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

Exploration in Reinforcement Learning (theory)

Lecturers: *A. Lazaric, M. Pirotta* *(January 10, 2021)*

Solution by Clément Bonnet

1 UCB

We find ourselves in the setting of multi-arm bandits.

$$S_{j,t} = \sum_{k=1}^{t} X_{i_k,k} \cdot \mathbb{1}(i_k = j)$$
$$N_{j,t} = \sum_{k=1}^{t} \mathbb{1}(i_k = j)$$
$$\widehat{\mu}_{j,t} = \frac{S_{j,t}}{N_{j,t}}$$

The question is to prove whether or not $\widehat{\mu}_{j,t}$ is an unbiased estimator of μ_j.
At first sight, one could interpret $\widehat{\mu}_{j,t}$ as the simple mean estimate of μ_j and thus would be unbiased. However, this would only apply if samples $X_{i_k,k}$ were independent and identically distributed (iid), which is not the case here in the online on-policy learning of UCB. Whether an arm is pulled or not depends on previous samples and therefore one can expect the estimate to rather have some bias.
To prove the biasedness of $\widehat{\mu}_{j,t}$, or rather to show that it is not unbiased in the general case, we will consider a simple case and compute its analytical bias. Let us consider the setting of Bernoulli bandits as in section 3 with $k = 2$ binary arms of parameters μ_1 and μ_2. One pulls the arm i_t such that

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

We assume here that arms are pulled **randomly** in case of a tie. The UCB exploration term is infinite for $t \in \{1, 2\}$ where both arms are pulled successively. At $t = 3$, both arms have been pulled once and one of them is going to be pulled again. We look at the sample mean estimates $\widehat{\mu}_{1,3}$ and $\widehat{\mu}_{2,3}$ after the third action.

$$\begin{aligned}
\mathbb{P}\left(\widehat{\mu}_{1,3} = \frac{1}{2}\right) &= (1-\mu_1)(1-\mu_2)\frac{\mu_1}{2} + \mu_1(1-\mu_1)(1-\mu_2) + \mu_1\mu_2\frac{1-\mu_1}{2} \\
&= \mu_1\left(\frac{3}{2} - \frac{3}{2}\mu_1 - \mu_2 + \mu_1\mu_2\right) \\
\mathbb{P}\left(\widehat{\mu}_{1,3} = 1\right) &= \mu_1^2(1-\mu 2) + \mu_1\mu_2(1 - \frac{1+\mu_1}{2}) \\
&= \mu_1\left(\mu_1 + \frac{1}{2}\mu_2 - \frac{1}{2}\mu_1\mu_2\right)
\end{aligned}$$

This leads to the calculation of the expected value of $\widehat{\mu}_{1,3}$,

$$\begin{aligned}\mathbb{E}_{UCB}\left[\widehat{\mu}_{1,3}\right] &= \frac{1}{2}\mathbb{P}\left(\widehat{\mu}_{1,3} = \frac{1}{2}\right) + \mathbb{P}\left(\widehat{\mu}_{1,3} = 1\right) \\ &= \mu_1\left(1 - \frac{1}{4}\left(1 - \mu_1\right)\right)\end{aligned}$$

The bias of arm 1 is therefore:

$$\boxed{\text{bias}_1 \equiv \mathbb{E}_{UCB}\left[\widehat{\mu}_{1,3}\right] - \mu_1 = -\frac{1}{4}\mu_1(1-\mu_1)}$$

Since the arms play a symmetrical role in the derivation of the bias, one can derive the bias for arm 2:

$$\boxed{\text{bias}_2 \equiv \mathbb{E}_{UCB}\left[\widehat{\mu}_{2,3}\right] - \mu_2 = -\frac{1}{4}\mu_2(1-\mu_2)}$$

These biases are strictly negative if $0 < \mu_1, \mu_2 < 1$. Therefore, $\boxed{\widehat{\mu}_{j,t} \text{ is not an unbiased estimator of } \mu_j \text{ in general}}$.

