mu}_{i,t}-\mu_i)\\
&= (\widehat{\mu}_{i^*,t}-\mu^*)+\Delta_i-(\widehat{\mu}_{i,t}-\mu_i)\\
&\geq (\widehat{\mu}_{i^*,t}-\mu^*)+C_1U(t,\delta')-(\widehat{\mu}_{i,t}-\mu_i)
\end{aligned} \tag{13}$$

Since $\widehat{\mu}_{i^*,t}-\mu^* \geq -|\widehat{\mu}_{i^*,t}-\mu^*| \geq -U(t,\delta')$ and $-(\widehat{\mu}_{i,t}-\mu_i) \geq -|\widehat{\mu}_{i,t}-\mu_i| \geq -U(t,\delta')$ then,

$$\widehat{\mu}_{i,t}-\widehat{\mu}_{i^*,t} \geq (C_1-2)U(t,\delta') > 2U(t,\delta') \tag{14}$$

Therfore,we conclude $\boxed{\text{Under } \neg\mathcal{E}, \text{ when } \Delta_i \geq C_1U(t,\delta'), \text{ the arm } i \text{ is removed from the active set}}$
Let $\tau_i$ denote the time required to have such condition for each non-optimal arm $i \neq i^*$. Then , $\tau_i = min\{t \in \mathbb{N}^* | U(t,\delta') \leq \frac{\Delta_i}{C_1}\}$.
Recall that $x \mapsto U(x,\delta')$ is a strictly decreasing function and $\forall t \geq 1, U(t,\delta') < \underset{x\to\infty}{lim} U(x,\delta') = 1$

- If $\frac{\Delta_i}{C_1} \geq 1$,then, $\tau_i = +\infty$

- If $\frac{\Delta_i}{C_1} < 1$:

$$\begin{aligned} U(x,\delta') = \frac{\Delta_i}{C_1} &\iff \sqrt{1 + \frac{log(\frac{4}{\delta})}{2x}} = \frac{\Delta_i}{C_1} \\ &\iff \frac{log(\frac{4}{\delta})}{2x} = \left(\frac{\Delta_i}{C_1}\right)^2 - 1 \\ &\iff x = \frac{log(\frac{4}{\delta})}{2\left(\frac{\Delta_i}{C_1}\right)^2 - 2} \end{aligned} \tag{15}$$

Therefore, $\tau_i = Ent\left(\frac{log(\frac{4}{\delta})}{2\left(\frac{\Delta_i}{C_1}\right)^2-2}\right)$ where $Ent(a)$ is the smallest integer $m$ such that $m \geq a$.
Hence, under $\neg\mathcal{E}$:

$$\boxed{\tau_i = \begin{cases} Ent\left(\frac{log(\frac{4}{\delta})}{2\left(\frac{\Delta_i}{C_1}\right)^2-2}\right) & \text{if } \frac{\Delta_i}{C_1} < 1 \\ +\infty & \text{otherwise} \end{cases}} \tag{16}$$

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

- Solution: In this question, I assume that $\forall i, \Delta_i < C_1$.
Now, let us express the stopping time $\tau$of the algorithm.

$$\begin{aligned} \{\tau \leq t\} &= \{\forall i \neq i^*, i \text{ has been removed before the iteration } t\} \\ &\supseteq \{\forall i \neq i^*, \tau_i \leq t\} \end{aligned} \tag{17}$$

We notice that:

$$\begin{aligned} P(\forall i \neq i^*, \tau_i \leq t) &= 1 - P(\bigcup_{i\neq i^*} \tau_i > t) \\ &\geq 1 - \sum_{i\neq i^*} P(\tau_i > t) \end{aligned} \tag{18}$$

with :

$$\begin{aligned} P(\tau_i > t) &= P(\tau_i > t|\neg\mathcal{E})P(\neg\mathcal{E}) + P(\tau_i > t|\mathcal{E})P(\mathcal{E}) \\ &= \mathbf{1}_{\left\{Ent\left(\frac{log(\frac{4}{\delta})}{2\left(\frac{\Delta_i}{C_1}\right)^2-2}\right) > t\right\}} P(\neg\mathcal{E}) + P(\tau_i > t|\mathcal{E})P(\mathcal{E}) \\ &\leq \mathbf{1}_{\left\{Ent\left(\frac{log(\frac{4}{\delta})}{2\left(\frac{\Delta_i}{C_1}\right)^2-2}\right) > t\right\}} P(\neg\mathcal{E}) + \delta \end{aligned} \tag{19}$$

