Best Arm Identification in Sample-path Correlated Bandits

R Sri Prakash
IIT Bombay
sriprakash@ee.iitb.ac.in

Nikhil Karamchandani
IIT Bombay
nikhilk@ee.iitb.ac.in

Sharayu Moharir
IIT Bombay
sharayum@ee.iitb.ac.in

***Abstract*—We consider the problem of best arm identification in the fixed confidence setting for a variant of the multi-arm bandit problem. In our problem, each arm is associated with two attributes, a known deterministic cost, and an unknown stochastic reward. In addition, it is known that arms with higher costs have higher rewards across every sample path. The net utility of each arm is defined as the difference between its expected reward and cost. We consider two information models, namely, the full information feedback and sequential bandit feedback. We derive a fundamental lower bound on the sample complexity of any policy and also propose policies with provable performance guarantees that exploit the structure of our problem. We supplement our analytical results by comparing the performance of various candidate policies via synthetic and data-driven simulations.**

I. Introduction

The widely studied multi-armed bandit (MAB) problem is a sequential decision-making problem. The classical MAB problem has K independent arms with unknown reward distributions. In each round, one of the K arms is chosen. Multiple performance metrics have been considered including maximizing long-term cumulative reward [1] and identifying the arm with the maximum expected reward [2], [3].

Many variants of the MAB have been studied, see [4] for a detailed exposition. Closest to this work, [5] focuses on the setting where the rewards across arms are correlated. In this setting, the reward obtained from arm i can provide information about the reward that one might have received if they sampled another arm j. In [5], this correlation is captured through apriori known pseudo-rewards. These pseudo-rewards provide us an upper bound on the expected reward from arm j, given that the response from arm i was r. The key take-away from [5] is that the knowledge of the pseudo rewards can be exploited to reduce the sample complexity of best-arm identification in multi-armed bandits in the fixed confidence setting. In this setting, the goal is to identify the arm with the maximum expected reward, with probability at least $1-\delta$ for some $\delta > 0$.

In this work, we focus on the best-arm identification in multi-arm bandits with additional structure. Each arm is associated with two attributes: *(i)* a known cost and *(ii)* reward with an unknown distribution. Further, it is known that the arms with higher costs have higher rewards for each sample-path. Note that this sample-path correlation is different from the correlation across arms studied in [5], which captures the correlation across arms only through the conditional expectation of the rewards. The goal of this work is to identify the arm with the highest utility, defined as the difference between its expected reward and cost, with probability $1-\delta$ for some $\delta > 0$.

The problem is motivated by Software as a Service (SaaS) applications like Video on Demand (VoD) services, online shopping platforms, etc. Many SaaSs offer a range of their versions to each potential customer, where the more expensive/resource-intensive versions provide better Quality of Service (QoS) and/or Quality of Experience (QoE) to the customer. An example of this is VoD services like YouTube/Netflix where one possible measure of QoS/QoE is the resolution of the videos streamed. The utility of the version is defined as the difference between the reward accrued by the customer and the cost of the version. Given the trade-off between the cost of a version and the reward obtained by the customers, identifying the version which provides the right balance between the two is an important problem. While the cost of a version is very often known/predictable, the reward received by the customers from a version is often subjective/unknown and has to be estimated by collecting feedback from the customers. This motivates the learning problem of identifying the "best" version of the service efficiently, which matches the correlated best-arm identification problem studies in this work with each arm representing a different version of the service.

A seemingly related variant of the classical MAB is that of *structured bandits* wherein the mean rewards of the arms are related to each other. The best-arm identification problem has been studied for various special cases of structured bandits such as *linear* [6], [7], *generalized linear* [8], and *graphical* [9] bandits. A key difference between structured bandits and the correlated bandits model studied in this work is that in the former, while the mean rewards of the various arms are related to each other, the reward realizations in any given round are not necessarily correlated unlike the latter which considers a specific model for such sample-path correlation.

A. Our Contributions

We consider multiple observation models. Let Δ denote the gap between the expected utility of the best and the second best arm.

1) Full information feedback: In this setting, the agent observes the reward of all K arms in each round. We first characterize a lower bound of $\Omega\left(\frac{\log(1/\delta)}{\Delta^2}\right)$ on the number of rounds needed by any algorithm to output the best arm with probability at least $1-\delta$. We propose a UCB-based policy to identify this best arm for this setting. We show that with probability at least $1-\delta$ the proposed policy identifies the best arm in $O\left(\frac{\log(K/\delta)}{\Delta^2}\right)$ rounds.

