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

Exploration in Reinforcement Learning (theory)

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

Solution by Raphael Reme

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 $S_{j,t} = \sum_{k=1}^{t} X_{i_k,k} \cdot \mathbb{1}(i_k = a)$ 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.

Answer

The first intuition could be that it's independent of $N_{j,t}$ and that we could consider that $\forall n, \mathbb{E}[\widehat{\mu}_{j,t}|N_{j,t} = n] = \mathbb{E}[\frac{S_{j,t}}{N_{j,t}}|N_{j,t} = n] = \frac{1}{n}\mathbb{E}[\sum_k X_{j,k}\mathbb{1}(i_k = j)|N_{j,t} = n] = \frac{n}{n}\mathbb{E}[X_{j,0}] = \mu_j$. And we could be tempted to conclude that therefore it's unbiased. But I will try to show that it's much more complicated. Indeed the information $N_{j,t} = n$ is not innocent at all. And the intuition that we should rather have, is that we resample more if we have high values of previous $(X_{j,k})_k$.
Let's focus with A actions $= [\![1, A]\!]$ and assume that that $U(0, \delta) = \infty$. Then the first A steps of the algorithm will choose all the actions once by definition of i_t. (Note that for $t < A$ then $\exists j, N_{j,t} = 0$ and then $\widehat{\mu}_{j,t}$ is not well defined). The definition of i is also not very clear. I will assume that $i_1 = 1$ and that $i_{t+1} \in \text{argmax}_{j\in[1,A]}\, \widehat{\mu}_{j,t} + U(N_{j,t}, \delta)$. (Otherwise i_t would depend on informations we don't have at the moment!)
Let's consider $t = A$. We have $\forall j, N_{j,t} = 1$. (The order of the i_k is irrelevant and I will suppose them ordered: $\forall k \in [1, A], i_k = k$). Then $i_{t+1} \in \text{argmax}_{j\in[1,A]}\, \widehat{\mu}_{j,t} + U(N_{j,t}, \delta) = \text{argmax}_{j\in[1,A]} X_{j,j}$.
This shows that we sample the action $t = A + 1$ according to the best rewards we got. And this could lead to negative bias: indeed if we get a low value for an action (lower than the expectation for instance), we will less resample it than if we had got a high value for this action. Low values of the laws of our rewards are over represented in our empirical mean.

Example: Let's use $A = 2, t = 3$ and $X_{1,k} \sim 2\mathbb{B}(p)+1$, $X_{2,k} \sim 2\mathbb{B}(q)$. We have $\mathbb{P}(X_{1,k} = 3) = p, \mathbb{P}(X_{1,k} = 1) = 1 - p, \mathbb{P}(X_{2,k} = 2) = q, \mathbb{P}(X_{2,k} = 0) = 1 - q$. And $\mu_1 = 2p + 1, \mu_2 = 2q$. Now we can see that $\widehat{\mu}_{1,3} = X_{1,1} + \frac{1}{2}\mathbb{1}_{i_3=1}(X_{1,3} - X_{1,1})$.

Then $\mathbb{E}[\widehat{\mu}_{1,3}] = \mu_1 + \frac{1}{2}\mathbb{E}[\mathbb{1}_{i_3=1}(X_{1,3} - X_{1,1})]$. And as $(i_3 = 1) = (X_{1,1} = 3)\bigcup(X_{1,1} = 1) \cap (X_{2,2} = 0)$:

