MVA: Reinforcement Learning (2021/2022) **Assignment 3**

Exploration in Reinforcement Learning (theory)

Lecturers: *M. Pirotta* *(December 16, 2021)*

Solution by FILL fullname command at the beginning of latex document

Instructions

- The deadline is **January 16, 2022. 23h59**
- 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 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. Assume that the best arm $i^\star$ is *unique* (i.e., there exists only one arm with maximum mean reward).

Notation

- I_t: the arm chosen at round t.
- $X_{i,t} \in [0,1]$: reward observed for arm i at round t.
- μ_i: the expected reward of arm i.
- $\mu^\star = \max_i \mu_i$.
- $\Delta_i = \mu^\star - \mu_i$: suboptimality gap.

Consider the following algorithm
The algorithm maintains an active set S and an estimate of the empirical reward of each arm $\widehat{\mu}_{i,t} = \frac{1}{t}\sum_{j=1}^{t} X_{i,j}$.

- Compute the function $U(t,\delta)$ that satisfy the any-time confidence bound. Let

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

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

Using Hoeffding's inequality and union bounds, 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 $\Delta_i \geq C_1 U(t, \delta')$ for some constant $C_1 \in \mathbb{N}$. Compute the time required to have such condition for each non-optimal arm. Use the condition of arm elimination applied to arm $i^\star$.[1]
- Compute a bound on the sample complexity (after how many *pulls* the algorithm stops) for identifying the optimal arm w.p. $1-\delta$.
- We assumed that the optimal arm $i^\star$ is unique. Would the algorithm still work if there exist multiple best arms? Why?

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

2 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

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

[1] Note that $at \geq \log(bt)$ can be solved using Lambert W function. We thus have $t \geq \frac{-W_{-1}(-a/b)}{a}$ since, given $a = \Delta_i^2$ and $b = 2k/\delta$, $-a/b \in (-1/e, 0)$. We can make the bound more explicit by noticing that $-1 - \sqrt{2u} - u \leq W_{-1}(-e^{-u-1}) \leq -1 - \sqrt{2u} - 2u/3$ for $u > 0$ [Chatzigeorgiou, 2016]. Then $t \geq \frac{1+\sqrt{2u}+u}{a}$ with $u = \log(b/a) - 1$.

- 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 $\delta_{hk}(s) = V_{hk}(s) - V^{\pi_k}_h(s)$ 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 $V^{\pi_k}_h(s_{hk}) = r(s_{hk}, a_{hk}) + \mathbb{E}_p[V_{h+1,k}(s')] - \delta_{h+1,k}(s_{h+1,k}) - m_{h,k}$
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 [Domingues et al., 2021]

$$\sum_{h,k} \frac{1}{\sqrt{N_{hk}(s_{hk}, a_{hk})}} \lesssim H^2S^2A + 2\sum_{h=1}^{H}\sum_{s,a}\sqrt{N_{hK}(s,a)}$$

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

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

Ioannis Chatzigeorgiou. Bounds on the lambert function and their application to the outage analysis of user cooperation. *CoRR*, abs/1601.04895, 2016.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. Kernel-based reinforcement learning: A finite-time analysis. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 2783–2792. PMLR, 2021.

Initialize $Q_{h1}(s,a) = 0$ for all $(s,a) \in S \times A$ and $h = 1, \ldots, H$

for $k = 1, \ldots, K$ **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 $h = H, \ldots, 1$ **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 $h = 1, \ldots, H$ **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

Reinforcement Learning - Homework 3

Nicolás Vivolante

(1) Best Arm Identification

a) Let $\delta > 0$

First, by Hoeffding's inequality

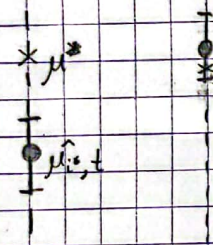
$$\mathbb{P}\left(|\hat{\mu}_{i,t} - \mu_i| > U\right) \leq 2e^{-2tU^2} = \delta \Leftrightarrow U(t,\delta) = \sqrt{\frac{1}{2t}\log\left(\frac{2}{\delta}\right)}$$

