AI-3001: Advanced Topics in Machine Learning **Aug-Nov 2022**

Lecture 14: Bandit Optimization

Lecturer: Ganesh Ghalme *Scribes: Ganesh Ghalme*

14.1 Stochastic MAB problem

We will study the classical version of the Stochastic MAB problem. In this, we have K arms. We consider K to be finite. Each arm i has an associated probability distribution $\mathcal{D}_i(\mu_i)$. Here, μ_i is the mean of $\mathcal{D}_i$.

The reward distributions are unknown to the algorithm. Each time an arm i is selected (or pulled) an algorithm obtained a reward sampled independently from (an unknown) distribution $\mathcal{D}_i$. The algorithm must carefully balance the exploration (pulling different arms to learn their expected reward) and exploitation (pulling arms having largest average reward according to available estimates).

We will consider that the probability distribution $\mathcal{D}_i$ is bounded and fixed. Let $\mathcal{D} = \times_{i=1}^{K} \mathcal{D}_i$ be the product distribution. We will also consider that the reward distributions are not related across arms i.e. the unstructured reward setting. One way to think about it is to assume that the reward distribution parameters μ_is are sampled independently from some distribution. However, the restrictions on the number of arms as well as the boundedness of reward distributions can be relaxed and study of specialized algorithms for those settings is beyond our scope. The stichastic MAB algorithm follows the below template

- At each time $t = 1, 2, 3, \cdots$
  1. ALG pulls an arm i_t
  2. Environment samples $r_t \sim \mathcal{D}$
  3. Environment reveals $r_{i_t,t}$ to the algorithm

14.2 Regret

So far, we studied the following notion of regret in adversarial setting(s).

$$\mathcal{R}_T(\textsc{Alg}) = \mathbb{E}[\max_i \sum_{t=1}^{T} r_{i,t} - \sum_{t=1}^{T} r_{i_t,t}] \tag{14.1}$$

However, this notion is seldom used in stochastic MAB setting. A primary reason for this is that the arm $i^* = \arg\max_i \mu_i$ may not be the same as the arm $i^{\#} = \arg\max_i \sum_{t=1}^{T} r_{i,t}$. To see this, consider a two armed bandits setting with $\mu_1 = 0.5 + \varepsilon$ and $\mu_1 = 0.5 - \varepsilon$ and let the reward realizations be $(1, 0, 0)$ and $(0, 1, 1)$ for arms 1 and 2 respectively. Note that for $T = 1$ the we have $i^{\#} = 1$ whereas for $T = 3$ we have $i^{\#} = 2$ whereas $i^* = 1$ for all stopping times T. Hence for stochastic MAB we will use the notion of pseudo-regret

given as

$$\begin{aligned}\mathcal{R}_T(\textsc{Alg}) &= \max_i \mathbb{E}[\sum_{t=1}^{T} r_{i,t} - \sum_{t=1}^{T} r_{i_t,t}] \\ &= \sum_{t=1}^{T} \mu_{i^*} - \mathbb{E}[\sum_{t=1}^{T} r_{i_t,t}] && \text{(Here, } i^* = \arg\max_i \mu_i\text{)} \\ &= \mu_{i^*} T - \mathbb{E}[\sum_{t=1}^{T} r_{i_t,t}]\end{aligned}$$

Lemma 14.1. *Let $\Delta_i = \mu_{i^*} - \mu_i$ be the optimality gap of arm i and $N_{i,t}$ be the number of pulls of arm i till (and including) time t then $\mathcal{R}_T(\textsc{Alg}) = \sum_{i=1}^{K} \Delta_i \cdot \mathbb{E}[N_{i,T}]$*

The proof of this lemma is left as an exercise.

14.3 Preliminaries

Recall from previous lecture the following.

