Multi-Armed Bandits With Costly Probes

Eray Can Elumar, *Student Member, IEEE*, Cem Tekin, *Senior Member, IEEE*,
and Osman Yağan, *Senior Member, IEEE*

***Abstract*— Multi-armed bandits is a sequential decision-making problem where an agent must choose between multiple actions to maximize its cumulative reward over time, while facing uncertainty about the rewards associated with each action. The challenge lies in balancing the exploration of potentially higher-rewarding actions with the exploitation of known high-reward actions. We consider a multi-armed bandit problem with probes, where before pulling an arm, the decision-maker is allowed to probe one of the K arms for a cost $c \geq 0$ to observe its reward. We introduce a new regret definition that is based on the expected reward of the optimal action. We develop UCBP, a novel algorithm that utilizes this strategy to achieve a gap-independent regret upper bound that scales with the number of rounds T as $O(\sqrt{KT \log T})$, and an order optimal gap-dependent upper bound of $O(K \log T)$. As a baseline, we introduce UCB-naive-probe, a naive UCB-based approach which has a gap-independent regret upper bound of $O(K\sqrt{T \log T})$, and gap-dependent regret bound of $O(K^2 \log T)$; and TSP, the Thompson sampling version of UCBP. In empirical simulations, UCBP outperforms UCB-naive-probe, and performs similarly to TSP, verifying the utility of UCBP and TSP algorithms in practical settings.**

***Index Terms*— Multi-armed bandits, online learning, sequential decision-making, probing.**

I. Introduction

Multi-armed bandits (MAB) is a widely studied framework for sequential decision-making under uncertainty. In the standard MAB formulation, an agent chooses one of K actions (often referred to as *arms*) in each round and receives a random *reward* that follows an *unknown* distribution associated with the selected action. The objective of the agent is to *maximize* the mean reward received in total over T rounds. To this end, the agent must balance exploration of the different actions to learn more about their rewards, and exploitation of the actions that have provided the highest rewards so far. The seminal work of [2] showed that the *regret*, defined as the difference in expected total rewards between a given policy and the *optimal policy in hindsight*, has to grow at least logarithmically in the number of plays, and developed asymptotically optimal decision policies. Thereafter, many other asymptotically efficient policies have been proposed, including [3], [4], and used in applications in many fields, such as online advertising [5], [6], clinical trials [7], [8], and recommendation systems [9], [10].

Fueled by the explosion of data and the need for efficient and effective decision-making in various domains in recent years, there has been a surge of interest in multi-armed bandits. This interest has led to many new developments and insights, spanning algorithmic design, theoretical analysis, and practical applications. One area of recent development is bandits with side information, which allows the agent to receive side information before making a decision [11], [12], [13]. The side information can be in the form of partial observations, expert advice, context of the arms, or prior knowledge about the reward distributions. Recent work has shown that bandits with side information can improve the learning rate and robustness of MAB algorithms, and can be useful in various practical settings, such as clinical trials and online auctions.

The idea of probing to reduce uncertainty in a decision-making process has been studied in many areas of research, such as wireless communication systems [14], stochastic probing [15], online learning [16], and multi-armed bandits [17], [18]. In settings that utilize costly expert advice, where either humans or machine learning models are experts, probing can be interpreted as getting a *prediction* of the reward of an arm from the expert without pulling the arm. In this paper, we consider a specific variant of this problem, namely multi-armed bandits with *probes*. In this problem, the decision-maker is allowed to probe one arm for a cost $c \geq 0$ to observe its reward for that round. Based on the information obtained from the probe, the decision-maker can then pull that arm, or any other arm. The decision-maker can also pull an arm in a round without probing an arm. This variation of the MAB problem introduces an additional level of complexity and challenge, as probing considerably expands the action space, and the agent must balance exploration and exploitation while incorporating the decision about whether to probe an arm in its decision-making process. The main goal of our work is to develop new algorithms for this framework that achieve as much *cumulative* reward as possible. Towards this end, we propose the UCBP algorithm, and provide its theoretical analysis. We also consider the extension of this setting to multiple probes under binary rewards, and propose

Received 6 February 2024; revised 13 November 2024; accepted 17 November 2024. Date of publication 26 November 2024; date of current version 26 December 2024. This work was supported in part by the National Science Foundation under Grant CCF-2007834, in part by the Office of Naval Research under Grant N00014-23-1-2275, and in part by the CyLab Enterprise Security Initiative. The work of Cem Tekin was supported by Turkish National Academy of Sciences Distinguished Young Scientist Award Program under Grant TUBA-GEBIP-2023. An earlier version of this paper was presented in part at the 2024 IEEE International Symposium on Information Theory (ISIT) [DOI: 10.1109/ISIT57864.2024.10619238]. *(Corresponding author: Eray Can Elumar.)*

Eray Can Elumar and Osman Yağan are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: eelumar@andrew.cmu.edu; oyagan@ece.cmu.edu).

Cem Tekin is with the Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Türkiye (e-mail: cemtekin@ee.bilkent.edu.tr).

Communicated by V. Y. F. Tan, Associate Editor for Machine Learning and Statistics.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2024.3506866.

Digital Object Identifier 10.1109/TIT.2024.3506866

the UCBMP algorithm. Related work for these settings are provided in detail in §III.

A. Applications

The formulation considered here has numerous applications across different fields. A good example is online learning with machine learning (ML) advice. In this setting, ML models are used to predict the outcomes of actions before deciding on an action [19], [20], [21] to characterize improved performance bounds compared to the case without predictions in settings such as when the predictions are perfect [22], when the predictions are adversarial [23], or when there is an upper limit on the error of the predictions [24]. While in this work we assume that a probe reveals the exact outcome of an arm, we associate a cost to probing that may be used to model the computational complexity of using ML predictions. This work is also useful in the sense that it may serve as a reference point for future work that relaxes this assumption to include the cases where probes are *noisy* reward predictions.

In hyperparameter optimization for machine learning models, one approach is to have human experts routinely inspect the learning curves to quickly terminate runs with poor hyperparameter settings [25]. Our work can be incorporated into this setting by defining pulling an arm as running the hyperparameter setting without human expert supervision, and probing an arm as running it with supervision. Since poor runs will be quickly terminated, regret will not be incurred from probes, and only probing cost which reflects the cost of having a human expert will be incurred. In fraud detection, probing can represent running a particular check on a given transaction to estimate the likelihood of fraudulent activity, while pulling can represent blocking or confirming a transaction.

Another possible application of our work is in wireless communications. Probes in wireless communications mainly involve sending small data packets to observe some channel properties at that time. Prior work generally assume knowing the distributions of the rewards of channels [26]. Our work can be especially useful when these distributions are unknown. One other application is the cold-start problem in recommender systems [27], where, when a new item, or arm is added to the system, it is needed to learn its reward without suffering too much regret. The general approach is to generate reward predictions for this new arm from rewards of similar arms [28]. The probes in our work can be used to model predictions from such systems and the cost of probe can model the cost of making predictions. Also, our work can be used to model some test, or incentivized users that reveal or predict the reward of the arm without suffering the regret. Then, the cost of probe can reflect the cost of incentivizing such users. We also believe our work can be useful in other areas where bandits are used such as drug trials and ad recommendations.

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 $c \geq 0$.[1] This is an intricate problem different from most previous bandit formulations as the action set is larger, and the decision to pull an arm after probing depends on the probe outcome, which makes the analysis harder.
2) **UCBP Algorithm:** We identify the optimal strategy to whether to pull or probe an arm, and if we probe an arm, we also identify which arm to probe, and which arm to pull after the probe by evaluating the expected reward of each action. We provide an order-optimal algorithm based on UCB that evaluates the value of each action and uses upper confidence bounds to explore and choose the optimal action.
3) **Regret Upper Bound for UCBP:** We provide upper bounds on the expected cumulative regret of UCBP through a novel decomposition of regret for this problem setting. We establish that the gap-independent regret upper bound scales with $O(\sqrt{KT \log T})$, and that when the reward distribution is discrete, the gap-dependent regret upper 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 $\Omega(K \log T)$.
4) **Simulations:** To demonstrate the empirical performance of UCBP, we provide two baseline algorithms for comparison. We provide UCB-naive-probe, a naive UCB-based algorithm that does not employ the optimal strategy of the UCBP algorithm; and TSP, a Thompson sampling version of UCBP. We perform simulations of UCBP, TSP, UCB-naive-probe, and the standard UCB algorithm on the MOVIELENS and the Open Bandit datasets.
5) **Extension to Multiple Probes:** To demonstrate how our problem setting can be extended to multiple probes, we provide UCBMP, the multiple probe version of our algorithm under Bernoulli arm rewards.

II. Problem Statement

In this section we define our problem setting of the multi-armed bandit model with probes and derive the optimal action for this setting. The notations of some of the terms used throughout the paper are given in Table I.

A. Multi-Armed Bandit Model With Probes

We consider a K-armed stochastic bandit problem with the set of arms $[K]$. When pulled, arm $i \in [K]$ generates a random reward from a distribution Γ_i with mean μ_i and support a subset of $[0, 1]$. Arm rewards are independent of each other and across time. At each round, the agent first selects one of the following two types of actions. The first type of action, called *pull*, is where the agent pulls a particular arm $i \in [K]$ to receive its reward $r(t) = r_i(t)$. In the second type of action, called *probe*, the agent selects a *probe arm* i and a *backup arm* $j \neq i$. First, the *probe arm* is probed, and its reward $r_i(t)$ is observed. Based on this, the agent can choose to pull the *probe arm* to receive reward $r(t) = r_i(t) - c$ or the *backup arm* to receive reward $r(t) = r_j(t) - c$. Here, $c \geq 0$ represents the known cost of probing.

[1] Note that this work can easily be extended to the setting where cost of probing arm i is $c_i \geq 0$.

TABLE I
NOTATIONS

K	Number of arms
$[K]$	Set of arms
$\mathcal{A}$	Set of actions
$\mathcal{A}_p$	Set of probe actions
$\mathcal{A}_s$	Set of pull actions
$\mathcal{A}_{p,i}$	Set of probe actions that involve arm i
c	Cost of probing an arm
$\mathcal{D}$	Discrete support of arm rewards
$r_i(t)$	Reward of arm i at round t
$a = (i, j)$	Probe action with i as the probe and j as the backup arm
$a = (i, \emptyset)$	Pull action of pulling arm i
a_t	The action taken at round t
a^*	The optimal action
μ_i	Mean reward of arm i
ν^*	Mean reward of the optimal action
ν_a	Mean reward of action a
Δ_a	Gap of action a
$\Delta_{\text{min},i}$	$\Delta_{\text{min},i} = \min_{a \in \mathcal{A}_p \setminus \{a^*\} \ s.t. \ i \in a} (\Delta_a)$
ρ_i	$\rho_i = \min_{a \in \mathcal{A}_{p,i} \setminus \{a^*\}} \left(\frac{\epsilon \Delta_a}{4}\right)$
$C_a(t)$	The confidence interval of action a at round t
γ_i	$\gamma_i := \min_l \lvert d_l - \mu_i \rvert$ if $\mu_i \notin \mathcal{D}$, and $\gamma_i := \min_{1 \le l \le \lvert\mathcal{D}\rvert - 1} \lvert d_l - d_{l+1} \rvert$ if $\mu_i \in \mathcal{D}$
$o(t)$	Set of arms whose reward is observed in round t
$U_i(t)$	Upper confidence bound of arm i in round t
$U_a(t)$	Upper confidence bound of action a in round t
$N_i(t)$	Number of times arm i is sampled (pull or probe)
$N_a(t)$	Number of times action a is taken

We define $\mathcal{A} = \mathcal{A}_s \cup \mathcal{A}_p$ as the action set where elements of $\mathcal{A}$ are tuples. $\mathcal{A}_p$ is the set of actions that involve probing, and $\mathcal{A}_s$ is the set of actions that do not involve probing. 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. It can be seen that $|\mathcal{A}| = K^2$. Further, the set of actions that include arm (either as probe or backup arm) i are denoted as $\mathcal{A}_i := \{a \in \mathcal{A} : i \in a\}$, and similarly $\mathcal{A}_{p,i} := \{a \in \mathcal{A}_p : i \in a\}$ is the set of probe actions that include arm i.

We also denote the action taken in round t by $a_t \in \mathcal{A}$. When $a_t = (i, j)$ in round t, after observing reward $r_i(t)$, the agent needs to decide whether to pull arm i or j. Since the reward of arm j is unobserved, only its expectation μ_j can be used. Hence, optimal decision is pulling arm i if $r_i(t) > \mu_j$, and arm j otherwise. We call this the *optimal reference point decision*. Note that the true μ_j needs to be known to be able to employ the optimal point decision strategy. Using this reference point strategy, it can be seen that the expected reward of playing action (i, j) is:

$$v_{(i,j)} = \mathbb{E}[\max(r_i, \mu_j)] - c \ .$$

Note that r_i here represents a generic reward value sampled from the arm reward distribution Γ_i and does not represent the reward at any specified time t. Hence, the expression $\mathbb{E}[\max(r_i, \mu_j)]$ represents expectation over the reward distribution of arm i. The calculated $v_{(i,j)}$ values for some example arm distributions and action choices are given in Table II.

Without loss of generality, we assume that the mean rewards of the arms are ordered such that $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$. For simplicity, we assume there is a unique arm with the highest mean, which we refer to as the *best arm*. In standard K-armed stochastic bandit, the only option available to the learner is the *pull* option. Hence, the optimal action is to choose the best arm in all rounds, leading to the standard definition of expected regret given as

TABLE II
EXAMPLE OF EXPECTED ACTION REWARDS UNDER DIFFERENT ARM DISTRIBUTIONS. (LEFT) DISTRIBUTIONS OF ARM REWARDS IN A SETTING WITH 3 DIFFERENT ARMS. THE 3/5 FRACTION IN FRONT OF THE BINOMIAL DISTRIBUTION IS USED TO SCALE THE REWARDS INTO RANGE [0,1]. (RIGHT) EXPECTED REWARD $v_{(i,j)}$ VALUES FOR SEVERAL DIFFERENT ACTIONS WHEN $c = 0$

Arm	Distribution
1	$\frac{1}{5} \cdot \text{Binomial}(n = 5, p = 0.4)$
2	$\text{Bernoulli}(p = 0.5)$
3	$\text{Beta}(\alpha = 3, \beta = 2)$

Action	Expected reward
$(1, 2)$	0.551
$(1, 3)$	0.619
$(3, 1)$	0.639
$(2, 3)$	0.8

$$R_T^{\text{std}} = T \cdot \mu_1 - \mathbb{E}\left[\sum_{t=1}^{T} r(t)\right] .$$

Unlike standard K-armed bandit, in our setup, the *probe* option makes the optimal action non-trivial. Since achieving even negative regret is straightforward under *probe* option if $\exists (i, j)$ s.t. $\mathbb{E}[\max(r_i, \mu_j)] - c > \mu_1$, it can be seen that $T \cdot \mu_1$ is a very weak benchmark. When $\Gamma_i, \ \forall i \in [K]$ are known *a priori*, the maximum expected reward that can be achieved in a round (the optimal reward) is

$$\nu^* = \max(\mu_1, \max_{i \in [K] \setminus \{1\}} \{-c + \mathbb{E}[\max(r_i, \mu_1)]\}, -c + \mathbb{E}[\max(r_1, \mu_2)]) \ . \quad (1)$$

This leads to the optimal action, which for simplicity we assume to be unique, being expressed as

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

We focus on achieving non-trivial sublinear regret bounds with respect to the optimal benchmark $T\nu^*$. Hence, we define the empirical cumulative regret with respect to the optimal reward as

$$\hat{R}_T = T\nu^* - \sum_{t=1}^{T} r(t) \ , \quad (2)$$

and the expected cumulative regret as

$$R_T = \mathbb{E}[\hat{R}_T] \ .$$

To define the gaps of actions $a \in \mathcal{A}$, we let ν_a represent the expected reward of action a. For $a = (i, \emptyset)$ such that $i \in [K]$, we have $\nu_a = \mu_i$. For $a = (i, j)$ such that $i, j \in [K]$ and $i \neq j$, we have $\nu_{(i,j)} = \mathbb{E}[\max(r_i, \mu_j)] - c$. The gaps of actions without probing are defined as $\Delta_{(i,\emptyset)} := \nu^* - \nu_{(i,\emptyset)}$. Gaps of actions with probing are defined as $\Delta_{(i,j)} := \nu^* - \nu_{(i,j)}$, and the gaps of arms are defined as $\Delta_i := \mu_1 - \mu_i$.

An important remark is that with this regret definition, and in view of (1), 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 $r_i \leq \mu_1$. This optimal action can only be taken with the exact knowledge of the mean reward of arm 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

$$r_{\text{ref}}(t) := |\tilde{\mu}_1(t) - \mu_1| \\ \cdot \mathbb{P}(r_i \in [\min(\mu_1, \tilde{\mu}_1(t)), \max(\mu_1, \tilde{\mu}_1(t))]).$$

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 *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: Before presenting the UCBP Algorithm, we present a naive UCB-based algorithm that will serve as baseline. In this algorithm, as will be seen, the reference point is also a part of the decision process, so we define actions different than the UCBP algorithm and treat each action triple as a super arm where actions of the form $a = (i, j, d_l) \in \mathcal{A}_N,\ i \in [K],\ j \in [K] \setminus \{i\}$ denote 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.

Clearly, for the set of super arms to be finite, we need to have finitely many reference point values; i.e., the UCB-naive-probe algorithm can only be used when the reward distributions of the arms are discrete with finite support. To this end, we assume that the rewards of the arms are distributed over a discrete support $\mathcal{D}$ in $[0, 1]$, and assume that $d_l \in \mathcal{D}$, are the elements of this discrete support (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 regular UCB indices for all super arms, and the arm with the highest UCB index is pulled each round. When a *super arm* (i, j, d_l) is selected for probing, and $r_i(t)$ is observed through probe, arm i is pulled if $r_i(t) \geq d_l$, and j is pulled otherwise. The pseudo-code is provided in Algorithm 1.

It can be seen that there are K arms for pull action, and $|\mathcal{D}| \cdot (K^2 - K)$ arms for probe action, hence the gap-independent and gap-dependent regret of this algorithm will scale with K and $|\mathcal{D}|$ as $O(\sqrt{|\mathcal{D}|K^2 T \log T})$, and $O(|\mathcal{D}|K^2 \log T)$, respectively. This demonstrates the complexity of the problem as the action space scales with $\tilde{O}(|\mathcal{D}|K^2)$.

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 probe and backup arm selection of the optimal strategy during probing. Our algorithm that achieves this reduction in regret is presented in §IV.

Algorithm 1 UCB-Naive-Probe

```
Initialize: $N_a = 0,\ a \in \mathcal{A}$
 Sample each super arm once
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 $r(t) = r_{i_t}(t)$
    else
        Probe arm $i_t$, observe reward $r_{i_t}(t)$
        if $r_{i_t}(t) \geq d(t)$ then
            Pull arm $i_t$, get $r(t) = r_{i_t}(t) - c$
        else
            Pull arm $j_t$, get $r(t) = r_{j_t}(t) - c$
        end if
    end if
    Update UCB indices and mean estimates
end for
```

III. Related Works

A. Bandits With Probes

To highlight the novelty in our work, we present prior work on bandits with probes that are similar to our problem setting. To our knowledge, probes were first studied in the setting of bandits with expert advice in [18], where there are multiple experts and after pulling an arm, the agent can observe the reward of any subset of arms by paying cost c for each observed arm. In [17], there is a limit on the number of queries allowed. In [31], advice-efficient multiarmed bandits with experts are studied where only a limited number of experts can be used at each round.

Recently, the bandit with probes problem for Bernoulli reward distribution is considered in [29], where an unlimited number of probes are allowed per round, but each probe has a cost. They propose an algorithm that achieves $O(K^2 \log T)$ gap-dependent regret by utilizing a strategy that orders arms from highest UCB value to lowest, and probes arms in this order until observing an arm with a reward of $'1'$. In our work, while we allow only one probe, we consider a more general bounded reward distribution which requires a more intricate strategy, and we achieve $O(K \log T)$ regret instead of $O(K^2 \log T)$.

