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

Exploration in Reinforcement Learning (theory)

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

Solution by Samuel Diai

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 : UCB

I'm going to detail an example in which the estimator $\widehat{\mu}_{j,t}$ is biased.
Let consider the following example :

- The number of arms $K = 2$
- The number of rounds $T = 3$
- The reward from the arm 1 is drawn from a Bernoulli distribution of parameter μ_1
- The reward from the arm 2 is drawn from a Bernoulli distribution of parameter μ_2
- During the algorithm, if there is a tie, both arm can be chosen with an equal probability of $\frac{1}{2}$

Let's start by an observation, if one arm is drawn in the first round. Necessarily , the other arm will be pulled in the second round. Indeed, let's say the first arm is pulled, then $N_{1,T=0} = 1$ and $N_{2,T=0} = 0$. As the UCB bound is proportinal to $\frac{1}{N_{2,T=0}}$, this value will be infinity for the second arm. Thus it will be automatically pulled.
Let's compute the expectation of the estimator $\widehat{\mu}_{1,T=3}$. This estimator can take the value 0, $\frac{1}{2}$, 1. We can compute the probabilities by enumerating all the possibilities.
For example, for $\widehat{\mu}_{1,T=3} = 0$, there are four possibilities to get a zero reward for the arm 1:

- The first arm pulled was the arm 1, then the reward drawn was 0 (with probability $1-\mu_1$), then we pull the arm 2 (necessarily), then we draw the reward 1 from the arm 2 (with probability μ_2), at the step 3, the arm pulled will be the arm 2, since the estimated reward is higher and the ucb bound is the same.
- The first arm pulled was the arm 1, then the reward drawn was 0 (with probability $1-\mu_1$), then we pull the arm 2 (necessarily), then we draw the reward 0 from the arm 2 (with probability $1-\mu_2$), as there is a tie, with probability $\frac{1}{2}$, we pull the arm 1, and with probability $1-\mu_1$, the reward drawn is 0.
- The first arm pulled was the arm 2, then the reward drawn was 0 (with probability $1-\mu_2$), then we pull the arm 1 (necessarily), then we draw the reward 0 from the arm 1 (with probability $1-\mu_1$), as there is a tie, with probability $\frac{1}{2}$, we pull the arm 1, and with probability $1-\mu_1$, the reward drawn is 0.
- The first arm pulled was the arm 2, then the reward drawn was 1 (with probability μ_2), then we pull the arm 1 (necessarily), then we draw the reward 0 from the arm 1 (with probability $1-\mu_1$), at the step 3, the arm pulled will be the arm 2, since the estimated reward is higher and the ucb bound is the same.

If we do this same kind of reasoning, we get :

- $\mathbb{P}(\widehat{\mu}_{1,T=3}=0) = (1-\mu_1)\mu_2 + (1-\mu_1)(1-\mu_2)[\frac{1}{2}+\frac{1}{2}(1-\mu_1)] = (1-\mu_1)[\mu_2+(1-\mu_2)\frac{1}{2}(2-\mu_1)]$
- $\mathbb{P}(\widehat{\mu}_{1,T=3}=\frac{1}{2}) = \mu_1\mu_2\frac{1}{2}(1-\mu_1) + \mu_1(1-\mu_2)(1-\mu_1) + (1-\mu_1)(1-\mu_2)\frac{1}{2}\mu_1 = (1-\mu_1)\mu_1[\frac{1}{2}\mu_2 + (1-\mu_2) + \frac{1}{2}(1-\mu_2)] = (1-\mu_1)\mu_1(\frac{3}{2}-\mu_2)$
- $\mathbb{P}(\widehat{\mu}_{1,T=3}=1) = \mu_1(1-\mu_2)\mu_1 + \mu_1\mu_2[\frac{1}{2}+\frac{1}{2}\mu_1] = \mu_1(\mu_1 - \frac{1}{2}\mu_1\mu_2 + \frac{1}{2}\mu_2)$