Definition 14.2 (Markov's Inequality)**.** *Let X be a non-negative random variable. We have*

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{\mathbb{E}[x]}{\varepsilon} \tag{14.2}$$

Definition 14.3 (Chebyshev's Inequality)**.** *Let X be a random variable with finite second moment,*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathtt{Var}(X)}{\varepsilon^2} \tag{14.3}$$

Definition 14.4 (Hoeffding's Lemma)**.** *Let X be any random variable with bounded support in the interval $[a, b]$ and let $\eta \in \mathbb{R}$ then we have*

$$\mathbb{E}[e^{\eta[X - \mathbb{E}[X]]}] \leq e^{\frac{\eta^2 \mathtt{Var}(X)}{2}} \tag{14.4}$$

14.4 Deriving Hoeffding's Inequality

First let us look at the Chebyshev's Inequality given above. Let $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the r.v. denoting the average of the iid r.v.s $X_1, X_2, \cdots X_n$. Furthermore, observe that $\mathtt{Var}(A_n) = \mathtt{Var}(X_1)/n$ hence we have

$$\mathbb{P}(|A_n - \mathbb{E}[A_n]| \geq \varepsilon) \leq \frac{\mathtt{Var}[X_1]}{n\varepsilon^2}$$

This already gives us some convergence of random variable around its mean. However the convergence rate is $1/n$ and this can be further improved. Towards that we will first introduce α-subgaussian random variables.

Definition 14.5 (α-subgaussian Random Variables)**.** *Let a a random variable X be such that $\mathbb{E}[X] = 0$. We call X, σ-subgaussian if for all $\eta \in \mathbb{R}$ we have*

$$\mathbb{E}[e^{\eta X}] \leq e^{\eta^2 \sigma^2 / 2} \tag{14.5}$$

We first notice that the RHS is the moment generating function of the mean zero Gaussian random variable with variance σ^2. This implies that the tail of subgaussian random variables falls of atleast as fast as that of the Gaussian random variable (which is exponential). This allows us to obtain the following upper bound on the probability of deviation of a random variable from its mean.

We now prove the following lemma for σ-subgaussian random variables.

Lemma 14.6. *Let X be a zero mean σ-sub Gaussian random variable. Then for any $\varepsilon > 0$ we have*

$$\mathbb{P}(X \geq \epsilon) \leq e^{-\varepsilon^2/2\sigma^2}$$
$$\mathbb{P}(X \leq -\epsilon) \leq e^{-\varepsilon^2/2\sigma^2}$$

Proof. We will begin with a common Cremer-Chernoffs method that allows us to get the bounds in terms of moment generating functions. Observe

$$\mathbb{P}(X \geq \varepsilon) = \mathbb{P}(e^{\eta X} \geq e^{\eta\varepsilon})$$

The above trick is called Cremer-Chernoff method and is used commonly to obtain exponential trail bounds. We use $\eta > 0$ above.

$$\begin{aligned}\mathbb{P}(X \geq \varepsilon) &\leq \frac{\mathbb{E}[e^{\eta X}]}{e^{\eta\varepsilon}} && \text{(Markov's Inequality)}\\ &\leq e^{\eta^2\sigma^2/2 - \eta\varepsilon}\end{aligned}$$

we now optimize over $\eta > 0$ to obtain $\eta^* = \varepsilon/\sigma^2$. This gives

$$\mathbb{P}(X \geq \varepsilon) \leq e^{-\varepsilon^2/2\sigma^2}$$

The second inequality follows in the same manner. Recall that the definition of the subgaussianity requires the inequality to hold for all $\eta \in \mathbb{R}$. □

Lemma 14.7 (Hoeffding's Inequality for bounded r.v.'s)**.** *Let $X_1, X_2, \cdots X_n$ be iid random variables taking values in $[0, 1]$ and let $X = \frac{1}{n}\sum_{i=1}^n X_i$. Then we have,*

$$\mathbb{P}(X - \mathbb{E}[X] \geq \varepsilon) \leq e^{-2n\varepsilon^2}$$
$$\mathbb{P}(X - \mathbb{E}[X] \leq -\varepsilon) \leq e^{-2n\varepsilon^2}$$

