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

Exploration in Reinforcement Learning (theory)

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

Solution by Mahdi KALLEL

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.

In what follows we will give an example where the the UCB algorithm is **Negatively biased**:

Let's consider a setting with two bernoulli arms. $X_1 \sim B(\mu_1), X_2 \sim B(\mu_2)$ where $\mu_1 > \mu_2$.
For simplicity we suppose that whenever the estimated values of both arms are equal we choose arm 1.
We consider taking only 3 steps in this setting.

$$\mathbb{E}[\widehat{\mu}_{1,3}] = P(i_3 = 1) * \mathbb{E}[\widehat{\mu}_{1,3}|i_3 = 1] + P(i_3 = 2) * \mathbb{E}[\widehat{\mu}_{1,3}|i_3 = 2]$$

$$= P(\widehat{\mu}_{1,1} \geq \widehat{\mu}_{2,2}) * \mathbb{E}[\widehat{\mu}_{1,3}|\widehat{\mu}_{1,1} \geq \widehat{\mu}_{2,2}] + P(\widehat{\mu}_{1,1} < \widehat{\mu}_{2,2}) * \mathbb{E}[\widehat{\mu}_{1,1}|\widehat{\mu}_{1,1} < \widehat{\mu}_{2,2}]$$

$$P(X_1 \geq X_2) * \mathbb{E}[\frac{2X_1}{2}] + P(X_2 > X_1) * 0\,[**]$$

$$(**)\widehat{\mu}_{1,1} < \widehat{\mu}_{2,2} \implies \widehat{\mu}_{1,1} = 0$$

$$\implies \mathbb{E}[\widehat{\mu}_{1,3}] = \mu_1 * (\mu_1(1 - \mu_1)(1 - \mu_2)$$

$$\implies \mathbb{E}[\widehat{\mu}_{1,3}] - \mu_1 = \mu_1\mu_2(\mu_1 - 1) < 0$$

This negative biased can be proved for more general settings that go even beyond UCB. In the (1) the authors provide a proof. The intuition is what follows :
We consider a setting with T time steps, suppose we are at time $t < T$ with a sample trajectory δ_t with corresponding sample evaluations $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_{1,t} \dots \widehat{\mu}_{n,t})$ we encounter two cases :

- $\widehat{\mu}_{1,t} > \mu_1$ in which case the hand 1 is more likely to be chosen in the next time steps since the next samples have the expected values $\mu_1 < \widehat{\mu}_{1,t}$ this makes our model decrease it's estimate towards the real values
- $\widehat{\mu}_{1,t} > \mu_1$ in which case hand 1 will be picked less often leading to less updates to the estimate of this arm, therefore there's a higher probability we get stuck with the negative bias.

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 δ
S = {1, ..., k}
for t = 1, ... do
    Pull all arms in S
    S = S \ { i ∈ S : ∃j ∈ S, μ̂_{j,t} − U(t,δ) ≥ μ̂_{i,t} + U(t,δ) }
    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 $i \in [k]$

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

  Use Hoeffding's inequality.

$$N_{i,t} = \sum_{j=1}^{t} \mathbf{1}_{i \in S_j},\ \tilde{\mu}_{i,t} = \frac{\sum_{j=1}^{t} X_{i,j} \mathbf{1}_{i \in S_j}}{N_{i,t}}$$

From hoeffding inequality we have that :

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

$$P(\bigcup_{t=1}^{\infty} |\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta)) \leq \sum_{t=1}^{\infty} P(|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta)) \leq \sum_{t=1}^{\infty} 2e^{-2N_{i,t}U(t,\delta)^2} = \boldsymbol{S}$$

Since we are picking all the remaining arms we have $N_{i,t} = \begin{cases} t \text{ if } i \in S_t \\ t' < t \text{ where } t' \text{ is the last time we picked i} \end{cases}$

We want to ensure the convergence of the series $\boldsymbol{S}$ and that $\lim_{\infty} U(t,\delta) = 0$