In [30], two different probing models are studied for probes without cost. In the first model, two arms are probed at each round, the probe reveals the arm with the higher reward, and that arm must be pulled. A UCB-based algorithm is proposed that treats the selection of two arms as a super arm. The regret is defined as $R_T = T \cdot \mu^* - E[\sum_{t=1}^{T} r(t)]$ where μ^* is the mean reward of the arm with the highest mean reward and $r(t)$ is the reward obtained by the algorithm at round t. Note that this reward is not defined based on the reward of the optimal super arm. $O(K^2 \log T)$ gap-independent regret is achieved under this definition, compared to the $O(\sqrt{KT})$ for the standard UCB algorithm. However, this result follows mainly due to the regret definition, since it is even possible to achieve negative regret with this definition as $\max(r_i, r_j)$, the reward of super arm (i, j), can be larger than μ^*. In the second model, three arms are probed each round to observe their rewards, and one of the probed arms is pulled. The provided algorithm achieves $O(K^2)$ regret with same regret definition. In this paper, we consider a similar scenario where it is allowed to probe at most one arm, but we allow any arm

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

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

to be pulled after probing. We also define our regret based on the *optimal action*. Comparison of our work with prior work is summarized in Table III.

B. Probes in Wireless Communications

While there are numerous prior work on probing in wireless communication systems [14], [32], [33], [34], one notable study related to our work is [26]. In this work, a wireless system is considered where each channel j is associated with a reward of transmission, X_j, whose distribution 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. Since the subsequent probing decisions depend on the outcome of probes, computing the optimal decision is nontrivial, and two different algorithms are proposed. The main difference of [26] from our work is that the reward distributions of the arms are unknown in our setting.

C. Combinatorial Bandits

Combinatorial bandits is an extension of the standard bandit framework where the action that can be taken in each round is composed of a combination of different base arms satisfying certain constraints, generally referred to as a *super arm* [35], [36]. Since the number of possible actions can be as high as the number of subsets of the arm set, estimating the optimal action in each round can be computationally challenging. To overcome this, assumptions like the existence of an oracle that can efficiently approximate the optimal action [37], the linearity of the rewards of super arms over the set of arms [38], or additional constraints that can reduce the size of the action set are commonly used [39]. Once the agent takes an action, a reward is received which is a function of the rewards of the base arms that compose the chosen super arm. There are two distinct categories of combinatorial bandits based on the feedback received. In semi-bandit feedback, both the received reward, and the rewards of the individual base arms that comprise the super arm are observed. In bandit feedback, only the received reward is observed.

Our work can be considered a special form of combinatorial semi-bandits based on our reward function and feedback model. In the semi-bandit literature, many different reward functions are studied, including linear [40], nonlinear [35], and some more distinct reward functions such as receiving the maximum reward of the selected arms and also observing which arm produces this max reward [41]. Our setting is also similar to this maximum reward feedback. In our setting, we can choose an action that consists of one arm as in $(i, \emptyset)$ or two arms as in (i, j). If a probing action (i, j) is selected, we first observe the reward of arm i, then pull arm i if $r_i(t) > U_j(t)$, and pull the *backup* arm j otherwise. Since we choose which arm to pull after the intermediate observation (after only observing arm i and not arm j), this introduces uncertainty in our setting as we might not be able to pull the arm with the highest reward in a round. Hence, our reward model can be considered a special case of the maximum reward function that includes this uncertainty.

D. Combinatorial Bandits and Probabilistic Triggering

Probabilistic triggering of arms is a special feedback model where when an action is played, a random subset of arms is triggered according to a triggering probability distribution [42]. The observed reward depends both on the set of arms in the chosen action, and the set of arms that are triggered. To aid in theoretical analysis, p^* is defined as the minimum positive probability that an arm is triggered by any action. It is shown in [42, Theorem 3] that the regret lower bound scales with the factor $\frac{1}{p^*}$ for the general combinatorial bandits with probabilistically triggered arms, which shows that the regret bounds scale with the factor $\frac{1}{p^*}$ when rewards of some arms in the chosen action are partially observed (observed only when that arm is triggered). Another variable used to analyze probabilistic triggering is p_i, which is the triggering probability of arm i.

In [43], a gap-dependent regret upper bound of $O(\sum_i \log T/(p_i \Delta_i))$ is derived for a combinatorial Thompson sampling based algorithm. To remove the dependency of regret on such factors, the *triggering probability modulated bounded smoothness* assumption is used in [42]. The main idea behind this assumption is that when an arm is unlikely to be triggered by an action, the importance of that arm also diminishes, and changing that arm's expected mean can only cause a small change in the expected reward of an action. Using this assumption, they prove regret bounds that do not depend on p^*; but do depend on B, the bounded smoothness constant, for combinatorial bandit problems that satisfy this assumption. This assumption is used in many other work, such as in [44] where a combinatorial Thompson sampling algorithm with regret bounds that do not depend on triggering probabilities is provided.

Our work is similar to this setting as the probability that the optimal reference point decision pulls the backup arm can be understood as the triggering probability. Similar to the triggering probability modulated bounded smoothness assumption, we show that the contribution of the backup arm to the reward of a probe action is upper bounded by the mean reward of the backup arm times its triggering probability. This lets us derive regret upper bounds that do not depend on the triggering probability of the backup arm by using some of the proof techniques in [41].

E. Cascading Bandits and Probabilistic Triggering

It is an extension of the combinatorial bandit framework where a list of items from an item pool is recommended to a user. The user observes the items in the order of the list and picks the first attractive item. This model presents additional challenges on analysis as the feedback is received only for the first attractive item and the items before it in the list which is referred to as the probabilistic triggering or the partial observability of the rewards. In [45], the amount of available feedback at each step is probabilistically estimated to overcome this challenge. In [46], a minimum probability of observing the rewards of all the items in the list, p^*, is assumed to help with the theoretical analysis. The given regret bounds scale with $\frac{1}{f^*}$, where f^* is a function that depends on p^*.

However, it was shown later in [42, Lemma 1] that this cascading bandit problem already satisfies the *triggering probability modulated bounded smoothness* assumption and that the $\frac{1}{f^*}$ factor in the regret upper bound is not needed. This is due to being able to express the expected rewards of actions using triggering probabilities. In [47], a Thompson Sampling based algorithm with a regret bound of $O\left(K \log T/\Delta + K/\Delta^2\right)$ is provided. This bound is achieved through a regret analysis that decomposes the regret in terms of the number of observations of the suboptimal items by using the properties of the reward in the cascading bandit setting.

F. Online Learning

In the classical online learning problem, an agent chooses an action, the loss function at that round is revealed, and the evaluation of the loss at the chosen action is incurred as regret. In [16], label efficient prediction with expert advice is studied, in which, the forecaster, after guessing the next element of the sequence, can only ask to observe its true value for a limited number of times. In [48], there are hints in an online linear optimization problem which are correlated with the cost function. An algorithm that achieves $O(\log T)$ regret with $O(\sqrt{T})$ hints is given.

G. Stochastic Probing

It is a problem where the distributions of a set of elements are known, but not the actual outcomes, and the aim is to maximize the expected utility by probing under certain constraints. This problem has applications such as database query optimization [49], radar systems [50], and Bayesian auctions [15]. In *Pandora's problem*, each probe has a cost, and the goal is to maximize the largest observed value minus the probing costs. While this problem was formulated and solved in [51], different settings of it are widely studied [52], [53], [54].

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 using the UCB strategy. The pseudocode of UCBP is provided in Algorithm 2. Recall that in UCB algorithm [3] for the stochastic K-armed bandit, at each round t, the arm with the highest UCB index $U_i(t)$ is pulled, i.e.,

$$i_t = \arg\max_i U_i(t), \quad U_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{\alpha \log t}{N_i(t)}}\ , \tag{3}$$

where $\hat{\mu}_i(t)$ is the empirical mean reward of arm i, $\alpha > 0$ is a constant (to be specified later), and i_t is the arm that is pulled in round t, and $N_i(t)$ is the number of times arm i is pulled until round t. The first term in $U_i(t)$, $\hat{\mu}_i(t)$ is to exploit the best performing arm, and the second term, also referred to as the exploration bonus, is used to explore other arms since it allows the algorithm pull the arms that have not been pulled too much. With this formulation, UCB algorithm balances exploration and exploitation to achieve optimal regret. We use similar ideas in our UCBP algorithm by appropriately defining the mean action rewards and the exploration bonuses. The UCBP algorithm works as follows. At each round t, first, the empirical mean rewards of arms are determined using

$$\hat{\mu}_i(t) = \sum_{\tau=1}^{t-1} \frac{r_i(\tau)\mathbb{1}\{i \in o(\tau)\}}{N_i(t)}\ ,$$

where $o(t)$ 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 index of each pull action $a = (i, \emptyset) \in \mathcal{A}_s$ is computed as

$$U_{(i,\emptyset)}(t) = \hat{\mu}_i(t) + C_i(t)\ ,$$

where $C_i(t) = \sqrt{3\log(t)/N_i(t)}$. The UCB index of each probe action $a = (i, j) \in \mathcal{A}_p$ is computed as

$$U_{(i,j)}(t) = \sum_{\tau=1}^{t-1} \frac{\max(r_i(\tau), \hat{\mu}_j(t) + C_j(t))\mathbb{1}\{i \in o(\tau)\}}{N_i(t)} - c + C_i(t)\ .$$

The claim that $U_{(i,j)}(t)$ is a valid UCB index for the probe action is proven in §B. After the UCB indices are computed for all actions, the action with the highest UCB index is selected. If the selected action is a pull action, i.e., $a_t = (i, \emptyset)$, then arm i is pulled, and reward $r_i(t)$ is collected. If the selected action is a probe action, i.e., $a_t = (i, j)$, first arm i is probed to observe $r_i(t)$. Then arm i is pulled if $r_i(t) > U_j(t)$, and reward $r_i(t) - c$ is collected. Otherwise, arm j is pulled, and reward $r_j(t) - c$ is collected. In other words, UCBP uses $U_j(t)$ as the reference point $\tilde{\mu}_j(t)$ at round t. When calculating $U_j(t)$, we use $\alpha = 27$ in (3) in order to guarantee that the backup arm is chosen sufficiently often to achieve low regret, hence $U_j(t) = \hat{\mu}_j(t) + \sqrt{27\log(t)/N_j(t)} = \hat{\mu}_j(t) + 3C_j(t)$.

Next, we explain why the reference point $U_j(t)$ used by UCBP is the right choice. Let $p_{(i,j)} := \mathbb{P}(r_i \leq \mu_j)$ denote the probability that the backup arm j in action (i, j) is pulled if the optimal reference point decision strategy is employed.

Algorithm 2 UCBP

```
Input: cost of probing $c$, action set $\mathcal{A}$
Initialize: $N_i = 0,\ 1 \le i \le K$
 Sample each arm $i \in [K]$ once
for each round $t$ do
    $a_t = \arg\max_{a\in\mathcal{A}} U_a(t)$
    if $a_t = (i_t, \emptyset)$ is a pull action then
        Pull arm $i_t$, get $r(t) = r_{i_t}(t)$
    else ($a_t = (i_t, j_t)$ is a probe action)
        Probe arm $i_t$, observe reward $r_{i_t}(t)$
        if $r_{i_t}(t) > U_{j_t}(t)$ then
            Pull arm $i_t$, get $r(t) = r_{i_t}(t) - c$
        else
            Pull arm $j_t$, get $r(t) = r_{j_t}(t) - c$
        end if
    end if
     Update UCB indices for all actions
end for
```

Note that since UCBP uses the condition $r_i(t) > U_j(t)$ to pull the backup arm, the probability of UCBP pulling the backup arm is $\mathbb{P}(r_i \le U_j(t))$, which is greater than $p_{(i,j)}$, i.e. $\mathbb{P}(r_i \le U_j(t)) \ge p_{(i,j)}$ when the confidence bounds hold (see §B). Employing this reference point decision strategy, the $p_{(i,j)}$ value in our setting is similar to p^* in combinatorial bandits with probabilistically triggered arms, where the p^* defined as the minimum positive probability that an arm is triggered by any action [42].

This lets us use the triggering probability modulated bounded smoothness assumption in [41]. The main idea behind this assumption is that when an arm is unlikely to be triggered by an action, the importance of that arm also diminishes, and changing that arm's expected mean can only cause a small change in the expected reward of an action. Using this assumption, regret bounds that do not depend on p^*, but do depend on B, the bounded smoothness constant, are proved for combinatorial bandit problems that satisfy this assumption. We also use techniques in their proof such as the reverse amortization trick to derive regret upper bounds that do not depend on $p_{(i,j)}$.

A. Analysis of UCBP

We now characterize the performance of the UCBP algorithm by providing theoretical upper and lower bounds on the expected cumulative regret. We refer the readers to the Appendix for detailed proofs of the results presented in this section.

B. Regret Decomposition

In order to prove gap-independent and gap-dependent regret results, we employ a divide and conquer approach. This section presents several results on decomposition of the regret that will be utilized in the gap-independent and gap-dependent analysis.

We start by defining an event under which UCB indices of UCBP concentrate sufficiently to guarantee low regret. Consider arm $i \in [K]$. Assume that arm i is observed u times up to round t. Denote u i.i.d. samples from arm i as $\tilde{r}_i(1), \ldots, \tilde{r}_i(u)$. Define

$$\hat{\mu}_i(t,u) := \sum_{\tau=1}^{u} \frac{\tilde{r}_i(\tau)}{u} .$$

Consider pull action $a = (i, \emptyset)$. Assume that arm i is observed u times up to round t. Define

$$\hat{\nu}_{(i,\emptyset)}(t,u) := \hat{\mu}_i(t,u)$$

as the empirical reward,

$$U_{(i,\emptyset)}(t,u) := \hat{\nu}_{(i,\emptyset)}(t,u) + \sqrt{\frac{3\log t}{u}}$$

as the upper confidence bound, and

$$L_{(i,\emptyset)}(t,u) := \hat{\nu}_{(i,\emptyset)}(t,u) - \sqrt{\frac{3\log t}{u}}$$

as the lower confidence bound of pull action $(i, \emptyset)$ at round t when arm i is observed u times.

Consider probe action $a = (i, j)$. Assume that arm i is observed r times and arm j is observed s times up to round t. Denote r i.i.d. samples from arm i and s i.i.d samples from arm j, by $\tilde{r}_i(1), \ldots, \tilde{r}_i(r)$ and $\tilde{r}_j(1), \ldots, \tilde{r}_j(s)$. Define

$$\hat{\nu}_{(i,j)}(t,r,s) := \sum_{\tau=1}^{r} \frac{\max(\tilde{r}_i(\tau), \hat{\mu}_j(t,s))}{r} - c$$

as the empirical probing reward,

$$U_{(i,j)}(t,r,s) := \sum_{\tau=1}^{r} \frac{\max\left(\tilde{r}_i(\tau), \hat{\mu}_j(t,s) + \sqrt{\frac{3\log t}{s}}\right)}{r} - c + \sqrt{\frac{3\log t}{r}}$$

as the upper confidence bound, and

$$L_{(i,j)}(t,r,s) := \sum_{\tau=1}^{r} \frac{\max\left(\tilde{r}_i(\tau), \hat{\mu}_j(t,s) - \sqrt{\frac{3\log t}{s}}\right)}{r} - c - \sqrt{\frac{3\log t}{r}}$$

as the lower confidence bound of probe action $a = (i, j)$ at round t when arm i is observed r times, and arm j is observed s times. For the sake of notation brevity, when $a = (i, \emptyset)$, $U_a(t, r, s) = U_a(t, r)$ and $L_a(t, r, s) = L_a(t, r)$ will be assumed.

Define the following events:

$$\mathcal{E}_{t,a} := \left\{ \min_{r \le t, s \le t} U_{a^*}(t,r,s) \ge \nu^* \wedge \max_{u \le t, v \le t} L_a(t,u,v) \le \nu_a \right\},$$

$$\mathcal{E}_t := \bigcap_{a\in\mathcal{A}} \mathcal{E}_{t,a},$$

$$\mathcal{E}(T) := \bigcap_{t=K+1}^{T} \mathcal{E}_t \ ,$$

where $\mathcal{E}_t$ is the good event at round t, the event where a suboptimal selection is not made due to confidence intervals not holding, and $\mathcal{E}(T)$ is the event that such a suboptimal selection is not made for all rounds $K+1 \le t \le T$.

Since we incur regret whenever a suboptimal action is taken, or when the decision to pull the probe arm or the backup arm after observing the outcome of the probe is incorrect, we upper bound the expected number of times each suboptimal action or decision is chosen by the UCBP Algorithm.

Let $a = (i,j)$ represent a probe action. First, we define $\hat{r}_a(t) = r_i(t)\mathbb{1}\{r_i(t) > U_j(t)\} + r_j(t)\mathbb{1}\{r_i(t) \le U_j(t)\}$ as the reward received from action a in round t when $U_j(t)$ is used as the reference point, and $r_a(t) = r_i(t)\mathbb{1}\{r_i(t) > \mu_j\} + r_j(t)\mathbb{1}\{r_i(t) \le \mu_j\}$ as the reward received from action a in round t when μ_j is used as the reference point, i.e., the optimal reference point decision is employed.

When a probe action $a = (i,j) \in \mathcal{A}_p$ is taken in round t, we call the regret incurred due to using $U_j(t)$ instead of μ_j as the *reference point error*, which is given as $d_a(t) = r_a(t) - \hat{r}_a(t)$. This regret $d_a(t)$ is additive to the regret of choosing a suboptimal action a, since $d_a(t)$ captures the additional regret of the incorrect decision compared to the correct decision when deciding to pull the probe arm or the backup arm. The definition of $d_a(t)$ can be expanded to include pull actions $a \in \mathcal{A}_s$ in the following way:

$$d_a(t) = \begin{cases} 0 & \text{if } a = (i,\emptyset) \\ r_a(t) - \hat{r}_a(t) & \text{if } a = (i,j) \end{cases}$$

Let $\mathcal{B}_a(t)$ denote the event that the reference point error of action a is zero at round t, i.e., $\mathcal{B}_a(t) = \{\hat{r}_a(t) = r_a(t)\}$. The following result decomposes R_T into multiple parts, which will be separately bounded in our gap-independent and gap-free analysis.

Lemma 1 (Regret Decomposition): When UCBP is run on the action set $\mathcal{A}$ and the cost of probing is $c \ge 0$, its cumulative expected regret can be decomposed as

$$R_T \le R_s(T) + R_{\text{ref}}(T) + \sum_{t=K+1}^{T} \mathbb{P}(\mathcal{E}_t^c) + K \ ,$$

where

$$R_{\text{ref}}(T) := \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}} \mathbb{1}\{a_t = a, \mathcal{B}_a^c(t), \mathcal{E}_t\}\cdot d_a(t)\right] \quad (4)$$

represents the cumulative regret incurred from the reference point error until round T, and

$$R_s(T) := \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}} \mathbb{1}\{a_t = a, \mathcal{E}_t\}\cdot(\nu^* - r_a(t))\right] \quad (5)$$

denotes the cumulative regret incurred until round T from suboptimal action choices (without the reference point error).

Proof: Regret given in (2) can be written as

$$\hat{R}_T \le \sum_{t=K+1}^{T} (\nu^* - \hat{r}_{a(t)}(t)) + K \ ,$$

where the summation starts from K due to the initialization phase of UCBP and since rewards are bounded in $[0,1]$.