Proof. We will prove one tail bound. The proof for the other bound is similar. As in the previous lemma we use Cremer-Chernoff's method to obtain

$$\begin{aligned}\mathbb{P}(X - \mathbb{E}[X] \geq \varepsilon) &= \mathbb{P}(e^{\eta(X - \mathbb{E}[X]} \geq e^{\eta\varepsilon})\\ &\leq \frac{\mathbb{E}(e^{\eta(X-\mathbb{E}[X]})}{e^{\eta\varepsilon}} && \text{(Markov's Inequality)}\\ &\leq e^{\eta^2/8n - \eta\varepsilon} && \text{(Hoeffding's Lemma)}\end{aligned}$$

Minimize the RHS to obtain $\eta^* = 4n\varepsilon$ to get optimal RHS bound.

$$\mathbb{P}(X - \mathbb{E}[X] \geq \varepsilon) \leq e^{-2n\varepsilon^2}$$

□

14.5 Initial attempts to obtain sublinear (pseudo-) regret guarantee

14.5.1 Exploration Separated Algorithm

Solutions to MAB problem involve designing strategies to trade-off between exploration (pulling the arm that has not been explored enough number of times) and exploitation (pulling the arm that has provided best reward so far). Exploration separated algorithms give a simple strategy for this trade-off wherein all the arms are pulled in round robin fashion for ϵT number of rounds and then, the arm with best empirical reward so far is pulled for the rest of the $(1-\epsilon)T$ number of rounds. Here, the parameter ϵ is set to minimize the regret.

For simplicity of exposition let us consider that the rewards distributions are Bernoulli.

Algorithm 1: Exploration Separated Algorithm

Input: Time horizon T, exploration parameter ϵ, number of arms K
Initialize: $t = 1, S_i = 0, N_i = 0$

- **for** $t = 1, 2, \cdots, \lfloor \frac{\epsilon T}{K} \rfloor K$ **do**
  - $i_t = t \mod (K) + 1$
  - $N_{i_t} = N_{i_t} + 1$
  - $S_i = S_i + \mathbb{1}(r_{i_t,t} = 1)$
- Let $\hat{\mu}_i = \frac{S_i}{N_i}$.
- **for** $t = \lfloor \frac{\epsilon T}{K} \rfloor K + 1, \cdots, T$ **do**
  - $i_t = i$ if $i = \arg\max_i \widehat{\mu}_i$

Let N_i denote the number of exploration rounds each agent faces. Thus, $N_i = \lfloor \frac{\epsilon T}{K} \rfloor \ \forall i = 1, 2, \cdots K$. Let $c_i = \sqrt{\frac{2 \ln T}{N_i}}$ and $j = \arg\max_i \hat{\mu}_i$. Also denote $i^* = \arg\max_i \mu_i$ and $\mu^* = \max_i \mu_i$ as the index and the mean reward of the optimal arm respectively. For notational convenience, we assume $\lfloor \frac{\epsilon T}{K} \rfloor = \frac{\epsilon T}{K}$. Regret in this setting is given as:

$$\begin{aligned} R_T(\mathcal{A}) = \mathbb{E}_{\mathcal{A}}[\sum_{t=1}^{T} \mu^* - \mu_{I_t}] &= \sum_{i=1}^{K} (\mu^* - \mu_i) \frac{\epsilon T}{K} + (1-\epsilon) T (\mu^* - \mu_j) \\ &\leq \epsilon T \mu^* + T(\mu^* - \mu_j). \end{aligned}$$

Here $j = \arg\max_i \hat{\mu}_i$. From Hoeffding's inequality, it can be seen that for any arm i:

$$\mathbb{P}\{\mu_i > \hat{\mu}_i + c_i\} \leq T^{-4}. \tag{14.6}$$

$$\mathbb{P}\{\mu_i < \hat{\mu}_i - c_i\} \leq T^{-4}. \tag{14.7}$$

Thus, we have:

$$\begin{aligned}
\mu^* - \mu_j &\le \mu^* - \hat{\mu}_j + c_j && \text{(with probability at least } 1 - T^{-4}) \\
&\le \mu^* - (\hat{\mu}_j + c_j) + 2c_j && \text{(adding and subtracting } c_j) \\
&\le \mu^* - (\hat{\mu}_{i^*} + c_{i^*}) + 2c_j && (\because j = \arg\max_i \hat{\mu}_i \text{ and } c_j = c_{i^*}) \\
&\le \hat{\mu}_{i^*} + c_{i^*} - (\hat{\mu}_{i^*} + c_{i^*}) + 2c_j. && \text{(with probability at least } 1 - T^{-4})
\end{aligned}$$

