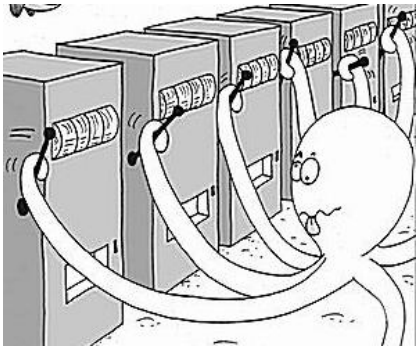
DS 102 Discussion 10

Wednesday, April 20, 2022

1. **Multi-Armed Bandits and the UCB Algorithm**

In the multi-armed bandits setting, we consider a decision-maker who is given K options to choose from. We refer to these options as "arms". Associated with each arm is a probability distribution over rewards. Initially, this distribution is unknown to the decision-maker. The decision-maker chooses an arm, usually referred to as pulling an arm, and receives a reward sampled from the corresponding reward distribution. This process is repeated over and over again.



Let's begin by setting up some mathematical notation. Suppose you have a set of K "arms", $\mathcal{A} = \{1, 2, ..., K\}$. Each arm $a \in \mathcal{A}$ has its own reward distribution $X_a \sim \mathbb{P}_a$ with mean $\mu_a = \mathbb{E}[X_a]$. Define the number of times arm a has been pulled up to and including time t as $T_a(t)$. In these problems we do not know μ_a but we would like to efficiently find the arm with the maximum mean by creating an algorithm that balances *exploration* of the arms with *exploitation* of the best possible arm. The efficiency of the algorithm is measured by a theoretical quantity known as regret, which measures how well the algorithm performs in expectation against an 'oracle' that knows the means of all the arms and always pulls the arm with highest mean. In this discussion, we will study the derivation of the UCB algorithm for the multi-armed bandit problem, which uses the following bound:

$$\mathbb{P}\left[\mu_a < \hat{\mu}_{a,T_a(t)} + C_a(T_a(t), \delta)\right] > 1 - \delta$$

where $C_a(T_a(t), \delta)$ is the upper confidence bound.

To derive this bound, we will make use of Hoeffding's Inequality, which tells us that if random variables $X_1, X_2, ..., X_n$ are independent and bounded between $[a, b]$, then

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq \epsilon\right] \leq \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

(a) *Applying Hoeffding's Inequality for Bounded Random Variables*

Suppose that you know that the reward of any arm is between 0 and 1: $X_a \in [0, 1]$. Find a bound on the difference between $\hat{\mu}_{a,T_a(t)}$ and μ_a using Hoeffding's Inequality.

(b) *Connecting Hoeffding's Inequality to UCB*

Use the inequality derived in Part (a) to show that

$$\mathbb{P}\left[\hat{\mu}_{a,T_a(t)} - \mu_a \leq -C_a(T_a(t), \delta)\right] \leq e^{-2T_a(t)C_a(T_a(t),\delta)^2}$$

(c) *Solving for the Upper Confidence Bound*

Find the value of $C_a(T_a(t), \delta)$ in terms of δ so that $\mathbb{P}\left[\mu_a < \hat{\mu}_{a,T_a(t)} + C_a(T_a(t), \delta)\right] > 1 - \delta$ holds. Plug this result into the inequality to derive the upper confidence bound.

(d) *Arm Selection with UCB*

Suppose we set $\delta = \frac{1}{t^3}$. This controls the probability that the true mean μ_a is greater than our upper confidence bound $C_a(T_a(t), \delta)$ on the estimated mean $\hat{\mu}_{a,T_a(t)}$. What rule does the UCB algorithm use to choose an arm A_t at each iteration t?

2. **Regret of Explore-Then-Commit**

Now, we will analyze the regret of the Explore-then-Commit algorithm for the multi-armed-bandit (MAB) problem. We consider a stochastic MAB problem with a set of $K = 2$ arms $\mathcal{A} = \{1, 2\}$. Recall that each arm $A \in \mathcal{A}$ is associated with a reward distribution $X_A \sim \mathbb{P}_A$, with corresponding mean $\mu_A = \mathbb{E}[X_A]$. **We will assume throughout this problem that the first arm has higher average reward, i.e. $\mu_1 > \mu_2$.** At each round $t = 1, \ldots, n$ our algorithm chooses an arm $A_t \in \mathcal{A}$ and receives a corresponding reward $X_{A_t}^{(t)} \sim \mathbb{P}_{A_t}$, independent of all previous rewards.

