CS 234: Assignment #3

1 Best Arm Identification in Multiarmed Bandit (35pts)

In this problem we focus on the Bandit setting with rewards bounded in $[0, 1]$. A Bandit problem instance is defined as an MDP with just one state and action set $\mathcal{A}$. Since there is only one state, a "policy" consists of the choice of a single action: there are exactly $A = |\mathcal{A}|$ different deterministic policies. Your goal is to design a simple algorithm to identify a near-optimal arm with high probability.
We recall Hoeffding's inequality below, where $\overline{x}$ is the expected value of a random variable, $\widehat{x}$ is the sample mean (under the assumption that the random variables are in the interval [0,1]) n is the number of samples and $\delta > 0$ is a scalar:

$$\Pr\left(|\widehat{x} - \overline{x}| > \sqrt{\frac{\log(2/\delta)}{2n}}\right) < \delta. \tag{1}$$

Assuming that the rewards are bounded in $[0, 1]$, we propose this simple strategy: allocate an identical number of samples $n_1 = n_2 = ... = n_A = n_{des}$ to every action and return the action with the highest average payout $\widehat{r}_a$. The purpose of this exercise is to study the number of samples required to output an arm that is at least ϵ-optimal with high probability. Intuitively, as n_{des} increases the empirical average of the payout $\widehat{r}_a$ converges to its expected value $\overline{r}_a$ for every action a, and so choosing the arm with the highest empirical payout $\widehat{r}_a$ corresponds to approximately choosing the arm with the highest expected payout $\overline{r}_a$.

(a) (15 pts) We start by defining a "good event". Under this "good event" the empirical mean of each arm is not too far from its expected return. Starting from Hoeffding inequality with n_{des} samples allocated to every action show that:

$$\Pr\left(\exists a \in \mathcal{A} \quad s.t. \quad |\widehat{r}_a - \overline{r}_a| > \sqrt{\frac{\log(2/\delta)}{2n_{des}}}\right) < A\delta. \tag{2}$$

In other words, the "bad event" is that at least one arm has an empirical mean that differs significantly from its expected value and this has probability at most $A\delta$.

Solution By union bound:

$$\Pr\left(\exists a \in \mathcal{A} \quad s.t. \quad |\widehat{r}_a - \overline{r}_a| > \sqrt{\frac{\log(2/\delta)}{2n_{des}}}\right) \tag{3}$$

$$= \Pr\left(\left(|\widehat{r}_1 - \overline{r}_1| > \sqrt{\frac{\log(2/\delta)}{2n_{des}}}\right) \cup \dots \cup \left(|\widehat{r}_A - \overline{r}_A| > \sqrt{\frac{\log(2/\delta)}{2n_{des}}}\right)\right) \tag{4}$$

$$\leq \Pr\left(|\widehat{r}_1 - \overline{r}_1| > \sqrt{\frac{\log(2/\delta)}{2n_{des}}}\right) + \dots + \Pr\left(|\widehat{r}_A - \overline{r}_A| > \sqrt{\frac{\log(2/\delta)}{2n_{des}}}\right) < A\delta. \tag{5}$$

(b) (20 pts) After pulling each arm (action) n_{des} times our algorithm returns the arm with the highest empirical mean:

$$a^\dagger = argmax_a \widehat{r}_a \tag{6}$$

Notice that $a^\dagger$ is a random variable. Define as $a^\star$ the optimal arm (that yields the highest average reward $a^\star = argmax_a \overline{x}_a$). Suppose that we want our algorithm to return at least an ϵ optimal arm with probability $1-\delta'$, as follows:

$$\Pr\left(\overline{r}_{a^\dagger} \geq \overline{r}_{a^\star} - \epsilon\right) \geq 1 - \delta'. \tag{7}$$

How many samples are needed to ensure this? Express your result as a function of the number of actions, the required precision ϵ and the failure probability δ'.

Solution Notice that if

$$\sqrt{\frac{\log(2A/\delta)}{2n_{des}}} < \epsilon/2 \tag{8}$$

then

$$\Pr\left(\forall a \in \mathcal{A} \quad |\widehat{r}_a - \overline{r}_a| \leq \epsilon/2\right) \geq 1 - \delta' \tag{9}$$

directly from part (a). Under this "good" event, the choice $a^\dagger = argmax_a \widehat{r}_a$ ensures

$$\overline{r}_{a^\dagger} \geq \widehat{r}_{a^\dagger} - \epsilon/2 \geq \widehat{r}_{a^\star} - \epsilon/2 \geq \overline{r}_{a^\star} - \epsilon. \tag{10}$$

Solving equation 8 provides the desired number of samples:

$$n_{des} \geq \frac{2log(2A/\delta)}{\epsilon^2} \tag{11}$$

which yields a final sample complexity of:

$$\frac{2Alog(2A/\delta)}{\epsilon^2} \tag{12}$$