Multi-armed Bandits with Probing

Eray Can Elumar
Dept. Electrical & Computer Engineering
Carnegie Mellon University
Pittsburgh, PA, USA
Email: eelumar@andrew.cmu.edu

Cem Tekin
Dept. Electrical & Electronics Engineering
Bilkent University
Ankara, Turkey
Email: cemtekin@ee.bilkent.edu.tr

Osman Yağan
Dept. Electrical & Computer Eng.
Carnegie Mellon University
Pittsburgh, PA, USA
Email: oyagan@andrew.cmu.edu

***Abstract*—We examine a K-armed multi-armed bandit problem involving probes, where the agent is permitted to probe one arm for a cost $c \geq 0$ to observe its reward before making a pull. We identify the optimal strategy for deciding whether to probe or pull an arm. In the case of probing an arm, we also make a decision on which arm to pull after observing the probe's outcome. Additionally, we introduce a novel regret definition based on the expected reward of the optimal action. We propose UCBP, a novel algorithm that utilizes this strategy. UCBP achieves a gap-independent regret upper bound in T rounds that scales with $\mathcal{O}(\sqrt{KT \log T})$, and an order optimal gap-dependent upper bound that scales with $\mathcal{O}(K \log T)$. We provide UCB-naive-probe, a naive UCB-based approach which has a gap-independent regret upper bound on the order of $O(K\sqrt{T \log T})$, and gap-dependent regret on the order of $O(K^2 \log T)$ as a baseline. We provide empirical simulations to verify the utility of the UCBP algorithms in practical settings, and show that UCBP outperforms UCB-naive-probe in simulations.**

I. Introduction

The multi-armed bandit problem is a classic dilemma in decision-making under uncertainty. The metaphorical *bandit* refers to a set of slot machine arms, and the *multi-armed* aspect indicates that there are multiple such arms available for the decision-maker to choose from. Each arm of the bandit is associated with an unknown probability distribution of rewards. The goal of the decision-maker, often referred to as an agent, is to *maximize* the cumulative reward over a sequence of trials or pulls. The seminal work of [1] showed that the *regret*, which is the difference in expected total rewards between a given policy and the *optimal policy*, grows at least logarithmically in the number of plays, and developed asymptotically optimal decision policies. Since then, other policies have also been devised, including [2], [3], and used in applications in many fields, such as online advertising [4], [5], clinical trials [6], [7], recommendation systems [8].

In the past few years, there has been a notable increase in interest surrounding MABs, driven by the demand for proficient and impactful decision-making across diverse domains. A recent avenue of exploration involves bandits with side information, enabling the agent to obtain additional information before making a decision. [9]–[11]. The side information may manifest as partial observations, expert guidance, context information, or prior knowledge pertaining to the reward distributions. In this paper, we consider a specific variant of this problem, namely multi-armed bandits with *probes*.

The concept of employing probing to diminish uncertainty in the decision-making process has been explored across various research domains, such as wireless communication systems [12], stochastic probing [13], online learning [14], and multi-armed bandits [15], [16]. In our scenario, the agent is provided with the opportunity to probe an arm by incurring a cost $c \geq 0$ to observe its reward before actually pulling that arm. This modification to the MAB problem introduces an added layer of complexity and difficulty, given that probing significantly enlarges the set of possible actions. In scenarios employing costly expert advice, whether from humans or machine learning models acting as experts, the act of probing can be understood as obtaining a reward prediction from the expert without actually pulling the arm. The main goal of our work is to develop new algorithms for this framework that achieve as much *cumulative* reward as possible.

A. Applications

The problem setting under consideration has diverse applications across various fields. One example is hyperparameter optimization for machine learning models, where a common strategy involves having human experts regularly examine learning curves to promptly end runs with suboptimal settings [17]. In this setting, "pulling an arm" can be understood as executing the hyperparameter setting without human expert oversight and "probing an arm" can be understood as executing it with supervision. This approach ensures that runs with poor settings are quickly terminated, preventing regret from probes and incurring only the probing cost, which reflects the expense of involving a human expert.

Another example is online learning with machine learning (ML) advice where ML models are used to predict the outcomes of actions before actually taking an action [18]–[20] in order to improve the quality of action taken in settings such as when the predictions are perfect [21], when the predictions are adversarial [22], or when there is an upper limit on the error of the predictions [23]. While in this work we make the assumption that a probe provides the precise outcome of an arm, we assign a cost to probing that can be employed to capture the computational complexity associated with using ML predictions. This work is also valuable as a foundation for future work that may relax this assumption, incorporating scenarios where probes yield noisy reward predictions.

