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 $O(\sqrt{KT \log T})$, and gap-dependent regret bound of $O(K^2 \log T)$; and TSP, the Thompson Sampling version of UCBP. In empirical simulations, the UCBP algorithm outperforms the UCB-naive-probe algorithm, 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 [1] 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 [2], [3], and used in applications in many fields, such as online advertising [4], [5], clinical trials [6], [7], and recommendation systems [8], [9].

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

Parts of the material is submitted to the 2024 IEEE International Symposium on Information Theory (ISIT)

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 [10]–[12]. 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 [13], stochastic probing [14], online learning [15], and multi-armed bandits [16], [17]. 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 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 [18]–[20] to characterize improved performance bounds compared to the case without predictions in settings such as when the predictions are perfect [21], when the predictions are adversarial [22], or when there is an upper limit on the error of the predictions [23]. While in this work we 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 [24]. 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 [25]. Our work can be especially useful when these distributions are unknown. One other application is the cold-start problem in recommender systems [26], 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 [27]. 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, and UCB-naive-probe 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.

[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 base arms
$\mathcal{A}$	Action set
$\mathcal{A}_p$	Set of probe actions
$\mathcal{A}_s$	Set of direct pull actions
$\mathcal{A}_{p,i}$	Set of 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 time t
$a = (i, j)$	Action with i as the probe and j as the backup arm
$a = (i, \emptyset)$	Action of pulling arm i
$a(t)$	The action taken at round t
a^*	Optimal action
ν^*	Mean reward of optimal action
μ_i	Mean reward of arm i
ν_a	Mean reward of action a
p^*	Maximum expected reward of probing actions
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
ϵ	The minimum probability of pulling backup arm
γ_i	$\gamma_i := \min_j \lvert d_j - \mu_i \rvert$
$S(t)$	Set of arms whose reward is observed in round t
$U_i(t)$	Upper confidence index of arm i in round t
$U_a(t)$	Upper confidence index of action a t
$P_{i,j}(t)$	Probing upper confidence index $a = (i, j)$
$N_i(t)$	Number of times arm i is sampled (pull or probe)
$N_a(t)$	Number of times action a is taken

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 base arms $[K]$. When pulled, arm $i \in [K]$ generates a random reward from a distribution Γ_i with mean μ_i. 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.

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 base arm (either as probe or backup arm) i are denoted as $\mathcal{A}_{p,i} := \{a \in \mathcal{A}_p : i \in a\}$. 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

TABLE II

EXAMPLE OF EXPECTED ACTION REWARDS UNDER DIFFERENT ARM DISTIBUTIONS. (LEFT) DISTRIBUTIONS OF ARM REWARDS IN A SETTING WITH 3 DIFFERENT ARMS. THE 1/5 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

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

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

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*. Using this reference point strategy, it can be seen that the expected reward of playing action (i,j) is:

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

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

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

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

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

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

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

Note that while it is not explicitly written, all probe actions above in this paper employ the *optimal reference point decision* described above. 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

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

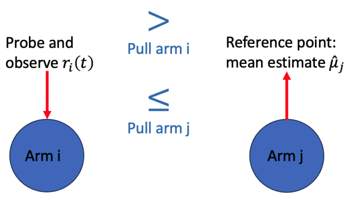

Fig. 1. Illustration of the decision rule for action (i, j) if $\hat{\mu}_j$ is used as the reference point.

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 base 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. We instead use an estimate of μ_1, e.g., the current empirical average $\hat{\mu}_1(t)$, as a *reference* value to compare against r_i, which will lead to incurring a *regret* of up to $|\hat{\mu}_1(t) - \mu_1|\mathbb{P}(r_i \in [\min(\mu_1, \hat{\mu}_1(t)), \max(\mu_1, \hat{\mu}_1(t))])$. This decision of choosing which arm to pull in action (i, j) when empirical estimate of arm j is used as the reference point is illustrated in Fig. 1.

We call the decision to pull arm i or j using $\hat{\mu}_j(t)$ as a reference point is called the *reference point decision*, and the regret it introduces as *reference point regret*. We denote the cumulative regret incurred until round T due to the *reference point error* as $R_{\text{ref}}(T)$.

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 countable, we need to have countably many reference point values; i.e., the UCB-naive-probe algorithm can only be used when the reward distributions of the arms are discrete. 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

TABLE III

COMPARISON OF OUR WORK WITH PRIOR WORK ON BANDITS WITH PROBES

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

$r_i(t) \geq d_l$, 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 $\tilde{O}(|\mathcal{D}|K^2)$.

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 $r(t) = r_t(i_t)$
    else
        Probe arm $i_t$, observe reward $r_t(i_t)$
        if $r_t(i_t) \geq d(t)$ then
            Pull arm $i_t$, get $r(t) = r_t(i_t) - c$
        else
            Pull arm $j_t$, get $r(t) = r_t(j_t) - c$
        end if
    end if
    Update UCB indices and mean estimates
end for
```

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.

III. Related Works

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 [17], 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 [16], there is a limit on the number of queries allowed. In [30], 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 [28], 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 [29], 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 base 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 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.

Probes in Wireless Communications: While there are numerous prior work on probing in wireless communication systems [13], [31]–[33], one notable study related to our work is [25]. 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 [25] from our work is that the reward distributions of the arms are unknown in our setting.

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* [34], [35]. 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 [36], the linearity of the rewards of super arms over the set of arms [37], or additional constraints that can reduce the size of the action set are commonly used [38]. 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 [39], nonlinear [34], 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 [40]. 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 $r_i(t) > \hat{\mu}_j(t)$, 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 max reward function that includes this uncertainty.

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 [41]. 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 [41, 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 [42], 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 [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, 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 [43] 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 we also have partial observability, or probabilistic triggering of the rewards due to the possibility of having a low probability of observing the reward of the backup arm, which is described in more detail in Assumption 1. Different from this work, we cannot use the *triggering probability modulated bounded smoothness* assumption in our work, as observing an arm depends on the choice of the algorithm in our setting, and hence triggering probabilities of arms cannot be expressed as constant values. As a result of this, as will be described in more detail in §IV, we assume that the probability of observing the backup arm is at least ϵ, and our regret upper bounds scale with the $1/\epsilon$ factor.

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 [44], the amount of available feedback at each step is probabilistically estimated to overcome this challenge. In [45], 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 [41, 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 [46], 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.

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 [15], 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 [47], 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.

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 [48], radar systems [49], and Bayesian auctions [14]. 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 [50], different settings of it are widely studied [51]–[53].

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. In UCB [2], at each round t, the arm with the highest UCB index $U_i(t)$ is pulled, i.e.,

$$i(t) = \arg\max_i U_i(t), \quad U_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{2\log t}{N_i(t)}}$$

where $\hat{\mu}_i(t)$ is the empirical mean reward of arm i, $i(t)$ 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 arm 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

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

where $S(t)$ denotes the set of arms whose reward is observed (by either pulling or probing) in round t and $N_i(t)$ denotes the number of times arm i is observed by round t. The UCB indexes for each arm i are computed as

$$U_i(t) = \hat{\mu}_i(t) + C_{(i,\emptyset)}(t)$$

where $C_{(i,\emptyset)}(t) = \sqrt{2\log(t)/N_i(t)}$. Then, the probe UCB indexes are evaluated for probing actions by using $P_{i,j}(t) = \hat{\nu}_{(i,j)}(t) + C_{(i,j)}(t)$, where

$$\hat{\nu}_{(i,j)}(t) = \sum_{\tau=1}^{t-1} \frac{\max(r_i(\tau), \hat{\mu}_j(t))\mathbb{1}\{i \in S(\tau)\}}{N_i(t)} - c, \text{ and}$$

$$C_{(i,j)}(t) = \sqrt{\frac{2\log(t)}{N_j(t)}} + \sqrt{\frac{2\log(t)}{N_i(t)}}.$$

Here $\hat{\nu}_{(i,j)}$ represents the empirical mean reward of action (i,j), and $C_{(i,j)}(t)$ is the exploration bonus associated with action (i,j). The pseudo-code is provided in Algorithm 2. The derivation of the exploration bonuses is in §B. Lastly, the UCB indexes of the actions $U_i(t),\ \forall i \in [K]$; and $P_a(t),\ \forall a \in \mathcal{A} \setminus [K]$ are compared and the one with highest UCB index is chosen. If this action is probing, i.e. $a = (i,j)$, arm i is probed to observe $r_i(t)$, then arm i is pulled if $r_i(t) > \hat{\mu}_j(t)$, and arm j otherwise.

Algorithm 2 UCBP

