Reinforcement Learning WS 2024
TAs: René Geist and Mikel Zhobro
Lecturer for this topic: Claire Vernade
Due-date: **19.11.2024, 14:00** (upload to ILIAS as one file)
Filename: homework5-NAME1-NAME2-...zip
Homework Sheet 5
November 11, 2024

This short exercise sheet will guide you through the implementation of bandit algorithms for the simple K-armed bandit problem seen in class. We provide a notebook with code that has to be completed.

1 Code attached

- Notebook: `MultiarmedBandits-TBC.ipynb`

Questions are divided into either practical Coding and Theory and are overall independent so you can address them as you please. The bonus section helps you derive and implement Thompson Sampling, a clever Bayesian way of dealing with the explore-exploit dilemma. Additional material in the slides of the class may help you with the theory questions.

2 Questions

Coding question 1: In the second cell of the notebook, complete the bandit environment by implementing a Gaussian reward generator and a Bernoulli reward generator[1].

Coding question 2: The first bandit agent is ϵ-Greedy. Recall that at each round, ϵ-Greedy explore uniformly at random with probability ϵ and *exploits* with probability $1-\epsilon$. Implement the `get action` function accordingly. Observe how the policy updates its own counts for each arm, and then only updates its reward counters when the `receive reward` function is called: the agent cannot call the environment by itself (no simulations allowed) because each call to the reward generator will count in our regret...

Theory question 1: : **Linear regret for ϵ-Greedy**.
We saw in class that

$$\mathcal{R}_\nu(T) \geq \epsilon \frac{K}{K-1} \Delta_{\min} T$$

Can you prove it by bounding from below $\mathbb{E}[N_a(T)]$ for each suboptimal arm $a \in [K] \backslash \{a^*\}$? Recall that $\Delta_{\min} = \min_{a \neq a^*} \Delta_a$, sum your result over suboptimal arms to obtain the result.

[1] Note that the means are given as input, and the variance too but is fixed to 1 for now.

Theory question 1: : **Explore-Then-Commit (ETC)**.
ETC is a simpler version of ϵ-Greedy, where all the exploration rounds are 'gathered' at the beginning. Fix an exploration parameter $m < T$ and pull each arm m times. At the end of these $m \times K$ rounds, compute

$$\forall k \in [K],\ \hat{\mu}_k = \frac{\textbf{sum of rewards from arm } k}{m}$$

and choose arm $\hat{a} \arg\max_k \hat{\mu}_k$. Then, *commit* to arm $\hat{a}$ for the remaining $T - mK$ rounds.
The goal of this exercise is to prove an upper bound on the regret of ETC.
Recall that for a suboptimal arm $a \in [K]$, we define

$$\mathbb{E}[N_a(T)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1} A_t = a\right] = \sum_{t=1}^{T} \mathbb{P}(A_t = a)$$

(a) Can you write $\mathbb{E}[N_a(T)]$ as a function of m, K and $\mathbb{P}(\hat{a} = a)$, i.e. the probability that the chosen arm is wrong?

(b) Notice that
$$\mathbb{P}(\hat{a} = a) \leq \mathbb{P}(\hat{\mu}_a \geq \hat{\mu} a^\star).$$
Choosing action a means that $a^\star$ is not chosen. The "bad" event $\{\hat{a} \neq a^\star\}$ happens only when the empirical mean of arm a is larger than that of arm $a^\star$ after collecting exactly m samples from each of them. How likely is this event to happen? Can you use Hoeffding's inequality to upper bound the probability of this event?

(c) How can you choose m to minimize the upper bound you just proved?

Coding question 3: Implement Explore-Then-Commit (ETC) and compare it with ϵ-greedy. You can choose the number of arms, the gaps between them, and the noise model (you implemented two: Gaussian and Bernoulli).
Coding question 4: Implement UCB as seen in class and compare it with your two other baselines.

3 Thompson Sampling – Bonus

This section is optional. We will see how Thompson Sampling can be implemented as a nice alternative to UCB.
Theory question TS1: Assume $X_1, ..., X_t \sim \mathcal{N}(\theta, \sigma^2)$. You want to estimate θ (unknown, but σ^2 is assumed to be known) using Bayesian statistics. You choose a prior distribution $\pi_0 = \mathcal{N}(0, \sigma^2)$. Can you compute the posterior (of the form $\pi_t(\theta) = \mathcal{N}(?, ?)$) using Bayes' rule?
Theory question TS2: Now I give you just one more $X_{t+1} \sim \mathcal{N}(\theta, \sigma^2)$. Can you write a simple update of your posterior above to include this new piece of evidence?

Coding question TS1: Thompson sampling is a very simple algorithm: at each round t, sample $\tilde{\theta}_k \sim \pi_{t,k}$, the current posterior on the mean of arm k (as computed above). This sample *hallucinates* the true mean and you can use it to make decisions: $A_t = \arg\max_k \tilde{\theta}_k$. Implement this algorithm as a new agent and test it against UCB. What do you think?