Another possible application is in wireless communications. Probing is mainly used there to send small data packets to observe some channel properties at that time. Prior work generally assume knowing the distributions of the rewards of channels [24]. Our work can be especially useful when these distributions are unknown.

In queuing, such as when queuing patients in the emergency room of a hospital according to urgency of the problem, probing can be represented as the hospital staff doing small tests to determine the urgency of the situation, and arm pulls can be represented as the patient actually being treated by the doctor.

B. Contributions

1) **Formulation:** To our knowledge, this work is the first to consider a multi-armed bandit setting with bounded reward distributions where before pulling an arm, the agent is allowed to probe one arm to observe its reward for a cost $c \geq 0$. This is an intricate problem different from most previous bandit formulations as the action set is larger.
2) **UCBP Algorithm:** We identify the optimal strategy, and provide an order-optimal algorithm based on UCB that evaluates the value of each action and uses upper confidence bounds to balance exploration and exploitation.
3) **Regret Upper Bound for UCBP:** Through a novel decomposition of regret, we establish that the gap-independent regret upper bound scales with $O(\sqrt{KT\log T})$, and when the reward distribution is discrete, the gap-dependent regret bound scales with $O(K\log T)$. We also show that the gap-dependent regret upper bound is order-optimal by showing that the regret lower bound also scales with $\Omega(K\log T)$.

II. Multi-Armed Bandit Model with Probes

We consider a K-armed stochastic bandit problem with the set of base arms $[K]$, where arm $i \in [K]$ is associated with a fixed reward distribution Γ_i that is independent of the reward distributions of other arms with mean μ_i. At each round, the agent selects one of the following two types of actions. The first type of action, referred to as *pull*, is where the agent pulls arm $i \in [K]$ to receive its reward $r(t) = r_i(t)$ drawn from Γ_i. In the second type of action, referred to as *probe*, the agent selects a *probe arm* i and a *backup arm* $j \neq i$. The *probe arm* is probed to observe its reward $r_i(t)$, then; the agent can choose to pull the *probe arm* to receive reward $r(t) = r_i(t) - c$ or the *backup arm* to receive $r(t) = r_j(t) - c$ where $c \geq 0$ is the known cost of probing.

$\mathcal{A}_p$ is defined the set of actions that involve probing, and $\mathcal{A}_s$ as the set of actions that do not involve probing, and $\mathcal{A} := \mathcal{A}_s \cup \mathcal{A}_p$. The ordered tuple $(i,j) \in \mathcal{A}_p$ for $i,j \in [K],\ i \neq j$ indicates arm i is the probe arm and arm j the backup arm, while $(i,\emptyset) \in \mathcal{A}_s$ for $i \in [K]$ indicates pulling arm i. Clearly, $|\mathcal{A}| = K^2$. Further, $\mathcal{A}_{p,i} := \{a \in \mathcal{A}_p : i \in a\}$ is the set of actions that include base arm (either as probe or backup arm) i. When $a(t) = (i,j)$, after observing reward $r_i(t)$, the agent needs to decide whether to pull arm i or j. The optimal decision here is to pull arm i if $r_i(t) > \mu_j$, and arm j otherwise. We call this the *optimal reference point decision*. It can be seen that when this optimal strategy is used, the expected reward of playing action (i,j) is: $v_{(i,j)} = \mathbb{E}[\max(r_i, \mu_j)] - c$.

For simplicity, we assume there is a unique arm with the highest mean, referred to as the *best arm*. Without loss of generality, we assume that the mean rewards of the arms are ordered such that $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$. When $\Gamma_i,\ \forall i \in [K]$ are known *a priori*, the optimal reward is $\nu^* = \max(\mu_1, \max_{i\in[K]\setminus\{1\}}\{-c + \mathbb{E}[\max(r_i,\mu_1)]\}, -c + \mathbb{E}[\max(r_1,\mu_2)])$. Then, the optimal action is

$$a^* = \begin{cases} (1,\emptyset) & \text{if } \nu^* = \mu_1 \\ (i,1) & \text{if } \nu^* = -c + \mathbb{E}[\max(r_i,\mu_1)] \\ (1,2) & \text{if } \nu^* = -c + \mathbb{E}[\max(r_1,\mu_2)] \end{cases}$$

Unlike standard K-armed bandit, in our setup, the *probe* option makes the optimal action non-trivial as achieving negative regret is possible under the *probe* option if $\exists(i,j)$ s.t. $\mathbb{E}[\max(r_i,\mu_j)] - c > \mu_1$ when $T\mu_1$, the reward of the *best* arm is used as a benchmark. Hence, we define the empirical cumulative regret with respect to the optimal reward. Using this, the expected cumulative regret can be written as

$$\hat{R}_T := T\nu^* - \sum_{t=1}^{T} r(t), \text{ and } R_T := \mathbb{E}[\hat{R}_T].$$