$\hat{R}_T$ can be further decomposed based on the good event $\mathcal{E}_t$ as

$$\begin{aligned} \hat{R}_T \le &\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}} \mathbb{1}\{a_t = a, \mathcal{E}_t\}\cdot(\nu^* - \hat{r}_a(t)) \\ &+ \sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}} \mathbb{1}\{a_t = a, \mathcal{E}_t^c\}\cdot(\nu^* - \hat{r}_a(t)) + K \\ \le &\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}} \mathbb{1}\{a_t = a, \mathcal{E}_t\}\cdot(\nu^* - \hat{r}_a(t)) \\ &+ \sum_{t=K+1}^{T} \mathbb{1}\{\mathcal{E}_t^c\}\cdot 1 + K \ . \end{aligned}$$

Since regret incurred from the reference point error when an action involving probing is chosen is additive to the regret from the suboptimality of the chosen action, conditioning on $\mathcal{B}_a(t)$, we write

$$\begin{aligned} \hat{R}_T = &\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}} [\mathbb{1}\{a_t = a, \mathcal{B}_a(t), \mathcal{E}_t\}\cdot(\nu^* - r_a(t)) \\ &+\mathbb{1}\{a_t = a, \mathcal{B}_a^c(t), \mathcal{E}_t\}\cdot(\nu^* - r_a(t) + d_a(t))] \\ &+ \sum_{t=K+1}^{T} \mathbb{1}\{\mathcal{E}_t^c\} + K \ . \end{aligned} \quad (6)$$

We continue bounding the expected regret by taking the expectation of (6).

$$\begin{aligned} R_T \le &\ \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}} [\mathbb{1}\{a_t = a, \mathcal{B}_a(t), \mathcal{E}_t\}\cdot(\nu^* - r_a(t))\right. \\ &\left. +\mathbb{1}\{a_t = a, \mathcal{B}_a^c(t), \mathcal{E}_t\}\cdot(\nu^* - r_a(t) + d_a(t))]\right] \\ &+ \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{\mathcal{E}_t^c\}\right] + K \\ = &\ \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}} [\mathbb{1}\{a_t = a, \mathcal{E}_t\}\cdot(\nu^* - r_a(t))\right. \\ &\left. +\mathbb{1}\{a_t = a, \mathcal{B}_a^c(t), \mathcal{E}_t\}\cdot d_a(t)]\right] \\ &+ \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{\mathcal{E}_t^c\}\right] + K \\ = &\ \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}} [\mathbb{1}\{a_t = a, \mathcal{E}_t\}\cdot(\nu^* - r_a(t))\right. \\ &\left. +\mathbb{1}\{a_t = a, \mathcal{B}_a^c(t), \mathcal{E}_t\}\cdot d_a(t)]\right] \\ &+ \sum_{t=K+1}^{T} \mathbb{P}(\mathcal{E}_t^c) + K \ . \end{aligned} \quad (7)$$

Observing (7), we obtain the final result:

$$R_T \le R_s(T) + R_{\text{ref}}(T) + \sum_{t=K+1}^{T} \mathbb{P}(\mathcal{E}_t^c) + K \ .$$

□

Each term in Lemma 1 can be further decomposed to facilitate regret analysis. Next, we state how $R_s(T)$ in (5) can be decomposed. Recall that $o(t) \subset a_t$ is the set of arms whose reward is observed in round t. Define $\mathcal{H}_t = (a_1, r(1), o(1), \cdots, a_{t-1}, r(t-1), o(t-1))$ as the history of UCBP up to choosing action a_t, and let $\mathbb{E}[\cdot|\mathcal{H}_t]$ be the conditional expectation given this history.

Lemma 2 ($R_s(T)$ Decomposition): When UCBP is run on the action set $\mathcal{A}$ and the cost of probing is $c \geq 0$, $R_s(T)$ can be decomposed as $R_s(T) = 2R_{s,1}(T) + 2R_{s,2}(T)$, where

$$R_{s,1}(T) := \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{(i,\cdot)\in\mathcal{A}} \mathbb{1}\{a_t = (i,\cdot), \mathcal{E}_t\} \cdot C_i(t)\right] ,$$

$$R_{s,2}(T) := \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{(i,j)\in\mathcal{A}_p} \mathbb{1}\{a_t = (i,j), \mathcal{E}_t\} \right.$$
$$\left. \cdot \mathbb{P}\left(r_i(t) \leq \mu_j + 2C_j(t)|\mathcal{H}_t\right) \cdot C_j(t)\right] .$$

Proof: Let $\tilde{R}_t := \sum_{a\in\mathcal{A}} \mathbb{1}\{a_t = a, \mathcal{E}_t\}(\nu^* - r_a(t))$, and note that $R_s(T) = \mathbb{E}\left[\sum_{t=K+1}^{T} \tilde{R}_t\right]$. $\mathbb{E}[\tilde{R}_t]$ can be expressed as:

$$\begin{aligned}
\mathbb{E}[\tilde{R}_t] &= \mathbb{E}\left[\mathbb{E}[\tilde{R}_t|\mathcal{H}_t]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{a\in\mathcal{A}} \mathbb{1}\{a_t = a, \mathcal{E}_t\}(\nu^* - r_a(t))|\mathcal{H}_t\right]\right] \\
&= \mathbb{E}\left[\sum_{a\in\mathcal{A}} \mathbb{1}\{a_t = a, \mathcal{E}_t\}\mathbb{E}\left[(\nu^* - r_a(t))|\mathcal{H}_t\right]\right] \\
&= \mathbb{E}\left[\sum_{a\in\mathcal{A}} \mathbb{1}\{a_t = a, \mathcal{E}_t\}(\nu^* - \nu_a)\right] \\
&= \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_t\}(\nu^* - U_{a_t}(t) + U_{a_t}(t) - \nu_{a_t})\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_t\}(\nu^* - U_{a^*}(t) + U_{a_t}(t) - \nu_{a_t})\right] && (8) \\
&\leq \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_t\}(U_{a_t}(t) - \nu_{a_t})\right] && (9) \\
&= \mathbb{E}\left[\sum_{a\in\mathcal{A}} \mathbb{1}\{a_t = a, \mathcal{E}_t\}(U_a(t) - \nu_a)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{a\in\mathcal{A}} \mathbb{1}\{a_t = a, \mathcal{E}_t\}(U_a(t) - \nu_a)|\mathcal{H}_t\right]\right] && (10)
\end{aligned}$$

where (8) follows from $U_{a_t}(t) \geq U_{a^*}(t)$ since the UCBP algorithm selects the action with the highest UCB index, and (9) follows from $\nu^* \leq U_{a^*}(t)$ under the event $\mathcal{E}_t$.

We will upper bound $U_a(t) - \nu_a$ separately for pull actions and probe actions under event $\mathcal{E}_t$ given $\mathcal{H}_t$. First, if $a = (i,j)$ is a probe action, using Corollary 18, it can be seen that

$$U_a(t) - \nu_a \leq 2C_i(t) + 2\mathbb{P}\left(r_i(t) \leq \mu_j + 2C_j(t)|\mathcal{H}_t\right) \cdot C_j(t) . \tag{11}$$

Second, if $a = (i,\emptyset)$ is a probe action, then

$$\begin{aligned}
U_a(t) - \nu_a &= U_i(t) - \mu_i \\
&= \hat{\mu}_i(t) + C_i(t) - \mu_i \\
&\leq \mu_i + C_i(t) + C_i(t) - \mu_i \\
&\leq 2C_i(t) . && (12)
\end{aligned}$$

Plugging (11) and (12) into (10), and summing over rounds,

$$\begin{aligned}
R_s(T) = \mathbb{E}&\left[\sum_{t=K+1}^{T} \sum_{(i,j)\in\mathcal{A}_p} \mathbb{1}\{a_t = (i,j), \mathcal{E}_t\}\right. \\
&\left. \cdot 2\left(C_i(t) + \mathbb{P}\left(r_i(t) \leq \mu_j + 2C_j(t)|\mathcal{H}_t\right) \cdot C_j(t)\right)\right] \\
&+ \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{(i,\emptyset)\in\mathcal{A}_s} \mathbb{1}\{a_t = (i,\emptyset), \mathcal{E}_t\} \cdot 2C_i(t)\right] \\
= \;& 2R_{s,1}(T) + 2R_{s,2}(T) .
\end{aligned}$$

□

C. Gap-Independent Expected Regret Upper Bound

Theorem 3 (Gap-Independent Expected Regret Upper Bound): 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

$$R_T \leq 8(\sqrt{6} + \sqrt{3})\sqrt{KT\log T} + \frac{2\pi^2 K^2}{3} + K .$$

This theorem shows that the gap-independent regret of UCBP is $O(\sqrt{KT\log T})$, which has the same order as the gap-independent regret of the stochastic K-armed bandit problem.

Proof: We will upper bound the regret by upper bounding each term in Lemma 1. First, we consider the decomposition of $R_s(T)$ given in Lemma 2. We start by upper bounding $R_{s,1}(T)$. Recall that $\mathcal{A}_i$ denotes the set of all actions that involve arm i as the probe or pull arm.

$$\begin{aligned}
R_{s,1}(T) &= \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{i=1}^{K} \mathbb{1}\{a_t \in \mathcal{A}_i, \mathcal{E}_t\} \cdot C_i(t)\right] \\
&= \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{i=1}^{K} \mathbb{1}\{a_t \in \mathcal{A}_i, \mathcal{E}_t\} \cdot \sqrt{\frac{3\log t}{N_i(t)}}\right] \\
&\leq \sqrt{3\log T} \cdot \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{i=1}^{K} \mathbb{1}\{a_t \in \mathcal{A}_i\} \cdot \sqrt{\frac{1}{N_i(t)}}\right] \\
&\leq \sqrt{3\log T} \cdot \mathbb{E}\left[\sum_{i=1}^{K} \sum_{t=K+1}^{T} \mathbb{1}\{i \in o(t)\} \cdot \sqrt{\frac{1}{N_i(t)}}\right] && (13) \\
&\leq \sqrt{3\log T} \cdot \mathbb{E}\left[\sum_{i=1}^{K} \sum_{x=1}^{N_i(T)} \sqrt{\frac{1}{x}}\right] && (14) \\
&\leq 2\sqrt{3\log T} \cdot \mathbb{E}\left[\sqrt{K\sum_{i=1}^{K} N_i(T)}\right] , && (15)
\end{aligned}$$

where (13) follows from the fact that whenever $a_t \in \mathcal{A}_i$, we have $i \in o(t)$, (14) follows from the fact that whenever $i \in o(t)$ then $N_i(t)$ is incremented by 1, and (15) follows from Cauchy-Schwarz inequality. Since up to two arms may be observed in a round, $\sum_{i=1}^{K} N_i(T) \leq 2T$, and using this we obtain

$$R_{s,1}(T) \leq 2\sqrt{6KT\log T} . \tag{16}$$

Next, we upper bound $R_{s,2}(T)$. Note that due to the reference point decision strategy employed by UCBP, the following holds under the good event $\mathcal{E}_t$

$$\begin{aligned}
&\mathbb{P}\left(j \in o(t) | a_t = (i,j), \mathcal{H}_t\right)\\
&\qquad = \mathbb{P}\left(r_i(t) \le U_j(t) | a_t = (i,j), \mathcal{H}_t\right)\\
&\qquad = \mathbb{P}\left(r_i(t) \le \hat{\mu}_j(t) + \sqrt{\frac{27 \log t}{N_i(t)}} \middle| \mathcal{H}_t\right)\\
&\qquad = \mathbb{P}\left(r_i(t) \le \hat{\mu}_j(t) + 3C_j(t) | \mathcal{H}_t\right)\\
&\qquad \ge \mathbb{P}\left(r_i(t) \le \mu_j + 2C_j(t) | \mathcal{H}_t\right)
\end{aligned}$$

Using this, $R_{s,2}(T)$ can be upper bounded as

$$\begin{aligned}
&R_{s,2}(T)\\
&= \mathbb{E}\Bigg[\sum_{t=K+1}^{T} \sum_{(i,j)\in\mathcal{A}_p} \mathbb{1}\left\{a_t = (i,j), \mathcal{E}_t\right\}\\
&\qquad \cdot \mathbb{P}\left(r_i(t) \le \mu_j + 2C_j(t) | \mathcal{H}_t\right) \cdot C_j(t)\Bigg]\\
&\le \mathbb{E}\Bigg[\sum_{t=K+1}^{T} \sum_{(i,j)\in\mathcal{A}_p} \mathbb{1}\left\{\mathcal{E}_t\right\}\\
&\qquad \cdot \mathbb{P}\left(r_i(t) \le \mu_j + 2C_j(t) | \mathcal{H}_t\right) \cdot C_j(t)\\
&\qquad \cdot \mathbb{E}\left[\mathbb{1}\left\{a_t = (i,j)\right\} \cdot \frac{\mathbb{1}\left\{j \in o(t)\right\}}{\mathbb{P}\left(r_i(t) \le U_j(t) | \mathcal{H}_t\right)} \middle| \mathcal{H}_t\right]\Bigg]
\end{aligned} \tag{17}$$

$$\begin{aligned}
&= \mathbb{E}\Bigg[\sum_{t=K+1}^{T} \mathbb{E}\Bigg[\sum_{(i,j)\in\mathcal{A}_p} \mathbb{1}\left\{a_t = (i,j), j \in o(t), \mathcal{E}_t\right\}\\
&\qquad \cdot C_j(t) \cdot \frac{\mathbb{P}\left(r_i(t) \le \mu_j + 2C_j(t) | \mathcal{H}_t\right)}{\mathbb{P}\left(r_i(t) \le U_j(t) | \mathcal{H}_t\right)} \Bigg| \mathcal{H}_t\Bigg]\Bigg]\\
&\le \mathbb{E}\Bigg[\sum_{t=K+1}^{T} \mathbb{E}\Bigg[\sum_{(i,j)\in\mathcal{A}_p} \mathbb{1}\left\{a_t = (i,j), j \in o(t), \mathcal{E}_t\right\}\\
&\qquad \cdot C_j(t) | \mathcal{H}_t\Bigg]\Bigg]\\
&= \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{(i,j)\in\mathcal{A}_p} \mathbb{1}\left\{a_t = (i,j), j \in o(t), \mathcal{E}_t\right\} \cdot C_j(t)\right]\\
&= \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{j=1}^{K} \sum_{a\in\mathcal{A}_{(\cdot,j)}} \mathbb{1}\left\{a_t = a, j \in o(t), \mathcal{E}_t\right\} \cdot C_j(t)\right]\\
&= \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{j=1}^{K} \mathbb{1}\left\{a_t \in \mathcal{A}_{(\cdot,j)}, j \in o(t), \mathcal{E}_t\right\} \cdot C_j(t)\right]\\
&\le \mathbb{E}\left[\sum_{j=1}^{K} \sum_{t=K+1}^{T} \mathbb{1}\left\{j \in o(t)\right\} \cdot C_j(t)\right]\\
&= \sqrt{3 \log T} \cdot \mathbb{E}\left[\sum_{i=1}^{K} \sum_{t=K+1}^{T} \mathbb{1}\left\{i \in o(t)\right\} \cdot \sqrt{\frac{1}{N_i(t)}}\right]\\
&\le \sqrt{3 \log T} \cdot \mathbb{E}\left[\sum_{i=1}^{K} \sum_{x=1}^{N_i(T)} \sqrt{\frac{1}{x}}\right]\\
&\le 2\sqrt{3 \log T} \cdot \mathbb{E}\left[\sqrt{K \sum_{i=1}^{K} N_i(T)}\right],
\end{aligned}$$

where (17) follows from the tower rule and the fact that under the event $a_t = (i,j)$, the event $j \in o(t)$ can happen if and only if $r_i(t) \le U_j(t)$. The last two inequalities follow the same reasoning as in (12) and (14). Using $\sum_{i=1}^{K} N_i(T) \le 2T$, we get

$$R_{s,2}(T) \le 2\sqrt{6KT \log T}\ . \tag{18}$$

Combining (16) and (18) with $R_s(T) = 2R_{s,1}(T) + 2R_{s,2}(T)$ from Lemma 2, we get

$$R_s(T) \le 8\sqrt{6}\sqrt{KT \log T}\ .$$

It can be seen from Lemma 19 that,

$$\sum_{t=K+1}^{T} \mathbb{P}(\mathcal{E}_t^c) \le \frac{2\pi^2 K^2}{3}\ .$$

Also, using the fact that $R_{\text{ref}}(T) \le 8\sqrt{3KT \log T}$ from Lemma 21, we arrive at the final bound on R_T:

$$\begin{aligned}
R_T &= R_s(T) + R_{\text{ref}}(T) + \sum_{t=K+1}^{T} \mathbb{P}(\mathcal{E}_t^c) + K\\
&\le 8(\sqrt{6} + \sqrt{3})\sqrt{KT \log T} + \frac{2\pi^2 K^2}{3} + K\ .
\end{aligned}$$

□

D. Gap-Dependent Expected Regret Upper Bound

Theorem 4 (Gap-dependent expected regret upper bound): When UCBP is run on the action set $\mathcal{A}$ and the cost of probing is $c \ge 0$, its expected cumulative regret is upper bounded as

$$R_T \le \sum_{i=1}^{K} \frac{192 \log T}{\Delta_{\min,i}} + R_{\text{ref}}(T) + \frac{2\pi^2 K^2}{3} + K\ ,$$

where $\Delta_{\min,i} = \min_{a\in\mathcal{A}_i:\Delta_a>0} \Delta_a$, and $R_{\text{ref}}(T)$ is the reference point regret.

Further assuming that the distributions Γ_i for each $i \in [K]$ are defined over a *discrete* support $\mathcal{D}$ in $[0,1]$, using the upper bound of $R_{\text{ref}}(T)$ in Lemma 5, regret can be upper bounded as

$$R_T \le \sum_{i=1}^{K} \frac{192 \log T}{\Delta_{\min,i}} + \sum_{i=1}^{K} \frac{96 \log T}{\gamma_i} + \frac{2\pi^2 K^2}{3} + K\ ,$$

where we use $d_l \in \mathcal{D}$, $1 \le l \le |\mathcal{D}|$ 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 $\mu_i \in \mathcal{D}$.

Note that the cost of probing c is included in the gap of actions. This shows that the gap-dependent regret of UCBP is $O(K \log T)$ when the reward distribution is discrete. This result is order optimal, as it matches the lower bound given in Lemma 5, and also the regret lower bound of the standard UCB algorithm. If the reward distribution is not discrete, then the regret upper bound is $O(\sqrt{KT \log T})$.

Proof Sketch. The proof follows some of the steps in the proof of Theorem 4 in [41]. We first decompose regret due to suboptimal action selections and regret due to reference point error. After this step, one key idea is to show that

$$U_a(t) - \nu_a \leq 2C_i(t) + 2\mathbb{P}\left(r_i(t) \leq \mu_j + 2C_j(t)|\mathcal{H}_t\right) \cdot C_j(t),$$

where $\mathbb{P}\left(r_i(t) \leq \mu_j + 2C_j(t)|\mathcal{H}_t\right)$ is the triggering probability of the backup arm. This property is combined with the fact that UCBP pulls the backup arm if $r_i(t) \leq U_j(t)$ for an action (i, j), to conjecture that the backup arm is pulled with a probability greater than or equal to the triggering probability. This lets regret be upper bounded by the expected regret of the round in which the backup arm is actually triggered. Reverse amortization trick in [41, Theorem 4] is used to upper bound this expected regret, which can than be summed over the rounds of the algorithm to obtain the regret upper bound due to suboptimal action selections. The regret due to reference point decision error is upper bounded by using the fact that reference point decision regret cannot be incurred when $C_j(t) \leq \gamma_j/4$ for a backup arm j. The detailed proof is in Appendix D.

We now provide upper bounds on *reference point regret*, which is incurred since the algorithm does not have information on the true means, and only uses the estimated 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 5: a) Regret due to the reference point error is upper bounded as:

$$R_{\text{ref}}(T) \leq 8\sqrt{3KT\log T}\ .$$

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

$$R_{\text{ref}}(T) \leq \sum_{i=1}^{K} \frac{96\log T}{\gamma_i}\ ,$$

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. Proof of Lemma 5 is given in Appendix E.

Theorem 6 (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 [2], 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:

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

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

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 the optimal action, and $D_{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence.

Proof of this result is given in Appendix F. It can be seen that the lower bound on regret of UCBP is $\Omega(K\log T)$ since $C(\Gamma)$ is $\Omega(K)$. Since the upper bound on expected regret is also $O(K\log T)$ under discrete rewards in Theorem 4, we can conclude that the gap-dependent upper bound of the UCBP algorithm is order-wise optimal.

Finally, we state in the corollaries below regret upper bounds for UCB-naive-probe, which was introduced in §II.

Corollary 7: If the rewards of the arms are distributed over the discrete support $\mathcal{D}$, when UCB-naive-probe is run on $\mathcal{A}$ and the cost of probing is $c \geq 0$, the gap-independent upper bound for the expected regret, denoted as $R_U(T)$, is:

$$\begin{aligned} R_U(T) &\leq 4\sqrt{2|\mathcal{D}|K^2 T\log T} \\ &\quad + \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} + |\mathcal{D}|K^2 \\ &= O(\sqrt{|\mathcal{D}|K^2 T\log T}) + O(1)\ . \end{aligned}$$

Corollary 8: If the rewards of the arms are distributed over the discrete support $\mathcal{D}$, the gap-dependent upper bound for $R_U(T)$ is:

$$\begin{aligned} R_U(T) &\leq \sum_{a \in \mathcal{A}_N\setminus\{u^*\}} \frac{8\log T}{\Delta_a} + |\mathcal{D}|K^2 \\ &\quad + \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} \\ &= O(|\mathcal{D}|K^2\log T) + O(1)\ . \end{aligned}$$

where $\Delta_{(i,\emptyset,\emptyset)} = \Delta_i$,

$$\Delta_{(i,j,d_l)} = c + \nu^* - \mathbb{E}\left[r_i \cdot \mathbb{1}\left\{r_i \geq d_l\right\} + \mu_j \cdot \mathbb{1}\left\{r_i < d_l\right\}\right],$$

and u^* is the optimal action in this setting.

