Multi-armed Bandits with Missing Outcomes

Ilia Mahrooghi[†1] **Mahshad Moradi**[†1] **Sina Akbari**[1] **Negar Kiyavash**[1]

[1]EPFL, Switzerland

Abstract

While significant progress has been made in designing algorithms that minimize regret in online decision-making, real-world scenarios often introduce additional complexities, with missing outcomes perhaps among the most challenging ones. Overlooking this aspect or simply assuming random missingness invariably leads to biased estimates of the rewards and may result in linear regret. Despite the practical relevance of this challenge, no rigorous methodology currently exists for systematically handling missingness, especially when the missingness mechanism is not random. In this paper, we address this gap in the context of multi-armed bandits (MAB) with missing outcomes by analyzing the impact of different missingness mechanisms on achievable regret bounds. We introduce algorithms that account for missingness under both missing at random (MAR) and missing not at random (MNAR) models. Through both analytical and simulation studies, we demonstrate the drastic improvements in decision-making by accounting for missingness in these settings.

1 INTRODUCTION

Multi-armed bandit (MAB) algorithms have emerged as indispensable tools for decision-making under uncertainty, balancing the trade-off between exploring different options and exploiting the best-known action. These algorithms have achieved success in various domains ranging from personalized online advertisement and recommender systems [Li et al., 2010, Xu et al., 2020, Ban et al., 2024] to clinical trials [Villar et al., 2015, Aziz et al., 2021, Varatharajah and Berry, 2022] and adaptive routing in communication systems [Maghsudi and Hossain, 2016, Li et al., 2020a]. For instance, in online advertising, advertisers need to continuously select which ad to show to a user to maximize click-through rates. Similarly, in clinical trials, researchers must decide which treatment to administer to patients to optimize recovery rates. MAB algorithms guide decision-makers in such scenarios to learn actions that minimize regret.

Significant progress has been made in developing MAB algorithms that minimize regret in various settings [Lai and Robbins, 1985, Auer et al., 2002, Bubeck et al., 2012, Lattimore and Szepesvári, 2020, Slivkins, 2019]. However, the real world often introduces challenges that deviate from the assumptions of the classical MAB framework or its current extensions. One of the most critical challenges is that of missing outcomes – situations where the results of certain actions are not always observed. This challenge arises more often than not in practice and can fundamentally undermine the decision-making process if left unaddressed. To illustrate this, consider an example of a large-scale clinical trial for a new cancer treatment. Patients are randomly assigned to different treatment arms, and their health outcomes are monitored over time. In practice, not all patients will complete the trial. Some may drop out early due to side effects, while others may stop reporting outcomes for personal reasons, and some could pass away during the trial due to reasons not related to the treatment (competing events). Crucially, the missingness of the outcome may not be random. Patients experiencing severe side effects or poor health are more likely to drop out, meaning that the missingness mechanism is correlated with the unobserved outcome itself. This introduces systematic bias into the estimation of the rewards, and if not accounted for, would lead to poor decision-making.

The issue of missingness is not confined to healthcare. In a recommendation system that suggests articles to users on an online platform, if users who find the content irrelevant are less likely to provide feedback (e.g., they leave the site without interacting), the system could overestimate the value of the recommended articles, assuming that the

[†]This work was completed during the authors' tenure as research interns at EPFL.

missing feedback is independent of user satisfaction. Here too, missingness is correlated with the unobserved outcome, leading to biased reward estimates and sub-optimal recommendations.

The problem of missing data is a fundamental challenge in causal inference. This issue has been extensively studied over the past decades, with seminal works such as [Rubin, 1976, Little and Rubin, 2019, Bang and Robins, 2005] laying the foundation for dealing with biased estimations in the presence of missing data. These methods, along with more recent developments in graphical models for handling missing data [Mohan and Pearl, 2021, Nabi et al., 2020], have become standard approaches in causal inference. Missing data has also been extensively explored in specific contexts such as instrumental variables, [Tchetgen Tchetgen and Wirth, 2017, Sun et al., 2018, Kennedy et al., 2019] and mediation analysis [Zhang and Wang, 2013, Zhang et al., 2015, Kidd et al., 2023], among others. By contrast, the challenge of missing outcomes has received relatively little attention in multi-armed bandit problems, although some progress has been made in related areas. For instance, the problem of delayed feedback in bandits bears some similarity to our setting, as both involve incomplete information at decision time. Several works have addressed stochastic bandits with unrestricted delays [Joulani et al., 2013, Vernade et al., 2017], and delays dependent on stochastic rewards [Pike-Burke et al., 2018, Lancewicki et al., 2021]. In contextual bandits, [Bouneffouf et al., 2017] studied linear contextual bandits with missing (restricted) contexts. While this work addresses missing data in bandits, it focuses on missing contexts rather than outcomes and assumes a linear reward model. Others have explored bandit problems with variable costs or restricted observations. For example, [Ding et al., 2013, Seldin et al., 2014] studied MAB problems with variable costs, where the outcome is observable only after paying the associated cost.

There are two lines of research closely related to our work. The first includes works such as [Chen et al., 2022] and [Bouneffouf et al., 2020], which consider the problem of MAB with missing outcomes. [Chen et al., 2022] settles for some empirical considerations and suggestions, without formally studying the problem or providing tailored algorithms. [Bouneffouf et al., 2020] employs unsupervised learning techniques to impute the missing rewards in a contextual bandit setting. Both of these works assume that the missingness mechanism is random, possibly after conditioning on the context. In this paper, we do a thorough study from a formal perspective, characterizing the best achievable regret bounds under multiple scenarios with missing outcomes. We also provide novel regret lower bounds and algorithms that are guaranteed to achieve optimal regrets. The second line of related research concerns bandits with graph feedback [Mannor and Shamir, 2011], where pulling an arm provides feedback about the rewards of other connected arms, where connections are represented by a graph structure. Typically, these models assume that each arm has a self-loop, ensuring its own reward is always observed Mannor and Shamir [2011], Alon et al. [2017], Li et al. [2020b], Cortes et al. [2020], Dai et al. [2024]. Esposito et al. [2022] extended this framework by allowing for missing self-loops, aligning with the missing outcome setting. However, their model assumes that missingness depends only on the chosen action, whereas we explicitly analyze cases where the missingness mechanism is outcome-dependent. We are able to achieve unbiased estimates of the expected rewards in this setting through using a mediator variable as auxiliary information.

Addressing the problem of missing outcomes is both practically relevant and theoretically challenging. In applications such as healthcare, education, and e-commerce, accounting for missing data could lead to better treatment policies, more personalized learning experiences, and more effective product recommendations, potentially affecting millions of individuals. In this paper, we undertake the first formal study of multi-armed bandits with missing outcomes and provide tailored algorithms that explicitly handle different types of missingness. Our main contributions are two-fold. First, we provide an analysis of the impact of missing outcomes on achievable regret (the loss of optimality). Second, we introduce provably good upper confidence bound (UCB) algorithms that are tailored to handle both missing at random and missing not at random mechanisms. Our algorithms are designed to adjust reward estimates based on the observed data and the missingness mechanism, ensuring unbiased estimation. Finally, we extend our analysis to settings where not only outcomes but also mediators (e.g., users providing feedback) are prone to missingness, to further broaden the applicability of our approach.

The remainder of this paper is structured as follows. In Section 2, we review the relevant background and formalize the problem of multi-armed bandits with missing outcomes. In Section 3 we present our algorithms in the settings of missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), respectively. Additionally, we provide the corresponding achievable regret lower bounds. The technical proofs are postponed to Appendix B due to space limitations. In Section 4, we extend our approach and present algorithms for the case when the mediator is also prone to missingness. A discussion of the limitations of our work and our concluding remarks appear in Section 7.

2 FORMALIZATION AND PROBLEM SETUP

We begin by reviewing the classic multi-armed bandit (MAB) setup and then extend it to incorporate missing outcomes. The MAB problem involves an agent (decision-maker) who interacts with an environment over a sequence of T time steps. At each time step $t \in \{1, \ldots, T\}$, the agent pulls an arm a_t from a set of n available actions indexed by $\mathcal{A} = \{1, \ldots, n\}$. Upon pulling this arm, the agent receives a stochastic reward $Y_t \in \mathcal{Y}$ drawn from a fixed (but unknown) probability distribution associated with arm a_t. The goal of the agent is to minimize the *cumulative regret* over the time horizon T, which is defined as the cumulative difference between the rewards of the optimal arm and the chosen arms. Specifically, let $\mu_a := \mathbb{E}[Y \mid A = a]$ for every $a \in \mathcal{A}$. The optimal arm, denoted by a^*, is the arm that maximizes the expected reward, i.e., $a^* := \arg\max_{a \in \mathcal{A}} \mu_a$. The regret at time t is defined as $R_t := \mu_{a^*} - \mathbb{E}[Y \mid A = a_t]$, and the cumulative regret over T rounds, denoted by R_T, is the sum of the latter instantaneous regrets over the horizon T:

$$R_T := \sum_{t=1}^{T} (\mu_{a^*} - \mathbb{E}[Y \mid A = a_t]) \tag{1}$$

In the classical setting, it is assumed that after pulling an arm a_t, the agent always observes the true reward Y_t without any missingness.

We extend the classic MAB model to accommodate missingness. We assume that pulling each arm $a_t \in \mathcal{A}$, draws a stochastic tuple $\left(Y_t, O_t^Y, M_t, O_t^M\right)$ from a fixed but unknown probability distribution associated with arm a_t. In this tuple, $Y_t \in \mathcal{Y}$ represents the true reward (as before), whereas $O_t^Y \in \{0, 1\}$ is an indicator denoting whether this reward is observed. $M_t \in \mathcal{M}$ is a possible mediator[1] or an auxiliary variable, with $O_t^M \in \{0, 1\}$ indicating whether this auxiliary variable is observed.

For example, in online recommendations, auxiliary information could include metrics such as the time a user spends on a webpage before navigating away, or other data points gathered from browser cookies, such as past browsing behavior, device type, or location. The agent has access to the 'observed' tuple $\left(Y_t^o, O_t^Y, M_t^o, O_t^M\right)$, where the observed values Y_t^o and M_t^o are defined as follows:

$$Y_t^o = \begin{cases} Y_t; & \text{if } O_t^Y = 1, \\ ?; & \text{o.w.} \end{cases}, M_t^o = \begin{cases} M_t; & \text{if } O_t^M = 1, \\ ?; & \text{o.w.} \end{cases}, \tag{2}$$

where ? denotes a missing value. We define μ_a as the expected value of Y_t given $A_t = a$ as before, with the crucial difference that samples of Y_t are missing when $O_t^Y = 0$.

Clearly, without imposing further structure, it is not possible to construct unbiased estimators for the expected rewards of each arm. In fact, these expectations are not 'identifiable,' meaning that they are not uniquely determinable functionals of the probability measure over observable variables. In what follows, we begin with the case where the mediator is fully observed ($O_t^M = 1$ with probability 1). We first consider the case where the missingness mechanism of the outcome is independent of everything else. Subsequently, we analyze the more realistic cases where this missingness is correlated with the missing outcome Y_t. Later in 4 we extend our findings further to the case where even the mediator is prone to missingness.

[1]Our use of the term 'mediator' is broader than in traditional causal inference. In this context, it refers to any auxiliary variable potentially correlated with the reward or the missingness mechanism, not necessarily one on a specific causal pathway. The inclusion of this variable is also without loss of generality, as it can be a degenerate variable ($M_t \equiv 0$) that carries no information.

3 MAB WITH MISSING OUTCOME

Throughout, we assume that the outcomes are not 'always missing.'

Assumption 1 (Positivity). *For every action $a \in \mathcal{A}$ and mediator $m \in \mathcal{M}$, $\mathbb{P}\big(O_t^Y = 1 \mid M_t = m, A_t = a\big) > 0$. Moreover, $\mathbb{P}(M_t \mid A_t)$ is positive everywhere*[2].

Assumption 1 is reasonable as otherwise there exists an arm for which the agent observes no reward samples. For the rest of this section, we assume that the auxiliary variable M_t is always observed.

3.1 MISSING COMPLETELY AT RANDOM (MCAR)

We begin with the case where the outcome missingness mechanism is independent of the other variables (including the outcome itself). This case is studied for the sake of completeness, and we acknowledge that, unlike the other cases to follow, it can be accommodated by most existing approaches.

Assumption 2 (MCAR). *The outcome is missing completely at random. That is, $O_t^Y \perp\!\!\!\perp (A_t, Y_t, M_t)$ for $t \in \{1, \ldots, T\}$.*

This assumption holds, for instance, when data gets erased by say an independent mechanism such as a power outage. The graph of Figure 1a represents this missingness mechanism, whereby the missingness indicator O^Y is an isolated node. As there is no information conveyed by the missingness indicator, the missing chunk of the data can be discarded without any need for extra care. As such, the classic upper confidence bound (UCB) algorithms are expected to achieve (near-)optimal regret. We formalize these claims through the next two propositions. For the sake of completeness, we have included the adapted UCB algorithm (Alg. 1) for this scenario in Appendix A. Let $\gamma = \mathbb{P}(O_t^Y = 1)$ be the probability of observing the output in each round.

[2]With sufficient caution, the second part of this assumption could be omitted. However, we include it here for the sake of simplicity in the presentation.

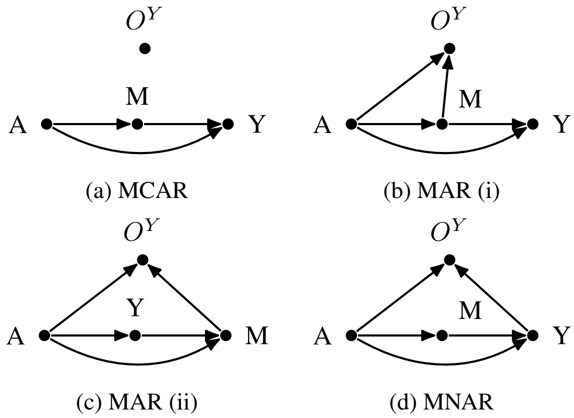


Figure 1: Graphical representations of the missing data mechanisms considered in this paper.

Theorem 1. *(MCAR regret guarantee) Under Assumption2, for every $\alpha > 1$, the cumulative regret of the adapted UCB (Alg. 1) is bounded as follows:*

$$\mathbb{E}[R_T] = O\left(\sqrt{\frac{\alpha n T \log(T)}{\gamma}}\right).$$

The proof of Theorem 1, which provides a regret bound similar to that of the classic UCB algorithm, but adapted to our setting, is included in Appendix B. The following result indicates that this regret bound is (near-)optimal.

Theorem 2. *(Minimax lower bound for MCAR) For any policy π, there exists an MCAR instance ν s.t.*

$$\mathbb{E}[R_T(\pi,\nu)] = \Omega\left(\sqrt{\frac{nT}{\gamma}}\right),$$

where $\mathbb{E}[R_T(\pi,\nu)]$ represents the expected regret of policy π in instance ν.

See Appendix B for the proof of Theorem 2 as well as the rest of the results of this paper.