Therefore choose the terms of the series such that $2e^{-2N_{i,t}U(t,\delta)^2} \leq \frac{\alpha}{t^2}$

$$\implies U(t,\delta) \geq \sqrt{\frac{ln(\frac{t}{\sqrt{\alpha}})}{t}}$$

$$\sum_{t=1}^{\infty} \frac{\alpha}{t^2} = \frac{2\alpha\pi^2}{6} = \delta \implies \alpha = \frac{3\delta}{\pi^2}$$

If we choose $U(t,\delta) = \sqrt{\frac{ln(\frac{t\pi}{\sqrt{3\delta}})}{t}}$ we ensure that $\mathbb{P}\left(\bigcup_{t=1}^{\infty} \{|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta)\}\right) \leq \delta$ and that $lim_{\infty} U(t,\delta) = 0$

- 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.

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

$$\leq \sum_{i=1}^{k} \delta' \leq k\delta'$$

$$\implies k\delta' = \delta$$

- Show that with probability at least $1-\delta$, the optimal arm $i^\star = \arg\max_i \{\mu_i\}$ remains in the active set S. Use your definition of δ' and start from the condition for arm elimination. From this, use the definition of $\neg\mathcal{E}$.

According to the algorithm, the event of choosing a wrong arm is $A = \{\exists(i,t) \text{ st. } \mu_{i,t} \geq \mu_{i^\star,t}+2U(t,\delta')\}$

We have two cases :

–
$$\text{If } |\mu_{i,t} - \mu_i| \leq U(t,\delta') \implies U(t,\delta') \geq \mu_{i,t} - \mu_i \geq \mu_{i^\star,t} - \mu_i + 2U(t,\delta')$$
$$\text{Since } \mu_{i^\star} \geq \mu_i \implies U(t,\delta') \geq \mu_{i,t} - \mu_{i^\star} \geq \mu_{i^\star,t} - \mu_{i^\star} + 2U(t,\delta')$$

$$\implies \mu_{i^\star,t} - \mu_i^\star \leq -U(t,\delta') \implies \mathcal{E}$$

–
$$\text{If } |\mu_{i,t} - \mu_i| > U(t,\delta') \implies \mathcal{E}$$

We proved that $A \implies \mathcal{E}$ therefore $P(A) \leq P(\mathcal{E}) \leq \delta \implies 1-\delta \leq 1-P(\mathcal{E}) \leq P(\neg A)$ which is the wanted result.

- 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 $i^\star$.

A non optimal arm can be removed if $\mu_{i^\star,t} \geq \mu_{i,t} + 2U(t,\delta)$

$$\Delta_i = \mu_{i^\star} - \mu_i = (\mu_{i^\star} - \mu_{i^\star,t}) + (\mu_{i,t} - \mu_i) + (\mu_{i^\star,t} - \mu_{i,t})$$
$$if \Delta_i \geq 4U(t,\delta) \implies (\mu_{i^\star} - \mu_{i^\star,t}) + (\mu_{i,t} - \mu_i) + (\mu_{i^\star,t} - \mu_{i,t}) \geq 4U(t,\delta)$$

$$\text{Under the event } \neg\mathcal{E}, (\mu_{i,t} - \mu_i) + (\mu_{i^\star,t} - \mu_{i,t}) \leq 2U(t,\delta)$$
$$\text{and thus we have that if } \Delta_i \geq 4U(t,\delta) \text{ ,then} (\mu_{i^\star,t} - \mu_{i,t}) \geq 2U(t,\delta)$$

$$\text{Therefore under } \neg\mathcal{E} \text{we have that if } \Delta_i \geq 4U(t,\delta) \implies \text{arm i is eliminated}$$

Let $f(t) = 4\sqrt{\frac{ln(\frac{t\pi}{\sqrt{3\delta}})}{t}}$ this is a strictly decreasing function for $t \geq 1$ and thus it's inverse $f^{-1}(y)$ is well defined. After $t_i = f^{-1}(\Delta_i)$ steps we are almost sure to eliminate arm "i".

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

$f^{-1\prime}(y) = \frac{1}{f'(f^{-1}(y))}$ since f is increasing for $t \geq 1$ it's derivative is negative on the domain and so is the derivative of it's inverse. $\implies f^{-1\prime}$ is strictly decreasing.

$$\implies t_{max} = \max_i f^{-1}(\Delta_i) = f^{-1}(\Delta_{min}) \text{ with confidence } 1-\delta$$

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.

