Notes for 2022 exam

abandon

q1)

states

C1 - not good

C2 - not good

C0

C1 - suitable

C2 - suitable

C1 - perfect

C2 - perfect

- C0 is starting state
- abandon and recruit are terminal states

recruit

actions

- A (abandon): available in all states but the terminal ones, lead to abandon state with prob. 100% and reward 0
- R (recruit): available in C1, C2 - suitable and - perfect states, lead to recruit state with prob. 100% and reward 200 for C1, C2 - perfect and 50 for C1, C2 - suitable
- I (interview): available in C0 and C1 states, leads to C1 - not good wp. 25%, C1 - suitable wp. 50% or C1 - perfect wp. 25% with reward - 30 from C0 and to C2 - not good wp. 25%, C2 - suitable wp. 50% or C2 - perfect wp. 25% with reward - 30 from any of the C1 states.

Q2) Bellman equations for v_π

$$\begin{cases} v_\pi(s_0) = (-1 \times 0.3) + 0.9\left[0.3\, v_\pi(s_0) + 0.7\, v_\pi(s_1)\right] \\ v_\pi(s_1) = (-1 \times 0.85) + 0.9\left[0.85\, v_\pi(s_1) + 0.15\, v_\pi(s_2)\right] \\ v_\pi(s_2) = 10 \end{cases}$$

Q3) Policy improvement: in s_1 the greedy action wrt. v_π is

$$\arg\max_a \left[r(s_1,a) + \gamma \sum_{s'} p(s'|s_1,a)\, v_\pi(s')\right]$$

is obviously achieved when a = east for which

$$v_{\tilde{\pi}}(s_1) = (-1 \times 0.3) + 0.9\left[0.3\, v_{\tilde{\pi}}(s_1) + 0.7 \times 10\right]$$
$$= 6.575 > 2.128$$

Hence $\tilde{\pi}$ which differs from π by chosing action east in s_1 has higher value function

Q4) This time it holds both in s_0 and s_1 that

$$v_{\hat{\pi}}(s) = \arg\max_a \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a)\, v_{\tilde{\pi}}(s')\right]$$

showing that $v_{\hat{\pi}} = v^*$ and hence that $\tilde{\pi}$ is optimal

---

Q5) Usualy this would be shown by backward induction (Bellman equations for optimal policy in finite horizon case) but here there are only 8 possible sequence of actions (see Q7 below) and it is easily checked that the best ones are (2, 1, *) both yielding a reward of 2.

Q6) $\begin{cases} \pi(1|s) = 1 & \text{stays in state 1} \\ \pi(2|s) = 1 & \text{stays in states } \{1,2\} \end{cases}$

and neither of them reaches the third state which provides rewards

Q7) See part of course on REINFORCE algorithm

Q8)

Sequence	$\sum_{i=0}^{2} R_{i+1}$	$\sum_{i=0}^{2} \frac{d \log \pi_\theta(A_i \mid S_1)}{d\theta}$	prob. sequence
(2,1,1)	2	$1/\theta - 2/(1-\theta)$	$\theta(1-\theta)^2$
(2,1,2)	2	$2/\theta - 1/(1-\theta)$	$\theta^2(1-\theta)$
(1,2,1)	1	$1/\theta - 2/(1-\theta)$	$\theta(1-\theta)^2$

Q9) The REINFORCE formula yields

$$2\left[(1-\theta)^2 - 2\theta(1-\theta)\right] + 2\left[2\theta(1-\theta) - \theta^2\right] + 1\left[(1-\theta)^2 - 2\theta(1-\theta)\right]$$

$$= 3\theta^2 - 8\theta + 3$$ for the gradient of the policy

The optimal policy corresponds to the root that is in $(0,1)$:

$$\theta^* = \frac{4 - \sqrt{7}}{6} \simeq 0{,}451$$

Q10) There are $K(K-1)$ actions consisting of the (ordered) pairs (j,k) where $j \in \{1,\dots,K\}$, $k \in \{1,\dots K\}$ with $k \neq j$

The best action is (j^*, k^*) where

$$\begin{cases} \theta_{j^*} > \theta_j & \text{for } j \neq j^* \text{ (highest } \theta_i) \\ \theta_{k^*} > \theta_k & \text{for } k \neq j^*, k^* \text{ (second highest } \theta_i) \end{cases}$$

Q11) See course for the definition of regret and its fundamental rewriting. Here

$$\mathbb{E}[R_T] = \sum_{(j,k) \neq (j^*,k^*)} [(\theta_{j^*} - \theta_j) + \alpha(\theta_{k^*} - \theta_k)] \, \mathbb{E}[N_{j,k}(T)]$$

where $N_{j,k}(T) = \sum_{t=1}^{T} \mathbb{1}\{A_t = (j,k)\}$

Q12) $$\overline{X}_{j,k}(t) = \frac{1}{N_{j,k}(t)} \sum_{s=1}^{t} X_s \, \mathbb{1}\{A_s = (j,k)\}$$

UCB plays

$$\arg\max_{(j,k)} \overline{X}_{j,k}(t) + \sqrt{\frac{\gamma \log t}{2 N_{j,k}(t)}}$$

Q13) $\Delta_{j,k} = (\theta_{j^*} - \theta_j) + \alpha(\theta_{k^*} - \theta_k)$ and the result from the course is that for UCB

$$\mathbb{E}[R_T] \leq \sum_{(j,k) \neq (j^*,k^*)} C \frac{\log T}{\Delta_{j,k}} + O(1)$$

which grows as $\log T$, as expected, but

also as K^2 (depending on the gaps) which is surprising (and in fact not optimal) as there are only K unknown values of θ_i (and not really $K(K-1)$ unrelated arms).