```
Input: cost of probing $c$, action set $\mathcal{A}$
Initialize: $N_i = 0,\ 1 \le i \le K$
Sample each base arm once
for each round $t$ do
    $i_t^* = \arg\max_i U_i(t)$
    $a_t = (j_t, k_t) = \arg\max_{a \in \mathcal{A}_p} P_a(t)$
    if $U_{i_t^*}(t) > P_{a_t}(t)$ then
        Pull arm $i_t^*$, get $r(t) = r_t(i_t^*)$
    else
        Probe arm $j_t$, observe reward $r_t(j_t)$
        if $r_t(j_t) > \hat{\mu}_{k_t}(t)$ 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 UCB indices for all arms
end for
```

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 first state a mild assumption on the reward distributions of the arms that are required for the theoretical analysis. We refer the readers to the Appendix for detailed proofs of the results presented in this section.

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

Assumption 1 ensures the backup arm is pulled at least ϵ fraction of the time in expectation when action (i,j) is chosen. This assumption is needed in our setting, since if for some arm $j \in [K]$ the gap of actions $(j,\cdot)$ and $(j,\emptyset)$ are much larger than the gap of the actions $(\cdot,j)$; then the algorithm will mostly choose actions of the form $(\cdot,j)$, meaning arm j will only be selected as the backup arm, which might not produce enough samples for arm j without this assumption. This assumption 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 [41]. In this work, the *triggering probability modulated bounded smoothness* assumption is used to remove the dependency of the regret bounds on p^*. 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. However, they also prove in [41, Theorem 3] that for the general combinatorial bandit settings that do not necessarily satisfy this assumption, the regret lower bound scales with the factor $\frac{1}{p^*}$, demonstrating that the $\frac{1}{p^*}$ factor in regret bound cannot be avoided without making additional assumptions. In our setting, this *triggering probability modulated bounded smoothness* assumption cannot be used, as observing the backup arm in a probe action is not an event with a constant probability, but rather a choice of the algorithm that depends on the reward distribution of the probe arm, and on the estimated mean of the backup arm. As a result of this, our regret bounds scale with the $\frac{1}{\epsilon}$ factor.

B. Gap-independent Expected Regret Upper Bound

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

$$R_T \leq \frac{8\sqrt{KT\log T}}{\epsilon} + R_{\mathit{ref}}(T) + \frac{5\pi^2 K}{3} + K$$

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

In Lemma IV.4, we show that $R_{\text{ref}}(T) = O(\sqrt{KT\log T})$, which together with Theorem IV.1 shows that the gap-independent regret of UCBP is $O(\sqrt{KT\log T})$.

Proof. 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. The proof follows some of the steps in the proof of Lemma 1 in [54], and Lemma A.2 in [55].

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, letting $\mathcal{B}_a(t)$ denote the event that the decision to pull the probe or backup arm is correct, i.e. $\mathcal{B}_a(t) = \mathbb{1}\{r_{\hat{a}}(t) = r_a(t)\}$, the empirical regret can be decomposed as

$$\hat{R}_T = \sum_{t=K+1}^{T} \sum_{a\in\mathcal{A}} \left[\mathbb{1}\{a(t) = a, \mathcal{B}_a(t)\} \cdot (\nu^* - \nu_{a(t)}(t)) + \mathbb{1}\{a(t) = a, \mathcal{B}_a^c(t)\} \cdot (\nu^* - \nu_{a(t)}(t) + d_a(t))\right] + K$$

The summation in time starts from $t = K + 1$ due to the UCBP algorithm sampling each arm once in the first K rounds, and this can contribute at most K to regret since the rewards are bounded. Expected regret can be obtained by taking the expectation of this expression

$$R_T = \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}}\left[\mathbb{1}\{a(t)=a,\mathcal{B}_a(t)\}\cdot(\nu^*-\nu_{a(t)}(t))+\mathbb{1}\{a(t)=a,\mathcal{B}_a^c(t)\}\cdot(\nu^*-\nu_{a(t)}(t)+d_a(t))\right]\right]+K$$

$$= \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}}\left[\mathbb{1}\{a(t)=a\}\cdot(\nu^*-\nu_{a(t)}(t))+\mathbb{1}\{a(t)=a,\mathcal{B}_a^c(t)\}\cdot d_a(t)\right]\right]+K$$

Define the following events

$$\mathcal{E}_t := \{|\hat{\mu}_i(t)-\mu_i|\le C_{(i,\emptyset)}(t)\wedge|\hat{\nu}_{(i,j)}(t)-\nu_{(i,j)}|\le C_{(i,j)}(t),\ \forall i,j\in[K],\ i\neq j\}, \text{ and}$$

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

where $\mathcal{E}_t$ is the event that all confidence intervals hold in round t, and $\mathcal{E}(T)$ is the event that all confidence intervals hold for all rounds $K+1\le t\le T$. Regret can be decomposed based on this event $\mathcal{E}(T)$ as:

$$R_T \le \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}}\left[\mathbb{1}\{a(t)=a\}\cdot(\nu^*-\nu_{a(t)}(t))+\mathbb{1}\{a(t)=a\}\cdot d_a(t)\right]\Big|\mathcal{E}(T)\right]+\sum_{t=K+1}^{T}\mathbb{P}(\mathcal{E}_1^c(t))+K$$

Define

$$R_a(T) := \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}}\mathbb{1}\{a(t)=a\}\cdot(\nu^*-\nu_{a(t)}(t))\Big|\mathcal{E}(T)\right] = \mathbb{E}\left[\sum_{t=K+1}^{T}R_t(a)\Big|\mathcal{E}(T)\right]$$

$$R_{\text{ref}}(T) := \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}}\mathbb{1}\{a(t)=a\}\cdot d_a(t)\Big|\mathcal{E}(T)\right]$$

where $R_t(a) := \sum_{a\in\mathcal{A}}\mathbb{1}\{a(t)=a\}\cdot(\nu^*-\nu_{a(t)}(t))$ is the regret of choosing action $a(t)$ at round t (without the reference point error) under the event that the confidence intervals hold. $R_{\text{ref}}(T)$ denotes the cumulative regret incurred from reference point error until round T, and $R_a(T)$ denotes the cumulative regret incurred until round T from suboptimal action choices (without the reference point error). We start by upper bounding $R_t(a)$. For this, the regret in round t can be written in terms of the upper confidence bound as:

$$\nu^*-\nu_{a(t)}(t) = \nu^*-U_{a(t)}(t)+U_{a(t)}(t)-\nu_{a(t)}(t)\le(\nu^*-U_{a^*}(t))+U_{a(t)}(t)-\nu_{a(t)}(t)$$

where the inequality $U_{a(t)}(t)\ge U_{a^*}(t)$ holds since the UCBP algorithm selects the action with the highest UCB index. Under the event $\mathcal{E}(T)$, we have that $\nu^*\le U_{a^*}(t)$. Hence, under $\mathcal{E}(T)$, it holds that

$$\nu^*-\nu_{a(t)}(t)\le U_{a(t)}(t)-\nu_{a(t)}(t)\le U_{a(t)}(t)-L_{a(t)}(t)\le 2C_{a(t)}$$

It can be seen that $C_{a(t)} = \sum_{i\in a(t)}C_{(i,\emptyset)}(t)$ since if $a(t)=(i,j)$, $C_{(i,j)}(t) = C_{(i,\emptyset)}(t)+C_{(j,\emptyset)}(t)$; and if $a(t)=(i,\emptyset)$, $C_{a(t)} = C_{(i,\emptyset)}(t)$.

Define $o(t)\subset a(t)$ as the set of arms whose reward is observed in round t; and $\mathcal{H}_t = (a(1),r(1),o(1),\cdots,a(t-1),r(t-1),o(t-1),a(1))$ 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. Also let $p_i(a(t),t)$ denote the conditional probability of observing the reward of arm

i at round t when the chosen action is $a(t)$ given $\mathcal{H}_t$. Following the analysis in [55], regret can be decomposed in the following way if the upper confidence bounds hold:

$$R_a(T) = \mathbb{E}\left[\sum_{t=K+1}^{T} R_t(a) \middle| \mathcal{E}(T)\right]$$

$$= \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}[R_t(a)|\mathcal{H}_t] \middle| \mathcal{E}(T)\right] \tag{2}$$

$$\leq \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}\left[\sum_{i\in a(t)} 2C_{(i,\emptyset)}(t)|\mathcal{H}_t\right] \middle| \mathcal{E}(T)\right]$$

$$= \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}\left[\sum_{i\in a(t)} 2C_{(i,\emptyset)}(t) \cdot \mathbb{E}\left[\frac{\mathbb{1}\{i \in o(t)\}}{p_i(a(t),t)}|\mathcal{H}_t\right]|\mathcal{H}_t\right] \middle| \mathcal{E}(T)\right] \tag{3}$$

$$\leq \frac{2}{\epsilon} \cdot \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}\left[\sum_{i\in a(t)} C_{(i,\emptyset)}(t)\mathbb{1}\{i \in o(t)\}|\mathcal{H}_t\right] \middle| \mathcal{E}(T)\right] \tag{4}$$

$$\leq \frac{2}{\epsilon} \cdot \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}\left[\sum_{i\in a(t)} \sqrt{\frac{2\log t}{N_i(t)}}\mathbb{1}\{i \in o(t)\}|\mathcal{H}_t\right] \middle| \mathcal{E}(T)\right]$$

$$= \frac{2\sqrt{2\log T}}{\epsilon} \cdot \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}\left[\sum_{i=1}^{K} \sqrt{\frac{1}{N_i(t)}}\mathbb{1}\{i \in o(t)\}|\mathcal{H}_t\right] \middle| \mathcal{E}(T)\right]$$

$$\leq \frac{2\sqrt{2\log T}}{\epsilon} \cdot \mathbb{E}\left[\sum_{i=1}^{K} \sum_{x=1}^{N_i(T)} \sqrt{\frac{1}{x}}\right] \tag{5}$$

$$\leq \frac{4\sqrt{2\log T}}{\epsilon} \cdot \mathbb{E}\left[\sum_{i=1}^{K} \sqrt{N_i(T)}\right]$$

$$\leq \frac{4\sqrt{2\log T}}{\epsilon} \cdot \mathbb{E}\left[\sqrt{K\sum_{i=1}^{K} N_i(T)}\right] \tag{6}$$

$$\leq \frac{8\sqrt{KT\log T}}{\epsilon} \tag{7}$$

Eq. (2) is due to the tower rule. In Eq. (3) we used the fact that given $\mathcal{H}_t$, the probability of $\mathbb{1}\{i \in o(t)\}$ is $p_i(a(t),t)$. In Eq. (4), we used $p_i(a(t),t) \geq \epsilon$; in Eq. (5), we used the fact that the summation of the confidence intervals of a base arm i from round $t = K+1$ to T can be expressed using the total number of times the arm is sampled to remove the dependency on individual rounds where the arm is sampled. Cauchy-Schwarz inequality is used to obtain Eq. (6); and for Eq. (7), the fact that $T \leq \sum_{i=1}^{K} N_i(T) \leq 2T$ is used.

It can be seen from Lemma IV.2 that,

$$\sum_{t=K+1}^{T} \mathbb{P}(\mathcal{E}_1^c(t)) \leq \frac{5\pi^2 K}{3}.$$

Combining this with (7), it can be concluded that:

$$R_T = R_a(T) + R_{\text{ref}}(T) + K^2 \leq \frac{8\sqrt{KT\log T}}{\epsilon} + R_{\text{ref}}(T) + \frac{5\pi^2 K}{3} + K$$

□

Also, using the fact that $R_{\text{ref}}(T) \leq \frac{2\sqrt{2KT\log T}}{\epsilon}$ from Lemma D.2, it can also be seen that

$$R_T = R_a(T) + R_{\text{ref}}(T) + K^2 \leq \frac{(8+2\sqrt{2})\sqrt{KT\log T}}{\epsilon} + \frac{5\pi^2 K}{3} + K$$

Lemma IV.2. The expected number of times the event $\mathcal{E}(T)$ does not happen can be upper bounded as

$$\sum_{t=K+1}^{T} \mathbb{P}[\mathcal{E}_1^c(t)] \leq \frac{5\pi^2 K}{3}$$

Proof. Through a union bound over all the probabilities of each upper confidence bound not holding, we have that

$$\sum_{t=K+1}^{T} \mathbb{P}[\mathcal{E}_1^c(t)] \leq \sum_{i=1}^{K} \sum_{t=K+1}^{T} 2t^{-3} + \sum_{i=1}^{K} \sum_{t=K+1}^{T} 4t^{-3} + \sum_{i=1}^{K} \sum_{t=K+1}^{T} 4t^{-3} \tag{8}$$

$$= 10K \sum_{t=K+1}^{T} t^{-3} \leq \frac{5\pi^2 K}{3} \tag{9}$$

where for $i \in [K]$, the first summation term in the right side of (8) is for the exploration bonus of $\hat{\mu}_i(t)$, the second term is for the exploration bonus of $\hat{\nu}_{(i,1)}(t)$ and the last term is for the exploration bonus of $\hat{\nu}_{(1,i)}(t)$. Note that in (9), we used the fact that $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$ to upper bound the summation. □

C. Gap-dependent Expected Regret Upper Bound

Theorem IV.3 (Gap-dependent Expected Regret Upper Bound)**.** *Under Assumption 1, when UCBP is run on the action set $\mathcal{A}$ and the cost of probing is $c \geq 0$, its expected cumulative regret is upper bounded as*

$$R_T \leq \sum_{i=1}^{K} \frac{16\log T}{\delta_i} + R_{ref}(T) + \frac{5\pi^2 K}{3} + K, \textit{ where}$$

$$\delta_i = \begin{cases} \rho_i & \textit{if } a^* = (i,\emptyset) \\ \frac{2\min(\rho_i, \Delta_{(i,\emptyset)})}{3} & \textit{if } \frac{\rho_i}{2} \leq \Delta_{(i,\emptyset)} \leq \frac{2\rho_i}{\epsilon} \\ \min(\rho_i, \Delta_{(i,\emptyset)}) & \textit{otherwise} \end{cases}$$

and $\rho_i = \min_{a \in \mathcal{A}_{p,i}\setminus\{a^*\}} \left(\frac{\epsilon\Delta_a}{4}\right)$.

Note that the cost of probing c is included in the gap of actions. In Lemma IV.4, we show that the reference point regret is $R_{\text{ref}}(T) = O(K\log T)$ when the reward distributions are *discrete*. Together with Theorem IV.3, this shows that the gap-dependent regret of UCBP is $O(K\log T)$ when the reward distribution is discrete.

Proof Sketch. The proof follows some of the steps in the proof of Theorem 3 in [56]. The key idea is to use the event $\mathcal{G}_t = \{a(t) \in \mathcal{A}_p$, at least one of the base arms in $a(t)$ was observed at most $\frac{32\log t}{\Delta^2_{a(t)}}$ times$\}$ to upper bound the number of times actions $a \in \mathcal{A}_p$ can happen, as it can be seen that the probing action $a(t)$ can only be chosen when $\mathcal{G}_t$ happens. The upper bound on the number of times actions $a \in \mathcal{A}_s$ can happen is obtained from the analysis of the standard UCB algorithm. Combining these upper bounds for all $a \in \mathcal{A}$ while also considering that probe actions provide samples from both base arms with at least ϵ probability, we derive an upper bound on the number of times each base arm is sampled. The result follows by upper bounding the regret when we consider the worst case the samples for the base arms can be obtained (when we assume samples of the base arms are obtained from possible actions with highest gap). The detailed proof is in Appendix C.

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 IV.4. a) Regret due to the reference point error can be upper bounded as:

$$R_{\text{ref}}(T) \leq \frac{2\sqrt{2KT\log T}}{\epsilon}$$

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

Under this assumption, it can be seen that the gap-dependent regret upper bound is $O(K\log T)$. Proof for these results is given in Appendix D.

Theorem IV.5 (Lower Bound on Expected Regret)**.** *For the multi-armed bandit setting with costly probes where the optimal action is unique, the lower bound on the expected cumulative regret for any* uniformly good *algorithm, as defined in [1], is:*

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

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

$$\min_{b_a \geq 0,\ \forall a \in \mathcal{A}\setminus\{a^*\}} \sum_{a\in\mathcal{A}\setminus\{a^*\}} b_a \Delta_a \qquad s.t. \quad \forall i \in [K], \sum_{a\in\mathcal{A}_i, a\neq a^*} b_a \geq \left[\min_{a\in\mathcal{A}_i, a\neq a^*}\{D_{KL}(\Gamma_a||\Gamma^*)\}\right]^{-1}$$

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 E. 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 IV.3, excluding the ϵ term, we can conclude that the gap-dependent upper bound of the UCBP algorithm is order-wise optimal.

Corollary IV.6. 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:

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

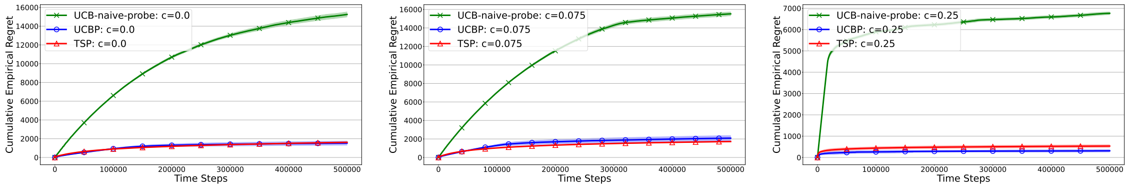

Fig. 2. Plots of the cumulative empirical regret of the UCBP, TSP and UCB-naive-probe algorithms for recommending the best genre in the MovieLens dataset.

Corollary IV.7. If the rewards of the arms are distributed over the discrete support $\mathcal{D}$, the gap-dependent upper bound for $R_U(T)$ is:

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

where $\Delta_{(i,\emptyset,\emptyset)} = \Delta_i$, $\Delta_{(i,j,d_l)} = c + \nu^* - \mathbb{E}\left[r_i \cdot \mathbb{1}\{r_i \geq d_l\} + \mu_j \cdot \mathbb{1}\{r_i < d_l\}\right]$, and u^* is the optimal action in this setting.

The proofs of both Corollary IV.6 and IV.7 are provided in Appendix F.

D. 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.

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. Due to the partial observability of the backup arm, we incur an additional $1/\epsilon$ term in regret, but this is in line with the lower bound given in [41, Theorem 3]. UCB-naive-probe, on the other hand, incurs 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 theoretic 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.

E. 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; and with TSP, the Thompson Sampling

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 base 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
```