Let ν_a represent the expected reward of action a. Then, $\nu_a = \mu_i$ when $a = (i,\emptyset),\ i \in [K]$. For $a = (i,j)$ such that $i,j \in [K]$ and $i \neq j$, $\nu_{(i,j)} = \mathbb{E}[\max(r_i,\mu_j)] - c$. The gaps of actions are defined as $\Delta_{(i,\emptyset)} := \nu^* - \nu_{(i,\emptyset)}$, and $\Delta_{(i,j)} := \nu^* - \nu_{(i,j)}$, and the gaps of base arms are defined as $\Delta_i := \mu_1 - \mu_i$. We remark that with this regret definition, identifying the probe arm and the backup arm correctly may not be sufficient to receive the optimal reward ν^*. To illustrate this, assume that $a^* = (i,1)$ for some $i \neq 1$. To receive $\nu^* = -c + \mathbb{E}[\max(r_i,\mu_1)]$, after probing arm i and observing r_i, the agent needs to pull arm i if $r_i > \mu_1$ or pull arm 1 if $r_i \leq \mu_1$. However, this optimal action can only be taken with the exact knowledge of the reference point μ_1, which the agent does *not* have. If one uses an estimate $\tilde{\mu}_1(t)$ of the reference point at round t, this will lead to an additional regret of up to $r_{\text{ref}}(t) := |\tilde{\mu}_1(t) - \mu_1|\mathbb{P}(r_i \in [\min(\mu_1,\tilde{\mu}_1(t)), \max(\mu_1,\tilde{\mu}_1(t))])$. We call the decision to pull arm i using $\tilde{\mu}_i(t)$ as the *reference point decision*, and the regret it introduces as the *reference point regret*. $R_{\text{ref}}(T) := \sum_{t=1}^{T} r_{\text{ref}}(t)$ is used to denote the regret incurred until round T due to the *reference point error*. We first present a naive UCB-based algorithm, which treats the reference point as part of the action it takes to serve as baseline.

UCB-naive-probe algorithm: We treat each action triple as a super arm. $a = (i,j,d_l) \in \mathcal{A}_N,\ i \in [K],\ j \in [K] \setminus \{i\}$ denotes that the probe arm is arm i, the backup arm is arm j, and the reference point is d_l. $\mathcal{A}_N$ denotes the action set for this algorithm. UCB-naive-probe algorithm can only be used when the reward distributions of the arms are discrete for the set of super arms to be finite. Hence, we assume that $\mathcal{D}$ in $[0,1]$ is the finite support of the rewards of the arms, and that $d_l \in \mathcal{D}$ are the elements of $\mathcal{D}$ (excluding the smallest one) where $2 \leq l \leq |\mathcal{D}|$. The actions $a = (i,\emptyset,\emptyset),\ i \in [K]$ denote pulling arm i. We use standard UCB indices for all

Algorithm 1 UCB-naive-probe

```
for each round $t$ do
    $a_t = (i_t, j_t, d(t)) = \arg\max_{a \in \mathcal{A}} U_a(t)$
    if $j_t = \emptyset$ then
        Pull arm $i_t$, get $r(t) = r_t(i_t)$
    else
        Probe arm $i_t$, observe reward $r_t(i_t)$
        if $r_t(i_t) \geq d(t)$ then
            Pull arm $i_t$, get $r(t) = r_t(i_t) - c$
        else
            Pull arm $j_t$, get $r(t) = r_t(j_t) - c$
        end if
    end if
    Update UCB indices and mean estimates
end for
```

TABLE I
COMPARISON OF OUR WORK WITH PRIOR WORK ON BANDITS WITH PROBES

Work	Probe Model	Reward Distr.	Regret Defn.
[26]	Can probe multiple arms, can pull any arm, $c \geq 0$	Bernoulli	Opt. policy
[27]	Probe 2 arms, pull the one with highest reward, $c = 0$	Bounded	Best arm
[27]	Probe 3 arms, pull the one with highest reward, $c = 0$	Bounded	Best arm
Ours	Can probe one arm, can pull any arm, $c \geq 0$	Bounded	Opt. action

super arms, and the arm with the highest UCB index is pulled each round. When a *super arm* (i, j, d_l) is chosen and $r_i(t)$ is observed by probing; arm i is pulled if $r_i(t) \geq d_l$, and j is pulled otherwise. The pseudo-code is given in Algorithm 1.