Thus, if we take $t = max_{i\neq i^*} Ent\left(\frac{log(\frac{4}{\delta})}{2\left(\frac{\Delta_i}{C_1}\right)^2-2}\right) = Ent\left(\frac{log(\frac{4}{\delta})}{2\left(\frac{min_{i\neq i^*}\Delta_i}{C_1}\right)^2-2}\right)$:

$$\begin{aligned} P\left(\tau \leq Ent\left(\frac{log(\frac{4}{\delta})}{2\left(\frac{\Delta_i}{C_1}\right)^2 - 2}\right)\right) &\geq 1 - \sum_{i\neq i^*} [0 \times P(\neg\mathcal{E}) + \delta] \\ &= 1 - k\delta \end{aligned} \tag{20}$$

Unless we replace $\delta$ by $\frac{\delta}{k}$, we get:

$$P\left(\tau \leq Ent\left(\frac{log(\frac{4k}{\delta})}{2\left(\frac{min_{i\neq i^*}\Delta_i}{C_1}\right)^2 - 2}\right)\right) \geq 1-\delta \tag{21}$$

Thus, a possible bound on the sample complexity for identifying the optimal arm w.p. $1-\delta$ is

$$\boxed{Ent\left(\frac{log(\frac{4k}{\delta})}{2\left(\frac{min_{i\neq i^*}\Delta_i}{C_1}\right)^2 - 2}\right) = O\left(log\left(\frac{k}{\delta}\right)\right)}$$

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

## 3 Bernoulli Bandits

In this exercise, you compare KL-UCB and UCB empirically with Bernoulli rewards $X_t \sim Bern(\mu_{I_t})$.

- Implement KL-UCB and UCB

**KL-UCB:**

$$I_t = \arg\max_i \max\left\{\mu \in [0,1] : d(\widehat{\mu}_{i,t}, \mu) \leq \frac{\log(1 + t\log^2(t))}{N_{i,t}}\right\}$$

where $d$ is the Kullback–Leibler divergence (see closed form for Bernoulli). A way of computing the inner max is through bisection (finding the zero of a function).

**UCB:**

$$I_t = \arg\max_i \widehat{\mu}_{i,t} + \sqrt{\frac{\log(1 + t\log^2(t))}{2N_{i,t}}}$$

that has been tuned for 1/2-subgaussian problems.

- The function $d$ is given by:

$$d(p,q) = plog\left(\frac{p}{q}\right) + (1-p)log\left(\frac{1-p}{1-q}\right) \tag{22}$$

Let $f_p = d(p,.)$ a function over $]0,1[$.
$f_p'(q) = -\frac{p}{q} + \frac{1-p}{1-q} = \frac{q-p}{q(1-q)}$.

| $q$ | $0$ | | $p$ | | $1$ |
|---|---|---|---|---|---|
| $f_p'(q)$ | | $-$ | $0$ | $+$ | |
| $f_p'(q)$ | $+\infty$ | $\searrow$ | $0$ | $\nearrow$ | $+\infty$ |

Therefore, $\forall b > 0$, there exists $q_1, q_2$ such that $0 < q_1 < p < q_2 < 1$ and $f_p(q_1) = f_p(q_2) = b$.
Thus, $max\{q | f_p(q) \leq b\} = q_2$.
Since $q_2 > p$, we can do a dichotomous search over $]p, 1[$ to get the zero of $f_p - b$.
Please find the code in the notebook `RL3imenAyadi.ipynb`.

