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

Exploration in Reinforcement Learning (theory)

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

Solution by AMEKOE Kodjo Mawuena

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.

UCB Answers

$\widehat{\mu}_{j,t}$ is not an unbiased estimator of μ_j.

Counter example (Inspired from the paper [1]):

Suppose that we continuously alternate between drawing reward from Bernoulli distribution of parameters $\mu_1, \mu_2 \in]0, 1[$. Define t as the first time we observe 1 from the first arm (μ_1). In this case case, $N_{1,t}$ follows geometric distribution of parameter μ_1.
We have then $\mathbb{E}_{UCB}[\widehat{\mu}_{1,t}] = \mathbb{E}[\frac{S_{1,t}}{N_{1,t}}] = \mathbb{E}[\frac{1}{N_{1,t}}]$ because $S_{1,t} = 1$.

Moreover if X follows geometric distribution of parameter μ_1 then $\mathbb{E}[\frac{1}{X}] = \dfrac{\mu_1 \log(\frac{1}{\mu_1})}{1 - \mu_1}$

[1] Are sample means in multi-armed bandits positively or negatively biased? Example 1

Proof:

$$
\begin{aligned}
\mathbb{E}[\frac{1}{X}] &= \sum_{k=1}^{\infty} \frac{1}{k}\mathbb{P}(X = k) \\
&= \sum_{k=1}^{\infty} \frac{1}{k}(1-\mu_1)^{k-1}\mu_1 \\
&= \frac{\mu_1}{1-\mu_1} \sum_{k=1}^{\infty} \frac{1}{k}(1-\mu_1)^{k} \\
&= \frac{\mu_1}{1-\mu_1} \sum_{k=0}^{\infty} \frac{1}{k+1}(1-\mu_1)^{k+1} \\
&= \frac{\mu_1}{1-\mu_1} \sum_{k=0}^{\infty} \int -(1-\mu_1)^{k} d\mu_1 \\
&= \frac{\mu_1}{1-\mu_1} \int -\sum_{k=0}^{\infty} (1-\mu_1)^{k} d\mu_1 \\
&= \frac{\mu_1}{1-\mu_1} \int -\frac{1}{\mu_1} d\mu_1 \\
&= -\frac{\mu_1}{1-\mu_1} \log(\mu_1) \\
\mathbb{E}[\frac{1}{X}] &= \frac{\mu_1 \log(\frac{1}{\mu_1})}{1-\mu_1}
\end{aligned}
$$

So we have $\mathbb{E}_{UCB}[\widehat{\mu}_{1,t}] = \dfrac{\mu_1 \log(\frac{1}{\mu_1})}{1-\mu_1} \neq \mu_1$, hence this estimation is not 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 $\mathbb{P}_{\mu_1,\ldots,\mu_k}(\widehat{i} \neq i^\star) \leq \delta$ and $\tau_\delta < \infty$ almost surely for any $\mu_1, \ldots, \mu_k$. Our goal is to find a δ-correct algorithm that minimizes the sample complexity, that is, $\mathbb{E}[\tau_\delta]$ the expected number of sample needed to predict an answer.

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.

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

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 $i \in [k]$

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

  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.

Best Arm Identification Answers

- Computation of $U(t, \delta)$ that satisfy $\mathbb{P}\left(\{|\widehat{\mu}_{i,t} - \mu_i| > U(t, \delta)\}\right) \leq \frac{\delta}{2t^2}$

  We will use Hoeffding's inequality :

  If $X_1, ..., X_n$ are independent random variables bounded by the interval $[0, 1]$ and $\bar{X} = \frac{1}{n}(X_1 + .. + X_n)$ then $P(|\bar{X} - E[\bar{X}]| \geq u) \leq 2e^{-2nu^2}$. Hence :

  $\mathbb{P}\left(\{|\widehat{\mu}_{i,t} - \mu_i| > U(t, \delta)\}\right) \leq 2e^{-2tU^2}$ where $U = U(t, \delta)$.