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_{i_3=1}(X_{1,3} - X_{1,1})] &= \sum_{\substack{x_1\in\{1,3\}\\ x_2\in\{0,2\}\\ x_3\in\{1,3\}}} \mathbb{P}(X_{1,1} = x_1)\mathbb{P}(X_{2,2} = x_2)\mathbb{P}(X_{1,3} = x_3)\mathbb{1}_{i_3=3}(x_3 - x_1) \\
&= \sum_{\substack{x_1\in\{1,3\}\\ x_2\in\{0,2\}\\ x_3\in\{1,3\}}} p_1(x_1)p_2(x_2)p_1(x_3)(\mathbb{1}_{x_1=3} + \mathbb{1}_{x_1=1}\mathbb{1}_{x_2=0})(x_3 - x_1) \\
&= p \sum_{\substack{x_2\in\{0,2\}\\ x_3\in\{1,3\}}} p_2(x_2)p_1(x_3)(x_3 - 3) + (1-p)(1-q) \sum_{x_3\in\{1,3\}} p_1(x_3)(x_3 - 1) \\
&= p \times 1 \times (\mu_1 - 3) + (1-p)(1-q)(\mu_1 - 1) \\
&= p(2p - 2) - (1-p)(1-q)(2p) \\
&= -2p(1-p)q \\
&< 0
\end{aligned}$$

Hence $\boxed{\mathbb{E}[\widehat{\mu}_{1,3}] < \mu_1}$. (This also holds for μ_2) In this example we have a negative bias. This shows that it can't be an unbiased estimator in the general case .

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 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 $\mathbb{P}_{\mu_1,\dots,\mu_k}(\widehat{i} \neq i^\star) \le \delta$ 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.

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. For any arm $i \in [k]$

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

  Use Hoeffding's inequality.

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

- 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 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 $\Delta_i \geq C_1 U(t,\delta')$ 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 $i^\star$.
- Compute a bound on the sample complexity (after how many rounds the algorithm stops) for identifying the optimal arm w.p. $1-\delta$.

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

Answers

1-

First the Hoeffding's inequality can be stated as $\forall \delta, t, i,\ \mathbb{P}\left(|\widehat{\mu}_{i,t} - \mu_i| > \sqrt{\frac{\log \frac{2}{\delta}}{2t}} \right) \leq \delta$.

Now let's define $\forall t \geq 1, B_t = \{|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta)|\}$, $\forall n \geq 1, A_n = \bigcup_{t=1}^{n} B_t$. We are trying to find $U(t,\delta)$ such that $\mathbb{P}\left(\bigcup_{t=1}^{\infty} B_t\right) \leq \delta$.
With our notations we have $A_n \subset A_{n+1}$ and therefore

$$\begin{aligned} \mathbb{P}\left(\bigcup_{t=1}^{\infty} B_t\right) &= \mathbb{P}\left(\bigcup_{t=1}^{\infty} A_n\right) \\ &= \lim_{n\to\infty} \mathbb{P}\left(A_n\right) \end{aligned}$$

Moreover $\mathbb{P}\left(A_n\right) = \mathbb{P}\left(\bigcup_{t=1}^{n} B_t\right) \leq \sum_{t=1}^{n} \mathbb{P}\left(B_t\right) \leq \sum_{t=1}^{\infty} \mathbb{P}\left(B_t\right)$.

Let's define $U(t,\delta) = \sqrt{\frac{\log \frac{2}{\frac{1}{2^t}\delta}}{2t}} = \sqrt{\frac{\log \frac{2^{t+1}}{\delta}}{2t}}$ then the Hoeffding's inequality applied to B_t gives

$$\begin{aligned} \mathbb{P}\left(B_t\right) &= \mathbb{P}\left(|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta)\right) \\ &\leq \frac{1}{2^t}\delta \end{aligned}$$

Thus we have $\forall n \geq 1$:

$$\mathbb{P}\left(A_n\right) \leq \sum_{t=1}^{\infty} \mathbb{P}\left(B_t\right)$$
$$\mathbb{P}\left(A_n\right) \leq \sum_{t=1}^{\infty} \frac{1}{2^t}\delta$$
$$\mathbb{P}\left(A_n\right) \leq \delta$$
$$\lim_{n \leftarrow \infty} \mathbb{P}\left(A_n\right) \leq \delta$$
$$\mathbb{P}\left(\bigcup_{t=1}^{\infty} B_t\right) \leq \delta$$
$$\mathbb{P}\left(\bigcup_{t=1}^{\infty} \{|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta)\}\right) \leq \delta \quad \text{with } \boxed{U(t,\delta) = \sqrt{\frac{\log \frac{2^{t+1}}{\delta}}{2t}}}$$

