CHAUVIN Paul

MVA-RL-HW3

I. Best Arm identification :

1). Let $\mathcal{E} = \bigcup_{i=1}^{n} \bigcup_{t=1}^{+\infty} \left\{ |\hat{\mu}_{i,t} - \mu_i| > U(t,\delta') \right\}$

Then, $P(\mathcal{E}) \leq \sum_{i=1}^{n} P\left(\bigcup_{t=1}^{+\infty} |\hat{\mu}_{i,t} - \mu_i| > U(t,\delta') \right)$

Using the any time confidence bound, we have :

$$P(\mathcal{E}) \leq \sum_{i=1}^{n} P\left(\bigcup_{t=1}^{+\infty} |\hat{\mu}_{i,t} - \mu_i| > U(t,\delta') \right) \leq \sum_{i=1}^{n} \delta'$$

$$\leq n\delta'$$

$$\leq \delta$$

with $\boxed{\delta' = \frac{\delta}{n}}$

2). With the arm elimination condition for the optimal arm :

$$\exists j \,/\, \hat{\mu}_{j,t} - U(t,\delta') \geq \hat{\mu}^*_t + U(t,\delta')$$

In order to delete the optimal arm, μ^* should be outside the confidence interval : $|\hat{\mu}^*_t - \mu^*| > U(t,\delta')$

$$\Rightarrow P(|\hat{\mu}^*_t - \mu^*| > U(t,\delta'), \text{for one } t) \leq P(\mathcal{E}) \leq \delta$$

$$\Rightarrow \forall t,\ P(|\hat{\mu}^*_t - \mu^*| \leq U(t,\delta')) \geq 1-\delta$$

I. 3

Let $\hat{\mu}^*_t$ be the estimated reward of the arm with the largest expected reward μ^*

Under $\neg\mathcal{E}$: $\hat{\mu}^*_t \geq \mu^* - U(t, \delta')$

$$\hat{\mu}_{it} \leq \mu_i + U(t, \delta')$$

and with elimination condition

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

$\Rightarrow$ arm i will be deleted if

$$\mu^* - 2U(t, \delta') \geq \mu_i + 2U(t, \delta')$$

$$\Rightarrow \underbrace{\mu^* - \mu_i}_{\Delta_i} \geq \underbrace{4}_{C_1} U(t, \delta')$$

$$U(t, \delta') = \sqrt{\frac{1}{2t} \log\left(4t^2/\delta'\right)}$$

$$U(t, \delta') = \sqrt{\frac{1}{2t} \log\left(\frac{4nt^2}{\delta}\right)} \quad \text{because } \delta' = \frac{\delta}{n}$$

Hence, $\Delta_i^2 \geq \frac{16}{2t} \log\left(\frac{4nt^2}{\delta}\right)$

$$\Rightarrow \frac{\Delta_i^2}{16} t \geq \log\left(\frac{4nt}{\delta}\right)$$

$$\Rightarrow at \geq \log(bt)$$

with $\boxed{a = \frac{\Delta_i^2}{16} \; ; \; b = \frac{4n}{\delta}}$

$$\Rightarrow \boxed{t \geq \frac{1 + \sqrt{2u} + u}{a}} \;, \boxed{u = \log\left(\frac{b}{a}\right) - 1}$$

using the footnote.

CHAUVIN Paul

I. 4). The arm will be removed after sampling each sub-optimal arm.

The sample complexity is then :

$$O\left(\sum_{i \neq i^*} \frac{\log(b/a)}{a}\right) = O\left(\sum_{i \neq i^*} \frac{\log\left(\frac{n}{\delta \Delta_i^2}\right)}{\Delta_i^2}\right)$$

I. 5).

If multiple best arm exist, the algorithm would never stop as it would not be able to find "bad" arms and S would never be equal to one, so the algorithm would not stop.