Thus we can compute the expectation :

$$\mathbb{E}(\widehat{\mu}_{1,T=3}) = \frac{1}{2}\mathbb{P}(\widehat{\mu}_{1,T=3}=\frac{1}{2}) + \mathbb{P}(\widehat{\mu}_{1,T=3}=1)$$

If we do the computations :

$$\mathbb{E}(\widehat{\mu}_{1,T=3}) = \frac{1}{2}(1-\mu_1)\mu_1(\frac{3}{2}-\mu_2) + \mu_1(\mu_1 - \frac{1}{2}\mu_1\mu_2 + \frac{1}{2}\mu_2) = \frac{1}{4}\mu_1(3+\mu_1)$$

As the example is purely symmetric (equal probability if there is a tie), we also get for the arm 2:

$$\mathbb{E}(\widehat{\mu}_{2,T=3}) = \frac{1}{4}\mu_2(3+\mu_2)$$

Thus we can compute the biases :

$$\begin{cases} b_1 = \mathbb{E}(\widehat{\mu}_{1,T=3}) - \mu_1 = \frac{1}{4}\mu_1(\mu_1-1) \\ b_2 = \mathbb{E}(\widehat{\mu}_{2,T=3}) - \mu_2 = \frac{1}{4}\mu_2(\mu_2-1) \end{cases}$$

Thus we can see that the estimators are clearly biased even for a simple example, thus in the general case, there is absolutely no reason why the estimators should be unbiased.

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

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

```
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
```

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 $i \in [k]$

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

  Use Hoeffding's inequality.

- 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.

Solution : Best Arm Identification

- We want to find $U(t,\delta)$ such that :

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

  We can develop the empirical mean and mean as :

  $$\mathbb{P}\left(\left\{|\frac{1}{t}\sum_{j=1}^{t} X_{i,j} - \frac{1}{t}\sum_{j=1}^{t}\mathbb{E}(X_{i,j})| > U(t,\delta)\right\}\right) \leq \delta$$

  Thus :

  $$\mathbb{P}\left(\left\{|\sum_{j=1}^{t} X_{i,j} - \sum_{j=1}^{t}\mathbb{E}(X_{i,j})| > tU(t,\delta)\right\}\right) \leq \delta$$

  Using Hoeffding inequality we find :

  $$\mathbb{P}\left(\left\{|\sum_{j=1}^{t} X_{i,j} - \sum_{j=1}^{t}\mathbb{E}(X_{i,j})| > tU(t,\delta)\right\}\right) \leq 2\exp\left\{-\frac{2t^2U(t,\delta)^2}{t}\right\} = 2\exp\left\{-2tU(t,\delta)^2\right\}$$

  Thus if we take $U(t,\delta)$ such that :

  $$U(t,\delta) = \sqrt{-\frac{1}{2t}\ln(\frac{\delta}{2})} = \sqrt{\frac{1}{2t}\ln(\frac{2}{\delta})}$$

  We have by construction :

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

- Let $\mathcal{E} = \bigcup_{i=1}^{k}\bigcup_{t=1}^{\infty}\{|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta')\}$. We want to find δ such that $\mathbb{P}(\mathcal{E}) \leq \delta$

  We have

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

  If we take $\delta' = \frac{6\delta}{kt^2\pi^2}$, we have :

  $$\sum_{i=1}^{k}\sum_{t=1}^{\infty}\delta' = \sum_{i=1}^{k}\sum_{t=1}^{\infty}\frac{6\delta}{kt^2\pi^2} = \delta$$

  This gives : $\mathbb{P}(\mathcal{E}) \leq \delta$.

- With probability $1-\delta$, $\neg\mathcal{E}$ is realized. That is

  $$\forall i \in S, \forall t, |\widehat{\mu}_{i,t} - \mu_i| \leq U(t,\delta')$$

  This implies in particular for $i^\star$:

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

  Let's suppose $i^\star$ is removed from the set S, that is, it exists j such that :

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

  Using the above inequality and that $\mu_{i^\star} \geq \mu_j$ for all $j \in S$:

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

  This is in contradiction with the event $\neg\mathcal{E}$ being realized. Thus by contradiction, With probability $1-\delta$, $i^\star$ remains in the active set S.

