(1)

1. Best Arm Identification

- Let's compute $\mathcal{U}(t,\delta)$:

We want $\mathcal{U}(t,\delta)$ such that:

$$\mathbb{P}\left(\bigcap_{t=1}^{+\infty}\{|\hat{\mu}_{i,t}-\mu_i| \leqslant \mathcal{U}(t,\delta)\}\right) \geqslant 1-\delta$$

Thus $\mathbb{P}\left(\bigcup_{t=1}^{+\infty}\{|\hat{\mu}_{i,t}-\mu_i| > \mathcal{U}(t,\delta)\}\right) \leqslant \delta$

Let's note that

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

$$\leqslant \sum_{t=1}^{+\infty} \mathbb{P}(|\hat{\mu}_{i,t}-\mu_i| > \mathcal{U}(t,\delta)\}) \text{ by Union bond}$$

$$= \sum_{t=1}^{+\infty} \mathbb{P}(|\sum_{j=1}^{t} X_{i,j} - t\mu_i| > t\,\mathcal{U}(t,\delta)\})$$

$$\leqslant \sum_{t=1}^{+\infty} 2\exp\left(-\frac{2\mathcal{U}(t,\delta)^2 t^2}{t}\right) \text{ by Hoeffding.}$$

If we set $2\exp\left(-\frac{2\mathcal{U}(t,\delta)^2 t^2}{t}\right) = \frac{\delta}{2t^2}$, we will get:

$$\sum_{t=1}^{+\infty} 2\exp\left(-\frac{2\mathcal{U}(t,\delta)^2 t^2}{t}\right) = \sum_{t=1}^{+\infty}\frac{\delta}{2t^2} = \frac{\delta}{2}\frac{\pi^2}{6} < \frac{\delta}{2}.2 = \delta.$$

From that we can deduce $\mathcal{U}(t,\delta)$:

$$2\exp\left(-\frac{2\mathcal{U}(t,\delta)^2 t^2}{t}\right) = \frac{\delta}{2t^2}$$

$$\iff 2\,\mathcal{U}(t,\delta)^2 t = -\ln\left(\frac{\delta}{4t^2}\right)$$

$$\iff \mathcal{U}(t,\delta) = \sqrt{\frac{1}{2t}\ln\left(\frac{4t^2}{\delta}\right)}$$

(2) Let's show that $\mathbb{P}(\mathcal{E}) \leqslant \delta$ for a certain δ':

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

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

$$\leqslant \sum_{i=1}^{k} \delta' \quad \text{by definition of } \mathcal{U}(t, \delta')$$

$$= k\,\delta'$$

So, if we set $\delta' = \frac{\delta}{k}$ we get $\mathbb{P}(\mathcal{E}) \leqslant \delta$.

- Let's show that $\mathbb{P}($" i^* remains in the active set S"$) \geqslant 1-\delta$.

i^* is eliminated from S if:

$$\exists j \in S,\ \hat{\mu}_{j,t} - \mathcal{U}(t,\delta') \geqslant \hat{\mu}_{i^*,t} + \mathcal{U}(t,\delta').\ (*)$$

If $\neg\mathcal{E}$ holds, $\forall j \in S$ we have:

$$|\hat{\mu}_{j,t} - \mu_j| \leqslant \mathcal{U}(t, \delta').$$

$$\neg\mathcal{E} = \bigcap_{j=1}^{k} \bigcap_{t=1}^{+\infty} \{|\hat{\mu}_{j,t} - \mu_j| \leqslant \mathcal{U}(t,\delta')\}.$$

So we have:

$$|\hat{\mu}_{j,t} - \mu_j| \leqslant \mathcal{U}(t, \delta') \iff \hat{\mu}_{j,t} - \mathcal{U}(t,\delta') \leqslant \mu_j \leqslant \hat{\mu}_{j,t} + \mathcal{U}(t,\delta').$$
$$\forall j \in S.$$

From that and $(*)$, we have.

$$\mu_j \geqslant \mu_{i^*}.$$

③ By definition of i^*, $\mu_j \geq \mu_{i^*}$ is impossible hence i^* can't be eliminated from S when $\neg \mathcal{E}$ holds. Moreover, $\mathbb{P}(\neg \mathcal{E}) \geq 1-\delta$. and thus $\mathbb{P}($"i^* remains in the active set S"$) \geq 1-\delta$.