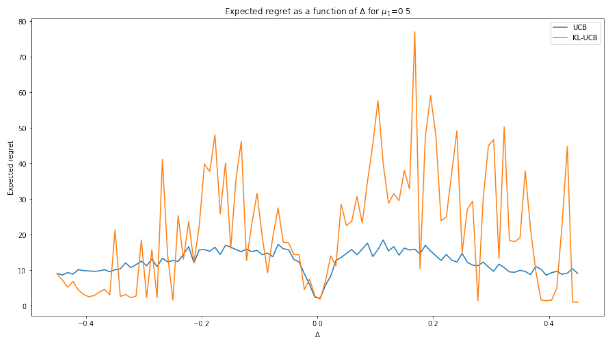


Figure 1: $\mu_1 = 0.5$

- Let $n = 10000$ and $k = 2$. Plot the expected regret of each algorithm as a function of $\Delta$ when $\mu_1 = 1/2$ and $\mu_2 = 1/2 + \Delta$.
- Solution : To obtain an expectation of the regret, I average over several executions.See 1
- Repeat the above experiment with $\mu_1 = 1/10$ and $\mu_1 = 9/10$.

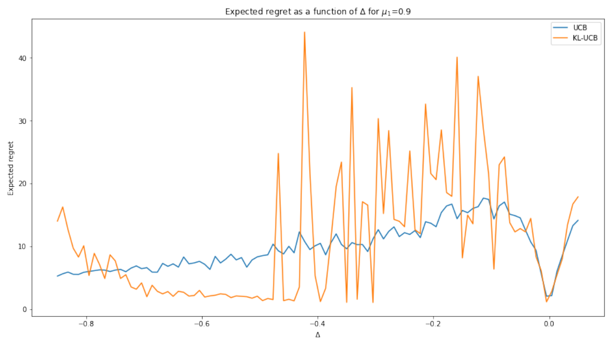


Figure 2: $\mu_1 = 0.9$

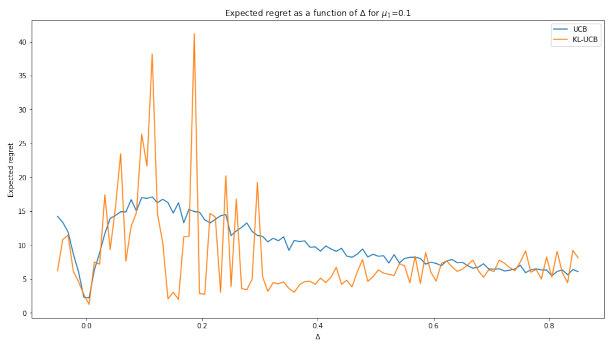


Figure 3: $\mu_1 = 0.1$

- Discuss your results.
- Solution:
  *When $\mu_1$ and $\mu_2$ are close (i.e $\Delta \approx 0$ ), we have a little value for the expected regret. The UCB and KL-UCB yields almost the same values of the expected regret.

*When $\mu_1$ and $\mu_2$ are far from each other, we have a big value for the expected regret
*This is homogeneous with the theoretical bound: papers stated that $E(R(n))$ is bounded by $O(log(n)/\sum$ *In theory, KL-UCB should outperform UCB since : $d_{Kl}(p,q) \geq 2(p-q)^2$. But, my plots show the opposite : there must be an error in my code. Actually, I think there is an error in my implementation. The evolution of the expected regret with UCB as a function of $\Delta$ seems to be invariant of $\mu_1$ : I get almost the same shape of the curse. This is not the case of KL-UCB, where the expected regret seems to be larger when $\mu_1 =$ . Furthermore, the curve of the KL-UCB presents a lot of fluctuations. these fluctuations seem to be smaller for $\mu_1 = 0.9,\ \mu_2 \leq 0.4$ and $\mu_1 = 0.1,\ \mu_2 \geq 0.3$. For these cases, KL-UCB outperforms the UCB as seen in the plots.
*Remark: Maybe the error induced by the bisection algorithm impacts on the accuracy of the KL-UCB algorithm.
* We notice that the situation when $\mu_1 = 0.1$ and $\mu_2$ is low but not very close to $\mu_1$ is the most difficult scenario.
*Maybe I had to average on more runs (here I took 100 executions) to smooth the fluctuations. It could be more rigorous to compute the estimated standard deviation for each algorithm.