It is shown in the long version of the manuscript that the gap-dependent regret scales with $\mathcal{O}(|\mathcal{D}|K^2 \log T)$, and the gap-independent regret scales with $O(\sqrt{|\mathcal{D}|K^2 T \log T})$ [25]. This dependence of regret on $\tilde{O}(|\mathcal{D}|K^2)$ demonstrates the complexity of the problem. The main goal of our paper is to decrease this dependency of regret on K and $|\mathcal{D}|$ from $\tilde{O}(|\mathcal{D}|K^2)$ to $\tilde{O}(K)$ by utilizing the optimal action and optimal reference point decision strategies described above. Our algorithm that achieves this reduction in regret is in §IV.

III. Related Works

Bandits with Probes: To our knowledge, probes were first studied in the context of bandits with expert advice in [16] featuring multiple experts, where the agent, after pulling an arm, has the ability to observe the reward of any chosen subset of arms by incurring a cost c for each observed arm. More recently, there has been consideration of the bandit with probes problem specifically for Bernoulli reward distributions [26], where an unlimited number of probes are allowed per round, but each probe has a cost. They suggest an algorithm that attains $O(K^2 \log T)$ gap-dependent regret by employing a strategy that probes arms in order of their highest UCB value to the lowest. In our work, while we permit only one probe, we consider a more general bounded reward distribution which necessitates a more intricate strategy, and we achieve $O(K \log T)$ regret. In [27], two different models are studied for probes without cost. In the first model, at each round, two arms are probed, and the probe reveals the arm with the higher reward, which must be pulled. The proposed UCB-based algorithm achieves $O(K^2 \log T)$ gap-independent regret. However, the regret is defined on the base arm with highest mean, and not on the optimal super arm. This regret bound follows mainly due to this regret definition, since it is even possible to achieve negative regret when $\max(r_i, r_j)$, the reward of super arm (i, j), is larger than μ_1. In the second model, three arms are probed each round, and one of them is pulled. The provided algorithm achieves $O(K^2)$ regret. In this paper, we explore a comparable scenario where the option to probe is limited to at most one arm, but any arm is permitted to be pulled after probing. We also define regret based on the *optimal action*. Comparison of our work with prior work is summarized in Table I.

Probes in Wireless Communications: One notable prior work related to ours is [24]. Here, a wireless system is considered where the reward distribution of each channel is known *a priori*. It is allowed to probe multiple channels to reveal its reward before selecting a channel, but there is a cost for each probe. The main difference of [24] from our work is that the reward distributions of the arms are unknown in our setting. Other prior work on probing in wireless communication systems include [12], [28]–[30].

Combinatorial bandits: It is an extension of the standard bandit framework where an action is composed of a combination of different base arms satisfying certain constraints [31], [32]. One work that is of interest is the combinatorial bandits with probabilistically triggered arms [33], where when an action is played, a random subset of arms is triggered according to a triggering probability distribution. They show in [33, Theorem 3] that the regret lower bound scales with the factor $\frac{1}{p^*}$ for the general combinatorial bandits with probabilistically triggered arms. They also show that regret bounds that do not depend on p^*, but do depend on B, the bounded smoothness constant, if the problem setting satisfies the *triggering probability modulated bounded smoothness* assumption.

IV. The UCBP Algorithm

We propose an algorithm called *Upper Confidence Bound with Probes* (UCBP) that utilizes the structure of the action set and expected rewards to minimize the regret. The pseudo-code is provided in Algorithm 2. At each round t, first, the empirical mean rewards of arms are determined using $\hat{\mu}_i(t) = \sum_{\tau=1}^{t-1} r_i(\tau)\mathbb{1}\{i \in S(\tau)\}/N_i(t)$ where $S(t)$ denotes the set of arms whose reward is observed (by either pulling or probing) in round t and $N_i(t)$ denotes the number of times arm i is observed by round t. The UCB indexes for each arm i are computed as $U_i(t) = \hat{\mu}_i(t) + C_{(i,\emptyset)}(t)$ where $C_{(i,\emptyset)}(t) = \sqrt{3\log(t)/N_i(t)}$. Then, the probe UCB indexes are evaluated for probe actions by using $P_{i,j}(t) = \hat{\nu}_{(i,j)}(t) + C_{(i,j)}(t)$, where $\hat{\nu}_{(i,j)}(t) = \sum_{\tau=1}^{t-1}(\max(r_i(\tau), \hat{\mu}_j(t))\mathbb{1}\{i \in S(\tau)\})/N_i(t) - c$, and $C_{(i,j)}(t) = \sqrt{3\log(t)/N_j(t)} + \sqrt{3\log(t)/N_i(t)}$. Here $\hat{\nu}_{(i,j)}$ represents the empirical mean reward of action (i, j),