Thus, expected regret can be bounded by:

$$\begin{aligned}
\mathbb{E}[R_T(\mathcal{A})] &\le \epsilon T\mu^* + 2Tc_j(1 - T^{-4}) + T^{-4}T \\
&\le \epsilon T\mu^* + 2T\sqrt{\frac{2K\ln T}{\epsilon T}} + T^{-3} \\
&\le \epsilon T\mu^* + 2T^{1/2}\sqrt{\frac{2K\ln T}{\epsilon}} + T^{-3}.
\end{aligned}$$

If $\epsilon = T^{-1/3}$, then we get $\mathbb{E}[R_T(\mathcal{A})] = O(T^{2/3})$.

14.5.2 Another Attempt: ε-greedy Algorithm

The previous algorithm has two shortcomings. First, we assume that algorithm knows the time horizon beforehand (HW: check where do we use this assumption.), and second that the algorithm does not use the partially learned parameters during the exploration phase and does not update the values in exploitation phase; hence it is less flexible.

Next we attempt the following algorithm which does not make any assumption. However, it can be proved that this algorithm also gives similar regret bound i.e. $O(T^{2/3})$.

Algorithm 2: ε-greedy Algorithm

Input: exploration parameter sequence $(\varepsilon_t)_{t\ge1}$, number of arms K
Initialize: $t = 1, S_i = 0, N_i = 0, \widehat{\mu}_i = 0$ for all i;
for $t = 1, 2, \cdots$ **do**
- Toss a coin with bias ε_t ;
- **if** *heads* **then**
  - explore: Choose an arm uniformly at random ;
- **else**
  - exploit: Choose an arm $i_t = \arg\max_i \widehat{\mu}_i$
- update:
  - $S_{i_t} = S_{i_t} + \mathbb{1}(r_{i_t,t=1})$
  - $N_{i_t} = N_{i_t} + 1$
  - $\widehat{\mu}_{i_t} = \frac{S_{i_t}}{N_{i_t}}$ (the empirical averages of arms $j \ne i_t$ remain the same)

Theorem 14.8. *The ε-greedy algorithm achieves the following upper bound on regret*

$$\mathbb{E}[R_T(\varepsilon\text{-Greedy})] \le O(T^{2/3}(K\log(T))^{1/3}).$$

Proof Intuition: We consider ε_t as a decreasing sequence. We have that till any time t, the expected number of exploration rounds are at-least $t\varepsilon_t$. Hence, each arm $i$is pulled for at-least $t\varepsilon_t/K$ times. We then use Hoeffding's ineuality to obtain the following regret bound at time i

$$\begin{aligned}\mathbb{P}(heads) + \mathbb{P}(tails)\Delta_{i_t} &\leq \varepsilon_t + (1-\varepsilon_t)\sqrt{\frac{2K\log(t)}{t\varepsilon_t}} \\ &\leq \varepsilon_t + \sqrt{\frac{2K\log(t)}{t\varepsilon_t}}\end{aligned}$$

we now bound the overall regret as

$$\begin{aligned}\mathcal{R}_T(\varepsilon\text{-}\textsc{Greedy}) &\leq \sum_{t=1}^{T}\varepsilon_t + \sqrt{\frac{2K\log(t)}{t\varepsilon_t}} \\ &\leq \sum_{t=1}^{T}(3t^{-1/3}(K\log(t))^{1/3}) && (\text{Choose } \varepsilon_t = t^{-1/3}(K\log(t))^{1/3}) \\ &\leq T^{2/3}(K\log(T))^{1/3}\end{aligned}$$

Note that this is a loose bound; one need to conidtion on exact number of arm pulls to use Hoeffding's inequality. However, the bound we get is not optimal and hence we will not explore the exact bounds in detail. We will not turn to the (asymptotically) optimal algorithms that give $O(\log(T))$ regret guarantee.