EE737 Project
Distribution Centric Approach to Correlated Multi-Armed Bandits

Neharika Jali 160040101
Ishank Juneja 16D070012

July 01, 2020

1 Introduction

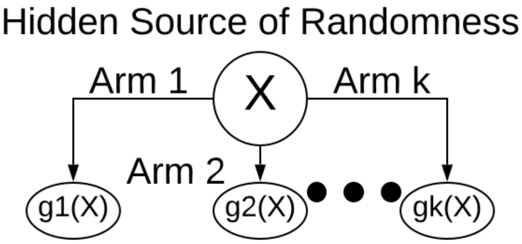


Figure 1: A model for correlated bandits

This work considers a variant of the usual Multi Armed Bandit (MAB) problem that models correlation between rewards obtained from the arms. This correlated bandit framework assumes that each arm k is associated with a reward function g_k. Further the reward received on sampling a certain arm k depends on the realization of the underlying random state X. Say if we sample the arm k at time step t and the realization of X is x_i, then the reward received will be $g_k(x_i)$.
Hence each bandit instance with K arms is composed of a discrete random variable X with an unknown distribution over an alphabet $\{x_1, x_2, \dots, x_n\}$ and a collection of arm functions $g_1(X), g_2(X) \dots, g_K(X)$.
An example of such a bandit instance, with 3 arm functions and the distribution unspecified is given in Figure 2

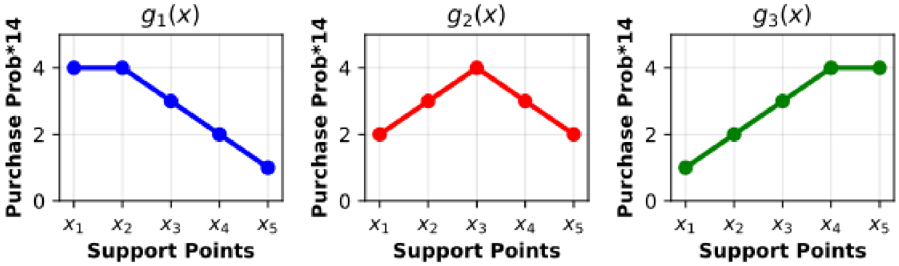


Figure 2: An example. On the Y axis we have purchase probability and income brackets on the x axis

2 Existing Work - The C Bandit Algorithm

The framework described in the previous section is from the reference [2]. The work of [2] proposes a strategy to exploit the correlation present between the rewards of various arms by identifying some arms as 'non competitive'. The arm to be sampled in a certain time slot is still chosen using a traditional bandit algorithm like UCB or Thompson sampling but arms that are non-competitive at a particular time instant are not considered for sampling.

The C-Bandit algorithm uses distribution agnostic side information gathered for arm l using the pulls of arm k. It skips sampling arms based on pairwise comparisons between the mean side reward of arm l and the true mean reward of certain reference arms k.

The C-Bandit algorithm has guarantees on giving orderwise logarithmic regret for all bandit instances and orderwise constant regret for certain bandit instances. However, the work of [3] perceived two limitations with the C-Bandit approach.

1. The arm exclusion criteria is agnostic to underlying the distribution of random variable X
2. Comparisons are pairwise between an arm being inspected as competitive and certain special reference arms

3 Our previous work - The UCUCB Algorithm

In an earlier work [3], we asked whether it was possible to outperform or match the regret performance of the C-Bandit arm exclusion approach using a distribution learning based method. In [3] we proposed such an algorithm, called UCUCB, that outperformed the C-Bandit approach on some bandit instances and appeared to match it orderwise on most other bandit instance examples that were considered. Operating under the same bandit framework as the one described in Section 1, we define the components of UCUCB next

Pseudo Distribution

As described in [3], UCUCB obtains an estimate of the distribution of random variable X and uses this estimate to skip sampling certain bad arms. For the notion of non competitive arms under UCUCB, we first define a quantity called the pseudo distribution $\tilde{P}_X = [\tilde{p}_1(t), \tilde{p}_2(t), ..., \tilde{p}_n(t)]^T$, as opposed to the true unknown distribution $P_X = [p_1, p_2, ..., p_n]^T$.
Each component $\tilde{p}_i(t)$ of $\tilde{P}_X$ is computed as follows,