- Let's show that under event $\neg \mathcal{E}$, an arm $j \neq i^*$ will be removede from the active set when $\Delta_j \geq C_1 \mathcal{U}(t, \delta')$ for $C_1 = 4$.

Indeed, we have:

$$\Delta_j \geq 4\,\mathcal{U}(t, \delta') \iff \mu_{i^*} - 2\,\mathcal{U}(t, \delta') \geq \mu_j + 2\,\mathcal{U}(t, \delta')$$

Moreover, under $\neg \mathcal{E}$, we have:

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

$$\iff \mu_i - \mathcal{U}(t, \delta') \leq \hat{\mu}_{i,t} \leq \mu_i + \mathcal{U}(t, \delta') \; \forall i.$$

From that we can deduce that

$$\mu_{i^*} - 2\,\mathcal{U}(t, \delta') \geq \mu_j + 2\,\mathcal{U}(t, \delta')$$

$$\iff \hat{\mu}_{i^*,t} - \mathcal{U}(t, \delta') \geq \hat{\mu}_{j,t} + \mathcal{U}(t, \delta')$$

And thus j is elimated from the active set.

④

2. Regret Minimization in RL

- We want:

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

$$\iff \mathbb{P}\Big(\exists k, h, s, a : |\hat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a) \vee \|\hat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 > \beta^p_{hk}(s,a)\Big) \leq \frac{\delta}{2}.$$

Let's note that:

$$\mathbb{P}\Big(\exists k, h, s, a : |\hat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a) \vee \|\hat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 > \beta^p_{hk}(s,a)\Big)$$

$$\leq \sum_{k,h,s,a} \mathbb{P}\big(|\hat{r}_{hk}(s,a) - r_h(s,a)| > \beta^r_{hk}(s,a)\big) + \mathbb{P}\big(\|\hat{p}_{hk}(\cdot|s,a) - p_h(\cdot|s,a)\|_1 > \beta^p_{hk}(s,a)\big)$$

$$\leq \sum_{k,h,s,a} 2\exp\left(-2\,\frac{\beta^r_{hk}(s,a)^2 N_{hk}(s,a)^2}{N_{hk}(s,a)}\right) + (2^S-2)\exp\left(-\frac{1}{2} N_{hk}(s,a)\,\beta^p_{hk}(s,a)^2\right)$$

So we want:

$$2\exp\left(-2\,\beta^r_{hk}(s,a)^2 N_{hk}(s,a)\right) = \frac{\delta}{4HKSA}$$

and

$$(2^S-2)\exp\left(-\frac{1}{2} N_{hk}(s,a)\,\beta^p_{hk}(s,a)^2\right) = \frac{\delta}{4HKSA}.$$

5

$$2\exp\left(-2\,\beta_{hk}^{r\,2}(s,a)\,N_{hk}(s,a)\right) = \frac{\delta}{4HKSA}$$

$$\iff \beta_{hk}^{r^2}(s,a) = \frac{1}{2N_{hk}(s,a)} \ln\left(\frac{8HKSA}{\delta}\right)$$

$$\iff \beta_{hk}^{r}(s,a) = \sqrt{\frac{1}{2N_{hk}(s,a)} \ln\left(\frac{8HKSA}{\delta}\right)}$$

$$(2^S-2)\exp\left(-\frac{1}{2} N_{hk}(s,a)\,\beta_{hk}^{p}(s,a)^2\right) = \frac{\delta}{4HKSA}$$

$$\iff \beta_{hk}^{p}(s,a)^2 = \frac{2}{N_{hk}(s,a)} \ln\left(\frac{(2^S-2)\,4HKSA}{\delta}\right)$$

$$\iff \beta_{hk}^{p}(s,a) = \sqrt{\frac{2}{N_{hk}(s,a)} \ln\left(\frac{(2^S-2)\,4HKSA}{\delta}\right)}$$

So if we set $\beta_{hk}^{r}(s,a) = \sqrt{\frac{1}{2N_{hk}(s,a)} \ln\left(\frac{8HKSA}{\delta}\right)}$

and $\beta_{hk}^{p}(s,a) = \sqrt{\frac{2}{N_{hk}(s,a)} \ln\left(\frac{(2^S-2)\,4HKSA}{\delta}\right)}$

we get $\mathbb{P}(\neg \mathcal{E}) \leqslant \frac{\delta}{2}$.