- Suppose that $\neg\mathcal{E}$ is realized, and that $\Delta_i \geq 4U(t,\delta')$. With the definition of $\neg\mathcal{E}$:

  $$U(t,\delta') \geq \mu_{i^\star} - \widehat{\mu}_{i^\star,t}, \text{ and } U(t,\delta') \geq \widehat{\mu}_{i,t} - \mu_i$$

  This leads to

  $$2U(t,\delta') \geq \mu_{i^\star} - \widehat{\mu}_{i^\star,t} - \mu_i + \widehat{\mu}_{i,t} = \widehat{\mu}_{i,t} - \widehat{\mu}_{i^\star,t} + \Delta_i$$

  As $\Delta_i \geq 4U(t,\delta')$,

  $$2U(t,\delta') \leq \widehat{\mu}_{i^\star,t} - \widehat{\mu}_{i,t}$$

  Which can be rewritten as :

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

  Thus, by definition, the arm i is rejected from the set of active arm S (with probability $1-\delta$).

  To have such a condition, we need to have $\Delta_i \geq 4U(t,\delta')$. This means :

  $$\Delta_i \geq 4\sqrt{\frac{1}{2t}\ln(\frac{2}{\delta'})} = 4\sqrt{\frac{1}{2t}\ln(\frac{kt^2\pi^2}{3\delta})}$$

  Which leads to :

  $$\Delta_i^2 \geq 16\frac{1}{2t}\ln(\frac{kt^2\pi^2}{3\delta})$$

  $$t\Delta_i^2 \geq 16\ln(t) + 8\ln(\frac{k\pi^2}{3\delta})$$

  If we use the typical inequality $log(1+x) \leq x$, for $x \geq 0$, we will get a negative bound for t (which is useless). Instead, let's use the concavity of the logarithm at the point $\alpha\Delta_i^2 x$. We get :

  $$\ln(x) \leq \alpha\Delta_i^2 x - \ln(\alpha\Delta_i^2), \text{ for } x \geq 1$$

  Using this result, we have :

  $$t\Delta_i^2 \geq 16\alpha\Delta_i^2 t - 16\ln(\alpha\Delta_i^2) + 8\ln(\frac{k\pi^2}{3\delta})$$

  Thus the following inequality is a sufficient condition for rejecting the arm i (under $\neg\mathcal{E}$):

  $$t_i \geq \frac{1}{1-16\alpha}\frac{8\ln(\frac{k\pi^2}{3\delta}) - 16\ln(\alpha\Delta_i^2)}{\Delta_i^2}$$

  We can tune α (positive) to get a small enough bound.

- To compute the final bound, we need to have the previous inequality true for every arm $i \neq i^\star$. Indeed, under $\neg\mathcal{E}$, we will reject every non optimal arm once we satisfy the above inequality.

  So we need to have :

  $$T \geq \max_i \left\{ \frac{1}{1-16\alpha}\frac{8\ln(\frac{k\pi^2}{3\delta}) - 16\ln(\alpha\Delta_i^2)}{\Delta_i^2} \right\}$$

  As the function $\Delta_i \mapsto -\ln(\Delta_i^2)/\Delta_i^2$ is decreasing.

  $$T \geq \frac{1}{1-16\alpha}\frac{8\ln(\frac{k\pi^2}{3\delta}) - 16\ln(\alpha\Delta_{i^\star}^2)}{\Delta_i^{\star 2}}$$

  Where $\Delta_i^\star$ is the smallest gap. This gives us the result, with a probability $1-\delta$,

  $$T = \widetilde{O}(\frac{1}{1-16\alpha}\frac{8\ln(\frac{k\pi^2}{3\delta}) - 16\ln(\alpha\Delta_{i^\star}^2)}{\Delta_i^{\star 2}})$$

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.

