Reinforcement Learning - HW 3

Elías Masquil

1) Best Arm Identification

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

Any-time bound:

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

$$P\left\{ |\hat{\mu}_{i,t} - \mu_i| > U(t,\delta') \right\} \leq \delta' \quad \text{for a single arm and time}$$

By Hoeffding's inequality

$$P\left(|\hat{\mu}_{i,t} - \mu_i| \geq U(t,\delta')\right) \leq 2e^{-2tU(t,\delta')} = \delta'$$

$$\Rightarrow \boxed{U(t,\delta') = \sqrt{\frac{1}{2t} \log \frac{2}{\delta'}}}$$

By Union's bound

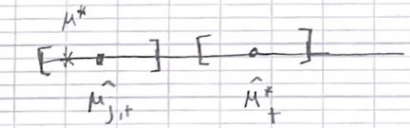
$$P(\varepsilon) \leq \sum_{i=1}^{K} \sum_{t=1}^{\infty} \delta' \qquad \left(\sum_t \frac{1}{t^2} = \frac{\pi^2}{6} \right)$$

If we consider $\boxed{\delta' = \frac{6}{\pi^2} \frac{1}{K} \cdot \frac{1}{t^2} \cdot \delta}$ $\Rightarrow P(\varepsilon) \leq \delta$

ii) Show that with $p \geq 1-\delta$, the optimal arm remains in the active set

Arm elimination condition (for the optimal arm)

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



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

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

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

Reinforcement Learning – HW 3

Elías Masquil

1) iii) Let's define $\hat{\mu}_t^*$ as the estimated reward of the arm with the largest expected reward μ^*

Under event $\mathcal{E}$:

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

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

And the elimination condition is:

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

Then, the arm i will be dropped if

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

$$\boxed{\Delta_i \geq 4\,U(t,\delta')}$$

$$U(t,\delta') = \sqrt{\frac{1}{2t}\log\left(\frac{2\pi^2 t^2 K}{6\delta}\right)}$$

$$\Delta_i^2 \geq 16\,\frac{1}{2t}\log(bt) \quad \text{with} \quad b = \pi\sqrt{\frac{K}{3\delta}}$$

$$at \geq \log(bt) \quad \text{with} \quad a = \frac{\Delta_i^2}{16}$$

Using the suggestion

$$t \geq \frac{1 + \sqrt{2v} + v}{a} \quad \text{with} \quad v = \log\frac{b}{a} - 1$$

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

iv) After sampling each sub-optimal arm T_i times, that arm will be eliminated. Then the sample complexity is just the sum of all T_i $\forall\, i=1,\dots,n$:

$$O\left(\sum_{i\neq i^*} \frac{\log\left(\frac{1}{\Delta_i^2}\sqrt{\frac{K}{\delta}}\right)}{\Delta_i^2}\right)$$

Reinforcement Learning - HW 3

Elias Hasquil

1) Best Arm Identification

v) I think it won't work since it won't converge. After eliminating all sub-optimal arms, the algorithm might keep iterating across the optimal ones. Since $\Delta_i = 0$ for optimal arms, they won't be eliminated and then the time of removal will be ∞.

2) Regret minimization in RL

i) For fixed s, a, h, k we have two confidence intervals, for the rewards and for the transitions:

- Using Hoeffding's inequality:

$$\mathbb{P}(\mathcal{E}_r) = \mathbb{P}\left(|\hat{r}_{hk}(s,a) - r_{hk}(s,a)| \geq B^r_{hk}(s,a)\right) \leq 2e^{-2N_{hk}(s,a)\,B^r_{hk}(s,a)^2}$$

$$2e^{-2N_{hk}(s,a)\,B^r_{hk}(s,a)^2} = \delta_r$$

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

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

- Using Weissmain inequality:

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

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

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

Then $\mathbb{P}(\mathcal{E}) \leq \mathbb{P}(\mathcal{E}_r) + \mathbb{P}(\mathcal{E}_p)$

$\mathbb{P}(\mathcal{E})$: Probability of having at least one of the true $p_{hk}(\cdot|s,a)$ or $r_{hk}(s,a)$ outside of the confidence bounds in a fixed s,a,h,k

Reinforcement Learning - HW 3

Elías Masquil

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

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

We have a bound for any k, h, s, a.

Then $\mathbb{P}(\mathcal{E}) = 1 - \mathbb{P}(\neg\mathcal{E}) = 1 - \mathbb{P}\left(\bigcup_{sahk} \neg\mathcal{E}_{sahk}\right)$

By union bound

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

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

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

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

Now we can correctly define the original confidence bounds, using the δ' value.