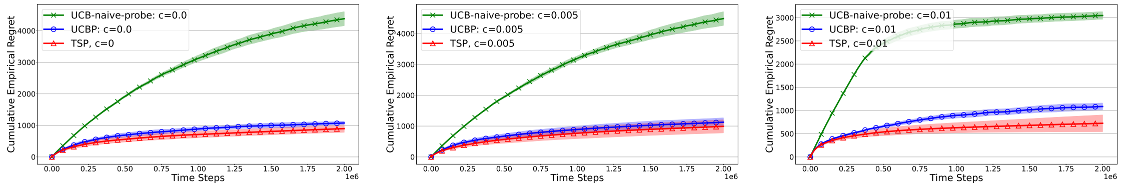

Fig. 3. Plots of the cumulative empirical regret of the UCBP and UCB-naive-probe algorithms for recommending the best item in the Open Bandit dataset.

based version of UCBP. The TSP algorithm operates as follows. First, samples $\theta_i(t)$ for mean arm rewards are generated for base 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 base 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

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

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, and TSP are provided below for the MovieLens and the Open Bandit datasets.

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 [57]. 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 average reward of the best arm is around 0.7925. Our experimental results for this setting are shown in Figure 2, where we plot the cumulative regret averaged over 100 independent trials for $500,000$ rounds when the cost of probing is $c = 0$ (Left), $c = 0.075$ (Middle) and $c = 0.25$ (Right). The shaded area represents error bars with one standard deviation. It can be seen that both algorithms have a logarithmic regret curve, and both the UCBP and the TSP algorithm outperforms the baseline UCB-naive-probe algorithm. 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 TSP performs slightly better than UCBP when $c = 0.25$. While Thompson Sampling based algorithms are known to perform better empirically than UCB based algorithms in general, it was shown in [58] that Thompson Sampling might perform suboptimally in combinatorial bandits or in settings with high dimensions; hence these results are not unexpected.

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 [59]. The dataset includes data from three different campaigns, and we selected the campaign from "Men's" 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 average reward of the best arm is around 0.0087. The goal is to recommend the best item to a cold (new) user. Our experimental results for this setting are shown in Figure 3, where we plot the cumulative regret averaged over 100 independent trials for $2,000,000$ rounds when the cost of probing an arm is $c = 0$ (Left), $c = 0.005$ (Middle) and $c = 0.01$ (Right). The shaded area represents error bars with one standard deviation. Again it can be seen that both 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.

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 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

Algorithm 4 UCBMP

```
Input: cost of probing $c$, action set $\mathcal{A}$
Initialize: $N_i = 0,\ 1 \le i \le K$
Sample each base arm once
for each round $t$ do
    $S(t) = \text{argsort}_i - U_i(t)$
    Evaluate $P_i(t)$ values using Eq. (10)
    $s(t) = \arg\max_i P_i(t)$
    $j \leftarrow 0$
    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
```

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 $P_i(t)$,

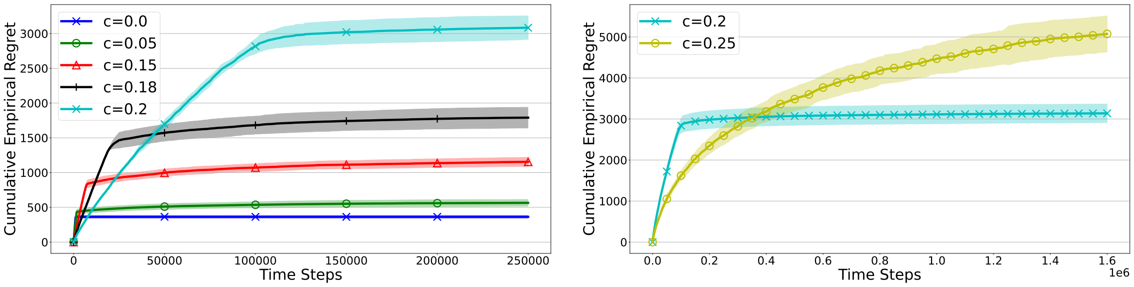

Fig. 4. Plots of the cumulative empirical regret of the UCBMP algorithm in a Bernoulli reward bandit setting with $K = 10$ arms for various probing cost values.

the upper bound on the expected reward when at most i probes are allowed. These estimated $P_i(t)$ values can be found as:

$$P_0(t) = U_{S_1(t)}(t)$$

$$P_1(t) = U_{S_1(t)}(t) + (1 - U_{S_1(t)}(t)) \cdot U_{S_2(t)}(t) - c$$

$$P_2(t) = U_{S_1(t)}(t) + (1 - U_{S_1(t)}(t)) \cdot U_{S_2(t)}(t) + (1 - U_{S_1(t)}(t)) \cdot (1 - U_{S_2(t)}(t)) \cdot U_{S_3(t)}(t) - c \cdot (2 - U_{S_1(t)}(t))$$

$$\begin{aligned} P_i(t) = U_{S_1(t)}(t) \cdot (1-c) + \sum_{j=2}^{i+1} U_{S_j(t)}(t) \cdot \prod_{k=1}^{j-1} (1 - U_{S_k(t)}(t)) - c \cdot \sum_{j=2}^{i-1} j \cdot U_{S_j(t)}(t) \cdot \prod_{k=1}^{j-1} (1 - U_{S_k(t)}(t)) \\ - i \cdot c \cdot \prod_{j=1}^{i-1} (1 - U_{S_j(t)}(t)), \quad 3 \le i \le K-1 \end{aligned} \tag{10}$$

The maximum number of probes that are allowed in round t, $s(t)$, can then be 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. 4. 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 $c \geq 0$. 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 $O(\sqrt{K^2T\log T})$, 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.