Solution : Bernoulli Bandits

I implemented the KL-UCB and UCB algorithm using python.
For the KL-UCB, I needed to compute :

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

In the bernouilli case :

$$d(p,q) = p\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q})$$

Moreover, to compute the maximum (as the problem is convex, the inequality constraints will be saturated at the optimum), I simply computed the value $\mu_\star$, such that

$$d(\widehat{\mu}_{i,t}, \mu_\star) = \frac{\log(1 + t\log^2(t))}{N_{i,t}}$$

Indeed, as the function $\mu \mapsto d(\widehat{\mu}_{i,t}, \mu)$ is strictly convex, continuous, and positive, and as $\frac{\log(1+t\log^2(t))}{N_{i,t}}$ is also positive there are only two intersecting points.
As $d(\widehat{\mu}_{i,t}, \widehat{\mu}_{i,t}) = 0$, there is one solution $\mu_\star \in [0, \widehat{\mu}_{i,t}]$ and the other solution $\mu_\star \in [\widehat{\mu}_{i,t}, 1]$.
As we are looking for the maximum value w.r.t μ I did a bisection search on the interval $[\widehat{\mu}_{i,t}, 1]$.
Let's see the regret we have for the different experiments. For each experiment, we run 100 the same algorithm and we average the regret to get an approximated value of the average true regret.

Figure 1: Average estimated regret $\mu_1 = 1/2$ and $\mu_2 = 1/2 + \Delta$

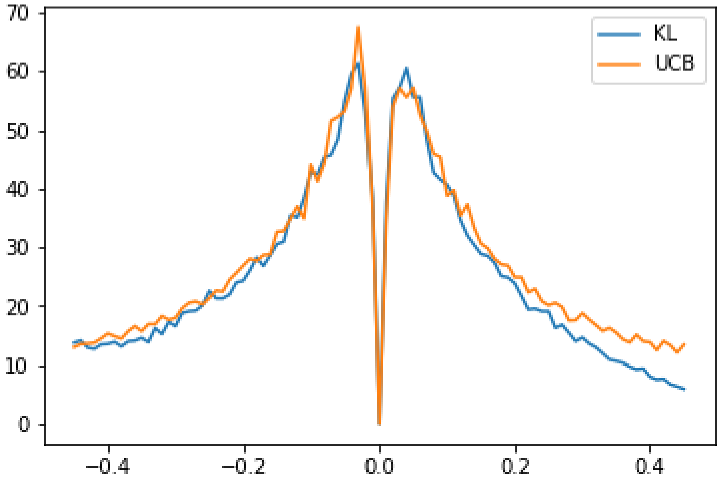


Figure 2: Average estimated regret $\mu_1 = 1/10$ and $\mu_2 = 1/10 + \Delta$

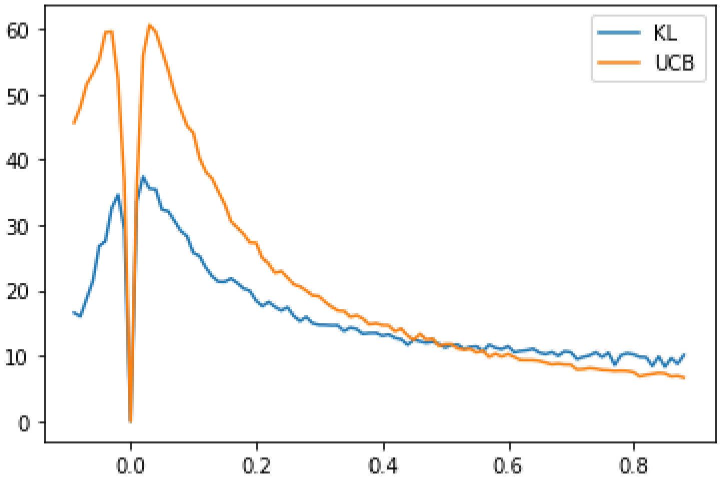
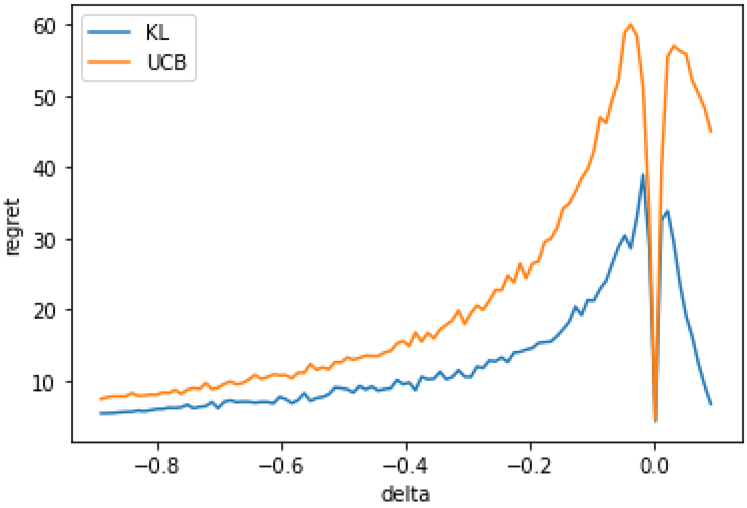