  Then for $2e^{-2tU^2} = \frac{\delta}{2t^2}$, we have $\underline{U(t, \delta) = \sqrt{\frac{\log(4t^2\frac{1}{\delta})}{2t}}}$

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

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &= \mathbb{P}(\bigcup_{i=1}^{k} \bigcup_{t=1}^{\infty} \{|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta')\}) \\
&\leq \sum_{i=1}^{k} \mathbb{P}(\bigcup_{t=1}^{\infty} \{|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta')\}) \\
&\leq \sum_{i=1}^{k} \sum_{t=1}^{\infty} \mathbb{P}(\{|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta')\}) \\
&\leq \sum_{i=1}^{k} \sum_{t=1}^{\infty} \frac{\delta'}{2t^2} \quad \text{because} \quad \mathbb{P}(\{|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta')\}) \leq \frac{\delta'}{2t^2} \\
&= \sum_{i=1}^{k} \delta' \frac{\pi^2}{12} \quad \text{because} \quad \sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6} \\
\mathbb{P}(\mathcal{E}) &\leq k\delta' \frac{\pi^2}{12}
\end{aligned}$$

  For $k\delta' = \delta$, we already have $\mathbb{P}(\mathcal{E}) \leq \delta$. So $\underline{\delta' = \frac{\delta}{k}}$

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

  Let's assume that $\mathbb{P}(\mathcal{E}) \leq \delta$ (then $\mathbb{P}(\neg\mathcal{E}) > 1-\delta$).

  Hence $|\widehat{\mu}_{i,t} - \mu_i| \leq U(t,\delta')$ $\forall i,t$, with probability at least $1-\delta$

  $\implies \widehat{\mu}_{i,t} - \mu_i \geq -U(t,\delta')$

  $\implies \underline{\widehat{\mu}_{i,t} + U(t,\delta') \geq \mu_i}$, $\forall i$, in particular, $\underline{\widehat{\mu}_{i^\star,t} + U(t,\delta') \geq \mu^\star}$

  Let's assume that arm $i^\star$ is removed from the active set. Then $\exists j \neq i^\star \in S$, $\widehat{\mu}_{j,t} - U(t,\delta') \geq \widehat{\mu}_{i^\star,t} + U(t,\delta')$

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

  $\implies \widehat{\mu}_{j,t} - U(t,\delta') > \mu_j$

  $\implies \widehat{\mu}_{j,t} - \mu_j > U(t,\delta')$. This is contradictory to the fact that $|\widehat{\mu}_{i,t} - \mu_i| \leq U(t,\delta')$ $\forall i$.

  So $i^\star$ remains in the active set at least we probability $1-\delta$.

- Under event $\neg\mathcal{E}$, 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 :

  Let's show that $\exists t \geq 1$ such that for $i \neq i^\star$ will be removed from S.

  Let's assume that the event $\neg\mathcal{E}$ holds. Then $\forall t, i^\star \in S$ with probability at least $1-\delta$.

  Since $i^\star \in S$, we want to prove that $\widehat{\mu}_{i^\star,t} - U(t,\delta') \geq \widehat{\mu}_{i^\star,t} + U(t,\delta')$ (So that the arm i will be removed from S)

  Under $\neg\mathcal{E}$, we have $|\widehat{\mu}_{i,t} - \mu_i| \leq U(t,\delta') \implies \underline{\widehat{\mu}_{i,t} - \mu_i \leq U(t,\delta')}$
  we have also $|\widehat{\mu}_{i^\star,t} - \mu^\star| \leq U(t,\delta') \implies \underline{\mu^\star - \widehat{\mu}_{i\star,t} \leq U(t,\delta')}$