VII. ACKNOWLEDGMENTS

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

REFERENCES

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

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

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

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

[5] D. Chakrabarti, R. Kumar, F. Radlinski, and E. Upfal, "Mortal multi-armed bandits," *Advances in neural information processing systems*, vol. 21, 2008.

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

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

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

[9] J. Mary, R. Gaudel, and P. Preux, "Bandits and recommender systems," in *Machine Learning, Optimization, and Big Data: First International Workshop, MOD 2015, Taormina, Sicily, Italy, July 21-23, 2015, Revised Selected Papers 1*, pp. 325–336, Springer, 2015.

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

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

[12] J. Langford and T. Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information," *Advances in neural information processing systems*, vol. 20, 2007.

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

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

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

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

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

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

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

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

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

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

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

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

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

[26] 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, p. e113457, 2014.

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

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

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

[30] S. Kale, "Multiarmed bandits with limited expert advice," in *Conference on Learning Theory*, pp. 107–122, PMLR, 2014.

[31] S. Guha, K. Munagala, and S. Sarkar, "Optimizing transmission rate in wireless channels using adaptive probes," in *Proceedings of the joint international conference on Measurement and modeling of computer systems*, pp. 381–382, 2006.

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

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

[34] W. Chen, Y. Wang, and Y. Yuan, “Combinatorial multi-armed bandit: General framework and applications,” in *International conference on machine learning*, pp. 151–159, PMLR, 2013.

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

[36] M. Jourdan, M. Mutnỳ, J. Kirschner, and A. Krause, “Efficient pure exploration for combinatorial bandits with semi-bandit feedback,” in *Algorithmic Learning Theory*, pp. 805–849, PMLR, 2021.

[37] Q. Liu, W. Xu, S. Wang, and Z. Fang, “Combinatorial bandits with linear constraints: Beyond knapsacks and fairness,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 2997–3010, 2022.

[38] F. Li, J. Liu, and B. Ji, “Combinatorial sleeping bandits with fairness constraints,” *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 3, pp. 1799–1813, 2019.

[39] Y. Gai, B. Krishnamachari, and R. Jain, “Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations,” *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1466–1478, 2012.

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

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

[42] A. Huyuk and C. Tekin, “Analysis of thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms,” in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1322–1330, PMLR, 2019.

[43] A. Hüyük and C. Tekin, “Thompson sampling for combinatorial network optimization in unknown environments,” *IEEE/ACM Transactions on Networking*, vol. 28, no. 6, pp. 2836–2849, 2020.

[44] Z. Zhong, W. C. Cheung, and V. Tan, “Best arm identification for cascading bandits in the fixed confidence setting,” in *International Conference on Machine Learning*, pp. 11481–11491, PMLR, 2020.

[45] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, “Combinatorial cascading bandits,” *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[46] Z. Zhong, W. C. Chueng, and V. Y. Tan, “Thompson sampling algorithms for cascading bandits,” *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 9915–9980, 2021.

[47] A. Bhaskara, A. Cutkosky, R. Kumar, and M. Purohit, “Logarithmic regret from sublinear hints,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 28222–28232, 2021.

[48] A. Goel, S. Guha, and K. Munagala, “Asking the right questions: Model-driven optimization using probes,” in *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 203–212, 2006.

[49] A. A. Mogyla, “Application of stochastic probing radio signals for the range-velocity ambiguity resolution in doppler weather radars,” *Radioelectronics and Communications Systems*, vol. 57, no. 12, pp. 542–552, 2014.

[50] M. Weitzman, *Optimal search for the best alternative*, vol. 78. Department of Energy, 1978.

[51] S. Chawla, E. Gergatsouli, Y. Teng, C. Tzamos, and R. Zhang, “Pandora’s box with correlations: Learning and approximation,” in *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1214–1225, IEEE, 2020.

[52] S. Boodaghians, F. Fusco, P. Lazos, and S. Leonardi, “Pandora’s box problem with order constraints,” in *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 439–458, 2020.

[53] H. Beyhaghi and R. Kleinberg, “Pandora’s problem with nonobligatory inspection,” in *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 131–132, 2019.

[54] D. Russo and B. Van Roy, “Learning to optimize via posterior sampling,” *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.

[55] S. Li, B. Wang, S. Zhang, and W. Chen, “Contextual combinatorial cascading bandits,” in *International conference on machine learning*, pp. 1245–1253, PMLR, 2016.

[56] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, “Tight regret bounds for stochastic combinatorial semi-bandits,” in *Artificial Intelligence and Statistics*, pp. 535–543, PMLR, 2015.

[57] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *Acm transactions on interactive intelligent systems (tiis)*, vol. 5, no. 4, pp. 1–19, 2015.

[58] R. Zhang and R. Combes, "On the suboptimality of thompson sampling in high dimensions," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8345–8354, 2021.

[59] Y. Saito, A. Shunsuke, M. Megumi, and N. Yusuke, "Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation," *arXiv preprint arXiv:2008.07146*, 2020.

[60] B. Kveton, Z. Wen, A. Ashkan, H. Eydgahi, and B. Eriksson, "Matroid bandits: Fast combinatorial optimization with learning," in *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pp. 420–429, 2014.

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, Turkey, in 2019. He is currently pursuing the Ph.D. degree in electrical and computer engineering at Carnegie Mellon University, PA, USA. His main research interests include multi-armed bandits, information theory, machine learning, and optimization. He is 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 the Middle East Technical University, Ankara, Turkey, in 2008, and the M.S.E. degree in electrical engineering: systems, M.S. degree in mathematics, and 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 Postdoctoral 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, Turkey. His research interests include multiarmed bandit problems, reinforcement learning, and cognitive communications. He was the recipient of the numerous awards including the Fred W. Ellersick Award for the Best Paper in MILCOM 2009 and the Distinguished Young Scientist (BAGEP) Award of the Science Academy Association of Turkey in 2019, and 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 (Turkey) in 2007, and the Ph.D. degree in Electrical and Computer Engineering from University of Maryland, College Park, MD in 2011. In August 2013, he joined the faculty of the Department of Electrical and Computer Engineering at Carnegie Mellon University, where he is currently a Research Professor. Dr. Yagan's research is on 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 is a recipient of the CIT Dean's Early Career Fellowship, IBM Faculty Award, and ICC 2021 Best Paper Award.

APPENDIX A
PRELIMINARIES

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

Fact A.1 (Hoeffding's Inequality)**.** Let $Z_1, Z_2, \cdots, Z_n$ be i.i.d. random variables bounded between $a_i \leq Z_i \leq b_i$, then for any $\delta > 0$, we have

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

Lemma A.2 ([1, 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

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

Fact A.3 (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 A.4. We include the following trivial bounds for the $\max$ function when $b > 0,\ c > 0$:

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

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

$$\mathbb{E}[\max(r_i, a)] + b\mathbb{P}(r_i < a) \leq \mathbb{E}[\max(r_i, a + b)] \leq\ \mathbb{E}[\max(r_i, a)] + b\mathbb{P}(r_i \leq a + b)\ . \tag{11}$$

Appendix B
DERIVATION OF CONFIDENCE INTERVALS FOR PROBING ACTIONS

In this section we derive the high probability confidence intervals used in the algorithm when the arms are probed. First, we start by recalling the upper confidence bound used in the standard UCB algorithm [2].

$$U_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{2\log t}{N_i(t)}}$$

Fact B.1. At any time t, the probability that the true mean of the reward of arm i is within its confidence interval interval can be upper bounded with the following two inequalities:

$$\mathbb{P}\left(\mu_i - \hat{\mu}_i > \sqrt{\frac{2\log t}{N_i(t)}}\right) \leq e^{-4\log t} = t^{-3} \tag{12}$$

$$\mathbb{P}\left(\hat{\mu}_i - \mu_i > \sqrt{\frac{2\log t}{N_i(t)}}\right) \leq e^{-4\log t} = t^{-3} \tag{13}$$

Proof.

$$\mathbb{P}\left(\mu_i - \hat{\mu}_i > \sqrt{\frac{2\log t}{N_i(t)}}\right) \leq \sum_{N_i(t)=1}^{t} \mathbb{P}\left(\mu_i - \hat{\mu}_i > \sqrt{\frac{2\log t}{N_i(t)}}\right) \leq \sum_{n=1}^{t} t^{-4} \leq t^{-3}$$

where we used Fact A.1. The second inequality is also proved similarly. □

We now derive similar upper confidence bounds for the case where the arms are probed. In particular, we will show the following result.

Lemma B.2. For the action $a = (i, j)$, and the confidence interval defined as $C_{(i,j)}(t) = \sqrt{\frac{2\log t}{N_j(t)}} + \sqrt{\frac{2\log t}{N_i(t)}}$, the probability that the empirical probing reward being outside the confidence interval can be upper bounded as:

$$\mathbb{P}\left(|\hat{\nu}_{(i,j)}(t) - \nu_{(i,j)}| \geq \sqrt{\frac{2\log t}{N_j(t)}} + \sqrt{\frac{2\log t}{N_i(t)}}\right) \leq 4t^{-3}$$

Proof. First, using (11) with $a = \hat{\mu}_j(t)$ and $b = \sqrt{\frac{2\log t}{N_1(t)}}$, along with (12) and (13), it can be seen that each of the following occurs with probability at least $1 - t^{-3}$. Note that the rather loose bounds $P\left(r_i \leq a + b\right) \leq 1$ is used here since the actual probabilities are unknown.

$$\mathbb{E}[\max(r_i, \mu_j)] \leq \ \mathbb{E}[\max(r_i, \hat{\mu}_j(t))] + \sqrt{\frac{2\log t}{N_j(t)}} \tag{14}$$

$$\mathbb{E}[\max(r_i, \hat{\mu}_j(t))] - \sqrt{\frac{2\log t}{N_j(t)}} \leq \ \mathbb{E}[\max(r_i, \mu_j)] \tag{15}$$

Using (A.1), each of the following holds with probability at least $1 - t^{-3}$:

$$\mathbb{E}[\max(r_i, \hat{\mu}_j(t))] - \sqrt{\frac{2\log t}{N_i(t)}} \leq \hat{\nu}_{(i,j)}(t) \tag{16}$$

$$\hat{\nu}_{(i,j)}(t) \leq \ \mathbb{E}[\max(r_i, \hat{\mu}_j(t))] + \sqrt{\frac{2\log t}{N_i(t)}} \tag{17}$$

Combining (14) and (15) with (16) and (17), also noting that $\nu_{(i,j)} = \mathbb{E}[\max(r_i, \mu_j)]$, it can be seen that with probability at least $1 - 2t^{-4}$, each of the following occur:

$$\nu_{(i,j)} - \sqrt{\frac{2\log t}{N_j(t)}} - \sqrt{\frac{2\log t}{N_i(t)}} \leq \hat{\nu}_{(i,j)}(t)$$

$$\hat{\nu}_{(i,j)}(t) \leq \nu_{(i,j)} + \sqrt{\frac{2\log t}{N_j(t)}} + \sqrt{\frac{2\log t}{N_i(t)}}$$

Hence, $\mathbb{P}\left(|\hat{\nu}_{(i,j)}(t) - \nu_{(i,j)}| \geq \sqrt{\frac{2\log t}{N_j(t)}} + \sqrt{\frac{2\log t}{N_i(t)}}\right) \leq 4t^{-3}$, which concludes the analysis for the upper bound on the probability of the estimated probing reward falling outside the confidence interval. □

Appendix C
Proof of Theorem IV.3

We provide the gap-dependent regret analysis of the UCBP algorithm in this section. 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. The proof follows some of the steps in the proof of Theorem 3 in [56], and the proof of Lemma A.2 in [55].