978-1-6654-5136-9/22/$31.00 ©2022 IEEE

2) Sequential bandit feedback: In this setting, in each round, the agent sequentially selects arms and observes their corresponding rewards. Thus, based on the history of the observed rewards, the agent can either select the next arm to sample in the current round, or decide to end this round and move to the next round. Note that the lower bound of $\Omega\left(\frac{\log(1/\delta)}{\Delta^2}\right)$ for the full information case is also a lower bound on the number of samples needed by any algorithm to output the best arm with probability at least $1-\delta$ for the sequential bandit feedback setting. We propose a policy which exploits the sample-path correlation in rewards and show that, with probability at least $1-\delta$, the proposed policy identifies the best arm in $O\left(\log K \frac{\log(K/\delta)}{\Delta^2}\right)$ rounds.

We supplement our analytical results by comparing the performance of various candidate policies via simulations using real-world datasets.

II. Problem setup

We consider a variant of the popular Multi-Armed Bandit (MAB) problem. Consider K arms where each arm i is associated with a fixed known cost $c_i > 0$. We assume that $\{c_i\}_{i=1}^K$ form an increasing sequence, i.e., $c_1 < c_2 < \ldots < c_K$. In each round t, a random reward vector $R^t = (R_1^t, R_2^t, \ldots, R_K^t) \in \{0,1\}^K$ is generated where R_i^t denotes the reward corresponding to arm i in round t. Furthermore, we assume that the reward vector R^t satisfies the additional constraint that arms with higher cost have higher reward in all sample paths, i.e, R^t is composed of a sequence of zeroes followed by a sequence of ones. For example, for $K=4$, the reward vector in a round can be one of the following $K+1=5$ possibilities: $\{(0,0,0,0),(0,0,0,1),(0,0,1,1),(0,1,1,1),(1,1,1,1)\}$. In order to specify the probability distribution for the reward vector R^t, let $X^t \in \{1,\ldots,K+1\}$ denote the position[1] of the first one in the vector R^t. Note that there is a one-to-one correspondence between the reward vector R^t and X^t. Let r_i denote the expected value of the reward from arm i, i.e., $r_i = \mathbb{E}[R_i^t]$. It follows that $0 \le r_1 < r_2 < \ldots < r_K \le 1$. Let $r_0 = 0$ and $r_{K+1} = 1$. Therefore, $\mathbb{P}(X^t = i) = r_i - r_{i-1}$ for $1 \le i \le K+1$.

From the above description, it follows that the different arms generate correlated rewards in each given round. The reward vectors across rounds are assumed to be identically and independently distributed (i.i.d.). The expected net utility of arm i is defined as $\mu_i = r_i - c_i$ and the arm with the highest net utility is referred to as the best arm, i.e., $i^\star = \arg\max_i \mu_i$. Let (i) denote the index of the i-th best arm in terms of net utility, i.e., $\mu_{(1)} > \mu_{(2)} > \ldots > \mu_{(K)}$. From above, the index $i^\star$ is the same as (1). The utility gap between sub optimal arm i and best arm $i^\star$ is defined as $\Delta_i = (r_{(1)} - c_{(1)}) - (r_i - c_i)$. The minimum utility gap is then given by $\Delta_{\min} \triangleq \Delta_{(2)} = \left(r_{(1)} - c_{(1)}\right) - \left(r_{(2)} - c_{(2)}\right)$.

Consider an agent which does not apriori know the underlying reward vector distribution, and whose goal is to use the reward vector observations (or part thereof) to identify the best arm $i^\star$ with probability at least $1-\delta$, for any given $\delta > 0$. We consider two different observation models for the agent.

(i) *Full information feedback*: Here, in each round t, the agent observes the entire reward vector R^t and the goal is to analyze the minimum number of rounds τ^F needed by the agent to identify the best arm with desired confidence $1-\delta$.

(ii) *Sequential multi-arm feedback*: In each round, the agent sequentially selects arms and observes their corresponding rewards. Thus, based on the history of the observed rewards, the agent can either select the next arm to sample in the current round, or decide to end this round and move to the next round. Note that in any round t, the number of observed components of the reward vector R^t lies in the interval $[1,K]$ and is potentially random. The agent continues until it can successfully identify the best arm and we are interested in the total number of samples κ^S summed across all rounds.

III. Results

A. Full information feedback

We consider the full information feedback setting here and begin by providing an instance-dependent lower bound on the number of rounds τ^F needed by any scheme which can identify the best arm with probability at least $1-\delta$.

Theorem 1: Under the full information feedback setting, the expected number of rounds needed for any Algorithm $\mathcal{A}$ to output the best arm with probability at least $1-\delta$ is lower bounded as

$$\mathbb{E}[\tau^F] \ge \frac{\log(1/2.4\delta)}{\Delta_{\min}^2} \frac{(c_{i^\star} - c_{i^\star-1})(r_{i^\star+1} - r_{i^\star})}{r_{i^\star+1} - r_{i^\star-1}}.$$

Proof: Here, in each round t, the agent observes the entire reward vector R^t and the goal is to analyze the minimum number of rounds τ^F needed by the agent to identify the best arm with desired confidence $1-\delta$. Also, recall that the reward vector R^t is uniquely determined by X^t which denotes the position of the first one in the vector. Let P represent the distribution of X^t which is a function of the underlying instance $\{r_i\}$ and recall that $i^\star = (1)$ denotes the index of the best arm for the given instance.