Algorithm 2 UCBP

```
Sample each base arm once
for each round $t$ do
    $i_t^* = \arg\max_i U_i(t)$
    $a_t = (j_t, k_t) = \arg\max_{a \in \mathcal{A}_p} P_a(t)$
    if $U_{i_t^*}(t) > P_{a_t}(t)$ then
        Pull arm $i_t^*$, get $r(t) = r_t(i_t^*)$
    else
        Probe arm $j_t$, observe reward $r_t(j_t)$
        if $r_t(j_t) > U_{k_t}(t)$ then
            Pull arm $j_t$, get $r(t) = r_t(j_t) - c$
        else
            Pull arm $k_t$, get $r(t) = r_t(k_t) - c$
        end if
    end if
    Update UCB indices for all arms
end for
```

and $C_{(i,j)}(t)$ is the exploration bonus associated with action (i, j). Lastly, the UCB indexes of the actions $U_i(t),\ \forall i \in [K]$; and $P_a(t),\ \forall a \in \mathcal{A} \setminus [K]$ are compared and the one with highest UCB index is chosen. If this action is probing, i.e. $a = (i, j)$, arm i is probed to observe $r_i(t)$, then arm i is pulled if $r_i(t) > U_j(t)$, and arm j otherwise. In other words, UCBP uses $U_j(t)$ as the reference point $\tilde{\mu}_j(t)$ at round t.

A. Analysis of UCBP

To characterize the performance of UCBP, we provide theoretical upper and lower bounds on the expected cumulative regret. We first state a mild assumption on the reward distributions of the arms that are required for the theoretical analysis. We refer the readers to the longer version of the manuscript for proofs of the presented results [25].

Assumption 1. For each Γ_i and Γ_j, $i, j \in [K],\ i \neq j$, we have $\mathbb{P}(r_i \leq \mu_j) \geq \epsilon$ for some $\epsilon > 0$.

Assumption 1 ensures that the backup arm is pulled at least ϵ fraction of the time by UCBP in expectation when a probe action is taken since $\epsilon \leq \mathbb{P}(r_i \leq \mu_j) \leq \mathbb{P}(r_i \leq U_j(t))$ if the confidence bounds hold. This assumption is needed, since if for some arm $j \in [K]$ the gap of actions $(j, \cdot)$ and $(j, \emptyset)$ are much larger than the gap of the actions $(\cdot, j)$; then the algorithm will predominantly choose actions $(\cdot, j)$, meaning arm j will only be selected as the backup arm, which might not produce enough samples for arm j. This assumption is similar to p^* in combinatorial bandits with probabilistically triggered arms, where p^* is defined as the minimum positive probability that an arm is triggered by any action [33]. In [33], to remove the dependency of regret on p^*, the *triggering probability modulated bounded smoothness* assumption is used. This assumption states that the change in the reward of an action if the mean vector is perturbed by a given amount is upper bounded by the triggering probability times the bounded smoothness coefficient B and the amount of perturbation, which enables regret to be upper bounded by an expression that contains the triggering probability. Using this assumption, regret bounds that do not depend on p^*, but do depend on B, are proved for combinatorial bandit problems that satisfy this assumption. However, they also prove in [33, Theorem 3] that in settings where this assumption is not necessarily satisfied, regret lower bound scales with the factor $\frac{1}{p^*}$, demonstrating that the $\frac{1}{p^*}$ factor in regret bound cannot be avoided without making additional assumptions. In our setting, this *triggering probability modulated bounded smoothness* assumption cannot be used, as observing the backup arm in a probe action is not an event with a constant probability, but rather a choice of the algorithm that depends on the reward distribution of the probe arm, and on the estimated mean of the backup arm.

Theorem IV.1 (Gap-independent Expected Regret Upper Bound)**.** *Under Assumption 1, when UCBP is run on the action set $\mathcal{A}$ and the cost of probing is $c \geq 0$, its cumulative expected regret is upper bounded as*

$$R_T \leq \frac{4\sqrt{6KT\log T}}{\epsilon} + R_{\textit{ref}}(T) + \frac{2\pi^2 K^2}{3} + K$$

where $R_{\textit{ref}}(T)$ is the reference point regret.

In Lemma IV.3, we show that $R_{\text{ref}}(T) = O(\sqrt{KT\log T})$, which together with Theorem IV.1 shows that the gap-independent regret of UCBP is $O(\sqrt{KT\log T})$.

Theorem IV.2 (Gap-dependent Expected Regret Upper Bound)**.** *Under Assumption 1, when UCBP is run on $\mathcal{A}$ and given $c \geq 0$, the expected cumulative regret is upper bounded as*

$$R_T \leq \sum_{i=1}^{K} \frac{12\log T}{\delta_i} + R_{\textit{ref}}(T) + \frac{2\pi^2 K^2}{3} + K, \textit{ where}$$

$$\delta_i = \begin{cases} \rho_i/8 & \textit{if } a^* = (i, \emptyset) \\ \min(\rho_i, \Delta_{(i,\emptyset)})/9 & \textit{otherwise} \end{cases}$$

where $\rho_i = \min_{a \in \mathcal{A}_{p,i} \setminus \{a^*\}} (\epsilon\Delta_a)$.