The proofs of both Corollary 7 and 8 are provided in Appendix G.

E. 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 one arm to observe its reward for a cost $c \geq 0$. This is a complex problem setting different from most previous bandit formulations both due to the large action space of K^2 actions, and the possibility of still incurring regret due to the *reference point error* even when the chosen action is optimal. Further, the use of a stronger regret benchmark that uses the optimal action rather than μ^* makes the analysis rather intricate. We provide a gap-dependent regret bound of $O(K\log T)$ and a gap-independent bound of $O(\sqrt{KT\log T})$ for UCBP that match the order of the regret bounds of the standard UCB algorithm for the standard UCB problem.

Compared to UCB-naive-probe, and to the prior work for slightly different settings whose regrets scale with $\tilde{O}(K^2)$ on the number of arms, the regret of UCBP scales with

$\tilde{O}(K)$ since UCBP narrows down the action space by utilizing the structure of the problem. UCB-naive-probe further incurs an additional $\mathcal{D}$ term in regret as the reference point value affects the mean reward of a super arm. We would like to note that we assume cost of probing c as a constant for simplicity of the theoretical analysis, but this work can easily be extended to the setting where c is time dependent or cost of probing is different for each arm.

Algorithm 3 TSP

```
Input: cost of probing $c$, action set $\mathcal{A}$, exploration parameter $\beta$
Initialize: $N_i = 0,\ 1 \leq i \leq K$
 Sample each arm once
for each round $t$ do
    Sample $\theta_i(t) \sim N\left(\hat{\mu}_i(t), \frac{\beta}{N_i(t)}\right)$
    $i_t^* \leftarrow \arg\max_j \theta_j(t)$
    $i_t^{**} \leftarrow \arg\max_{j \neq i_t^*} \theta_j(t)$
    $\gamma_i(t) \sim N\left(\hat{\psi}_{(i,i_t^*)}(t), \frac{\beta}{N_i(t)} + \frac{\beta}{N_{i_t^*}(t)}\right),\ \forall i \neq i_t^*$
    $\gamma_{i_t^*}(t) \sim N\left(\hat{\psi}_{(i_t^*,i_t^{**})}(t), \frac{\beta}{N_{i^*(t)}(t)} + \frac{\beta}{N_{i_t^{**}}(t)}\right)$
    $j_t^* = \arg\max_{i \in [K]} \gamma_i(t)$
    $k_t = i_t^*$ if $j_t^* \neq i_t^*$, else $k_t = i_t^{**}$
    if $\theta_{i_t^*}(t) > \gamma_{j_t^*}(t)$ then
        Pull arm $i_t^*$, get $r(t) = r_t(i_t^*)$
    else
        Probe arm $j_t^*$, observe reward $r_t(j_t^*)$
        if $r_t(j_t^*) > \hat{\mu}_{k_t}(t)$ 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 $\hat{\mu}_i(t)$, and $N_i(t) = N_i(t-1) + 1$ for all observed arms $i \in o(t)$
end for
```

F. Simulations

We now evaluate the performance of the proposed UCBP Algorithm 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 as a baseline in §II; with TSP, the Thompson sampling based version of UCBP; and standard UCB that does not do any probing. The TSP algorithm operates as follows. First, samples $\theta_i(t)$ for mean arm rewards are generated for arms using a Gaussian distribution with mean $\hat{\mu}_i(t)$ and variance $\frac{\beta}{N_i(t)}$, where $\beta > 1$ is the exploration parameter. To estimate the mean probe reward, the backup arm will either be $i_t^* = \arg\max_j \theta_j(t)$ or $i_t^{**} = \arg\max_{j \neq i_t^*} \theta_j(t)$ depending on the probe arm. Note that this step is not done explicitly in the UCBP algorithm as the backup arm for the probing action with the highest UCB value is already either the arm with highest or second highest UCB value. After this step, the mean probe reward for action (i, j) can be calculated using these samples as

$$\hat{\psi}_{(i,j)}(t) = \sum_{\tau=1}^{t-1} \frac{\max(r_i(\tau), \theta_j(t)) \mathbb{1}\{i \in o(\tau)\}}{N_i(t)} - c.$$

We generate samples for the mean probe action reward using a Gaussian distribution with mean $\hat{\psi}_{(i,i_t^*)}(t)$ and variance $\frac{\beta}{N_i(t)} + \frac{\beta}{N_{i_t^*}(t)}$ for $i \neq i_t^*$, and using a Gaussian distribution with mean $\hat{\psi}_{(i_t^*,i_t^{**})}(t)$ and variance $\frac{\beta}{N_{i^*}(t)} + \frac{\beta}{N_{i_t^{**}}(t)}$ for $i \neq i_t^*$ when the probe arm is i_t^*. The action that has the largest sample value is chosen. If this action is probing, i.e. $a = (i, j)$, similar to the UCBP algorithm, arm i is probed to observe $r_i(t)$, then arm i is pulled if $r_i(t) > \hat{\mu}_j(t)$, and arm j otherwise. The pseudo-code of TSP is provided in Algorithm 3. The simulation results of UCB-naive-probe, UCBP, TSP, and UCB are provided below for the MovieLens and the Open Bandit datasets. We would like to note that we also clip the UCB indexes of UCB, UCBP, and UCB-naive-probe algorithms at 1 in the simulations. This reduces the number of probes in UCBP and UCB-naive-probe algorithms when the probing cost is high.

1) The MovieLens *Dataset:* The MovieLens dataset contains a total of 1M ratings on a total of 3883 movies, where a total of 6040 users rated the movies on a scale of 1 to 5 [55]. Using this dataset, we aim to provide the best genre recommendations to a population with an unknown demographic. To fit each movie into one genre, we pick one genre uniformly at random from the genres associated with each movie. We model each genre as an arm, where there are $K = 18$ arms, and the reward of an arm is obtained by sampling the rating of one of the users for a movie in that genre, chosen uniformly at random. The rewards of the arms are normalized to be between $[0, 1]$, and the mean reward of best action is around 0.864, and the reward of the best action without probing (the mean reward of the best arm) is around 0.792.

Our experimental results for this setting are shown in Figure 1, where we plot the cumulative regret averaged over 100 independent trials for $500,000$ rounds when the cost of probing is $c = 0$ (Upper Left), $c = 0.075$ (Upper Middle), $c = 0.25$ (Upper Right), $c = 0.5$ (Lower Left), and $c = 1$ (Lower Right). When $c = 0$, optimal action involves probing, and when $c = 0.075$ or higher; optimal action is pulling an arm without probing. The shaded area represents error bars with one standard deviation.

It can be seen that UCB has a linear regret curve when $c = 0$, and logarithmic regret curve when $c = 0.075$ or higher. This is as expected since UCB does not probe and hence cannot pull the optimal action when $c = 0$, which leads to linear regret. All other algorithms have a logarithmic regret curve. As expected, it can be seen that the UCBP algorithm outperforms the baseline UCB-naive-probe algorithm in all cost values tested. Comparing UCBP and TSP, it can be seen that both have very similar regret curves. UCBP performs slightly better than TSP when $c = 0$ and $c = 0.075$; and UCBP performs better with a larger difference than TSP when $c = 0.25$. While Thompson Sampling based algorithms are known

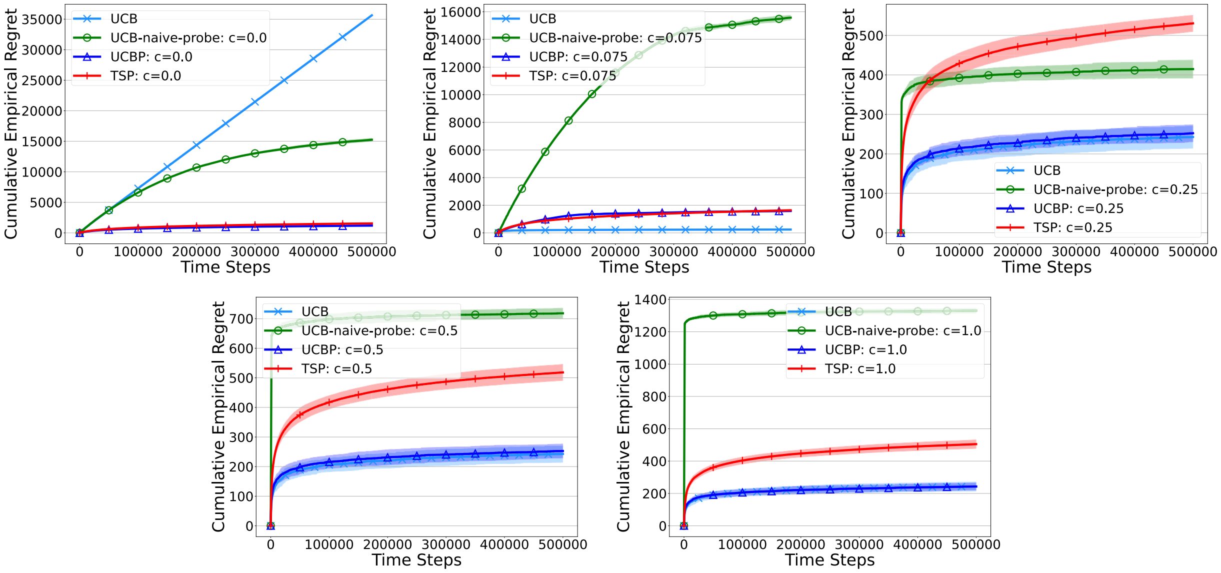

Fig. 1. Plots of the cumulative empirical regret of the UCB, UCBP, TSP and UCB-naive-probe algorithms for recommending the best genre in the MovieLens dataset.

to perform better empirically than UCB based algorithms in general, it was shown in [56] that Thompson Sampling might perform suboptimally in combinatorial bandits or in settings with high dimensions; hence these results are not unexpected. Comparing UCB with UCBP and TSP, it can be seen that UCB has much higher regret when $c = 0$ as it cannot choose the optimal action. When $c = 0.075$, the optimal action is pulling an arm directly, and probing is only slightly suboptimal. UCB has less regret in this case, as both UCBP and TSP need to spend some pulls on learning that probe actions are optimal. When $c = 0.25$ and higher, plots of UCB and UCBP overlap as UCBP does not choose to probe when the probe cost is too high and its behavior converges to UCB. UCB-naive-probe also does not choose to probe, but has higher regret than UCB or UCBP due to a larger regret from the initialization phase of the algorithm where all arms need to be pulled once.

2) The Open Bandit Dataset: Open Bandit Dataset is a public real-world logged bandit dataset provided by ZOZO, Inc., the largest fashion e-commerce company in Japan [57]. The dataset includes data from three different campaigns, and we selected the campaign from "Men" items which contains a total of $4,077,727$ data points showing whether the user clicked on the item or not when an item is recommended in one of the three positions, left, middle, or right. To make the clicks independent from the position, we only select the $1,358,878$ data points recommended in the left position. We model each item as an arm, there are $K = 34$ arms in total, and the rewards are binary indicating whether the user clicked on the item. The mean reward of best action is around 0.01697, and the reward of the best action without probing (the mean reward of the best arm) is around 0.00872. The goal is to recommend the best item to a cold (new) user. Our experimental results for this setting are shown in Figure 2, where we plot the cumulative regret averaged over 20 independent trials for $2,000,000$ rounds when the cost of probing an arm is $c = 0$ (Upper Left), $c = 0.005$ (Upper Middle), $c = 0.01$ (Upper Right), $c = 0.5$ (Lower Left), and $c = 1$ (Lower Right). The shaded area represents error bars with one standard deviation.

Again it can be seen that UCB has a linear regret curve when $c = 0$ or $c = 0.005$, and logarithmic regret curve when $c = 0.01$ or higher since the optimal action involves probing in the former case. It can be seen that all other algorithms have a logarithmic regret curve, and both the UCBP and the TSP algorithm outperforms the baseline UCB-naive-probe algorithm. This validates the usefulness of UCBP in practical settings. Comparing UCBP and TSP, it can be seen that TSP performs better than UCBP in all plots. Since this dataset has Bernoulli distribution, it may be argued that the performance difference between UCBP and TSP depends on the distribution of arm rewards.

Comparing UCB with UCBP and TSP, it can be seen that UCB has much higher regret when $c = 0$ or $c = 0.005$ as it cannot choose the optimal action. When $c = 0.01$, the optimal action is pulling an arm directly, and probing is only slightly suboptimal. UCB has less regret in this case, as both UCBP and TSP need to spend some pulls on these suboptimal probe actions. When $c = 0.5$ and higher, plots of UCB and UCBP are very close as UCBP does not choose to probe when the probe cost is too high and its behavior converges to UCB. The slight difference in the regret plots can be explained by the fact that a probe action obtains samples from two arms in a single round. Regarding UCB-naive-probe, it can be seen again that it has higher regret than UCB or UCBP due to a larger regret from the initialization phase of the algorithm.

V. Extension to Multiple Probes

One natural extension of our work is allowing multiple probes. Since the multiple probe setting is a much more complicated problem, here we study it only for Bernoulli arm rewards, and leave the consideration of more general

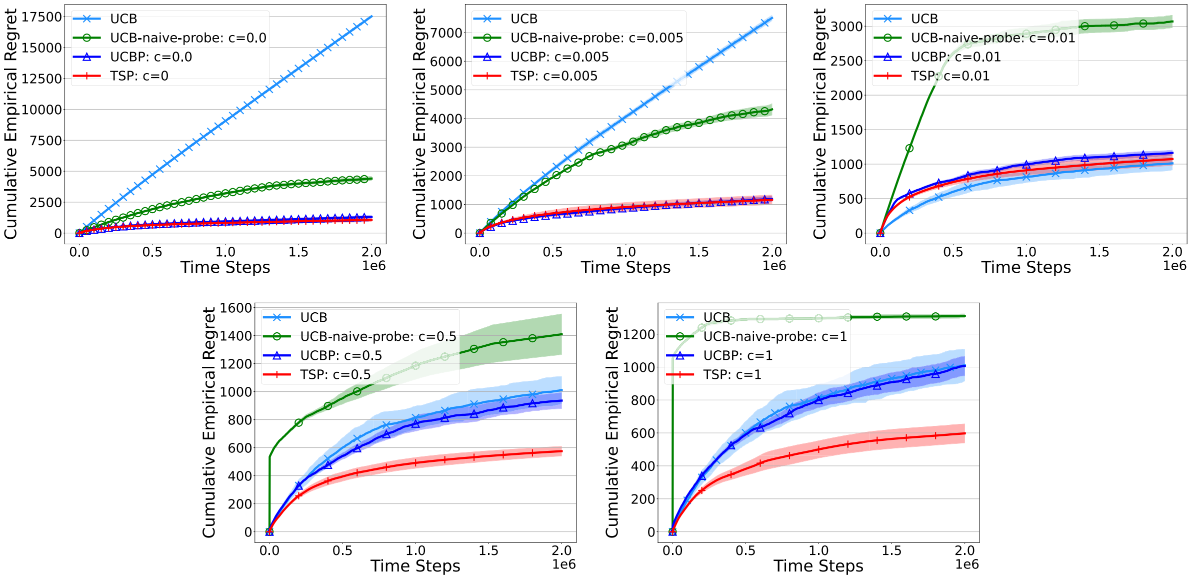

Fig. 2. Plots of the cumulative empirical regret of the UCBP and UCB-naive-probe algorithms for recommending the best item in the Open Bandit dataset.

Algorithm 4 UCBMP

```
Input: cost of probing $c$, action set $\mathcal{A}$
Initialize: $N_i = 0,\ 1 \le i \le K$
Sample each arm once
for each round $t$ do
    $S(t) = \text{argsort}_i - U_i(t)$
    Evaluate $P_i(t)$ values using Eq. (19)
    $s(t) = \arg\max_i P_i(t)$
    $j \leftarrow 0$
    for $i = 1$ to $s(t)$ do
        Probe arm $S_i(t)$, observe reward $r_i(t)$
        if $r_i(t) = 1$ then
            $j \leftarrow i$
            break
        end if
    end for
    If $j = s(t)$ and $r_j(t) = 0$, $j \leftarrow K$
    Pull arm $S_j(t)$, receive reward $r_j(t)$
    Update UCB indices for all observed arms
end for
```

bounded arm reward distributions for future work. Under Bernoulli arm rewards, the optimal strategy is to order the arms from highest to lowest mean reward, and probe the arms in this order until obtaining a reward of 1 if the cost to probe arms is ignored. But since probes have a cost, the optimal strategy also needs to terminate probing if the cost of probing exceeds the expected increase in reward through probing. Hence, the optimal action will have an upper limit on how many arms are allowed to be probed. For this end, we define R_i as the expected reward when at most i probes are allowed. It can be seen that R_i values can be evaluated as follows:

$$R_0 = \mu_1$$

$$R_1 = \mu_1 + (1-\mu_1)\cdot\mu_2 - c$$

$$R_2 = \mu_1 + (1-\mu_1)\cdot\mu_2 + (1-\mu_1)\cdot(1-\mu_2)\cdot\mu_3 - c\cdot(2-\mu_1)$$

$$R_i = \mu_1\cdot(1-c) + \sum_{j=2}^{i+1}\mu_j\cdot\prod_{k=1}^{j-1}(1-\mu_k) - c\cdot\sum_{j=2}^{i-1} j\cdot\mu_j\cdot\prod_{k=1}^{j-1}(1-\mu_k) - i\cdot c\cdot\prod_{j=1}^{i-1}(1-\mu_j), \quad 3 \le i \le K-1$$

Using these expected reward values, the upper limit on the number of allowed probes in the optimal action can then be found as:

$$s^* = \arg\max_{0\le i\le K-1} R_i$$

The optimal action can then be represented with the tuple $a^* = (1,\cdots,s^*)$, i.e. if $s^* \neq 0$ to probe arms from arm 1 to arm s^* in the given order until observing a reward of 1 and then pulling that arm. If no arm is probed or none of the probed arms produce a reward of 1, then the arm with $(s^*+1)^{\text{th}}$ highest mean reward is pulled. The optimal reward can be written as $\nu^* = R_{s^*}$.

We propose an algorithm called *Upper Confidence Bound with Multiple Probes* (UCBMP) that utilizes this optimal strategy to choose the optimal action. Since only the empirical mean estimates of the arms are known, UCBMP uses the UCB upper bound of empirical arm mean rewards to determine in which order arms should be probed. For this end, let $S(t)$ denote the ordered K-tuple whose elements are ordered by decreasing upper confidence values of arm rewards $U_i(t)$. At each round t, UCBMP first constructs this K-tuple $S(t)$, and then uses it to evaluate $P_i(t)$, the upper bound on the expected reward when at most i probes are allowed. These estimated $P_i(t)$ values can be

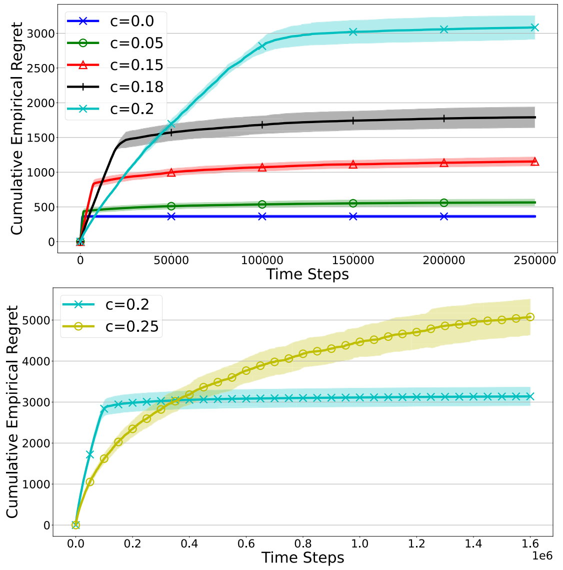

Fig. 3. Plots of the cumulative empirical regret of the UCBMP algorithm in a Bernoulli reward bandit setting with $K = 10$ arms for various probing cost values.

found as:

$$
\begin{aligned}
P_0(t) &= U_{S_1(t)}(t)\\
P_1(t) &= U_{S_1(t)}(t) + (1 - U_{S_1(t)}(t)) \cdot U_{S_2(t)}(t) - c\\
P_2(t) &= U_{S_1(t)}(t) + (1 - U_{S_1(t)}(t)) \cdot U_{S_2(t)}(t)\\
&\quad + (1 - U_{S_1(t)}(t)) \cdot (1 - U_{S_2(t)}(t)) \cdot U_{S_3(t)}(t)\\
&\quad - c \cdot (2 - U_{S_1(t)}(t))\\
P_i(t) &= U_{S_1(t)}(t) \cdot (1 - c)\\
&\quad + \sum_{j=2}^{i+1} U_{S_j(t)}(t) \cdot \prod_{k=1}^{j-1} (1 - U_{S_k(t)}(t))\\
&\quad - c \cdot \sum_{j=2}^{i-1} j \cdot U_{S_j(t)}(t) \cdot \prod_{k=1}^{j-1} (1 - U_{S_k(t)}(t))\\
&\quad - i \cdot c \cdot \prod_{j=1}^{i-1} (1 - U_{S_j(t)}(t)), \quad 3 \le i \le K-1
\end{aligned} \tag{19}
$$

The maximum number of probes that are allowed in round t, $s(t)$; is found as $s(t) = \arg\max_{0 \le i \le K-1} P_i(t)$. Arms are probed in the order of $S(t)$ until observing a reward of 1, and then that arm is pulled. If a reward of 1 is not observed in $s(t)$ probes, then arm $S_{s(t)+1}(t)$ is pulled. The reward $r(t)$ is received from the arm that is pulled. The pseudo-code is provided in Algorithm 4.