Correction after Q4: This choice of U is valid but not good enough because here U doesn't converge to 0 when t goes to infinity (in this case it converges to $\sqrt{\frac{\log 2}{2}}$).
We have to choose U such that $\sum_{t=1}^{\infty} \mathbb{P}(B_t) = \delta$ and so that U converges to 0. As we have seen the choice of $\mathbb{P}(B_t)$ gives U. And an exponential choice in t leads to a function U that doesn't converge towards 0. It seems obvious that a polynomial one would be good. ($\mathbb{P}(B_t) \propto \frac{1}{t^n}$). Let's prove it! (I will choose $n = 2$ here)
Let's defined $C = \sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6}$. And with the Hoeffding's inequality let's choose U such that $\mathbb{P}(B_t) = \frac{\delta}{Ct^2}$: $U(t,\delta) = \sqrt{\frac{\log \frac{2Ct^2}{\delta}}{2t}} = \sqrt{\frac{2\log t + \log \frac{2C}{\delta}}{2t}}$.
With this U we have $\forall \delta, \lim_{t\to\infty} U(t,\delta) = 0$ and $\sum_{t=1}^{\infty} \mathbb{P}(B_t) = \delta$.
Therefore we have:

$$\mathbb{P}\left(\bigcup_{t=1}^{\infty} \{|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta)\}\right) \leq \delta \quad \text{with } \boxed{U(t,\delta) = \sqrt{\frac{\log \frac{2Ct^2}{\delta}}{2t}}}$$

2-

With $C_i = \bigcup_{t=1}^{\infty} \{|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta')\}$, we have:

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{k} C_i\right) &\leq \sum_{i=1}^{k} \mathbb{P}\left(C_i\right) \\ &\leq k\delta' \qquad\qquad (\text{As } \forall i, \mathbb{P}(C_i) < \delta')\end{aligned}$$

Therefore with $\boxed{\delta' = \frac{\delta}{k}}$, then $\mathbb{P}(\mathcal{E}) < \delta$.

3-

We have shown that $\mathbb{P}(\mathcal{E}) < \delta$. Therefore we have $\mathbb{P}(\neg\mathcal{E}) = 1 - \mathbb{P}(\mathcal{E}) > 1 - \delta$.
Now let's denote by A the event stating that the optimal arm remains in S:

$$A = \bigcap_{t=1}^{\infty} \left\{ i^\star \notin \left\{ i \in S \ : \ \exists j \in S,\ \widehat{\mu}_{j,t} - U(t,\delta') \geq \widehat{\mu}_{i,t} + U(t,\delta') \right\}\right\}$$

(Note: I'm using δ' as input of the algorithm rather than δ otherwise it won't work. And I will assume that $i^\star$ is unique. If it exists $j^\star$ s.t. $mu_{j^\star} = \mu^\star$, we could have a case were $i^\star$ is removed. But $j^\star$ would stay.)

I will show that $\neg\mathcal{E} \subset A$: if $\neg\mathcal{E}$ occurs then so does A.
Assuming that $\neg\mathcal{E}$ has happened. Then $\forall i, t, |\widehat{\mu}_{i,t} - \mu_i| < U(t, \delta')$. Let $t \geq 1$, $j \in S \setminus \{i^\star\}$, we have :

$$\begin{aligned} |\widehat{\mu}_{j,t} - \mu_j| \leq U(t,\delta') \quad &\text{and} \quad |\widehat{\mu}_{i^\star,t} - \mu^\star| \leq U(t,\delta') \\ \widehat{\mu}_{j,t} - \mu_j \leq U(t,\delta') \quad &\text{and} \quad \mu^\star - \widehat{\mu}_{i^\star,t} \leq U(t,\delta') \quad (\text{As } \pm x \leq |x|) \end{aligned}$$