$$\tilde{p}_i(t) = \sum_{\tau=1}^{t} \frac{\beta_i(\tau)}{t}. \tag{1}$$

Where, $\beta_i(\tau)$ is given by,

$$\beta_i(\tau) = \begin{cases} \frac{1}{|\text{inv}_k(r_\tau)|} & i \in \text{inv}_k(r_\tau) \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Here $\text{inv}_k(r_\tau)$ is the pre-image of the reward r_τ under the function g_k (assuming arm k pulled at time t) and $|\text{inv}_k(r)|$ is the cardinality of the set $\text{inv}_k(r_\tau)$.
Hence the pseudo distribution works by essentially diving a totally probability mass of 1 into $|\text{inv}(r_\tau)|$ equal parts among all the possible realizations of x_i for which a reward of r_τ could have been obtained from a pull of arm k_τ.

The Confidence Set

Intuitively, the confidence set is the set of realizations of random variable X having a 'high enough' probability of occurrence. At time instant t, we denote this set as C_t^*.
Say, the realizations of random variable X lie in the alphabet $\{x_1, x_2, \ldots, x_n\}$ which has n indices. To find C_t^*, we do the following,

(i) First sort the empirical p.m.f. in descending order and obtain the sorted indices as $q_1(t), q_2(t), \ldots, q_n(t)$

(ii) Then we pick $C^* = \{q_1(t), q_2(t), \ldots, q_j(t)\}$ where j is the smallest m s.t

$$\sum_{i=1}^{m} \tilde{p}_i(t) > 1 - \epsilon \tag{3}$$

Thus, the confidence interval is the minimal interval whose cumulative probability is greater than $(1 - \epsilon)$ where ϵ is a small number, and is modeled as a Hyper-Parameter.

Competitive and Non Competitive arms

Under UCUCB, if an arm lies uniformly below another arm for all likely realizations of X, then, with high probability the arm would be sub-optimal or non-competitive. Hence the name of our Algorithm U-C-UCB (U for Uniform

and C for Correlated). To quantify the notion of "high probability" we have previously defined the Confidence Set and to estimate these probability values we have defined the Pseudo-Distribution.
Using this structure, we say arm k is non-competitive if $\exists$ an arm j s.t.,

$$g_k(x) < g_j(x) \; \forall \; x \in C^* - \text{ Clause 1 } \text{ and } \; \tilde{g_k} < \tilde{g_j} - \text{ Clause 2 }, \tag{4}$$

where $\tilde{g_k}$ is the empirical expectation of the reward of arm g_k i.e. the expectation calculated assuming X is distributed according to $\tilde{P}_X$.
Here, $x \in C_t^*$ ensures that the criteria is applied only over realizations of X that have high probability. Hence existence of an arm j s.t. $g_k(x) < g_j(x)$ is true over C_t^* suggests sub-optimality of arm k.
For examples of the working of UCUCB, please see Section 4 of the earlier work [3]

4 Comments on the analysis of UCUCB

Need for the Two Clauses

Clause 1 is the $C_t^* - \epsilon_t$ clause which compares arms based on their function values over a set of highly probable points. Whereas Clause 2 compares the expected reward under the Pseudo Distribution. The condition for excluding an arm has been cast as the intersection of these two clauses. Clause 1 by itself is certainly not enough to ensure logarithmic (or even sub-linear) regret. Since even if the estimated pseudo distribution $\tilde{P}_X$ were a perfect estimate of the true distribution, that is $P_X = \tilde{P}_X$, there would still be a constant probability of excluding the optimal arm k^* since the reward function relationship between any pair of arms over a set of points with likelihood $\leq \epsilon$ will be ignored. A possible workaround to the problem of there being a constant ϵ probability of the optimal arm k^* being excluded might be to select a decreasing sequence $\{\epsilon_t\}_{t=1}^{\infty}$ instead of the constant ϵ. This sequence could create a diminishing probability of making a mistake. However bias present in the Pseudo Distribution $\tilde{P}_X$ would make such an approach ineffective as we shall see in Section 5.