# 4 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

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

- Solution :
We have

$$\begin{aligned}\neg\mathcal{E} &= \Big\{\exists k,h,s,a : |r_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a) \text{ or } \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 > \beta^p_{hk}(s,a)\Big\} \\ &= \bigcup_{k,h,s,a} \{|r_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)\} \cup \{\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 > \beta^p_{hk}(s,a)\}\end{aligned} \tag{23}$$

Therefore,

$$P(\neg\mathcal{E}) \leq \sum_{k,h,s,a} P(|r_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)) + P(\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 > \beta^p_{hk}(s,a)) \tag{24}$$

Using the Hoeffding's inequality (the rewards are in $[0,1]$), we get:

$$\begin{aligned}
P(|r_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)) &= P\left(\left|\frac{\sum_{i=1}^{k-1} r_{hi} \cdot \mathbb{1}_{\{(s_{hi},a_{hi})=(s,a)\}}}{N_{hk}(s,a)} - r_h(s,a)\right| > \beta^r_{hk}(s,a)\right) \\
&= P\left(\left|\sum_{i=1}^{k-1} [r_{hi} - r_h(s,a)] \cdot \mathbb{1}_{\{(s_{hi},a_{hi})=(s,a)\}}\right| > N_{hk}(s,a)\beta^r_{hk}(s,a)\right) \\
&\leq 2exp\left(\frac{-2\ \ [N_{hk}(s,a)\beta^r_{hk}(s,a)]^2}{\sum_{i=1}^{k-1}[(1-0)^2 \mathbb{1}_{\{(s_{hi},a_{hi})=(s,a)\}}]}\right) \\
&= 2exp\left(\frac{-2\ \ [N_{hk}(s,a)\beta^r_{hk}(s,a)]^2}{N_{hk}(s,a)}\right) \\
&= 2exp(-2\ N_{hk}(s,a)\ \beta^r_{hk}(s,a)^2)
\end{aligned} \tag{25}$$

If we choose $\beta^r_{h,k}(s,a) = \sqrt{\frac{\log(8\ SAHK/\delta)}{2N_{h,k}(s,a)}}$,then:

$$P(|r_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)) \leq 2exp(-\log(8\ SAHK/\delta)) = \frac{\delta}{4SAHK} \tag{26}$$

Using the Hoeffding and Weissmain's inequality, we get :

$$P(\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 > \beta^p_{hk}(s,a)) \leq (2^S-2)\exp\Big(-\frac{N_{h,k}(s,a)\beta^p_{hk}(s,a)^2}{2}\Big) \tag{27}$$

If we choose $\beta^p_{hk} = \sqrt{\frac{2\log(4(2^S-2)SAHK/\delta)}{N_{h,k}(s,a)}}$,we get:

$$\begin{aligned}
P(\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 > \beta^p_{hk}(s,a)) &\leq (2^S-2)\exp\Big(-\frac{N_{h,k}(s,a)\beta^p_{hk}(s,a)^2}{2}\Big) \\
&= (2^S-2)\exp\Big(-log\left(\frac{4(2^S-2)SAHK}{\delta}\right)\Big) \\
&= \frac{\delta}{4\ SAHK}
\end{aligned} \tag{28}$$

Using eq:11, eq:12 and eq:13, we have:

$$P(\neg\mathcal{E}) \leq \sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{s\in S}\sum_{a\in A} \frac{\delta}{4\ SAHK} + \frac{\delta}{4\ SAHK} = \frac{\delta}{2} \tag{29}$$

Thus, $\boxed{P(\neg\mathcal{E}) \leq \frac{\delta}{2}}$

- 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 $h$.