3.2 MISSING AT RANDOM (MAR)

We now focus on more realistic settings where the missingness mechanism provides information about the missing outcomes. This is the case, for instance, when the unsatisfied customers are more likely to leave comments on an online platform, or in health applications, the patients with severe side effects are more likely to drop out of the study. We first consider the case when missingness is at random, i.e., independent of Y given M and A. The graphs of Figure 1b and Figure 1c illustrate two possible representations of the MAR mechanism, under which Assumption 3 holds.

Assumption 3 (MAR). $O_t^Y \perp\!\!\!\perp Y_t \mid (A_t, M_t)$ *for* $t \in \{1,\dots,T\}$.

Under Assumption 3, the expected reward is identifiable as follows:

$$\begin{aligned}\mu_a = \mathbb{E}[Y_t \mid A_t = a] &= \mathbb{E}\big[\mathbb{E}[Y_t \mid M, a] \mid A_t = a\big] \\ &\overset{(a)}{=} \mathbb{E}\big[\mathbb{E}[Y_t \mid M, a, O_t^Y = 1] \mid A_t = a\big] \qquad (3) \\ &\overset{(b)}{=} \mathbb{E}\big[\mathbb{E}[Y_t^o \mid M, a, O_t^Y = 1] \mid A_t = a\big],\end{aligned}$$

where (a) follows from Assumption 3 and (b) holds due to consistency (see Equation 2.)

Accordingly, we will use the following estimator for μ_a:

$$\hat{\mu}_a = \frac{1}{|T_a|}\sum_{t\in T_a}\Big(\sum_{m\in\mathcal{M}}\Big(\frac{\mathbb{1}\{M_t = m\}}{|T_{m,a,o}|}\sum_{t'\in T_{m,a,o}} Y_{t'}^o\Big)\Big), \quad (4)$$

where $T_a, T_{m,a,o} \subseteq \{1,\dots,T\}$ are the sets of iterations where $A_t = a$, and iterations where $A_t = a$, $M_t = m$ and $O_t^Y = 1$, respectively. In what follows, for brevity, we define $p_{m,a} := \mathbb{P}(M_t = m \mid A_t = a)$. To build intuition for the general case, we first consider the simplified theoretical setting where the conditional probabilities $p_{m,a}$ are known. We then adapt our algorithm to the case where these probabilities are unknown. Recall that $n = |\mathcal{A}|$ is the number of arms. We assume that $\mathbb{E}[Y_t \mid m, a] \in [0,1]$ for all arms and that the reward Y_t is sub-Gaussian. Algorithm 2 presents the pseudo-code for the first case. The algorithm is based on UCB, but with an initial step where the agent pulls each arm $\log(T)^2$ times. At the subsequent rounds, both the expected rewards and the associated confidence bounds are estimated based on Equation (4). In order to present the regret bounds, we need the following definitions. Let $P_a = \sum_{m\in\mathcal{M}} \frac{p_{m,a}}{\gamma_{m,a}}$ where $\gamma_{m,a} = \mathbb{P}(O^Y = 1 \mid m, a)$. Further, define S and H as the arithmetic mean and the harmonic mean of the P_a values, respectively:

$$S := \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} P_a, \quad H := \frac{|\mathcal{A}|}{\sum_{a\in\mathcal{A}}\frac{1}{P_a}}.$$

Theorem 3. *(Regret guarantee for Alg. 2) Under Assumption 3, for every $\alpha > 1$, there exists a constant c such that the following regret bound holds for $T \geq c$:*

$$\mathbb{E}[R_T] = O\left(\sqrt{\alpha T \log(T) n S}\right).$$

Next, we show that Algorithm 2 can be adapted to the case where the conditional probabilities $p_{m,a}$ are not known and must be estimated – see Algorithm 3. The following theorem shows that this algorithm achieves the same regret bound as Algorithm 2, i.e., the estimation of $p_{m,a}$ does not affect the cumulative regret.

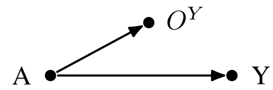


Figure 2: Special case of MAR.

Theorem 4. *(Regret guarantee for Alg. 3) Under Assumptions 3, for every $\alpha > 1$, there exists a constant c such that the following regret bound holds for $T \geq c$:*

$$\mathbb{E}[R_T] = O\left(\sqrt{\alpha T \log(T) n S}\right).$$

Since P_a is a weighted average of $1/\gamma_{m,a}$ over $m \in \mathcal{M}$ (with weights $p_{m,a}$), the regret bounds depend not on the cardinality of the mediator set, $|\mathcal{M}|$, but rather on the heterogeneity of the $\gamma_{m,a}$ values. The following theorem provides the minimax lower bound, demonstrating near-optimality of Algorithms 2 and 3.

Theorem 5. *(Minimax lower bound for MAR) For any policy π, there exists a MAR instance ν such that:*

$$\mathbb{E}[R_T(\pi, \nu)] = \Omega\left(\sqrt{TnH}\right).$$

Note that when $\gamma_{m,a}$ values are identical (and equal to γ) then S and H coincide. Further, the upper and lower bounds in this case match those of MCAR.

A special case of the MAR environment (depicted in Figure 2) pertains to when there is no mediator. In this case, Assumption 3 reduces to the following:

Assumption 4. $O_t^Y \perp\!\!\!\perp Y_t \mid A_t$ *for all* $t \in \{1, \dots, T\}$.

Theorems 3 and 5 with a degenerate mediator ($|\mathcal{M}| = 1$) imply the following corollary.

Corollary 1. *Under Assumption 4, Algorithm 3 induces cumulative regret $\mathbb{E}[R_T] = O\left(\sqrt{\alpha T \log(T) n S}\right)$ and the cumulative regret of any policy is lower bounded by $\mathbb{E}[R_T] = \Omega\left(\sqrt{TnH}\right)$, where $S = \frac{\sum_a \frac{1}{\gamma_a}}{n}$ and $H = \frac{n}{\sum_a \gamma_a}$.*

Discussion 1. We used estimators that explicitly use the mediator values in this section. As we pointed out earlier, the size of $\mathcal{M}$ (the alphabet of M) does not affect the regret bounds. However, one might wonder whether the use of the mediator can be avoided, resulting in simpler algorithms and/or estimation schemes. We show next that any such algorithm can induce linear regret in the worst case. As a corollary, this result implies that naïvely employing the classical UCB algorithm also induces linear regret.

Theorem 6. *For any mediator-agnostic policy π (a policy that does not have access to mediator values), there exists a MAR instance ν which satisfies Assumption 3 and its regret grows linearly*

$$\mathbb{E}[R_T(\pi, \nu)] = \Omega(T).$$

Discussion 2. The expected reward μ_a can also be estimated using a Horvitz-Thompson (HT) type estimator [Horvitz and Thompson, 1952]. Specifically, the conditional expectation terms in Equation (3) can be expressed as follows:

$$\mathbb{E}[Y_t^o \mid m, a, O_t^Y = 1] = \mathbb{E}[\frac{Y_t^o \mathbb{1}\{M_t = m, O_t^Y = 1\}}{p_{m,a}\gamma_{m,a}} \mid A_t = a],$$

and after plugging it into Eq. (3),

$$\mu_a = \mathbb{E}[\sum_{m \in \mathcal{M}} \frac{Y_t^o \mathbb{1}\{M_t = m, O_t^Y = 1\}}{\gamma_{m,a}} \mid A_t = a]. \quad (5)$$

An estimator based on the latter does not require estimating the conditional outcome means (in contrast to Eq. 3), but it rather needs the estimates of the missingness probability $\gamma_{m,a}$. Using such an estimator is particularly beneficial when the missingness probabilities are known in advance, or a parametric model can be justified. However, if the missingness probabilities are small or estimated imprecisely, the HT estimator can exhibit high variance, leading to instability. One can take a step further and construct augmented inverse propensity weighted (AIPW) estimators for μ_a:

$$\mu_a = \mathbb{E}\Big[\sum_{m \in \mathcal{M}} \frac{\mathbb{1}\{M_t = m\}}{\gamma_{m,a}}\big(Y_t^o \mathbb{1}\{O_t^Y = 1\} - (\mathbb{1}\{O_t^Y = 1\} - \gamma_{m,a})\mathbb{E}[Y_t^o \mid m, a, O_t^Y = 1]\big) \mid A_t = a\Big],$$

which is doubly robust (DR) in the sense that it is consistent as long as either the missingness probabilities $\gamma_{m,a}$ or the conditional outcome means $\mathbb{E}[Y_t^o \mid m, a, O_t^Y = 1]$ (but not necessarily both) can be consistently estimated. We prove this claim formally in Appendix B for the sake of completeness. In this paper, we consider discrete-valued mediators, and estimate all the quantities of interest through empirical means. Therefore, all three estimators (outcome-based, HT, and DR) coincide. However, the HT and DR estimators can prove beneficial for extending our approach to incorporate continuous mediators, or in problems with high-dimensional actions and/or mediators where (semi)parametric models can help improve estimation efficiency.

3.3 MISSING NOT AT RANDOM (MNAR)

Finally, we consider the case where the missingness mechanism directly depends on the outcome value Y. Here, we follow the identification strategy of [Zuo et al., 2024] for MNAR. However, we are interested only in identifying the expected rewards, rather than conducting mediation analysis. We begin with the following assumption.

Assumption 5 (MNAR). $O_t^Y \perp\!\!\!\perp M_t \mid (A_t, Y_t)$ *for* $t \in \{1, \ldots, T\}$.

In other words, the missingness is independent of the mediator when conditioned on the action and the actual outcome. Figure 1d graphically represents this scenario. This situation commonly arises in environments where the reward is missing due to its value. For example, if the outcome of interest is the income of an individual, they may not be inclined to report it if the value is too high or too low.

We further make the following assumption, which is the minimal assumption required for identifiability.

Assumption 6 (Completeness). *The distribution* $\mathbb{P}(M, Y, O^Y = 1 \mid a)$ *is complete in* M*, that is, for any* $a \in \mathcal{A}$*, and for any function* $g : \mathcal{Y} \to \mathbb{R}$*,*

$$\int_{y \in \mathcal{Y}} \mathbb{P}(M, Y = y, O^Y = 1 \mid a) g(y)\, dy = 0$$

implies that $g(Y) = 0$ *with probability one.*

Below we show how μ_a is identified under these assumptions. The identification strategy outlined here is akin to [Zuo et al., 2024].

$$\begin{aligned}\mathbb{P}(m, O^Y = 0 \mid a) &= \int_{y \in \mathcal{Y}} \mathbb{P}(m, y, O^Y = 0 \mid a)\, dy \\ &\overset{(a)}{=} \int_{y \in \mathcal{Y}} \mathbb{P}(m, y, O^Y = 1 \mid a) \frac{\mathbb{P}(O^Y = 0 \mid y, a, m)}{\mathbb{P}(O^Y = 1 \mid y, a, m)}\, dy \\ &\overset{(b)}{=} \int_{y \in \mathcal{Y}} \mathbb{P}(m, y, O^Y = 1 \mid a) \frac{\mathbb{P}(O^Y = 0 \mid y, a)}{\mathbb{P}(O^Y = 1 \mid y, a)}\, dy,\end{aligned}$$

where (a) and (b) follow from Bayes' rule and Assumption 5, respectively. Since $\mathbb{P}(M, Y, O^Y = 1 \mid a)$ is complete in M, solving this integral equation uniquely determines the inverse odds ratio $\mathrm{OR}_{y,a} = \frac{\mathbb{P}(O^Y=0|y,a)}{\mathbb{P}(O^Y=1|y,a)}$, allowing us to identify $\mathbb{P}(y \mid a)$ as follows:

$$\begin{aligned}\mathbb{P}(y \mid a) &= \sum_{m \in \mathcal{M}} \mathbb{P}(y, m \mid a) = \sum_{m \in \mathcal{M}} \frac{\mathbb{P}(y, m \mid O^Y = 1, a)}{\mathbb{P}(O^Y = 1 \mid y, a)} \\ &= \sum_{m \in \mathcal{M}} (1 + \mathrm{OR}_{y,a}) \mathbb{P}(y, m \mid O^Y = 1, a).\end{aligned} \tag{6}$$

Finally, $\mu_a = \mathbb{E}[Y_t \mid A_t = a]$ is identified as $\mu_a = \int_{y \in \mathcal{Y}} y \mathbb{P}(y \mid a)\, dy$.

In the remainder of this section, we assume Y is discrete with $|\mathcal{Y}| = L$, and the outcomes are normalized so that $\sum_{y \in \mathcal{Y}} |y| = 1$. Define $K = |\mathcal{M}|$, and $\Theta_a = [\mathbb{P}(m, y, O^Y = 1 \mid a)]_{K \times L}$. Additionally, we assume that these matrices are not ill-conditioned.[3]

[3] A problem is considered ill-conditioned if small changes to the input can cause large changes in the output solution. Bounding the condition number ensures the problem is well-conditioned.

Assumption 7. *[Bounded condition number] For each arm* $a \in \mathcal{A}$*, the condition number of* Θ_a *is bounded by:*

$$\kappa(\Theta_a) \leq C_a,$$

where $\kappa(\Theta_a)$ *denotes the condition number of* Θ_a *with respect to* ∞*-norm, defined as* $\kappa(\Theta_a) = \|\Theta_a\|_\infty \|\Theta_a^\dagger\|_\infty$*, with* $\Theta_a^\dagger$ *being the pseudo-inverse of* Θ_a*.*

We present Algorithm 4 for minimizing cumulative regret under the MNAR assumptions. The key intuition behind this algorithm is to construct an estimator based on Eq. (6) and build upper confidence bounds under Assumption 7. In order to present the regret bound of this algorithm, define $p_{y,a} = \mathbb{P}(Y = y \mid A = a)$, and $\gamma_{y,a} = \mathbb{P}(O^Y = 1 \mid Y = y, A = a)$.

Theorem 7. *(Regret guarantee for Alg. 4) Under Assumptions 5, 6, and 7, for every* $\alpha > 1$*, there exists a constant* c *such that the following regret bound holds for* $T \geq c$*:*

$$\mathbb{E}[R_T] = O\Big(\sqrt{\alpha T \log(T) \sum_a S_a^2}\Big),$$

with $S_a = \max\{\frac{LC_a}{\gamma_a \|\Theta_a\|_\infty}, \frac{K}{\gamma_a \sqrt{\sum_{y \in \mathbb{Y}} p_{y,a}\gamma_{y,a}}}\}, \gamma_a = \min_y \gamma_{y,a}$*.*

Remark 1. *With* $\gamma_{\min} = \min_{y,a} \gamma_{y,a}$ *and* $\kappa_{\max} = \max_a \frac{C_a}{\|\Theta_a\|_\infty}$*, Theorem 7 implies the following bound:*

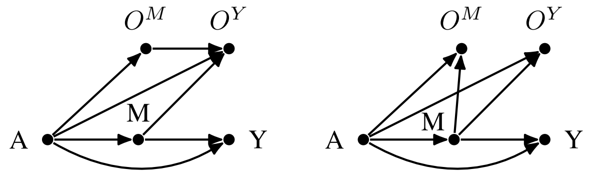
$$\mathbb{E}[R_T] = O\Big(\sqrt{\alpha T \log(T) N \max\{\frac{L\kappa_{\max}}{\gamma_{\min}}, \frac{K}{\gamma_{\min}^{3/2}}\}^2}\Big).$$

4 MAB WITH MISSING OUTCOME AND MEDIATOR

So far we considered cases where the mediator was fully observable. We now discuss how our results extend to scenarios involving missing data in both Y and M. Here, we assume that the outcome is MAR, and discuss the cases where the mediator is MAR and MNAR separately. For the case where both outcome and mediator are MNAR, refer to Appendix C. We begin by outlining each scenario, providing identification schemes and estimators for μ_a. The corresponding algorithms, theoretical results, and proofs are postponed to Appendix C. Throughout this section, we work under Assumption 3.

