Assignment 2
Reinforcement Learning
Prof. B. Ravindran

1. Which of the following statements is NOT true about Thompson Sampling or Posterior Sampling?

   (a) After each sample is drawn, the q_* distribution for that sampled arm is updated to be closer to the true distribution.

   (b) Thompson sampling has been shown to generally give better regret bounds than UCB.

   (c) In Thompson sampling, we do not need to eliminate arms each round to get good sample complexity.

   (d) The algorithm requires that we use Gaussian priors to represent distributions over q_* values for each arm.

   Sol. (d)
   (d) is NOT true. We are not constrained to Gaussian priors. We can assume a prior distribution of any type over the q_* values for each arm.

2. In UCB, the term $\sqrt{\frac{2\ln(n)}{n_j}}$ is added to each arm's Q value and the arm with the highest value of this sum is chosen. Which one of the following would definitely happen to the frequency of picking sub-optimal arms when adding $\sqrt{\frac{2\ln(n)}{n_j^2}}$ instead of $\sqrt{\frac{2\ln(n)}{n_j}}$?

   (a) Sub-optimal arms would be chosen more frequently.

   (b) Sub-optimal arms would be chosen less frequently.

   (c) Makes no change to the frequency of picking sub-optimal arms.

   (d) Sub-optimal arms could be chosen less or more frequently, depending on the samples.

   Sol. (d)
   The uncertainty term in the UCB algorithm shrinks much faster as actions are selected. This means that we have higher confidence that our estimates are correct with fewer samples. If we are able to make good estimates of the true action-values within a few samples, this would mean that we waste less time selecting sub-optimal arms.
   However, fewer samples typically mean worse estimates -especially for bandit problems with noisy reward distributions. Low uncertainty terms would increase the importance of potentially worse Q-value estimates after a few trials, and it is possible that we would select sub-optimal actions more often as a result.

   Since either (a) or (b) could be True, depending on the problem setting, the correct answer is (d).

3. In the proof of the UCB Theorem, when $l = \lceil 8\ln m/\Delta_i^2 \rceil$, $q_*(a^*) - q_*(i) - 2C_{m,l} = 0$. Which of the following correctly explains $q_*(a^*) - q_*(i) - 2C_{m,T_i(s_i)}$ being less than 0 when we replace this fixed value of l by the value $T_i(s_i)$: the number of times the arm i is selected. ?

   Note: $C_{m,l} = \sqrt{2\ln(m)/l}$,$q_*(a^*) - q_*(i) = \Delta_i$

(a) $T_i(s_i) < l$ thus $C_{m,T_i(s_i)} < C_{m,l}$.

(b) $T_i(s_i) > m$ thus $C_{m,T_i(s_i)} < C_{m,l}$.

(c) $T_i(s_i) > l$ thus $C_{m,T_i(s_i)} < C_{m,l}$.

(d) $T_i(s_i) < l$ thus $C_{m,T_i(s_i)} > C_{m,l}$.

Sol. (d)
If the number of times that the arm is selected is less than l, it implies that our confidence interval is larger. i.e $C_{m,T_i(s_i)} > C_{m,l} \implies q_*(a^*) - q_*(i) - 2C_{m,T_i(s_i)} < q_*(a^*) - q_*(i) - 2C_{m,l} = 0$

4. Which of the following is true about the Median Elimination algorithm?

   (a) It is a regret minimizing algorithm.

   (b) The probability of the ϵ_l-optimal arms of round l being eliminated is less than δ_l for the round.

   (c) It is guaranteed to provide an ϵ-optimal arm at the end.

   (d) Replacing ϵ with $\frac{\epsilon}{2}$ doubles the sample complexity.

   Sol. (b)
   Look at the derivation for Median Elimination.

5. We need 8 rounds of median-elimination to get an $(\epsilon, \delta) - PAC$ arm. Approximately how many samples would have been required using the naive $(\epsilon, \delta) - PAC$ algorithm given $(\epsilon, \delta) = (1/2, 1/e)$? (Choose the value closest to the correct answer)

   (a) 15000

   (b) 10000

   (c) 500

   (d) 20000

   Sol. (a)
   no. of rounds in median elimination = $\log_2$(no. of arms) = 8
   $\implies$ no. of arms = 256
   No. of samples required by the naive algorithm = $\frac{2k}{\epsilon^2} \ln(2k/\delta)$
   Substituting,
   No. of samples required = 14824

6. In case of UCB, the regret is generally of the form $f(n) + c$, where, for the confidence bound given by $\sqrt{\frac{2\ln(n)}{n_j}}$, $f(n) = \sum_{j=1}^{k} \frac{8\ln(n)}{\Delta_j}$ and $c = (1 + \frac{\pi^2}{3}) \sum_{j=1}^{k} \Delta_j$. What is the form of $f(n)$ for the confidence bound given by $\sqrt{\frac{\alpha \ln(n)}{n_j}}$, where $\alpha \geq 2$?