2 Best Arm Identification

- Let us compute a function $U(t,\delta)$ that satisfies the any-time confidence bound. For any arm $i \in [k]$

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

If one chooses $\boxed{U(t,\delta) = \sqrt{\frac{1}{2t}\log\frac{\pi^2 t^2}{3\delta}}}$,

$$\begin{aligned}\mathbb{P}\left(\bigcup_{t=1}^{\infty}\left\{|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta)\right\}\right) &\leq \sum_{t=1}^{\infty}\mathbb{P}\left(\left\{|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta)\right\}\right) \\ &\leq \sum_{t=1}^{\infty} 2\exp(-2tU(t,\delta)^2) \quad \text{(Hoeffding's inequality)} \\ &= \sum_{t=1}^{\infty} 2\exp\left(-\log\frac{\pi^2 t^2}{3\delta}\right) \\ &= \sum_{t=1}^{\infty}\frac{6\delta}{\pi^2}\frac{1}{t^2} \\ &= \delta\end{aligned}$$

- Let $\mathcal{E} = \bigcup_{i=1}^{k}\bigcup_{t=1}^{\infty}\left\{|\widehat{\mu}_{i,t} - \mu_i| > U(t,\delta')\right\}$. For $\boxed{\delta' = \frac{\delta}{k}}$,

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

Therefore, $\mathbb{P}\left(\mathcal{E}\right) \leq \delta$. This is a bad event since the confidence intervals do not hold.

- Let us show that with probability at least $1-\delta$, the optimal arm $i^\star = \arg\max_i\{\mu_i\}$ remains in the active set S.

  Let us assume $\neg\mathcal{E}$. Under such conditions,

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

  Therefore,

  $$\forall t, \forall i \neq i^\star, \begin{cases} \widehat{\mu}_{i^\star,t} \geq \mu^\star - U(t,\delta') \\ \widehat{\mu}_{i,t} \leq \mu_i + U(t,\delta') \end{cases}$$

  $$\forall t, \forall i \neq i^\star, \quad \widehat{\mu}_{i^\star,t} - \widehat{\mu}_{i,t} \geq \Delta_i - 2U(t,\delta') \tag{1}$$

  Let us now show that this implies that in such conditions, arm $i^\star$ remains in the active set S. Under such conditions, let us assume the opposite and prove by contradiction. Assume the arm $i^\star$ is eliminated at time t_0. Using δ' instead of δ in the algorithm, this means:

  $$\exists i_0 \neq i^\star, \widehat{\mu}_{i^\star,t_0} \leq \widehat{\mu}_{i_0,t_0} - 2U(t_0,\delta') \tag{2}$$

  Using equation (1) for $t = t_0$ and $i = i_0$ combined with equation (2), one finds the following inequality:

  $$\Delta_{i_0} - 2U(t_0,\delta') \leq \widehat{\mu}_{i^\star,t_0} - \widehat{\mu}_{i_0,t_0} \leq -2U(t_0,\delta')$$

  This implies $\Delta_{i_0} \leq 0$ which is a contradiction since $\Delta_{i_0} = \mu^\star - \mu_{i_0} > 0$ (we assume for simplicity there is only one best arm). Therefore, by contradiction, we prove that under $\neg\mathcal{E}$ conditions, the arm $i^\star$ remains in the active set S.

  $$\neg\mathcal{E} \subset \{\text{arm } i^\star \text{ remains in the active set}\}$$

  $$\mathbb{P}(\neg\mathcal{E}) \leq \mathbb{P}(\{\text{arm } i^\star \text{ remains in the active set}\})$$

  $$\boxed{\mathbb{P}(\{\text{arm } i^\star \text{ remains in the active set}\}) \geq 1 - \mathbb{P}(\mathcal{E}) \geq 1 - \delta}$$

- Under event $\neg\mathcal{E}$, let us find C_1 such that for an arm $i \neq i^\star$, if $\Delta_i \geq C_1 U(t,\delta')$, then the arm i will be removed from the active set.