If we knew arm 1 has higher average reward, we would choose $A_t = 1$ each round in order to maximize the expected total reward. In practice, however, we do not know which arm is better since the means $\{\mu_1, \mu_2\}$ are unknown. The expected reward of our algorithm will always be less than $n\mu_1$, and we quantify the price we pay for not knowing the better arm via the *regret*

$$R_n := n\mu_1 - \mathbb{E}\left[\sum_{t=1}^{n} X_{A_t}^{(t)}\right]$$

In this problem we analyze the regret of the explore-then-commit (ETC) algorithm, which is outlined as follows:

Algorithm 1 Explore-then-Commit (ETC) Algorithm

Inputinput Number of initial pulls c per arm $t = 1, \ldots, cK$: Choose arm $A_t = (t \bmod K) + 1$
Let $\hat{A} \in \{1, \ldots, K\}$ denote the arm with the highest average reward so far
$t = cK + 1, cK + 2, \ldots, n$: Choose arm $A_t = \hat{A}$

As we can observe from the algorithm above, ETC proceeds in two phases. In the exploration phase, each arm $A \in \mathcal{A}$ is pulled c times in order to produce an estimate $\hat{\mu}_A = \frac{1}{c}\sum_{t \le cK : A_t = A} X_A^{(t)}$ of the mean reward for that arm. In the commit phase, i.e. for every $t > cK$, we choose $A_t = \hat{A}$, where $\hat{A} := \text{argmax}_{A \in \mathcal{A}} \hat{\mu}_A$ is the apparent best arm at the end of the exploration phase. In the first part of our analysis, we evaluate the probability that we incorrectly identify arm 2 as the best arm, i.e. $\mathbb{P}(\hat{A} = 2)$.

(a) *Bounding the Chance of Selecting a Sub-optimal Arm*

Assume each reward is in the unit interval $[0, 1]$, i.e. $0 \le X_A \le 1$ for $A \in \{1, 2\}$. Show that

$$\mathbb{P}\left(\hat{A} = 2\right) \le \exp\left(-\frac{c\Delta^2}{2}\right),$$

where $\Delta = \mu_1 - \mu_2$.

(b) *Regret Decomposition*

Let m denote the number of times arm 2 has been pulled, up to and including time n. Show

$$R_n = \Delta \mathbb{E}[m]$$

Hint: Start from the following:

$$\begin{aligned} R_n &:= n\mu_1 - \mathbb{E}\left[\sum_{t=1}^{n} X_{A_t}^{(t)}\right] = \mathbb{E}\left[\sum_{t=1}^{n} \left(\mu_1 - X_{A_t}^{(t)}\right)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\{A_t = 1\}\left(\mu_1 - X_1^{(t)}\right)\right] + \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\{A_t = 2\}\left(\mu_1 - X_2^{(t)}\right)\right]. \end{aligned}$$

Note also that for all t, A_t is independent of $X_A^{(t)}$ for $A \in \{1, 2\}$.

(c) *Finding the Expected number of Sub-optimal Pulls*
Show that if $n > 2c$, then:

$$\mathbb{E}[m] = c + (n - 2c)\mathbb{P}\left(\hat{A} = 2\right)$$

(d) *Bounding the Regret of ETC*
Show that:

$$R_n \leq \Delta\left(c + (n - 2c)\exp\left(-\frac{c\Delta^2}{2}\right)\right)$$

(e) *Analyzing the Regret of ETC*

Suppose you knew the sub-optimality gap Δ. Solve for a value of c which guarantees that:

$$\exp\left(-\frac{c\Delta^2}{2}\right) \leq \frac{1}{n}$$

For this number of exploratory pulls c, what is the upper bound on the regret from Part (d)? Your answer should be in terms of n and Δ. Does this bound grow linearly in n, or does it do better (i.e. is it sublinear)?