Approach for Analysis

A requirement to ensure logarithmic regret while we skip sampling some arms is that the probability of the optimal arm k^* being excluded from the set of competitive arms be 'small enough'. Formally, let $E_1(t)$ be the event that the optimal arm k^* is excluded from sampling (because of being deemed non competitive) at time step $t + 1$. We can lower bound the expected number of sub-optimal arm pulls using the following construction with elements borrowed from Theorem 2 of [2].

$$\mathbb{E}[\sum_{t=0}^{T-1} \mathbb{1}_{k_{t+1} \neq k^*}] = \sum_{t=1}^{T} \mathbb{P}(k_{t+1} \neq k^*). \tag{5}$$

$$\mathbb{P}(k_t \neq k^*) \geq \mathbb{P}(E_1(t)). \tag{6}$$

Hence we need the sum,

$$\sum_{t=1}^{T} \mathbb{P}(E_1(t)) \tag{7}$$

to be sub-linear if not bounded.

$$\mathbb{P}(E_1(t)) = \mathbb{P}(\text{Clause 1} \cap \text{Clause 2}). \tag{8}$$

Here, 5 follows from the property that $\mathbb{E}[\mathbb{1}_E] = \mathbb{P}(E)$. The relation 6 follows from the fact that one of the many reasons why arm k^* is not sampled at time step t is that it was deemed non competitive.

The truth of Clause 1 depends on the parameter ϵ and the particular functions which are part of the bandit instance. Since no distribution is available over these family of functions it is not possible to utilize the confidence interval criteria in analysis. Further, the learned pseudo distribution depends on the bandit instance. The accuracy of the $1-\epsilon$ confidence set in turn depends on this learned Pseudo-Distribution. This removes the possibility of any analysis based on Clause 1, that does not first consider Clause 1, since Clause 1 uses the Pesudo distribution in a more direct manner. Hence we focus our attention to analyzing solely based on Clause 2. That is, we consider the analysis of a simplified UCUCB which has solely the criteria of Clause 2 as the non-competitiveness criteria. Hence under the new scheme the event $E_2(t)$ of the arm k^* being non competitive at time $t+1$ would be identical to Clause 2, leading to,

$$\mathbb{P}(k_t \neq k^*) \geq \mathbb{P}(E_2(t)) = \mathbb{P}(\text{Clause 2}). \tag{9}$$

In the next section we present a counter example to illustrate the problem with any Pseudo Distribution approach, such as the one in 9, for the purpose of regret minimization.

5 Problems with Distribution Centric Approach - A counter example

In this section we see that in the absence of strong restrictions on the allowed family of bandit instances, it is not possible to guarantee sub-linear regret. The following comments can be viewed as general for any distribution centric approach.

- The regret minimization framework does not give the freedom to actively explore the distribution of the underlying random variable since the aim of reducing regret is compromised by choosing low return arms that provide information about the distribution

- The pseudo distribution is information captured through the sampling of all arms, each arm function introduces its own instance dependent bias onto the distribution. So it is incorrect to compare two arms using such a distribution estimate

- There is no instance independent way to combine information from the pulls of arms l and m to estimate the return of arm k hence any global comparison basis is inherently flawed

- The C-Bandit approach of [2] overcomes the problem of instance dependence using comparisons with special reference arms. However it makes an optimistic estimate of the side reward obtained from other arms. This is the property that introduces the possibility of instance independent analysis

Next we present a counter example which refutes the possibility of analysis of a distribution centric approach such as UCUCB in the absence of strong restrictions on the bandit instances. This counter example demonstates that there can always be a constant probability of linear regret for certain bandit instances using a distribution learning based approach in the general case.

A Counter Example

Random variable X lies in the alphabet $\{x_1, x_2, x_3, x_4\}$ and $P_X = [0.1, 0.7, 0.1, 0.1]$ (Rightmost in Figure 3) and the arm functions are as shown in Figure 3. The expected return of $g_1(X)$ (left most) is $\mu_1 = 0.1\times2+0.7\times4+0.1\times1+0.1\times1 = 3.2$ and the expected return of $g_2(X)$ is $\mu_2 = 0.1\times2+0.7\times3+0.1\times3+0.1\times2 = 2.8$. This makes arm 1 the optimal arm. The regret minimization algorithm could