  Let $i \neq i^\star$ and apply $\neg\mathcal{E}$ conditions on i and $i^\star$.

  $$\begin{cases} |\widehat{\mu}_{i^\star,t} - \mu^\star| \leq U(t,\delta') \\ |\widehat{\mu}_{i,t} - \mu_i| \leq U(t,\delta') \end{cases}$$

  $$\begin{cases} -U(t,\delta') + \mu^\star \leq \widehat{\mu}_{i^\star,t} \leq \mu^\star + U(t,\delta') \\ -U(t,\delta') - \mu_i \leq -\widehat{\mu}_{i,t} \leq -\mu_i + U(t,\delta') \end{cases}$$

  $$\neg\mathcal{E} \implies \Delta_i - 2U(t,\delta') \leq \widehat{\mu}_{i^\star,t} - \widehat{\mu}_{i,t} \leq \Delta_i + 2U(t,\delta')$$

  According to the algorithm (using δ' and not δ in the pseudo-code), if $\widehat{\mu}_{i^\star,t} - \widehat{\mu}_{i,t} \geq 2U(t,\delta')$, the arm i will be removed from the active set.

  Therefore, if $\Delta_i - 2U(t,\delta') \geq 2U(t,\delta') \iff \Delta_i \geq 4U(t,\delta')$, the arm i will be removed for sure from the active set, under $\neg\mathcal{E}$ conditions.

  Under event $\neg\mathcal{E}$, an arm $i \neq i^\star$ will be removed from the active set when $\boxed{\Delta_i \geq C_1 U(t,\delta') \text{ with } C_1 = 4}$.

  With our definition of $U(t,\delta')$,

  $$\Delta_i \geq 4U(t,\delta') \iff \Delta_i^2 \geq \frac{8}{t}\left(2\log t + \log\frac{\pi^2}{3\delta}\right)$$

  By minimizing $\log t$ by 0 (since $t \geq 1$), for every arm $i \neq i^\star$,

  $$\boxed{t \geq \frac{8\log\frac{\pi^2}{3\delta}}{\Delta_i^2}} \implies \Delta_i \geq 4U(t,\delta') \implies \text{arm } i \text{ will be removed}$$

- Let us compute a lower bound on the sample complexity for identifying the optimal arm with probability $1-\delta$.

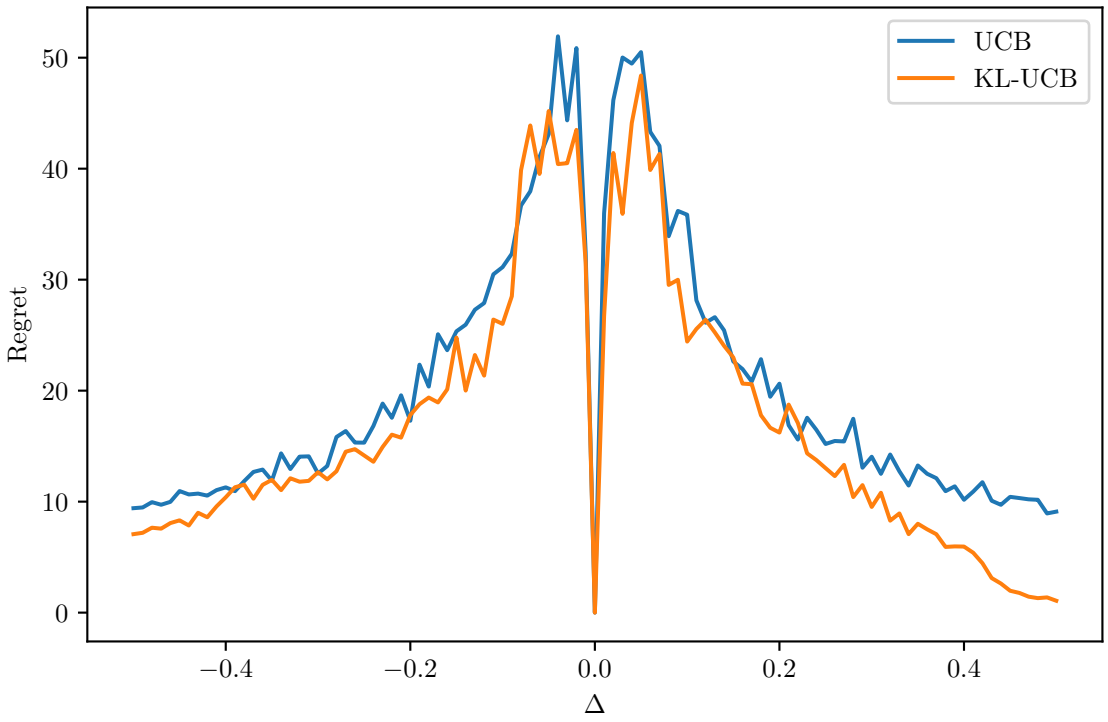
$$\left(\forall i \neq i^\star,\ t \geq \frac{8\log\frac{\pi^2}{3\delta}}{\Delta_i^2}\right) \iff t \geq \frac{8\log\frac{\pi^2}{3\delta}}{\Delta_{i^\star}^2}$$

With $\Delta_{i^\star} = \min_{i \neq i^\star} \Delta_i$.