Now let us consider an alternate instance of rewards $\{r_i'\}$ for which the index of the best arm is different from the best arm (1) in the original setting. Let P' represent the corresponding distribution of X^t under this alternate instance. By definition, any feasible scheme $\mathcal{A}$ should be able to identify the best arm under both instances with probability at least $1-\delta$. Let τ^F represent the stopping time and let $\widehat{i^\star}$ denote the scheme output. Then, by definition, we have

$$P\left(\widehat{i^\star} = (1)\right) \ge 1-\delta,\ P'\left(\widehat{i^\star} = (1)\right) \le \delta. \tag{1}$$

Using Wald's lemma, we have

$$\begin{aligned}\mathbb{E}_P\left[\sum_{t=1}^{\tau^F} \log\frac{P(X_t)}{P'(X_t)}\right] &= \mathbb{E}[\tau^F]\mathbb{E}_P\left[\log\frac{P(X_t)}{P'(X_t)}\right] \\ &= \mathbb{E}[\tau^F] D(P||P'), \end{aligned} \tag{2}$$

[1] For the all-zero reward vector, we will set $X^t = K+1$.

where $D(P||P')$ represents the KL divergence between P and P'. Also,

$$\begin{aligned}
&\mathbb{E}_P\left[\sum_{t=1}^{\tau^F}\log\frac{P(X_t)}{P'(X_t)}\right]\\
&= D(P(X_1,\cdots,X_{\tau^F})||P'(X_1,\cdots,X_{\tau^F})),\\
&\overset{(a)}{\geq} D\left(Ber\left(P(\widehat{i^\star}=(1)\right)||Ber\left(P'(\widehat{i^\star}=(1)\right)\right)\\
&\overset{(b)}{\geq} D\left(Ber(1-\delta)||Ber(\delta)\right)\geq\log(1/2.4\delta), \qquad (3)
\end{aligned}$$

where $Ber(x)$ denotes the Bernoulli distribution with parameter $x\in(0,1)$; (a) follows from the data-processing inequality [10]; and (b) follows from (1). Combining (2) and (3), we have that for any alternate instance of rewards $\{r_i'\}$ and the corresponding induced distribution P' on X^t

$$\mathbb{E}[\tau^F]\geq\frac{\log(1/2.4\delta)}{D(P||P')}. \qquad (4)$$

Consider an alternate instance $\{r'\}$ given by $r_i'=r_i$ for all $i\neq(1)$ and $r'_{(1)}=r_{(2)}-c_{(2)}+c_{(1)}-\epsilon$, for some $0<\epsilon<r_{(2)}-c_{(2)}-r_{(1)-1}+c_{(1)}$. Note that for this choice of alternate instance, we have $r_{(1)-1}<r'_{(1)}<r_{(1)}$ and $r'_{(1)}-c_1<r_{(2)}-c_2$. Thus, the best arm under the alternate instance is different from the original instance. The induced distribution P' on X^t differs from P at $\{(1),(1)+1\}$, and we have

$$\begin{aligned}
&D(P||P')\\
&=(r_{(1)}-r_{(1)-1})\log\frac{r_{(1)-1}-r_{(1)}}{r_{(1)-1}-r'_{(1)}}\\
&\quad+(r_{(1)+1}-r_{(1)})\log\frac{r_{(1)}-r_{(1)+1}}{r_{(1)}-r'_{(1)+1}}\\
&=(r_{(1)+1}-r_{(1)-1})D\left(\frac{r_{(1)-1}-r_{(1)}}{r_{(1)-1}-r_{(1)+1}}\middle|\middle|\frac{r_{(1)-1}-r'_{(1)}}{r_{(1)-1}-r_{(1)+1}}\right)\\
&\leq\frac{(r_{(1)+1}-r_{(1)-1})(r_{(1)}-r'_{(1)})^2}{(r_{(1)-1}-r'_{(1)})(r'_{(1)}-r_{(1)+1})} \qquad (5)\\
&=(\Delta_{\min}+\epsilon)^2\frac{(r_{(1)+1}-r_{(1)-1})}{(r'_{(1)}-r_{(1)-1})(r_{(1)+1}-r'_{(1)})}\\
&\leq(\Delta_{\min}+\epsilon)^2\frac{(r_{(1)+1}-r_{(1)-1})}{(r'_{(1)}-r_{(1)-1})(r_{(1)+1}-r_{(1)})} \qquad (6)\\
&\leq(\Delta_{\min}+\epsilon)^2\frac{(r_{(1)+1}-r_{(1)-1})}{(c_{(1)}-c_{(1)-1})(r_{(1)+1}-r_{(1)})}, \qquad (7)
\end{aligned}$$