⑥ • Let $b_{h,k}(s,a) := \beta^{r}_{hk}(s,a) + H\beta^{P}_{hk}(s,a)$.

Let's note that $\hat{r}_{H,k}(s,a) + b_{Hk}(s,a) \geqslant r_{H,k}(s,a)$

under event $\mathcal{E}$.

Under event $\mathcal{E}$, we have for $(s,a) \in S \times A$:

$$|\hat{r}_{H,k}(s,a) - r_{Hk}(s,a)| \leqslant \beta^{r}_{Hk}(s,a).$$

Hence $r_{H,k}(s,a) \leqslant \beta^{r}_{Hk}(s,a) + \hat{r}_{H,k}(s,a)$.

So
$$\hat{r}_{H,k}(s,a) + b_{Hk}(s,a) = \hat{r}_{H,k}(s,a) + \beta^{r}_{hk}(s,a) + \underbrace{H\beta^{P}(s,a)}_{\geqslant 0}$$
$$\geqslant r_{H,k}(s,a).$$

and thus
$$Q_{H,k}(s,a) = \hat{r}_{h,k}(s,a) + b_{h,k}(s,a)$$
$$\geqslant r_{H,k}(s,a) = Q^*_{h}(s,a) \; \forall s,a.$$

Let's suppose that $Q_{h,k}(s,a) \geqslant Q^*_{h}(s,a) \; \forall s,a$ for $h \leqslant H-1$

Let's show that $Q_{h-1,k}(s,a) \geqslant Q^*_{h-1}(s,a) \; \forall s,a$.

Let $(s,a) \in S \times A$.

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

⑦

$$\geqslant r_{h-1,k}(s,a) + \sum_{s'} P_{h-1,k}(s,a) V_{h,k}(s')$$

since $b_{h-1,k}(s,a) = \beta^r_{h-1,k}(s,a) + H\beta^P_{h-1,k}(s,a)$.

and $\mathcal{E}$ holds so:

$$|\hat{r}_{h-1,k}(s,a) - r_{h-1,k}(s,a)| \leqslant \beta^r_{h-1,k}(s,a)$$

$$\Leftrightarrow r_{h-1,k}(s,a) \leqslant \beta^r_{h-1,k}(s,a) + \hat{r}_{h-1,k}(s,a).$$

and

$$\sum_{s'} (\hat{P}_{h-1,k}(s'|s,a) - P_{h-1,k}(s'|s,a)) \underbrace{V_{h,k}(s')}_{\leq H}$$

$$\leqslant H \|\hat{P}_{h-1,k}(\cdot|s,a) - P_{h-1,k}(\cdot|s,a)\|_1$$

$$\leqslant H \beta^P_{h-1,k}(s,a)$$

$$\Leftrightarrow \sum_{s'} P_{h-1,k}(s'|s,a) V_{h,k}(s')$$

$$\leqslant H\beta_{h-1,k}(s,a) + \sum_{s'} \hat{P}_{h-1,k}(s'|s,a) V_{h,k}(s').$$

So we have:

$$Q_{h-1,k}(s,a) \geqslant r_{h-1,k}(s,a) + \mathbb{E}_{s'\sim p(\cdot|s,a)}[V_{h,k}(s')]$$

if $V_{h,k}(s') = H$ for a $s' \in S$, we have

$$V_{h,k}(s') = H \geqslant \cancel{Q^*_h(s')} \max_a Q^*_h(s',a).$$

if $V_{h,k}(s') = \max_a Q_{h,k}(s',a)$, we have.

(8)

$V_{k,k}(s') = \max_a Q_{k,k}(s',a) \geq \max_a Q^*_k(s',a)$?

by assumption.

so $\forall s' \in S$, $V_{k,k}(s') \geq \max_a Q^*_k(s',a)$.