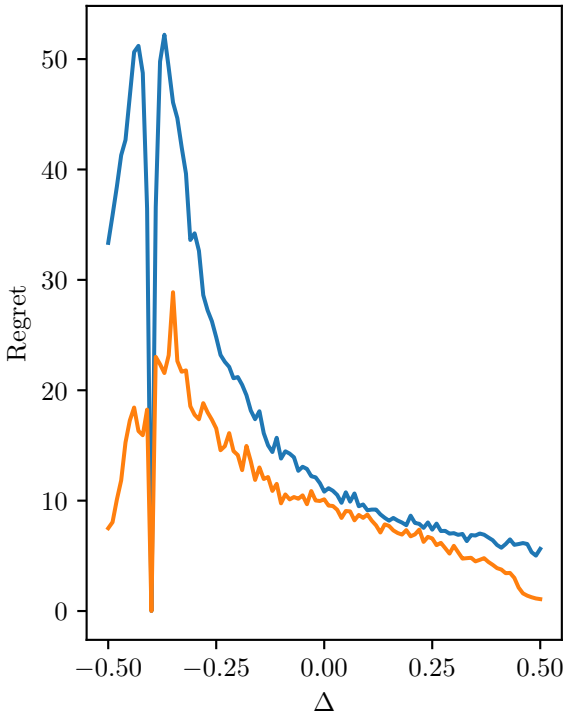
$$\boxed{\tau_\delta \geq \frac{8\log\frac{\pi^2}{3\delta}}{\Delta_{i^\star}^2}} \quad \text{with probability } 1-\delta.$$

3 Bernoulli Bandits

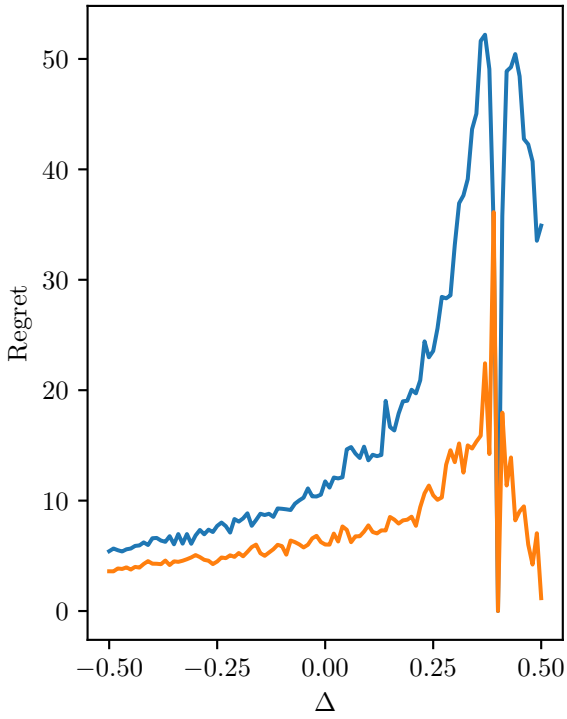
UCB and KL-UCB algorithms for Bernoulli Bandits have been implemented in Python, using NumPy and Matplotlib libraries. Expected regret of both algorithms are plotted in figure 1 in the case of $k = 2$ Bernoulli arms of means $\mu_1 \in \{0.1, 0.5, 0.9\}$ and $\mu_2 = 0.5 + \Delta$ with $\Delta \in [-0.5, 0.5]$.
First, one must observe that both algorithms always have a regret of 0 when $\mu_1 = \mu_2$ (corresponding to $\Delta = 0$ in (a), $\Delta = -0.4$ in (b) and $\Delta = 0.4$ in (c)). Indeed this is explained by both arms having the same expected value and thus they are equally good in average. This means that whatever choice one makes, there is no regret from it.
Then, both algorithms perform the worst when $\mu_1 \approx \mu_2$ but $\mu_1 \neq \mu_2$. This is straightforward to understand since when their expected values are very close to each other, it is harder or at least it takes longer to distinguish them. Therefore, one makes many more errors in choosing the wrong arm, increasing the regret.
Finally, one can see in figure 1 that the KL-UCB algorithm performs better than the UCB one. Although for $\mu_1 = 0.5$, KL-UCB and UCB tend to have rather similar performances when μ_2 remains close to μ_1, KL-UCB performs significantly better when $\mu_1 \in \{0.1, 0.9\}$. The Kullback-Leibler divergence is indeed better at distinguishing between two Bernoulli distributions that would have their means close to 0 or close to 1. This is why KL-UCB performs better than UCB in figure 1 (b) and (c) when $\mu_1 \approx \mu_2$ whereas it performs closely to UCB in (a) since the KL divergence is smaller when both means are around 0.5.