Let $\boxed{\delta' = \frac{6}{\pi^2}\frac{\delta}{K}\frac{1}{t^2}}$

$$\mathbb{P}\left(\bigcup_{i=1}^{K}\bigcup_{t=1}^{\infty}\left\{|\hat{\mu}_{i,t} - \mu_i| > U(t,\delta')\right\}\right) \leq \sum_{i=1}^{K}\sum_{t=1}^{\infty}\frac{6}{\pi^2}\frac{\delta}{K}\frac{1}{t^2}$$

$$= \frac{\delta}{\pi^2}\frac{6}{K}\left(\sum_{i=1}^{K}\right)\left(\sum_{t=1}^{\infty}\frac{1}{t^2}\right)$$

$$= \frac{\delta}{\pi^2}\frac{6}{K}\cdot K\cdot\frac{\pi^2}{6} = \delta$$

b) 

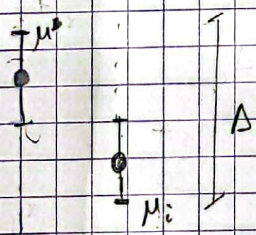
If $\exists j \in S / \left\{\hat{\mu}_{j,t} - U(t,\delta') \geq \hat{\mu}_{i^*,t} + U(t,\delta')\right\}$

then μ^* is outside the confidence interval

$\Rightarrow \mathbb{P}(i^* \text{ removed from } S) \leq \mathbb{P}(\mathcal{E}) \leq \delta$

$\Rightarrow \mathbb{P}(i^* \text{ remains on } S) \geq 1 - \delta$

c) Under $\mathcal{E}^c$, $\Delta_i = \mu^* - \mu_i \geq C_1 U(t,\delta')$

The worst case is 

then we need $C_1 = 4$ at least

Then, if $\Delta_i \geq 4U(t,\delta)$ the arm i will be removed

Analitically, since $\mathcal{E}^c$ holds $\mu^* \leq \hat{\mu}^* + U(t,\delta)$ and $\mu_i \geq \hat{\mu}_i - U(t,\delta)$

The arm elimination condition is

$$\hat{\mu}^* - U(t,\delta') \geq \hat{\mu}_i + U(t,\delta)$$

$$\Leftrightarrow \mu^* - U(t,\delta) - U(t,\delta') \geq \mu_i + U(t,\delta) + U(t,\delta')$$

$$\Leftarrow \mu^* - \mu_i \geq 4U(t,\delta')$$

(continue)

Reinforcement learning - Homework 3 — Nicolás Violante

(1) (c) $\Delta_i \geq 4\sqrt{\frac{1}{2t}\log\left(b^2t^2\right)}$ with $b^2 = \frac{9\pi^2 K}{6\delta} = \frac{\pi^2 K}{3\delta}$

$\Delta_i^2 \geq \frac{16}{2t}\log(b^2t^2) = \frac{16}{t}\log(bt) \Leftrightarrow \left(\frac{\Delta_i^2}{16}\right)\cdot t \geq \log(bt)$

a

Then, using Lambert W $t \geq \frac{1 + \sqrt{2u}}{a} + u$ with $u = \log(b/a) - 1$

$t \geq \frac{1 + \sqrt{2(\log(b/a)-1)}}{a} + \log(b/a) - 1$

$t \geq \frac{16}{\Delta_i^2}\left\{\sqrt{2\log\left(\frac{16}{\Delta_i^2}\pi\sqrt{\frac{K}{3\delta}}\right) - 2} + \log\left(\frac{16}{\Delta_i^2}\pi\sqrt{\frac{K}{3\delta}}\right)\right\}$ (*)

(d) The equation (*) is for the arm i and has complexity

$$\Theta\left(\frac{1}{\Delta_i^2}\log\left(\frac{1}{\Delta_i^2}\sqrt{\frac{K}{\delta}}\right)\right)$$ (simplifying constants and lower-order terms)

If we consider all the pulls we have a sum over arms $i \neq i^*$, so

$$\Theta\left(\sum_{i \neq i^*}\frac{1}{\Delta_i^2}\log\left(\frac{1}{\Delta_i^2}\sqrt{\frac{K}{\delta}}\right)\right)$$ is the sample complexity

(e) The algorithm won't work correctly. Between two optimal arms we have $\Delta = 0$, hence an infinit time for the removal of the arm, as given by equation (*).