Then we have:

$$\begin{aligned} \widehat{\mu}_{j,t} - \mu_j + \mu^\star - \widehat{\mu}_{i^\star,t} &\leq 2U(t,\delta') \\ \widehat{\mu}_{j,t} - \mu_j + \mu^\star - \widehat{\mu}_{i^\star,t} &< 2U(t,\delta') + \Delta_j \quad (\Delta_j < 0) \\ \widehat{\mu}_{j,t} - \mu_j + \mu^\star - \widehat{\mu}_{i^\star,t} &< 2U(t,\delta') + \mu^\star - \mu_j \\ \widehat{\mu}_{j,t} - \widehat{\mu}_{i^\star,t} &< 2U(t,\delta') \\ \widehat{\mu}_{j,t} - U(t,\delta') &< \widehat{\mu}_{i^\star,t} + U(t,\delta') \end{aligned}$$

And as $U(t,\delta') > 0$ we also have this inequality for $j = i^\star$. Therefore $\forall t \geq 1, i^\star \notin \Big\{i \in S \ : \ \exists j \in S,\ \widehat{\mu}_{j,t} - U(t,\delta') \geq \widehat{\mu}_{i,t} + U(t,\delta')\Big\}$. And thus we are in A.
We've shown that $\neg\mathcal{E} \subset A$ (any w realising $\neg\mathcal{E}$ will also realised A) and therefore $\boxed{\mathbb{P}(A) \geq \mathbb{P}(\neg\mathcal{E}) \geq 1-\delta}$, that is to say that the optimal arm remains in S with probability at least $1-\delta$.

4-

Let's call $D(S) = \Big\{i \in S \ : \ \exists j \in S,\ \widehat{\mu}_{j,t} - U(t,\delta') \geq \widehat{\mu}_{i,t} + U(t,\delta')\Big\}$. (Note that S depends on t)
Let $i \neq i^\star$ we want to show that $\exists t \geq 1$ such that $i \in D(S)$ and find this t.
Now as in the previous question, let's assume that $\neg\mathcal{E}$ has happened. Then we've shown that $\forall t \geq 1, i^\star \in S$. And we will use that knowledge to show that $\widehat{\mu}_{i^\star,t} - U(t,\delta') \geq \widehat{\mu}_{i,t} + U(t,\delta')$ under a condition on t. (And thus for that $t, i \in D(S)$ and is removed from the active set.)
As before we have:

$$\begin{aligned} \widehat{\mu}_{i,t} - \mu_i + \mu^\star - \widehat{\mu}_{i^\star,t} &\leq 2U(t,\delta') \\ \widehat{\mu}_{i,t} + \Delta_i - \widehat{\mu}_{i^\star,t} &\leq 2U(t,\delta') \\ \widehat{\mu}_{i^\star,t} &\geq \widehat{\mu}_{i,t} - 2U(t,\delta') + \Delta_i \\ \widehat{\mu}_{i^\star,t} - U(t,\delta') &\geq \widehat{\mu}_{i,t} + U(t,\delta') - 4U(t,\delta') + \Delta_i \\ \widehat{\mu}_{i^\star,t} - U(t,\delta') &\geq \widehat{\mu}_{i,t} + U(t,\delta') \qquad \text{If } \boxed{\Delta_i \geq 4U(t,\delta')} \end{aligned}$$

Therefore $\boxed{C_1 = 4}$.
Asusming that $i^\star$ is unique then, $\Delta_i > 0$ and as $\lim_{t\to\infty} U(t,\delta') = 0$ from Q1 (after correction):

$$\boxed{\exists t_i \geq 1,\ C_1 U(t_i,\delta') \geq \Delta_i}$$

Let's fix $t_i = \inf_t\{t \geq 1, C_1 U(t,\delta') \geq \Delta_i\}$. Then the sub-optimal arm i is removed after at most $\lceil t_i \rceil$ iterations.
Let's try to express this t_i w.r.t. Δ_i and δ:

$$\begin{aligned} 4U(t_i,\delta) &\leq \Delta_i \\ \sqrt{\frac{\log \frac{\pi^2 t_i^2}{3\delta}}{2t_i}} &\leq \frac{\Delta_i}{4} \\ \frac{\log \frac{\pi^2 t_i^2}{3\delta}}{2t_i} &\leq \left(\frac{\Delta_i}{4}\right)^2 \\ \frac{\log t_i}{t_i} + \frac{\log \frac{\pi^2}{3\delta}}{2t} &\leq \left(\frac{\Delta_i}{4}\right)^2 \\ \log t_i + A \leq Bt_i \quad &\text{With } A = \frac{\log\frac{\pi^2}{3\delta}}{2}, B = \left(\frac{\Delta_i}{4}\right)^2 \\ Bt_i - \log t_i - A &\geq 0 \end{aligned}$$

Let's analyse this function $f(t) = Bt - \log t - A$. First as $\Delta_i \in]0,1]$, we have $0 < B \leq \frac{1}{16}$. And I will suppose than $\delta < \frac{1}{3}$ (High values for δ have no interest and this will help to characterize t_i) then, $A > \log \pi > 1 > B$.
We can compute the derivative of f : $f'(t) = B - \frac{1}{t}$. $f'(t) = 0 \Leftrightarrow t = \frac{1}{B}$. As f is convexe f reaches a minimum at $t = \frac{1}{B}$. And $f(1) = B - A < 0$. Therefore f is negative on $[1, \frac{1}{B}]$ ($\frac{1}{B} > 1$). And $f(t) \longrightarrow_{t\to\infty} \infty$.
Therefore there is a unique $t_0 \in [\frac{1}{B}, +\infty]$ such that $f(t_0) = 0$. And by definition of t_i we have $t_i = t_0$ (because t_0 is the first $t \geq 1$ such that $f(t) \geq 0$). We thus have a first bound for t_i:

$$\begin{aligned} t_i &\geq \frac{1}{B} \\ &\geq \left(\frac{4}{\Delta_i}\right)^2 \end{aligned}$$

It's not an upperbound and therefore we can't deduce anything for the time needed to eliminate the sub-optimal arm i from this equation (but still it gives informations on t_i)
As log is concave it's below any of its tangent. I will use this to find an upperbound:

$$\begin{aligned} \forall x_0 > 0, t > 0,\ \log(t) &\leq \log(x_0) + \frac{1}{x_0}(t - x_0) \\ -\log(t) &\geq -\frac{t}{x_0} + 1 - \log(x_0) \\ f(t) &\geq (B - \frac{1}{x_0})t + 1 - A - \log x_0 \\ f(t) &\geq \frac{Bx_0 - 1}{x_0}t + 1 - A - \log x_0 \end{aligned}$$

Let's use $x_0 = \frac{A}{B}$ (in order to use the fact that $A > 1$)

$$f(t) \geq B\frac{A-1}{A}t + 1 - A - \log A + \log B$$

Now as $A > 1$ we have:

$$\begin{aligned} \forall t \geq A\frac{A - 1 + \log A - \log B}{B(A-1)},\ B\frac{A-1}{A}t + 1 - A - \log A + \log B \geq 0 \\ f(t) \geq 0 \end{aligned}$$

(If we don't assume that $\delta < \frac{1}{3}$ then we could still prove that $A > 0.5$ and use directly $x_0 = 2\frac{A}{B}$, which leads to a similar results)
We have therefore $t_i \leq A\frac{A-1+\log A - \log B}{B(A-1)}$ (as t_i is the smallest t that verify this equation.)
Assuming that δ (and Δ_i) are small enough we can simplify the expression keeping only dominant terms:

$$A\frac{A - 1 + \log A - \log B}{B(A-1)} + 1 \sim 16\frac{\log\frac{1}{\delta} - 2\log\Delta_i}{\Delta_i^2}$$

Finally $\boxed{t_i \leq \alpha\frac{\log\frac{1}{\delta} - 2\log\Delta_i}{\Delta_i^2}}$ (With α a constante).