(a) $\mu_1 = 0.5$



(b) $\mu_1 = 0.1$



(c) $\mu_1 = 0.9$

Figure 1: Expected regret after $n = 10000$ steps for Bernoulli bandits with $k = 2$ arms and means $\mu_1 = 0.5$ in (a), 0.1 in (b) and 0.9 in (c), and $\mu_2 = 0.5 + \Delta$ with $\Delta \in [-0.5, 0.5]$. The plots were averaged over 50 runs for each Δ.

4 Regret Minimization in RL

We consider a finite-horizon MDP $M^\star = (S, A, p_h, r_h)$ with stage-dependent transitions and rewards.

- We define the event $\mathcal{E} = \{\forall k, M^\star \in \mathcal{M}_k\}$ and $\mathcal{M}_k = \{M = (S, A, p_{h,k}, r_{h,k}) \ : \ r_{h,k}(s,a) \in \mathcal{B}^r_{h,k}(s,a), p_{h,k}(\cdot|s,a) \in \mathcal{B}^p_{h,k}(s,a)\}$. Let us define confidence intervals $\beta^r_{hk}(s,a)$ and $\beta^p_{hk}(s,a)$ as a function of δ such that $\mathbb{P}(\neg\mathcal{E}) \leq \delta/2$.

  Let us choose:

$$\boxed{\beta^r_{hk}(s,a) = \sqrt{\frac{\log\left(\frac{8HSAK}{\delta}\right)}{2N_{h,k}(s,a)}}} \quad \text{and} \quad \boxed{\beta^p_{hk}(s,a) = \sqrt{\frac{2\log\left(\frac{4HSAK(2^S-2)}{\delta}\right)}{N_{h,k}(s,a)}}}$$

$$\begin{aligned}
\mathbb{P}(\neg\mathcal{E}) &= \mathbb{P}\left(\bigcup_{k=1}^{K}\{M^\star \notin \mathcal{M}_k\}\right) \\
&= \mathbb{P}\left(\bigcup_{k=1}^{K}\bigcup_{h=1}^{H}\bigcup_{s=1}^{S}\bigcup_{a=1}^{A}\{|\widehat{r}_{hk}(s,a) - r_h(s,a)| \geq \beta^r_{hk}(s,a)\} \cup \{\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \geq \beta^p_{hk}(s,a)\}\right) \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{S}\sum_{a=1}^{A}\mathbb{P}\{|\widehat{r}_{hk}(s,a) - r_h(s,a)| \geq \beta^r_{hk}(s,a)\} + \mathbb{P}\{\|\widehat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 \geq \beta^p_{hk}(s,a)\} \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{S}\sum_{a=1}^{A}\left[2\exp\left(-2N_{h,k}(s,a)\beta^r_{hk}(s,a)^2\right) + (2^S-2)\exp\left(-\frac{N_{h,k}(s,a)\beta^p_{hk}(s,a)^2}{2}\right)\right] \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{S}\sum_{a=1}^{A}\left[\frac{\delta}{4HSAK} + \frac{\delta}{4HSAK}\right] \\
&= \frac{\delta}{2}
\end{aligned}$$

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

- Let us be under the event $\mathcal{E}$ and let $b_{h,k}(s,a)$ be a bonus to define.