The regret of UCBMP can be written as $R(T) = T \cdot \nu^* - \sum_{t=1}^{T} r(t)$. To evaluate the performance of UCBMP in real world applications, we ran simulations for a Bernoulli bandit setting with $K = 10$ arms, where their mean reward vector is $\mu = [0.7, 0.69, 0.68, 0.67, 0.66, 0.65, 0.63, 0.6, 0.5, 0.4]$. The simulation results for this setting for cost values $c = [0, 0.05, 0.15, 0.18, 0.2, 0.25]$ are provided in Fig. 3. The optimal number of probes is $s^* = 9$ when cost is $0, 0.05, 0.15,$ or 0.18; is $s^* = 7$ when cost is 0.2; and is $s^* = 0$ when $c = 0.25$. As can be seen from the plots, regret of UCBMP scales sublinearly with t. While the plots can not be directly compared as the optimal reward value changes with cost, it can still be seen that in general regret increases with cost. This is because the number of arms that can be probed is higher when cost is low, which provides more reward observations per round. Also note that the plot for $c = 0.25$ converges slower because of this effect, since the optimal action is not to make any probes, arm reward observations are collected slower in time. The theoretical analysis of UCBMP is much more intricate, hence we leave the regret analysis of UCBMP as future work.

VI. Concluding Remarks

In this paper, we introduce a previously unexplored setting for the multi-armed bandit problem with probes, where before pulling an arm, the agent is allowed to probe one arm to observe its reward, which is sampled from a bounded distribution, for a cost $c \ge 0$. We introduce a new regret definition that is based on the expected reward of the optimal action, and we identify the optimal strategy. We provide UCBP, a novel algorithm that utilizes this strategy to achieve a gap-independent regret upper bound that scales with $O(\sqrt{KT \log T})$, and a gap-dependent bound that scales with $O(K \log T)$ if rewards are discrete. To demonstrate the empirical performance of UCBP, we provide a naive UCB-based approach that has a gap-independent regret upper bound on the order of $O(\sqrt{K^2 T \log T})$, and a gap-dependent bound on the order of $O(K^2 \log T)$. We use this algorithm as a baseline in our simulations, and simulation results corroborate the better performance of UCBP over the UCB-naive-probe algorithm, and validate the utility of UCBP in practical settings.

Our work opens multiple directions for future research. In Section V, we extend our setting to multiple probes for each round when the reward distributions of arms are Bernoulli, and we provide the UCBMP algorithm. This can be further extended by providing the theoretical analysis of UCBMP, and extending UCBMP to more general bounded arm reward distributions in future work. Another interesting future direction is to extend our bandit results to the case with imperfect probes. We believe this can be accomplished by deriving confidence intervals for the probe reward since the upper confidence index of the probe outcome can be used to decide whether to pull the probe arm or the backup arm. We anticipate the regret analysis for this case to be challenging since the uncertainty of the actions with probes will induce further suboptimal actions to be taken by the algorithm. Lastly, the case where the rewards of different arms are correlated can also be considered. 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.

Appendix A
Preliminaries

Before presenting the regret analysis of the UCBP algorithm, we start by presenting some well-known properties.

Fact 9 (Hoeffding's Inequality): Let $Z_1, Z_2, \cdots, Z_n$ be independent random variables bounded between $a_i \leq Z_i \leq b_i$, then for any $\delta > 0$, we have

$$\mathbb{P}\left(\frac{\sum_{i=1}^n Z_i}{n} - \ \mathbb{E}[Z] \geq \delta\right) \leq e^{-\frac{2n^2\delta^2}{\sum_{i=1}^n (b_i - a_i)^2}} \ .$$

Lemma 10 ([2, Theorem 2]): Consider a K-armed bandit problem with reward distributions $\Gamma = (\Gamma_1, \cdots, \Gamma_K),\ \Gamma \in \Theta$ where $\Gamma_i,\ i \in [K]$ is the reward distribution of arm i. Also define $\Theta^i = \{\Gamma : \mu(\Gamma_i) > \max_{j \neq i} \mu(\Theta_j)\}$ as the parameter set where arm i is the unique optimal arm. An algorithm $\pi \in \Pi$ is defined as *uniformly good* if for all $\Gamma \in \Theta^i$, $R^\pi(T) = o(T^a)$, for all $a > 0$. Let $D_{KL}(\cdot||\cdot)$ denote the Kullback-Leibler divergence. Assume that $D_{KL}(\Gamma||\lambda)$, satisfies the following two conditions:

a) $0 < D_{KL}(\Gamma, \lambda) < \infty$ whenever $\mu(\lambda) > \mu(\Gamma)$, and
b) $\forall \epsilon > 0$ and $\forall \epsilon > 0$ and $\forall\ \Sigma, \lambda \in \Theta$ such that $\mu(\lambda) > \mu(\Sigma), \exists \delta = \delta(\epsilon, \Sigma, \lambda) > 0$ for which $|D_{KL}(\Gamma, \lambda) - D_{KL}(\Gamma, \lambda')| < \epsilon$ whenever $\mu(\lambda) \leq \mu(\lambda') \leq \mu(\lambda) + \delta$

Also assume that Θ is such that $\forall \lambda \in \Theta$ and $\forall \delta > 0, \exists \lambda' \in \Gamma$ such that $\mu(\lambda) < \mu(\lambda') < \mu(\lambda) + \delta$.

Let $\pi \in \Pi$ be a uniformly good algorithm. Under these assumptions, for any $\Gamma \in \Theta^j$, it holds that

$$\liminf_{T \to \infty} \frac{\mathbb{E}[N_i(T)]}{\log T} \geq \frac{1}{D_{KL}(\Gamma_i, \Gamma^*)},\ \forall i \neq j$$

Fact 11 (Conditional Probabilities): The probability of an event A can be upper bounded by conditioning on an event B as follows

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A, B) + \mathbb{P}(A, B^c)\\ &= \mathbb{P}(A|B)\,\mathbb{P}(B) + \mathbb{P}(A|B^c)\,\mathbb{P}(B^c)\\ &\leq \mathbb{P}(A|B) + \mathbb{P}(B^c)\ .\end{aligned}$$

Upper bounds of similar form are used throughout the proof.

Fact 12: We include the following trivial bounds for the max function when $b > 0,\ c > 0$:

$$\begin{aligned}&(i) \max(a, b + c) \leq \max(a, b) + c\\ &(ii) \max(a, b - c) \geq \max(a, b) - c\\ &(iii) \max(a, b) \pm c = \max(a \pm c, b \pm c)\ .\end{aligned}$$

(iv) We also note the following inequality when $a, b > 0$:

$$\begin{aligned}&\mathbb{E}[\max(r_i, a)] + b\mathbb{P}(r_i < a)\\ &\qquad\qquad \leq \mathbb{E}[\max(r_i, a + b)]\\ &\qquad\qquad \leq\ \mathbb{E}[\max(r_i, a)] + b\mathbb{P}(r_i \leq a + b)\ .\end{aligned}$$

Corollary 13: The mean reward of a probe action $a = (i, j)$ can be upper bounded as

$$\nu_{(i,j)} \leq \mu_i + p_{(i,j)} \cdot \mu_j$$

Proof:

$$\begin{aligned}\nu_{(i,j)} &= \mathbb{E}[\max(r_i, \mu_j)]\\ &= \mathbb{E}[\max(r_i, \mu_j)\mathbb{1}\{r_i > \mu_j\}\\ &\qquad + \max(r_i, \mu_j)\mathbb{1}\{r_i \leq \mu_j\}]\\ &= \mathbb{E}[r_i\mathbb{1}\{r_i > \mu_j\} + \mu_j\mathbb{1}\{r_i \leq \mu_j\}]\\ &= \int_{r > \mu_j} r f_i(r)dr + \int_{r \leq \mu_j} \mu_j f_i(r)dr\\ &= \int_{r > \mu_j} r f_i(r)dr + \int_{r \leq \mu_j} r f_i(r)dr\\ &\qquad + \int_{r \leq \mu_j} (\mu_j - r) f_i(r)dr\\ &= \int r f_i(r)dr + \int_{r \leq \mu_j} (\mu_j - r) f_i(r)dr\\ &\leq \mu_i + \int_{r \leq \mu_j} \mu_j f_i(r)dr\\ &= \mu_i + \mathbb{P}(r_i \leq \mu_j)\,\mu_j = \mu_i + p_{(i,j)} \cdot \mu_j\end{aligned}$$

where $f_i(r)$ is used to denote the probability distribution function of the reward of arm i. □

Corollary 14: Letting $\nu(i, \mu) := \mathbb{E}[\max(r_i, \mu)]$, $\nu(i, \mu + \delta) - \nu(i, \mu)$ can be upper bounded as

$$\nu(i, \mu + \delta) - \nu(i, \mu) \leq \delta \cdot \mathbb{P}(r_i \leq \mu + \delta)$$

Proof:

$$\begin{aligned}\nu(i, \mu) =&\mathbb{E}[\max(r_i, \mu) \cdot \mathbb{I}(r_i > \mu) + \max(r_i, \mu) \cdot \mathbb{I}(r_i \leq \mu)]\\ =&\mathbb{E}[r_i \cdot \mathbb{I}(r_i > \mu) + \mu \cdot \mathbb{I}(r_i \leq \mu)]\\ =&\mathbb{E}[r_i \cdot \mathbb{I}(r_i > \mu)] + \mu \cdot \mathbb{P}(r_i \leq \mu)\\ =&\int_{r > \mu} r f_i(r)dr + \int_{r \leq \mu} \mu f_i(r)dr\\ =&\int_{\mu < r \leq \mu + \delta} r f_i(r)dr + \int_{r > \mu + \delta} r f_i(r)dr\\ &\qquad + \int_{r \leq \mu} \mu f_i(r)dr.\end{aligned}$$

where $f_i(r)$ is used to denote the probability distribution function of the reward of arm i. Similarly,

$$\begin{aligned}&\nu(i, \mu + \delta)\\ &\quad =\mathbb{E}[\max(r_i, \mu + \delta) \cdot \mathbb{I}(r_i > \mu + \delta)\\ &\qquad\quad + \max(r_i, \mu + \delta) \cdot \mathbb{I}(r_i \leq \mu + \delta)]\\ &\quad =\mathbb{E}[r_i \cdot \mathbb{I}(r_i > \mu + \delta) + (\mu + \delta) \cdot \mathbb{I}(r_i \leq \mu + \delta)]\\ &\quad =\mathbb{E}[r_i \cdot \mathbb{I}(r_i > \mu + \delta)] + (\mu + \delta) \cdot \mathbb{P}(r_i \leq \mu + \delta)\\ &\quad = \int_{r > \mu + \delta} r f_i(r)dr + \int_{r \leq \mu + \delta} (\mu + \delta) f_i(r)dr\\ &\quad = \int_{r > \mu + \delta} r f_i(r)dr + \int_{r \leq \mu + \delta} (\mu + \delta) f_i(r)dr\\ &\quad = \int_{r > \mu + \delta} r f_i(r)dr + \int_{\mu < r_i \leq \mu + \delta} (\mu + \delta) f_i(r)dr\\ &\qquad\qquad + \int_{r \leq \mu} (\mu + \delta) f_i(r)dr.\end{aligned}$$

Then,

$$\begin{aligned}\nu(i, \mu + \delta) - \nu(i, \mu) &= \delta \int_{r \leq \mu} f_i(r)dr\\ &\qquad + \int_{\mu < r_i \leq \mu + \delta} (\mu + \delta - r) f_i(r)dr\\ &\leq \delta \int_{r \leq \mu} f_i(r)dr\end{aligned}$$

$$
\begin{aligned}
&+\delta \int_{\mu<r_i\le\mu+\delta} f_i(r)dr\\
&\le \delta \int_{r\le\mu+\delta} f_i(r)dr\\
&\le \delta\cdot\mathbb{P}\left(r_i\le\mu+\delta\right)
\end{aligned}
$$

□

Appendix B
Derivation of Confidence Intervals For Actions

In this section we derive the high probability confidence intervals for the pull and probe actions of UCBP.

Fact 15: The following holds for all $i\in[K]$ and $u<t$.

$$\mathbb{P}\left(\hat{\mu}_i(t,u)-\mu_i>\sqrt{\frac{3\log t}{u}}\right)\le e^{-6\log t}=t^{-6}\ , \tag{20}$$

$$\mathbb{P}\left(\mu_i-\hat{\mu}_i(t,u)>\sqrt{\frac{3\log t}{u}}\right)\le e^{-6\log t}=t^{-6}\ . \tag{21}$$

Proof: We prove (20).

$$
\begin{aligned}
&\mathbb{P}\left(\hat{\mu}_i(t,u)-\mu_i>\sqrt{\frac{3\log t}{u}}\right)\\
&\qquad\le\mathbb{P}\left(\frac{\sum_{\tau=1}^{u}\tilde{r}_i(\tau)}{u}-\mu_i\ge\sqrt{\frac{3\log t}{u}}\right)\\
&\qquad\le e^{-\frac{2u^2\left(\sqrt{\frac{3\log t}{u}}\right)^2}{u}}=t^{-6}\ ,
\end{aligned}
$$

where the last line follows from Fact 9. (21) can be proven via a similar argument. □

Corollary 16: The following holds for all $a=(i,\emptyset)\in\mathcal{A}_s$ and $u<t$.

$$
\begin{aligned}
&\mathbb{P}\left(\hat{\nu}_{(i,\emptyset)}(t,u)>U_{(i,\emptyset)}(t,u)\right)\le t^{-6}\ ,\\
&\mathbb{P}\left(\hat{\nu}_{(i,\emptyset)}(t,u)<L_{(i,\emptyset)}(t,u)\right)\le t^{-6}\ .
\end{aligned}
$$

Proof: The result follows from Fact 15 by observing that $\hat{\nu}_{(i,\emptyset)}(t,u):=\hat{\mu}_i(t,u)$. □

Corollary 17: The following two inequalities hold for all $a=(i,j)\in\mathcal{A}_p$ and $r,s<t$.

$$
\begin{aligned}
&\mathbb{P}\left(\nu_{(i,j)}>U_{(i,j)}(t,r,s)\right)\\
&\le\mathbb{P}\left(\nu_{(i,j)}\ge\sum_{\tau=1}^{r}\frac{\max\left(\tilde{r}_i(\tau),\hat{\mu}_j(t,s)+\sqrt{\frac{3\log t}{s}}\right)}{r}-c+\sqrt{\frac{3\log t}{r}}\right)\le 2t^{-6}\ ,
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{P}\left(\nu_{(i,j)}<L_{(i,j)}(t,r,s)\right)\\
&\le\mathbb{P}\left(\nu_{(i,j)}\le\sum_{\tau=1}^{r}\frac{\max\left(\tilde{r}_i(\tau),\hat{\mu}_j(t,s)-\sqrt{\frac{3\log t}{s}}\right)}{r}-c-\sqrt{\frac{3\log t}{r}}\right)\le 2t^{-6}\ .
\end{aligned}
$$

Proof: We have

$$
\begin{aligned}
&\mathbb{P}\left(\nu_{(i,j)}\ge\sum_{\tau=1}^{r}\frac{\max\left(\tilde{r}_i(\tau),\hat{\mu}_j(t,s)+\sqrt{\frac{3\log t}{s}}\right)}{r}-c+\sqrt{\frac{3\log t}{r}}\right)\\
&=\mathbb{P}\left(\nu_{(i,j)}\ge\sum_{\tau=1}^{r}\frac{\max\left(\tilde{r}_i(\tau),\hat{\mu}_j(t,s)+\sqrt{\frac{3\log t}{s}}\right)}{r}-c+\sqrt{\frac{3\log t}{r}},\ \hat{\mu}_j(t,s)\ge\mu_j-\sqrt{\frac{3\log t}{s}}\right)\\
&d+\mathbb{P}\left(\nu_{(i,j)}\ge\sum_{\tau=1}^{r}\frac{\max\left(\tilde{r}_i(\tau),\hat{\mu}_j(t,s)+\sqrt{\frac{3\log t}{s}}\right)}{r}-c+\sqrt{\frac{3\log t}{r}},\ \hat{\mu}_j(t,s)<\mu_j-\sqrt{\frac{3\log t}{s}}\right)\\
&\le\mathbb{P}\left(\nu_{(i,j)}\ge\sum_{\tau=1}^{r}\frac{\max\left(\tilde{r}_i(\tau),\mu_j\right)}{r}-c+\sqrt{\frac{3\log t}{r}}\right)\\
&\qquad+\mathbb{P}\left(\hat{\mu}_j(t,s)<\mu_j-\sqrt{\frac{3\log t}{s}}\right)\\
&\le\mathbb{P}\left(\mathbb{E}[\max(r_i,\mu_j)]-c\ge\sum_{\tau=1}^{r}\frac{\max\left(\tilde{r}_i(\tau),\mu_j\right)}{r}-c+\sqrt{\frac{3\log t}{r}}\right)+t^{-6} \qquad (22)\\
&=\mathbb{P}\left(\sum_{\tau=1}^{r}\frac{\max\left(\tilde{r}_i(\tau),\mu_j\right)}{r}-\mathbb{E}[\max(r_i,\mu_j)]\le-\sqrt{\frac{3\log t}{r}}\right)+t^{-6}\\
&\le 2\,t^{-6}\ , \qquad (23)
\end{aligned}
$$

where (22) uses Fact 15, and (23) follows from the fact that $\{\max(\tilde{r}_i(\tau),\mu_j)\}_{\tau=1}^{r}$ forms a sequence of i.i.d. random variables with mean $\mathbb{E}[\max(r_i,\mu_j)]$ together with the Hoeffding bound in Fact 15 for i.i.d. random variables.