II. Regret Minimization.

II. 1). For fixed s, a, h, K we have :

- Hoeffding's inequality :

$$P(\neg \mathcal{E}_r) = P\left(|\hat{r}_{hK}(s,a) - r_{hK}(s,a)| \geq B^r_{h,K}(s,a)\right)$$
$$\leq \underbrace{2\exp\left(-2N_{h,K}(s,a)\, B^r_{h,K}(s,a)^2\right)}_{\delta_r}$$

$$\Rightarrow 2N_{h,K}(s,a)\, B^r_{hK}(s,a)^2 = \log\left(\frac{2}{\delta_r}\right)$$

$$\Rightarrow \qquad B^r_{hK}(s,a) = \sqrt{\frac{\log\left(\frac{2}{\delta_r}\right)}{2N_{h,K}(s,a)}}$$

- Weissman inequality:

$$P(\neg \mathcal{E}_p) = P(\|\hat{P}_{hk}(\cdot|s,a) - P_h(\cdot|s,a)\|_1 \geq B^P_{hk}(s,a))$$
$$\leq (2^S - 2)\exp\left(\frac{-N_{hk}(s,a) B^P_{hk}(s,a)^2}{2}\right)$$

$$N_{hk}(s,a) B^P_{hk}(s,a)^2 = 2\log\left(\frac{2^S-2}{\delta_p}\right)$$

$$\Rightarrow B^P_{hk}(s,a)^2 = \sqrt{\frac{2}{N_{hk}(s,a)}\log\left(\frac{2^S-2}{\delta_p}\right)}$$

- Both inequalities give:

$$P(\neg \mathcal{E}_{s,a,k,h}) \leq P(\neg \mathcal{E}_r) + P(\neg \mathcal{E}_p)$$

we set $\delta_r = \delta_a = \frac{\delta'}{2}$

$$P(\neg \mathcal{E}_{s,a,k,h}) \leq \delta'$$

the bound for any s,a,k,h.

Then, $P(\mathcal{E}) = 1 - P(\neg \mathcal{E}) = 1 - P\left(\bigcup_{s,a,k,h} \neg \mathcal{E}_{s,a,k,h}\right)$

By Union bound:

$$1 - P\left(\bigcup_{s,a,k,h} \neg \mathcal{E}_{s,a,k,h}\right) \geq 1 - \sum_{s,a,k,h} P(\neg \mathcal{E}_{s,a,k,h})$$

we want $P(\mathcal{E}) \geq 1 - \delta/2$

$$\sum_{s,a,h,k} P(\neg \mathcal{E}_{s,a,h,k}) \leq \frac{\delta}{2}$$

$SAHK\,\delta' = \frac{\delta}{2} \Rightarrow \delta' = \frac{\delta}{2SAHK}$

Hence, the confidence bounds are:

$$B^r_{hk}(s,a) = \sqrt{\frac{\log\left(\frac{8SAHK}{\delta}\right)}{2N_{hk}(s,a)}}$$

$$B^P_{hk}(s,a) = \sqrt{\frac{2}{N_{hk}(s,a)} \log\left(\frac{(2^S-2)4SAHK}{\delta}\right)}$$

II.2)

Base case: $h = H$.

$$Q_{H,k}(s,a) = \hat{r}_{H,k}(s,a) + b_{H,k}(s,a)$$

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

we are under $\mathcal{E}$ event, hence:

$$\hat{r}_{H,k}(s,a) \geq r_{H,k}(s,a) - B^r_{H,k}(s,a)$$

then: $Q_{H,k}(s,a) \geq r_{Hk}(s,a) + b_{H,k}(s,a) - B^r_{H,k}(s,a)$

with bonus: $b_{H,k}(s,a) \geq B^r_{H,k}(s,a)$, base case is true.

Inductive step:

Assume $Q_{h,k}(s,a) \geq Q^*_h(s,a)$

let's prove $Q_{h-1,k}(s,a) \geq Q^*_{h-1}(s,a)$

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

$$= \hat{r}_{h-1,k}(s,a) + b_{h-1,k}(s,a) + \sum_{s'} \hat{P}_{h-1,k}(s'|s,a) \min\left(H, \max Q_{h,k}(s,a)\right)$$

$$Q^*_{h-1,k}(s,a) = r_{h-1,k}(s,a) + \sum_{s'} p_{h-1,k}(s'|s,a) \max_a Q^*_{h,k}(s,a)$$

$$\Rightarrow Q_{h-1,k}(s,a) - Q^*_{h-1,k}(s,a) = \hat{r}_{h-1,k}(s,a) + b_{h-1,k}(s,a)$$

$$- r_{h-1,k}(s,a) + \sum_{s'} \hat{p}_{h-1,k}(s'|s,a) \min\left(H, \max_a Q(s,a)\right)$$

$$- p_{h-1,k}(s'|s,a) \max_a Q^*_{h,k}(s,a)$$

$$\Rightarrow Q_{h-1,k} - Q^*_{h-1,k} \geq \hat{r}_{h-1,k}(s,a) + b_{h-1,k}(s,a) - r_{h-1,k}(s,a)$$