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

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

  Induction step

  Let $h \in [1, H-1]$ and let us assume the following: $\forall s, a, k, \ Q_{h+1,k}(s,a) \geq Q^\star_{h+1}(s,a)$ (inductive assumption).

Let us show that: $\forall s, a, k,\ Q_{h,k}(s,a) \geq Q_h^\star(s,a)$.

$$
\begin{aligned}
Q_{h,k}(s,a) - Q_h^\star(s,a) &= \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) + \sum_{s'} \widehat{p}_{h,k}(s'|s,a) V_{h+1,k}(s') - \left(r_h(s,a) + \sum_{s'} p_h(s'|s,a) V_{h+1}^\star(s') \right) \\
&= \sum_{s'} \left(\widehat{p}_{h,k}(s'|s,a) \min\left\{H, \max_{a'} Q_{h+1,k}(s',a')\right\} - p_h(s'|s,a) \max_{a'} Q_{h+1}^\star(s',a') \right) \\
&\quad + \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) - r_h(s,a) \\
&\geq \sum_{s'} \left(\widehat{p}_{h,k}(s'|s,a) \min\left\{H, \max_{a'} Q_{h+1,k}(s',a')\right\} - p_h(s'|s,a) \min\left\{H, \max_{a'} Q_{h+1,k}(s',a')\right\} \right) \\
&\quad + \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) - r_h(s,a) \\
&= \sum_{s'} \min\left\{H, \max_{a'} Q_{h+1,k}(s',a')\right\} \left(\widehat{p}_{h,k}(s'|s,a) - p_h(s'|s,a) \right) \\
&\quad + \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) - r_h(s,a) \\
&\geq - \sum_{s'} \min\left\{H, \max_{a'} Q_{h+1,k}(s',a')\right\} \left| \widehat{p}_{h,k}(s'|s,a) - p_h(s'|s,a) \right| \\
&\quad + \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) - r_h(s,a) \\
&\geq -H \sum_{s'} \left| \widehat{p}_{h,k}(s'|s,a) - p_h(s'|s,a) \right| + \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) - r_h(s,a) \\
&= -H \left\| \widehat{p}_{h,k}(s'|s,a) - p_h(s'|s,a) \right\|_1 + \widehat{r}_{h,k}(s,a) + b_{h,k}(s,a) - r_h(s,a) \\
&\geq -H\beta_{h,k}^p(s,a) + b_{h,k}(s,a) - \beta_{h,k}^r(s,a) + \underbrace{\widehat{r}_{h,k}(s,a) + \beta_{h,k}^r(s,a) - r_h(s,a)}_{\geq 0} \\
&\geq b_{h,k}(s,a) - \beta_{h,k}^r(s,a) - H\beta_{h,k}^p(s,a)
\end{aligned}
$$

Indeed, the induction step works if $b_{h,k}(s,a)$ is chosen such that

$$b_{h,k}(s,a) \geq \beta_{h,k}^r(s,a) + H\beta_{h,k}^p(s,a)$$

Let us define $b_{h,k}(s,a)$ to ensure $Q_{h,k}$ is optimistic.

$$\boxed{b_{h,k}(s,a) = \beta_{h,k}^r(s,a) + H\beta_{h,k}^p(s,a)}$$

With this choice of $b_{h,k}(s,a)$,

$$Q_{h,k}(s,a) - Q_h^\star(s,a) \geq 0$$

The induction step is now proved.

Base case

Since we are under the event $\mathcal{E}$, we have:

$$\widehat{r}_{H,k}(s,a) + b_{Hk}(s,a) \geq \widehat{r}_{H,k}(s,a) + \beta_{Hk}^r(s,a) \geq r_H(s,a).$$

Then, $\forall s',\ V_{H+1,k}(s') = V_{H+1}^\star(s') = 0$. Therefore, $\forall s, a, k,\ Q_{H,k}(s,a) - Q_H^\star(s,a)$. The base case is proven.