- Let $n = 10000$ and $k = 2$. Plot the expected regret of each algorithm as a function of Δ when $\mu_1 = 1/2$ and $\mu_2 = 1/2 + \Delta$.

Please find the code in the following jupyter notebook.
We report on the expected regret as a function of Δ we chose to take n=1000 as the algorithm takes much time to run.

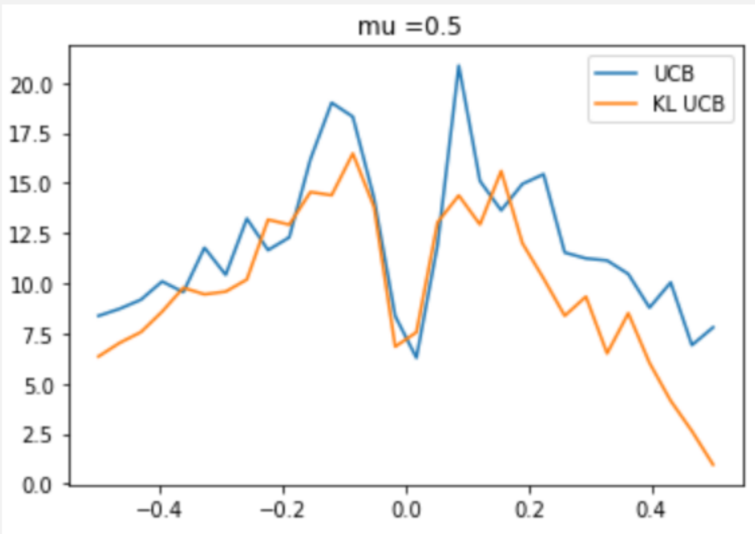
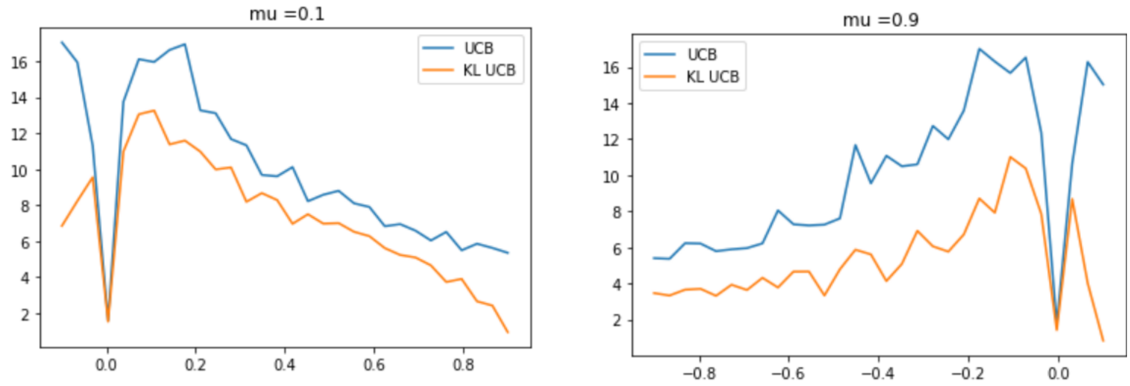


Figure 1: Regret as function of $\Delta, \mu = 0.5$

- Repeat the above experiment with $\mu_1 = 1/10$ and $\mu_1 = 9/10$.



(a) (b)

Figure 2: regret as function of $\Delta, \mu = 0.1/0.9$

- Discuss your results.

We see that the KL-UCB algorithm has lower expected regret than the standard UCB for almost all cases. The regret gap between these two becomes negligeable when Δ itself becomes negligeable which is normal. For small but non negligeable Δ in the range of 0.05 to 0.15 KL-UCB the gap between KL-UCB and UCB is remarkable.
We also notice a strong tendency of KL-UCB to increase the regret gap for when both arms have high values . For example for the same $\Delta = 0.4$ for $\mu_1 = 0.1 and \mu_2 = 0.5$ we notice a gap of around 4. Whereas for the same $\Delta = 0.4$ for $\mu_1 = 0.1 and \mu_2 = 0.5$ we notice a gap of around 1.