where (5) follows since the relative entropy between $Ber(p),Ber(q)$ is upper bounded as $D(p||q)\leq(p-q)^2/(q(1-q))$; (6) follows since $r'_{(1)}<r_{(1)}$; and (7) follows since $r'_{(1)}-r_{(1)-1}=r_{(2)}-c_{(2)}+c_{(1)}-\epsilon-r_{(1)-1}=\left[\left(r_{(2)}-c_{(2)}\right)-\left(r_{(1)-1}-c_{(1)-1}\right)\right]+c_{(1)}-c_{(1)-1}\geq 0$. Since (7) holds for any $\epsilon>0$, we can choose it to be arbitrarily small and combining with (4), we have

$$\mathbb{E}[\tau^F]\geq\frac{\log(1/2.4\delta)}{\Delta^2_{\min}}\frac{(c_{(1)}-c_{(1)-1})(r_{(1)+1}-r_{(1)})}{r_{(1)+1}-r_{(1)-1}}.$$

■

Next, we present a scheme in Algorithm 1 which recovers the best arm under full information feedback and derive a bound on the number of rounds required. In Algorithm 1 we calculate the Lower Confidence Bound (LCB) and Upper Confidence Bound (UCB) of all arm initially and eliminate the arm j if UCB of arm j is less than LCB of any other arm i i.e., $i\neq j$. We call the set of non eliminated arms as active arms and calculate LCB and UCB of active arms only in each round. The elimination continues till one arm is left in the active arms set, which is the output of algorithm.

Algorithm 1: Best arm identification under full information

```
Input: $c_i$ for all $i$, $\delta$
Output: Best arm with probability 1-$\delta$
Initialization: $\hat{r}_i(0)=0$, for all $i\in\{1\cdots K\}$,
  $\mathcal{A}_1=\{1\cdots K\}$, t=1;
while $|\mathcal{A}_t|>1$ do
    for $i\in\mathcal{A}_t$ do
        update sample mean:
          $\hat{r}_i(t)=((t-1)\hat{r}_i(t-1)+R_i^t)/t$;
        LCB$_i=\hat{r}_i(t)-c_i-\sqrt{\frac{1}{2t}\ln(K/\delta)}$;
        UCB$_i=\hat{r}_i(t)-c_i+\sqrt{\frac{1}{2t}\ln(K/\delta)}$;
    end
    for $j\in\mathcal{A}_t$ do
        if $LCB_i>UCB_j$, for any $i\neq j$ then
            eliminate arm $j$: $\mathcal{A}_{t+1}=\mathcal{A}_t/\{j\}$;
        else
            continue;
        end
    end
    t++;
    if $|\mathcal{A}_t|=1$ then
        return $\mathcal{A}_t$;
    else
        continue;
    end
end
```

Theorem 2: With probability at least $1-\delta$, Algorithm 1 outputs the best arm under the full information feedback agent model and the number of rounds required τ^F satisfies

$$\tau^F\leq\mathrm{O}\left(\log(K/\delta)/\Delta^2_{\min}\right)$$

with probability at least $1-\delta$.

Proof: Let $\hat{\mu}_i(n)=\left(\sum_{t=1}^n R_i^t\right)/n-c_i$ denote the empirical estimate of the expected net utility of arm i after n rounds and let $\mathcal{E}$ be the event that for all $1\underline{\leq i\leq K,}\ n\geq 1$, we have $|\mu_i-\hat{\mu}_i(n)|\leq\epsilon_n$, where $\epsilon_n=\sqrt{\frac{\log(K/\delta)}{2n}}$. Using Hoeffding's inequality and the union bound, we get $\mathbb{P}(\mathcal{E}^c)\leq\delta$. Hereafter, we will assume that the event $\mathcal{E}$ holds,and show that the algorithm will output the best arm.