4.1 MAR

Since the mediator values are missing, neither the conditional outcome means nor the probabilities $p_{m,a}$ are identifiable. We require further structure to make progress. One such structure is when the mediator missingness can be assumed to be at random, i.e., $O_t^M \perp\!\!\!\perp (M_t, Y_t, O_t^Y) \mid A_t$.



(a) MAR (b) MNAR

Figure 3: Graphical representations of the missing data mechanisms with missing outcome and mediator.

This assumption is valid for instance when the missingness mechanism for the mediator depends only on the action. A less stringent alternative can be formalized as:

Assumption 8. *$O_t^M \perp\!\!\!\perp M_t \mid A_t$, and $O_t^M \perp\!\!\!\perp Y_t \mid (A_t, M_t, O_t^Y)$ for all $t \in \{1, \ldots, T\}$.*

See Fig. 3a for a graph representation satisfying Assumptions 3 and 8. Under these two assumptions, analogous to Eq. (3), μ_a can be identified as follows.

$$\begin{aligned}\mu_a &= \mathbb{E}\big[\mathbb{E}[Y_t \mid M, a, O_t^Y = 1] \mid A_t = a\big] \\ &= \mathbb{E}\big[\mathbb{E}[Y_t^o \mid M, a, O_t^Y = 1, O_t^M = 1] \mid A_t = a, O_t^M = 1\big],\end{aligned}$$

where the second equation is due to Assumption 8.

4.2 MNAR

When the mediator is missing not at random, stronger assumptions are necessary to identify the expected rewards. Analogous to Section 3.3, we will use a completeness assumption. Here too, the identification strategy follows the approach of [Zuo et al., 2024].

Assumption 9. *Y_t, O_t^Y and O_t^M are mutually independent conditioned on (A_t, M_t) for all $t \in \{1, \ldots, T\}$.*

Assumption 10. *For every $a \in \mathcal{A}, m \in \mathcal{M}$, $\mathbb{P}(M = m, Y, O^M = 1, O^Y = 1 \mid a)$ is complete in Y. That is, for any function $g : \mathcal{Y} \to \mathbb{R}$,*

$$\int_{\mathcal{Y}} \mathbb{P}(M = m, Y = y, O^M = 1, O^Y = 1 \mid a) g(y) dy = 0$$

implies $g(Y) = 0$ with probability one.

Under Assumption 9, μ_a can be expressed as:

$$\begin{aligned}\mu_a &= \sum_{m \in \mathbb{M}} \mathbb{E}[Y \mid a, m] p_{m,a} \\ &= \sum_{m \in \mathbb{M}} \mathbb{E}[Y \mid a, m, O^Y = 1, O^M = 1] p_{m,a}.\end{aligned}$$

To proceed, we need to identify $p_{m,a} = \mathbb{P}(M = m \mid A = a)$. This is achieved through Assumption 10:

$$\begin{aligned}&\mathbb{P}(Y = y, O^M = 0, O^Y = 1 \mid a) \\ &= \sum_{m \in \mathbb{M}} \mathbb{P}(M = m, Y = y, O^M = 0, O^Y = 1 \mid a) \\ &= \sum_{m \in \mathbb{M}} \mathbb{P}(M = m, Y = y, O^M = 1, O^Y = 1 \mid a) \\ &\quad \times \frac{\mathbb{P}(O^M = 0 \mid M = m, A = a)}{\mathbb{P}(O^M = 1 \mid M = m, A = a)},\end{aligned}$$

where we used Assumption 9 in the last equation. By Assumption 10, the inverse odds ratios $\mathrm{OR}_{m,a} = \frac{\mathbb{P}(O^M = 0 \mid m, a)}{\mathbb{P}(O^M = 1 \mid m, a)}$ can be uniquely determined. Finally,

$$\begin{aligned}p_{m,a} &= \frac{\mathbb{P}(M = m, O^M = 1 \mid A = a)}{\mathbb{P}(O^M = 1 \mid A = a, M = m)}, \\ &= (1 + \mathrm{OR}_{m,a}) \mathbb{P}(M = m, O^M = 1 \mid A = a).\end{aligned}$$

We use a two-step estimation process, whereby in the first step, $\hat{p}_{m,a}$ is estimated, and in the second step, the expected reward is estimated as

$$\hat{\mu}_a = \sum_{m \in \mathcal{M}} \hat{p}_{m,a} \hat{\mu}_{m,a}$$

where $\hat{\mu}_{m,a}$ is the empirical mean of the samples Y_t, obtained after pulling arm a, conditioned on $M_t = m$ with both $O^M = 1$ and $O^Y = 1$. Here, we require Y_t to be finite-valued, analogous to Section 3.

5 SUMMARY OF ASSUMPTIONS AND RESULTS

To facilitate comparison, the following table summarizes the core assumptions and regret bounds for the MCAR, MAR, and MNAR frameworks under the setting where the mediator variable, M, is always observed.

Setting	Core Assumptions	Regret Bounds
MCAR	$O_t^Y \perp\!\!\!\perp (A_t, Y_t, M_t)$	**Upper (Alg. 1):** $O\left(\sqrt{\frac{\alpha n T \log(T)}{\gamma}}\right)$ **Lower:** $\Omega\left(\sqrt{\frac{nT}{\gamma}}\right)$
MAR	$O_t^Y \perp\!\!\!\perp Y_t \mid (A_t, M_t)$	**Upper (Alg. 3):** $O\left(\sqrt{\alpha T \log(T) n S}\right)$ **Lower:** $\Omega\left(\sqrt{TnH}\right)$
MNAR	$O_t^Y \perp\!\!\!\perp M_t \mid (A_t, Y_t)$ **Completeness** (6) **Condition number** (7)	**Upper (Alg. 4):** $O\left(\sqrt{\alpha T \log(T) \sum_a S_a^2}\right)$

6 EMPIRICAL EVALUATION

Here, we provide an empirical evaluation of our MAB algorithms across different missing data scenarios – MCAR, MAR(i,ii), and MNAR. All our simulations were run on Google Colab[4] with Intel Xeon CPUs. Python implementations for reproducing the results of this paper are available on GitHub[5].

Python code to reproduce our results is attached as supplementary material. We model the MAB environment in all the aforementioned settings with $n = 10$ arms. More comprehensive simulation results are provided in Appendix D.

6.1 EXPERIMENT SETUP OF MCAR

Each arm $a \in \{1, \ldots, n\}$ has an associated mean reward μ_a, sampled independently from a uniform distribution over the interval $[0, 1]$. The observation probability γ is randomly drawn from a uniform distribution over $[0.5, 1.0]$. At each time t, when arm a is pulled, the reward Y_t is generated from a normal distribution $\mathcal{N}(\mu_a, 1)$. Algorithm 1's performance is reported across 20 independent runs in the MCAR environment over a time horizon of $T = 10{,}000$ iterations, with a fixed parameter $\alpha = 2$. Figure 4a depicts the cumulative regret for different γ values. As expected, when γ decreases, the regret grows more rapidly as a consequence of lower observation likelihood.

6.2 EXPERIMENT SETUP OF MAR

The MAB environment is modeled with $n = 10$ arms but with $K = 5$ possible mediator values. The expected reward for all arms is determined by $\{\mu_{m,a}\}_{m,a} \in \mathbb{R}^{n\times K}$, where $\mu_{m,a}$ represents the mean reward for arm a when the mediator takes value m. The latter reward matrix is chosen by sampling each $\mu_{m,a}$ independently from a uniform distribution over $[0, 0.4]$. To ensure the first arm is the optimal one, an additional 0.6 is added to its corresponding mean. The observation mechanism is defined by a matrix $\{\gamma_{m,a}\}_{m,a} \in \mathbb{R}^{n\times K}$, where each $\gamma_{m,a}$ is sampled independently from a uniform distribution over $[0.8, 1]$.

For each arm a, a categorical probability distribution $\{p_{m,a}\}_m \in \mathbb{R}^K$ is defined over the K values of M. This distribution is drawn from a Dirichlet distribution, i.e., $\{p_{m,a}\}_m \sim$ Dirichlet$(\mathbf{1}_K)$. Upon pulling arm a and the mediator taking value m, reward Y_t is drawn from a normal distribution $\mathcal{N}(\mu_{m,a}, 1)$, where $\mu_{m,a}$ is the mean reward for arm a when mediator takes value m. The reward is observed with probability $\gamma_{m,a}$. We ran Algorithms 3 and 4 over a time horizon of $T = 100,000$. Their cumulative regret was averaged across 10 independent runs. As shown in Fig. 4b., knowing conditional probabilities $p_{m,a}$ in advance improves the cumulative regret, as expected.

Fig. 4c demonstrates the average cumulative regret of the MAR algorithm with different probability distributions over the mediator. In particular, two mediator value selection strategies were tested: (i) uniform, where each mediator value has an equal probability, and (ii) a peaked distribution, where one mediator per arm has a higher probability, using a Dirichlet distribution biased by $\alpha = 5$ for the chosen mediator. The peaked distribution results in a higher cumulative regret, which aligns with the result from Theorem 5, since S is maximized when the probability distribution is concentrated on the largest $\gamma_{m,a}$.

In Figure 4d, we compare the performance of the UCB and MAR algorithms in the MAR bandit environment. The results illustrate that the cumulative regret of the UCB algorithm is consistently higher than that of the MAR algorithm. Notably, the regret of the UCB algorithm exhibits a near-linear growth as a result of the bias in its estimation of the reward. This bias is due to the failure to account for the mediator structure. In contrast, the MAR algorithm, which explicitly utilizes mediators to handle missingness, achieves accurate reward estimation and a significantly lower regret.

6.3 EXPERIMENT SETUP OF MNAR

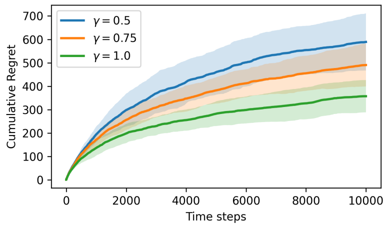
The MNAR algorithm was evaluated in an environment with $n = 10$ arms, $K = 5$ mediators, and $|\mathcal{Y}| = 5$ possible outcomes, over a horizon of $T = 100{,}000$, repeated 10 times. For each arm a and mediator m, the reward function followed a categorical distribution sampled from a Dirichlet distribution, except for one arm which was sampled from a biased Dirichlet distribution. The bias was applied to the largest $y \in \mathcal{Y}$, ensuring that this arm had a higher expected reward. The observation probabilities $\gamma_{y,a}$ were drawn from a uniform distribution over $[0.5, 1.0]$, while the mediator probabilities were sampled from a Dirichlet distribution. Fig. 4e shows that the algorithm successfully adapts to the MNAR setup, effectively minimizing the cumulative regret.

7 LIMITATIONS AND CONCLUDING REMARKS

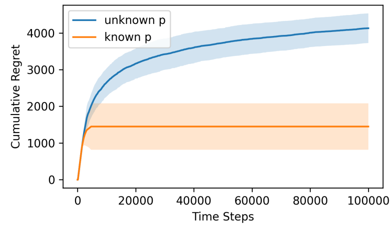
We studied multi-armed bandits with missing outcomes and adapted UCB algorithms to incorporate missingness. Our approaches extend the applicability of MAB algorithms to a wider range of real-world online decision-making problems. We expect that the insights given by this paper will help researchers to develop and adapt other existing decision-making algorithms to take missingness into account. We assumed that the auxiliary (mediator) M takes values in

[4]https://colab.google

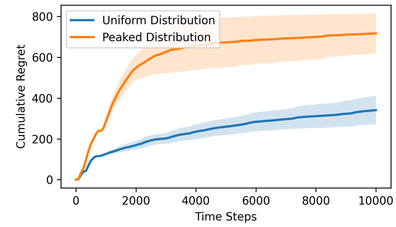
[5]https://github.com/ilia-mahrooghi/Multi-armed-Bandits-with-Missing-Outcome



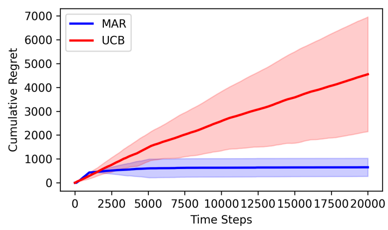
(a) MCAR algorithm on MCAR bandit with various γ values.



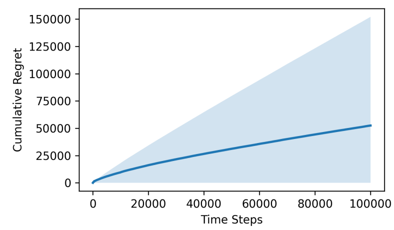
(b) MAR algorithms with known and unknown p for comparison.



(c) MAR algorithm with different p initializations on MAR environment.



(d) MAR and UCB algorithms in the MAR bandit environment.



(e) Performance of the MNAR algorithm in the described environment.

Figure 4: Results corresponding to MCAR, MAR, and MNAR settings. The shaded regions represent the error bars, showing one standard deviation across multiple runs of the simulations.

a finite set. Parametric (or semiparametric) models can be adopted to relax this assumption in the future. We further acknowledge that estimating the odds ratios through integral equations in the MNAR setting presents significant challenges, both in terms of computational complexity and sample efficiency. Hence, we have postponed the problem of MAB with continuous outcomes missing not at random to future work.

A practical challenge arises when choosing between several plausible settings for the missingness mechanism. To address this model selection problem, dynamic balancing offers a principled solution Cutkosky et al. [2021]. The technique involves running an instance of our algorithm for each candidate setting and using a meta-algorithm to dynamically arbitrate between them based on performance. This approach is analogous to recent methods for handling model uncertainty in causal bandits Liu et al. [2024]. Its inclusion extends our framework to a more robust version capable of automatically selecting the most appropriate model, enhancing its practical applicability

ACKNOWLEDGMENTS

This research was in part supported by the Swiss National Science Foundation under NCCR Automation II, grant agreement 51NF40_225155.

References

Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002. doi: 10.1023/A:1013689704352. URL `https://doi.org/10.1023/A:1013689704352`.