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

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

Let's denote by $R_k = \{\exists h, s, a : |r_{hk}(s,a) - r_h(s,a)| \geq \beta^r_{hk}(s,a)\}$
and $\mathrm{P}_k = \{\exists h, s, a : \|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \geq \beta^p_{hk}(s,a)\}$

$P(\neg\mathcal{E}) = P(\bigcup_{k=1}^{\infty} R_k \cup P_k) \leq \sum_{k=1}^{\infty} P(R_k) + P(P_k) \leq \frac{\delta}{2}$

$P(R_k) = P(\cup_{h=0}^{H} \cup_s \cup_a \{|r_{hk}(s,a) - r_h(s,a)| \geq \beta^r_{hk}(s,a)\}$
First we want that $P(P_k) \leq \frac{\delta}{4}$
$P(P_k) \leq \frac{\delta}{4}$
$\mathrm{P}(\mathrm{P}_k) \leq \sum_h \sum_s \sum_a P(\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \geq \beta^p_{hk}(s,a)) \leq \frac{\delta}{4}$
If we choose $\beta^p_{hk}(s,a)$ s.t $P(\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \geq \beta^p_{hk}(s,a)) \leq \frac{\delta}{4HSA}$ then we ensure the above property.
From Weissman ineq we get :
$P(\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \geq \beta^p_{hk}(s,a))) \leq (2^S - 2)\exp\left(-\frac{N_{h,k}(s,a)\beta^p_{hk}(s,a)^2}{2}\right)$
Same as exercice 2, we choose $\beta^p_{hk}(s,a)$ s.t $\lim_{h=\infty} = \beta^p_{hk}(s,a) = 0$ and
$(2^S - 2)\exp\left(-\frac{N_{h,k}(s,a)\beta^p_{hk}(s,a)^2}{2}\right) \sim \frac{\alpha}{N_{h,k}(s,a)^2}$
After developping this expression developping and choosing α s.t $\sum \frac{\alpha}{t^2} = \frac{\delta}{4HSA}$ We find
$\beta^p_{hk}(s,a) = \sqrt{\frac{2\ln(2\pi^2 HSANh,k(s,a)^2(2^S-2)/3\delta)}{N_{h,k}(s,a)}}$
Second want that $P(R_k) \leq \frac{\delta}{4}$
$\mathrm{P}(\mathrm{R}_k) \leq \sum_h \sum_s \sum_a P(|r_{hk}(s,a) - r_h(s,a)| \geq \beta^r_{hk}(s,a)) \leq \frac{\delta}{4}$
We choose $\beta^r_{hk}(s,a)$ s.t $P(|r_{hk}(s,a) - r_h(s,a)| \geq \beta^r_{hk}(s,a) \leq \frac{\delta}{4HSA} = \delta'$
Following the same reasoning as exercice 2 by replacing δ with $\frac{\delta}{4HSA}$:
$\beta_{hk}r(s,a) = \sqrt{\frac{ln(4\pi^2 HSAN_{h,k}(s,a)^2/3\delta)}{2N_{h,k}(s,a)}}$

- 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.

$$\hat{Q}_{h,k}(s,a) = \max_{\beta^r_{hk}(s,a)} r_h(s,a) + \max_{P \in \beta^p_{hk}(s,a)} P\hat{V}_{h+1}$$

$$\max_{\beta^r_{hk}(s,a)} r_h(s,a) \leq \hat{r}(s,a) + \beta^r_{hk}(s,a)$$

Using Holder inequality: $\max_{P \in \beta^p_{hk}(s,a)} P\hat{V}_{h+1} \leq \hat{P}_h\hat{V}_{h+1} + ||P - \hat{P}_h||_1||\hat{V}_{h+1}||_\infty$

$$\leq \hat{P}_h\hat{V}_{h+1} + (H-h)\beta^p_{hk}(s,a)$$

Summing up we get that : $b_{h,k}(s,a) = \beta^r_{hk}(s,a) + (H-h)\beta^p_{hk}(s,a)$

By induction let's prove that $Q_{h,k}(s,a) \geq Q^\star_{h,k}(s,a) \forall h$

For h = H we have :