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, again letting $\mathcal{B}_a(t)$ denote the event that the decision to pull the probe or backup arm is correct, i.e. $\mathcal{B}_a(t) = \mathbb{1}\{r_{\hat{a}}(t) = r_a(t)\}$, the empirical regret can be decomposed as

$$\hat{R}_T = \sum_{t=K+1}^{T} \sum_{a\in\mathcal{A}} \left[\mathbb{1}\{a(t) = a, \mathcal{B}_a(t)\} \cdot (\nu^* - \nu_{a(t)}(t)) + \mathbb{1}\{a(t) = a, \mathcal{B}_a^c(t)\} \cdot (\nu^* - \nu_{a(t)}(t) + d_a(t))\right] + K$$

The summation in time starts from $t = K+1$ due to the UCBP algorithm taking each action once in the first K rounds, and this can contribute at most K to regret since the rewards are bounded. Expected regret can be obtained by taking the expectation of this expression

$$R_T = \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{a\in\mathcal{A}} \left[\mathbb{1}\{a(t) = a, \mathcal{B}_a(t)\} \cdot (\nu^* - \nu_{a(t)}(t)) + \mathbb{1}\{a(t) = a, \mathcal{B}_a^c(t)\} \cdot (\nu^* - \nu_{a(t)}(t) + d_a(t))\right]\right] + K$$

$$= \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{a\in\mathcal{A}} \left[\mathbb{1}\{a(t) = a\} \cdot (\nu^* - \nu_{a(t)}(t)) + \mathbb{1}\{a(t) = a, \mathcal{B}_a^c(t)\} \cdot d_a(t)\right]\right] + K$$

Define the following events

$$\mathcal{E}_t := \{|\hat{\mu}_i(t) - \mu_i| \leq C_{(i,\emptyset)}(t) \wedge |\hat{\nu}_{(i,j)}(t) - \nu_{(i,j)}| \leq C_{(i,j)}(t),\ \forall i,j \in [K],\ i \neq j\}, \text{ and}$$

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

where $\mathcal{E}_t$ is the event that all confidence intervals hold in round t, and $\mathcal{E}(T)$ is the event that all confidence intervals hold for all rounds $K+1 \leq t \leq T$. Regret can be decomposed based on this event $\mathcal{E}(T)$ as:

$$R_T \leq \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}}\left[\mathbb{1}\{a(t)=a\}\cdot(\nu^*-\nu_{a(t)}(t))+\mathbb{1}\{a(t)=a\}\cdot d_a(t)\right]\Big|\mathcal{E}(T)\right]+\sum_{t=K+1}^{T}\mathbb{P}(\mathcal{E}_1^c(t))+K$$

Define

$$R_a(T):=\mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}}\mathbb{1}\{a(t)=a\}\cdot(\nu^*-\nu_{a(t)}(t))\Big|\mathcal{E}(T)\right]=\mathbb{E}\left[\sum_{t=K+1}^{T}R_t(a)\Big|\mathcal{E}(T)\right]$$

$$R_{\text{ref}}(T):=\mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}}\mathbb{1}\{a(t)=a\}\cdot d_a(t)\Big|\mathcal{E}(T)\right]$$

where $R_t(a):=\sum_{a\in\mathcal{A}}\mathbb{1}\{a(t)=a\}\cdot(\nu^*-\nu_{a(t)}(t))$ is the regret of the action (without the reference point regret) under the event $\mathcal{E}(T)$. We start by upper bounding $R_a(T)$.

Similar to the proof of gap-independent upper bound in Section **??**, define $o(t)\subset a(t)$ as the set of arms whose reward is observed in round t; $\mathcal{H}_t=(a(1),r(1),o(1),\cdots,a(t-1),r(t-1),o(t-1),a(1))$ 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. Let $\mathcal{B}_a(t),\ a\in\mathcal{A}_p$ denote the event that the reward of both the probe arm and the backup arm is observed in round t. Also let $p(a(t),t)$ denote the conditional probability of observing the reward of arm i at round t when the chosen action is $a(t)$ given $\mathcal{H}_t$. Following the analysis in [55], regret can be decomposed in the following way if the confidence intervals hold:

$$\begin{aligned}
R_a(T)&=\mathbb{E}\left[\sum_{t=K+1}^{T}R_t(a)\Big|\mathcal{E}(T)\right]\\
&=\mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{E}[R_t(a)|\mathcal{H}_t]\Big|\mathcal{E}(T)\right] \qquad (18)\\
&=\mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{E}\left[\sum_{a\in\mathcal{A}_p}\mathbb{1}\{a(t)=a\}\cdot(\nu^*-\nu_{a(t)}(t))+\sum_{a\in\mathcal{A}_s}\mathbb{1}\{a(t)=a\}\cdot(\nu^*-\nu_{a(t)}(t))|\mathcal{H}_t\right]\Big|\mathcal{E}(T)\right]\\
&=\mathbb{E}\Bigg[\sum_{t=K+1}^{T}\mathbb{E}\Bigg[\sum_{a\in\mathcal{A}_p}\mathbb{1}\{a(t)=a\}\cdot(\nu^*-\nu_{a(t)}(t))\cdot\mathbb{E}\left[\frac{\mathbb{1}\{\mathcal{B}_a(t)\}}{p(a(t),t)}|\mathcal{H}_t\right] \qquad (19)\\
&\qquad\qquad+\sum_{a\in\mathcal{A}_s}\mathbb{1}\{a(t)=a\}\cdot(\nu^*-\nu_{a(t)}(t))|\mathcal{H}_t\Bigg]\Bigg|\mathcal{E}(T)\Bigg]\\
&\leq\mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{E}\left[\sum_{a\in\mathcal{A}_p}\mathbb{1}\{a(t)=a,\mathcal{B}_a(t)\}\cdot\frac{\nu^*-\nu_{a(t)}(t)}{\epsilon}+\sum_{a\in\mathcal{A}_s}\mathbb{1}\{a(t)=a\}\cdot(\nu^*-\nu_{a(t)}(t))|\mathcal{H}_t\right]\Big|\mathcal{E}(T)\right]\\
&\qquad\qquad (20)\\
&=\mathbb{E}\left[\sum_{t=K+1}^{T}\left(\sum_{a\in\mathcal{A}_p}\mathbb{1}\{a(t)=a,\mathcal{B}_a(t)\}\cdot\frac{\Delta_a}{\epsilon}+\sum_{a\in\mathcal{A}_s}\mathbb{1}\{a(t)=a\}\cdot\Delta_a\right)\Big|\mathcal{E}(T)\right]
\end{aligned}$$

Eq. (18) is due to the tower rule. In Eq. (19) we used the fact that given $\mathcal{H}_t$, the probability of the event $\mathcal{B}_a(t)$ is $p(a(t),t)$. In Eq. (20), we used $p(a(t),t)\geq\epsilon$.

To upper bound this expression, we note the following condition for an action to be chosen at round t. 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 arm can only be pulled in round t if

$$\hat{\nu}_{a^*}(t) + C_{a^*}(t) \leq \hat{\nu}_a(t) + C_a(t)$$

is satisfied. Using the fact that $\nu^* \leq \hat{\nu}_{a^*}(t) + C_{a^*}(t)$, we have

$$\nu^* \leq \nu_a + 2C_a(t)$$
$$\frac{\Delta_a}{2} \leq C_a(t)$$

If $a = (i, \emptyset) \in \mathcal{A}_s$, this condition can be written as:

$$\frac{\Delta_{(i,\emptyset)}}{2} \leq C_{(i,\emptyset)}(t) = \sqrt{\frac{2\log t}{N_i(t)}}$$
$$N_i(t) \leq \frac{8\log t}{\Delta^2_{(i,\emptyset)}}$$

Further, if $a = (i, j) \in \mathcal{A}_p$, this condition can be written as:

$$\frac{\Delta_{(i,j)}}{2} \leq C_{(i,j)}(t) = \sqrt{\frac{2\log t}{N_i(t)}} + \sqrt{\frac{2\log t}{N_j(t)}}$$

Define event $\mathcal{R}_a(t) = \{2C_a(t) \geq \Delta_a\}$. It can be seen that action a can be chosen in round t only when $\mathcal{R}_a(t)$ happens. Also define

$$\mathcal{G}_t = \left\{ a(t) \in \mathcal{A}_p, \text{ at least one of the base arms in } a(t) \text{ was observed at most } \frac{32\log t}{\Delta^2_{a(t)}} \text{ times} \right\}$$

It can be seen that action $a(t) \in \mathcal{A}_p$ cannot be chosen under the event $\mathcal{G}^c_t$. This is since for $\mathcal{G}^c_t$ to happen, both of the base arms need to be sampled more than $\frac{32\log t}{\Delta^2_{a(t)}}$ times. Then, under $\mathcal{G}^c_t$ when $a(t) \in \mathcal{A}_p$, we have that

$$C_{a(t)} = \sum_{i \in a(t)} \sqrt{\frac{2\log t}{N_i(t)}} \leq 2\sqrt{\frac{2\log t}{\frac{32\log t}{\Delta^2_{a(t)}}}} = \frac{\Delta_{a(t)}}{2}$$

Since the event $\mathcal{R}_a(t)$ does not hold when $\mathcal{G}^c_t$ happens, it can be seen that action $a(t)$ can only be chosen in round t when $\mathcal{G}_t$ happens. Further, define

$$\mathcal{G}_{i,t} := \mathcal{G}_t \cap \left\{ i \in a(t), N_i(t) \leq \frac{32\log t}{\Delta^2_{a(t)}} \right\}$$

as the event that the base arm i is not observed *sufficiently often* under event $\mathcal{G}_t$. Then, it can be seen that

$$\mathbb{1}\{\mathcal{G}_t, \Delta_{a(t)} > 0\} \leq \sum_{i=1}^{K} \mathbb{1}\{\mathcal{G}_{i,t}, \Delta_{a(t)} > 0\}.$$

