Online Learning - Assignment 2

Mohammad Mohammadi - 402208592

November 8, 2024

Assignment introduction:

This assignment is assignment # 2 of Online Learning course, instructed by Dr. Alishahi, Fall 2024, at Sharif University of Technology. The questions of this assignment are mostly from exercises of the book Bandit Algorithms, from Tor Lattimore and Csaba Szepesv ´ari.

Question 1

Problem: Show that the pseudo-regret $\bar{R}_n = \sum_{t=1}^{n} \Delta_{A_t}$ and the random regret $\hat{R}_n = n\mu^* - \sum_{t=1}^{n} X_t$ are both non skewed estimators for regret. Which of these two estimators is more accurate? Why?

Solution:
For the pseudo-regret we have:

$$\bar{R}_n = \sum_{t=1}^{n} \Delta_{A_t}$$

where $\Delta_{A_t} = \mu^* - \mu_{A_t}$ and A_t is the arm selected at time t.

Since Δ_{A_t} are deterministic differences based on the chosen arms, $\bar{R}_n$ is a sum of deterministic quantities. Therefore, the distribution of $\bar{R}_n$ does not have variability; it is a constant given the arm selections. And as a constant variable has zero skewness because there is no variability or asymmetry in its distribution. Hence , $\bar{R}_n$ is a non-skewed estimator.

For the random regret we have:

$$\hat{R}_n = n\mu^* - \sum_{t=1}^{n} X_t$$

where X_t is the reward obtained from the arm A_t at time t.

For the skewness of $\hat{R}_n$, we consider its distribution. The term $\sum_{t=1}^{n} X_t$ is a sum of random variables. Assuming that the rewards X_t are independent and identically distributed (i.i.d.) with finite third moments, the distribution of $\sum_{t=1}^{n} X_t$ tends towards a normal distribution as n increases, by the Central Limit Theorem. The normal distribution is symmetric and thus has zero skewness.

Therefore, for large n, $\hat{R}_n$ is approximately normally distributed with zero skewness, making it a non-skewed estimator in the asymptotic sense.

While both $\bar{R}_n$ and $\hat{R}_n$ are non-skewed estimators for regret, $\bar{R}_n$ is more accurate in the sense that it is deterministic and does not introduce variability. On the other hand, $\hat{R}_n$ involves randomness due to the stochastic nature of the rewards X_t. This randomness can lead to higher variance in the estimation of regret. Therefore, $\bar{R}_n$ provides a more precise estimate of regret compared to $\hat{R}_n$.

Question 2

Problem: Show that equation below implies the regret of an optimally tuned ETC for subgaussian two-armed bandits satisfies $R_n \leq \Delta + C\sqrt{n}$ where $C > 0$ is a universal constant.

$$R_n \leq \min\left\{n\Delta,\ \Delta + \frac{4}{\Delta}\left(1 + \max\left\{0,\ \log\left(\frac{n\Delta^2}{4}\right)\right\}\right)\right\}$$

Solution:
For $\Delta \leq \frac{1}{\sqrt{n}}$: since $R_n \leq n\Delta$, we obtain $R_n \leq \sqrt{n}$.
For $\Delta > \frac{1}{\sqrt{n}}$:

$$R_n \leq \Delta + \frac{4}{\Delta}\left(1 + \log_+\left(\frac{n\Delta^2}{4}\right)\right) \leq \Delta + 4\sqrt{n} + \max_{x>0}\frac{1}{x}\log_+\left(\frac{nx^2}{4}\right).$$

We know that we can rewrite as: $\max_{x>0}\frac{1}{x}\log_+\left(\frac{nx^2}{4}\right) = e^{-2}\sqrt{n}$.

Combining these results, we find that $R_n \leq \Delta + (4 + e^{-2})\sqrt{n}$, which is true regardless of the value of $\Delta > 0$.

Question 3

Try to modify the ETC algorithm so that it does not require knowledge of the suboptimality gap Δ. For simplicity, assume we have only two arms, each following 1-SubGaussian distributions. In the new algorithm, the level of exploration should not be predetermined and should depend on the obtained results. Using a confidence-bound logic, establish the following bound:

$$R_n \leq \Delta + \frac{C\log n}{\Delta}$$

Solution:

Consider that our ETC algorithm interacts with a two-armed, 1-subgaussian bandit $v \in \epsilon$, where the arms have means μ_1 and μ_2 (both in $\mathbb{R}$), and the gap is defined as $\Delta_v = |\mu_1 - \mu_2|$.

Let's take an arbitrary m:

$$\hat{\mu}_i(2m) - \mu_i \text{ is } \sqrt{\frac{1}{m}}\text{-subgaussian.}$$

Now, let's define the event $G = \{|\hat{\mu}_i(2m) - \mu_i| \leq \sqrt{2\log(n/\delta)/m}, i = 1, 2, m = 1, 2, \ldots, \lfloor n/2 \rfloor\}$. Using union bounds with $n \geq 2\lfloor n/2 \rfloor$, we obtain $\mathbb{P}(G) \geq 1 - \delta$.

Next, we introduce $w(m) = \sqrt{2\log(n/\delta)/m}$.

Define $M = \min\{1 \leq m \leq \lfloor n/2 \rfloor : |\hat{\mu}_1(2m) - \hat{\mu}_2(2m)| > 2w(m)\}$ (note that $M = \infty$ if the condition is never met).