- Solution: The inductive hypothesis is $\mathcal{H}_h := Q_{h,k}(s,a) \geq Q_h^\star(s,a), \forall s,a$
  We use backward induction over $h$ to prove the result.
  *Initialisation: For $h = H+1$, we have already $V_{H+1,k}(s) = V_{H+1}^\star(s) = 0$, which implies that $max_a Q_{H+1,k}(s,a) = max_a Q_{H+1}^*(s,a) = 0 \, \forall s$. Since the Q-function is positive, then, $Q_{H+1,k}(s,a) = Q_{H+1}^*(s,a) = 0 \, \forall s,a$. In particular ,$Q_{H+1,k}(s,a) \geq Q_{H+1}^*(s,a) = 0 \, \forall s,a$. We could start the induction from $H$ since it is proved in the question.
  *Heredity: Suppose that $Q_{h+1,k}(s,a) \geq Q_{h+1}^\star(s,a), \forall s,a$. Let us prove that$Q_{h,k}(s,a) \geq Q_h^\star(s,a), \forall s,a$.
  From the inductive hypothesis at stage $h+1$, we deduce that $max_a Q_{h+1,k}(s) \geq max_a Q_{h+1}^\star(s)$.
  Therefore,$V_{h+1,k}(s) \geq V_{h+1}^\star(s) \;\; \forall s$.
  Let $s \in S$, $a \in A$ and $k \in \{1,..K\}$. We have :

$$\begin{aligned} Q_{h,k}(s,a) - Q_h^*(s,a) &= \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \sum_{s'} \widehat{p}_{h,k}(s'|s,a) V_{h+1,k}(s') - \widehat{r}_h(s,a) - \sum_{s'} p_h(s'|s,a) V_{h+1}^*(s') \\ &\geq \widehat{r}_{h,k}(s,a) - r_h(s,a) + b_{h,k}(s,a) + \sum_{s'} \widehat{p}_{h,k}(s'|s,a) V_{h+1}^*(s') - \sum_{s'} p_h(s'|s,a) V_{h+1}^*(s') \\ &= \widehat{r}_{h,k}(s,a) - r_h(s,a) + b_{h,k}(s,a) + \sum_{s'} [\widehat{p}_{h,k}(s'|s,a) - p_h(s'|s,a)] V_{h+1}^*(s') \end{aligned} \tag{30}$$

Using Holder's inequality,

$$|\sum_{s'} [\widehat{p}_{h,k}(s'|s,a) - p_h(s'|s,a)] V_{h+1}^*(s')| \leq ||\widehat{p}_{h,k}(.|s,a) - p_h(.|s,a)||_1 \; ||V_{h+1}^*(.)||_\infty \leq H ||\widehat{p}_{h,k}(.|s,a) - p_h(.|s,a)||_1 \tag{31}$$

Under the event $\mathcal{E}$,

$$\forall k,h,s,a : |r_{hk}(s,a) - r_h(s,a)| \leq \beta_{hk}^r(s,a) \wedge \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \leq \beta_{hk}^p(s,a)$$

Then, if we choose the bonus properly (for example, $b_{h,k}(s,a) := \beta_{hk}^r(s,a) + H\beta_{hk}^p(s,a)$), we get that:

$$|r_{hk}(s,a) - r_h(s,a)| + H \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \leq b_{h,k}^p(s,a) \tag{32}$$

Thus,

$$\begin{aligned} (r_{hk}(s,a) - r_h(s,a)) + \sum_{s'} [\widehat{p}_{h,k}(s'|s,a) - p_h(s'|s,a)] V_{h+1}^*(s') &\geq -|r_{hk}(s,a) - r_h(s,a)| - H \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \\ &\geq -b_{h,k}^p(s,a) \end{aligned} \tag{33}$$