  By summing the two underline in-equations, and for $U = U(t,\delta')$ we get :

$$\begin{aligned}
&\widehat{\mu}_{i,t} - \mu_i + \mu^\star - \widehat{\mu}_{i^\star,t} \leq 2U \\
&\implies \widehat{\mu}_{i,t} + \Delta_i - \widehat{\mu}_{i^\star,t} \leq 2U \\
&\implies \widehat{\mu}_{i^\star,t} \geq -2U + \widehat{\mu}_{i,t} + \Delta_i \\
&\implies \widehat{\mu}_{i^\star,t} - U + U \geq -2U + \widehat{\mu}_{i,t} + U - U + \Delta_i \\
&\implies \widehat{\mu}_{i^\star,t} - U \geq \widehat{\mu}_{i,t} + U + (-4U + \Delta_i) \\
&\implies \underline{\widehat{\mu}_{i^\star,t} - U \geq \widehat{\mu}_{i,t} + U} \quad \text{for} \quad -4U + \Delta_i \geq 0
\end{aligned}$$

Hence we have $\Delta_i \geq 4U$ (ie $C_1 = 4$).

Since $U(t,\delta') = \sqrt{\frac{\log(4t^2\frac{1}{\delta})}{2t}}$, $\lim_{t\to\infty} U(t,\delta') = 0$.

Then $\exists t$ such that $i \neq i^\star$ will be removed.

The time required is obtained by solving :

$$\begin{aligned}\Delta_i \geq 4U(t,\delta') &\implies \Delta_i/4 \geq \sqrt{\frac{\log(4t^2\frac{1}{\delta})}{2t}} \\ &\implies (\Delta_i/4)^2 \geq \frac{\log(4t^2\frac{1}{\delta})}{2t}\end{aligned}$$

- Computation of the bound on the sample complexity :

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.