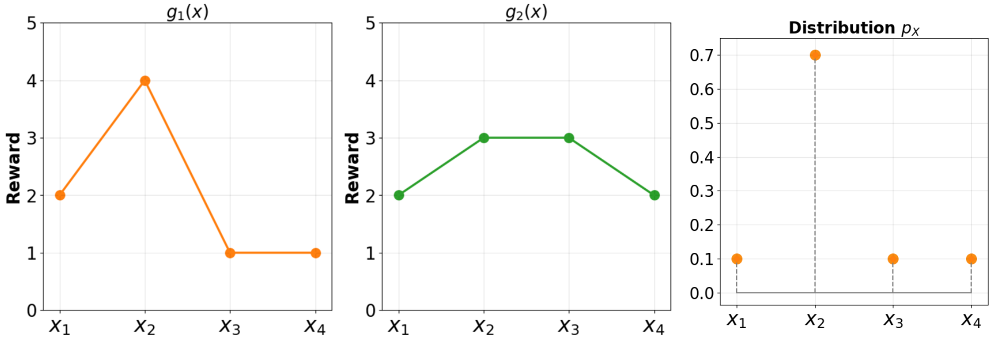


Figure 3: Bandit Instance used for the counter example

start off picking any one of the arms, let us say it breaks ties by picking the arm with the higher index - arm 2. Further assume that it receives a reward $r_1 = 3$. Based on the observed reward, the Pseudo Distribution at the start of time step 2 would then be, $\tilde{P} = [0, 0.5, 0.5, 0]$. This is the case since a reward of 3 from arm 2 would mean that X can only come from $\{x_2, x_3\}$, hence a probability mass of 1 would be distributed equally between the both of them as is done in the Pseudo Distribution estimation described in 1. Deciding which arm to sample next based on this estimate $\tilde{P}$ would make arm 2 seem better than arm 1. So using the known arm functions for arms 1 and 2 we will choose to sample arm 2 again. Over numerous consecutive samples of arm 2, the estimated probability masses of positions x_1 and x_4, $\tilde{p}_1(t)$ and $\tilde{p}_4(t)$, will remain close to $\frac{1-(0.7+0.1)}{2} = 0.1$. Further through samples of arm 2 x_2 and x_3 will remain indistinguishable and their individual probability masses will never exceed 0.5. As arm 2 is repeatedly sampled due to its perceived superiority, the Pseudo Distribution $\tilde{P}(X)$ would remain in a 1 dimensional set containing with its end points as $[0, 0.5, 0.5, 0]$ and $[0.1, 0.4, 0.4, 0.1]$. It can be verified that for all distributions lying in this set, arm 2 will be perceived as superior to arm 1. This will lead to linear cumulative expected regret as is formalized next.

Linear Regret for Bandit Instance

Let $E_3(t)$ be the event that the algorithm breaks ties by pulling arm 2 first and obtains 3 as the reward. Then the expected cumulative regret can be written as,

$$\sum_{t=1}^{T} \mathbb{E}[R(t)] = \sum_{t=1}^{T} \mathbb{E}[R(t)|E_3(t)]\mathbb{P}(E_3(t)) + \sum_{t=1}^{T} \mathbb{E}[R(t)|E_3^c(t)]\mathbb{P}(E_3^c(t)). \tag{10}$$

The first term on the RHS contributes linear regret. The second term is expected cumulative regret when $E_3(t)$ does not occur, which would still be non-negative.

UCUCB Conclusion

From the discussed counter example, it is clear that any temporary bias in the distribution is problematic and can lead to linear regret. Due to the problems pointed out with the $C_t^* - \epsilon_t$ clause, this problem will hold for certain bandit instances even if consider the more restrictive compound condition of Clause 1 $\cap$ Clause 2 or think of ϵ_t as a time varying sequence. This is the case since, as we saw actions taken at the start can have an irreversible effect on the perceived ordering between the various arms. Hence, to use a distribution based approach for regret minimization, the restriction that all arms be invertible is required.

All arm functions being invertible means that the problem is no longer a partial observability bandit problem. The problem becomes similar to experts setting with the restriction that rewards are drawn from a distribution. Also there is no exploration-exploitation trade off like in the usual bandit setting.

Under this setting, constant expected cumulative regret can be achieved using a distribution estimation based method and we propose and analyze an algorithm for the same.

6 Restricted Bandit setup and an Algorithm under it

The restriction being imposed is that either all arms should be invertible or the bandit instance should be reducible to an instance with all arms invertible.
The algorithm selects the arm that has the highest expectation under the learned distribution. We call this the **Regret Minimization with Distribution** (RMD) algorithm.
Computing the expectation under the estimated distribution is mathematically the same as updating the mean reward of arm k using the samples of all other arms. However, thinking of it as an expectation over the estimated distribution makes analysis easier as we see next.
Define the quantity $\tilde{g}_k(t)$, as

$$\tilde{g}_k(t) = \sum_{i=1}^{n} g_k(x_i)\tilde{p}_i(t) \tag{11}$$

Where the notation followed is, True discrete distribution of random variable X,

$$P_X = [p_1, p_2, \ldots, p_n]^T \tag{12}$$

The discrete random variable X is derived from the discrete alphabet $\{x_1, x_2, \ldots, x_n\}$.
Our estimate of P_X at time step t is,