Similarly, we have

$$
\begin{aligned}
&\mathbb{P}\left(\nu_{(i,j)}\le\sum_{\tau=1}^{r}\frac{\max\left(\tilde{r}_i(\tau),\hat{\mu}_j(t,s)-\sqrt{\frac{3\log t}{s}}\right)}{r}-c-\sqrt{\frac{3\log t}{r}}\right)\\
&=\mathbb{P}\left(\nu_{(i,j)}\le\sum_{\tau=1}^{r}\frac{\max\left(\tilde{r}_i(\tau),\hat{\mu}_j(t,s)-\sqrt{\frac{3\log t}{s}}\right)}{r}\right.
\end{aligned}
$$

$$
\begin{aligned}
&\qquad -c-\sqrt{\frac{3\log t}{r}},\ \tilde{\mu}_j(s)\le \mu_j+\sqrt{\frac{3\log t}{s}}\Bigg) \\
&+\mathbb{P}\left(\nu_{(i,j)}\le \sum_{\tau=1}^{r}\frac{\max\left(\tilde{r}_i(\tau),\hat{\mu}_j(t,s)-\sqrt{\frac{3\log t}{s}}\right)}{r}\right. \\
&\qquad \left.-c-\sqrt{\frac{3\log t}{r}},\ \hat{\mu}_j(t,s)>\mu_j+\sqrt{\frac{3\log t}{s}}\right) \\
&\le \mathbb{P}\left(\nu_{(i,j)}\le \sum_{\tau=1}^{r}\frac{\max\left(\tilde{r}_i(\tau),\mu_j\right)}{r}-c-\sqrt{\frac{3\log t}{r}}\right) \\
&\qquad +\mathbb{P}\left(\hat{\mu}_j(t,s)>\mu_j+\sqrt{\frac{3\log t}{s}}\right) \\
&\le \mathbb{P}\left(\mathbb{E}[\max(r_i,\mu_j)]-c\le \sum_{\tau=1}^{r}\frac{\max\left(\tilde{r}_i(\tau),\mu_j\right)}{r}-c\right. \\
&\qquad \left.-\sqrt{\frac{3\log t}{r}}\right)+t^{-6} \qquad (24) \\
&= \mathbb{P}\left(\sum_{\tau=1}^{r}\frac{\max\left(\tilde{r}_i(\tau),\mu_j\right)}{r}-\mathbb{E}[\max(r_i,\mu_j)]\ge\sqrt{\frac{3\log t}{r}}\right) \\
&\qquad +t^{-6} \\
&\le 2\ t^{-6}\ , \qquad (25)
\end{aligned}
$$

where again (24) uses Fact 15, and (25) follows from the fact that $\{\max(\tilde{r}_i(\tau),\mu_j)\}_{\tau=1}^{r}$ forms a sequence of i.i.d. random variables with mean $\mathbb{E}[\max(r_i,\mu_j)]$ together with the Hoeffding bound in Fact 15 for i.i.d. random variables. □

Corollary 18: For probe action $a=(i,j)\in\mathcal{A}_p$, the following holds at round t given $\mathcal{H}_t$ and under event $\mathcal{E}_t$.

$$
\begin{aligned}
&U_{(i,j)}(t)-\nu_{(i,j)} \\
&\qquad \le 2C_i(t)+2\mathbb{P}\left(r_i(t)\le\mu_j+2C_j(t)|\mathcal{H}_t\right)\cdot C_j(t)\ .
\end{aligned}
$$

Proof: We have

$$
\begin{aligned}
U_{(i,j)}(t) &= \sum_{\tau=1}^{N_i(t)}\frac{\max\left(\tilde{r}_i(\tau),\hat{\mu}_j(t)+C_j(t)\right)}{N_i(t)}-c+C_i(t) \\
&\le \sum_{\tau=1}^{N_i(t)}\frac{\max\left(\tilde{r}_i(\tau),\mu_j+2C_j(t)\right)}{N_i(t)}-c+C_i(t) \qquad (26) \\
&\le \mathbb{E}\left[\max(r_i,\mu_j+2C_j(t))|\mathcal{H}_t\right]-c+2C_i(t) \qquad (27)
\end{aligned}
$$

where (26) is due to $\hat{\mu}_j(t)\le\mu_j+C_j(t)$ on $\mathcal{E}_t$. (27) is due to

$$
\sum_{\tau=1}^{N_i(t)}\frac{\max\left(\tilde{r}_i(\tau),\mu_j+2C_j(t)\right)}{N_i(t)}\le\mathbb{E}\left[\max(r_i,\mu_j+2C_j(t))|\mathcal{H}_t\right]
$$

on $\mathcal{E}_t$. Note that given $\mathcal{H}_t$, $C_j(t)$ is deterministic so we can use Fact 15 as in the proof of Corollary 17. To proceed, note that

$$
\begin{aligned}
&\mathbb{E}\left[\max(r_i,\mu_j+2C_j(t))|\mathcal{H}_t\right] \\
&= \int_r \max(r,\mu_j+2C_j(t))f(r)dr \\
&= \int_{r\le\mu_j}\max(r,\mu_j+2C_j(t))f(r)dr \\
&\quad +\int_{\mu_j<r\le\mu_j+2C_j(t)}\max(r,\mu_j+2C_j(t))f(r)dr \\
&\quad +\int_{\mu_j+2C_j(t)<r}\max(r,\mu_j+2C_j(t))f(r)dr \\
&= \int_{r\le\mu_j}(\mu_j+2C_j(t))f(r)dr \\
&\quad +\int_{\mu_j<r\le\mu_j+2C_j(t)}(\mu_j+2C_j(t))f(r)dr \\
&\quad +\int_{\mu_j+2C_j(t)<r}rf(r)dr \\
&\le \int_{r\le\mu_j}(\mu_j+2C_j(t))f(r)dr \\
&\quad +\int_{\mu_j<r\le\mu_j+2C_j(t)}(r+2C_j(t))f(r)dr \\
&\quad +\int_{\mu_j+2C_j(t)<r}rf(r)dr \\
&= \int_{r\le\mu_j}\mu_jf(r)dr+\int_{r>\mu_j}rf(r)dr \\
&\quad +2\int_{r\le\mu_j+2C_j(t)}C_j(t)f(r)dr \\
&= \int_r\max(r,\mu_j)f(r)dr+2\mathbb{P}(r_i(t)\le\mu_j+2C_j(t))\cdot C_j(t) \\
&= \mathbb{E}\left[\max(r_i,\mu_j)\right]+2\mathbb{P}(r_i(t)\le\mu_j+2C_j(t))\cdot C_j(t)\ .
\end{aligned}
$$

Using this, it can be seen that

$$
\begin{aligned}
U_{(i,j)}(t) &\le \mathbb{E}\left[\max(r_i,\mu_j)\right] \\
&\qquad +2\cdot\mathbb{P}\left(r_i(t)\le\mu_j+2C_j(t)|\mathcal{H}_t\right)\cdot C_j(t) \\
&\qquad -c+2C_i(t) \\
&= \nu_{(i,j)}+2C_i(t) \\
&\qquad +2\cdot\mathbb{P}\left(r_i(t)\le\mu_j+2C_j(t)|\mathcal{H}_t\right)\cdot C_j(t).
\end{aligned}
$$

□

Appendix C
Expected Number of Violations of Good Event

Lemma 19: We have $\sum_{t=K+1}^{T}\mathbb{P}(\mathcal{E}_t^c)\le\frac{2\pi^2K^2}{3}$.

Proof: Note that

$$
\begin{aligned}
\sum_{t=1}^{T}\mathbb{1}\left\{\mathcal{E}_{t,a}^c\right\} &= \sum_{t=1}^{T}\mathbb{1}\left\{\min_{r\le t,s\le t}U_{a^*}(t,r,s)\ge\nu^*\right. \\
&\qquad \left.\wedge\max_{u\le t,v\le t}L_a(t,u,v)\le\nu_a\right\} \\
&\le \sum_{t=1}^{T}\sum_{r=1}^{t}\sum_{s=1}^{t}\sum_{u=1}^{t}\sum_{v=1}^{t}\mathbb{1}\left\{U_{a^*}(t,r,s)\ge\nu^*\right. \\
&\qquad \left.\wedge L_a(t,u,v)\le\nu_a\right\}\ .
\end{aligned}
$$

By the monotonicity of expectation, it holds that

$$
\begin{aligned}
\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{E}_{t,a}^c\right) \le \sum_{t=1}^{T}\sum_{r=1}^{t}\sum_{s=1}^{t}\sum_{u=1}^{t}\sum_{v=1}^{t}\mathbb{P}\left(U_{a^*}(t,r,s)\ge\nu^*\right. \\
\left.\wedge L_a(t,u,v)\le\nu_a\right)
\end{aligned}
$$

$$\le \sum_{t=1}^{T}\sum_{r=1}^{t}\sum_{s=1}^{t}\sum_{u=1}^{t}\sum_{v=1}^{t} 4t^{-6} \tag{28}$$

$$\le \sum_{t=1}^{T} 4t^{-2} \le \frac{2\pi^2}{3} \ ,$$

where we used Corollary 17 in (28). The result follows using $\mathcal{E}_t = \bigcap_{a\in\mathcal{A}} \mathcal{E}_{t,a}$.

$$\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{E}_t^c\right) = \sum_{t=1}^{T}\mathbb{P}\left(\bigcup_{a\in\mathcal{A}}\mathcal{E}_{t,a}^c\right) \le \sum_{a\in\mathcal{A}}\sum_{t=1}^{T}\mathbb{P}\left(\mathcal{E}_{t,a}^c\right) = \frac{2\pi^2K^2}{3} \ .$$

□

Appendix D
Proof of Theorem 4

Recall the regret decomposition in Lemma 1 and Lemma 2.

Define $o(t) \subset a_t$ as the set of arms whose reward is observed in round t; and $\mathcal{H}_t = (a_1, r(1), o(1), \cdots, a_{t-1}, r(t-1), o(t-1))$ as the history of UCBP up to choosing action a_t, and let $\mathbb{E}[\cdot|\mathcal{H}_t]$ be the conditional expectation given this history. Using Lemma 2, $R_s(T)$ can be decomposed as

$$R_s(T) = 2\mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{(i,\cdot)\in\mathcal{A}}\mathbb{1}\left\{a_t=(i,\cdot),\mathcal{E}_t\right\}\cdot C_i(t)\right] + 2\mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{(i,j)\in\mathcal{A}_p}\mathbb{1}\left\{a_t=(i,j),\mathcal{E}_t\right\} \cdot\mathbb{P}\left(r_i(t)\le\mu_j+2C_j(t)|\mathcal{H}_t\right)\cdot C_j(t)\right] \ .$$

To proceed, we note the following condition for an action to be chosen at round t by UCBP. Given event $\mathcal{E}_t$, for action a to occur in round t, the upper confidence index of action a needs to be above the upper confidence index of the optimal action a^* at round t. Hence, the action a can only be chosen in round t if

$$U_{a^*}(t) \le U_a(t)$$

is satisfied. If $a=(i,j)$ is a probe action, using Corollary 18, and the fact that $\nu^* \le U_{a^*}(t)$, we have

$$\begin{aligned}\nu^* &\le U_{a^*}(t)\\ &\le U_a(t)\\ &\le \nu_a + 2C_i(t) + 2\mathbb{P}\left(r_i(t)\le\mu_j+2C_j(t)|\mathcal{H}_t\right)\cdot C_j(t) \ .\end{aligned}$$

Hence, action $a=(i,j)$ can be chosen only if

$$\Delta_a \le 2C_i(t) + 2\mathbb{P}\left(r_i(t)\le\mu_j+2C_j(t)|\mathcal{H}_t\right)\cdot C_j(t) \ .$$

Similarly, if $a=(i,\emptyset)$ is a pull action, it can be chosen only if

$$\Delta_a \le 2C_i(t) \ .$$

Defining $\Delta_{\min,i} = \min_{a\in\mathcal{A}_i:\Delta_a>0}\Delta_a$, we apply the reverse amortization trick that is used in [42, Theorem 4] for an action $a_t=(i_t,j_t)$ as follows.

$$\begin{aligned}\Delta_{a_t} &\le 2C_{i_t}(t) + 2\mathbb{P}\left(r_{i_t}(t)\le\mu_{j_t}+2C_{j_t}(t)|\mathcal{H}_t\right)\cdot C_{j_t}(t)\\ &\le -\Delta_{a_t} + 4C_{i_t}(t)\\ &\quad + 4\mathbb{P}\left(r_{i_t}(t)\le\mu_{j_t}+2C_{j_t}(t)|\mathcal{H}_t\right)\cdot C_{j_t}(t) \qquad (29)\\ &\le \left(-\frac{\Delta_{a_t}}{2}+4C_{i_t}(t)\right)\\ &\quad + \mathbb{P}\left(r_{i_t}(t)\le\mu_{j_t}+2C_{j_t}(t)|\mathcal{H}_t\right)\\ &\qquad \cdot\left(-\frac{\Delta_{a_t}}{2}+4C_{j_t}(t)\right) \qquad (30)\\ &\le 4\left(-\frac{\Delta_{\min,i_t}}{8}+\sqrt{\frac{3\log T}{N_{i_t}(t-1)}}\right)\\ &\quad + 4\mathbb{P}\left(r_{i_t}(t)\le\mu_{j_t}+2C_{j_t}(t)|\mathcal{H}_t\right)\\ &\quad \cdot\left(-\frac{\Delta_{\min,j_t}}{8}+\sqrt{\frac{3\log T}{N_{j_t}(t-1)}}\right) \ ,\end{aligned}$$

where Eq. (29) is one of the main observations for the reverse amortization trick which brings the gap to the right side of the equation, and the fact that $\mathbb{P}\left(r_{i_t}(t)\le\mu_j+2C_{j_t}(t)|\mathcal{H}_t\right)\le 1$ is used in Eq. (30). Similarly, for the case where $a_t=(i_t,\emptyset)$ the following can be obtained

$$\Delta_{a_t} \le 4\left(-\frac{\Delta_{\min,i_t}}{8}+\sqrt{\frac{3\log T}{N_{i_t}(t-1)}}\right) \ .$$

Define

$$\kappa_{i,T}(\ell) = \begin{cases} 4\sqrt{\frac{3\log T}{\ell}}, & \text{if } 1\le\ell\le L_{i,T},\\ 0, & \text{if } \ell > L_{i,T},\end{cases}$$

where $L_{i,T} = \frac{192\log T}{(\Delta_{\min,i})^2}$. It can be seen that the reverse amortization trick greatly simplifies upper bounding the regret since regret is not incurred when $\ell > L_{i,T}$. Regret can be written in terms of this $\kappa_{i,T}(\ell)$ term as

$$\begin{aligned} &R_s(T)\\ &\le \mathbb{E}\Bigg[\sum_{t=K+1}^{T}\sum_{(i,j)\in\mathcal{A}_p}\mathbb{1}\left\{a_t=(i,j),\mathcal{E}_t\right\}\cdot \ [\kappa_{i,T}\left(N_i(t)\right)\\ &\quad +\mathbb{P}\left(r_i(t)\le\mu_j+2C_j(t)|\mathcal{H}_t\right)\cdot\kappa_{j,T}\left(N_j(t)\right)]\Bigg]\\ &\quad + \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{(i,\emptyset)\in\mathcal{A}_s}\mathbb{1}\left\{a_t=(i,\emptyset),\mathcal{E}_t\right\} \cdot\kappa_{i,T}\left(N_i(t)\right)\right]\\ &= \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{(i,j)\in\mathcal{A}_p}\mathbb{1}\left\{a_t=(i,j),\mathcal{E}_t\right\} \cdot\mathbb{P}\left(r_i(t)\le\mu_j+2C_j(t)|\mathcal{H}_t\right)\cdot\kappa_{j,T}\left(N_j(t)\right)\right]\\ &\quad + \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{(i,j)\in\mathcal{A}}\mathbb{1}\left\{a_t=(i,j),\mathcal{E}_t\right\}\cdot\kappa_{i,T}\left(N_i(t)\right)\right]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\Bigg[\sum_{t=K+1}^{T}\sum_{(i,j)\in\mathcal{A}_p} \mathbb{1}\left\{\mathcal{E}_t\right\}\cdot\mathbb{P}\left(r_i(t)\le \mu_j+2C_j(t)|\mathcal{H}_t\right)\\
&\qquad\cdot\kappa_{j,T}\left(N_j(t)\right)\cdot\mathbb{E}\Bigg[\mathbb{1}\left\{a_t=(i,j)\right\}\\
&\qquad\cdot\frac{\mathbb{1}\left\{j\in o(t)\right\}}{\mathbb{P}\left(r_i(t)\le U_j(t)|\mathcal{H}_t\right)}\Bigg|\mathcal{H}_t\Bigg]\Bigg]\\
&\quad+\mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{(i,j)\in\mathcal{A}} \mathbb{1}\left\{a_t=(i,j),\mathcal{E}_t\right\}\cdot\kappa_{i,T}\left(N_i(t)\right)\right] \qquad (31)\\
&= \mathbb{E}\Bigg[\sum_{t=K+1}^{T}\mathbb{E}\Bigg[\sum_{(i,j)\in\mathcal{A}_p} \mathbb{1}\left\{a_t=(i,j),j\in o(t),\mathcal{E}_t\right\}\\
&\qquad\cdot\kappa_{j,T}\left(N_j(t)\right)\cdot\frac{\mathbb{P}\left(r_i(t)\le \mu_j+2C_j(t)|\mathcal{H}_t\right)}{\mathbb{P}\left(r_i(t)\le U_j(t)|\mathcal{H}_t\right)}\Bigg|\mathcal{H}_t\Bigg]\Bigg]\\
&\quad+\mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{(i,j)\in\mathcal{A}} \mathbb{1}\left\{a_t=(i,j),\mathcal{E}_t\right\}\cdot\kappa_{i,T}\left(N_i(t)\right)\right]
\end{aligned}
$$

where (31) follows from the tower rule and the fact that the under event $a_t = (i,j)$, the event $j\in o(t)$ can happen if and only if $r_i(t)\le U_j(t)$. To proceed, we use the fact that $\mathbb{P}\left(r_i(t)\le \mu_j+2C_j(t)|\mathcal{H}_t\right)\le\mathbb{P}\left(r_i(t)\le U_j(t)|\mathcal{H}_t\right)$. Then,

$$
\begin{aligned}
&R_s(T)\\
&\le \mathbb{E}\Bigg[\sum_{t=K+1}^{T}\mathbb{E}\Bigg[\sum_{(i,j)\in\mathcal{A}_p} \mathbb{1}\left\{a_t=(i,j),j\in o(t),\mathcal{E}_t\right\}\\
&\qquad\cdot\kappa_{j,T}\left(N_j(t)\right)\Bigg|\mathcal{H}_t\Bigg]\Bigg]\\
&\quad+\mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{(i,j)\in\mathcal{A}} \mathbb{1}\left\{a_t=(i,j),\mathcal{E}_t\right\}\cdot\kappa_{i,T}\left(N_i(t)\right)\right]\\
&= \mathbb{E}\Bigg[\sum_{t=K+1}^{T}\sum_{(i,j)\in\mathcal{A}_p} \mathbb{1}\left\{a_t=(i,j),j\in o(t),\mathcal{E}_t\right\}\\
&\qquad\cdot\kappa_{j,T}\left(N_j(t)\right)\Bigg]\\
&\quad+\mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{(i,j)\in\mathcal{A}} \mathbb{1}\left\{a_t=(i,j),\mathcal{E}_t\right\}\cdot\kappa_{i,T}\left(N_i(t)\right)\right]\\
&= \mathbb{E}\Bigg[\sum_{t=K+1}^{T}\sum_{(i,j)\in\mathcal{A}} \mathbb{1}\left\{a_t=(i,j),\mathcal{E}_t\right\}\cdot\left[\mathbb{1}\left\{i\in o(t)\right\}\right.\\
&\qquad\left.\cdot\kappa_{i,T}\left(N_i(t)\right)+\mathbb{1}\left\{j\in o(t)\right\}\cdot\kappa_{j,T}\left(N_j(t)\right)\right]\Bigg]\\
&= \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{i=1}^{K}\mathbb{1}\left\{i\in o(t),\mathcal{E}_t\right\}\cdot\kappa_{i,T}\left(N_i(t)\right)\right]\\
&= \mathbb{E}\left[\sum_{i=1}^{K}\sum_{s=0}^{N_i(T-1)}\kappa_{i,T}(s)\right] \qquad (32)\\
&\le \sum_{i=1}^{K}\sum_{s=0}^{L_{i,T}}\kappa_{i,T}(s)\\
&= \sum_{i=1}^{K}\sum_{s=1}^{L_{i,T}}4\sqrt{\frac{3\log T}{s}}\\
&\le 4\sqrt{3\log T}\sum_{i=1}^{K}\sum_{s=1}^{L_{i,T}}\sqrt{\frac{1}{s}}\\
&\le 8\sqrt{3\log T}\sum_{i=1}^{K}\sqrt{L_{i,T}}\\
&\le \sum_{i=1}^{K}\frac{192\log T}{\Delta_{\min,i}} \qquad (33)
\end{aligned}
$$

where Eq. (32) follows from the fact that $N_i(t)$ increases by 1 when $i\in o(t)$. It can be seen from Lemma 19 that,

$$\sum_{t=K+1}^{T}\mathbb{P}(\mathcal{E}_t^c)\le\frac{2\pi^2K^2}{3}.$$

Combining this with (33), it can be concluded that:

$$
\begin{aligned}
R_T &= R_s(T)+R_{\text{ref}}(T)+\sum_{t=K+1}^{T}\mathbb{P}(\mathcal{E}_t^c)+K\\
&\le\sum_{i=1}^{K}\frac{192\log T}{\Delta_{\min,i}}+R_{\text{ref}}(T)+\frac{2\pi^2K^2}{3}+K
\end{aligned}
$$

Also, using the fact that $R_{\text{ref}}(T)\le\frac{12\log T}{\gamma_i}$ from Lemma 21, it can also be seen that

$$
\begin{aligned}
R_T &= R_s(T)+R_{\text{ref}}(T)+\sum_{t=K+1}^{T}\mathbb{P}(\mathcal{E}_t^c)+K\\
&\le\sum_{i=1}^{K}\frac{192\log T}{\Delta_{\min,i}}+\sum_{i=1}^{K}\frac{96\log T}{\gamma_i}+\frac{2\pi^2K^2}{3}+K
\end{aligned}
$$

Appendix E
Upper Bound on Reference Point Regret

Lemma 20: Given $\mathcal{H}_t$, under the event that the confidence intervals hold in round t, the upper bound on reference point regret if action $(i,j),\forall i\in[K]$ is chosen at round t is

$$d_{(i,j)}(t)\le 4C_j(t)\mathbb{1}\left\{U_j(t)\ge r_i(t)\ge\mu_j, j\in o(t)\right\}.$$

Proof: To upper bound $d_{(i,j)}(t)$, notice that when $r_i(t)$ is not between the values of $U_j(t)$ and μ_j, $d_{(i,j)}(t)=0$ will hold since the decision will not be incorrect in these instances. Hence, the decision to pull the probe arm or the backup arm can be incorrect only when $r_i(t)$ is between the values of $U_j(t)$ and μ_j, and this can be analyzed in two different cases. Assuming the observed reward from the probe is $r_i(t)$, the first case is when $U_j(t)\ge r_i(t)\ge\mu_j$. Then, the UCBP algorithm will decide to pull the backup arm j, and hence $j\in o(t)$, and UCBP will get expected reward μ_j even though the optimal decision is to pull arm i and get reward $r_i(t)$. The gap in reward compared to the optimal decision is $d_{(i,j)}(t)=r_i(t)-\mu_j\le U_j(t)-\mu_j\le 4C_j(t)$ in this case. The second case is when $\mu_j\ge r_i(t)\ge U_j(t)$, but this cannot happen when the confidence bounds hold. □

Lemma 21: The cumulative reference point regret $R_{\text{ref}}(T)$ given in (4) is upper bounded as

$$R_{\text{ref}}(T)\le 8\sqrt{3KT\log T}\ .$$

Proof: To derive an upper bound on the reference point regret, it can be seen from Lemma 20 that $d_{(i,j)}(t) \le 4C_j(t)$. We define $N_a(T) := \sum_{t=K+1}^{T} \mathbb{1}\left\{a_t = a\right\}$ as the total number of times action a is taken until round T; and $B_j(T) := \sum_{i=1, i \neq j}^{K} N_{(i,j)}(T)$ as the total number of times action $(\cdot, j)$ is taken until round T. We also let $\mathcal{A}_{(\cdot,j)}$ denote the set of probe actions with backup arm j. Let $\mathcal{T}_j = \{K+1 \le t \le T : j \in o(t)\}$. Let $\mathcal{T}_{\mathcal{A},j} = \{K+1 \le t \le T : a_t \in \mathcal{A}_{(\cdot,j)}, j \in o(t)\}$.