$Q_{H,k}(s,a) = r_H(s,a) + \beta^r_{hk}(s,a) \geq r^\star_H(s,a) = Q^\star_{H,k}(s,a)$ *Since the real reward falls in the confidence interval*

Suppose that for h $\leq H, Q^\star_{h,k}(s,a) \leq Q_{h,k}(s,a)$ and lets prove this property for h-1.

$$Q^\star_{h,k}(s,a) = r_h(s,a) + P_{h,k}V^{\pi\star}_{h+1,k}$$

$$V^{\pi\star}_{h+1,k}(s) = \max_a Q^{\pi\star}_{h+1,k}(s,a) \leq \max_a Q_{h+1,k}(s,a) = V_{h+1,k}(s)$$

$$V^{\pi\star}_{h+1,k} \leq V_{h+1,k}$$

$$\implies P_{h,k}V^{\pi\star}_{h+1,k} \leq P_{h,k}V_{h+1,k} \leq \max_p PV_{h+1,k} = P_{h+1,k}V_{h+1,k}$$

$$\implies Q^\star_{h,k}(s,a) = r_h(s,a) + P_{h,k}V^{\pi\star}_{h+1,k} \leq r_h(s,a) + P_{h+1,k}V_{h+1,k} = Q_{h,k}(s,a)$$

- In class we have seen that

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

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

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

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

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

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

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

$$= V^{\pi^k}_h(s_{hk})$$

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

$$\begin{aligned} V_{h,k}(s_{hk}) &= min\{H-h, \max_a Q_{h,k}(s,a)\} \\ &\leq \max_a Q_{h,k}(s,a) = Q_{h,k}(s_{hk}, a_{hk}) \text{ (because we take the greedy action)} \end{aligned}$$

3. Putting everything together prove Eq. 1.

$$\begin{aligned} &\delta_{1k}(s_{1k}) = V_{1k} - V_{1k}^{\pi^k}(s_{1k}) \\ &= V_{1k} - r(s_{1k}, a_{1k}) - \mathbb{E}_p[V_{2,k}(s')] + \delta_{2,k}(s_{2,k}) + m_{1,k} \\ &\textit{By extending over } \delta_{h,k}(s_{h,k}) \textit{ for } 2 \leq h \leq H: \\ &\leq \delta_{H+1,k}(s_{H+1,k}) + \sum_{h=1}^{H} V_{hk}(s_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + m_{hk} \\ &\delta_{H+1,k}(s_{H+1,k}) = V_{H+1,k}(s_{H+1,k}) - V_{H+1,k}^{\pi^k}(s_{H+1,k}) = 0 - 0 \\ &V_{hk}(s_{hk}) \leq Q_{hk}(s_{hk}, a_{hk}) \\ &\implies \delta_{hk}(s_{1,k}) \leq \sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + m_{hk} \end{aligned}$$

- 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

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

$$\begin{aligned} R(T) &= \sum_{k=1}^{K} \delta_{1k}(s_{1k}) \\ &\leq \sum_{k=1}^{K}\sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y\sim p(\cdot|s_{hk},a_{hk})}[V_{h+1,k}(Y)]) + \sum_{k,h} m_{hk} \\ &\leq \sum_{k,h} r(s_{hk}, a_{hk}) + b_{hk}(s_{hk}, a_{hk}) + (\hat{P} - P^{\mathbf{true}})V_{h+1,k} - r(s_{hk}, a_{hk}) + \dots \\ &\leq \sum_{k,h} b_{hk}(s_{hk}, a_{hk}) + b_{hk}^{p}(H-h) + \dots \\ &\leq \sum_{k,h}^{K} 2 * b_{hk}(s_{hk}, a_{hk}) + \sum_{k,h}^{K} m_{hk} \\ &\leq 2\sum_{k,h} b_{hk}(s_{hk}, a_{hk}) + 2H\sqrt{KH \log(2/\delta)} \end{aligned}$$

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

- Finally, we have that

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

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

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

References

[1] Why Adaptively Collected Data Have Negative Bias and How to Correct for It *https://arxiv.org/abs/1708.01977*