   (a) $\sum_{j=1}^{k} \frac{4\alpha \ln(n)}{\Delta_j}$

   (b) $\sum_{j=1}^{k} \frac{2\alpha^2 \ln(n)}{\Delta_j}$

   (c) $\sum_{j=1}^{k} \frac{\alpha^3 \ln(n)}{\Delta_j}$

   (d) $\sum_{j=1}^{k} \frac{\alpha^4 \ln(n)}{2\Delta_j}$

Sol. (a)
Re-derive the proof for UCB in class with a general α.

7. In the naive (ϵ, δ)-PAC algorithm, suppose we draw $\frac{2}{\epsilon^2} ln(\frac{k}{\delta})$ samples for each arm instead of $\frac{2}{\epsilon^2} ln(\frac{2k}{\delta})$ samples. Using the same analysis presented in the lectures, with what probability can we guarantee that the arm a' returned will have $q_*(a')$ value ϵ close to the q_* value of the optimal arm?

   (a) $1 - \delta$
   (b) $1 - 4\delta$
   (c) $1 - 2\delta$
   (d) $1 - \frac{\delta}{2}$

   Sol. (c)
   From the lectures, assuming that a' is an arm with $q_*(a')$ value at least ϵ away from $q_*(a^*)$:

   $P(Q(a') \geq Q(a^*)) \leq P(Q(a') \geq q_*(a') + \epsilon/2) + P(Q(a^*) < q_*(a^*) - \epsilon/2)$

   Using the Chernoff-Hoeffding bounds introduced in the lectures:

   $P(Q(a') \geq Q(a^*)) \leq 2e^{\frac{-\epsilon^2 l}{2}}$

   Substituting the new number of samples drawn $= l$, we get:

   $P(Q(a') \geq Q(a^*)) \leq \frac{2\delta}{k}$
   Summing over all k arms, the probability of picking an arm a' that does not meet our criteria is bounded by 2δ, and the probability that the arm returned has $q_*(a')$ value ϵ close to the q_* value of the optimal arm is, therefore, $\geq 1 - 2\delta$

8. In median elimination method for obtaining (ϵ, δ)-PAC bounds, which of the following is a plausible update rule for ϵ_l and δ_l if $\epsilon_1 = \frac{\epsilon}{3}$ and $\delta_1 = \frac{\delta}{4}$?

   (a) $\epsilon_{l+1} = \frac{3\epsilon_l}{4}$ and $\delta_{l+1} = \frac{4\delta_l}{5}$
   (b) $\epsilon_{l+1} = \frac{3\epsilon_l}{4}$ and $\delta_{l+1} = \frac{3\delta_l}{4}$
   (c) $\epsilon_{l+1} = \frac{2\epsilon_l}{3}$ and $\delta_{l+1} = \frac{3\delta_l}{4}$
   (d) $\epsilon_{l+1} = \frac{2\epsilon_l}{3}$ and $\delta_{l+1} = \frac{4\delta_l}{5}$

   Sol. (c)
   For getting (ϵ, δ)-PAC bound ,$\sum_l \epsilon_l \leq \epsilon$ and $\sum_l \delta_l \leq \delta$

9. Suppose we are facing a non-stationary bandit problem. We want to use posterior sampling for picking the correct arm. What is the likely change that needs to be done to the algorithm so that it can adapt to non-stationarity?

   (a) Update the posterior rarely.
   (b) Randomly shift the posterior drastically from time to time.
   (c) Keep adding a slight noise to the posterior to prevent its variance from going down quickly.

(d) No change is required.

Sol. (c)
It is necessary to prevent the posterior from collapsing, while simultaneously allowing it to learn from the current samples.

10. In median elimination method for (ϵ, δ)-PAC bounds, we claim that for every phase l, $Pr[A \leq B + \epsilon_l] > 1 - \delta_l$. ($S_l$ – is the set of arms remaining in the l^{th} phase)

Consider the following statements:

(i) A – is the maximum of rewards of true best arm in S_l, i.e. in l^{th} phase
(ii) B – is the maximum of rewards of true best arm in S_{l+1}, i.e. in $l + 1^{th}$ phase
(iii) B – is the minimum of rewards of true best arm in S_{l+1}, i.e. in $l + 1^{th}$ phase
(iv) A – is the minimum of rewards of true best arm in S_l, i.e. in l^{th} phase
(v) A – is the maximum of rewards of true best arm in S_{l+1}, i.e. in $l + 1^{th}$ phase
(vi) B – is the maximum of rewards of true best arm in S_l, i.e. in l^{th} phase

Which of the statements above are correct?

(a) i and ii
(b) iii and iv
(c) iii and iv
(d) v and vi
(e) i and iii

Sol. (a)
Refer Lemma 1 in the proof for the Median Elimination Algorithm