Q14) Action (j,k) is associated with vector

$$A_t = \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{pmatrix} \begin{matrix} \\ \leftarrow j^{th} \text{ position} \\ \\ \leftarrow k^{th} \text{ position} \\ \end{matrix}$$

such that

$$\mathbb{E}[X_t \mid A_t] = A_t^T \theta = \theta_j - \theta_k$$

where

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_K \end{pmatrix}$$

Q15) $\bar{X}_1(T)$ and $\bar{X}_2(T)$ are independent and follow, respectively, $N\left(\mu_1, \frac{\sigma^2}{(T/2)}\right)$ and $N\left(\mu_2, \frac{\sigma^2}{(T/2)}\right)$ distributions, hence

$$\bar{X}_1(T) - \bar{X}_2 \sim N\left(\Delta, \frac{4\sigma^2}{T}\right)$$

Q16) The error (ie. selecting arm 2) occurs when $\bar{X}_2(T) > \bar{X}_1(T)$, that is,

when $\bar{X}_1(T) - \bar{X}_2(T) < 0$

Hence

$$\mathbb{P}(\text{error}) = \mathbb{P}\underbrace{(\bar{X}_1(T) - \bar{X}_2(T) - \Delta < -\Delta)}_{\sim \mathcal{N}(0, \frac{4\sigma^2}{T})} \leq e^{-\frac{\Delta^2 T}{8\sigma^2}}$$

which is less than δ when

$$\frac{\Delta^2 T}{8\sigma^2} \geq \log \frac{1}{\delta}$$

that is

$$T \geq \frac{8\sigma^2}{\Delta^2} \log \frac{1}{\delta}$$

Q17) The error occurs when

$$\bar{X}_1(T) - \bar{X}_2(T) < \frac{4\sigma^2 \log(1/\delta)}{\Delta T}$$

Hence

$$\mathbb{P}(\text{error}) \leq e^{-\underbrace{\left(\Delta + \frac{4\sigma^2 \log(1/\delta)}{\Delta T}\right)^2 \frac{T}{8\sigma^2}}_{\geq \frac{8\sigma^2 \log(1/\delta)}{T}}}$$

thus

$$\mathbb{P}(\text{error}) \leq e^{-\log\left(\frac{1}{\delta}\right)} = \delta$$

Q18) The exact probability of correct (ie. choose arm 1) is given by

$$\mathbb{P}\left(\bar{X}_1(T) - \bar{X}_2(T) > \frac{4\sigma^2 \log(1/\delta)}{\Delta T}\right)$$

where $\bar{X}_1(T) - \bar{X}_2(T) \sim \mathcal{N}\left(\Delta, \frac{2\sigma^2}{T}\right)$

By symetry of the Gaussian distribution this is larger than 1/2 iff.

$$\frac{4\sigma^2 \log(1/\delta)}{\Delta T} \leq \Delta$$

that is $$T \geq \frac{4\sigma^2}{\Delta^2} \log\left(\frac{1}{\delta}\right)$$

Q19) $KL(p\|q) = \mathbb{E}_{X\sim p}\left[\log \frac{p(X)}{q(X)}\right]$, here

$$= \mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[\frac{1}{2\sigma^2}\underbrace{\{(X-\mu')^2 - (X-\mu)^2\}}_{= 2\left(X - \frac{\mu'+\mu}{2}\right)(\mu-\mu')}\right]$$

$$= \frac{(\mu-\mu')^2}{2\sigma^2}$$

Q20) (Hard!) Using the general inequality with

<u>model ν</u>: $\mu_1, \mu_2 < \mu_1$

<u>model ν'</u>: $\mu_1' = \mu_1, \mu_2' > \mu_1$

yields

$$KL(\mu_2, \mu_2')\, \mathbb{E}_\nu[N_2(T)] \geq d(\delta, 1-\delta)$$

by selecting E as the event that arm 2 get selected, which is an error under model ν which occurs w.p. at most δ whereas it is the correct outcome under model ν' (as $\mu'_2 > \mu'_1$) and hence occurs w.p. at least $1-\delta$

(Notice that when $\delta < 1/2$ $d(p,q) > d(\delta, 1-\delta)$ when $p < \delta$ and $q > 1-\delta$)

Hence

$$\frac{(\mu_2 - \mu'_2)^2}{2\sigma^2} \mathbb{E}_\nu[N_2(T)] \geq d(\delta, 1-\delta)$$

for all $\mu'_2 > \mu_1$ so that it also holds when $\mu'_2 = \mu_1$ by continuity and thus

$$\mathbb{E}_\nu[N_2(T)] \geq \frac{2\sigma^2}{\Delta^2} d(\delta, 1-\delta)$$

Proceed similarly with the second change of distribution to obtain the inequality for $\mathbb{E}_\nu[N_1(T)]$ (when $\mu_1 < \mu_2$).