Maryam Aziz, Emilie Kaufmann, and Marie-Karelle Riviere. On multi-armed bandit designs for dose-finding trials. *Journal of Machine Learning Research*, 22(14):1–38, 2021.

Yikun Ban, Yunzhe Qi, and Jingrui He. Neural contextual bandits for personalized recommendation. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1246–1249, 2024.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Djallel Bouneffouf, Irina Rish, Guillermo A Cecchi, and Raphaël Féraud. Context attentive bandits: Contextual bandit with restricted context. *arXiv preprint arXiv:1705.03821*, 2017.

Djallel Bouneffouf, Sohini Upadhyay, and Yasaman Khazaeni. Contextual bandit with missing rewards. *arXiv preprint arXiv:2007.06368*, 2020.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Xijin Chen, Kim May Lee, Sofia S Villar, and David S Robertson. Some performance considerations when using multi-armed bandit algorithms in the presence of missing data. *Plos one*, 17(9):e0274272, 2022.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Ningshan Zhang. Online learning with dependent stochastic feedback graphs. In *International Conference on Machine Learning*, pages 2154–2163. PMLR, 2020.

Ashok Cutkosky, Christoph Dann, Abhimanyu Das, Claudio Gentile, Aldo Pacchiano, and Manish Purohit. Dynamic balancing for model selection in bandits and rl. In *International Conference on Machine Learning*, pages 2276–2285. PMLR, 2021.

Jessica Dai, Bailey Flanigan, Meena Jagadeesan, Nika Haghtalab, and Chara Podimata. Can probabilistic feedback drive user impacts in online platforms? In *International Conference on Artificial Intelligence and Statistics*, pages 2512–2520. PMLR, 2024.

Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. Multi-armed bandit with budget constraint and variable costs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 232–238, 2013.

Emmanuel Esposito, Federico Fusco, Dirk van der Hoeven, and Nicolò Cesa-Bianchi. Learning on the edge: Online learning with stochastic feedback graphs. *Advances in Neural Information Processing Systems*, 35:34776–34788, 2022.

Nicholas J Higham. *A survey of componentwise perturbation theory*, volume 48. American Mathematical Society, 1994.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483446`.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International conference on machine learning*, pages 1453–1461. PMLR, 2013.

Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Conference on Learning Theory*, pages 1066–1100. PMLR, 2015.

Edward H Kennedy, Jacqueline A Mauro, Michael J Daniels, Natalie Burns, and Dylan S Small. Handling missing data in instrumental variable methods for causal inference. *Annual review of statistics and its application*, 6(1):125–148, 2019.

John Kidd, Chelsea K Raulerson, Karen L Mohlke, and Dan-Yu Lin. Mediation analysis of multiple mediators with incomplete omics data. *Genetic epidemiology*, 47 (1):61–77, 2023.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tal Lancewicki, Shahar Segal, Tomer Koren, and Yishay Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In *International Conference on Machine Learning*, pages 5969–5978. PMLR, 2021.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Feng Li, Dongxiao Yu, Huan Yang, Jiguo Yu, Holger Karl, and Xiuzhen Cheng. Multi-armed-bandit-based spectrum scheduling algorithms in wireless networks: A survey. *IEEE Wireless Communications*, 27(1):24–30, 2020a.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Shuai Li, Wei Chen, Zheng Wen, and Kwong-Sak Leung. Stochastic online learning with probabilistic graph feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4675–4682, 2020b.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Ziyi Liu, Idan Attias, and Daniel M Roy. Causal bandits: The pareto optimal frontier of adaptivity, a reduction to linear bandits, and limitations around unknown marginals. *arXiv preprint arXiv:2407.00950*, 2024.

Setareh Maghsudi and Ekram Hossain. Multi-armed bandits with application to 5g small cells. *IEEE Wireless Communications*, 23(3):64–73, 2016.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. *Advances in Neural Information Processing Systems*, 24, 2011.

Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021.

Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. Full law identification in graphical models of missing data: Completeness results. In *International conference on machine learning*, pages 7153–7163. PMLR, 2020.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113. PMLR, 2018.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Yevgeny Seldin, Peter Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *International Conference on Machine Learning*, pages 280–287. PMLR, 2014.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2): 1–286, 2019. doi: 10.1561/2200000068.

BaoLuo Sun, Lan Liu, Wang Miao, Kathleen Wirth, James Robins, and Eric J Tchetgen Tchetgen. Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28(4):1965, 2018.

Eric J Tchetgen Tchetgen and Kathleen E Wirth. A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics*, 73(4): 1123–1131, 2017.

Yogatheesan Varatharajah and Brent Berry. A contextual-bandit-based approach for informed decision-making in clinical trials. *Life*, 12(8):1277, 2022.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. *arXiv preprint arXiv:1706.09186*, 2017.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2): 199, 2015.

Xiao Xu, Fang Dong, Yanghua Li, Shaojian He, and Xin Li. Contextual-bandit based personalized recommendation with time-varying user interests. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6518–6525, 2020.

Zhiyong Zhang and Lijuan Wang. Methods for mediation analysis with missing data. *Psychometrika*, 78:154–184, 2013.

Zhiyong Zhang, Lijuan Wang, and Xin Tong. Mediation analysis with missing data through multiple imputation and bootstrap. In *Quantitative Psychology Research: The 79th Annual Meeting of the Psychometric Society, Madison, Wisconsin, 2014*, pages 341–355. Springer, 2015.

Shuozhi Zuo, Debashis Ghosh, Peng Ding, and Fan Yang. Mediation analysis with the mediator and outcome missing not at random. *Journal of the American Statistical Association*, pages 1–21, 2024.

Appendix

This appendix is organized as follows. Section A includes the omitted algorithms referred to in the main text. Section B includes the technical proofs of our results. Section C provides our algorithm for the case where both the outcome and the mediator are missing not at random (MNAR) along with the regret analysis and proofs. Finally, Section D includes additional empirical evaluation results.

A MAIN ALGORITHMS

Algorithm 1 MCAR algorithm

Input: Number of arms n, time horizon T, $\alpha \geq 1$
Initialize: $\hat{\mu}_a = 0$ for all arms $a = 1, 2, \ldots, n$ ▷ initial mean reward estimate for each arm
Set: $T_{a,o} = 0$ for all arms $a = 1, 2, \ldots, n$ ▷ number of times each arm is pulled and reward is observed
for each round $t = 1, 2, \ldots, T$ **do**
  for each arm $a = 1, 2, \ldots, n$ **do**
    $\text{UCB}_a(t) = \hat{\mu}_a + \sqrt{\frac{\alpha \log(T)}{2T_{a,o}}}$
  end for
  Select arm $a_t = \arg\max_a \text{UCB}_a(t)$
  Pull arm a_t and observe reward r_t
  if reward is observed **then**
    Update T_{a_t} and $\hat{\mu}_{a_t}$
  end if
end for

Algorithm 2 MAR Algorithm with known $p_{m,a}$

Input: Number of arms n, time horizon T, exploration parameter α
Initialize:
for each arm $a \in [n]$ and $m \in \mathbb{M}$ **do**
  $\hat{\mu}_{m,a} = 0$ ▷ estimated mean reward for arm a when $M = m$
  $T_{m,a,o} = 0$ ▷ number of times arm a is pulled with $M = m$ and reward observed
  $T_{m,a} = 0$ ▷ number of times $M = m$ was observed for arm a
end for
for each arm $a \in [n]$ **do**
  for $\log(T)^2$ rounds **do**
    Pull arm a, observe m and reward r
    Update $T_{m,a}$ for observed $M = m$
    if reward is observed **then**
      Update $T_{m,a,o}$ and $\hat{\mu}_{m,a}$
    end if
  end for
end for
$T_1 = n\log(T)^2$, $T_2 = T - T_1$
for each round $t = 1, \ldots, T_2$ **do**
  for each arm $a \in [n]$ **do**
    Compute $\hat{\mu}_a = \sum_{m \in [K]} p_{m,a}\hat{\mu}_{m,a}$ ▷ estimated mean reward for arm a
    Compute $\text{UCB}_a(t) = \hat{\mu}_a + \sqrt{\frac{\alpha \log(T)}{2} \sum_{m \in \mathbb{M}} \frac{p_{m,a}^2}{T_{m,a,o}}}$ ▷ Upper Confidence Bound for arm a
  end for
  Select arm $a_t = \arg\max_a \text{UCB}_a(t)$
  Pull arm a_t, observe m and reward r_t
  Update T_{m,a_t} and, if reward is observed, update $T_{m,a_t,o}$ and $\hat{\mu}_{m,a_t}$
end for

Algorithm 3 MAR Algorithm with unknown $p_{m,a}$

Input: Number of arms n, time horizon T, exploration parameter α
Initialize:
for each arm $a \in [n]$ and $m \in \mathbb{M}$ **do**
    $\hat{\mu}_{m,a} = 0$ ▷ estimated mean reward for arm a when $M = m$
    $T_{m,a,o} = 0$ ▷ number of times arm a is pulled with $M = m$ and reward observed
    $T_{m,a} = 0$ ▷ number of times $M = m$ was observed for arm a
end for
for each arm $a \in [n]$ **do**
    for $\log(T)^2$ rounds **do**
        Pull arm a, observe m and reward r
        Update $T_{m,a}$ for observed $M = m$
        if reward is observed **then**
            Update $T_{m,a,o}$ and $\hat{\mu}_{m,a}$
        end if
    end for
end for
$T_1 = n \log(T)^2$, $T_2 = T - T_1$
for each round $t = 1, \ldots, T_2$ **do**
    for each arm $a \in [n]$ **do**
        Estimate $\hat{p}_{m,a} = \frac{T_{m,a}}{T_a}$ for each $m \in \mathbb{M}$
        Compute $\hat{\mu}_a = \frac{1}{T_a} \sum_t \sum_{m \in \mathcal{M}} \hat{\mu}_{m,a} \mathbb{1}\{M_t = m\} \mathbb{1}\{A_t = a\}$
        Compute $\mathrm{UCB}_a(t) = \hat{\mu}_a + 8\sqrt{\frac{\alpha \log(T)}{2} \sum_{m \in \mathbb{M}} \frac{\hat{p}^2_{m,a}}{T_{m,a,o}}}$
    end for
    Select arm $a_t = \arg\max_a \mathrm{UCB}_a(t)$
    Pull arm a_t, observe m and reward r_t
    Update T_{m,a_t} and, if reward is observed, update $T_{m,a_t,o}$ and $\hat{\mu}_{m,a_t}$
end for

Algorithm 4 MNAR Algorithm

Input: Number of arms n, time horizon T, exploration parameter α
Initialize:
for each arm $a \in [n]$ and $m \in \mathbb{M}$ **do**
  Set $b_{m,0|a} = 0$ ▷ Estimation of $\mathbb{P}(M = m, O^Y = 0 \mid a)$
  Set $\hat{\theta}_a = [0]_{k \times L}$ ▷ Estimation of matrix $\theta_a[m, y] = \mathbb{P}(m, y, O^Y = 1 \mid a)$
  Set $q_{m,y|1,a} = 0$ ▷ Estimation of $\mathbb{P}(M = m, Y = y \mid a, O^Y = 1)$
  Set $T_a = 0$ ▷ Count of pulls of arm a
  Set $T_{a,o} = 0$ ▷ Count of pulls of arm a with observed reward
end for
for each arm $a \in [n]$ **do**
  for $\log(T)^2$ rounds **do**
    Pull arm a, observe mediator m and reward y
    Update T_a
    if reward is observed **then**
      Update $T_{a,o}$, $\hat{\theta}_a[m, y]$, and $q_{m,y|1,a}$
    else
      Update $b_{m,0|a}$
    end if
  end for
end for
Set $T_1 = n \log(T)^2$ and $T_2 = T - T_1$
for each round $t = 1, \dots, T_2$ **do**
  for each arm $a \in [n]$ **do**
    Solve $x_a = \hat{\theta}_a^{\dagger} b_a$
    Update $x_a = x_a + [1]_{L \times 1}$
    Compute $\hat{p}(m, y) = x_a[y] \times q_{m,y|1,a}$
    Compute $\hat{p}(y) = \sum_{m \in \mathbb{M}} \hat{p}(m, y)$
    Compute $\hat{\mu}_a = \sum_{y \in \mathbb{Y}} y \times \hat{p}(y)$
    Compute $\hat{\gamma}_a = \frac{1}{\max_{y \in \mathbb{Y}}(x_a[y])}$
    Compute $\text{UCB}_a(t) = \hat{\mu}_a + 8 \frac{L C_a}{\|\hat{\theta}_a\|_\infty \hat{\gamma}_a} \sqrt{\frac{\alpha \log(T)}{T_a}} + \frac{K}{\hat{\gamma}_a} \sqrt{\frac{\alpha \log(T)}{T_{a,o}}}$
  end for
  Select arm $a_t = \arg\max_a \text{UCB}_a(t)$
  Pull arm a_t, observe m and reward y_t
  Update T_{a_t}
  if reward is observed **then**
    Update $T_{a_t,o}$, $\hat{\theta}_{a_t}[m, y]$, and $q_{m,y|1,a_t}$
  else
    Update $b_{m,0|a_t}$
  end if
end for

B TECHNICAL PROOFS

Double Robustness of the AIPW estimator. Following Discussion 2 in Section 3.2, let $\hat{\gamma}_{m,a}$ and $\hat{\mu}_{m,a}$ be models for $\gamma_{m,a}$ and $\mathbb{E}[Y_t^o \mid m, a, O_t^Y = 1]$, respectively. Define

$$\hat{\mu}_a = \mathbb{E}\big[\sum_{m\in\mathcal{M}} \frac{\mathbb{1}\{M_t = m\}}{\hat{\gamma}_{m,a}}\big(Y_t^o \mathbb{1}\{O_t^Y = 1\} - (\mathbb{1}\{O_t^Y = 1\} - \hat{\gamma}_{m,a})\hat{\mu}_{m,a}\big) \mid A_t = a\big], \tag{7}$$

as an estimator for μ_a of Eq. (3.2). Herein, we prove that $\hat{\mu}_a$ is *doubly robust*, in the sense that if either of the missingness probability models ($\hat{\gamma}_{m,a}$) or the outcome regression models ($\hat{\mu}_{m,a}$), but not necessarily both, are correctly specified, then $\hat{\mu}_a$ of Eq. (7) is consistent for μ_a of Eq. (3.2). We discuss the two cases separately:

Case (i): the missingness probabilities are correctly specified; i.e., $\hat{\gamma}_{m,a} = \gamma_{m,a}$. In this case,

$$\begin{aligned}
&\mathbb{E}[\sum_{m\in\mathcal{M}} (\frac{\mathbb{1}\{O_t^Y = 1\}}{\hat{\gamma}_{m,a}} - 1)\hat{\mu}_{m,a}\mathbb{1}\{M_t = m\} \mid A_t = a] \\
&\overset{(a)}{=} \sum_{m\in\mathcal{M}} \mathbb{E}[(\frac{\mathbb{1}\{O_t^Y = 1\}}{\gamma_{m,a}} - 1)\hat{\mu}_{m,a}\mathbb{1}\{M_t = m\} \mid A_t = a] \\
&\overset{(b)}{=} \sum_{m\in\mathcal{M}} \mathbb{E}[(\frac{\mathbb{1}\{O_t^Y = 1\}}{\gamma_{m,a}} - 1) \mid A_t = a, M_t = m]\hat{\mu}_{m,a}p_{m,a} \\
&\overset{(c)}{=} \sum_{m\in\mathcal{M}} (\frac{\gamma_{m,a}}{\gamma_{m,a}} - 1)\hat{\mu}_{m,a}p_{m,a} \\
&= 0,
\end{aligned}$$

where (a) is due to $\hat{\gamma}_{m,a}$ being correctly specified, (b) is an application of the law of total expectation, and (c) is by definition of $\gamma_{m,a} = \mathbb{E}[O_t^Y \mid A_t = a, M_t = m]$. As a result, we get

$$\hat{\mu}_a = \mathbb{E}\big[\sum_{m\in\mathcal{M}} \frac{\mathbb{1}\{M_t = m\}}{\gamma_{m,a}} Y_t^o \mathbb{1}\{O_t^Y = 1\} \mid A_t = a\big],$$

which matches Eq. (5), and therefore $\hat{\mu}_a$ is consistent for μ_a.