Bernoulli Bandits Answers

- The expected regret as function of Δ

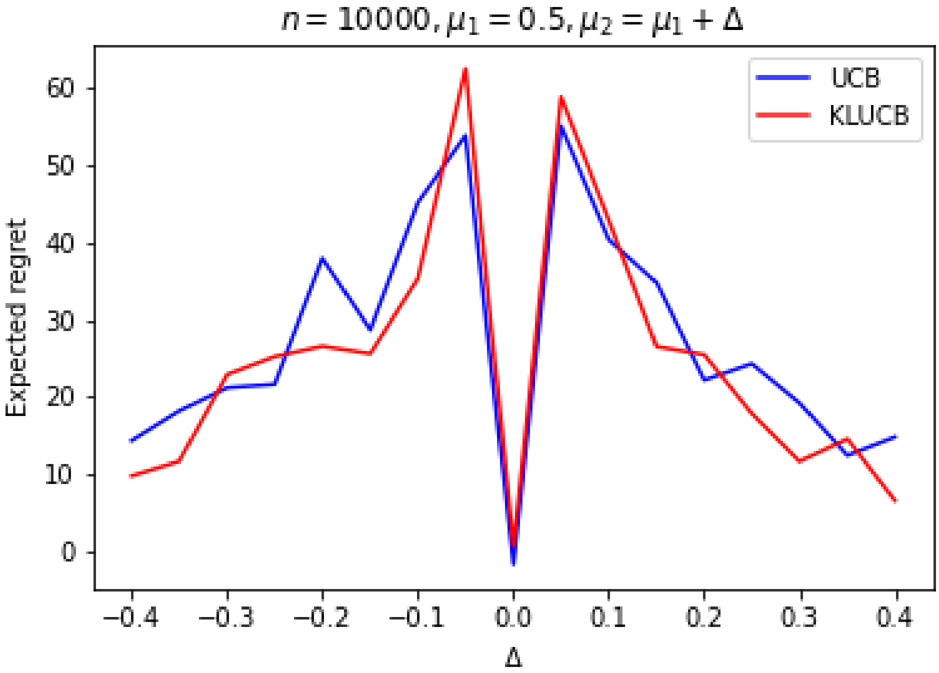

Figure 1: The expected reward over 100 simulations for UCB and KLUCB algorithms

- Now we take $\mu_1 = 0.1, \quad \mu_1 = 0.9$

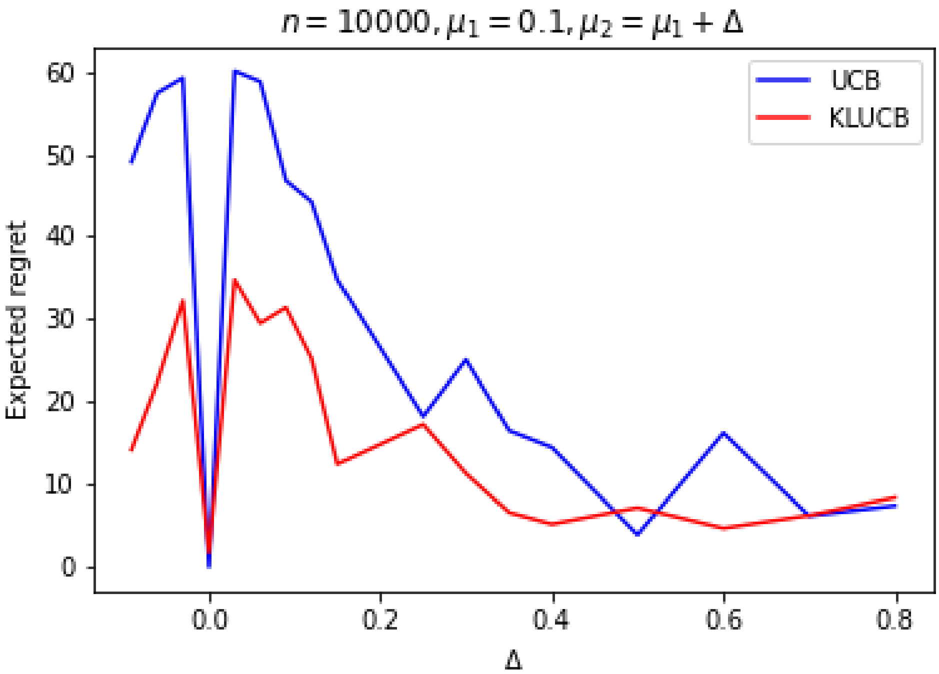

Figure 2: The expected reward over 100 simulations for UCB and KLUCB algorithms

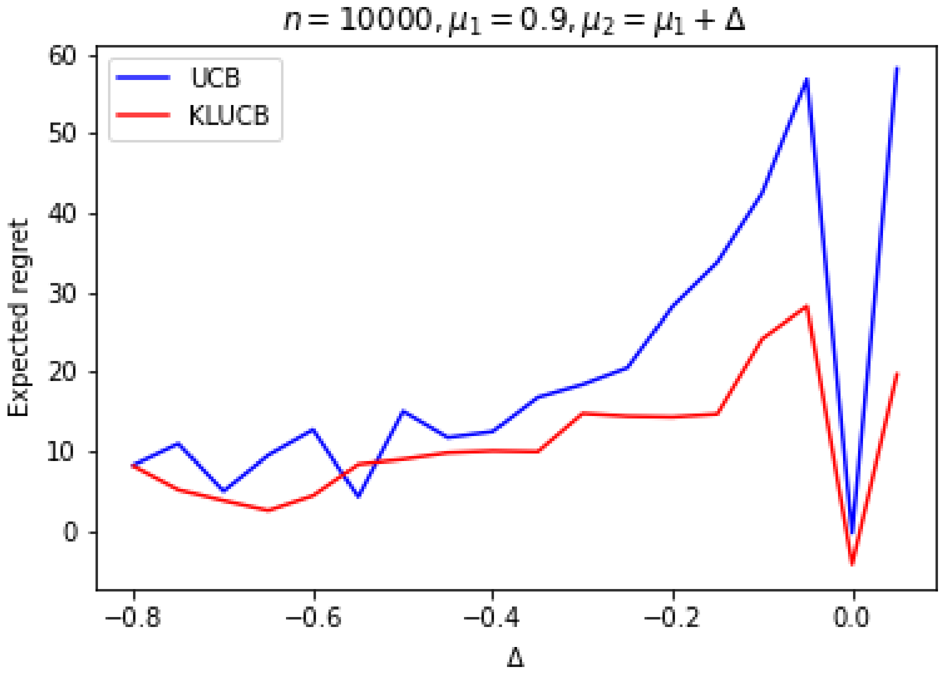

Figure 3: The expected reward over 100 simulations for UCB and KLUCB algorithms

- Discussion

  For all the values of μ_1 considered, we can first notice that the regret becomes great the closer Δ gets to zero without taking the value zero. This can be explained by the fact that the two arms are very similar and therefore the algorithms have a hard time finding the best one.

  For $\mu_1 = 0.1, \quad \mu_1 = 0.9$ we can see that KLUCB is much more efficient (precise) than UCB.

  There is a decreasing trend in regret with Delta for $\mu_1 = 0.1$ (see Figure 2) . This can be explained by the fact that the bigger Δ is, the more differentiated the arms are (so easy to find the best one). It is the same for $\mu_1 = 0.9$, there is increasing trend of the regret Figure 3. See the notebook Code_A3_Amekoe for the code.

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

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

- 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. 5.

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

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

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

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

Regret Minimization in RL Answers

- We define the event:

$$\begin{aligned}\mathcal{E} &= \{\forall k, M^\star \in \mathcal{M}_k\} \\ &= \{(S, A, p_h, r_h), \quad |\widehat{r}_{hk}(s,a) - r_h(s,a)| \le \beta^r_{hk}(s,a) \wedge \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \le \beta^p_{hk}(s,a), \forall k, \forall (s,a) \in S \times A\}\end{aligned}$$

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

Proof:

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

By Hoeffding's inequality , $P(|\bar{X} - E[\bar{X}]| \ge u) \le 2e^{-2nu^2}$, hence $P\left(|\widehat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)\right) \le 2exp\left(-2N_{hk}\beta^r_{hk}(s,a)^2\right)$.

So if we set : $2exp\left(-2N_{hk}(s,a)\beta^r_{hk}(s,a)^2\right) = \dfrac{\delta}{4SAHK}$, we will have:

$$\begin{aligned}-2N_{hk}(s,a)\beta^r_{hk}(s,a)^2 = \log\left(\frac{\delta}{8SAHK}\right) &\implies \beta^r_{hk}(s,a)^2 = \frac{\log(8SAHK\frac{1}{\delta})}{2N_{hk}(s,a)} \\ &\implies \beta^r_{hk}(s,a) \le \sqrt{\frac{\log(8SAHK\frac{1}{\delta})}{2N_{hk}(s,a)}}\end{aligned}$$

So we have

$$\beta^r_{hk}(s,a) = \sqrt{\frac{\log(8SAHK\frac{1}{\delta})}{2N_{hk}(s,a)}} \tag{2}$$

and $\underline{P\left(|\widehat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)\right) \le \dfrac{\delta}{4SAHK} \le \dfrac{\delta}{4}}$.

Moreover by Weissmain inequality, we have :

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

If we set $(2^S - 2)\exp\Big(-\frac{N_{hk}(s,a)\beta^p_{hk}(s,a)^2}{2}\Big) = \dfrac{\delta}{4SAHK}$ we will get :

$$\beta^p_{hk}(s,a) = \sqrt{\frac{2\log((2^S-2)4SAHK\frac{1}{\delta})}{N_{hk}(s,a)}} \tag{3}$$

and $\underline{\mathbb{P}(\|\widehat{p}_h(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \geq \beta^p_{hk}(s,a)) \leq \dfrac{\delta}{4SAHK} \leq \dfrac{\delta}{4}}$

Finally we have : $\mathbb{P}(\neg\mathcal{E}) \leq \delta/4 + \delta/4 = \delta/2$.