We update the empirical estimates and confidence intervals of the expected net utility of all arms after each round and eliminate arm j if UCB$_j$ $\leq$ LCB$_i$ for any $i\neq j$. This elimination continues till we are left with one arm which is the estimate for the best arm. We demonstrate the correctness of the scheme by showing that the true best arm (1) is not eliminated. Assuming to the contrary, say arm (1) is eliminated which implies there exists an arm j such that $\mu_{(1)} \leq \hat{\mu}_{(1)}(n) + \epsilon_n \leq \hat{\mu}_j(n) - \epsilon_n \leq \mu_j$ which is a contradiction. To derive an upper bound on the stopping time of the algorithm, we note that a sufficient condition for any arm $j \neq (1)$ to have been eliminated by round (n) is $\hat{\mu}_{(1)} - \epsilon_n \geq \hat{\mu}_j + \epsilon_n$. Under the event $\mathcal{E}$, we have $\hat{\mu}_{(1)} \geq \mu_{(1)} - \epsilon_n$ and $\hat{\mu}_j \leq \mu_j + \epsilon_n$, and so the above condition is satisfied if $\mu_{(1)} - 2\epsilon_n \geq \mu_j + 2\epsilon_n$ or equivalently, $\epsilon_n \leq (\mu_{(1)} - \mu_j)/4$, which is true for any $n \geq 8\log(K/\delta)/(\mu_{(1)} - \mu_j)^2$. Considering the worst-case requirement across all sub-optimal arms j, we have that the algorithm terminates by $8\log(K/\delta)/(\mu_{(1)} - \mu_{(2)})^2 = 8\log(K/\delta)/\Delta^2_{\text{min}}$ rounds, which concludes the proof. ■

Comparing Theorems 1 and 2, we see that the optimal sample complexity τ^F under the full information feedback setting depends inversely on Δ^2_{min}.

B. Sequential multi-arm feedback

Recall that in the sequential multi-arm feedback model, the agent can sequentially select arms in each round t and observe the corresponding elements of the reward vector R^t. A naive scheme is to simulate the scheme in Algorithm 1, proposed for the full feedback model before, by sampling all the K arms in each round t so that the entire reward vector R^t is observed. From Theorem 2, this gives an upper bound on the total number of samples required of the form

$$\kappa^S \leq \text{O}\left(K\log(K/\delta)/\Delta^2_{\text{min}}\right). \tag{8}$$

In what follows, we utilize the special structure of the reward vector in any round to propose a policy whose sample complexity is much lower than the above upper bound for the naive strategy. Since the reward vector in any round is a sequence of zeroes followed by a sequence of ones, the agent might be able to recover the reward of an arm from the observations of other arms. For e.g., if in round t the agent observed the reward of arm i, R^t_i as 1, then the agent will also be able to infer that the reward of any arm $j \geq i$ in the same round, R^t_j will also be 1. Iterating using a simple binary search procedure, for any subset of arms S we can recover the rewards of all the arms in S in any round by sampling at most $\log_2(|S|+1)$ arms.

In Algorithm 2, we propose a scheme which proceeds in rounds and in each round t, uses the above binary search procedure to infer the components of the reward vector R^t corresponding to a subset of arms A_t referred to as 'active arms'. As the algorithm proceeds, the active set is updated by eliminating arms based on the reward samples observed thus far. For each active arm, we maintain an upper and lower confidence bound on its net utility and eliminate it whenever its upper confidence bound (UCB) becomes lower than the lower confidence bound (LCB) for another arm, thus indicating that it is not the best arm. Finally, the scheme terminates when a single active arm is left and this arm is declared as the estimate for the best arm. The detailed pseudocode for the scheme is provided in Algorithm 2. Note that we denote by GetSamples($\mathcal{A}_t$), the binary search procedure which recovers all the reward values in a particular round for arms in the set $\mathcal{A}_t$. The following result provides an upper bound on the sample complexity of the scheme in Algorithm 2.

Algorithm 2: Best arm identification with correlation

```
Input: $c_i$ for all $i$, $\delta$
Output: Best arm with probability 1-$\delta$
Initialization: $\hat{r}_i(0) = 0$, for all $i \in \{1 \cdots K\}$,
  $\mathcal{A}_1 = \{1 \cdots K\}$, t=1;
while $|\mathcal{A}_t| > 1$ do
    $\mathbf{R}^{\mathbf{t}}_{\mathcal{A}_\mathbf{t}}$ =GetSamples($\mathcal{A}_t$);
    for $i \in \mathcal{A}_t$ do
        update sample mean:
          $\hat{r}_i(t) = ((t-1)\hat{r}_i(t-1) + R^t_i)/t$;
        LCB$_i = \hat{r}_i(t) - c_i - \sqrt{\frac{1}{2t}\ln(K/\delta)}$;
        UCB$_i = \hat{r}_i(t) - c_i + \sqrt{\frac{1}{2t}\ln(K/\delta)}$;
    end
    for $j \in \mathcal{A}_t$ do
        if $LCB_i > UCB_j$, for any $i \neq j$ then
            eliminate arm $j$: $\mathcal{A}_{t+1} = \mathcal{A}_t/\{j\}$;
        else
            continue;
        end
    end
    t++;
    if $|\mathcal{A}_t| = 1$ then
        return $\mathcal{A}_t$;
    else
        continue;
    end
end
```

Theorem 3: Algorithm 2 outputs the best arm with probability at least $1-\delta$ under the sequential multi-arm feedback model and the total number of samples across rounds, κ^S, is upper bounded as

$$\kappa^S \leq \frac{8}{\Delta^2_{(2)}}\log(K/\delta) + \sum_{i=1}^{\gamma-1}\frac{8}{\Delta^2_{(2^i)}}\log(K/\delta).$$

with probability at least $1-\delta$, where $\gamma = \lceil \log_2(K+1) \rceil$.