Using this, regret can be bounded as:

$$R_a(T) \leq \mathbb{E}\left[\sum_{t=K+1}^{T} \left(\sum_{a \in \mathcal{A}_p} \mathbb{1}\{a(t) = a, \mathcal{B}_a(t)\} \cdot \frac{\Delta_a}{\epsilon} + \sum_{i=1}^{K} \mathbb{1}\left\{ a(t) = (i, \emptyset), N_i(t) \leq \frac{8\log t}{\Delta^2_{(i,\emptyset)}} \right\} \cdot \Delta_{(i,\emptyset)} \right) \middle| \mathcal{E}(T) \right]$$

$$\leq \mathbb{E}\left[\sum_{t=K+1}^{T}\left(\sum_{i=1}^{K}\mathbb{1}\{\mathcal{G}_{i,t},\mathcal{B}_a(t)\}\cdot\frac{\Delta_{a(t)}}{\epsilon}+\sum_{i=1}^{K}\mathbb{1}\left\{a(t)=(i,\emptyset),N_i(t)\leq\frac{8\log t}{\Delta^2_{(i,\emptyset)}}\right\}\cdot\Delta_{(i,\emptyset)}\right)\middle|\mathcal{E}(T)\right]$$

Let $\Delta_{\text{min},i} := \min_{a\in\mathcal{A}_p\setminus\{a^*\}\ s.t.\ i\in a}(\Delta_a)$ be the smallest gap of suboptimal probing actions that include base arm i, and let

$$\mathcal{S}_i := \begin{cases}\{\Delta_a : a\in\mathcal{A}_p\setminus\{a^*\}, i\in a\}\cup\{2\Delta_{i,\emptyset}\} & \text{, if } a^*\neq(i,\emptyset)\\ \{\Delta_a : a\in\mathcal{A}_p\setminus\{a^*\}, i\in a\} & \text{, if } a^*=(i,\emptyset)\end{cases}$$

be the set of gaps of suboptimal actions that involve probing and also 2 times the gap of the action $(i,\emptyset)$ if $(i,\emptyset)$ is not the optimal action. Also let $\sigma_{i,1}\geq\cdots\geq\sigma_{i,M_i}$ be the gaps of the actions in $\mathcal{S}_i$ ordered from the one with largest gap to the smallest one, and let η_i be the index where $\sigma_{i,\eta_i}=2\Delta_{i,\emptyset}$ (the index of action $(i,\emptyset)$) if action $(i,\emptyset)$ is suboptimal; for the case where action $(i,\emptyset)$ is optimal, let $\eta_i=0$. Using this, define $\tau_{i,j}=\Delta_{i,\emptyset}$ if $j=\eta_i$, and $\tau_{i,j}=\frac{\sigma_{i,j}}{\epsilon}$ otherwise. Note that M_i is equal to either $2K-1$ or $2K-2$ depending on whether the optimal action contains the base arm i. Then,

$$R_a(T)\leq\mathbb{E}\left[\sum_{t=K+1}^{T}\left(\sum_{i=1}^{K}\sum_{j=1}^{M_i}\mathbb{1}\left\{\mathcal{G}_{i,t},\mathcal{B}_a(t),\Delta_{a(t)}=\sigma_{i,j}\right\}\cup\mathbb{1}\left\{a(t)=(i,\emptyset),N_i(t)\leq\frac{8\log t}{\Delta^2_{(i,\emptyset)}}\right\}\cdot\tau_{i,j}\right)\middle|\mathcal{E}(T)\right]$$
$$\leq\mathbb{E}\left[\sum_{t=K+1}^{T}\left(\sum_{i=1}^{K}\sum_{j=1}^{M_i}\mathbb{1}\left\{i\in a(t),N_i(t)\leq\frac{32\log t}{\sigma^2_{i,j}}\right\}\cdot\tau_{i,j}\right)\middle|\mathcal{E}(T)\right]$$

Here, the terms $\sigma_{i,j}$ are related to how many times a base arm should be observed, and the terms $\tau_{i,j}$ are related to mean error received in expectation until the rewards observed from the action lead to an observation of the rewards of all the base arms involved in that action. To proceed, as in [56], we consider the worst case the samples for the base arms can be obtained. The key idea is observing that highest regret is possible when we assume $\frac{32\log T}{\sigma^2_{i,1}}$ samples are obtained by sampling the action with highest gap $\sigma_{i,1}$ for an expected regret of $\tau_{i,1}$ per sample; and then when we sample the action with $\sigma_{i,2}$ gap $\frac{32\log T}{\sigma^2_{i,2}}-\frac{32\log T}{\sigma^2_{i,1}}$ times and receive an expected regret of $\tau_{i,2}$ per sample; and so on. This key idea can only be used when $\sigma_{i,j}$ and $\tau_{i,j}$ have the same ordering, i.e. when $\tau_{i,j}$ is also the j^{th} largest $\tau_{i,\cdot}$ for all possible j values. It can be seen that except the pull action $(i,\emptyset)$, the $\sigma_{i,j}$ and $\tau_{i,j}$ terms will follow the same ordering, so we consider the following five cases.

Case 1: If $a^*=(i,\emptyset)$, we do not need to consider η_i in the upper bound, the regret can simply be upper bounded as:

$$R_a(T)\leq R_{a,i}(T)$$
$$R_{a,i}(T)\leq\sum_{i=1}^{K}32\log T\cdot\left[\tau_{i,1}\cdot\frac{1}{\sigma^2_{i,1}}+\sum_{k=2}^{M_i}\tau_{i,k}\cdot\left(\frac{1}{\sigma^2_{i,k}}-\frac{1}{\sigma^2_{i,k-1}}\right)\right]$$
$$=32\log T\cdot\left[\frac{\sigma_{i,1}}{\epsilon}\cdot\frac{1}{\sigma^2_{i,1}}+\sum_{k=2}^{M_i}\frac{\sigma_{i,k}}{\epsilon}\cdot\left(\frac{1}{\sigma^2_{i,k}}-\frac{1}{\sigma^2_{i,k-1}}\right)\right]$$

The following upper bound can be derived to upper bound this expression as in Lemma 4 of [60]:

$$\sigma_{i,1}\cdot\frac{1}{\sigma^2_{i,1}}+\sum_{k=2}^{M_i}\sigma_{i,k}\cdot\left(\frac{1}{\sigma^2_{i,k}}-\frac{1}{\sigma^2_{i,k-1}}\right)=\frac{1}{\sigma_{i,M_i}}+\sum_{k=1}^{M_i-1}\frac{\sigma_{i,k}-\sigma_{i,k+1}}{\sigma^2_{i,k}}$$

$$\leq \frac{1}{\sigma_{i,M_i}} + \sum_{k=1}^{M_i-1} \frac{\sigma_{i,k} - \sigma_{i,k+1}}{\sigma_{i,k}\sigma_{i,k+1}}$$

$$\leq \frac{1}{\sigma_{i,M_i}} + \sum_{k=1}^{M_i-1} \frac{1}{\sigma_{i,k+1}} - \frac{1}{\sigma_{i,k}} \leq \frac{2}{\sigma_{i,M_i}}$$

Using this, it can be seen that

$$R_{a,i}(T) \leq 32 \log T \cdot \frac{2}{\epsilon \sigma_{i,M_i}} = \frac{64 \log T}{\epsilon \Delta_{\text{min},i}}$$

Case 2: If $\Delta_{i,\emptyset} \leq \frac{\epsilon \Delta_{\text{min},i}}{8}$, we will have $\eta_i = M_i$, and $\tau_{i,M_i} \leq \tau_{i,M_i-1}$, so the regret can be upper bounded as:

$$R_{a,i}(T) \leq \sum_{i=1}^{K} 32 \log T \cdot \left[\tau_{i,1} \cdot \frac{1}{\sigma_{i,1}^2} + \sum_{k=2}^{M_i-1} \tau_{i,k} \cdot \left(\frac{1}{\sigma_{i,k}^2} - \frac{1}{\sigma_{i,k-1}^2} \right) + \tau_{i,M_i} \cdot \left(\frac{1}{\sigma_{i,M_i}^2} - \frac{1}{\sigma_{i,M_i-1}^2} \right) \right]$$

$$= 32 \log T \cdot \left[\frac{\sigma_{i,1}}{\epsilon} \cdot \frac{1}{\sigma_{i,1}^2} + \sum_{k=2}^{M_i-1} \frac{\sigma_{i,k}}{\epsilon} \cdot \left(\frac{1}{\sigma_{i,k}^2} - \frac{1}{\sigma_{i,k-1}^2} \right) + \frac{\sigma_{i,M_i}}{2} \cdot \left(\frac{1}{\sigma_{i,M_i}^2} - \frac{1}{\sigma_{i,M_i-1}^2} \right) \right]$$

Similar to case 1, we have:

$$\sigma_{i,1} \cdot \frac{1}{\sigma_{i,1}^2} + \sum_{k=2}^{M_i-1} \sigma_{i,k} \cdot \left(\frac{1}{\sigma_{i,k}^2} - \frac{1}{\sigma_{i,k-1}^2} \right) \leq \frac{2}{\sigma_{i,M_i-1}}$$

Using this, it can be seen that

$$R_{a,i}(T) \leq 32 \log T \cdot \left[\frac{2}{\epsilon \sigma_{i,M_i-1}} + \frac{1}{2\sigma_{i,M_i}} - \frac{\sigma_{i,M_i}}{2\sigma_{i,M_i-1}^2} \right] \tag{21}$$

Noting that $\sigma_{i,M_i-1} \geq \frac{4}{\epsilon}\sigma_{i,M_i}$, we have

$$R_{a,i}(T) \leq 32 \log T \cdot \left[\frac{1}{2\sigma_{i,M_i}} + \frac{1}{2\sigma_{i,M_i}} - \frac{\sigma_{i,M_i}}{2\sigma_{i,M_i-1}^2} \right] \leq \frac{32 \log T}{\sigma_{i,M_i}} = \frac{32 \log T}{2\Delta_{(i,\emptyset)}} = \frac{16 \log T}{\Delta_{(i,\emptyset)}}$$

Case 3: If $\frac{\epsilon \Delta_{\text{min},i}}{8} \leq \Delta_{i,\emptyset} \leq \frac{\epsilon \Delta_{\text{min},i}}{4}$, using Eq. (21) with $\sigma_{i,M_i-1} \geq \frac{2}{\epsilon}\sigma_{i,M_i}$, we have

$$R_{a,i}(T) \leq 32 \log T \cdot \left[\frac{1}{\sigma_{i,M_i}} + \frac{1}{2\sigma_{i,M_i}} - \frac{\sigma_{i,M_i}}{2\sigma_{i,M_i-1}^2} \right] \leq \frac{48 \log T}{\sigma_{i,M_i}} = \frac{24 \log T}{\Delta_{(i,\emptyset)}}$$