Reinforcement Learning - Homework 3 — Nicolas Violante

② Regret Minimization in RL

1) $\mathbb{P}\left(|\hat{r}_{nk}(s,a) - r_n(s,a)| \geq \beta^r_{nk}(s,a)\right) \leq 2e^{-2N_{nk}(s,a)\beta^r_{nk}(s,a)^2} = \delta_1$

$$\Leftrightarrow \beta^r_{nk}(s,a) = \sqrt{\frac{1}{2N_{nk}(s,a)} \log\left(\frac{2}{\delta_1}\right)} \quad (1)$$

$$\mathbb{P}\left(\|\hat{p}_{nk}(\cdot|s,a) - p_n(\cdot|s,a)\|_1 \geq \beta^p_{nk}(s,a)\right) \leq (2^S - 2)\, e^{-\frac{1}{2}N_{nk}(s,a)\beta^p_{nk}(s,a)^2} = \delta_2$$

$$\frac{1}{2}N_{nk}(s,a)\beta^p_{nk}{}^2 = \log\left(\frac{2^S - 2}{\delta_2}\right)$$

$$\Rightarrow \beta^p_{nk} = \sqrt{\frac{2}{N_{nk}(s,a)} \log\left(\frac{2^S - 2}{\delta_2}\right)} \quad (2)$$

- Let h, k, s, a be fixed and let $\delta' > 0$:

$$\mathbb{P}\left(|\hat{r}_{nk}(s,a) - r_n(s,a)| \geq \beta^r_{nk}(s,a) \text{ or } \|\hat{p}_{nk}(s,a) - p_n(\cdot|s,a)\|_1 \geq \beta^p_{nk}(s,a)\right)$$
$$\leq \mathbb{P}\left(|\hat{r}_{nk}(s,a) - r_n(s,a)| \geq \beta^r_{nk}(s,a)\right) + \mathbb{P}\left(\|\hat{p}_{nk}(s,a) - p_n(\cdot|s,a)\|_1 \geq \beta^p_{nk}(s,a)\right)$$
$$\leq \frac{\delta'}{2} + \frac{\delta'}{2} = \delta' \text{ for a proper } \beta^r_{nk}(s,a) \text{ and } \beta^p_{nk}(s,a)$$

$$\mathbb{P}\left(\forall k,h,s,a: |\hat{r}_{nk}(s,a) - r_n(s,a)| \geq \beta^r_{nk}(s,a) \text{ and } \|\hat{p}_{nk}(\cdot|s,a) - p_n(\cdot|s,a)\|_1 \geq \beta^p_{nk}(s,a)\right)$$
$$= 1 - \mathbb{P}\left(\bigcup_{k=1}^{K}\bigcup_{h=1}^{H}\bigcup_{s=1}^{S}\bigcup_{a=1}^{A}\left\{|\hat{r}_{nk}(s,a) - r_n(s,a)| \geq \beta^r_{nk}(s,a) \text{ or } \|\hat{p}_{nk}(\cdot|s,a) - p_n(\cdot|s,a)\|_1 \geq \beta^p_{nk}(s,a)\right\}\right)$$
$$\geq 1 - \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{S}\sum_{a=1}^{A}\mathbb{P}\left(|\hat{r}_{nk}(s,a) - r_n(s,a)| \geq \beta^r_{nk}(s,a) \text{ or } \|\hat{p}_{nk}(\cdot|s,a) - p_n(\cdot|s,a)\|_1 \beta^p_{nk}(s,a)\right)$$

$$= 1 - \underbrace{KHSA\delta'}_{:=\frac{\delta}{2}} \Rightarrow \delta = 2KHSA\delta' \Rightarrow \frac{\delta'}{2} = \frac{\delta}{4KHSA}$$

$$\Rightarrow \beta^r_{nk}(s,a) = \sqrt{\frac{1}{2N_{nk}(s,a)} \log(8KHSA)}$$

$$\beta^p_{nk} = \sqrt{\frac{2}{N_{nk}(s,a)}\left[\log\left(\frac{(2^S-2)\,4KHSA}{\delta}\right)\right]}$$

then $\mathbb{P}(\neg\mathcal{E}) \geq 1 - \frac{\delta}{2}$

Reinforcement Learning - Homework 3 Nicolás Violante.