Proof: Let $\hat{\mu}_i(n) = \left(\sum_{t=1}^n R^t_i\right)/n - c_i$ denote the empirical estimate of the expected net utility of arm i after n rounds and let $\mathcal{E}$ be the event that for all $\underline{1 \leq i \leq} K$, $n \geq 1$, we have $|\mu_i - \hat{\mu}_i(n)| \leq \epsilon_n$, where $\epsilon_n = \sqrt{\frac{\log(K/\delta)}{2n}}$. Using Hoeffding's inequality and the union bound, we get $\mathbb{P}(\mathcal{E}^c) \leq \delta$. Hereafter, we will assume that the event $\mathcal{E}$ holds, and show that the algorithm will output the best arm.

In each round of the algorithm, the scheme uses the GetSamples($\mathcal{A}_t$) module to recover the reward vector components $\{R^t_i\}_{i\in\mathcal{A}_t}$ using at most $\log_2(|\mathcal{A}_t|+1)$ samples. Thus, the number of samples in a round will be at most m once the number of active arms $|\mathcal{A}_t| < 2^m$. From the proof of Theorem 2, we have that the 2^m-th best arm and all the arms worse than that will be eliminated by $\frac{8}{\Delta^2_{(2^m)}}\log(K/\delta)$ rounds. Thus, for any round t such that $\frac{8}{\Delta^2_{(2^{m+1})}}\log(K/\delta) < t < \frac{8}{\Delta^2_{(2^m)}}\log(K/\delta)$, the number of samples needed by the GetSamples($\mathcal{A}_t$) module to recover the reward vector components of all the active arms will be at most $m+1$. Setting $\lceil \log_2(K+1) \rceil = \gamma$, the total number of samples across all rounds, κ^S, is given by

$$\kappa^S \leq \gamma \frac{8\log(K/\delta)}{\Delta^2_{(2^{\gamma-1})}} + \sum_{i=1}^{\gamma-2} (\gamma - i) \left[\frac{8\log(K/\delta)}{\Delta^2_{(2^{\gamma-i-1})}} - \frac{8\log(K/\delta)}{\Delta^2_{(2^{\gamma-i})}} \right]$$

$$= \frac{8}{\Delta^2_{(2)}} \log(K/\delta) + \sum_{i=1}^{\gamma-1} \frac{8}{\Delta^2_{(2^i)}} \log(K/\delta).$$

■

Note that the sample complexity of the proposed algorithm can be much smaller than that of the naive scheme in (8) which samples all the K arms in each round. Also, notice that the lower bound on the sample complexity in Theorem 1 for the full information feedback setting is also a valid lower bound for the sequential multi-arm feedback setting. Comparing the expression in the lower bound and the upper bound on the sample complexity of the proposed scheme derived above, we see that Algorithm 2 is order-wise optimal when $\sum_{i=2}^{\gamma-1} \frac{1}{\Delta^2_{(2^i)}} \ll \frac{1}{\Delta^2_{(2)}}$.

IV. Simulations

In this section, we compare the performance of our proposed policy in Algorithm 2 with various other policies mentioned below via simulations.

A. Alternate policies

Active-P: This scheme is similar to the naive strategy, but instead of sampling all the K arms in each round, we only sample the arms in the active set. While this scheme samples fewer arms in each round, it still does so in parallel unlike our proposed scheme in Algorithm 2 which uses a sequential strategy to further reduce the number of sampled arms.

LUCB: For the standard MAB setting a popular scheme for best arm identification is the LUCB, which in each round samples two arms: one with the highest sample mean and the other with the highest upper confidence bound amongst the remaining arms. For a target error probability δ, the total sample complexity of LUCB is given by $O\left(\sum_{i\neq i^\star} \frac{1}{\Delta_i^2} \log\left(\frac{K}{\delta}\log(1/\Delta_i^2)\right)\right)$ [2]. This scheme does not explicitly use the correlation amongst the rewards of different arms in a round, and provides a useful baseline to compare our proposed scheme against.