$$\tilde{P}_X = [\tilde{p}_1(t), \ldots, \tilde{p}_n(t)]^T \tag{13}$$

Algorithm 1 RMD Algorithm

```
Input: Alphabet $\{x_1, ..., x_n\}$, Functions $\{g_1, ..., g_K\}$, All Invertible
Initialize : $t = 0, \tilde{g}_k = \infty$ (like Vanilla UCB)
for Every round t do
    $\tilde{g}_k(t) \leftarrow \sum_{i=1}^{n} g_k(x_i)\tilde{p}_i(t)$
    $k_t = \arg\max_k \tilde{g}_k(t)$
    Receive reward $r_t$ by sampling arm $k_t$
    Record the realization of $x$, $x \leftarrow g_k^{-1}(r_t)$
    $t \leftarrow t + 1$
    $\tilde{p}_i(t+1) \leftarrow (\tilde{p}_i(t) \times t + \mathbb{1}_{x=x_i})/t$
end for
```

Interlayed Active Exploration for Regret Minimization

A distribution learning based approach that can provide sub linear expected cumulative regret could be active exploration with a successively reducing (with time step t) probability interlayed active exploration component. That is we actively explore the distribution at every time step with probability ϵ_t. If we think of a decreasing series of the form $\frac{1}{t}$, then in the limit we actively explore the distribution an infinite number of times. However, as is well known the sum of the harmonic sequence,

$$\sum_{t=1}^{T} \frac{1}{t}, \tag{14}$$

though sub linear is super logarithmic, with the difference,

$$\lim_{T\to\infty} \sum_{t=1}^{T} \Big(\frac{1}{t} - \log{(t)}\Big) = \gamma. \tag{15}$$

Where, γ is the positive Euler–Mascheroni constant. Further, it is known from the study of the Riemann Zeta function that the sum,

$$\lim_{T\to\infty} \sum_{t=1}^{T} \frac{1}{t^{\alpha}}, \tag{16}$$

converges to a finite constant $\forall\, \alpha > 1$. Hence any diminishing sequence with $\alpha > 1$, would not be sufficient exploration since even in the limit it would actively explore only a constant number of times. Any constant number of samplings to estimate the distribution would be insufficient to ensure sub linear regret as can be verified easily. Other families of decreasing sequences could be constructed using known functions like $\log{(T)}$ and $\exp{(t)}$, however we could not find a sequence that in the sum would not only be sub linear, but also sub logarithmic.

7 Analysis of RMD

Lemma 1: Hoeffding Inequality
For a random variable $X \in (a, b)$

$$\mathbb{P}\Big(\frac{\sum_{i=1}^{t} X_i}{t} - \mu \geq \epsilon\Big) \leq \exp{\Big(\frac{-2t\epsilon^2}{(b-a)^2}\Big)} \tag{17}$$

Lemma 2: The Expected reward under the distribution $\tilde{P}_X$ is a random variable averaged over t samples and bounded between 0 and B.

Proof. Our estimates of the distributions $\tilde{p}_i(t)$ are each random variables averaged over t samples since,

$$\tilde{p}_i(t) = \frac{\sum_{\tau=1}^{t} \mathbb{1}_{g^{-1}_{k_\tau}(x_\tau)=x_i}}{t} \tag{18}$$

Since $\tilde{g}_k(t) = \sum_{i=1}^{n} g_k(x_i)\tilde{p}_i(t)$, we have the following,

$$\tilde{g}_k(t) = \frac{\sum_{\tau=1}^{t} \sum_{i=1}^{n} g_k(x_i) \mathbb{1}_{g_{k_\tau}^{-1}(x_\tau)=x_i}}{t}, \tag{19}$$

this can be written as,

$$\tilde{g}_k(t) = \frac{\sum_{\tau=1}^{t} Y_\tau^k}{t}. \tag{20}$$

Where the random variable Y_τ^k is a linear combination of the random variables $\mathbb{1}_{g_{k_\tau}^{-1}(x_\tau)}$. Now, the random variable Y_τ is itself bounded between 0 and B.

□

Lemma 3: The expected number of sub-optimal pulls is upper bounded by a constant.