Thus $r_{k-1,k}(s,a) + \mathbb{E}_{s' \sim p(\cdot|s,a)}[V_{k,k}(s')]$

$\geq r_{k-1,k}(s,a) + \mathbb{E}_{s' \sim p(\cdot|s,a)}[\max_{a'} Q^*_k(s',a')]$

$= Q^*_{k-1}(s,a)$

Finally, $Q_{k-1,k}(s,a) \geq Q^*_{k-1}(s,a)$ $\forall (s,a)$.

and thus $Q_{k,k}(s,a) \geq Q^*_k(s,a)$ $\forall (s,a)$
$\forall k = 1, \dots, H$

- 1. Let's show that $V_k^{\pi_k}(s_{kk}) = r(s_{kk}, a_{kk}) + \mathbb{E}_{s' \sim p(\cdot|s_{kk}, a_{kk})}[V_{k+1,k}(s')] - \delta_{k+1,k}(s_{k+1,k}) - m_{k,k}$.

We have:

$V_k^{\pi_k}(s_{kk}) = Q_k^{\pi_k}(s_{kk}, \pi_{k,k}(s_{kk}))$

$= Q_k^{\pi_k}(s_{kk}, a_{kk}) = r(s_{kk}, a_{kk}) + \mathbb{E}_{s' \sim p(\cdot|s_{kk}, a_{kk})}[V_{k+1}^{\pi_k}(s')]$

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

9

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

$$+ \delta_{h+1,k}(s_{h+1,k}) - \delta_{h+1,k}(s_{h+1,k})$$

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

$$- m_{hk}.$$

2. Let's show that $V_{h,k}(s_{hk}) \leq Q_{h,k}(s_{hk}, a_{hk})$

We have:

$$V_{h,k}(s_{hk}) = \min\{H, \max_a Q_{h,k}(s_{hk}, a_{hk})\}$$

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

$$= Q_{h,k}(s_{hk}, a_{hk}) \text{ because } a_{hk} = \operatorname{argmax}_a Q_{h,k}(s_{hk}, a)$$

We do have $V_{h,k}(s_{hk}) \leq Q_{h,k}(s_{hk}, a_{hk})$

3. Let's prove eq. 1.

Let's not that for $h = 1, \dots, H$, we have:

$$\delta_{hk}(s_{hk}) = V_{hk}(s) - V_h^{\pi k}(s) \leq Q_{hk}(s_{hk}, a_{hk})$$

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

$$+ \delta_{h+1,k}(s_{h+1,k}) + m_{hk}.$$

by q. 1 and q. 2.

(10)

Hence:

$$\delta_{1,k}(s_{1,k}) \leqslant Q_{1k}(s_{1,k}, a_{1k}) - r(s_{1k}, a_{1k}) - \mathbb{E}_{s' \sim p(\cdot|s_{1k}, a_{1k})}[V_{2,k}(s')] + \delta_{2k}(s_{2k}) + m_{1k}$$

$$\leqslant Q_{1k}(s_{1,k}, a_{1k}) - r(s_{1k}, a_{1k}) - \mathbb{E}_{s' \sim p(\cdot|s_{1k}, a_{1k})}[V_{2,k}(s')] + m_{1k} + Q_{2k}(s_{2k}, a_{2k}) - r(s_{2k}, a_{2k}) - \mathbb{E}_{s' \sim p(\cdot|s_{2k}, a_{2k})}[V_{3,k}(s')] + \delta_{3k}(s_{3k}) + m_{2k}$$

$$\leqslant \dots$$

$$\leqslant Q_{1k}(s_{1k}, a_{1k}) - r(s_{1k}, a_{1k}) - \mathbb{E}[V_{2,k}(s')] + m_{1k} + Q_{2k}(s_{2k}, a_{2k}) - r(s_{2k}, a_{2k}) - \mathbb{E}_{s' \sim p(\cdot|s_{2k}, a_{2k})}[V_{3,k}(s')] + m_{2k} + \dots + Q_{Hk}(s_{Hk}, a_{Hk}) - r(s_{Hk}, a_{Hk}) \underbrace{- \mathbb{E}_{s' \sim p(\cdot|s_{Hk}, a_{Hk})}[V_{H+1,k}(s')]}_{=0} + \underbrace{\delta_{H+1\,k}(s_{H+1\,k})}_{= V_{H+1\,k}(s_{H+1\,k}) - V^{\pi_k}_{H+1}(s_{H+1\,k}) = 0} + m_{Hk}.$$

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

(11) Finally,

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

- Let's suppose that $\mathcal{E}$ and $\left\{\sum_{k,h} m_{hk} \leqslant 2H\sqrt{KH\log\left(\frac{2}{\delta}\right)}\right.$

hold.