(2) b) $Q_{nk}(s,a) = \hat{r}_{nk}(s,a) + b_{nk}(s,a) + \sum_{s'} \hat{p}_{nk}(s'|s,a) V_{n+1,k}(s')$ under $\mathcal{E}$

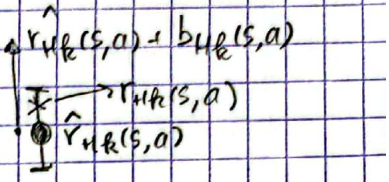
- Base case $h = H$

$$Q_{Hk}(s,a) = \hat{r}_{Hk}(s,a) + b_{Hk}(s,a) + \sum_{s'} \hat{p}_{Hk}(s'|s,a) \underbrace{V_{H+1,k}(s')}_{=0}$$

$$= \hat{r}_{Hk}(s,a) + b_{Hk}(s,a)$$

$$Q^*_H(s,a) = r_{Hk}(s,a)$$

$\Rightarrow Q_{Hk}(s,a) \geq Q^*_H(s,a)$ if $b_{Hk}(s,a) \geq \beta^r_{Hk}(s,a)$ $\Rightarrow$



- Inductive step

Assume $Q_{n+1,k}(s,a) \geq Q^*_{n+1}(s,a)$ $\forall s,a$

$$Q_{nk}(s,a) - Q^*_n(s,a) = \hat{r}_{nk}(s,a) + b_{nk}(s,a) + \sum_{s'} \hat{p}_{nk}(s'|s,a) \min\{H, \max_a Q_{n+1,k}(s,a)\}$$
$$- r_n(s,a) - \sum_{s'} p_{n+1}(s'|s,a) \max_a Q^*_{n+1}(s,a)$$
$$\geq \hat{r}_{nk}(s,a) + b_{nk}(s,a) - r_n(s,a)$$
$$+ \sum_{s'} \min\{H, \max_a Q_{n+1}(s',a)\} \left(\hat{p}_{nk}(s'|s,a) - p_n(s'|s,a)\right)$$
$$\geq \hat{r}_{nk}(s,a) + b_{nk}(s,a) - r_n(s,a)$$
$$- \sum_{s'} \min\{H, \max_a Q_{n+1}(s',a)\} |\hat{p}_{nk}(s'|s,a) - p_n(s'|s,a)|$$
$$\geq \hat{r}_{nk}(s,a) + b_{nk}(s,a) - r_n(s,a) - H \underbrace{\sum_{s'} |\hat{p}_{nk}(s'|s,a) - p_n(s'|s,a)|}_{\|\hat{p}_{nk}(s'|s,a) - p_n(s'|s,a)\|_1}$$
$$\geq \hat{r}_{nk}(s,a) + b_{nk}(s,a) - r_n(s,a) - H\beta^p_{nk}(s,a)$$
$$\geq -\beta^r_{nk}(s,a) + b_{nk}(s,a) - H\beta^p_{nk}(s,a) \geq 0$$
$$\geq -$$

$\Rightarrow b_{nk}(s,a) \geq \beta^r_{nk}(s,a) + H\beta^p_{nk}(s,a)$

Reinforcement Learning – Homework 3 — Nicolás Violante

② 3) 1. $V_h^{\pi_k}(s) = r(s_{hk}, a_{hk}) + \sum_{s'} p_{hk}(s' | s_{hk}, a_{hk}) \underbrace{V_{h+1}^{\pi_k}(s')}_{= -\delta_{h+1,k}(s') + V_{h+1,k}(s')}$

$\Leftrightarrow \sum_{s'} p(s' | s_{hk}, a_{hk}) \left(V_{h+1,k}(s') - \delta_{h+1,k}(s') \right) =$

$= \mathbb{E}_{s' \sim p}\left(V_{h+1,k}(s') \right) - \mathbb{E}_{s' \sim p}\left[\delta_{h+1,k}(s') \right] = \mathbb{E}_{s' \sim p}\left[V_{h+1,k}(s') \right] - \delta_{h+1,k}(s_{h+1,k}) - m_{hk}$

$\Rightarrow V_h^{\pi_k}(s) = r(s_{hk}, a_{hk}) + \mathbb{E}_{s' \sim p}\left(V_{h+1,k}(s') \right) - \delta_{h+1,k}(s_{h+1,k}) - m_{hk}$

2. $V_{h,k}(s_{h,k}) = \min\left\{ H, \max_a Q_{h,k}(s_{h,k}, a) \right\} = \min\left\{ H, Q_{h,k}(s_{hk}, a_{hk}) \right\}$

$\leq Q_{h,k}(s_{h,k}, a_{hk})$

3. $\delta_{1k}(s_{1,k}) = V_{1k}(s_{1k}) - V_1^{\pi_k}(s_{1k})$

$\leq Q_{1k}(s_{1k}, a_{1k}) - r(s_{1k}, a_{1k}) - \mathbb{E}_p\left[V_{2k}(s') \right] + m_{1k} + \delta_{2k}(s_{2k})$