Finally, we plug 33 in 30 to obtain:

$$Q_{h,k}(s,a) - Q_h^*(s,a) \geq b_{h,k}^p(s,a) - b_{h,k}^p(s,a) = 0 \tag{34}$$

Consequently, $\boxed{Q_{h,k}(s,a) \geq Q_h^*(s,a) \; \forall s,a}$

- In class we have seen that

$$\delta_{1k}(s_{1,k}) \leq \sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y \sim p(\cdot|s_{hk}, a_{hk})}[V_{h+1,k}(Y)]) + m_{hk} \tag{35}$$

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

– Solution:

We recall that:

$$\begin{aligned}\delta_{h+1,k}(s_{h+1,k}) + m_{h,k} &= \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[\delta_{h+1,k}(Y)] \\ &= \sum_{s'} p_h(s'|s_{hk},a_{hk})\delta_{h+1,k}(s') \\ &= \sum_{s'} p_h(s'|s_{hk},a_{hk})[V_{h+1,k}(s') - V_h^{\pi_k}(s')]\end{aligned} \tag{36}$$

and

$$\mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)] = \sum_{s'} p_h(s'|s_{hk},a_{hk})V_{h+1,k}(s') \tag{37}$$

Then,

$$\begin{aligned}r(s_{hk},a_{hk}) + \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(s')] - \delta_{h+1,k}(s_{h+1,k}) - m_{h,k} \\ = r(s_{hk},a_{hk}) + \sum_{s'} p_h(s'|s_{hk},a_{hk})V_h^{\pi_k}(s') \\ = V_h^{\pi_k}(s_{hk})\end{aligned} \tag{38}$$

where the last inequality holds thanks to the Belleman equation.
Therefore, $\boxed{V_h^{\pi_k}(s_{hk}) = r(s_{hk},a_{hk}) + \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(s')] - \delta_{h+1,k}(s_{h+1,k}) - m_{h,k}}$

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

– Solution:

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

3. Putting everything together prove Eq. 35.

– Solution:

$$\begin{aligned}\delta_{1k}(s_{1,k}) - \delta_{H+1,k}(s_{H+1,k}) &= \sum_{h=1}^{H} \delta_{hk}(s_{h,k}) - \delta_{h+1,k}(s_{h+1,k}) \\ &= \sum_{h=1}^{H}[V_{hk}(s_{h,k}) - V_h^{\pi_k}(s_{hk})] - \delta_{h+1,k}(s_{h+1,k}) \\ &= \sum_{h=1}^{H} V_{hk}(s_{h,k}) - r(s_{hk},a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(s')] + m_{h,k} \\ &\leq \sum_{h=1}^{H} Q_{hk}(s_{h,k},a_{h,k}) - r(s_{hk},a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(s')] + m_{h,k}\end{aligned} \tag{39}$$

Recall that $\delta_{H+1,k}(s) = V_{H+1,k}(s) - V_{H+1}^{\pi_k}(s) = 0 - V_{H+1}^{\pi_k}(s) \leq 0$.
Then, $\delta_{1k}(s_{1,k}) \leq \delta_{1k}(s_{1,k}) - \delta_{H+1,k}(s_{H+1,k})$.
Hence, $\boxed{\delta_{1k}(s_{1,k}) \leq \sum_{h=1}^{H} Q_{hk}(s_{h,k},a_{h,k}) - r(s_{hk},a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(s')] + m_{h,k}}$

- 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

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

- Solution:
  Under the event $\mathcal{E}$:

$$\begin{aligned} R(T) &= \sum_{k=1}^{K} V_1^{\star}(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) \\ &= \sum_{k=1}^{K} (V_1^{\star}(s_{1,k}) - V_{1,k}(s_{1,k})) + (V_{1,k}(s_{1,k}) - V_1^{\pi_k}(s_{1,k})) \\ &\leq \sum_{k=1}^{K} V_{1,k}(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) = \sum_{k=1}^{K} \delta_{1,k}(s_{1,k}) \end{aligned} \tag{40}$$