- We consider the bonus function

$$b_{h,k}(s,a) = H\sqrt{\frac{2\log((2^S-2)4SAHK\frac{1}{\delta})}{N_{hk}(s,a)}} \tag{4}$$

and

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

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

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

Proof:

  - If $h = H$, we have $\widehat{r}_{H,k}(s,a) + b_{H,k}(s,a) \geq r_{H,k}(s,a)$ and thus $Q_{H,k}(s,a) \geq Q^\star_H(s,a)$.
  - for $h < H$, we will use induction : We suppose that, $Q_{h,k}(s,a) \geq Q^\star_h(s,a)$.
    Since $V_{h,k}(s) = \min\{H, \max_a Q_{h,k}(s,a)\}$ and $Q_{h,k}(s,a) \geq Q^\star_h(s,a)$,
    we have also $\underline{V_{h,k}(s) \geq V^\star_h(s) \quad, \forall k}$.
    Then :

$$\begin{aligned}
Q_{h-1,k}(s,a) &= \widehat{r}_{h-1,k}(s,a) + b_{h-1,k}(s,a) + \sum_{s'} \widehat{p}_{h-1,k}(s'|s,a)V_{h,k}(s') \\
&\geq \widehat{r}_{h-1,k}(s,a) + b_{h-1,k}(s,a) + \sum_{s'} \widehat{p}_{h-1,k}(s'|s,a)V^\star_h(s)(s') \quad \text{because} \quad V_{h,k}(s) \geq V^\star_h(s) \\
&\geq r_{h-1}(s,a) + \sum_{s'} p_{h-1}(s'|s,a)V^\star_h(s)(s') \quad \text{under} \quad \mathcal{E} \\
&= Q^\star_{h-1}(s,a)
\end{aligned}$$

    So $Q_{h-1,k}(s,a) \geq Q^\star_{h-1}(s,a)$.
    Finally we get $\underline{Q_{h,k}(s,a) \geq Q^\star_h(s,a), \forall s,a}$

- We have seen (In class) that :

$$\delta_{1k}(s_{1,k}) \leq \sum_{h=1}^{H} Q_{hk}(s_{hk},a_{hk}) - r(s_{hk},a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + m_{hk} \tag{5}$$

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

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

$$\begin{aligned}
&\implies \delta_{h+1,k}(s_{h+1,k}) = \mathbb{E}_p[\delta_{h+1,k}(s')] - m_{hk}\\
&\implies \delta_{h+1,k}(s_{h+1,k}) = \mathbb{E}_p\left[V_{h+1,k}(s') - V_{h+1}^{\pi_k}(s')\right] - m_{hk}\\
&\implies \delta_{h+1,k}(s_{h+1,k}) = \mathbb{E}_p\left[V_{h+1,k}(s')\right] - V_h^{\pi_k}(s_{hk}) + r(s_{hk}, a_{hk}) - m_{hk} \quad \text{(the Bellman equation)}\\
&\implies \underline{V_h^{\pi_k}(s_{hk}) = r(s_{hk}, a_{hk}) + \mathbb{E}_p\left[V_{h+1,k}(s')\right] - \delta_{h+1,k}(s_{h+1,k}) - m_{hk}}
\end{aligned}$$

2. $V_{h,k}(s_{hk}) \leq Q_{h,k}(s_{hk}, a_{hk})$.
   Proof : We have $V_{h,k}(s) = \min\{H, \max_a Q_{h,k}(s,a)\}$. Since a_{hk} is the greedy action, we have