Note that the cost of probing c is included in the gap of actions. In Lemma IV.3, we show that $R_{\text{ref}}(T) = O(K\log T)$ when the rewards are *discrete*. Together with Theorem IV.2, this shows that the gap-dependent regret of UCBP is $O(K\log T)$ under discrete rewards.

We now provide upper bounds on *reference point regret*, which is incurred as the algorithm only uses the estimated means instead of the true means in the *reference point decision*. We show that for arbitrary reward distributions, $R_{\text{ref}}(T) = O(\sqrt{KT\log T})$, while tighter upper bounds can be established with additional assumptions on reward distributions.

Lemma IV.3. a) $R_{\text{ref}}(T) \leq \frac{4\sqrt{3KT\log T}}{\epsilon}$.
b) If the distributions Γ_i for each $i \in [K]$ are defined over a *discrete* support $\mathcal{D}$ in $[0, 1]$, then $R_{\text{ref}}(T)$ is upper bounded as $R_{\text{ref}}(T) \leq \sum_{i=1}^{K} 24\log T/(\epsilon\gamma_i)$ where we use $d_l \in \mathcal{D},\ 1 \leq l \leq |\mathcal{D}|$ to denote the elements of the set $\mathcal{D}$; and we let $\gamma_i := \min_l |d_l - \mu_i|$ if $\mu_i \notin \mathcal{D}$, and $\gamma_i := \min_l |d_l - d_{l+1}|$ if $\mu_i \in \mathcal{D}$. It can be seen that $\gamma_i > 0$ always holds. Under this assumption, it can be seen that the gap-dependent regret upper bound is $O(K\log T)$.

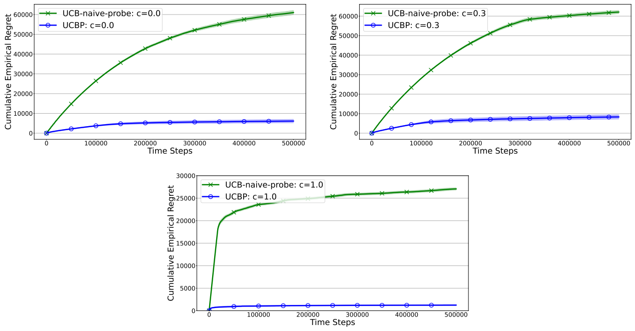

Fig. 1. Plots of the cumulative empirical regret of the UCBP and UCB-naive-probe algorithms for recommending the best genre in the MovieLens dataset.

Theorem IV.4 (Lower Bound on Expected Regret)**.** *For the multi-armed bandit setting with costly probes where the optimal action is unique, the lower bound on the expected cumulative regret for any* uniformly good *algorithm, as defined in [1], is* $\liminf_{T\to\infty} \frac{R_T}{\log T} \geq C(\Gamma)$*, where* $C(\Gamma)$ *is the minimal value of the following linear optimization problem:*

$$\min_{b_a \geq 0,\ \forall a \in \mathcal{A}\setminus\{a^*\}} \sum_{a\in\mathcal{A}\setminus\{a^*\}} b_a \Delta_a$$

$$\textit{s.t. } \forall i \in [K], \sum_{a\in\mathcal{A}_i, a\neq a^*} b_a \geq \left[\min_{a\in\mathcal{A}_i, a\neq a^*}\{D_{KL}(\Gamma_a||\Gamma^*)\}\right]^{-1}$$

where $\mathcal{A}_i = \{(i,j) : j \in ([K]\cup\{\emptyset\})\setminus\{i\}\}\cup\{(j,i) : j\in[K]\setminus\{i\}\}$*,* $\Gamma_{(i,\emptyset)} = \Gamma_i$*,* $\Gamma_{(i,j)} = \max(r_i, \mu_j) - c$ *is the distribution function of action* (i,j) *for* $i \neq j$*,* Γ^* *is the distribution function of* a^**, and* $D_{KL}(\cdot||\cdot)$ *is the KL divergence.*

It can be seen that UCBP regret lower bound is $\Omega(K\log T)$ since $C(\Gamma)$ is $\Omega(K)$. As the regret is also $O(K\log T)$ under discrete rewards in Theorem IV.2, excluding the ϵ term, we can conclude that the gap-dependent upper bound of UCBP is order-wise optimal.