Denote the ith element of each set $\mathcal{S}$ above by $\mathcal{S}(i)$. Note that

$$\begin{aligned} &N_j(\mathcal{T}_j(1)) = 1, N_j(\mathcal{T}_j(2)) = 2, \ldots, \\ &\qquad N_j(\mathcal{T}_j(k)) = k, \ldots, N_j(\mathcal{T}_j(|\mathcal{T}_j|)) = |\mathcal{T}_j| \ , \end{aligned}$$

and

$$\begin{aligned} &N_j(\mathcal{T}_{\mathcal{A},j}(1)) \ge 1, N_j(\mathcal{T}_{\mathcal{A},j}(2)) \ge 2, \ldots, \\ &\qquad N_{\mathcal{A},j}(\mathcal{T}_j(k)) \ge k, \ldots, N_j(\mathcal{T}_{\mathcal{A},j}(|\mathcal{T}_j|)) \ge |\mathcal{T}_{\mathcal{A},j}| \ . \end{aligned}$$

Using the above display, we obtain

$$\sum_{t \in \mathcal{T}_{\mathcal{A},j}} \sqrt{\frac{1}{N_j(t)}} \le \sum_{x=1}^{|\mathcal{T}_{\mathcal{A},j}|} \sqrt{\frac{1}{x}} \le 2\sqrt{|\mathcal{T}_{\mathcal{A},j}|} \ . \tag{34}$$

which will be used in the rest of the proof. With these definitions and properties, $R_{\text{ref}}(T)$ can be upper bounded as follows

$$\begin{aligned} &R_{\text{ref}}(T) \\ &= \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{a \in \mathcal{A}} \mathbb{1}\left\{a_t = a\right\} \cdot d_a(t) \middle| \mathcal{E}(T)\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{K} \sum_{a \in \mathcal{A}_{(\cdot,j)}} \sum_{t=K+1}^{T} \mathbb{1}\left\{a_t = a\right\} \cdot d_a(t) \middle| \mathcal{E}(T)\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{K} \sum_{a \in \mathcal{A}_{(\cdot,j)}} \sum_{t=K+1}^{T} \left(\mathbb{1}\left\{a_t = a, j \in o(t)\right\} \cdot d_a(t)\right.\right. \\ &\qquad\qquad \left.\left. + \mathbb{1}\left\{a_t = a, j \notin o(t)\right\} \cdot d_a(t)\right) \middle| \mathcal{E}(T)\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{K} \sum_{t=K+1}^{T} \left(\mathbb{1}\left\{a_t \in \mathcal{A}_{(\cdot,j)}, j \in o(t)\right\} \cdot d_{a_t}(t)\right.\right. \\ &\qquad\qquad \left.\left. + \mathbb{1}\left\{a_t \in \mathcal{A}_{(\cdot,j)}, j \notin o(t)\right\} \cdot d_{a_t}(t)\right) \middle| \mathcal{E}(T)\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{K} \sum_{t=K+1}^{T} \mathbb{1}\left\{a_t \in \mathcal{A}_{(\cdot,j)}, j \in o(t)\right\}\right. \\ &\qquad\qquad \left. \cdot\, d_{a_t}(t) \middle| \mathcal{E}(T)\right] \qquad (35) \\ &\le \mathbb{E}\left[\sum_{j=1}^{K} \sum_{t=K+1}^{T} \mathbb{1}\left\{a_t \in \mathcal{A}_{(\cdot,j)}, j \in o(t)\right\}\right. \\ &\qquad\qquad \left. \cdot\, 4C_j(t) \middle| \mathcal{E}(T)\right] \\ &= 4\sqrt{3\log T} \cdot \mathbb{E}\left[\sum_{j=1}^{K} \sum_{t=K+1}^{T}\right. \\ &\qquad \left. \mathbb{1}\left\{a_t \in \mathcal{A}_{(\cdot,j)}, j \in o(t)\right\} \cdot \sqrt{\frac{1}{N_j(t)}} \middle| \mathcal{E}(T)\right] \\ &\le 4\sqrt{3\log T} \cdot \mathbb{E}\left[\sum_{j=1}^{K} \sum_{t \in \mathcal{T}_{\mathcal{A},j}} \sqrt{\frac{1}{N_j(t)}} \middle| \mathcal{E}(T)\right] \\ &\le 8\sqrt{3\log T} \cdot \mathbb{E}\left[\sum_{j=1}^{K} \sqrt{|\mathcal{T}_{\mathcal{A},j}|}\right] \qquad (36) \\ &\le 8\sqrt{3\log T} \cdot \mathbb{E}\left[\sqrt{K \sum_{j=1}^{K} |\mathcal{T}_{\mathcal{A},j}|}\right] \\ &\le 8\sqrt{3KT\log T} \ , \end{aligned}$$

where (35) follows from the fact that when $a_t \in \mathcal{A}_{(\cdot,j)}$ and $j \notin o(t)$, $d_{a_t}(t) = 0$, (36) follows from (34), and the last inequality follows from $\sum_{j=1}^{K} |\mathcal{T}_{\mathcal{A},j}| \le T$ which holds since in each round only one action is chosen. □

Lemma 22: If the distributions Γ_i for each $i \in [K]$ are defined over a *discrete* support $\mathcal{D}$ in $[0,1]$, the cumulative reference point regret until round T can be upper bounded as:

$$R_{\text{ref}}(T) \le \sum_{i=1}^{K} \frac{96 \log T}{\gamma_i} \ ,$$

where we use $d_l \in \mathcal{D}$, $1 \le l \le |\mathcal{D}|$ 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 $\gamma_i := \min_{1 \le l \le |\mathcal{D}|-1} |d_l - d_{l+1}|$ if $\mu_i \in \mathcal{D}$.

Proof: We proceed bounding $R_{\text{ref}}(T)$ from (36) in Lemma 21. Note that r_i can only take discrete values under the discrete distribution, i.e. $r_i \in d_l$, $1 \le l \le |\mathcal{D}|$. Hence, regret can only be incurred when $\{\exists l : U_j(t) \ge d_l \ge \mu_j,\ 1 \le l \le |\mathcal{D}|\}$, since $d_{(i,j)}(t) = 0$ otherwise. From this, it can be seen that $d_{(i,j)}(t) = 0$ when $4C_j(t) < \gamma_j$ since it cannot be the case that $\mu_i \le d_l \le U_i(t)$ for $1 \le l \le |\mathcal{D}|$ when $4C_j(t) < \gamma_j$. Hence, for action (i,j), regret can only be incurred for the rounds where

$$4C_j(t) = 4\sqrt{\frac{3\log t}{N_j(t)}} \ge \gamma_j$$

happens. Rearranging the terms,

$$N_j(t) \le \frac{48 \log t}{\gamma_j^2} \le \frac{48 \log T}{\gamma_j^2}$$

which also means

$$|\mathcal{T}_{\mathcal{A},j}| \le \frac{48 \log T}{\gamma_j^2}$$

Using this, $R_{\text{ref}}(T)$ can be upper bounded as

$$R_{\text{ref}}(T) \le \sum_{j=1}^{K} \frac{96 \log T}{\gamma_j} \ .$$

□

Appendix F
Proof of Theorem 6

In the standard K-armed bandit problem, the reward distributions of arms are given by $\Gamma_i, \forall i \in [K]$. However, in our multi-armed bandit setting with probes, the agent chooses actions that are composed of one or more arms. To characterize the distributions of these actions, we define

$\Gamma_{(i,j)}$ as the distribution function of action (i, j), and Γ^* as the distribution function of the optimal action a^*.

We denote the distribution function of action $(i, \emptyset)$ as $\Gamma_{(i,\emptyset)}$, it can be seen that its distribution is the same as the distribution function of arm i, i.e. $\Gamma_{(i,\emptyset)} = \Gamma_i$. We also use $D_{KL}(\cdot||\cdot)$ to denote the Kullback-Leibler divergence function. From Lemma 10, we know that the following holds for the standard multi-armed bandit problem:

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[N_i(T)\right]}{\log T} \geq \frac{1}{D_{KL}(\Gamma_i, \Gamma^*)}$$

To expand this result into our problem setting of multi-armed bandits with probes, we note the dependency between different actions. First, it can be seen that taking action $a = (i, j)$ yields in a sample of arm i, and if the backup arm is pulled, it also yields in a sample of arm j. Therefore, letting $\mathcal{A}_i = \{(i,j) : j \in ([K] \cup \{\emptyset\}) \setminus \{i\}\} \cup \{(j,i) : j \in [K] \setminus \{i\}\}$, it can be seen that taking an action $a \in \mathcal{A}_i$ may possibly yield samples of arm i (it may not yield in a sample when arm i is the backup arm and the backup arm is not pulled). We let $s_i(t)$ denote the total number of samples obtained for arm i up to round t when the reward of arm i is observed through taking an action $a \in \mathcal{A}_i$. Further, also note that one reward sample of action (i, j) can be produced from one reward sample of arm i and one sample from arm j (these samples need not be from the same time instant as we assume the stochasticity of the reward samples across time). Let $s_{(i,j)}(t)$ denote the total number of samples obtained on action $a = (i, j)$ when all the information from samples of all actions up to round t are used to produce samples of other actions, i.e. when samples of arms i and j are used to obtain the maximum possible number of samples of action $a = (i, j)$, it can be seen that $s_{(i,j)}(t) = \min(s_i(t), s_j(t))$.

Now that we have seen that $s_{(i,j)}(t)$ captures the total amount of samples obtained from action $a = (i, j)$ (by also utilizing the information obtained for action $a = (i, j)$ when an action $a' \in \mathcal{A}_i \cup \mathcal{A}_j$ is taken), Lemma 10 can be used to lower bound the total number of samples (sampled or constructed from other samples) of an action a as:

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[s_{(i,j)}(T)\right]}{\log T} \geq \frac{1}{D_{KL}(\Gamma_{(i,j)}, \Gamma^*)} \tag{37}$$

Combining (37) with the fact that $s_{(i,j)}(t) = \min(s_i(t), s_j(t))$, we have that

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[s_i(T)\right]}{\log T} \geq \frac{1}{D_{KL}(\Gamma_{(i,j)}, \Gamma^*)}$$

Deriving similar inequalities for all actions that involve arm i, which are $(i, \emptyset)$, and for some $j \neq i$, (i, j) and (j, i), and excluding the optimal action a^*, we have

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[s_i(T)\right]}{\log T} \geq \left[\min_{a\in\mathcal{A}_i, a\neq a^*} \{D_{KL}(\Gamma_a||\Gamma^*)\}\right]^{-1} \tag{38}$$

where $\mathcal{A}_i = \{(i,j) : j \in ([K] \cup \{\emptyset\}) \setminus \{i\}\} \cup \{(j,i) : j \in [K] \setminus \{i\}\}$. It can be seen that $s_i(t)$ can be upper bounded by the following:

$$s_i(t) \leq \sum_{\substack{j\in[K]\cup\{\emptyset\} \\ (i,j)\neq a^*}} N_{(i,j)}(t) + \sum_{\substack{j=1 \\ j\neq i,\ (j,i)\neq a^*}}^{K} N_{(j,i)}(t) \tag{39}$$

since in the best case, when an action (i, j) is taken, the rewards of both arm i and arm j can be observed. Combining (38) and (39), we have

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[\sum_{a\in\mathcal{A}_i, a\neq a^*} N_a(t)\right]}{\log T} \geq \left[\min_{a\in\mathcal{A}_i, a\neq a^*} \{D_{KL}(\Gamma_a||\Gamma^*)\}\right]^{-1} \tag{40}$$

Denoting $\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} = b_a$, (40) can be rewritten as:

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

Using the number of samples of the suboptimal actions, the expected cumulative regret can be given as

$$R_T \geq \sum_{a\in\mathcal{A}\setminus\{a^*\}} \mathbb{E}\left[N_a(T)\right]\Delta_a$$

$$\liminf_{T\to\infty} \frac{R_T}{\log T} \geq \liminf_{T\to\infty} \frac{\sum_{a\in\mathcal{A}\setminus\{a^*\}} \mathbb{E}\left[N_a(T)\right]\Delta_a}{\log T}$$

$$\liminf_{T\to\infty} \frac{R_T}{\log T} \geq \sum_{a\in\mathcal{A}\setminus\{a^*\}} b_a\Delta_a$$

Therefore, we can conclude that for the multi-armed bandit setting with costly probes where there is a unique optimal action, the expected cumulative regret for any *uniformly good* algorithm, as defined in [2], is lower bounded as

$$\liminf_{T\to\infty} \frac{R_T}{\log T} \geq C(\Gamma),$$

where $C(\Gamma)$ is the minimal value of the following linear optimization problem:

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

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

$\Gamma_{(i,\emptyset)} = \Gamma_i$, $\Gamma_{(i,j)} = \max(r_i, \mu_j) - c$ is the distribution function of action (i, j) for $i \neq j$, Γ^* is the distribution function of the optimal action, and $D_{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence.

Also note that $C(\Gamma)$ is $\Omega(K)$. This can be seen by summing all the constraint equations:

$$\sum_{i=1}^{K}\left(\sum_{a\in\mathcal{A}_i, a\neq a^*} b_a\right) \geq \sum_{i=1}^{K}\left[\min_{a\in\mathcal{A}_i, a\neq a^*} \{D_{KL}(\Gamma_a||\Gamma^*)\}\right]^{-1} \tag{41}$$

TABLE IV
NOTATIONS FOR THE UCB-NAIVE-PROBE ALGORITHM

$\mathcal{A}_N$	Action set
$a = (i, j, d_l)$	Super arm of selecting i as probe and j as backup arm and using d_l as the reference
$a = (i, \emptyset, \emptyset)$	Super arm of pulling arm i
$N_a(t)$	Number of times super arm a is sampled until round t
$U_a(t)$	UCB index of action a at round t
u^*	The optimal action
ν^*	Mean reward of the optimal action

We have that

$$\sum_{i=1}^{K}\left(\sum_{a\in\mathcal{A}_i,a\neq a^*} b_a\right) = \sum_{i=1}^{K}\left(\sum_{a\in\mathcal{A}_i,a\neq a^*}\liminf_{T\to\infty}\frac{\mathbb{E}\left[N_a(T)\right]}{\log T}\right)$$
$$\le 2\sum_{a\in\mathcal{A}\setminus\{a^*\}}\liminf_{T\to\infty}\frac{\mathbb{E}\left[N_a(t)\right]}{\log T}$$

We define $D_{KL}^i = \min_{a\in\mathcal{A}_i,a\neq a^*}\{D_{KL}(\Gamma_a||\Gamma^*)\}$. Then, (41) can be rewritten as:

$$\sum_{a\in\mathcal{A}\setminus\{a^*\}}\liminf_{T\to\infty}\frac{\mathbb{E}\left[N_a(t)\right]}{\log T} \ge \frac{1}{2}\sum_{i=1}^{K} D_{KL}^i$$

From this, it can be concluded that the lower bound on regret of UCBP is $\Omega(K\log T)$.

APPENDIX G
DERIVATION OF THE EXPECTED REGRET UPPER BOUND OF THE UCB-NAIVE-PROBE ALGORITHM

We provide the regret analysis of the UCB-naive-probe algorithm in this section. The table for the notations used in this section is provided in Table IV. Note that actions for this algorithm are defined over 3-tuples of the form (i, j, d_l) and $(i, \emptyset, \emptyset)$. The action $a = (i, j, d_l)$ denotes that the probe arm is arm i, the backup arm is arm j, and the reference point is d_l. While definitions of variables are the extensions of the variables defined for the 2-tuple actions in the UCBP algorithm to the setting with 3-tuple actions, we briefly define them for this setting for completeness. For pulling actions, ν is defined as $\nu_{(i,\emptyset,\emptyset)} = \mu_i,\ i \in [K]$, and for probing actions, $\nu_{(i,j,d_l)} = -c + \mathbb{E}\left[r_i \cdot \mathbb{1}\{r_i \ge d_l\} + r_j \cdot \mathbb{1}\{r_i < d_l\}\right],\ i, j \in [K],\ i \neq j,\ d_l \in \mathcal{D}, l \in [\mathcal{D}] \setminus \{1\}$ (to exclude the smallest possible discrete value). $\hat{\nu}_a(t)$ is the empirical estimation of ν_a. $N_a(t)$ is the number of times action a is chosen up to round t. The confidence interval can be defined as:

$$C_{(i,j,d_l)}(t) = \sqrt{\frac{2\log t}{N_{(i,j,d_l)}(t)}}$$

Using this, the UCB indices for super arms are defined as $U_a(t) = \hat{\nu}_a(t) + C_a(t)$. The optimal action is denoted as u^*. The gaps of actions are defined as $\Delta_{(i,j,d_l)} = \nu^* - \mathbb{E}\left[r_i \cdot \mathbb{1}\{r_i \ge d_l\} + \mu_j \cdot \mathbb{1}\{r_i < d_l\}\right] + c$, and $\Delta_{(i,\emptyset,\emptyset)} = \Delta_i$.

We first start with the gap-dependent upper bound as the gap-independent bound will be derived from the gap-dependent bound.