   $\max_a Q_{h,k}(s_{hk}, a) = Q_{h,k}(s_{hk}, a_{hk})$

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

3. Finally we have :

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

   We will use induction for this proof :

$$\begin{aligned}
\delta_{hk}(s_{hk}) &= V_{hk}(s_{hk}) - V_h^{\pi_k}(s_{hk})\\
&\leq Q_{hk}(s_{hk}, a_{hk}) - V_h^{\pi_k}(s_{hk})\\
&= Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_p\left[V_{h+1,k}(s')\right] + \delta_{h+1,k}(s_{h+1,k}) + m_{hk}.
\end{aligned}$$

   Then we have:

   - for $h = H$:

$$\begin{aligned}
\delta_{Hk}(s_{hk}) &\leq Q_{Hk}(s_{Hk}, a_{Hk}) - r(s_{Hk}, a_{Hk}) - \mathbb{E}_p\left[V_{H+1,k}(s')\right] + m_{Hk} \quad \text{because} \quad \delta_{H+1,k}(s_{H+1,k}) = 0\\
&= \sum_{h=H}^{H} Q_{hk}(s_{hk}, a_{Hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_p\left[V_{h+1,k}(s')\right] + m_{hk}
\end{aligned}$$

   - Now let assume the property is true for $h < H$: $\delta_{hk}(s_{hk}) = \sum_h^H Q_{hk}(s_{hk}, a_{Hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_p\left[V_{h+1,k}(s')\right] + m_{hk}$ and prove that it also true for $h-1$.
     We have :

$$\begin{aligned}
\delta_{h-1,k}(s_{h-1,k}) &\leq Q_{h-1,k}(s_{h-1,k}, a_{h-1,k}) - r(s_{h-1,k}, a_{h-1,k}) - \mathbb{E}_p\left[V_{h,k}(s')\right] + \delta_{h,k}(s_{h,k}) + m_{h-1,k}\\
&= Q_{h-1,k}(s_{h-1,k}, a_{h-1,k}) - r(s_{h-1,k}, a_{h-1,k}) - \mathbb{E}_p\left[V_{h,k}(s')\right] + m_{h-1,k}\\
&+ Q_{h,k}(s_{h,k}, a_{h,k}) - r(s_{h,k}, a_{h,k}) - \mathbb{E}_p\left[V_{h+1,k}(s'')\right] + m_{h,k}\\
&= \sum_{u=h-1}^{H} Q_{uk}(s_{uk}, a_{uk}) - r(s_{uk}, a_{uk}) - \mathbb{E}_p\left[V_{u+1,k}(s')\right] + m_{uk}
\end{aligned}$$

   Then $\forall h$ we have : $\delta_{h,k}(s_{h,k}) \leq \sum_{u=h}^{H} Q_{uk}(s_{uk}, a_{uk}) - r(s_{uk}, a_{uk}) - \mathbb{E}_p\left[V_{u+1,k}(s')\right] + m_{uk}$, in particular for $h = 1$, we get :
   $\underline{\delta_{1,k}(s_{1,k}) \leq \sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_p\left[V_{h+1,k}(s')\right] + m_{hk}}$

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

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

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

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

Proof :

$$\begin{aligned}
R(T) &= \sum_{k=1}^{K} V_1^{\star}(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) \\
&\leq \sum_{k}^{K} V_{1,k}(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) \quad \text{under the event} \quad \mathcal{E} \\
&\leq \sum_{k=1}^{K} \delta_{1,k}(s_{1,k}) \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_p\left[V_{h+1,k}(s')\right] + m_{hk} \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_p\left[V_{h+1,k}(s')\right] + 2H\sqrt{KH\log(2/\delta)} \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H} \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \mathbb{E}_{\widehat{p}}\left[V_{h+1,k}(s')\right] - r(s_{hk}, a_{hk}) - \mathbb{E}_p\left[V_{h+1,k}(s')\right] + 2H\sqrt{KH\log(2/\delta)} \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H} b_{h,k}(s,a) + \widehat{r}_{h,k}(s,a) - r(s_{hk}, a_{hk}) + \mathbb{E}_{\widehat{p}}\left[V_{h+1,k}(s')\right] - \mathbb{E}_p\left[V_{h+1,k}(s')\right] + 2H\sqrt{KH\log(2/\delta)} \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} b_{h,k}(s,a) + b_{h,k}(s,a) + 2H\sqrt{KH\log(2/\delta)} \quad \text{under the event} \quad \mathcal{E} \\
&= 2\sum_{k,h} b_{h,k}(s,a) + 2H\sqrt{KH\log(2/\delta)}
\end{aligned}$$