Combining the base case and the inductive step gives us:

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

- The aim in this question is to prove the following:

$$\delta_{1,k}(s_{1,k}) \leq \sum_{h=1}^{H} Q_{h,k}(s_{h,k}, a_{h,k}) - r(s_{h,k}, a_{h,k}) - \mathbb{E}_{Y \sim p(\cdot|s_{h,k}, a_{h,k})}[V_{h+1,k}(Y)]) + m_{h,k} \tag{3}$$

Where $\delta_{h,k}(s) = V_{h,k}(s) - V_h^{\pi_k}(s)$ and $m_{h,k} = \mathbb{E}_{Y\sim p(\cdot|s_{h,k},a_{h,k})}[\delta_{h+1,k}(Y)] - \delta_{h+1,k}(s_{h+1,k})$.

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

$$
\begin{aligned}
&r(s_{h,k},a_{h,k}) + \mathbb{E}_p[V_{h+1,k}(s')] - \delta_{h+1,k}(s_{h+1,k}) - m_{h,k} \\
&= r(s_{h,k},a_{h,k}) + \mathbb{E}_p[V_{h+1,k}(s')] - \delta_{h+1,k}(s_{h+1,k}) - \big(\mathbb{E}_{Y\sim p(\cdot|s_{h,k},a_{h,k})}[\delta_{h+1,k}(Y)] - \delta_{h+1,k}(s_{h+1,k})\big) \\
&= r(s_{h,k},a_{h,k}) + \mathbb{E}_p[V_{h+1,k}(s')] - \mathbb{E}_{Y\sim p(\cdot|s_{h,k},a_{h,k})}[\delta_{h+1,k}(Y)] \\
&= r(s_{h,k},a_{h,k}) + \mathbb{E}_p[V_{h+1,k}(s')] - \mathbb{E}_{Y\sim p(\cdot|s_{h,k},a_{h,k})}[V_{h+1,k}(Y) - V_{h+1}^{\pi_k}(Y)] \\
&= r(s_{h,k},a_{h,k}) + \mathbb{E}_p[V_{h+1,k}(s')] + \mathbb{E}_p[V_{h+1}^{\pi_k}(Y)] - \mathbb{E}_{Y\sim p(\cdot|s_{h,k},a_{h,k})}[V_{h+1,k}(Y)] \\
&= r(s_{h,k},a_{h,k}) + \mathbb{E}_p[V_{h+1}^{\pi_k}(Y)] \\
&= V_h^{\pi_k}(s_{h,k}) \qquad \text{(Bellman equation)}
\end{aligned}
$$

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

2. Let us prove that $V_{h,k}(s_{h,k}) \leq Q_{h,k}(s_{h,k},a_{h,k})$.

$$
\begin{aligned}
V_{h,k}(s_{h,k}) &= \min\{H, \max_{a'} Q_{h,k}(s_{h,k},a')\} \\
&\leq \max_{a'} Q_{h,k}(s_{h,k},a') \\
&\leq Q_{h,k}(s_{h,k},a_{h,k})
\end{aligned}
$$

Therefore, $\boxed{V_{h,k}(s_{h,k}) \leq Q_{h,k}(s_{h,k},a_{h,k})}$.

3. Let us prove equation 3.

$$
\begin{aligned}
\delta_{1,k}(s_{1,k}) &= V_{1,k}(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) \\
&\leq Q_{1,k}(s_{1,k},a_{1,k}) - (r(s_{1,k},a_{1,k}) + \mathbb{E}_p[V_{2,k}(s')] - \delta_{2,k}(s_{2,k}) - m_{1,k}) \\
&= \delta_{2,k}(s_{2,k}) + \big[Q_{1,k}(s_{1,k},a_{1,k}) - r(s_{1,k},a_{1,k}) - \mathbb{E}_p[V_{2,k}(s')] - m_{1,k}\big] \\
&= V_{2,k}(s_{2,k}) - V_2^{\pi_k}(s_{2,k}) + \big[Q_{1,k}(s_{1,k},a_{1,k}) - r(s_{1,k},a_{1,k}) - \mathbb{E}_p[V_{2,k}(s')] - m_{1,k}\big] \\
&\leq \dots \\
&\leq \sum_{h=1}^{H} Q_{h,k}(s_{h,k},a_{h,k}) - r(s_{h,k},a_{h,k}) - \mathbb{E}_p[V_{h+1,k}(s')] - m_{h,k}
\end{aligned}
$$