Proof. Let $E(t)$ be the event that the optimal arm k^* is not pulled in time slot $t+1$.

$$\begin{align}
\mathbb{P}(E(t)) &= \mathbb{P}(\max_k \tilde{g}_k(t) > \tilde{g}_{k^*}(t), k \neq k^*) \tag{21}\\
&\leq \sum_{k \neq k^*} \mathbb{P}(\tilde{g}_k(t) > \tilde{g}_{k^*}(t)) \tag{22}\\
&= \sum_{k \neq k^*} \mathbb{P}(\tilde{g}_k(t) - \tilde{g}_{k^*}(t) > 0) \tag{23}\\
&= \sum_{k \neq k^*} \mathbb{P}(\tilde{g}_k(t) - \tilde{g}_{k^*}(t) - (\mu_k - \mu^*) > \mu^* - \mu_k)) \tag{24}\\
&= \sum_{k \neq k^*} \mathbb{P}(\tilde{g}_k(t) - \tilde{g}_{k^*}(t) - (\mu_k - \mu^*)) > \Delta_k)). \tag{25}
\end{align}$$

Using the notation of Lemma 2, we have

$$= \sum_{k \neq k^*} \mathbb{P}\Big(\frac{\sum_{\tau=1}^{t} Y_\tau^k - \sum_{\tau=1}^{t} Y_\tau^{k^*}}{t} - (\mu_k - \mu^*)) > \Delta_k)\Big). \tag{26}$$

Applying the Hoeffding Inequality from Lemma 1 and using the fact that each of the Y variables are bounded between 0 and B, we have,

$$\begin{align}
&\leq \sum_{k \neq k^*} \exp\Big(\frac{-2t\Delta_k^2}{(2B)^2}\Big) \tag{27}\\
&\leq K \exp\Big(\frac{-t\Delta_{\min}^2}{2B^2}\Big). \tag{28}
\end{align}$$

From the property of indicator random variables, we know that,

$$\mathbb{E}[\mathbb{1}_{k_{\tau+1}\neq k^*}] = \mathbb{P}(E(t)) \tag{29}$$

Hence, the total number of sub optimal pulls,

$$\mathbb{E}[\sum_{\tau=0}^{t-1} \mathbb{1}_{k_{\tau+1}\neq k^*}] = \sum_{\tau=1}^{t} \mathbb{P}(E(\tau)) \tag{30}$$

$$\leq K \sum_{\tau=1}^{t} \exp\left(\frac{-\tau\Delta_{\min}^2}{2B^2}\right) \tag{31}$$

Which in turn is upper bounded by the constant infinte sum. □

Theorem 1. *The expected cumulative regret is upper bounded by,*

$$\mathbb{E}[R(t)] \leq K\Delta_{\max} \sum_{\tau=1}^{t} \exp\left(\frac{-\tau\Delta_{\min}^2}{2B^2}\right) \tag{32}$$

Which is a constant

Proof. The proof is easy to see from the definition of expected regret,

$$\mathbb{E}[R(t)] = \sum_{k\neq k^*} \Delta_k \mathbb{E}[n_k(t)] \tag{33}$$

$$\leq \Delta_{\max} \sum_{k\neq k^*} \mathbb{E}[n_k(t)] \tag{34}$$

$$= \Delta_{\max} \sum_{\tau=1}^{t} \mathbb{E}[\mathbb{1}_{k_\tau \neq k^*}] \tag{35}$$

Substituting the sum $\sum_{\tau=1}^{t} \mathbb{E}[\mathbb{1}_{k_\tau\neq k^*}]$ from Lemma 3 completes the proof. □

7.1 Restriction Relaxed

If the restrictions that the arms are invertible be relaxed, then the bound on the number of pulls of suboptimal arms which involves application of Hoeffding's Inequality (27) wouldn't hold true any longer due to the absence of unbaised estimates of the arm rewards. Hence, number of times a sub-optimal arm is pulled won't necessarily be a constant leading to the regret being not necessarily logarithmic.

8 $(0, \delta)$-PAC Algorithm for Correlated Bandits

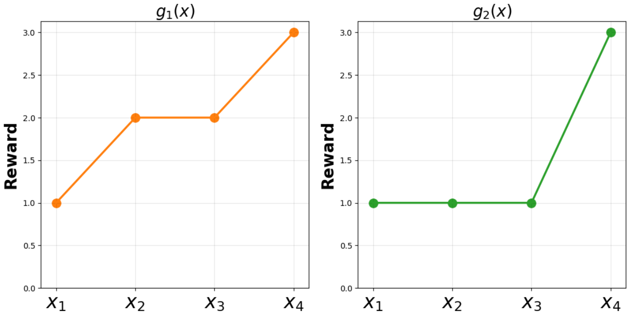
Since distribution learning is not suitable for regret minimization, we look at the pure exploration setting of Multi Armed Bandits.