Figure 3: Average estimated regret $\mu_1 = 9/10 + \Delta$ and $\mu_2 = 9/10$



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

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

- Define the bonus function and consider the Q-function computed at episode k

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

with $V_{h,k}(s) = \min\{H, \max_a Q_{h,k}(s,a)\}$. Recall that $V_{H+1,k}(s) = V^\star_{H+1}(s) = 0$. Prove that under event $\mathcal{E}$, Q_k is optimistic, i.e.,

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

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

- In class we have seen that

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

where $\delta_{hk}(s) = V_{hk}(s) - V_h^{\pi_k}(s)$ and $m_{hk} = \mathbb{E}_{Y \sim p(\cdot|s_{hk},a_{hk})}[\delta_{h+1,k}(Y)] - \delta_{h+1,k}(s_{h+1,k})$. We now want to prove this result. Denote by a_{hk} the action played by the algorithm (you will have to use the greedy property).

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

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

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

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

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

- Finally, we have that

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

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

```
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 $\pi_{hk}$ 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

Solution : Regret Minimization in RL

- Let $\mathcal{E} = \{\forall k, M^\star \in \mathcal{M}_k\}$. We want to find $\beta_{hk}^r(s,a)$ and $\beta_{hk}^p(s,a)$, such that $\mathbb{P}(\neg\mathcal{E}) \leq \delta/2$ or alternatively, $\mathbb{P}(\mathcal{E}) \geq 1 - \delta/2$.

  We have :

$$\mathbb{P}(\neg\mathcal{E}) = \mathbb{P}\Big(\forall k, h, s, a : |\widehat{r}_{hk}(s,a) - r_h(s,a)| \leq \beta_{hk}^r(s,a) \bigcup \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \leq \beta_{hk}^p(s,a)\Big)$$

Thus :

$$\mathbb{P}(\neg\mathcal{E}) \leq \mathbb{P}\Big(\forall k,h,s,a : |\widehat{r}_{hk}(s,a) - r_h(s,a)| \leq \beta^r_{hk}(s,a))\Big) + \mathbb{P}\Big(\forall k,h,s,a : \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \leq \beta^p_{hk}(s,a)\Big)$$

And :

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

For the first term, we can use Hoeffding inequality :

$$\mathbb{P}\Big(|\widehat{r}_{hk}(s,a) - r_h(s,a)| \leq \beta^r_{hk}(s,a))\Big) \leq 2\exp\left\{-\frac{2N_{h,k}(a,s)^2\beta^r_{hk}(s,a)^2}{N_{h,k}(a,s)}\right\} = 2\exp\left\{-2N_{h,k}(a,s)\beta^r_{hk}(s,a)^2\right\}$$

For the second term, we can use Weissmain inequality :

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

Thus :