B. Discussion of the Results

To our knowledge, this work is the first to consider a multi-armed bandit setting with arbitrary bounded reward distributions where before pulling an arm, the agent is allowed to probe an arm to observe its reward for a cost $c \geq 0$. This problem setting is intricate and distinct from the majority of prior bandit formulations, primarily owing to the extensive range of K^2 involved actions. Further, the possibility of still incurring regret due to the *reference point error* even when the chosen action is optimal creates additonal challenges.

Compared to UCB-naive-probe, regret of UCBP scales with $\tilde{O}(K)$ since UCBP narrows down the action space by utilizing the structure of the problem. Due to the probabilistic observability of the backup arm, we incur an additional $1/\epsilon$ term in regret, but this is in line with the lower bound given in [33, Theorem 3]. We note that even though we assume cost of probing c as a constant for simplicity of the theoretic analysis, this work can easily be extended to the setting where c is time dependent or cost of probing c_i is different for each arm i.

C. Simulations

We now provide simulation results to characterize the performance of UCBP in a real world setting. Since to our knowledge, there are no other bandit algorithms for our specific problem setting, we compare our results with the results from the UCB-naive-probe algorithm which we introduced in §II.

The MovieLens Dataset: The MovieLens dataset contains a total of 1M ratings on a total of 3883 movies, where users rated the movies on a scale of 1 to 5 [34]. We study the problem of providing the best genre recommendations to a population with an unknown demographic in this dataset. Each genre is modeled as an arm (there are $K = 18$ arms), and the reward of an arm is obtained by randomly sampling the rating of one of the users for a movie in that genre. The average reward of the best action is around 4.17. In Figure 1, we plot the cumulative regret averaged over 100 independent trials for $500,000$ rounds when the cost of probing is $c = 0$, $c = 0.3$, and $c = 1$. The shaded area represents error bars with one standard deviation. It can be seen that both algorithms have a logarithmic regret curve, and UCBP outperforms the baseline UCB-naive-probe algorithm. Further simulation results on the Open Bandit Dataset is provided in the long version of the manuscript [25].

V. Concluding Remarks

In this paper, we focus on problem setting of the multi-armed bandits with probes. We introduce a new regret definition that is based on the expected reward of the optimal action, and we identify the optimal strategy that attains this reward. We provide UCBP, a novel algorithm that utilizes this strategy to achieve a gap-independent regret upper bound that scales with $\mathcal{O}(\sqrt{KT\log T})$, and a gap-dependent bound that scales with $\mathcal{O}(K\log T)$ if rewards are discrete. To demonstrate the empirical performance of UCBP in simulations, we introduce a naive UCB-based algorithm as a baseline. Simulation results corroborate the better performance of UCBP over UCB-naive-probe, and validate the utility of UCBP in practical settings.

Our work opens multiple directions for future research. One interesting future direction is to extend our bandit results to the case with noisy probes. We believe that confidence intervals of probe rewards can be derived in this setting that can then be used to decide whether to pull the probe arm or the backup arm. Another interesting open direction is the case where the rewards of different arms are correlated. In this case, the correlation between arms can be used to predict the rewards of the other arms from the probe outcome, thereby providing more utility to the probes.

Acknowledgment

This work was supported in part by the National Science Foundation through Grant # CCF-2007834, by the Office of Naval Research through Grant # N00014-23-1-2275, and by the CyLab Enterprise Security Initiative. Cem Tekin acknowledges the support by the Turkish Academy of Sciences Distinguished Young Scientist Award Program (TÜBA-GEBİP-2023).

References

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

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, pp. 235–256, 2002.

[3] R. Agrawal, "Sample mean based index policies with o(log n) regret for the multi-armed bandit problem," *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054–1078, 1995.

[4] E. M. Schwartz, E. T. Bradlow, and P. S. Fader, "Customer acquisition via display advertising using multi-armed bandit experiments," *Marketing Science*, vol. 36, no. 4, pp. 500–522, 2017.

[5] D. Chakrabarti, R. Kumar, F. Radlinski, and E. Upfal, "Mortal multi-armed bandits," *Advances in Neural Information Processing Systems*, vol. 21, 2008.

[6] Y. Varatharajah and B. Berry, "A contextual-bandit-based approach for informed decision-making in clinical trials," *Life*, vol. 12, no. 8, 2022.

[7] W. H. Press, "Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research," *Proceedings of the National Academy of Sciences*, vol. 106, no. 52, pp. 22387–22392, 2009.

[8] N. Silva, H. Werneck, T. Silva, A. C. Pereira, and L. Rocha, "Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions," *Expert Systems with Applications*, vol. 197, p. 116669, 2022.