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

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

Reinforcement Learning - HW3

Elías Masquil

2) Regret minimization in RL

ii)

Base case

$h = H$

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

$$Q_H^*(s,a) = r_{H,K}(s,a)$$

Since we are under $\mathcal{E}$ event, we also know that

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

Then

$$Q_{H,K}(s,a) \geq r_{HK}(s,a) + b_{H,K}(s,a) - B^r_{H,K}(s,a)$$

If we define a bonus such that

$b_{H,K}(s,a) \geq B^r_{H,K}(s,a)$ the base case is true.

Inductive step

Assume that $Q_{h,K}(s,a) \geq Q^*_h(s,a)$, let's prove the inequality for $h-1$

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

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

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

Taking the difference

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

$$+ \sum_{s'} \hat{P}_{h-1,K}(s'|s,a) \min\{H, \max_a Q_{h,K}(s,a)\}$$

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

Reinforcement Learning - HW 3

Elías Masquil

By using the inductive assumption

$$Q_{h-1,k} - Q^*_{h-1,k} \geq \hat{r}_{h-1,k}(s,a) + b_{h-1,k}(s,a) - r_{h-1,k}(s,a) + \sum_{s'} \min\{H, \max_a Q_{hk}(s,a)\} \left(\hat{p}_{h-1,k}(s'|s,a) - p_{h-1,k}(s'|s,a)\right) \geq$$

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

$$\geq \hat{r}_{h-1,k}(s,a) + b_{h-1,k}(s,a) - r_{h-1,k}(s,a) - H \sum_{s'} \left|\hat{p}_{h-1,k}(s'|s,a) - p_{h-1,k}(s'|s,a)\right|$$

Since we are under event $\mathcal{E}$, both confidence intervals hold

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

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

$$\geq 0 \iff \boxed{b_{h-1,k}(s,a) \geq B^r_{h-1,k}(s,a) + H B^p_{h-1,k}}$$

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

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

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

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

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

$\leq Q_{h,k}(s_{h,k}, a_{hk})$ since a_{hk} was the greedy action ($a_{hk} = \arg\max Q_{h,k}(s_{h,k}, a)$)

Reinforcement Learning - HW 3

Elias Masquil

2) Regret minimization in RL

iii) 3)

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

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

$$+ \delta_{2k}(s_{2k}) + m_{1k} \quad \text{(develop this)}$$

$$\leq \dots$$

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

iv)

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

Since our estimates are optimistic

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

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

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

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

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

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

→ Note that by Azuma Hoeffding:

$$\sum_{k,h} m_{hk} \leq 2H\sqrt{kH \log(2/\delta)} \quad \text{with proba} \geq 1 - \delta/2$$

→ The other terms are bounded by the confidence intervals

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

$$\leq \sum_{k,h} \beta^r_{h,k}(s,a) + H\,\beta^p_{h,k}(s,a) \leq \sum_{k,h} b_{hk}(s_{kh}, a_{kh})$$

with proba $\geq 1 - \delta/2$

Reinforcement Learning - HW 3

Elías Masquil

Finally, combining the two previous bounds:

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

with probability $1-\delta$

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

using Jensen inequality for concave functions

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

since $\sum_{s,a} N_{hk}(s,a) \leq K$

$$\Rightarrow \leq \sqrt{HSA}\sqrt{\sum_{h} K} = H\sqrt{SAK}$$

Now making the inequalities between the bonus and the confidence bounds, hold with equality. We have

$$R(T) \leq 2 \sum_{h,k} \left(\sqrt{\frac{\log\left(\frac{8SAHK}{\delta}\right)}{2(N_{hk}(s,a))}} + H\sqrt{\frac{2}{N_{hk}(s,a)} \log\left[\frac{(2^3-2)\,4SAHK}{\delta}\right]} \right)$$

$$+ 2H\sqrt{KH\log(2/\delta)}$$

$$\leq \frac{2}{\sqrt{2}} \sqrt{\log\left(\frac{8SAHK}{\delta}\right)} \sum_{h,k} \sqrt{\frac{1}{N_{hk}(s,a)}} + H\sqrt{2\log\left[\frac{(2^3-2)\,4SAHK}{\delta}\right]} \sum_{h,k} \sqrt{\frac{1}{N_{hk}(s,a)}}$$

$$+ 2H\sqrt{KH\log(2/\delta)}$$

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

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