A. Gap-Dependent Regret Upper Bound For UCB-Naive-Probe

Regret is incurred whenever a suboptimal action is taken. Therefore, we upper bound the expected number of times each suboptimal super arm is pulled by the UCB-naive-probe algorithm. Similar to the regret analysis of UCBP, first, the regret is decomposed into components reflecting the regret of each suboptimal action. We condition the occurrence of suboptimal actions on the event that the confidence intervals hold to help upper bound the number of times each suboptimal action is chosen, and then we sum the regret from each to obtain the expected regret of the UCB-naive-probe algorithm. The empirical regret of the UCB-naive-probe algorithm can be written as:

$$\hat{R}_U(T) = \sum_{t=|\mathcal{D}|K^2+1}^{T}\sum_{a\in\mathcal{A}}\mathbb{1}\{a_t = a\}\cdot(\nu^* - r_a(t)) + |\mathcal{D}|K^2$$

Expected regret can be obtained by taking the expectation of this expression

$$R_U(T) = \mathbb{E}\left[\hat{R}_U(T)\right] = \mathbb{E}\left[\sum_{t=|\mathcal{D}|K^2+1}^{T}\mathbb{E}\left[\sum_{a\in\mathcal{A}}\mathbb{1}\{(a_t = a\}\right.\right.$$
$$\left.\left.\cdot(\nu^* - r_a(t))\middle|\mathcal{H}_t\right]\right] + |\mathcal{D}|K^2$$

We condition this expression using $\mathcal{E}(T) := \{|\hat{\nu}_a(t) - \nu_a| \le C_a(t),\ \forall a \in \mathcal{A}\}$, the event that all confidence intervals hold in round t. Then the expected regret can be upper bounded as:

$$\begin{aligned}
&R_U(T)\\
&\le \mathbb{E}\left[\sum_{t=|\mathcal{D}|K^2+1}^{T}\mathbb{E}\left[\sum_{a\in\mathcal{A}}\mathbb{1}\{(a_t = a)\}\right.\right.\\
&\qquad\left.\left.\cdot(\nu^* - r_a(t))\middle|\mathcal{H}_t\right]\middle|\mathcal{E}(T)\right]\\
&\qquad\qquad + \mathbb{E}\left[\sum_{t=|\mathcal{D}|K^2+1}^{T}\mathbb{P}(\mathcal{E}_1^c(t))\right]\\
&= \mathbb{E}\left[\sum_{t=|\mathcal{D}|K^2+1}^{T}\sum_{a\in\mathcal{A}\setminus\{u^*\}}\mathbb{1}\{(a_t = a)\}\middle|\mathcal{E}(T)\right]\cdot\Delta_a\\
&\qquad\qquad + \mathbb{E}\left[\sum_{t=|\mathcal{D}|K^2+1}^{T}\mathbb{P}(\mathcal{E}_1^c(t))\right]\\
&= \sum_{a\in\mathcal{A}\setminus\{u^*\}}\mathbb{E}\left[\sum_{t=|\mathcal{D}|K^2+1}^{T}\mathbb{1}\{(a_t = a)\}\middle|\mathcal{E}(T)\right]\cdot\Delta_a\\
&\qquad\qquad + \mathbb{E}\left[\sum_{t=|\mathcal{D}|K^2+1}^{T}\mathbb{P}(\mathcal{E}_1^c(t))\right]
\end{aligned}$$

Defining

$$\mathbb{E}\left[N_a(T)\right] := \mathbb{E}\left[\sum_{t=|\mathcal{D}|K^2+1}^{T}\mathbb{1}\{(a_t = a)\}\middle|\mathcal{E}(T)\right],$$

R_T can be upper bounded as:

$$R_U(T) \leq \sum_{a\in\mathcal{A}\setminus\{u^*\}} \mathbb{E}\left[N_a(T)\right]\cdot\Delta_a + \mathbb{E}\left[\sum_{t=|\mathcal{D}|K^2+1}^{T} \mathbb{P}(\mathcal{E}_1^c(t))\right] \quad (42)$$

To upper bound $\mathbb{E}\left[N_a(T)\right]$, we will show that the suboptimal action $a \neq u^*$ cannot occur at any round $t \leq T$ if the total number of times the super arm a has been sampled (pulled or probed) is $N_a(T) \geq \frac{8\log T}{\Delta_a^2}$. We start by noting that for action a to occur, the upper confidence index of action a needs to be above the upper confidence index of the optimal action u^* at round t. Hence, the arm can only be pulled if

$$\hat{\nu}_{u^*}(t) + C_{u^*}(t) < \hat{\nu}_a(t) + \sqrt{\frac{2\log t}{N_a(t)}}$$

is satisfied. Using the fact that $\nu^* \leq \hat{\nu}_{u^*}(t) + C_{u^*}(t)$, and $\hat{\nu}_a(t) \leq \nu_a + C_a(t)$ under the event $\mathcal{E}(T)$, we have

$$\nu^* < \nu_a + 2\sqrt{\frac{2\log t}{N_a(t)}}$$
$$N_a(t) \leq \frac{8\log t}{\Delta_a^2}$$

This means that action a can only be taken in rounds $t \leq T$ when $N_a(t) < \frac{8\log t}{\Delta_a^2}$ is satisfied. Noticing that this can happen at most $\frac{8\log T}{\Delta_a^2}$ times until round T upper bounds the expected number of times action a is taken, hence

$$\mathbb{E}\left[N_a(t)\right] \leq \frac{8\log T}{\Delta_a^2} \quad (43)$$

We now bound the term $\sum_{t=1}^{T}\mathbb{P}(\mathcal{E}_t^c)$ where $\mathcal{E}_t$ is the event that all confidence bounds hold in round t. Note that from (21) and (20), we have that the probability that the confidence interval for any arm a does not hold is upper bounded by $2t^{-3}$. Using this, through a union bound over all the probabilities of each confidence interval not holding, we have that

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{P}(\mathcal{E}_t^c)\right] \leq \sum_{i=1}^{K}\sum_{t=1}^{T}2t^{-3} + \sum_{l=2}^{|\mathcal{D}|}\sum_{i=1}^{K^2-K}\sum_{t=1}^{T}2t^{-3} \quad (44)$$
$$= 2((|\mathcal{D}|-1)(K^2-K)+K)\sum_{t=1}^{T}t^{-3}$$
$$\leq \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} \quad (45)$$

where the first summation term in the right side of (44) is for the actions of the form $(i,\emptyset,\emptyset)$, and the second term is for the actions of the form (i,j,d_l). In (45), we again use the fact that $\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}$.

Combining (43) and (45), it can be concluded that

$$R_U(T) \leq \sum_{a\in\mathcal{A}_N\setminus\{u^*\}} \frac{8\log T}{\Delta_a} + \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} + |\mathcal{D}|K^2$$
$$= O(|\mathcal{D}|K^2\log T) + O(1)$$

□

B. Gap-Independent Regret Upper Bound For UCB-Naive-Probe

The gap-independent upper bound can be obtained from the gap dependent upper bound by dividing the action set into two as follows

$$\mathcal{A}_{N,1} := \left\{a \in \mathcal{A}\setminus\{u^*\} : \Delta_a \geq \sqrt{\frac{8|\mathcal{D}|K^2\log T}{T}}\right\}$$
$$\mathcal{A}_{N,2} := \left\{a \in \mathcal{A}\setminus\{u^*\} : \Delta_a < \sqrt{\frac{8|\mathcal{D}|K^2\log T}{T}}\right\}$$

Using (42), we have

$$R_U(T) \leq \sum_{a\in\mathcal{A}_N\setminus\{u^*\}} \mathbb{E}\left[N_a(T)\right]\cdot\Delta_a + \mathbb{E}\left[\sum_{t=|\mathcal{D}|K^2+1}^{T}\mathbb{P}(\mathcal{E}_1^c(t))\right]$$
$$\leq \sum_{a\in\mathcal{A}_{N,1}\setminus\{u^*\}} \mathbb{E}\left[N_a(T)\right]\cdot\Delta_a + \sum_{a\in\mathcal{A}_{N,2}\setminus\{u^*\}} \mathbb{E}\left[N_a(T)\right]\cdot\Delta_a + \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} + |\mathcal{D}|K^2$$

For $a \in \mathcal{A}_{N,1}$, use $\mathbb{E}\left[N_a(t)\right] \leq \frac{8\log T}{\Delta_a^2}$, and for $a \in \mathcal{A}_{N,2}$, use $\Delta_a \leq \sqrt{\frac{8|\mathcal{D}|K^2\log T}{T}}$. Then

$$R_U(T) \leq \sum_{a\in\mathcal{A}_{N,2}\setminus\{u^*\}} \mathbb{E}\left[N_a(t)\right]\cdot\sqrt{\frac{8|\mathcal{D}|K^2\log T}{T}} + \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} + \sum_{a\in\mathcal{A}_{N,1}\setminus\{u^*\}} \frac{8\log T}{\Delta_a^2}\cdot\Delta_a + |\mathcal{D}|K^2$$

Using $\sum_{a\in\mathcal{A}_{N,2}\setminus\{u^*\}}\mathbb{E}\left[N_a(t)\right] \leq T$, and the fact that $|\mathcal{A}_{N,1}| \leq |\mathcal{D}|K^2$ we have

$$R_U(T) \leq \sum_{a\in\mathcal{A}_{N,1}\setminus\{u^*\}} \frac{8\log T}{\Delta_a} + T\sqrt{\frac{8|\mathcal{D}|K^2\log T}{T}} + \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} + |\mathcal{D}|K^2$$
$$\leq \sum_{a\in\mathcal{A}_{N,1}\setminus\{u^*\}} \sqrt{\frac{8T\log T}{|\mathcal{D}|K^2}} + T\sqrt{\frac{8|\mathcal{D}|K^2\log T}{T}} + \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} + |\mathcal{D}|K^2$$
$$\leq |\mathcal{D}|K^2\cdot\sqrt{\frac{8T\log T}{|\mathcal{D}|K^2}} + \sqrt{8|\mathcal{D}|K^2T\log T} + \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} + |\mathcal{D}|K^2$$

$$\leq 4\sqrt{2|\mathcal{D}|K^2T\log T} + \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} + |\mathcal{D}|K^2$$
$$= O(\sqrt{|\mathcal{D}|K^2T\log T}) + O(1)$$

□

References

[1] E. C. Elumar, C. Tekin, and O. Yağan, "Multi-armed bandits with probing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2024, pp. 2080–2085.

[2] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2, pp. 235–256, 2002.

[4] R. Agrawal, "Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem," *Adv. Appl. Probab.*, vol. 27, no. 4, pp. 1054–1078, 1995.

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

[6] D. Chakrabarti, R. Kumar, F. Radlinski, and E. Upfal, "Mortal multi-armed bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 21, 2008, pp. 273–280.

[7] Y. Varatharajah and B. Berry, "A contextual-bandit-based approach for informed decision-making in clinical trials," *Life*, vol. 12, no. 8, p. 1277, Aug. 2022.

[8] W. H. Press, "Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 52, pp. 22387–22392, Dec. 2009.

[9] N. Silva, H. Werneck, T. Silva, A. C. M. Pereira, and L. Rocha, "Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions," *Expert Syst. Appl.*, vol. 197, Jul. 2022, Art. no. 116669.

[10] J. Mary, R. Gaudel, and P. Preux, "Bandits and recommender systems," in *Proc. Int. Workshop Mach. Learn. Optim. Big Data*, Sicily, Italy. Cham, Switzerland: Springer, 2015, pp. 325–336.

[11] T. Lu, D. Pál, and M. Pál, "Contextual multi-armed bandits," in *Proc. 30th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 485–492.

[12] S. Mannor and O. Shamir, "From bandits to experts: On the value of side-observations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 684–692.

[13] J. Langford and T. Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 20, 2007, pp. 817–824.

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

[15] A. Gupta and V. Nagarajan, "A stochastic probing problem with applications," in *Proc. 16th Int. Conf. Integer Program. Combinat. Optim. (IPCO)*, Valparaíso, Chile, 2013, pp. 205–216.

[16] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz, "Minimizing regret with label efficient prediction," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 2152–2162, Jun. 2005.

[17] Y. Efroni, N. Merlis, A. Saha, and S. Mannor, "Confidence-budget matching for sequential budgeted learning," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2937–2947.

[18] Y. Seldin, P. Bartlett, K. Crammer, and Y. Abbasi-Yadkori, "Prediction with limited advice and multiarmed bandits with paid observations," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 280–287.

[19] S. Gollapudi and D. Panigrahi, "Online algorithms for rent-or-buy with expert advice," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2319–2327.

[20] E. Bamas, A. Maggiori, and O. Svensson, "The primal-dual method for learning augmented algorithms," in *Proc. NIPS*, 2020, pp. 20083–20094.

[21] K. Anand, R. Ge, A. Kumar, and D. Panigrahi, "Online algorithms with multiple predictions," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 582–598.

[22] S. Wang, J. Li, and S. Wang, "Online algorithms for multi-shop ski rental with machine learned advice," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 8150–8160.

[23] A. Rakhlin and K. Sridharan, "Online learning with predictable sequences," in *Proc. Conf. Learn. Theory*, 2013, pp. 993–1019.

[24] T. Lykouris and S. Vassilvitskii, "Competitive caching with machine learned advice," *J. ACM*, vol. 68, no. 4, pp. 1–25, Aug. 2021.

[25] A. Klein, S. Falkner, J. T. Springenberg, and F. Hutter, "Learning curve prediction with Bayesian neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2017.

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

[27] J.-H. Liu, T. Zhou, Z.-K. Zhang, Z. Yang, C. Liu, and W.-M. Li, "Promoting cold-start items in recommender systems," *PLoS ONE*, vol. 9, no. 12, Dec. 2014, Art. no. e113457.

[28] H.-N. Kim, A. El-Saddik, and G.-S. Jo, "Collaborative error-reflected models for cold-start recommender systems," *Decis. Support Syst.*, vol. 51, no. 3, pp. 519–531, Jun. 2011.

[29] J. Zuo, X. Zhang, and C. Joe-Wong, "Observe before play: Multi-armed bandit with pre-observations," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 4, pp. 7023–7030.

[30] A. Bhaskara, S. Gollapudi, S. Im, K. Kollias, and K. Munagala, "Online learning and bandits with queried hints," in *Proc. 14th Innov. Theor. Comput. Sci. Conf. (ITCS)*, Cambridge, MA, USA, Jan. 2023, pp. 16:1–16:24.

[31] S. Kale, "Multiarmed bandits with limited expert advice," in *Proc. Conf. Learn. Theory*, 2014, pp. 107–122.

[32] S. Guha, K. Munagala, and S. Sarkar, "Optimizing transmission rate in wireless channels using adaptive probes," in *Proc. Joint Int. Conf. Meas. Modeling Comput. Syst.*, Jun. 2006, pp. 381–382.

[33] 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 *Proc. 1st Int. Conf. Wireless Internet (WICON)*, 2005, pp. 156–163.

[34] A. Johnsson, B. Melander, and M. Björkman, "Bandwidth measurement in wireless networks," in *Proc. IFIP Annu. Mediterranean Ad Hoc Netw. Workshop*, Île de Porquerolles, France. Cham, Switzerland: Springer, 2005, pp. 89–98.

[35] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *Proc. 30th Int. Conf. Mach. Learn.*, Jun. 2013, pp. 151–159.

[36] N. Cesa-Bianchi and G. Lugosi, "Combinatorial bandits," *J. Comput. Syst. Sci.*, vol. 78, no. 5, pp. 1404–1422, Sep. 2012.

[37] M. Jourdan, M. Mutnỳ, J. Kirschner, and A. Krause, "Efficient pure exploration for combinatorial bandits with semi-bandit feedback," in *Proc. 32nd Int. Conf. Algorithmic Learn. Theory* (Proceedings of Machine Learning Research), vol. 132, V. Feldman, K. Ligett, and S. Sabato, Eds. PMLR, Mar. 2021, pp. 805–849. [Online]. Available: http://proceedings.mlr.press/v132/jourdan21a/jourdan21a.pdf

[38] Q. Liu, W. Xu, S. Wang, and Z. Fang, "Combinatorial bandits with linear constraints: Beyond knapsacks and fairness," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 2997–3010.

[39] F. Li, J. Liu, and B. Ji, "Combinatorial sleeping bandits with fairness constraints," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 3, pp. 1799–1813, Jul. 2020.

[40] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1466–1478, Oct. 2012.

[41] Y. Wang, W. Chen, and M. Vojnović, "Combinatorial bandits for maximum value reward function under max value-index feedback," 2023, *arXiv:2305.16074*.

[42] Q. Wang and W. Chen, "Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1161–1171.

[43] A. Huyuk and C. Tekin, "Analysis of Thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms," in *Proc. The 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 1322–1330.

[44] A. Hüyük and C. Tekin, "Thompson sampling for combinatorial network optimization in unknown environments," *IEEE/ACM Trans. Netw.*, vol. 28, no. 6, pp. 2836–2849, Dec. 2020.

[45] Z. Zhong, W. C. Cheung, and V. Tan, "Best arm identification for cascading bandits in the fixed confidence setting," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11481–11491.

[46] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Combinatorial cascading bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1450–1458.

[47] Z. Zhong, W. C. Chueng, and V. Y. Tan, "Thompson sampling algorithms for cascading bandits," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 9915–9980, 2021.

[48] A. Bhaskara, A. Cutkosky, R. Kumar, and M. Purohit, "Logarithmic regret from sublinear hints," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 28222–28232.

[49] A. Goel, S. Guha, and K. Munagala, "Asking the right questions: Model-driven optimization using probes," in *Proc. 25th ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, Jun. 2006, pp. 203–212.

[50] A. A. Mogyla, "Application of stochastic probing radio signals for the range-velocity ambiguity resolution in Doppler weather radars," *Radioelectronics Commun. Syst.*, vol. 57, no. 12, pp. 542–552, Dec. 2014.

[51] M. L. Weitzman, "Optimal search for the best alternative," *Econometrica*, vol. 47, no. 3, pp. 641–654, 1976. Accessed: Nov. 26, 2024. [Online]. Available: http://www.jstor.org/stable/1910412

[52] S. Chawla, E. Gergatsouli, Y. Teng, C. Tzamos, and R. Zhang, "Pandora's box with correlations: Learning and approximation," in *Proc. IEEE 61st Annu. Symp. Found. Comput. Sci. (FOCS)*, Nov. 2020, pp. 1214–1225.

[53] S. Boodaghians, F. Fusco, P. Lazos, and S. Leonardi, "Pandora's box problem with order constraints," in *Proc. 21st ACM Conf. Econ. Comput.*, Jul. 2020, pp. 439–458.

[54] H. Beyhaghi and R. Kleinberg, "Pandora's problem with nonobligatory inspection," in *Proc. ACM Conf. Econ. Comput.*, Jun. 2019, pp. 131–132.

[55] F. M. Harper and J. A. Konstan, "The movieLens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, pp. 1–19, 2015.

[56] R. Zhang and R. Combes, "On the suboptimality of Thompson sampling in high dimensions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 8345–8354.

[57] Y. Saito, S. Aihara, M. Matsutani, and Y. Narita, "Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation," 2020, *arXiv:2008.07146*.

Eray Can Elumar (Student Member, IEEE) received the B.S. degree in electrical and electronics engineering and the B.S. degree in physics from Boğaziçi University, Istanbul, Türkiye, in 2019. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Carnegie Mellon University, Pittsburgh, PA, USA. His research interests include multi-armed bandits, information theory, machine learning, and optimization. He was a recipient of the David H. Barakat and LaVerne Owen-Barakat College of Engineering Dean's Fellowship.

Cem Tekin (Senior Member, IEEE) received the B.Sc. degree in electrical and electronics engineering from Middle East Technical University, Ankara, Türkiye, in 2008, and the M.S.E. degree in electrical engineering: systems, the M.S. degree in mathematics, and the Ph.D. degree in electrical engineering: systems from the University of Michigan, Ann Arbor, MI, USA, in 2010, 2011, and 2013, respectively. From February 2013 to January 2015, he was a Post-Doctoral Scholar with the University of California, Los Angeles, CA, USA. He is currently an Associate Professor with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara. His research interests include multiarmed bandit problems, reinforcement learning, and cognitive communications. He was a recipient of the numerous awards, including the Fred W. Ellersick Award for the Best Paper in MILCOM 2009, the Distinguished Young Scientist (BAGEP) Award of the Science Academy Association of Turkey in 2019, and the TUBA-GEBIP Award in 2023.

Osman Yağan (Senior Member, IEEE) received the B.S. degree in electrical and electronics engineering from Middle East Technical University, Ankara, Türkiye, in 2007, and the Ph.D. degree in electrical and computer engineering from University of Maryland, College Park, MD, USA, in 2011. In August 2013, he joined as the Faculty Member with the Department of Electrical and Computer Engineering, Carnegie Mellon University, where he is currently a Research Professor. His research interests include modeling, analysis, and performance optimization of computing systems, and uses tools from applied probability, data science, machine learning, and network science. Specific topics include wireless communications, security, random graphs, social and information networks, and cyber-physical systems. He was a recipient of the CIT Dean's Early Career Fellowship, the IBM Faculty Award, and the Best Paper Awards in ICC 2021, IPSN 2022, and ASONAM 2023.