$$\mathbb{P}(\neg\mathcal{E}) \leq \sum_{k,h,s,a} 2\exp\left\{-2N_{h,k}(a,s)\beta^r_{hk}(s,a)^2\right\} + \sum_{k,h,s,a} (2^S - 2)\exp\left\{-\frac{N_{h,k}(a,s)\beta^p_{hk}(s,a)^2}{2}\right\}$$

If we ensure that :

$$\sum_{k,h,s,a} 2\exp\left\{-2N_{h,k}(a,s)\beta^r_{hk}(s,a)^2\right\} = \frac{\delta}{4} \text{ and } \sum_{k,h,s,a} (2^S - 2)\exp\left\{-\frac{N_{h,k}(a,s)\beta^p_{hk}(s,a)^2}{2}\right\} = \frac{\delta}{4}$$

We have the desired result.

Then taking :

$$\begin{cases} \beta^r_{hk}(s,a) = \sqrt{\frac{1}{2N_{h,k}(a,s)}\ln(\frac{8HKSA}{\delta})} \\ \beta^p_{hk}(s,a) = \sqrt{\frac{2}{N_{h,k}(a,s)}\ln(\frac{4HKSA(2^S-2)}{\delta})} \end{cases}$$

Ensures :

$$\mathbb{P}(\neg\mathcal{E}) \leq \frac{\delta}{2}$$

- Let's define the bonus function so that Q-function computed at episode k is optimistic. That is :

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

where $Q^\star$ is the optimal Q-function of the unknown MDP $M^\star$.

By definition of $Q^\star$:

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

And :

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

I define the bonus function as $b_{h,k}(s,a) = \beta^r_{hk}(s,a) + H\beta^p_{hk}(s,a)$.

Let's prove by induction that $Q_{h,k}(s,a) \geq Q^\star_h(s,a), \forall s,a$.

Base case :

$$\widehat{r}_{H,k}(s,a) + b_{H,k}(s,a) = \widehat{r}_{H,k}(s,a) + \beta^r_{hk}(s,a) + H\beta^p_{hk}(s,a)$$

And using that we are under $\mathcal{E}$ and that $\beta^p_{hk}(s,a) \leq 0$:

$$\widehat{r}_{H,k}(s,a) + b_{H,k}(s,a) \geq r_{H,k}(s,a) - \beta^r_{hk}(s,a) + \beta^r_{hk}(s,a) + H\beta^p_{hk}(s,a) \geq r_{H,k}(s,a)$$

Thus, $Q_{H,k}(s,a) \geq Q^\star_H(s,a)$.

Inductive step :

We can substract the two equations :

$$Q^\star_h(s,a) - Q_{h,k}(s,a) = r_h(s,a) - \widehat{r}_{h,k}(s,a) - b_{h,k}(s,a) + \sum_{s'} \left[p_h(s'|s,a)V^\star_{h+1}(s') - \widehat{p}_{h,k}(s'|s,a)V_{h+1,k}(s')\right]$$

As we are under $\mathcal{E}$: $|\widehat{r}_{hk}(s,a) - r_h(s,a)| \leq \beta^r_{hk}(s,a)$ and $\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \leq \beta^p_{hk}(s,a)$

So,

$$Q^\star_h(s,a) - Q_{h,k}(s,a) \leq \beta^r_{hk}(s,a) - b_{h,k}(s,a) + \sum_{s'} \left[p_h(s'|s,a)V^\star_{h+1}(s') - \widehat{p}_{h,k}(s'|s,a)V_{h+1,k}(s')\right]$$

Using the inductive hypothesis : $Q_{h+1,k}(s,a) \geq Q^\star_{h+1}(s,a)$ implies $V_{h+1,k}(s) \geq V^\star_{h+1}(s)$ (by taking the maximum).

Then :

$$Q^\star_h(s,a) - Q_{h,k}(s,a) \leq \beta^r_{hk}(s,a) - b_{h,k}(s,a) + \sum_{s'} \left[p_h(s'|s,a) - \widehat{p}_{h,k}(s'|s,a)\right] V_{h+1,k}(s')$$

Using Cauchy-Schwartz, and that we are under $\mathcal{E}$ and that we have $\|V_{h+1,k}\|_1 \leq H$ by definition of V_h:

$$\sum_{s'} \left[p_h(s'|s,a) - \widehat{p}_{h,k}(s'|s,a)\right] V_{h+1,k}(s') \leq \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \|V_{h+1,k}\|_1 \leq H\beta^p_{hk}(s,a)$$