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

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

$$
\begin{aligned}
R(T) &= \sum_{k=1}^{K} V_1^\star(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) \\
&= \sum_{k=1}^{K} V_1^{\pi_k^\star}(s_{1,k}) - V_{1,k}(s_{1,k}) + \sum_{k=1}^{K} V_{1,k}(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) \\
&= \sum_{k=1}^{K} -\delta_{1,k}^\star(s_{1,k}) + \delta_{1,k}(s_{1,k}) \\
&\leq \sum_{k=1}^{K} \delta_{1,k}(s_{1,k}) \quad \text{(Since } V \geq V^\star \geq V^{\pi_k}\text{)} \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} Q_{h,k}(s_{h,k}, a_{h,k}) - r(s_{h,k}, a_{h,k}) - \mathbb{E}_{Y\sim p(\cdot|s_{h,k},a_{h,k})}[V_{h+1,k}(Y)]) + m_{h,k} \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H} m_{h,k} + \sum_{k=1}^{K}\sum_{h=1}^{H} Q_{h,k}(s_{h,k}, a_{h,k}) - r(s_{h,k}, a_{h,k}) - \mathbb{E}_{Y\sim p(\cdot|s_{h,k},a_{h,k})}[V_{h+1,k}(Y)]) \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} m_{h,k} + \sum_{k=1}^{K}\sum_{h=1}^{H} Q_{h,k}(s_{h,k}, a_{h,k}) - r(s_{h,k}, a_{h,k}) - \mathbb{E}_{Y\sim p(\cdot|s_{h,k},a_{h,k})}[V_{h+1}^\star(Y)]) \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H} m_{h,k} + \sum_{k=1}^{K}\sum_{h=1}^{H} Q_{h,k}(s_{h,k}, a_{h,k}) - Q_{h,k}^\star(s_{h,k}, a_{h,k})
\end{aligned}
$$

  The first sum is bounded by Azuma with probability $1-\frac{\delta}{2}$ whereas the second one is bounded by the bonuses again with probability $1-\frac{\delta}{2}$. Therefore, with probability $1-\delta$, we have:

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

- Finally, let us show that $R(T) \lesssim H^2 S\sqrt{AK}$.

$$
\begin{aligned}
\sum_{h,k} \frac{1}{\sqrt{N_{h,k}(s_{h,k}, a_{h,k})}} &= \sum_{h=1}^{H}\sum_{s,a}\sum_{i=1}^{N_{h,K}(s,a)} \frac{1}{\sqrt{i}} \\
&\leq 2\sum_{h=1}^{H}\sum_{s,a}\sqrt{N_{h,K}(s,a)} \\
&\leq 2\sqrt{SAH}\sqrt{\sum_{h=1}^{H}\sum_{s,a} N_{h,K}(s,a)} \qquad \text{(Jensen)} \\
&\leq 2\sqrt{SAH}\sqrt{\sum_{h=1}^{H} K} \qquad \left(\forall h,\ \sum_{s,a} N_{h,K}(s,a) \leq K\right) \\
&\leq 2H\sqrt{SAK}
\end{aligned}
$$

  Therefore,

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

Let us conclude on the regret.

$$
\begin{aligned}
R(T) &\leq \sum_{k=1}^{K}\sum_{h=1}^{H} b_{h,k}(s_{h,k},a_{h,k}) + 2H\sqrt{KH\log(2/\delta)} \\
&= \left(\sqrt{\frac{\log\left(\frac{8HSAK}{\delta}\right)}{2}} + H\sqrt{2\log\left(\frac{4HSAK\left(2^S-2\right)}{\delta}\right)}\right)\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{N_{h,k}(s_{h,k},a_{h,k})}} + 2H\sqrt{KH\log(2/\delta)} \\
&\leq \left(\sqrt{\frac{\log\left(\frac{8HSAK}{\delta}\right)}{2}} + H\sqrt{2\log\left(\frac{4HSAK\left(2^S-2\right)}{\delta}\right)}\right)2H\sqrt{SAK} + 2H\sqrt{KH\log(2/\delta)} \\
&\leq f\left(H,S,A,K\right)H\sqrt{SAK} + cH\sqrt{KH}
\end{aligned}
$$

Where c is a constant (depends on δ) and $f\left(H,S,A,K\right) \lesssim H\sqrt{S}$. Therefore, we find the regret upper bound:

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