$$+ \sum_{s'} \min\left(H, \max_a Q\right)\left(\hat{p}_{h-1,k} - p_{h-1,k}\right)$$

$$\geq \hat{r}_{h-1,k} + b_{h-1,k} - r_{h-1,k}$$

$$- \sum_{s'} \min\left(H, \max Q\right)\left|\hat{p} - p\right|$$

$$\geq \hat{r}_{h-1,k} + b_{h-1,k} - r_{h-1,k}$$

$$- H \sum_{s'} \left|\hat{p} - p\right|$$

$$\geq \hat{r}_{h-1,k} + b_{h-1,k} - r_{h-1,k} - H B^p_{h-1,k}(s,a)$$

because confidence intervals holds as we're under event $\mathcal{E}$.

$$\geq - B^r_{h-1,k}(s,a) - H B^p_{h-1,k}(s,a) + b_{h-1,k}(s,a)$$

$$\geq 0$$

$$\Longleftrightarrow \quad b_{h-1,k}(s,a) \geq B^r_{h-1}(s,a) + H B^p_{h-1,k}$$

II. 3. 1)

$$V_h^{\pi_k}(s_{hk}) = r(s_{hk}, a_{hk}) + \sum_{s'} p_{h+1,k}(s'|(s,a))\, V_{h+1}^{\pi_k}(s')$$

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

$$= r(s_{hk}, a_{hk}) + E_p\left[V_{h+1k}(s')\right] - E_p\left[\delta_{h+1,k}(s')\right]$$

$$= r(s_{hk}, a_{hk}) + E_p\left[V_{h+1k}(s')\right] - \delta_{h+1,k}(s_{h+1,k}) - m_{h,k}$$

III. 3.2)

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

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

III. 3.3)

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

$$\leq Q_{1k}(s_{1k}, a_{1k}) - r(s_{1k}, a_{1k}) - E_p\left[V_{2k}(s')\right]$$

$$+ \delta_{2k}(s_{2k}) + m_{1k}$$

$\leq \ldots$ (we replace $\delta_{2k}(s_{2k})$ in the same way

$$\leq \sum_{h=1}^{H} Q_{h,k}(s_{h,k}, a_{h,k}) - r(s_{h,k}, a_{h,k}) - E_p\left(V_{h+1,k}(s')\right)$$

$$+ m_{hk}$$

II. 4).

$$R(T) = \sum_{k=1}^{K} V_1^*(s_{1,k}) - V_1^{\pi_k}(s_{1,k})$$

$$\leq \sum_{k=1}^{K} V_1(s_{1,k}) - V_1^{\pi_k}(s_{1,k}) = \sum_{k=1}^{K} \delta_{1k}(s_{1k})$$

$$\leq \sum_{k=1}^{K} \sum_{h=1}^{H} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk})$$

$$- E_{y\sim p}\left[V_{h+1,k}(y)\right] + m_{hk}$$

$$= \sum_{h,k} \hat{r}_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk})$$

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

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

$$\leq \sum_{k,h} |\hat{r}_{hk} - r| + H \sum_{s'} |\hat{p} - p| + b_{hk} + m_{hk}$$

- we have $\sum_{k,h} m_{hk} \leq 2H\sqrt{KH \log\left(\frac{2}{\delta}\right)}$ with probability $\geq 1 - \frac{\delta}{2}$ thanks to Azuma-Hoeffding
- Other terms are bounded by confidence intervals.

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

with probability $1-\delta$.

II . 5).

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

$$\leq HSA \sqrt{\sum_{h} \sum_{s,a} \frac{N_{hk}(s,a)}{HSA}}$$

$$= \sqrt{HSA \sum_{h} \sum_{SA} N_{hk}(s,a)}$$

$$\leq \sqrt{HSA} \sqrt{\sum_{h} K} \quad \text{because } \sum_{S,a} N_{k,k} \leq K$$

Hence :

$$R(T) \leq 2 \sum_{h,k} \sqrt{\frac{\log\left(\frac{8SAHK}{\delta}\right)}{2N_{hk}(s,a)}} + H\sqrt{\frac{2}{N_{hk}}\left(\log\frac{(2^S-2)4SAHK}{\delta}\right)}$$

$$+ 2H\, KH \log(2/\delta)$$

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

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