Putting back everything together, and using the definition of we obtain :

$$Q^\star_h(s,a) - Q_{h,k}(s,a) \leq \beta^r_{hk}(s,a) - b_{h,k}(s,a) + H\beta^p_{hk}(s,a) = 0$$

Which is the desired result.

- 1. Let's show the suggested equality :

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

We start on the right hand side :

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

Thus the right hand side is equal to:

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

And using that:

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

Thus the right hand side can be written :

$$r(s_{hk}, a_{hk}) + \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V^{\pi_k}_{h+1}(Y)]$$

We recognize the bellman equation :

$$V^{\pi_k}_h(s_{hk}) = r(s_{hk}, a_{hk}) + \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V^{\pi_k}_{h+1}(Y)]$$

This gives the desired result.

2. Let show that $V_{h,k}(s_{hk}) \leq Q_{h,k}(s_{hk}, a_{hk})$.
   By definition of V_h :
   $$V_{h,k}(s) = \min\{H, \max_a Q_{h,k}(s,a)\}$$
   Thus for the current state:
   $$V_{h,k}(s_{hk}) = \min\{H, \max_a Q_{h,k}(s_{hk},a)\}$$
   This implies, using the definition of a_{hk} (greedy action):
   $$V_{h,k}(s_{hk}) \leq \max_a Q_{h,k}(s_{hk},a) = Q_{h,k}(s_{hk},a_{hk})$$

3. Substracting 2) - 1) we have :
   $$V_{h,k}(s_{hk}) \leq Q_{h,k}(s_{hk},a_{hk}) = V_h^{\pi_k}(s_{hk})$$
   And using 1) we have :
   $$V_{h,k}(s_{hk})-V_h^{\pi_k}(s_{hk}) = \delta_{h,k}(s_{h,k}) \leq Q_{h,k}(s_{hk},a_{hk})-r(s_{hk},a_{hk})-\mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)])+\delta_{h+1,k}(s_{h+1,k})+m$$
   Thus :
   $$\delta_{h,k}(s_{h,k}) - \delta_{h+1,k}(s_{h+1,k}) \leq Q_{h,k}(s_{hk},a_{hk}) - r(s_{hk},a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + m_{h,k}$$
   By summing and noticing that $\delta_{H+1,k}(s_{H+1,k}) = 0$:
   $$\sum_{h=1}^{H} \delta_{h,k}(s_{h,k}) - \delta_{h+1,k}(s_{h+1,k}) = \delta_{1,k}(s_{1,k}) - \delta_{H+1,k}(s_{H+1,k}) = \delta_{1,k}(s_{1,k})$$
   Thus summing the ineaqulity over h gives us the result :
   $$\delta_{1k}(s_{1,k}) \leq \sum_{h=1}^{H} Q_{hk}(s_{hk},a_{hk}) - r(s_{hk},a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + m_{hk}$$

- Let's show that the regret is upper bounded with probability $1-\delta$ by
  $$R(T) \leq 2\sum_{kh} b_{hk}(s_{hk},a_{hk}) + 2H\sqrt{KH\log(2/\delta)}$$
  We start by the definition of the regret :
  $$R(T) = \sum_{k=1}^{K} V_1^{\star}(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) = \sum_{k=1}^{K} \delta_{1k}(s_{1,k})$$
  By using the inequality of the previous question :
  $$R(T) \leq \sum_{k=1}^{K}\sum_{h=1}^{H} Q_{hk}(s_{hk},a_{hk}) - r(s_{hk},a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + m_{hk}$$
  By using the definition of Q_h :