because we proved that under $\mathcal{E}$ , $\forall h, k, s, a$ $Q_{h,k}(s,a) \geq Q^*_{h,k}(s,a)$. So, by passing to the maximum over $a \in A$, we get $V_{h,k}(s) \geq V^*_{h,k}(s)$. In particular, $V_{1,k}(s_{1,k}) \geq V^*_{1,k}(s_{1,k})$.
Then, under $\mathcal{E}$ :

$$\begin{aligned} R(T) &\leq \sum_{k=1}^{K} \delta_{1,k}(s_{1,k}) \\ &= \sum_{k=1}^{K}\sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + m_{hk} \\ &\leq \sum_{k=1}^{K}\sum_{h=1}^{H} 2b_{hk}(s_{hk}, a_{hk}) + m_{hk} \end{aligned} \tag{41}$$

because

$$\begin{aligned} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) &= [\widehat{r}_{h,k}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk})] + b_{h,k}(s_{hk}, a_{hk}) \\ &+ \sum_{s'} [\widehat{p}_{h,k}(s'|s_{hk}, a_{hk}) - p_h(s'|s_{hk}, a_{hk})]V_{h+1,k}(s') \\ &\leq 2b_{h,k}(s_{hk}, a_{hk}) \end{aligned} \tag{42}$$

So,

$$\begin{aligned} P\left(R(T) \leq \sum_{k=1}^{K}\sum_{h=1}^{H} 2b_{hk}(s_{hk}, a_{hk}) + m_{hk}\right) &\geq P\left(\{R(T) \leq \sum_{k=1}^{K}\sum_{h=1}^{H} 2b_{hk}(s_{hk}, a_{hk}) + m_{hk}\}|\mathcal{E}\right) P(\mathcal{E}) \\ &= P(\mathcal{E}) \\ &\geq 1 - \frac{\delta}{2} \end{aligned} \tag{43}$$

Finally,

$$\begin{aligned}
p &:= P\left(R(T) \leq \sum_{k,h} 2b_{hk}(s_{hk}, a_{hk}) + 2H\sqrt{KH\log(2/\delta)}\right) \\
&\geq P\left(\{R(T) \leq \sum_{k,h} 2b_{hk}(s_{hk}, a_{hk}) + m_{hk}\} \cap \{\sum_{k,h} m_{hk} \leq 2H\sqrt{KH\log(2/\delta)}\}\right) \\
&\geq P\left(R(T) \leq \sum_{k,h} 2b_{hk}(s_{hk}, a_{hk}) + m_{hk}\right) + P\left(\sum_{k,h} m_{hk} \leq 2H\sqrt{KH\log(2/\delta)}\right) - 1 \\
&\geq \left(1 - \frac{\delta}{2}\right) + \left(1 - \frac{\delta}{2}\right) - 1 \\
&= 1 - \delta
\end{aligned} \tag{44}$$

where we use the fact that $P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1$ to get the second inequality.

Thus, $\boxed{R(T) \leq \sum_{k,h} 2b_{hk}(s_{hk}, a_{hk}) + 2H\sqrt{KH\log(2/\delta)} \text{ with probability } 1 - \delta}$

- Finally, we have that

$$\sum_{h,k} \frac{1}{\sqrt{N_{hk}(s_{hk}, a_{hk})}} = \sum_{h=1}^{H} \sum_{s,a} \sum_{i=1}^{N_{h,K}(s,a)} \frac{1}{\sqrt{i}} \leq 2 \sum_{h=1}^{H} \sum_{s,a} \sqrt{N_{hK}(s,a)}$$

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

- Solution:

First, let us prove the given upper bound of $\sum_{h,k} \frac{1}{\sqrt{N_{hk}(s_{hk},a_{hk})}}$. For that, we need to show by induction that $\forall n \in \mathbb{N}^*$, $\sum_{i=1}^{n} \frac{1}{\sqrt{i}} \leq 2\sqrt{n}$:

*Initialisation : For $n = 1$, the inequality is trivial

*Heredity : assume that $\sum_{i=1}^{n} \frac{1}{\sqrt{i}} \leq 2\sqrt{n}$. Prove that $\sum_{i=1}^{n+1} \frac{1}{\sqrt{i}} \leq 2\sqrt{n+1}$.

$$\begin{aligned}
\sum_{i=1}^{n+1} \frac{1}{\sqrt{i}} = \sum_{i=1}^{n} \frac{1}{\sqrt{i}} + \frac{1}{\sqrt{n+1}} &\leq 2\sqrt{n} + \frac{1}{\sqrt{n+1}} \\
&= 2\sqrt{n} + \frac{(n+1) - n}{\sqrt{n+1}} \\
&= 2\sqrt{n} + \frac{(\sqrt{n+1} + \sqrt{n})(\sqrt{n+1} - \sqrt{n})}{\sqrt{n+1}} \\
&= 2\sqrt{n} + (1 + \sqrt{\frac{n}{n+1}})(\sqrt{n+1} - \sqrt{n}) \\
&\leq 2\sqrt{n} + 2(\sqrt{n+1} - \sqrt{n}) = 2\sqrt{n+1}
\end{aligned} \tag{45}$$

Finally , we conclude that $\forall n \in \mathcal{N}^*$, $\sum_{i=1}^{n} \frac{1}{\sqrt{i}} \leq 2\sqrt{n}$.

In particular, $\sum_{i=1}^{N_{hK}(s,a)} \frac{1}{\sqrt{i}} \leq 2\sqrt{N_{hK}(s,a)}$ .

The function $f : x \mapsto \sqrt{x}$ is convex.

Then, $\sum_{s,a} \sqrt{N_{hK}(s,a)} = SA \sum_{s,a} \frac{1}{SA} f(N_{hK}(s,a)) \leq SA\, f(\frac{1}{SA} \sum_{s,a} N_{hK}(s,a)) = \sqrt{SA \sum_{s,a} N_{hK}(s,a)}$

Therefore,

$$\sum_{h,k} \frac{1}{\sqrt{N_{hk}(s_{hk}, a_{hk})}} \leq 2 \sum_{h=1}^{H} \sqrt{SA \sum_{s,a} N_{hK}(s,a)} \leq 2 \sum_{h=1}^{H} \sqrt{SAK} = 2H\sqrt{SAK}$$

Initialize $Q_{h1}(s,a) = 0$ for all $(s,a) \in S \times A$ and $h = 1, \ldots, H$

**for** $k = 1, \ldots, K$ **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 $\pi_{hk}$ using $\widehat{M}_k$
  **for** $h = H, \ldots, 1$ **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** $h = 1, \ldots, H$ **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

Recall that with probability $1-\delta$, $R(T) \leq \sum_{k,h} 2b_{hk}(s_{hk}, a_{hk}) + 2H\sqrt{KH\log(2/\delta)}$ with

$$b_{hk}(s_{hk}, a_{hk}) = \frac{H\sqrt{4log((2^S-2)SAHK/\delta)} + \sqrt{log(8SAHK/\delta))}}{\sqrt{2N_{hk}(s_{hk}, a_{hk})}} \leq (H+1)\sqrt{\frac{2S + 2log(SAHK/\delta)}{N_{hk}(s_{hk}, a_{hk})}}$$

We have

$$\begin{aligned} \sum_{h,k} b_{hk}(s_{hk}, a_{hk}) &\leq (H+1)\sqrt{2S + 2log(SAHK/\delta)} \sum_{h,k} \frac{1}{\sqrt{N_{hk}(s_{hk}, a_{hk})}} \\ &\leq 2\sqrt{2}H(H+1)S\sqrt{AK\left(1 + \frac{log(SAHK/\delta)}{S^2}\right)} \end{aligned} \tag{46}$$

We can set $\delta = \frac{1}{HK}$ , which is close to 0.
Then, with probability $1-\delta$, $\boxed{E(R(T)) \lesssim H^2 S\sqrt{AK}}$

# 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

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