9 Multi-Armed Bandit

Exercise 9.1

Tell if the following statements about MAB are true or false and provide motivations for your choice.

1. The MDP corresponding to a MAB setting has no state.
2. The MDP corresponding to a MAB setting has no transition probabilities.
3. An algorithm which always uses the greedy choice (maximize the empirical expected reward) might get stuck to suboptimal solutions.
4. We are able to solve the exploration/exploitation problem by considering either upper or lower bounds on the expected rewards.
5. In a frequentist framework, if we base our sequential policy on a MAB by considering only $\hat{R}(a_i)$ we are not using all the information we have about the estimates.
6. Uncertainty over quantities is usually handled by explorative choices in the MAB setting.

Exercise 9.2

Tell if the following problems can be modeled as MAB and explain why.

1. Maze escape;
2. Pricing goods with stock;
3. Pricing goods without stock;
4. Optimal bandwidth allocation (send the largest amount of information without congesting the band);
5. Web ads placement;
6. Weather prediction (with multiple experts).

If they fit the MAB setting, is the environment adversarial or stochastic?

Exercise 9.3

Provide an example for which the pure exploitation strategy is failing to converge to the optimum in a MAB setting with Bernoulli rewards. Recall that the pure exploitation algorithm chooses the arm by selecting the one with the largest:

$$\hat{R}_t(a_i) = \frac{1}{N_t(a_i)} \sum_{j=1}^{t} r_{i,j} \, \mathbb{I}\left\{a_i = a_{i_j}\right\},$$

where $\mathbb{I}$ is the indicator function.

How often does this occurs? Can we compute an lower bound over the regret of this strategy?

Exercise 9.4

The ε-greedy algorithm selects the best action except for small percentages of times ε, where all the actions are considered. Consider a MAB stochastic setting.

1. Is this algorithm converging to the optimal strategy (in some sense)?
2. If not, propose a scheme which has the chance of converging to the optimal solution.
3. Are we in a MAB perspective if we are using this algorithm?

Exercise 9.5

Write the formula for the minimum regret we might have on average over $T = \lceil e^{10} \rceil$ time steps in the case we have a stochastic MAB with 3 arms and expected rewards:

$$R(a_1) = 0.2 \tag{9.1}$$
$$R(a_2) = 0.4 \tag{9.2}$$
$$R(a_3) = 0.7 \tag{9.3}$$

and each distribution $\mathcal{R}(a_i)$ is Bernoulli.

Note that the KL divergence for Bernoulli variables with means p and q is:

$$KL(p, q) = p \log\left(\frac{p}{q}\right) + (1 - p) \log\left(\frac{(1 - p)}{(1 - q)}\right).$$

Hints: $\log(\frac{0.2}{0.7}) = -1.25$, $\log(\frac{0.8}{0.3}) = 0.98$, $\log(\frac{0.4}{0.7}) = -0.56$ and $\log(\frac{0.6}{0.3}) = 0.69$.

Is it possible that your algorithm achieves lower regret? If so, provide an example.

Exercise 9.6

Provide examples of either Bayesian or frequentist MAB algorithm showing the following properties:

1. It incorporates expert knowledge about the problem in the arms;
2. It provides tight theoretical lower bound on the expected regret in the stochastic setting;
3. It provides tight theoretical upper bound on the expected regret in the stochastic setting;
4. At each turn, it modifies only the statistics of the chosen arm.

Motivate your answers.

Exercise 9.7

Consider the following bounds (supposed to hold with probability at least δ) for a MAB stochastic setting:

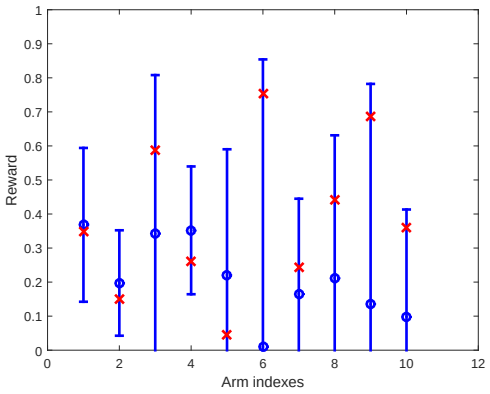

Figure 9.1: Bounds in a MAB with $N = 10$.

where the blue bars are the Hoeffding bounds for the expected rewards, the blue circles are the estimated expected rewards, and the red crosses are the real expected rewards.

1. Which arm would a UCB1 algorithm choose for the next round?
2. Do you think that Figure 9.1 might be the results obtained by running UCB1 for several rounds?

3. Which arm will UCB1 converge to if $T \to \infty$?
4. Which arm is the one which we pulled the most so far?

Motivate your answers.

Exercise 9.8

Consider the Thompson Sampling algorithm. Assume to have the following posterior distributions $Beta_i(\alpha_t, \beta_t)$ for arms $\mathcal{A} = \{a_1, \ldots, a_5\}$ rewards, which are distributed as Bernoulli r.v.:

$$
\begin{aligned}
a_1 &: \quad \alpha_t = 1 \quad \beta_t = 5 \\
a_2 &: \quad \alpha_t = 6 \quad \beta_t = 4 \\
a_3 &: \quad \alpha_t = 11 \quad \beta_t = 23 \\
a_4 &: \quad \alpha_t = 12 \quad \beta_t = 33 \\
a_5 &: \quad \alpha_t = 28 \quad \beta_t = 21
\end{aligned}
$$

From these distribution you extract the following samples for the current round:

$$
\begin{aligned}
\hat{r}(a_1) &= 0.63 \\
\hat{r}(a_2) &= 0.35 \\
\hat{r}(a_3) &= 0.16 \\
\hat{r}(a_4) &= 0.22 \\
\hat{r}(a_5) &= 0.7
\end{aligned}
$$

1. Which arm would the TS algorithm play for the next round?
2. What changes if the real distributions of the arm rewards are Gaussian.
3. Assume we started the TS algorithm with uniform $Beta(1, 1)$ priors. What would UCB1 have chosen in the case of Bernoulli rewards for the next round?