[9] T. Lu, D. Pál, and M. Pál, "Contextual multi-armed bandits," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 485–492, 2010.

[10] S. Mannor and O. Shamir, "From bandits to experts: On the value of side-observations," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[11] J. Langford and T. Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information," *Advances in Neural Information Processing Systems*, vol. 20, 2007.

[12] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1058–1076, 2010.

[13] A. Gupta and V. Nagarajan, "A stochastic probing problem with applications," in *Integer Programming and Combinatorial Optimization: 16th International Conference, IPCO 2013, Valparaíso, Chile, March 18-20, 2013*, pp. 205–216, 2013.

[14] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz, "Minimizing regret with label efficient prediction," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 2152–2162, 2005.

[15] Y. Efroni, N. Merlis, A. Saha, and S. Mannor, "Confidence-budget matching for sequential budgeted learning," in *International Conference on Machine Learning*, pp. 2937–2947, 2021.

[16] Y. Seldin, P. Bartlett, K. Crammer, and Y. Abbasi-Yadkori, "Prediction with limited advice and multiarmed bandits with paid observations," in *International Conference on Machine Learning*, pp. 280–287, 2014.

[17] A. Klein, S. Falkner, J. T. Springenberg, and F. Hutter, "Learning curve prediction with bayesian neural networks," in *International Conference on Learning Representations*, 2017.

[18] S. Gollapudi and D. Panigrahi, "Online algorithms for rent-or-buy with expert advice," in *International Conference on Machine Learning*, pp. 2319–2327, 2019.

[19] E. Bamas, A. Maggiori, and O. Svensson, "The primal-dual method for learning augmented algorithms," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20083–20094, 2020.

[20] K. Anand, R. Ge, A. Kumar, and D. Panigrahi, "Online algorithms with multiple predictions," in *International Conference on Machine Learning*, pp. 582–598, 2022.

[21] S. Wang, J. Li, and S. Wang, "Online algorithms for multi-shop ski rental with machine learned advice," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8150–8160, 2020.

[22] A. Rakhlin and K. Sridharan, "Online learning with predictable sequences," in *Conference on Learning Theory*, pp. 993–1019, 2013.

[23] T. Lykouris and S. Vassilvitskii, "Competitive caching with machine learned advice," *Journal of the ACM (JACM)*, vol. 68, no. 4, pp. 1–25, 2021.

[24] N. B. Chang and M. Liu, "Optimal channel probing and transmission scheduling for opportunistic spectrum access," *IEEE/ACM Transactions on Networking*, vol. 17, no. 6, pp. 1805–1818, 2009.

[25] E. C. Elumar, C. Tekin, and O. Yağan, "Multi-armed bandits with costly probes," 2024. Submitted to IEEE Transactions on Information Theory. Available at https://users.ece.cmu.edu/~oyagan/Journals/ProbingBandits.pdf.

[26] J. Zuo, X. Zhang, and C. Joe-Wong, "Observe before play: Multi-armed bandit with pre-observations," *ACM SIGMETRICS Performance Evaluation Review*, vol. 46, no. 2, pp. 89–90, 2019.

[27] A. Bhaskara, S. Gollapudi, S. Im, K. Kollias, and K. Munagala, "Online learning and bandits with queried hints," in *14th Innovations in Theoretical Computer Science Conference, ITCS 2023, January 10-13, 2023, MIT, Cambridge, Massachusetts, USA*, 2023.

[28] S. Guha, K. Munagala, and S. Sarkar, "Optimizing transmission rate in wireless channels using adaptive probes," in *Proceedings of the Joint international Conference on Measurement and Modeling of Computer Systems*, pp. 381–382, 2006.

[29] L.-J. Chen, T. Sun, G. Yang, M. Y. Sanadidi, and M. Gerla, "Ad hoc probe: path capacity probing in wireless ad hoc networks," in *First International Conference on Wireless Internet (WICON'05)*, pp. 156–163, 2005.

[30] A. Johnsson, B. Melander, and M. Björkman, "Bandwidth measurement in wireless networks," in *Challenges in Ad Hoc Networking: Fourth Annual Mediterranean Ad Hoc Networking Workshop, June 21–24, 2005, Île de Porquerolles, France*, pp. 89–98, Springer, 2006.

[31] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *International Conference on Machine Learning*, pp. 151–159, 2013.

[32] N. Cesa-Bianchi and G. Lugosi, "Combinatorial bandits," *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1404–1422, 2012.

[33] Q. Wang and W. Chen, "Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[34] F. M. Harper and J. A. Konstan, "The MovieLens datasets: History and context," *ACM Transactions on Interactive Intelligent Systems (TIIS)*, vol. 5, no. 4, pp. 1–19, 2015.