  $$R(T) \leq \sum_{k=1}^{K}\sum_{h=1}^{H} \widehat{r}_{h,k}(s_{hk},a_{hk})+b_{h,k}(s_{hk},a_{hk})+\sum_{s'}\widehat{p}_{h,k}(s'|s_{hk},a_{hk})V_{h+1,k}(s')-r(s_{hk},a_{hk})-\mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)$$
  $$R(T) \leq \sum_{k=1}^{K}\sum_{h=1}^{H} \widehat{r}_{h,k}(s_{hk},a_{hk})-r(s_{hk},a_{hk})+b_{h,k}(s_{hk},a_{hk})+\sum_{s'}(\widehat{p}_{h,k}(s'|s_{hk},a_{hk})-p_h(s'|s_{hk},a_{hk}))V_{h+1,k}(s')+m_{hk}$$

Using that we are under $\mathcal{E}$:

$$R(T) \leq \sum_{k=1}^{K} \sum_{h=1}^{H} \beta^r_{hk}(s_{hk}, a_{hk}) + b_{h,k}(s_{hk}, a_{hk}) + H\beta^p_{hk}(s_{hk}, a_{hk}) + m_{hk}$$

By the definition of the bonus :

$$R(T) \leq \sum_{k=1}^{K} \sum_{h=1}^{H} 2b_{h,k}(s_{hk}, a_{hk}) + m_{hk}$$

Thus, using Azuma-Hoeffding inequality :

$$R(T) \leq \sum_{k=1}^{K} \sum_{h=1}^{H} 2b_{h,k}(s_{hk}, a_{hk}) + 2H\sqrt{KH \log(2/\delta)}$$

Which is the desired result.

- Let's write the final regret bound. We start by showing an upper-bound of

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

Using the indication, we have :

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

Let's denote $\alpha_{a,s} = \frac{1}{SA}$, we have $\sum_{s,a} \alpha_{a,s} = 1$ We can use the concavity of the square root to write :

$$\sum_{s,a} \alpha_{a,s} \sqrt{N_{h,K}(s,a)} \leq \sqrt{\sum_{s,a} \alpha_{a,s} N_{h,K}(s,a)}$$

Thus we have :

$$\sum_{s,a} \sqrt{N_{h,K}(s,a)} \leq SA \sqrt{\sum_{s,a} \frac{N_{h,K}(s,a)}{SA}} \leq \sqrt{SA} \sqrt{\sum_{s,a} N_{h,K}(s,a)}$$

And as $\sum_{s,a} N_{h,K}(s,a) = K$, we have :

$$\sum_{s,a} \sqrt{N_{h,K}(s,a)} \leq \sqrt{SAK}$$

Putting back everything together, we obtain :

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

Let's compute the final bound on the regret :

$$\sum_{k=1}^{K} \sum_{h=1}^{H} 2b_{h,k}(s_{hk}, a_{hk}) = 2 \sum_{k=1}^{K} \sum_{h=1}^{H} \beta^r_{hk}(s,a) + H\beta^p_{hk}(s,a)$$

And :

$$\sum_{k=1}^{K}\sum_{h=1}^{H} 2b_{h,k}(s_{hk},a_{hk}) = 2\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{2N_{h,k}(a,s)}\ln(\frac{8HKSA}{\delta})}+H\sqrt{\frac{2}{N_{h,k}(a,s)}\ln(\frac{4HKSA(2^S-2)}{\delta})}$$

We have :

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{2N_{h,k}(a,s)}\ln(\frac{8HKSA}{\delta})} = \widetilde{O}(H\sqrt{SAK})$$

And :

$$\sum_{k=1}^{K}\sum_{h=1}^{H} H\sqrt{\frac{2}{N_{h,k}(a,s)}\ln(\frac{4HKSA(2^S-2)}{\delta})} = \widetilde{O}(H\sqrt{SAK}) \times \widetilde{O}(H\sqrt{S}) = \widetilde{O}(H^2 S\sqrt{AK})$$

Thus, as :

$$R(T) \leq \sum_{k=1}^{K}\sum_{h=1}^{H} 2b_{h,k}(s_{hk},a_{hk}) + 2H\sqrt{KH\log(2/\delta)}$$

We finally get :

$$R(T) = \widetilde{O}(H^2 S\sqrt{AK}) + \widetilde{O}(H\sqrt{SAK}) = \widetilde{O}(H^2 S\sqrt{AK})$$

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

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