Hence $\underline{R(T) \leq 2\sum_{k,h} b_{h,k}(s,a) + 2H\sqrt{KH\log(2/\delta)}}$

- Finally

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

By Cauchy-Schawrtz inequality, we have :

$$\begin{aligned}
\sum_{s,a}\sqrt{N_{hK}(s,a)} &= \sum_{s,a} 1 \times \sqrt{N_{hK}(s,a)} \\
&\leq \sqrt{\sum_{s,a} 1^2}\sqrt{\sum_{s,a}(\sqrt{N_{hK}(s,a)})^2} \\
&= \sqrt{SA}\sqrt{\sum_{s,a} N_{hK}(s,a)} \\
&= \sqrt{SA\sum_{s,a} N_{hK}(s,a)}
\end{aligned}$$

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

We have :
$R(T) \lesssim H^2S\sqrt{AK}$
By using the expression of our bonus (4), we have:

$$R(T) \leq 2\sum_{k,h} b_{h,k}(s,a) + 2H\sqrt{KH\log(2/\delta)}$$

$$= 2\sum_{k,h} H\sqrt{\frac{2\log((2^S-2)4SAHK\frac{1}{\delta})}{N_{hk}(s,a)}} + 2H\sqrt{KH\log(2/\delta)}$$

$$= 2H\sqrt{2\log((2^S-2)4SAHK\frac{1}{\delta})}\sum_{k,h}\frac{1}{\sqrt{N_{hk}(s,a)}} + 2H\sqrt{KH\log(2/\delta)}$$

$$\leq 2H\sqrt{2\log(2^S4SAHK\frac{1}{\delta})}\sum_{k,h}\frac{1}{\sqrt{N_{hk}(s,a)}} + 2H\sqrt{KH\log(2/\delta)}$$

$$\leq 2H\sqrt{2\log(2^S4SAHK\frac{1}{\delta})}(2H\sqrt{SAK}) + 2H\sqrt{KH\log(2/\delta)}$$

$$\leq 4H^2\sqrt{SAK}\sqrt{2\log(2^S(4SAHK\frac{1}{\delta})^S)} + 2H\sqrt{KH\log(2/\delta)} \quad \text{because} \quad 4SAHK\frac{1}{\delta} \geq 1, \quad \text{so}$$

$$4SAHK\frac{1}{\delta} \leq (4SAHK\frac{1}{\delta})^S$$

$$= 4H^2\sqrt{SAK}\sqrt{2\log((8SAHK\frac{1}{\delta})^S)} + 2H\sqrt{KH\log(2/\delta)}$$

$$= 4H^2\sqrt{SAK}\sqrt{2S\log(8SAHK\frac{1}{\delta})} + 2H\sqrt{KH\log(2/\delta)}$$

$$= 4H^2S\sqrt{2AK\log(8SAHK\frac{1}{\delta})} + 2H\sqrt{KH\log(2/\delta)}$$

$$= 6H^2S\sqrt{2AK\log(8SAHK\frac{1}{\delta})}$$

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

B conclusion

This assignment allowed us to review the theories behind LR and Bandits exploration, as well as Bernoulli Bandits experimentation. Unfortunately I could not find the solution to the last question of the Best Arm Identification part before the homework deadline.