Case (ii): the outcome regression models are correctly specified; i.e., $\hat{\mu}_{m,a} = \mathbb{E}[Y_t^o \mid m, a, O_t^Y = 1]$. Then,

$$\begin{aligned}
&\mathbb{E}\big[\sum_{m\in\mathcal{M}} \frac{\mathbb{1}\{M_t = m\}}{\hat{\gamma}_{m,a}}(Y_t^o - \hat{\mu}_{m,a})\mathbb{1}\{O_t^Y = 1\} \mid A_t = a\big] \\
&\overset{(a)}{=} \sum_{m\in\mathcal{M}} \mathbb{E}\big[\frac{\mathbb{1}\{O_t^Y = 1\}}{\hat{\gamma}_{m,a}}(Y_t^o - \hat{\mu}_{m,a}) \mid A_t = a, M_t = m\big]p_{m,a} \\
&\overset{(b)}{=} \sum_{m\in\mathcal{M}} \frac{\gamma_{m,a}}{\hat{\gamma}_{m,a}}\mathbb{E}\big[Y_t^o - \hat{\mu}_{m,a} \mid A_t = a, M_t = m, O_t^Y = 1\big]p_{m,a} \\
&\overset{(c)}{=} \sum_{m\in\mathcal{M}} \frac{\gamma_{m,a}}{\hat{\gamma}_{m,a}}\big(\mathbb{E}\big[Y_t^o \mid A_t = a, M_t = m, O_t^Y = 1\big] - \hat{\mu}_{m,a}\big)p_{m,a} \\
&\overset{(d)}{=} 0,
\end{aligned}$$

where (a) and (b) are due to the law of total expectations, (c) is by linearity of expectation, and (d) follows from the correctness of $\hat{\mu}_{m,a}$. From Eq. (7),

$$\begin{aligned}
\hat{\mu}_{m,a} &= \mathbb{E}[\sum_{m\in\mathcal{M}} \mathbb{1}\{M_t = m\}\hat{\mu}_{m,a} \mid A_t = a] \\
&= \mathbb{E}[\sum_{m\in\mathcal{M}} \mathbb{1}\{M_t = m\}\mathbb{E}[Y_t^o \mid m, a, O_t^Y = 1] \mid A_t = a] \\
&= \mathbb{E}[\mathbb{E}[Y_t^o \mid M, a, O_t^Y = 1] \mid A_t = a],
\end{aligned}$$

which matches Eq. (3), and therefore $\hat{\mu}_{m,a}$ is consistent for $\mu_{m,a}$. $\square$

Theorem 1. *(MCAR regret guarantee) Under Assumption2, for every $\alpha > 1$, the cumulative regret of the adapted UCB (Alg. 1) is bounded as follows:*

$$\mathbb{E}[R_T] = O\left(\sqrt{\frac{\alpha n T \log(T)}{\gamma}}\right).$$

Proof. Let $a^* = \arg\max_a \mu_a$ be the optimal arm. Using Hoeffding's inequality, we can derive the following bounds for any time step $1 \leq t \leq T$:

- If $a = a_t = \arg\max_a (\text{UCB}_a)$, we have:

$$|\hat{\mu}_a - \mu_a| \leq \sqrt{\frac{\alpha \log(t)}{2T_{a,o}}},$$

with probability $1 - 2t^{-\alpha}$. Name this "good event" A_t.

Now, define $\epsilon_a = \sqrt{\frac{\alpha \log(t)}{2T_{a,o}}}$. For $a = a_t = \arg\max_a (\text{UCB}_a)$, we get the following inequality:

$$\mu_a + 2\epsilon_a \geq \hat{\mu}_a + \epsilon_a = \text{UCB}_a \geq \text{UCB}_{a^*} = \hat{\mu}_{a^*} + \epsilon_{a^*} \geq \mu_{a^*} \quad \Rightarrow \quad \epsilon_a \geq \frac{\Delta_a}{2}, \tag{8}$$

where $\Delta_a = \mu_{a^*} - \mu_a$.

Now, if E_t represents the "good events" at time step t, then under $E = \bigcap_t E_t$, using (8) we obtain:

$$T_{a,o} \leq 4\alpha \log(T)\Delta_a^{-2}.$$

Thus, we have:

$$\begin{aligned}
\mathbb{E}[T_{a,o}] &= \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}(I_t = a, O_t^Y = 1)] \\
&\leq 4\alpha \log(T)\Delta_a^{-2} + \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}(E_t^c)] \\
&= 4\alpha \log(T)\Delta_a^{-2} + \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}((A_t^c))] \\
&\leq 4\alpha \log(T)\Delta_a^{-2} + \sum_{t=1}^{T} 2t^{-\alpha} \\
&\leq 4\alpha \log(T)\Delta_a^{-2} + \frac{2\alpha}{\alpha - 1}.
\end{aligned} \tag{9}$$

Since we observe the reward with probability γ, and $O^Y \perp\!\!\!\perp (A, Y)$, we have $\mathbb{E}[T_{a,o}] = \gamma \mathbb{E}[T_a]$. Therefore:

$$\mathbb{E}[T_a] \leq \frac{4\alpha \log(T)\Delta_a^{-2} + \frac{2\alpha}{\alpha-1}}{\gamma}.$$

Let $x = \sqrt{\frac{4\alpha n \log(T)}{T\gamma}}$. Then, we have:

$$
\begin{aligned}
\mathbb{E}[R_T] &= \sum_a \Delta_a \mathbb{E}[T_a] \\
&= \sum_{\Delta_a < x} \Delta_a \mathbb{E}[T_a] + \sum_{\Delta_a \geq x} \Delta_a \mathbb{E}[T_a] \\
&\leq Tx + \sum_{\Delta_a \geq x} \Delta_a \frac{4\alpha \log(T)\Delta_a^{-2} + \frac{2\alpha}{\alpha-1}}{\gamma} \\
&= Tx + \frac{4n\alpha \log(T)}{x\gamma} + \frac{2n\alpha}{(\alpha-1)\gamma} \\
&= 2\sqrt{\frac{4n\alpha T \log(T)}{\gamma}} + \frac{2n\alpha}{\gamma(\alpha-1)} \qquad (10) \\
&= O\left(\sqrt{\frac{\alpha n T \log(T)}{\gamma}}\right) \qquad (11)
\end{aligned}
$$

□

Theorem 2. *(Minimax lower bound for MCAR) For any policy π, there exists an MCAR instance ν s.t.*

$$\mathbb{E}[R_T(\pi, \nu)] = \Omega\left(\sqrt{\frac{nT}{\gamma}}\right),$$

where $\mathbb{E}[R_T(\pi, \nu)]$ represents the expected regret of policy π in instance ν.

Proof. Consider the following $n+1$ bandit instances, with n arms labeled $a_1, a_2, \ldots, a_n$. The reward distribution for each arm follows a Normal distribution with a variance of 1.

Bandit instance 0**:**

- $\mathbb{E}[Y(a)] = 0$ for all $a = a_1, \ldots, a_n$.

Bandit instance k **for** $k = 1, \ldots, n$**:**

- $\mathbb{E}[Y(a_k)] = \Delta$ for $a = a_k$.
- $\mathbb{E}[Y(a)] = 0$ for $a \neq a_k$.

Next, we present key lemmas adapted from Lattimore and Szepesvári [2020] to complete our analysis.

Divergence Decomposition: Let $\nu = (P(1), \ldots, P(k))$ and $\nu' = (P'(1), \ldots, P'(k))$ represent the reward distributions for two k-armed bandits. For a fixed policy π, let $P_\nu = P_{\nu,\pi}$ and $P_{\nu'} = P_{\nu',\pi}$ be the probability measures induced by the n-round interaction with ν and ν'. Then:

$$\mathrm{KL}(P_\nu, P_{\nu'}) = \sum_{i=1}^{k} \mathbb{E}_\nu[T_i(n)] \mathrm{KL}(P(i), P'(i)).$$

Pinsker's Inequality: For measures P and Q on the same probability space $(\Omega, \mathcal{F})$, the total variation distance is bounded by:

$$d_{\mathrm{TV}}(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)| \leq \sqrt{\frac{1}{2}\mathrm{KL}(P, Q)}.$$

Total Variation Bound: Let $(\Omega, \mathcal{F})$ be a measurable space, and let P and Q be probability measures on $\mathcal{F}$. For any $\mathcal{F}$-measurable random variable $X : \Omega \to [a, b]$, we have:

$$\left| \int_\Omega X(\omega) dP(\omega) - \int_\Omega X(\omega) dQ(\omega) \right| \leq (b - a) d_{\mathrm{TV}}(P, Q).$$

Now, in our setup with missing observations, so (O^Y, Y) represent the observation tuple. Hence, we have:

$$\mathrm{KL}(P_0, P_i) = \mathbb{E}_0[T_i]\mathrm{KL}(P_0(i), P_i(i)) = \mathbb{E}_0[T_i]\frac{\gamma\Delta^2}{2}.$$

From this, we can bound $\mathbb{E}_i[T_i(T)]$ as follows:

$$\begin{aligned}
\mathbb{E}_i[T_i(T)] &\leq \mathbb{E}_0[T_i(T)] + T d_{\mathrm{TV}}(P_0(i), P_i(i)) \\
&\leq \mathbb{E}_0[T_i(T)] + T\sqrt{\frac{1}{2}\mathrm{KL}(P_0(i), P_i(i))} \\
&= \mathbb{E}_0[T_i(T)] + T\sqrt{\frac{1}{2}\cdot\frac{\gamma\Delta^2}{2}\mathbb{E}_0[T_i(T)]} \\
&= \mathbb{E}_0[T_i(T)] + \frac{T}{2}\sqrt{\gamma\Delta^2\mathbb{E}_0[T_i(T)]}.
\end{aligned}$$

Let $R_i = R_T(\pi; i)$ denote the regret of applying policy π on the i-th bandit instance up to time T, where i refers to the i-th bandit instance.

Summing over all bandit instances, we have:

$$\begin{aligned}
\sum_{i=1}^{n}\mathbb{E}[R_i] &= \sum_{i=1}^{n}\Delta(T - \mathbb{E}_i[T_i(T)]) \\
&\geq \Delta T n - \Delta\sum_{i=1}^{n}\left(\mathbb{E}_0[T_i(T)] + \frac{T}{2}\sqrt{\gamma\Delta^2\mathbb{E}_0[T_i(T)]}\right) \\
&\geq \Delta T n - \Delta T - \frac{\Delta^2 T}{2}\sqrt{\gamma T n} \\
&\geq \frac{\Delta T n}{2} - \frac{\Delta^2 T}{2}\sqrt{\gamma T n} \quad \text{using } \Delta = \frac{n}{2\sqrt{\gamma T n}} \\
&\geq \frac{T n^2}{8\sqrt{\gamma T n}} = \frac{n}{8}\sqrt{\frac{Tn}{\gamma}}.
\end{aligned}$$

Thus, there exists an instance where $\mathbb{E}[R_i] \geq \Omega\left(\sqrt{\frac{Tn}{\gamma}}\right)$. $\square$

Theorem 3. *(Regret guarantee for Alg. 2) Under Assumption 3, for every $\alpha > 1$, there exists a constant c such that the following regret bound holds for $T \geq c$:*

$$\mathbb{E}[R_T] = O\left(\sqrt{\alpha T \log(T) n S}\right).$$

Proof. As before, let $a^* = \arg\max_a \mu_a$ denote the optimal arm, and define $T_1 = \sum_a T_{1,a}$ as the total number of times the agent samples each arm during the initial rounds. After the first T_1 rounds, we can derive the following bounds at any time step $1 \leq t \leq T_2 = T - T_1$.

For each arm a, let the reward samples observed when $M = m$ be denoted by $Y^o_{m,a}(1), \ldots, Y^o_{m,a}(T_{m,a,o})$. Applying Hoeffding's inequality, we obtain:

$$\mathbb{P}\left(\left|\sum_{m\in[K]} p_{m,a}\frac{\sum_{j=1}^{T_{m,a,o}} Y^o_{m,a}(j)}{T_{m,a,o}} - \mu_a\right| \geq \epsilon_a\right) \leq 2\exp\left(-\frac{2\epsilon_a^2}{\sum_{m\in[K]}\frac{p_{m,a}^2}{T_{m,a,o}}}\right)$$

This result holds because the sub-Gaussian norm of the random variable $p_{m,a}\frac{Y_{m,a}(j)}{T_{m,a,o}}$ is $\frac{p_{m,a}}{T_{m,a,o}}$.

By setting $\epsilon_a = \sqrt{\frac{\alpha \log(t)}{2} \sum_{m\in[K]} \frac{p_{m,a}^2}{T_{m,a,o}}}$, we obtain the following inequality, which holds with probability at least $1-2t^{-\alpha}$:

$$|\hat{\mu}_a - \mu_a| \leq \sqrt{\frac{\alpha \log(t)}{2} \sum_{m\in[K]} \frac{p_{m,a}^2}{T_{m,a,o}}}$$

Name the above “good event" $A_{t,a}$.

Also, like before for $a = a_t = \arg\max_a (\text{UCB}_a)$, we get the following inequality:

$$\mu_a + 2\epsilon_a \geq \hat{\mu}_a + \epsilon_a = \text{UCB}_a \geq \text{UCB}_{a^*} = \hat{\mu}_{a^*} + \epsilon_{a^*} \geq \mu_{a^*} \quad \Rightarrow \quad \epsilon_a \geq \frac{\Delta_a}{2}, \tag{12}$$

where $\Delta_a = \mu_{a^*} - \mu_a$.