5-

Finally if $i^\star$ is unique, then the algortihm will remove all sub-optimal arms $i \neq i^\star$ in at most $T = \max_{i\neq i^\star}\lceil t_i \rceil$ iterations!
Let's define $\Delta = \min_{i\neq i^\star}\Delta_i$. From the previous question we have:

$$\forall i \neq i^\star,\, t_i \leq \alpha \frac{\log \frac{1}{\delta} - 2\log \Delta_i}{\Delta_i^2} \leq \alpha \frac{\log \frac{1}{\delta} - 2\log \Delta}{\Delta^2}$$

$$\boxed{T \leq \alpha \frac{\log \frac{1}{\delta} - 2\log \Delta}{\Delta^2}}$$

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.

- Let $n = 10000$ and $k = 2$. Plot the expected regret of each algorithm as a function of Δ when $\mu_1 = 1/2$ and $\mu_2 = 1/2 + \Delta$.
- Repeat the above experiment with $\mu_1 = 1/10$ and $\mu_1 = 9/10$.
- Discuss your results.

Answers-

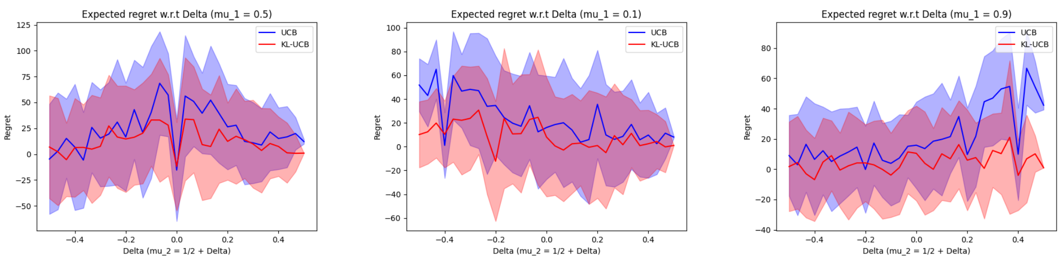

Figure 1: Expected regret for different μ_1 (0.5, 0.1, 0.9)

Those results have been obtained with 30 randoms runs of the algorithms. In order to reproduce them, you can run the code provided with:

```
$ python exercise3.py
$ python exercise3.py --mu-1 0.1
$ python exercise3.py --mu-1 0.9
```

It seems that even with 30 runs there are lots of uncertainty over our results and it's hard to conclude. But the KL-UCB method seems to slightly outperform the UCB method, specially when μ_1 is closed to μ_2.

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

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

- 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.,

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

  where $Q^\star$ is the optimal Q-function of the unknown MDP $M^\star$. Note that $\widehat{r}_{H,k}(s,a) + b_{h,k}(s,a) \ge r_{h,k}(s,a)$ and thus $Q_{H,k}(s,a) \ge Q^\star_H(s,a)$ (for a properly defined bonus). Then use induction to prove that this holds for all the stages h.

- In class we have seen that

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

  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 $V_{h,k}(s_{hk}) \le Q_{h,k}(s_{hk}, a_{hk})$.
  3. Putting everything together prove Eq. 1.

- Since $(m_{hk})_{hk}$ is an MDS, using Azuma-Hoeffding we show that with probability at least $1 - \delta/2$

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

  Show that the regret is upper bounded with probability $1 - \delta$ by

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

- Finally, we have that

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

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

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

Answers

1-

$$\begin{aligned}
\neg\mathcal{E} &= \bigcup_{k=1}^{K} \{M^\star \notin M_k\} \\
&= \bigcup_{k=1}^{K} \bigcup_{s \in S} \bigcup_{a \in A} \bigcup_{h=1}^{H} \{r_{h,k}(s,a) \notin B^r_{h,k}(s,a)\} \cup \{p_{h,k}(\cdot|s,a) \notin B^p_{h,k}(s,a)\}
\end{aligned}$$

With

$$\begin{aligned}
B^r_{h,k}(s,a) &= \{r \in \mathbb{R}, |r - r_h(s,a)| \le \beta^r_{h,k}(s,a)\} \quad \text{With } \beta^r \text{ a function to be expressed} \\
B^p_{h,k}(s,a) &= \{p \in \Delta(S), ||p - p_h(\cdot|s,a)||_1 \le \beta^p_{h,k}(s,a)\} \quad \text{With } \beta^p \text{ a function to be expressed}
\end{aligned}$$