CLUCB: Recently, [5] proposed the CLUCB algorithm for the correlated MAB setting where additional side-information is available in the form of upper bounds on the conditional expected rewards of the arms. In particular, in any round t and for any pair of arms l, k, we have $\mathbb{E}[R_l^t | R_k^t = r] \leq s_{l,k}(r)$ and $s_{l,k}(r)$ is a known upper bound on the expected reward of arm l conditioned on the event that the reward for arm k is r. This algorithm explicitly uses the correlation in the arm rewards to reduce the number of samples required for best arm identification, and achieves a sample complexity of $O\left(\sum_{k\in\mathcal{C}} \frac{1}{\Delta_k^2} \log\left(\frac{2K}{\delta}\log(1/\Delta_k^2)\right)\right)$ where $\mathcal{C} \subseteq \mathcal{K}$ represents a set of 'competitive' arms [5].

To use CLUCB in our setting, we define the upper bounds on conditional expected rewards as $s_{l,k}(r = 1) = 1$, and $s_{l,k}(r = 0)$ equals 0 for $l \leq k$ and 1 otherwise. The intuition behind this is that in any round, if the reward for arm index k is 0, then the reward for all arms with indices smaller than k will also be 0. However, there is no additional information on the rewards of the arms with indices higher than k.

To compare the performance of our proposed scheme (Algorithm 2) with these alternative policies, we consider the application of identifying the best version of a Software as a Service (SaaS) discussed in the introduction. Recall that the SaaS is offered in a range of versions, where Version i has a associated reward (with expected value r_i) and cost (c_i). The cost of each version is known while its expected reward is unknown and has to be estimated via user feedback. Further, it is known that on any sample path, a version with higher cost has higher reward. The goal is to identify the version with the highest utility ($r_i - c_i$) with probability at least $1 - \delta$. For our experiments, we fix $\delta = 0.01$.

We first discuss how we construct problem instances to evaluate the performance of various candidate policies. In Figure 1, each value of $\rho \in [0, 1]$ on the x-axis corresponds to a candidate arm. The cost of an arm is equal to $c \times \rho$ for a given constant $c > 0$. Further, $r(\rho)$ denotes the average reward corresponding to this arm. It follows that the cost and expected reward are increasing functions of ρ, thus satisfying the structure of our problem. We consider two types of simulation instances.

Synthetic parameters: In this set of experiments, the values of the r_is and c_is are obtained from a synthetically generated piece-wise linear function shown in Figure 1. For our experiment, we choose the arms such that they are equally spaced in the interval $[0, 1]$.

Data-driven parameters: In this set of experiments, the values of the r_is and c_is are obtained by using the GPS trajectory dataset [11]–[13] collected as a part of the Geolife Project by Microsoft Research Asia. More details on the dataset and how it is used to obtain these parameter values can be found in [14] and Section 7.2 of [15].[2] We choose the arms such that the best arm $\rho^*(= \arg\max_{\rho\in[0,1]} r(\rho) - c \times \rho)$ is always included and no arm is chosen in the interval $(\rho^* - 0.063, \rho^* + 0.063)$. The remaining arms are equally spaced in the intervals $(0, \rho^* - 0.063)$ and $(\rho^* + 0.063, 1)$.

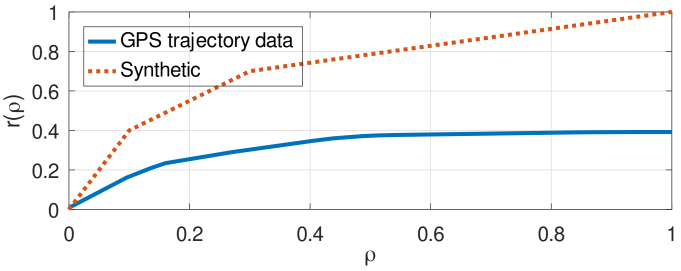

Fig. 1: Synthetic and data-driven reward functions

We begin by studying the impact of the value of c on the performance of the various policies. We select 33 values of c_is in $[0, 1]$ including 0 and 1. We plot the average sample complexity (over 10 runs) of Algorithm 2 (Adapt-P) and various alternate policies discussed in Section IV-A for the synthetic and data-driven parameter values in Figures 2 and

[2] The plot corresponding to the data-driven $r(\rho)$ in Figure 1 actually corresponds to the concave hull of the curve obtained in [14].

3 respectively. We observe that our proposed policy uses the lowest number of samples amongst all the policies.