Next, let $T_{m,a}$ represent the number of times arm a is pulled and $M = m$ is observed, and let T_a represent the total number of times arm a is pulled. Using Hoeffding's inequality, we can bound the deviation between $p_{m,a}$ (the probability of observing $M = m$) and the empirical ratio $\frac{T_{m,a}}{T_a}$ as follows:

$$p_{m,a} - \frac{T_{m,a}}{T_a} \leq \sqrt{\frac{\alpha \log(t)}{2T_a}} \leq \sqrt{\frac{\alpha \log(T)}{2T_a}},$$

with probability at least $1 - t^{-\alpha}$. Name this “good event" $B_{t,m,a}$.

Similarly, we bound the deviation between $\gamma_{m,a}$ and $\frac{T_{m,a,o}}{T_{m,a}}$, where $T_{m,a,o}$ is the number of times reward is observed for arm a and $M = m$:

$$\gamma_{m,a} - \frac{T_{m,a,o}}{T_{m,a}} \leq \sqrt{\frac{\alpha \log(t)}{2T_{m,a}}} \leq \sqrt{\frac{\alpha \log(T)}{2T_{m,a}}},$$

again with probability $1 - t^{-\alpha}$. Name this “good event" $C_{t,m,a}$.

For sufficiently large T, we have:

$$\log(T)^2 \geq 2\alpha \log(T) \frac{1}{p^2},$$

which implies $T_a \geq \log(T)^2 \geq 2\alpha \log(T) \frac{1}{p^2}$. This allows us to use inequality $\sqrt{\frac{\alpha \log(T)}{2T_a}} \leq \frac{p_{m,a}}{2}$ to derive a lower bound for $T_{m,a}$:

$$T_{m,a} \geq \frac{T_a p_{m,a}}{2}.$$

Furthermore, since $T_{m,a} \geq \frac{T_a p_{m,a}}{2}$ and For sufficiently large T, we know that $T_a \geq T_{1,a} = \log(T)^2 \geq 2\alpha \log(T) \frac{2}{\gamma_{m,a}^2 p}$. This gives us $\sqrt{\frac{\alpha \log(T)}{2T_{m,a}}} \leq \frac{\gamma_{m,a}}{2}$, which allows us to establish a lower bound for $T_{m,a,o}$:

$$T_{m,a,o} \geq \frac{\gamma_{m,a}}{2} T_{m,a}.$$

Combining this with $T_{m,a} \geq \frac{T_a p_{m,a}}{2}$, we derive:

$$T_{m,a,o} \geq \frac{T_a p_{m,a} \gamma_{m,a}}{4}.$$

Let E_t represent the intersection of "good events" at time step t. Under $E = \bigcap_t E_t$, we obtain:

$$
\begin{aligned}
\epsilon_a &= \sqrt{\frac{\alpha \log(t)}{2} \sum_{m \in [K]} \frac{p_{m,a}^2}{T_{m,a,o}}} \\
&\leq \sqrt{\frac{\alpha \log(T)}{2} \sum_{m \in [K]} \frac{4p_{m,a}^2}{T_a p_{m,a} \gamma_{m,a}}} \\
&= \sqrt{\frac{2\alpha \log(T)}{T_a} \sum_{m \in [K]} \frac{p_{m,a}}{\gamma_{m,a}}} \\
&= \sqrt{\frac{2\alpha \log(T)}{T_a} P_a}
\end{aligned}
$$

Using inequality (12), we have:

$$
T_a \leq \frac{8\alpha \log(T) P_a}{\Delta_a^2}
$$

Thus, we get:

$$
\begin{aligned}
\mathbb{E}[T_a] &= \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}(I_t = a)] \\
&\leq \frac{8\alpha \log(T) P_a}{\Delta_a^2} + \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}(E_t^c)] \\
&\leq \frac{8\alpha \log(T) P_a}{\Delta_a^2} + \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}(\bigcup_m \left(B_{t,m,a}^c \cup C_{t,m,a}^c\right) \cup A_{t,a})] \\
&\leq \frac{8\alpha \log(T) P_a}{\Delta_a^2} + \sum_{t=1}^{T} 4Kt^{-\alpha} \\
&\leq \frac{8\alpha \log(T) P_a}{\Delta_a^2} + \frac{4K\alpha}{\alpha - 1}. \qquad (13)
\end{aligned}
$$

To conclude, note that the regret of second part of algorithm is $\mathbb{E}[R_2] = \sum_a \Delta_a \mathbb{E}[T_a]$. We now split the arms into two groups: $\Delta_a \leq \sqrt{\frac{8\alpha \log(T) nS}{T}}$ and $\Delta_a \geq \sqrt{\frac{8\alpha \log(T) nS}{T}}$. Let R_2 be the regret for the second part, and let $x = \sqrt{\frac{8\alpha \log(T) nS}{T}}$. To conclude, note that $\mathbb{E}[R] = \sum_a \Delta_a \mathbb{E}[T_a]$. Since $S = \frac{\sum_a P_a}{n}$ Then:

$$
\begin{aligned}
\mathbb{E}[R_2] &= \sum_a \Delta_a \mathbb{E}[T_a] = \sum_{\Delta_a < x} \Delta_a \mathbb{E}[T_a] + \sum_{\Delta_a \geq x} \Delta_a \mathbb{E}[T_a] \\
&\leq Tx + \frac{8\alpha \log(T)}{x} S + \frac{4K\alpha n}{\alpha - 1} = 2\sqrt{8\alpha T \log(T) nS} + \frac{4K\alpha n}{\alpha - 1}.
\end{aligned}
$$

Finally, for the total regret $R = R_1 + R_2$, we have:

$$
\begin{aligned}
\mathbb{E}[R] &\leq 2\sqrt{8\alpha T \log(T) nS} + \frac{4K\alpha n}{\alpha - 1} + \sum_a T_{1,a} \\
&\leq 2\sqrt{8\alpha T \log(T) nS} + \frac{4K\alpha n}{\alpha - 1} + n \log(T)^2 \\
&= O\left(\sqrt{\alpha T \log(T) nS}\right). \qquad (14)
\end{aligned}
$$

□

Theorem 4. *(Regret guarantee for Alg. 3) Under Assumptions 3, for every $\alpha > 1$, there exists a constant c such that the following regret bound holds for $T \geq c$:*

$$\mathbb{E}[R_T] = O\left(\sqrt{\alpha T \log(T) n S}\right).$$

Proof. We follow the same approach as the previous proof. From the previous result, we know the following inequality holds with probability $1 - 2t^{-\alpha}$:

$$\left|\sum_{m\in\mathbb{M}} p_{m,a}\hat{\mu}_{m,a} - \mu_a\right| \leq \sqrt{\frac{\alpha \log(t)}{2} \sum_{m\in\mathbb{M}} \frac{p_{m,a}^2}{T_{m,a,o}}}.$$

Let this “good event" be denoted as $A_{t,a}$.

Now, we consider:

$$\left|\sum_{m\in\mathbb{M}} \hat{p}_{m,a}\hat{\mu}_{m,a} - \mu_a\right| = \left|\sum_{m\in\mathbb{M}} (\hat{p}_{m,a} - p_{m,a})\hat{\mu}_{m,a} + \sum_{m\in\mathbb{M}} p_{m,a}\hat{\mu}_{m,a} - \mu_a\right| \tag{15}$$

$$\leq \sum_{m\in\mathbb{M}} |\hat{p}_{m,a} - p_{m,a}|\, \hat{\mu}_{m,a} + \left|\sum_{m\in\mathbb{M}} p_{m,a}\hat{\mu}_{m,a} - \mu_a\right|. \tag{16}$$

From the previous proof, since $\hat{p}_{m,a} = \frac{T_{m,a}}{T_a}$, we have:

$$|p_{m,a} - \hat{p}_{m,a}| \leq \sqrt{\frac{\alpha \log(t)}{2T_a}},$$

with probability at least $1 - 2t^{-\alpha}$. Denote this "good event" as $B_{t,m,a}$. Under this event for $T_a \geq \log(T)^2$ and sufficient big T we will have $\frac{p_{m,a}}{2} \leq \hat{p}_{m,a} \leq 2p_{m,a}$

Additionally, we have:

$$\left|\gamma_{m,a} - \frac{T_{m,a,o}}{T_{m,a}}\right| \leq \sqrt{\frac{\alpha \log(t)}{2T_{m,a}}} \leq \sqrt{\frac{\alpha \log(T)}{2T_{m,a}}},$$

again with probability $1 - 2t^{-\alpha}$, denoted as "good event" $C_{t,m,a}$.

If these “good events" hold, we know:

$$T_{m,a,o} \geq \frac{\gamma_{m,a} p_{m,a} T_a}{4}.$$

We also know:

$$|\mu_{m,a} - \hat{\mu}_{m,a}| \leq \sqrt{\frac{\alpha \log(t)}{2T_{m,a,o}}},$$

which leads to:

$$\hat{\mu}_{m,a} \leq \mu_{m,a} + \sqrt{\frac{\alpha \log(t)}{2T_{m,a,o}}} \leq 1 + \sqrt{\frac{\alpha \log(t)}{2T_{m,a,o}}},$$

with probability at least $1 - 2t^{-\alpha}$, denoted as "good event" $D_{t,m,a}$.

Under all these “good events," we have:

$$\sum_{m\in\mathbb{M}} |\hat{p}_{m,a} - p_{m,a}|\, \hat{\mu}_{m,a} \leq \sum_{m\in\mathbb{M}} |\hat{p}_{m,a} - p_{m,a}| + \sum_{m\in\mathbb{M}} |\hat{p}_{m,a} - p_{m,a}| \times \sqrt{\frac{\alpha \log(t)}{2T_{m,a,o}}}$$

$$\leq \sum_{m\in\mathbb{M}} |\hat{p}_{m,a} - p_{m,a}| + \sum_{m\in\mathbb{M}} \sqrt{\frac{\alpha \log(T)}{2T_a}} \times \sqrt{\frac{2\alpha \log(T)}{p_{m,a}\gamma_{m,a} T_a}}.$$

Using Lemma 7 from Kamath et al. [2015], we get:

$$\sum_{m\in\mathbb{M}} |\hat{p}_{m,a} - p_{m,a}| \leq \sqrt{\frac{2(k-1)}{\pi T_a}} + \frac{4k^{\frac{1}{2}}(k-1)^{\frac{1}{4}}}{T_a^{\frac{3}{4}}}.$$

Now, combining everything with the initial inequality (15), we have:

$$\left|\sum_{m\in\mathbb{M}} \hat{p}_{m,a}\hat{\mu}_{m,a} - \mu_a\right| \leq \sqrt{\frac{2(k-1)}{\pi T_a}} + \frac{4k^{\frac{1}{2}}(k-1)^{\frac{1}{4}}}{T_a^{\frac{3}{4}}} + \sum_{m\in\mathbb{M}} \sqrt{\frac{\alpha\log(T)}{2T_a}} \times \sqrt{\frac{2\alpha\log(T)}{p_{m,a}\gamma_{m,a}T_a}} + \sqrt{\frac{\alpha\log(T)}{2}\sum_{m\in\mathbb{M}}\frac{p_{m,a}^2}{T_{m,a,o}}}.$$

For sufficiently large T, since $T_a > \log(T)^2$, we have:

$$\sqrt{\frac{\alpha\log(T)}{2}\sum_{m\in\mathbb{M}}\frac{p_{m,a}^2}{T_{m,a,o}}} \geq \sum_{m\in\mathbb{M}} \sqrt{\frac{\alpha\log(T)}{2T_a}} \times \sqrt{\frac{2\alpha\log(T)}{p_{m,a}\gamma_{m,a}T_a}}.$$

and

$$\sqrt{\frac{\alpha\log(T)}{2}\sum_{m\in\mathbb{M}}\frac{p_{m,a}^2}{T_{m,a,o}}} \geq \sqrt{\frac{2(k-1)}{\pi T_a}} \text{ and } \frac{4k^{\frac{1}{2}}(k-1)^{\frac{1}{4}}}{T_a^{\frac{3}{4}}}$$

Thus, we conclude:

$$\left|\sum_{m\in\mathbb{M}} \hat{p}_{m,a}\hat{\mu}_{m,a} - \mu_a\right| \leq 4\sqrt{\frac{\alpha\log(T)}{2}\sum_{m\in\mathbb{M}}\frac{p_{m,a}^2}{T_{m,a,o}}}.$$

Finally, since $\frac{p_{m,a}}{2} \leq \hat{p}_{m,a}$, we have:

$$\left|\sum_{m\in\mathbb{M}} \hat{p}_{m,a}\hat{\mu}_{m,a} - \mu_a\right| \leq 8\sqrt{\frac{\alpha\log(T)}{2}\sum_{m\in\mathbb{M}}\frac{\hat{p}_{m,a}^2}{T_{m,a,o}}}.$$

Following the exact reasoning in the proof of Theorem 3, and using the fact that $\hat{p}_{m,a} \leq 2p_{m,a}$ we conclude:

$$\mathbb{E}[R] = O\left(\alpha\sqrt{T\log(T)nS}\right).$$

$\square$

Theorem 5. *(Minimax lower bound for MAR) For any policy π, there exists a MAR instance ν such that:*

$$\mathbb{E}[R_T(\pi,\nu)] = \Omega\left(\sqrt{TnH}\right).$$

Proof. Consider the following $n+1$ bandit instances, with n arms labeled $a_1, a_2, \ldots, a_n$. The reward distribution for each arm follows a Normal distribution with variance one, and mean specified as follows.

Bandit instance 0:

- $\mu_{m,a} = 0$ for all arms $a = a_1, \ldots, a_n$ and for all $m = 1, \ldots, K$.

Bandit instance j for $j = 1, \ldots, n$:

- $\mu_{m,a} = \frac{\Delta}{P_a\gamma_{m,a}}$ for arm $a = a_j$, and for all $m = 1, \ldots, K$.

- $\mu_{m,a} = 0$ for all arms $a \neq a_j$, and for all $m = 1, \dots, K$.

For each instance $j \in \{1, \dots, n\}$:

- If $a = j$: $\mu_a = \sum_{m \in [K]} p_{m,a} \mu_{m,a} = \Delta$.
- If $a \neq j$: $\mu_a = 0$.

For instance 0:

- For all $a \in [1, \dots, n]$: $\mu_a = 0$.

Now like the previous we use the mentioned lemmas from Lattimore and Szepesvári [2020] to complete our proof.

Like before for every $a = 1, \dots, n$ we have:

$$
\begin{aligned}
\mathrm{KL}(P_0, P_a) &= \mathbb{E}_0[T_a] \mathrm{KL}(P_0(a), P_a(a)) \\
&= \mathbb{E}_0[T_a] \sum_{m \in [K]} p_{m,a} \gamma_{m,a} \frac{\Delta^2}{2 P_a^2 \gamma_{m,a}^2} \\
&= \mathbb{E}_0[T_a] \frac{\Delta^2}{P_a^2} \sum_{m \in [K]} \frac{p_{m,a}}{\gamma_{m,a}} \\
&= \mathbb{E}_0[T_a] \frac{\Delta^2}{P_a}
\end{aligned}
$$

From this, we can bound $\mathbb{E}_a[T_a(T)]$ as follows:

$$
\begin{aligned}
\mathbb{E}_a[T_a(T)] &\leq \mathbb{E}_0[T_a(T)] + T d_{\mathrm{TV}}(P_0(a), P_a(a)) \\
&\leq \mathbb{E}_0[T_a(T)] + T \sqrt{\frac{1}{2} \mathrm{KL}(P_0(a), P_a(a))} \\
&= \mathbb{E}_0[T_a(T)] + \frac{T}{2} \sqrt{\mathbb{E}_0[T_a] \frac{\Delta^2}{P_a}}
\end{aligned}
$$

Let $R_m = R_T(\pi; i)$ denote the regret of applying policy π on the i-th bandit instance up to time T, where i refers to the i-th bandit instance.