$\leq Q_{1k}(s_{1k}, a_{1k}) - r(s_{1k}, a_{1k}) - \mathbb{E}_p\left[V_{2k}(s') \right] + m_{1k}$

$+ Q_{2k}(s_{2k}, a_{2k}) - r(s_{2k}, a_{2k}) - \mathbb{E}_p\left[V_{2k}(s') \right] + m_{2k} + \delta_{3k}(s_{3k})$

$\vdots$

$\leq \sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_p\left[V_{hk}(s') \right] + m_{hk}$

4) $R(T) = \sum_{k=1}^{K} V_1^*(s_{1k}) - V_1^{\pi_k}(s_{1k}) \leq \sum_{k=1}^{T} \delta_{1k}(s_{1k})$ since $Q_{hk}(s,a) \geq Q_h^*(s,a) \; \forall s, a$

$\leq \sum_{k=1}^{K} \sum_{h=1}^{H} \left(Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_p\left[V_{h+1,k}(s') \right] + m_{hk} \right.$

$= \sum_{k=1}^{K} \sum_{h=1}^{H} \left\{ \hat{r}_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) + \sum_{s'} \left| \hat{p}_{hk}(s' | s_{hk}, a_{hk}) - p_{hk}(s' | s_{hk}, a_{hk}) \right| V_{h+1,k}(s') \right\} +$

$+ \sum_{hk} m_{hk} + \sum_{hk} b_{hk}(s_{hk}, a_{hk})$

using the bounds found of previous parts: $\leq \sum_{h=1}^{H} \sum_{k=1}^{K} \left\{ \left| \hat{r}_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) \right| + H \left\| \hat{p}_{hk}(\cdot | s_{hk}, a_{hk}) - p_h(\cdot | s_{hk}, a_{hk}) \right\|_1 \right\}$

$+ \sum_{hk} m_{hk} + \sum_{hk} b_{hk}(s_{hk}, a_{hk})$

with prob at least $1-\delta$ ↓

$\leq \sum_{h,k} \beta_{h,k}^r(s_{hk}, a_{hk}) + H \beta_{hk}^p(s_{h,k}, a_{hk}) + b_{hk}(s_{hk}, a_{hk}) + 2H\sqrt{KH \log\left(\frac{2}{\delta}\right)}$ → $\leq b_{hk}(s_{hk}, a_{hk})$

$\leq \sum_{hk} 2 b_{hk}(s_{hk}, a_{hk}) + 2H\sqrt{KH \log\left(\frac{2}{\delta}\right)}$

Reinforcement learning - Homework 3 — Nicolás Verdeja

5) Since $\sqrt{\cdot}$ is concave, by Jensen's inequality

$$\sum_{h=1}^{H}\sum_{s,a}\sqrt{N_{hk}(s,a)} = HSA\sum_{h=1}^{S}\sum_{s,a}\frac{1}{HSA}\sqrt{N_{hk}(s,a)}$$

$$\leq HSA\sqrt{\frac{1}{HSA}\sum_{h=1}^{H}\sum_{s,a}N_{hk}(s,a)} \qquad \left(\sum_{h=1}^{H}\sum_{s,a}N_{hk}(s,a) \leq K\right)$$

$$\leq \sqrt{HSA}\sqrt{\sum_{h=1}^{H}K} = H\sqrt{SAK} \qquad \left(\sum_{h=1}^{H}K = H\right)$$

Now we chose $b_{hk}(s_{hk},a_{hk}) = \beta^r_{hk}(s_{hk},a_{hk}) + H\beta^p_{hk}(s_{hk},a_{hk})$

Also since $S \gg 1 \Rightarrow \log(2^S-2) \approx \log 2^S = S\log 2$

$$R(T) \lesssim 2\sum_{hk}\sqrt{\frac{1}{2N_{hk}(s_{hk},a_{hk})}\log(8KHSA)} + H\sqrt{\frac{2}{N_{hk}(s_{hk},a_{hk})}\left[S\log 2 + \log\left(\frac{4KHSA}{\delta}\right)\right]}$$
$$+ 2H\sqrt{KH\log\left(\frac{2}{\delta}\right)}$$

$$\lesssim H^2S^2A + H\sqrt{SAK} + H\sqrt{S}\left(H^2S^2A + H\sqrt{SAK}\right) + 2H\sqrt{KH\log\left(\frac{2}{\delta}\right)}$$

$$\simeq H^2S\sqrt{AK}$$