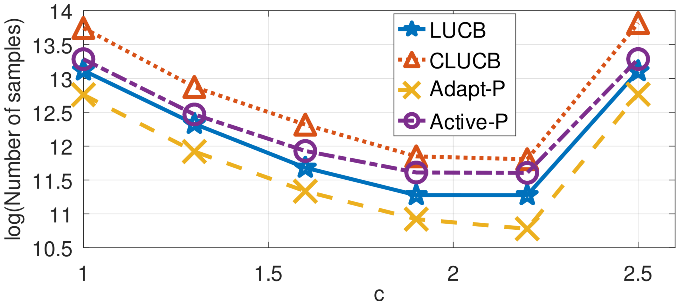

Fig. 2: Performance of various policies as a function of c for synthetic parameters

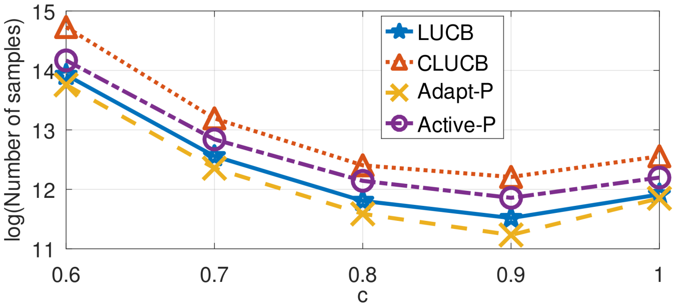

Fig. 3: Performance of various policies as a function of c for data-driven parameters

Next, we fix the value of c and consider the variation of the sample complexity of the various schemes as we change the number of arms. For Figure 4 we fix $c = 1$, note that as we increase the number of arms, the minimum gap between the optimal and a sub-optimal arms decreases. As expected from the sample complexity characterization in Theorem 2, we see in Figure 4 that the sample complexity grows with increasing number of arms. Again, amongst all the policies, our proposed scheme in Algorithm 2 achieves the best performance. On the other hand, in Figure 5, we fix $c = 0.8$. As observed in the previous experiments, the sample complexity of our proposed scheme is the lowest among the various strategies.

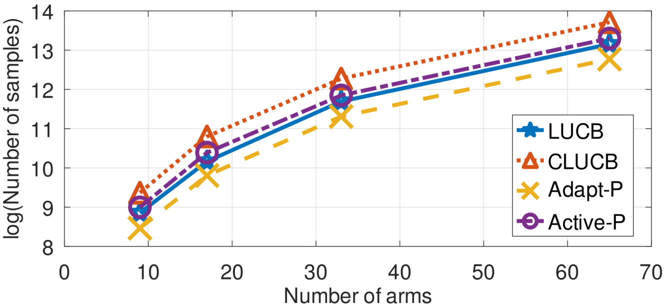

Fig. 4: Performance of various policies as a function of the number of arms for synthetic parameters

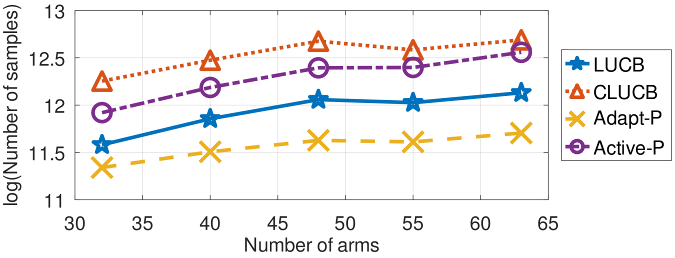

Fig. 5: Performance of various policies as a function of the number of arms for data-driven parameters

References

[1] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[2] K. Jamieson and R. Nowak, "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting," in *2014 48th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2014, pp. 1–6.

[3] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems," in *International conference on Algorithmic learning theory*. Springer, 2009, pp. 23–37.

[4] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[5] S. Gupta, G. Joshi, and O. Yağan, "Best-arm identification in correlated multi-armed bandits," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 2, pp. 549–563, 2021.

[6] M. Soare, A. Lazaric, and R. Munos, "Best-arm identification in linear bandits," *Advances in Neural Information Processing Systems*, vol. 27, pp. 828–836, 2014.

[7] C. Tao, S. Blanco, and Y. Zhou, "Best arm identification in linear bandits with linear dimension dependency," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4877–4886.

[8] A. Kazerouni and L. M. Wein, "Best arm identification in generalized linear bandits," *Operations Research Letters*, vol. 49, no. 3, pp. 365–371, 2021.

[9] T. Kocák and A. Garivier, "Best arm identification in spectral bandits," *arXiv preprint arXiv:2005.09841*, 2020.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006.

[11] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on gps data," in *Proceedings of the 10th international conference on Ubiquitous computing*, 2008, pp. 312–321.

[12] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 791–800.

[13] Y. Zheng, X. Xie, W.-Y. Ma *et al.*, "Geolife: A collaborative social networking service among user, location and trajectory." *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, 2010.

[14] M. Agarwal, url=https://github.com/mohit-iitb/mapDataCaching, [Online; accessed 23-January-2022].

[15] V. S. C. L. Narayana, M. Agarwala, R. S. Prakash, N. Karamchandani, and S. Moharir, "Online partial service hosting at the edge," *CoRR*, vol. abs/2103.00555, 2021. [Online]. Available: https://arxiv.org/abs/2103.00555