Case 4: If $2\Delta_{i,\emptyset} \leq \Delta_{\text{min},i} \leq \frac{4}{\epsilon}\Delta_{i,\emptyset}$, we will have $\eta_i = M_i$, so the regret can be upper bounded as:

$$R_{a,i}(T) \leq \sum_{i=1}^{K} 32 \log T \cdot \left[\tau_{i,1} \cdot \frac{1}{\sigma_{i,1}^2} + \sum_{k=2}^{M_i-1} \tau_{i,k} \cdot \left(\frac{1}{\sigma_{i,k}^2} - \frac{1}{\sigma_{i,k-1}^2} \right) + \tau_{i,M_i} \cdot \left(\frac{1}{\sigma_{i,M_i}^2} - \frac{1}{\sigma_{i,M_i-1}^2} \right) \right]$$

$$= 32 \log T \cdot \left[\frac{\sigma_{i,1}}{\epsilon} \cdot \frac{1}{\sigma_{i,1}^2} + \sum_{k=2}^{M_i-1} \frac{\sigma_{i,k}}{\epsilon} \cdot \left(\frac{1}{\sigma_{i,k}^2} - \frac{1}{\sigma_{i,k-1}^2} \right) + \frac{\sigma_{i,M_i}}{2} \cdot \left(\frac{1}{\sigma_{i,M_i}^2} - \frac{1}{\sigma_{i,M_i-1}^2} \right) \right]$$

$$\leq 32 \log T \cdot \left[\frac{2}{\epsilon \sigma_{i,M_i-1}} + \frac{1}{2\sigma_{i,M_i}} - \frac{\sigma_{i,M_i}}{2\sigma_{i,M_i-1}^2} \right]$$

Using $2\sigma_{i,M_i} \geq \epsilon \sigma_{i,M_i-1}$, it can be seen that:

$$R_{a,i}(T) \leq 32 \log T \cdot \left[\frac{3}{\epsilon \sigma_{i,M_i-1}} \right] = \frac{96 \log T}{\epsilon \sigma_{i,M_i-1}} = \frac{96 \log T}{\epsilon \Delta_{\text{min},i}}$$

Case 5: If $\Delta_{\text{min},i} \leq 2\Delta_{i,\emptyset}$, we will have $\eta_i \neq M_i$, so the regret can be upper bounded as:

$$\begin{aligned}
R_{a,i}(T) &\leq 32\log T \cdot \left[\tau_{i,1}\cdot\frac{1}{\sigma_{i,1}^2} + \sum_{k=2,k\neq\eta_i}^{M_i}\tau_{i,k}\cdot\left(\frac{1}{\sigma_{i,k}^2}-\frac{1}{\sigma_{i,k-1}^2}\right) + \frac{2}{\epsilon}\tau_{i,\eta_i}\cdot\left(\frac{1}{\sigma_{i,\eta_i}^2}-\frac{1}{\sigma_{i,\eta_i-1}^2}\right)\right] \\
&\leq \frac{32\log T}{\epsilon}\cdot\left[\sigma_{i,1}\cdot\frac{1}{\sigma_{i,1}^2} + \sum_{k=2}^{M_i}\sigma_{i,k}\cdot\left(\frac{1}{\sigma_{i,k}^2}-\frac{1}{\sigma_{i,k-1}^2}\right)\right] \\
&\leq \frac{64\log T}{\epsilon\sigma_{i,M_i}} = \frac{64\log T}{\epsilon\Delta_{\text{min},i}}
\end{aligned}$$

Combining all these cases, it can be concluded that

$$R_T \leq \sum_{i=1}^{K}\frac{16\log T}{\delta_i} + R_{\text{ref}}(T) + \frac{5\pi^2 K}{3} + K$$

where

$$\delta_i = \begin{cases} \rho_i & \text{if } a^* = (i,\emptyset) \\ \frac{2\min(\rho_i,\Delta_{(i,\emptyset)})}{3} & \text{if } \frac{\rho_i}{2} \leq \Delta_{(i,\emptyset)} \leq \frac{2\rho_i}{\epsilon} \\ \min(\rho_i,\Delta_{(i,\emptyset)}) & \text{otherwise} \end{cases}$$

, and $\rho_i = \min_{a\in\mathcal{A}_p\setminus\{a^*\}\ s.t.\ i\in a}\left(\frac{\epsilon\Delta_a}{4}\right)$. □

Appendix D
Upper Bound on Reference Point Regret

Recall that we call the error introduced due to an incorrect decision on pulling the probe arm or the backup arm the *reference point error*. We will denote the regret incurred from the reference point error when action a is taken in round t as $d_a(t)$. 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. Hence, $d_a(t)$ can be expressed as:

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

where $r_{\hat{a}}(t) = r_i(t)\mathbb{1}\{r_i(t) > \hat{\mu}_j(t)\} + r_j(t)\mathbb{1}\{r_i(t) \leq \hat{\mu}_j(t)\}$ is the reward received from action (i,j) in round t when $\hat{\mu}_j(t)$ is used as the reference point, and $r_a(t) = r_i(t)\mathbb{1}\{r_i(t) > \mu_j\} + r_j(t)\mathbb{1}\{r_i(t) \leq \mu_j\}$ is the reward received from the optimal decision rule, i.e. when μ_j is used as the reference point. Also recall that $\mathcal{B}_a(t) = \mathbb{1}\{r_{\hat{a}}(t) = r_a(t)\}$ denotes the event that the decision to pull the probe or backup arm is correct.

Let $o(t) \subset a(t)$ be the set of arms whose reward is observed in round t. Let $\mathcal{H}_t = (a(1), r(1), o(1), \cdots, a(t-1), r(t-1), o(t-1), a(1))$ be the history of UCBP up to choosing action $a(t)$.

Lemma D.1. 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 $d_{(i,j)}(t) \leq C_{(j,\emptyset)}(t)$.

Proof. To upper bound $d_{(i,j)}(t)$, notice that when $r_i(t)$ is not between the values of $\hat{\mu}_j(t)$ 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 $\hat{\mu}_j(t)$ 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 $\hat{\mu}_j(t) \geq r_i(t) \geq \mu_j$. Then, the UCBP algorithm will decide to pull the backup arm j to 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 $d_{(i,j)}(t) = r_i(t) - \mu_j \leq \hat{\mu}_j(t) - \mu_j = C_{(j,\emptyset)}(t)$ in this case. The second case is when $\mu_j \geq r_i(t) \geq \hat{\mu}_j(t)$, then it will be decided to pull arm i and get reward $r_i(t)$ even though the optimal decision is to pull the backup arm j to get expected reward μ_j. Again, the gap in reward compared to the optimal decision is $d_{(i,j)}(t) = \mu_j - r_i(t) \leq \mu_j - \hat{\mu}_j(t) = C_{(j,\emptyset)}(t)$. □

Lemma D.2. The cumulative reference point regret until round T can be upper bounded as:

$$R_{\text{ref}}(T) \leq \frac{2\sqrt{2KT\log T}}{\epsilon}$$

Proof. To derive an upper bound on the reference point regret, it can be seen from Lemma D.1 that $d_{(i,j)}(t) \leq C_{(j,\emptyset)}(t)$. Defining $N_a(T) := \sum_{t=K+1}^{T} \mathbb{1}\{a(t) = a\}$ 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, the reference point regret can be upper bounded as follows.

$$\begin{align*}
R_{\text{ref}}(T) &= \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{a\in\mathcal{A}}\mathbb{1}\{a(t)=a\}\cdot d_a(t)\Big|\mathcal{E}(T)\right] \\
&\leq \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{E}\left[\sum_{j=1}^{K}\mathbb{1}\{a(t)=(\cdot,j)\}\cdot C_{(j,\emptyset)}(t)\Big|\mathcal{H}_t\right]\Big|\mathcal{E}(T)\right] \tag{22} \\
&= \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{E}\left[\sum_{j=1}^{K}\mathbb{1}\{a(t)=(\cdot,j)\}\cdot C_{(j,\emptyset)}(t)\cdot\mathbb{E}\left[\frac{\mathbb{1}\{j\in o(t)\}}{p_j(a(t),t)}|\mathcal{H}_t\right]\Big|\mathcal{H}_t\right]\Big|\mathcal{E}(T)\right] \tag{23} \\
&\leq \frac{1}{\epsilon}\cdot\mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{E}\left[\sum_{j=1}^{K}\mathbb{1}\{a(t)=(\cdot,j)\}\cdot C_{(j,\emptyset)}(t)\cdot\mathbb{1}\{j\in o(t)\}\Big|\mathcal{H}_t\right]\Big|\mathcal{E}(T)\right] \tag{24} \\
&= \frac{1}{\epsilon}\cdot\mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{E}\left[\sum_{j=1}^{K}\mathbb{1}\{j\in o(t)\}\cdot\sqrt{\frac{2\log t}{N_j(t)}}\Big|\mathcal{H}_t\right]\Big|\mathcal{E}(T)\right] \\
&= \frac{\sqrt{2\log T}}{\epsilon}\cdot\mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{E}\left[\sum_{i=1}^{K}\mathbb{1}\{i\in o(t)\}\cdot\sqrt{\frac{1}{N_i(t)}}\Big|\mathcal{H}_t\right]\Big|\mathcal{E}(T)\right] \\
&\leq \frac{\sqrt{2\log T}}{\epsilon}\mathbb{E}\left[\cdot\sum_{i=1}^{K}\sum_{x=1}^{B_i(T)}\cdot\sqrt{\frac{1}{x}}\right] \\
&\leq \frac{2\sqrt{2\log T}}{\epsilon}\cdot\mathbb{E}\left[\sqrt{K\sum_{i=1}^{K}B_i(T)}\right] \tag{25}
\end{align*}$$

Again Eq. (22) is due to the tower rule. In Eq. (23) we used the fact that given $\mathcal{H}_t$, the probability of $j \in o(t)$ is $p_j(a(t), t)$. In Eq. (24), we used $p_j(a(t), t) \geq \epsilon$; and Cauchy-Schwarz inequality is used in Eq. (25). Using

$\sum_{i=1}^{K} B_i(T) \leq T$ concludes the proof. □

Lemma D.3. 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:

$$R_{\text{ref}}(T) \leq \sum_{i=1}^{K} \frac{4 \log T}{\epsilon \gamma_i}$$

where we use $d_l \in \mathcal{D},\ 1 \leq l \leq |\mathcal{D}|$ to denote the elements of the set $\mathcal{D}$; and we let $\gamma_i := \min_l |d_l - \mu_i|$ if $\mu_i \notin \mathcal{D}$, and $\gamma_i := |d_l - d_{l+1}|$ if $\mu_i \in \mathcal{D}$.