Therefore

$$\neg\mathcal{E} = \bigcup_{k=1}^{K} \bigcup_{s \in S} \bigcup_{a \in A} \bigcup_{h=1}^{H} \{|r_{h,k}(s,a) - r_h(s,a)| > \beta^r_{h,k}(s,a) \cup \{||p_{h,k}(\cdot|s,a) - p_h(\cdot|s,a)||_1 > \beta^p_{h,k}(s,a)\}$$

Let's consider the event $\mathcal{D}(s,a,h,k) = \{|r_{h,k}(s,a) - r_h(s,a)| > \beta^r_{h,k}(s,a) \cup \{||p_{h,k}(\cdot|s,a) - p_h(\cdot|s,a)||_1 > \beta^p_{h,k}(s,a)\}$
Using this we can rewrite $\neg\mathcal{E}$:

$$\neg\mathcal{E} = \bigcup_{k=1}^{K} \bigcup_{s \in S} \bigcup_{a \in A} \bigcup_{h=1}^{H} \mathcal{D}(s,a,h,k)$$

Thus we have:

$$\mathbb{P}(\neg\mathcal{E}) \le \sum_{k=1}^{K} \sum_{s \in S} \sum_{a \in A} \sum_{h=1}^{H} \mathbb{P}(\mathcal{D}(s,a,h,k))$$

Let's find β^r, β^p such that $\forall k, s, a, h,\ \mathbb{P}(\mathcal{D}(s,a,h,k)) < \frac{\delta}{2SAHK}$.

Using Hoeffding and Weissmain inequalities we can have such a bound:

$$\begin{aligned}\mathbb{P}(\mathcal{D}(s,a,h,k)) &\leq \mathbb{P}(|r_{h,k}(s,a) - r_h(s,a)| > \beta^r_{h,k}(s,a)) + \mathbb{P}(||p_{h,k}(\cdot|s,a) - p_h(\cdot|s,a)||_1 > \beta^p_{h,k}(s,a)) \\ &\leq \exp\left(-2N_{h,k}(s,a)\beta^r_{h,k}(s,a)^2\right) + (2^S - 2)\exp\left(-\frac{N_{h,k}(s,a)\beta^p_{h,k}(s,a)^2}{2}\right)\end{aligned}$$

Using a generic form for β^p and β^r we can simplify this expression:

$$\beta^r_{h,k}(s,a) = \sqrt{\frac{\log\frac{1}{\delta^r_{s,a,h,k}}}{2N_{h,k}(s,a)}}$$

$$\beta^p_{h,k}(s,a) = \sqrt{\frac{2\log\frac{1}{\delta^p_{s,a,h,k}}}{N_{h,k}(s,a)}}$$

We have then:

$$\mathbb{P}(\mathcal{D}(s,a,h,k)) \leq \delta^r_{s,a,h,k} + (2^S - 2)\delta^p_{s,a,h,k}$$

And thus with $\delta^r_{s,a,h,k} = \frac{\delta}{4SAHK}$ and $\delta^p_{s,a,h,k} = \frac{\delta}{4SAHK(2^S-2)}$ we have:

$$\mathbb{P}(\mathcal{D}(s,a,h,k)) \leq \frac{\delta}{2SAHK}$$

And we can conclude:

$$\begin{aligned}\mathbb{P}(\neg\mathcal{E}) &\leq \sum_{k=1}^{K}\sum_{s\in S}\sum_{a\in A}\sum_{h=1}^{H}\mathbb{P}(\mathcal{D}(s,a,h,k)) \\ &\leq \sum_{k=1}^{K}\sum_{s\in S}\sum_{a\in A}\sum_{h=1}^{H}\frac{\delta}{2SAHK} \\ &\leq \frac{\delta}{2}\end{aligned}$$

With

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

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

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

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