Under the event G, if $M < +\infty$ and we assume $1 = \arg\max_i \hat{\mu}_i(2M)$, then $\mu_1 \geq \hat{\mu}_1(2M) - w(m) > \hat{\mu}_2(2M) + 2w(M) - w(M) \geq \mu_2$. Here, the first and last inequalities follow from the assumption that we are on G, while the middle inequality uses the stopping condition and our assumption that arm one has the highest mean at stopping. Thus,

$$R_n = \mathbb{P}(G^c)\frac{n\Delta}{2} + \mathbb{E}[M\mathbb{I}\{G\}]\frac{\Delta}{2} \leq \delta n + \mathbb{E}[M\mathbb{I}\{G\}]\frac{\Delta}{2}.$$

To proceed, we'll now find a bound on M given G. To simplify, let's assume $\mu_1 > \mu_2$. If G holds and $m < M$, then $2w(m) \geq |\hat{\mu}_1(2m) - \hat{\mu}_2(2m)| \geq \hat{\mu}_1(2m) - \hat{\mu}_2(2m) \geq (\mu_1 - w(m)) - (\mu_2 + w(m)) = \Delta - 2w(m)$. Rearranging, we get $4w(m) \geq \Delta$, which, based on the definition of $w(m)$, implies $m \leq (4/\Delta)^2 2\log(n/\delta)$.

Therefore, on G, we have $M = 1 + \max\{m : 2w(i) \geq |\hat{\mu}_1(2i) - \hat{\mu}_i(2i)|, i = 1, 2, \ldots, m\} \leq 1 + (4/\Delta)^2 2\log(n/\delta)$.

Substituting this and setting $\delta = 1/n$, we conclude that

$$R_n \leq \Delta + \frac{16}{\Delta}\log(n)$$

Which is essentially what we wanted with C=16.

Question 4

Problem: For this exercise assume the rewards are 1-subgaussian and there are $k \geq 2$ arms. The ε -greedy algorithm depends on a sequence of parameters $\varepsilon_1, \varepsilon_2, \ldots$. First it chooses each arm once and subsequently chooses $A_t = \arg\max_i \hat{\mu}_i(t-1)$ with probability $1 - \varepsilon_t$ and otherwise chooses an arm uniformly at random.

Prove that if $\epsilon_t = \epsilon > 0$, then

$$\lim_{n\to\infty} \frac{R_n}{n} = \frac{\epsilon}{k}\sum_{i=1}^{k} \Delta_i.$$

Solution:

Given $\varepsilon_t = \varepsilon > 0$ for all t, the algorithm selects the optimal arm with probability $1 - \varepsilon$ and explores uniformly with probability ε.

The expected regret at each time step t can be decomposed into:

$$\mathbb{E}[R_t] = \varepsilon \cdot \mathbb{E}[\Delta_{A_t}]$$

Since exploration is uniform over k arms,

$$\mathbb{E}[\Delta_{A_t}] = \frac{1}{k}\sum_{i=1}^{k}\Delta_i$$

Thus,

$$\mathbb{E}[R_t] = \varepsilon \cdot \frac{1}{k}\sum_{i=1}^{k}\Delta_i$$

Summing over n time steps,

$$\mathbb{E}[R_n] = n \cdot \frac{\varepsilon}{k}\sum_{i=1}^{k}\Delta_i$$

Taking the limit,

$$\lim_{n\to\infty}\frac{R_n}{n} = \frac{\varepsilon}{k}\sum_{i=1}^{k}\Delta_i$$

Therefore,

$$\lim_{n\to\infty}\frac{R_n}{n} = \frac{\epsilon}{k}\sum_{i=1}^{k}\Delta_i.$$

Question 5

Problem: From definitions we have the pseudo-regret as:

$$\bar{R}_n = \sum_{t=1}^{n}\Delta_{A_t}.$$

The UCB policy in its algorithm depends on confidence parameter $\delta \in (0,1]$ that determines the level of optimism. State and prove a bound on the pseudo-regret of this algorithm that holds with probability $1 - f(n,k)\delta$, where $f(n,k)$ is a function that depends on n and k only. More precisely show that for bandit $\nu \in \mathcal{E}^k_{\mathrm{SG}}(1)$ that

$$\mathbb{P}\left(\bar{R}_n \geq g(n,\nu,\delta)\right) \leq f(n,k)\delta,$$

where g and f should be as small as possible (there are trade-offs – try and come up with a natural choice).

Solution:

Consider the UCB algorithm with confidence parameter δ. Let $\Delta_i = \mu^* - \mu_i$ for each suboptimal arm i. We aim to bound the pseudo-regret $\bar{R}_n$.

Using the properties of 1-subgaussian rewards and applying the Chernoff-Hoeffding inequality, we have that for each arm i, the probability that the UCB algorithm overestimates the mean reward is bounded by δ/k. Applying the union bound over all k arms, the probability that any arm's confidence bound fails is at most δ.

The number of times a suboptimal arm i is pulled can be bounded by:

$$T_i(n) \leq \left\lceil \frac{4\ln(n)}{\Delta_i^2} \right\rceil$$

Thus, the pseudo-regret can be bounded as:

$$\bar{R}_n \leq \sum_{i=1}^{k} \Delta_i T_i(n) \leq \sum_{i=1}^{k} \frac{4\ln(n)}{\Delta_i}$$

Therefore, setting $f(n,k) = k$ and $g(n,\nu,\delta) = \sum_{i=1}^{k} \frac{4\ln(n)}{\Delta_i}$, we obtain:

$$\mathbb{P}\left(\bar{R}_n \geq \sum_{i=1}^{k} \frac{4\ln(n)}{\Delta_i}\right) \leq k\delta$$

Hence, for the UCB algorithm, the pseudo-regret satisfies

$$\mathbb{P}\left(\bar{R}_n \geq g(n,\nu,\delta)\right) \leq f(n,k)\delta$$

where $f(n,k) = k$ and $g(n,\nu,\delta) = \sum_{i=1}^{k} \frac{4\ln(n)}{\Delta_i}$.