We propose a $(0, \delta)$-PAC Algorithm for a correlated bandit based on RRPULL + PIEST algorithm by Gupta et al. (2018) [1] that learns distributions of rewards from indirect samples and the Successive Elimination algorithm.

Given a set of arms with known reward distributions, based on a latent random variable X, $g_j(X) \forall j = \{1, 2, ..., K\}$

8.1 Preprocessing

If there any support points $x_i' s$ such that their reward is same for all support points for all arms, then merge them into a super support point $x_{i'}$. For example, in the distributions of two arms as show below, support points x_2 and x_3 can be merged into one superpoint. Further all rewards are normalised between $[0, 1]$.



Thus, after preprocessing, we have a set of arms, $S = \{1, 2, ..., K\}$ with rewards $g_j(X) \forall j = \{1, 2, ..., K\}$ which are a function of the latent random variable X with support points $X = \{1, 2, ..., n\}$.

8.2 Notation

Borrowing notation from Gupta et al. (2018) [1], we define the following. Let $\tilde{P}_X = [\tilde{p}_1(t), ..., \tilde{p}_n(t)]^T$ be the estimated probability distribution of latent random variable, X, $P_X = [p_1(t), ..., p_n(t)]^T$. For each arm k, let $\{z_{k,1}, ..., z_{k,m_k}\}$ denote the set of possible outcomes of the function g_k where m_k is the number of distinct outputs of g_k. The information about g_k required to estimate the probability distribution of X can be captured in matrix A_k with m_k rows and n columns called the Sample Generation Matrix for arm k. $A_k(i, j) = 1$ if output $z_{k,i}$ could have been generated by x_j in arm k. $A = [A_1^T, A_2^T, ..., A_K^T]$ of size $m \times n$ where $m = m_1 + m_2 + ... + m_K$. Let $q_{k,i}$ be the probability of observing $z_{k,i}$ each time arm k is pulled and define $Q = [q_{1,1}, ..., q_{1,m_1}, ..., q_{K,m_K}]^T$. For each arm k and output i, we hence have the following equations which when solved will give us the probability estimate of X.

$$q_{k,i} = \sum_{j=1}^{n} A_k(i, j) p_j \tag{36}$$

$$AP_X = Q \implies P_X = A^+Q \tag{37}$$

where A^+ is Moore-Penrose inverse of A.

From estimates of probability distribution of X, we can estimate the arm rewards at time t as,

$$\tilde{g}_s^t = \sum_{j=1}^{n} g_s(x_j)\tilde{p}_j(t) \tag{38}$$

8.3 Algorithm

The algorithm estimates the probability distribution of X indirectly by pulling arms and obtaining rewards dependent on a realisation of X. In every iteration, each arm is pulled once in a round robin manner and the estimates are updated. At the end of each round, we check for arms that are far enough from the optimal arm based on confidence intervals and eliminate them. The pseudocode of the algorithm is as follows.

Algorithm 2 Successive Elimination for Correlated Bandits

```
**Input:** Alphabet $\{x_1, ..., x_n\}$, Functions $\{g_1, ..., g_K\}$, Set of arms $S$
**Preprocess** by merging into superpoints
**Initialize:** $t = 0, t_k = 0 \ \forall \ k$, $t_{k,i} = 0 \ \forall \ i, k$, $\tilde{p}_j(0) = \frac{1}{n} \ \forall \ j$
**while** $|S| > 1$ **do**
  Pull arm $s_t = mod(t, |S|) + 1$, observe output $y_t$
  $t_{s_t} = t_{s_t} + 1$, $t = t + 1$
  **if** $y_t = z_{s_t,i}$ **then**
    $t_{s_t,i} = t_{s_t,i} + 1$
  **end if**
  $\hat{q}_{s,i} = \frac{t_{s,i}}{t_s} \ \forall \ i, s$
  Obtain estimates $\tilde{p}_j(t)$ as $\tilde{P}_X = A^+\hat{Q}$
  Let $\tilde{g}_{max}^t = \max_{s \in S} \tilde{g}_s^t$, $\alpha_t = \sqrt{\frac{\log(cKt^2/\delta)}{t}}$
  For every arm $s \in S$ s.t $\tilde{g}_{max}^t - \tilde{g}_s^t \geq 2\alpha_t$. Set $S = S \backslash s$
**end while**
**Return:** $S$
```

8.4 Analysis

Theorem 2 (Gupta et al., 2018 [1], Theorem 1)**.** *It is possible to achieve asymptotically consistent estimation of probability distribution of X if and only if $rank(A) = n$.*

Theorem 3 (Gupta et al., 2018 [1], Theorem 6)**.** *It is possible to achieve estimation error of probability distribution of X of $O(\frac{1}{t})$ if $rank(A) = n$.*

The preprocessing step ensures that all instances have $rank(A) = n$.