Summing over all bandit instances $1, \dots, n$, we have:

$$
\begin{aligned}
\sum_{i=1}^{n} \mathbb{E}[R_i] &= \sum_{a \in [n]} \Delta \left(T - \mathbb{E}_a[T_a(T)] \right) \\
&\geq \Delta T n - \Delta \sum_{a \in [n]} \left(\mathbb{E}_0[T_a(T)] + \frac{T}{2} \sqrt{\mathbb{E}_0[T_a] \frac{\Delta^2}{P_a}} \right) \\
&\geq \Delta T n - \Delta T - \frac{\Delta^2 T}{2} \sum_{a \in [n]} \sqrt{\frac{\mathbb{E}_0[T_a(T)]}{P_a}} \\
&\geq \Delta T n - \Delta T - \frac{\Delta^2 T}{2} \sqrt{T \sum_{a \in [n]} \frac{1}{P_a}} \\
&\geq \frac{\Delta T n}{2} - \frac{\Delta^2 T}{2} \sqrt{T \sum_{a \in [n]} \frac{1}{P_a}} \quad \text{using } \Delta = \frac{n}{2 \sqrt{T \sum_{a \in [n]} \frac{1}{P_a}}} \\
&\geq \frac{n}{8} \sqrt{\frac{T n^2}{\sum_{a \in [n]} \frac{1}{P_a}}}
\end{aligned}
$$

Thus, there exists an instance where $\mathbb{E}[R_i] \geq \Omega\left(\sqrt{\frac{Tn^2}{\sum\limits_{a\in[n]}\frac{1}{P_a}}}\right) = \Omega\left(\sqrt{TnH}\right)$. $\square$

Theorem 6. *For any mediator-agnostic policy π (a policy that does not have access to mediator values), there exists a MAR instance ν which satisfies Assumption 3 and its regret grows linearly*

$$\mathbb{E}[R_T(\pi,\nu)] = \Omega(T).$$

Proof. We construct two bandit problems, each with two arms, such that in the second bandit, the arms are swapped. The key observation is that the distributions of the observed outputs, when there is no mediator, are identical, but, the actual means of the arms differ, with $\mu_1 - \mu_2 = \Delta$.

First, let us assume that we have constructed two arms with distributions P_1 and P_2, and identical distribution P when there is no mediator. Now, let for instance 1, arms 1 and 2 have distributions P_1 and P_2, respectively, while for instance 2, arms 1 and 2 have distributions P_2 and P_1, respectively.

Now, consider the scenario without a mediator. Building on the ideas from previous proofs, we obtain:

$$\mathbb{E}_2[T_2(T)] \leq \mathbb{E}_1[T_2(T)] + Td_{TV}(P,P) = \mathbb{E}_1[T_2(T)],$$

By symmetry, we similarly have:

$$\mathbb{E}_1[T_2(T)] \leq \mathbb{E}_2[T_2(T)] + Td_{TV}(P,P) = \mathbb{E}_2[T_2(T)].$$

Therefore, we conclude that:

$$\mathbb{E}_1[T_2(T)] = \mathbb{E}_2[T_2(T)].$$

Applying the same symmetry argument again, it follows that:

$$\mathbb{E}_1[T_1(T)] = \mathbb{E}_2[T_1(T)].$$

Thus, the expected number of pulls for both arms in both cases remains identical.

But now with the actual means we would have:

$$\mathbb{E}[R_T(1)] + \mathbb{E}[R_T(2)] = \mathbb{E}_1[T_2(T)]\Delta + \mathbb{E}_2[T_1(T)]\Delta = \mathbb{E}_2[T_2(T)]\Delta + \mathbb{E}_2[T_1(T)]\Delta = T\Delta = \Omega(T)$$

Therefore, there exists $i \in \{1,2\}$ such that:

$$\mathbb{E}[R_T(1)] \geq \Omega(T)$$

Now lets construct arms with mentioned property. We assume that the rewards for each arm follow a discrete distribution. If $f_Y(y)$ represent the probability mass function of Y. We have:

$$f_Y(y \mid a, O^Y = 1) = \sum_m \mathbb{P}(m \mid a, O^Y = 1) f_Y(y \mid a, m, O^Y = 1) = \sum_m \mathbb{P}(m \mid a, O^Y = 1) f_Y(y \mid a, m),$$

and

$$\mathbb{P}(m \mid a, O^Y = 1) = \frac{\mathbb{P}(m, a, O^Y = 1)}{\sum\limits_m \mathbb{P}(m, a, O^Y = 1)} = \frac{\gamma_{m,a} p_{m,a}}{\sum\limits_m \gamma_{m,a} p_{m,a}}.$$

Now, if we let $\gamma_{m,a} = \frac{\frac{1}{p_{m,a}}}{\sum\limits_m \frac{1}{p_{m,a}}}$, we have:

$$f_Y(y \mid a, O^Y = 1) = \frac{\sum_m f_Y(y \mid a, m)}{K},$$

This expression is independent of $p_{m,a}$. Additionally, we have:

$$\mathbb{P}(O^Y = 1 \mid a) = \sum_m p_{m,a}\gamma_{m,a} = \frac{K}{\sum_m \frac{1}{p_{m,a}}},$$

Now, for any choice of $p_{m,a}$ such that both the set $P_a = \{p_{m,a} \mid \forall m \in \mathbb{M}\}$ and $f_Y(y \mid a, m)$ remain identical for both arms, the resulting arm distributions in the absence of a mediator are still identical. Now, we choose appropriate $p_{m,a}$ such that $\mu_1 - \mu_2 = \Delta$.

Since $\mu_a = \sum_m p_{m,a}\mu_{m,a}$, we let all $\mu_{m,a}$ be zero except for one, which we set to 1. Now, for arm $a = 1$, let $p_{m,a} = 1 - \epsilon$ for the m such that $\mu_{m,a} = 1$, and set the others equal to $\frac{\epsilon}{K-1}$. For arm $a = 2$, let $p_{m,a} = 1 - \epsilon$ for the m such that $\mu_{m,a} \neq 1$, and set the others equal to $\frac{\epsilon}{K-1}$.

In this way, $\mu_1 - \mu_2 = 1 - \epsilon - \frac{\epsilon}{K-1} = \Delta$, completing the construction for small ϵ.

□

Theorem 7. *(Regret guarantee for Alg. 4) Under Assumptions 5, 6, and 7, for every $\alpha > 1$, there exists a constant c such that the following regret bound holds for $T \geq c$:*

$$\mathbb{E}[R_T] = O\Big(\sqrt{\alpha T \log(T) \sum_a S_a^2}\Big),$$

with $S_a = \max\{\frac{LC_a}{\gamma_a \|\Theta_a\|_\infty}, \frac{K}{\gamma_a \sqrt{\sum_{y \in \mathbb{Y}} p_{y,a}\gamma_{y,a}}}\}, \gamma_a = \min_y \gamma_{y,a}$.

Proof. We follow the same approach used in previous upper bound proofs to derive a lower bound on the estimation error of μ_a.

At each time step $1 \leq t \leq T_2$, define:

$$\begin{aligned} b_a &= [\mathbb{P}(m, O^Y = 0 \mid a)]_{K\times 1} \\ \Theta_a &= [\mathbb{P}(m, y, O^Y = 1 \mid a)]_{K\times L} \\ x_a &= [\frac{\mathbb{P}(O^Y = 0 \mid y, a)}{\mathbb{P}(O^Y = 1 \mid y, a)}]_{L\times 1} \end{aligned}$$

Since we know that $\Theta_a x_a = b_a$, we now invoke Theorem 2.2 from Higham [1994], which states:

Theorem 2.2. Let $Ax = b$ and $(A + \Delta A)y = b + \Delta b$, where $\|\Delta A\| \leq \epsilon\|E\|$ and $\|\Delta b\| \leq \epsilon\|f\|$, and assume that $\epsilon\|A^{-1}\|\|E\| < 1$. Then:

$$\frac{\|x - y\|}{\|x\|} \leq \frac{\epsilon}{1 - \epsilon\|A^{-1}\|\|E\|}\left(\frac{\|A^{-1}\|\|f\|}{\|x\|} + \|A^{-1}\|\|E\|\right), \tag{17}$$

and this bound is attainable to first order in ϵ.

For each entry of b_a or Θ_a, we have T_a samples. By applying Hoeffding's inequality and following the same approach as in the proofs of previous theorems, we set $\epsilon = \sqrt{\frac{\alpha \log(T)}{2T_a}}$. Consequently, we obtain the following bounds (all norms are $\|.\|_\infty$):

$$
\|\hat{b}_a - b_a\| \leq \epsilon,
$$
$$
\|\hat{\Theta}_a - \Theta_a\| \leq L\epsilon,
$$

with probability at least $1 - 2K \times (L+1)t^{-\alpha}$.

Now, under the event described above and using (17), we have:

$$
\frac{\|x_a - \hat{x}_a\|}{\|x_a\|} \leq \frac{\epsilon}{1 - \epsilon L \|\Theta_a^{-1}\|} \left(\frac{\|\Theta_a^{-1}\|}{\|x_a\|} + L\|\Theta_a^{-1}\| \right).
$$

We have $x_a = \left[\frac{1-\gamma_{y,a}}{\gamma_{y,a}}\right]_{L\times 1}$, and therefore $\|x_a\| = \frac{1-\gamma_a}{\gamma_a}$. For sufficiently large T and for $T_a \geq \log(T)^2$, it follows that $\epsilon L \|\Theta_a^{-1}\| \leq \frac{1}{2}$, leading to:

$$
\|x_a - \hat{x}_a\| \leq 2\epsilon \left(\|\Theta_a^{-1}\| + L\|\Theta_a^{-1}\| \frac{1-\gamma_a}{\gamma_a} \right) = 2\epsilon \|\Theta_a^{-1}\| \left(\frac{L}{\gamma_a} - (L-1) \right) \leq 2\epsilon \frac{L}{\gamma_a} \|\Theta_a^{-1}\|.
$$

Now, since $\|\hat{\Theta}_a - \Theta_a\| \leq L\epsilon$, for sufficiently large T and $T_a \geq \log(T)^2$, we have $L\epsilon \leq \frac{\|\Theta_a\|}{2}$. Hence:

$$
\frac{\|\Theta_a\|}{2} \leq \|\hat{\Theta}_a\| \leq 2\|\Theta_a\|,
$$

which implies

$$
\|\Theta_a^{-1}\| = \frac{\kappa(\Theta_a)}{\|\Theta_a\|} \leq \frac{2\kappa(\Theta_a)}{\|\hat{\Theta}_a\|} \leq \frac{2C_a}{\|\hat{\Theta}_a\|}.
$$

Thus:

$$
\|x_a - \hat{x}_a\| \leq 4\epsilon \frac{LC_a}{\gamma_a \|\hat{\Theta}_a\|}. \tag{18}
$$

For sufficiently large T and $T_a \geq \log(T)^2$, we will have:

$$
\|x_a - \hat{x}_a\| \leq \frac{1}{2\gamma_a}
$$

so:

$$
\frac{1}{\hat{\gamma}_a} = \|\hat{x}_a + [1]_{L\times 1}\| \geq \|x_a + [1]_{L\times 1}\| - \frac{1}{2\gamma_a} = \|[\frac{1}{\gamma_{y,a}}]_{L\times 1}\| - \frac{1}{2\gamma_a} = \frac{1}{2\gamma_a}.
$$

Using (18), we have:

$$
\|x_a - \hat{x}_a\| \leq 8\epsilon \frac{LC_a}{\hat{\gamma}_a \|\hat{\Theta}_a\|}.
$$

Since $x_a + [1]_{L\times 1} = [\frac{1}{\gamma_{y,a}}]_{L\times 1}$, for every y, we have:

$$
\left| \frac{1}{\gamma_{y,a}} - \frac{1}{\hat{\gamma}_{y,a}} \right| \leq 8\epsilon \frac{LC_a}{\hat{\gamma}_a \|\hat{\Theta}_a\|}.
$$

Now let $p_{m,y|1,a} = \mathbb{P}(m, y \mid O^y = 1, a)$. By applying Hoeffding's inequality, we have the following inequality for all m, y:

$$|q_{m,y|1,a} - p_{m,y|1,a}| \leq \sqrt{\frac{\alpha \log(T)}{2T_{a,o}}}$$

with probability at least $1 - 2KLt^{-\alpha}$. Using the fact that $\frac{p_{m,y|1,a}}{\gamma_{y,a}} = \mathbb{P}(m, y \mid a) = p_{m,y|a}$, we have:

$$\begin{aligned}
\left|p_{m,y|a} - \hat{p}_{m,y|a}\right| &= \left|\frac{p_{m,y|a}}{\gamma_{y,a}} - \frac{q_{m,y|a}}{\hat{\gamma_{y,a}}}\right| \\
&\leq p_{m,y|a}\left|\frac{1}{\gamma_{y,a}} - \frac{1}{\hat{\gamma}_{y,a}}\right| + \frac{1}{\hat{\gamma}_a}\left|q_{m,y|1,a} - p_{m,y|1,a}\right| \\
&\leq 8p_{m,y|a}\epsilon\frac{LC_a}{\hat{\gamma}_a\|\hat{\Theta}_a\|} + \frac{1}{\hat{\gamma}_a}\sqrt{\frac{\alpha\log(T)}{T_{a,o}}}.
\end{aligned}$$

Summing up over m, we have:

$$\begin{aligned}
\left|p_{y|a} - \hat{p}_{y|a}\right| &\leq 8p_{y|a}\epsilon\frac{LC_a}{\hat{\gamma}_a\|\hat{\Theta}_a\|} + \frac{K}{\hat{\gamma}_a}\sqrt{\frac{\alpha\log(T)}{T_{a,o}}} \\
&\leq 8\epsilon\frac{LC_a}{\hat{\gamma}_a\|\hat{\Theta}_a\|} + \frac{K}{\hat{\gamma}_a}\sqrt{\frac{\alpha\log(T)}{T_{a,o}}}.
\end{aligned}$$