Proof. First, $d_{(i,\emptyset)}(t) = 0, \forall i \in [K]$ since this type of action does not involve probing. Hence, $R_{\text{ref}}(T)$ can be written as

$$\begin{aligned}
R_{\text{ref}}(T) &= \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{a \in \mathcal{A}} \mathbb{1}\{a(t) = a, \mathcal{B}^c_{a(t)}(t)\} \cdot d_a(t) \middle| \mathcal{E}(T)\right] \\
&\leq \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}\left[\sum_{j=1}^{K} \mathbb{1}\{a(t) = (\cdot, j), \mathcal{B}^c_{a(t)}(t)\} \cdot C_{(j,\emptyset)}(t) \middle| \mathcal{H}_t\right] \middle| \mathcal{E}(T)\right] && (26) \\
&= \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}\left[\sum_{j=1}^{K} \mathbb{1}\{a(t) = (\cdot, j), \mathcal{B}^c_{a(t)}(t)\} \cdot C_{(j,\emptyset)}(t) \cdot \mathbb{E}\left[\frac{\mathbb{1}\{j \in o(t)\}}{p_j(a(t),t)} | \mathcal{H}_t\right] \middle| \mathcal{H}_t\right] \middle| \mathcal{E}(T)\right] && (27) \\
&\leq \frac{1}{\epsilon} \cdot \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}\left[\sum_{j=1}^{K} \mathbb{1}\{a(t) = (\cdot, j), \mathcal{B}^c_{a(t)}(t)\} \cdot C_{(j,\emptyset)}(t) \cdot \mathbb{1}\{j \in o(t)\} \middle| \mathcal{H}_t\right] \middle| \mathcal{E}(T)\right] && (28) \\
&= \frac{1}{\epsilon} \cdot \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}\left[\sum_{j=1}^{K} \mathbb{1}\{j \in o(t), \mathcal{B}^c_{a(t)}(t)\} \cdot \sqrt{\frac{2 \log t}{N_j(t)}} \middle| \mathcal{H}_t\right] \middle| \mathcal{E}(T)\right] \\
&= \frac{\sqrt{2 \log T}}{\epsilon} \cdot \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}\left[\sum_{i=1}^{K} \mathbb{1}\{i \in o(t), \mathcal{B}^c_{a(t)}(t)\} \cdot \sqrt{\frac{1}{N_i(t)}} \middle| \mathcal{H}_t\right] \middle| \mathcal{E}(T)\right] && (29)
\end{aligned}$$

Again Eq. (26) is due to the tower rule. In Eq. (27) we used the fact that given $\mathcal{H}_t$, the probability of $j \in o(t)$ is $p_j(a(t),t)$; and in Eq. (28), we used $p_j(a(t),t) \geq \epsilon$. Recall from Lemma D.1 that for the cases where $r_i(t)$ is not between the values of $\hat{\mu}_j(t)$ and μ_j, $d_{(i,j)}(t) = 0$ will hold since $\mathcal{B}_{(i,j)}(t)$ will happen in these instances. This also entails $d_{(i,j)}(t) = 0$ when $C_{(j,\emptyset)}(t) < \gamma_j$ since it cannot be the case that $\hat{\mu}_i(t) \leq d_j \leq \mu_i$ or $\mu_i \leq d_j \leq \hat{\mu}_i(t)$ when $C_{(j,\emptyset)}(t) < \gamma_j$. Hence, for action (i,j), regret can only be incurred for the rounds where

$$C_{(j,\emptyset)}(t) = \sqrt{\frac{2 \log t}{N_j(t)}} \geq \gamma_j$$

happens. Rearranging the terms,

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

Hence, the event $\mathcal{B}^c_{(i,j)}(t)$ can happen at most $N_j(t) \leq \frac{2 \log T}{\gamma_j^2}$ times in expectation. Using this, the summation index in (29) can be changed from t to $N_j(t)$.

$$R_{\text{ref}}(T) \leq \frac{\sqrt{2 \log T}}{\epsilon} \cdot \sum_{j=1}^{K} \sum_{N_j(t)=1}^{\frac{2 \log T}{\gamma_j^2}} \sqrt{\frac{1}{N_j(t)}}$$

$$
\begin{aligned}
&= \frac{\sqrt{2\log T}}{\epsilon} \cdot \sum_{i=1}^{K} \sum_{x=1}^{\frac{2\log T}{\gamma_i^2}} \cdot \sqrt{\frac{1}{x}} \\
&\leq \sum_{i=1}^{K} \sqrt{2\log T} \left(1 + \int_{1}^{\frac{2\log T}{\gamma_i^2}} \sqrt{\frac{1}{x}} dx \right) \\
&= \sum_{i=1}^{K} \sqrt{2\log T} \left(1 + 2\sqrt{\frac{2\log T}{\gamma_i^2}} - 2 \right) \\
&\leq \sum_{i=1}^{K} \frac{4\log T}{\gamma_i}
\end{aligned}
$$

This completes the proof. □

Appendix E
Proof of Lemma IV.5

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 $\Gamma_{(i,j)} = \max(r_i, \mu_j) - c = r_i \cdot \mathbb{1}\{r_i > \mu_j\} + r_j \cdot \mathbb{1}\{r_i \leq \mu_j\} - c,\ i \neq j$ 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 A.2, 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 base arm i, and if the backup arm is pulled, it also yields in a sample of base 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 base 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 base 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 A.2 can be used to lower bound the total number of samples (sampled or constructed from other samples) of an action a as:

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[s_{(i,j)}(T)\right]}{\log T} \geq \frac{1}{D_{KL}(\Gamma_{(i,j)}, \Gamma^*)} \tag{30}$$

Combining (30) 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

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[s_i(T)\right]}{\log T} \geq \left[\min_{a\in\mathcal{A}_i, a\neq a^*} \{D_{KL}(\Gamma_a||\Gamma^*)\}\right]^{-1} \tag{31}$$

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:

$$s_i(t) \leq \sum_{\substack{j\in[K]\cup\{\emptyset\} \\ (i,j)\neq a^*}} N_{(i,j)}(t) + \sum_{\substack{j=1 \\ j\neq i,\ (j,i)\neq a^*}}^{K} N_{(j,i)}(t) \tag{32}$$

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 (31) and (32), we have

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[\sum_{a\in\mathcal{A}_i, a\neq a^*} N_a(t)\right]}{\log T} \geq \left[\min_{a\in\mathcal{A}_i, a\neq a^*} \{D_{KL}(\Gamma_a||\Gamma^*)\}\right]^{-1} \tag{33}$$

Denoting $\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} = b_a$, (33) 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

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

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 [1], 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:

$$\begin{aligned} &\min_{b_a \geq 0,\ \forall a\in\mathcal{A}\setminus\{a^*\}} \sum_{a\in\mathcal{A}\setminus\{a^*\}} b_a \Delta_a \\ &\text{s.t. } \forall i \in [K], \sum_{a\in\mathcal{A}_i, a\neq a^*} b_a \geq \left[\min_{a\in\mathcal{A}_i, a\neq a^*} \{D_{KL}(\Gamma_a||\Gamma^*)\}\right]^{-1} \end{aligned}$$

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

, $\Gamma_{(i,\emptyset)} = \Gamma_i$, $\Gamma_{(i,j)} = \max(r_i, \mu_j) - c$ 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:

$$\sum_{i=1}^{K} \left(\sum_{a \in \mathcal{A}_i, a \neq a^*} b_a \right) \geq \sum_{i=1}^{K} \left[\min_{a \in \mathcal{A}_i, a \neq a^*} \{D_{KL}(\Gamma_a||\Gamma^*)\} \right]^{-1} \tag{34}$$

We have that

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

We define $D^i_{KL} = \min_{a \in \mathcal{A}_i, a \neq a^*} \{D_{KL}(\Gamma_a||\Gamma^*)\}$. Then, (34) can be rewritten as:

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

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

APPENDIX F
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, $\nu_{(i,j,d_l)} = -c + \mathbb{E}\left[r_i \cdot \mathbb{1}\{r_i \geq d_l\} + r_j \cdot \mathbb{1}\{r_i < d_l\}\right],\ i, j \in [K],\ i \neq j,\ d_l \in \mathcal{D}, l \in [\mathcal{D}] \setminus \{1\}$ (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 $\Delta_{(i,j,d_l)} = \nu^* - \mathbb{E}\left[r_i \cdot \mathbb{1}\{r_i \geq d_l\} + \mu_j \cdot \mathbb{1}\{r_i < d_l\}\right] + c$, 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:

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

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

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

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

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

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:

$$R_U(T) \leq \sum_{a\in\mathcal{A}\setminus\{u^*\}} \mathbb{E}\left[N_a(T)\right] \cdot \Delta_a + \mathbb{E}\left[\sum_{t=|\mathcal{D}|K^2+1}^{T} \mathbb{P}(\mathcal{E}_1^c(t))\right] \tag{35}$$

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

$$\mathbb{E}\left[N_a(T)\right] \leq \frac{8\log T}{\Delta_a^2} \tag{36}$$

We now bound the term $\sum_{t=1}^{T} \mathbb{P}[\mathcal{E}_1^c(t)]$. Note that from (12) and (13), 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

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{P}[\mathcal{E}_1^c(t)]\right] &\leq \sum_{i=1}^{K}\sum_{t=1}^{T} 2t^{-3} + \sum_{l=2}^{|\mathcal{D}|}\sum_{i=1}^{K^2-K}\sum_{t=1}^{T} 2t^{-3} &(37)\\
&= 2((|\mathcal{D}|-1)(K^2-K)+K)\sum_{t=1}^{T} t^{-3}\\
&\leq \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} &(38)
\end{aligned}$$

where the first summation term in the right side of (37) 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 (38), we again use the fact that $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$.

Combining (36) and (38), it can be concluded that

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

□

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 (35), we have

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

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

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

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

$$
\begin{aligned}
R_U(T) &\leq \sum_{a\in\mathcal{A}_{N,1}\setminus\{u^*\}} \frac{8\log T}{\Delta_a} + T\sqrt{\frac{8|\mathcal{D}|K^2\log T}{T}} + \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} + |\mathcal{D}|K^2 \\
&\leq \sum_{a\in\mathcal{A}_{N,1}\setminus\{u^*\}} \sqrt{\frac{8T\log T}{|\mathcal{D}|K^2}} + T\sqrt{\frac{8|\mathcal{D}|K^2\log T}{T}} + \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} + |\mathcal{D}|K^2 \\
&\leq |\mathcal{D}|K^2\cdot\sqrt{\frac{8T\log T}{|\mathcal{D}|K^2}} + \sqrt{8|\mathcal{D}|K^2T\log T} + \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} + |\mathcal{D}|K^2 \\
&\leq 4\sqrt{2|\mathcal{D}|K^2T\log T} + \frac{\pi^2[(|\mathcal{D}|-1)(K^2-K)+K]}{3} + |\mathcal{D}|K^2 \\
&= O(\sqrt{|\mathcal{D}|K^2T\log T}) + O(1)
\end{aligned}
$$

□