Let's show that $R(T) \leqslant 2\sum_{kh} b_{hk}(s_{hk}, a_{hk}) + 2H\sqrt{KH\log\left(\frac{2}{\delta}\right)}$

with probability $1-\delta$:

$$R(T) = \sum_{k=1}^{K} V_1^*(s_{1k}) - V_1^{\pi_k}(s_{1k})$$
$$= \sum_{k=1}^{K} V_1^*(s_{1k}) - V_{1k}(s_{1k}) + \underbrace{V_{1k}(s_{1k}) - V_1^{\pi_k}(s_{1k})}_{=\delta_1(s_{1k})}$$

Let's note that $V_{1k}(s_{1k}) \geqslant \cancel{Q^*} \max_a Q^*(s_{1k}, a)$.

If $V_{1k}(s_{1,k}) = H$, then $V_{1k}(s_{1,k}) = H \geqslant \max_a Q_1^*(s_{1k}, a)$

If $V_{1k}(s_{1,k}) = \max_a Q_{1k}(s_{1k}, a) \geqslant \max_a Q_1^*(s_{1k}, a)$

because Q_{1k} is optimistic under $\mathcal{E}$.

Thus,

$$R(T) \leqslant \sum_{k=1}^{K} \underbrace{V_1^*(s_{1k}) - \underbrace{\max_a Q_1^*(s_{1,k}, a)}_{=V_1^*(s_{1k})}}_{=0} + \delta_1(s_{1k})$$

(12)

$$\leq \sum_{k,h} Q_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk}) - \mathbb{E}_{Y \sim p(\cdot|s_{hk}, a_{hk})}[V_{h+1,k}(Y)]$$
$$+ m_{hk} \quad \text{by equation (1).}$$

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

$$= \sum_{k,h} \hat{r}_{hk}(s_{hk}, a_{hk}) - r(s_{hk}, a_{hk})$$
$$+ \sum_{s'} \big(\hat{p}_{hk}(s'|s_{hk}, a_{hk}) - p(s'|s_{hk}, a_{hk})\big) V_{h+1,k}(s')$$
$$+ b_{hk}(s_{hk}, a_{hk}) + \sum_{k,h} m_{hk}.$$

$$\leq \sum_{k,h} \beta^r_{hk}(s_{hk}, a_{hk}) + H \beta^p_{hk}(s_{hk}, a_{hk})$$
$$+ b_{hk}(s_{hk}, a_{hk}) + 2H\sqrt{KH \log\left(\frac{2}{\delta}\right)}$$

because we are under $\mathcal{E}$ and $\left\{\sum_{k,h} m_{hk} \leq 2H\sqrt{KH\log\left(\frac{2}{\delta}\right)}\right\}$

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

by definition of the bonus.

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

Moreover, $\mathbb{P}\left(\neg \mathcal{E} \cup \left\{\sum_{k,h} m_{hk} > 2H\sqrt{KH\log\left(\frac{2}{\delta}\right)}\right\}\right)$
$$\leq \frac{\delta}{2} + \frac{\delta}{2} = \delta$$

thus $\mathbb{P}\left(\mathcal{E} \cap \left\{\sum_{k,h} m_{hk} \leq 2H\sqrt{KH\log\left(\frac{2}{\delta}\right)}\right\}\right)$
$$\geq 1 - \delta$$

(13)

$$R(T) \leqslant 2 \sum_{k,h} b_{hk}(s_{hk}, a_{hk}) + 2H\sqrt{KH \log\left(\frac{2}{\delta}\right)}$$

with probability at least $1-\delta$.

•