Theorem 4. *The empirical estimate of the distribution P_X, $\tilde{P}_X = A^+\hat{Q}$ is unbiased*

Proof:
By the definition of $\hat{q}_{s,i}$, $\hat{Q}$ is unbiased. Therefore, $\mathbb{E}[\hat{Q}] = Q$. Hence,

$$\mathbb{E}[\tilde{P}_X] = \mathbb{E}[A^+\hat{Q}] = A^+\mathbb{E}[\hat{Q}] = A^+Q = P_X \tag{39}$$

Theorem 5. *The Successive Elimination for Correlated Bandits is* $(0, \delta)$*-PAC algorithm, and with probability* $(1-\delta)$ *its arm complexity is bounded by* $O\left(\frac{\log(K/\delta\Delta_{min})}{\Delta_{min}^2}\right)$

Proof. **Part (A) - Proof of $(0, \delta)$-PAC**
For any time t and action $s \in S_t$, we have,

$$Pr[|\tilde{g}_s^t - \mu_s| \geq \alpha_t] \leq \exp^{-\alpha_t^2 t} \leq \frac{\delta}{cKt^2} \tag{40}$$

because $\hat{\mu}_s^t$ is an unbiased estimate of μ_s from [4].
With probability at least $(1-\frac{\delta}{K})$ for any time t and action $s \in S_t$, $|\tilde{g}_s^t - \mu_s| \leq \alpha_t$. Hence, with probability $(1-\delta)$, best arm is never eliminated as as $\alpha_t \to 0$ as t increases, eventually every non-best arm is eliminated. Hence the algorithm is $(0, \delta)$-PAC.

Part (B) - Sample Complexity
To eliminate a non-best arm s_i, we need to reach a time t_i such that,

$$\hat{\Delta}_{t_i} = \tilde{g}_{s^*}^{t_i} - \tilde{g}_{s_i}^{t_i} \geq 2\alpha_{t_i} \tag{41}$$

where s^* represents the best arm. Definition of α_t combined with the assumption that $|\hat{\mu}_s^t - \mu_s| \leq \alpha_t$ yields that,

$$\Delta_i - 2\alpha_t = (\mu_{s^*} - \alpha_t) - (\mu_{s_i} + \alpha_t) \geq \tilde{g}_{s^*} - \tilde{g}_{s_i} \geq 2\alpha_t \tag{42}$$

which holds with probability atleast $(1 - \frac{\delta}{K})$ for

$$t_i = O\left(\frac{\log(K/\delta\Delta_i)}{\Delta_i^2}\right) \tag{43}$$

The last non-best arm will thus be eliminated and the best arm output at

$$t = O\left(\frac{\log(K/\delta\Delta_{min})}{\Delta_{min}^2}\right) \tag{44}$$

where $\Delta = \min_j \Delta_j$, the sub-optimality gap between the best and the second best arm. The implicit improvement, due to the correlated bandit framework, in concentration of rewards of all arms by pulling an arm reduces the number of times that each arm has be pulled. This, thus, amounts to the gain we achieve in sample complexity over the vanilla Successive Elimination algorithm. □

References

[1] S. Gupta, G. Joshi, O. Yagan. "Active Distribution Learning from Indirect Samples." *Allerton Conference on Communication, Control and Computing, 2018*

[2] S. Gupta, G. Joshi and O. Yağan, "Correlated Multi-Armed Bandits with A Latent Random Source," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[3] I. Juneja, D. S. Gaharwar, D. Varshney and S. Moharir, "A New Approach to Correlated Multi Armed Bandits," 2020 COMSNETS

[4] E Even-Dar, S Mannor, Y Mansour. "PAC bounds for multi-armed bandit and Markov decision processes". International Conference on Computational Learning Theory 2002, 255-270

[5] Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014, May). lil'ucb: An optimal exploration algorithm for multi-armed bandits. In Conference on Learning Theory (pp. 423-439).