Thus, using $\sum_y |y| = 1$:

$$|\mu_a - \hat{\mu}_a| \leq \sum_y |y|\left|p_{y|a} - \hat{p}_{y|a}\right| \leq \sum_y |y|\left(8\epsilon\frac{LC_a}{\hat{\gamma}_a\|\hat{\Theta}_a\|} + \frac{K}{\hat{\gamma}_a}\sqrt{\frac{\alpha\log(T)}{T_{a,o}}}\right) = 8\epsilon\frac{LC_a}{\hat{\gamma}_a\|\hat{\Theta}_a\|} + \frac{K}{\hat{\gamma}_a}\sqrt{\frac{\alpha\log(T)}{T_{a,o}}} = \epsilon_a.$$

Hence, $\mathrm{UCB}(a) = \hat{\mu}_a + \epsilon_a$, and using previous proofs, we conclude that $\epsilon_a \geq \frac{\Delta_a}{2}$. To finalize the proof:

$$\|\hat{\Theta}_a\| \geq \frac{\|\Theta_a\|}{2}, \hat{\gamma}_a \geq \frac{\gamma_a}{2},$$

and we have $\mathbb{P}(O^Y = 1 \mid a) = \sum_y p_{y,a}\gamma_{y,a}$. Applying Hoeffding's inequality gives:

$$\left|\frac{T_{a,o}}{T_a} - \sum_y p_{y,a}\gamma_{y,a}\right| \leq \epsilon,$$

which for sufficiently large T and $T_a \geq \log(T)^2$, states:

$$T_{a,o} \geq T_a\left(\sum_y p_{y,a}\gamma_{y,a}\right).$$

Finally, we have:

$$\epsilon_a \leq \sqrt{\frac{\alpha\log(T)}{T_a}}8\frac{LC_a}{\gamma_a\|\Theta_a\|} + 2\frac{K}{\gamma_a}\sqrt{\frac{1}{\sum_y p_{y,a}\gamma_{y,a}}} \leq 8\sqrt{\frac{\alpha\log(T)}{T_a}}\max\left(\frac{LC_a}{\gamma_a\|\Theta_a\|}, \frac{K}{\gamma_a}\sqrt{\frac{1}{\sum_y p_{y,a}\gamma_{y,a}}}\right),$$

which, following the exact steps of previous proofs, leads to:

$$\mathbb{E}[R_T] = O\left(\sqrt{\alpha T \log(T) \sum_a S_a^2}\right)$$

□

C THEORETICAL RESULTS ON MISSING OUTCOME AND MISSING MEDIATOR

In this section, we present our theoretical results on Missing at Random (MAR) and Missing Not at Random (MNAR) environments for both Missing Outcome and Missing Mediator cases.

C.1 MISSING AT RANDOM (MAR)

As discussed earlier, the identification of $\mu_a = \mathbb{E}[Y \mid a]$ is given by:

$$\mu_a = \sum_{m \in \mathbb{M}} \mathbb{P}(M = m \mid a, O^M = 1)\mathbb{E}[Y \mid M = m, a, O^Y = 1, O^M = 1].$$

Using this identification, we will prove the following theorem. We define $p_{m,a} = \mathbb{P}(M = m, a), \gamma_{m,a} = \mathbb{P}(O^Y = 1 \mid M = m, a), \lambda_a = \mathbb{P}(O^M = 1 \mid a)$. Our algorithm is exactly like 3 where if M is missed we don't update anything.

Theorem 8. *(Regret bound for MAR with missing mediator and outcome) Under Assumption 8, for every $\alpha > 1$, the following regret bound holds for sufficiently large T:*

$$\mathbb{E}[R_T] = O\left(\sqrt{\alpha T \log(T) n S}\right),$$

where $P_a = \sum_{m \in \mathcal{M}} \frac{p_{m,a}}{\gamma_{m,a}\lambda_a}$ and $S := \frac{1}{|\mathcal{A}|} \sum\limits_{a \in \mathcal{A}} P_a$.

Proof. The proof follows a similar approach to the proof of Theorem 4. Using the same reasoning, we have (where T_{m,a,o_Y} is the number of times $M = m$ and the reward are observed when pulling arm a):

$$\left|\sum_{m \in \mathbb{M}} \hat{p}_{m,a}\hat{\mu}_{m,a} - \mu_a\right| \leq 8\sqrt{\frac{\alpha \log(T)}{2} \sum_{m \in \mathbb{M}} \frac{\hat{p}_{m,a}^2}{T_{m,a,o_Y}}}.$$

Similarly, we also have the following inequality (where T_{a,o_M} is the number of times $M = m$ is observed when pulling arm a):

$$|p_{m,a} - \hat{p}_{m,a}| \leq \sqrt{\frac{\alpha \log(t)}{2T_{a,o_M}}}.$$

Additionally, we have:

$$\left|\frac{T_{a,o_M}}{T_a} - \lambda_a\right| \leq \sqrt{\frac{\alpha \log(t)}{2T_a}}.$$

Under this event, for sufficiently large T, we have $T_a \geq \log(T)^2$ and $\frac{p_{m,a}}{2} \leq \hat{p}_{m,a} \leq 2p_{m,a}$. Following similar steps, we get:

$$T_{m,a,o_Y} \geq \frac{p_{m,a}\gamma_{m,a}\lambda_a T_a}{4}.$$

Therefore, using the same definition of ϵ_a, we obtain:

$$\epsilon_a \leq 8\sqrt{\frac{8\alpha \log(T)}{2} \sum_{m \in \mathbb{M}} \frac{p_{m,a}^2}{p_{m,a}\gamma_{m,a}\lambda_a T_a}} = 8\sqrt{\frac{8\alpha \log(T)}{2} \sum_{m \in \mathbb{M}} \frac{p_{m,a}}{\gamma_{m,a}\lambda_a T_a}}.$$

Finally, following the same steps as in previous proofs, we conclude:

$$\mathbb{E}[R_T] = O\left(\sqrt{\alpha T \log(T) n S}\right).$$

P.S.: By "sufficiently large" T, we mean a T that is large enough to satisfy $\log(T) > C$ for some constant C. □

C.2 MISSING NOT AT RANDOM (MNAR)

In this section, we use the identification formula discussed earlier to develop an algorithm and establish an upper bound for this environment. Assume that $\mathbb{P}(m \mid a) = p_{m,a}$, $\mathbb{P}(O^M = 1 \mid m, a) = \lambda_{m,a}$, $\mathbb{P}(O^Y = 1 \mid m, a) = \gamma_{m,a}$. Also define $\lambda_a = \min_m \lambda_{m,a}$. We assume a similar condition to Assumption 7, but for a different matrix. Let $\Theta_a = [\mathbb{P}(m, y, O^M = 1, O^Y = 1 \mid a)]_{K \times L}$:

Assumption 11 (Bounded condition number)**.** *For each arm $a \in \mathcal{A}$, the condition number of the matrix Θ_a is bounded by:*

$$\kappa(\Theta_a) \leq C_a,$$

where $\kappa(\Theta_a)$ denotes the condition number of Θ_a with respect to the ∞-norm, defined as

$$\kappa(\Theta_a) = \|\Theta_a\|_\infty \|\Theta_a^\dagger\|_\infty,$$

with $\Theta_a^\dagger$ being the pseudo-inverse of Θ_a.

In our algorithm we use the the given identification formula and

$$\text{UCB}(a) = \hat{\mu}_a + 2\sqrt{\frac{\alpha \log(T)}{2T_a}}\left(8\frac{C_a}{\|\hat{\Theta}_a\|}\frac{K}{\hat{\lambda}_a} + \frac{1}{\hat{\lambda}_a}\right) + \sqrt{\frac{\alpha \log(T)}{2}\sum_{m \in \mathbb{M}} \frac{4\hat{p}_{m,a}^2}{T_{a,m,o_Y}}}.$$

we will prove the following theorem.

Theorem 9. *(Regret bound for MNAR with missing mediator and outcome) Under Assumptions 9, 10, and 11, for every $\alpha > 1$, the following regret bound holds for sufficiently large T:*

$$\mathbb{E}[R_T] = O\left(\sqrt{\alpha T \log(T) n S}\right).$$

where $S_a = \max\left(\left(32\frac{C_a}{\|\Theta\|_a}\frac{K}{\lambda_a} + \frac{2}{\lambda_a}\right), \sqrt{\sum_{m \in \mathbb{M}} \frac{32 p_{m,a}}{\lambda_{m,a}\gamma_{m,a}}}\right)$, $S = \frac{\sum_a S_a^2}{n}$

Proof. The proof closely follows the reasoning from Theorem 7. Let:

$$\begin{aligned}
b_a &= [\mathbb{P}(y, O^M = 0, O^Y = 1 \mid a)]_{L \times 1}, \\
\Theta_a &= [\mathbb{P}(m, y, O^M = 1, O^Y = 1 \mid a)]_{K \times L}, \\
x_a &= \begin{bmatrix} \frac{\mathbb{P}(O^M = 0 \mid m, a)}{\mathbb{P}(O^M = 1 \mid m, a)} \end{bmatrix}_{K \times 1}.
\end{aligned}$$

We know that $\Theta_a x_a = b_a$. Using the same approach as in Theorem 7, we derive the following inequality for $\epsilon = \sqrt{\frac{\alpha \log(T)}{2T_a}}$:

$$\|x - \hat{x}\| \leq 2\epsilon\|\Theta_a^{-1}\|\left(1 + K\frac{1 - \lambda_a}{\lambda_a}\right) \leq 2\epsilon\|\Theta_a^{-1}\|\frac{K}{\lambda_a} \leq 4\epsilon\frac{C_a}{\|\hat{\Theta}_a\|}\frac{K}{\lambda_a} \leq 8\epsilon\frac{C_a}{\|\hat{\Theta}_a\|}\frac{K}{\hat{\lambda}_a}.$$

Additionally, since $x = \left[\frac{1 - \lambda_{m,a}}{\lambda_{m,a}}\right]_{K \times 1}$, we have:

$$\left|\frac{1}{\lambda_{m,a}} - \frac{1}{\hat{\lambda}_{m,a}}\right| \leq 8\epsilon\frac{C_a}{\|\hat{\Theta}_a\|}\frac{K}{\hat{\lambda}_a}.$$

Furthermore, for $p_{m,1|a} = \mathbb{P}(M = m, O^M = 1 \mid a)$, we have:

$$\left|p_{m,1|a} - \hat{p}_{m,1|a}\right| \leq \epsilon.$$

Using a similar approach to the proof of Theorem 7, we obtain:

$$\left|\frac{p_{m,1|a}}{\lambda_{m,a}} - \frac{\hat{p}_{m,1|a}}{\hat{\lambda}_{m,a}}\right| \leq p_{m,1|a}\left(8\epsilon\frac{C_a}{\|\hat{\Theta}_a\|}\frac{K}{\hat{\lambda}_a}\right) + \frac{1}{\hat{\lambda}_a}\epsilon = \epsilon\left(8p_{m,1|a}\frac{C_a}{\|\hat{\Theta}_a\|}\frac{K}{\hat{\lambda}_a} + \frac{1}{\hat{\lambda}_a}\right).$$

Therefore, we can conclude:

$$|\hat{p}_{m,a} - p_{m,a}| \leq \epsilon\left(8p_{m,1|a}\frac{C_a}{\|\hat{\Theta}_a\|}\frac{K}{\hat{\lambda}_a} + \frac{1}{\hat{\lambda}_a}\right).$$

Additionally, we have the following bound for T_{a,o_M,o_Y} (the number of times both M and Y are observed):

$$|\hat{\mu}_{m,a} - \mu_{m,a}| \leq \sqrt{\frac{\alpha\log(T)}{T_{a,o_M,o_Y}}}.$$

Using $\mathbb{P}(O^Y = 1, O^M = 1 \mid a) = \sum_{m\in\mathbb{M}} p_{m,a}\gamma_{m,a}\lambda_{m,a}$, we have:

$$\left|\frac{T_{a,o_M,o_Y}}{T_a} - \sum_{m\in\mathbb{M}} p_{m,a}\gamma_{m,a}\lambda_{m,a}\right| \leq \sqrt{\frac{\alpha\log(T)}{T_a}},$$

which gives $T_{a,o_M,o_Y} \geq \frac{T_a}{2}\left(\sum_{m\in\mathbb{M}} p_{m,a}\gamma_{m,a}\lambda_{m,a}\right)$. Thus, for sufficiently large T and $T_a \geq \log(T)^2$, we have $\hat{\mu}_{m,a} \leq 2\mu_{m,a} \leq 2$.

Therefore:

$$\begin{aligned}
\left|\sum_{m\in\mathbb{M}} \hat{p}_{m,a}\hat{\mu}_{m,a} - \mu_a\right| &= \left|\sum_{m\in\mathbb{M}} (\hat{p}_{m,a} - p_{m,a})\hat{\mu}_{m,a} + \sum_{m\in\mathbb{M}} p_{m,a}\hat{\mu}_{m,a} - \mu_a\right| && (19)\\
&\leq \sum_{m\in\mathbb{M}} |\hat{p}_{m,a} - p_{m,a}|\,\hat{\mu}_{m,a} + \left|\sum_{m\in\mathbb{M}} p_{m,a}\hat{\mu}_{m,a} - \mu_a\right| && (20)\\
&\leq 2\sum_{m\in\mathbb{M}} \epsilon\left(8p_{m,1|a}\frac{C_a}{\|\hat{\Theta}_a\|}\frac{K}{\hat{\lambda}_a} + \frac{1}{\hat{\lambda}_a}\right) + \left|\sum_{m\in\mathbb{M}} p_{m,a}\hat{\mu}_{m,a} - \mu_a\right| \quad \left(\text{using } \sum_{m\in\mathbb{M}} p_{m,1|a} \leq 1\right) && (21)\\
&\leq 2\epsilon\left(8\frac{C_a}{\|\hat{\Theta}_a\|}\frac{K}{\hat{\lambda}_a} + \frac{1}{\hat{\lambda}_a}\right) + \left|\sum_{m\in\mathbb{M}} p_{m,a}\hat{\mu}_{m,a} - \mu_a\right|. && (22)
\end{aligned}$$

Using the same technique as before, we have the following inequality for T_{a,m,o_Y} (the number of times $M = m$ and the reward are observed when pulling arm a):

$$\left|\sum_{m\in\mathbb{M}} p_{m,a}\hat{\mu}_{m,a} - \mu_a\right| \leq \sqrt{\frac{\alpha\log(T)}{2}\sum_{m\in\mathbb{M}}\frac{4\hat{p}_{m,a}^2}{T_{a,m,o_Y}}}.$$

Therefore:

$$\left|\sum_{m\in\mathbb{M}} \hat{p}_{m,a}\hat{\mu}_{m,a} - \mu_a\right| \leq 2\epsilon\left(8\frac{C_a}{\|\hat{\Theta}_a\|}\frac{K}{\hat{\lambda}_a} + \frac{1}{\hat{\lambda}_a}\right) + \sqrt{\frac{\alpha\log(T)}{2}\sum_{m\in\mathbb{M}}\frac{4\hat{p}_{m,a}^2}{T_{a,m,o_Y}}}$$

which proves our UCB upper bound.

Similarly, we know that $T_{a,m,o_Y} \geq \frac{T_a}{2} p_{m,a}\lambda_{m,a}\gamma_{m,a}$, which gives:

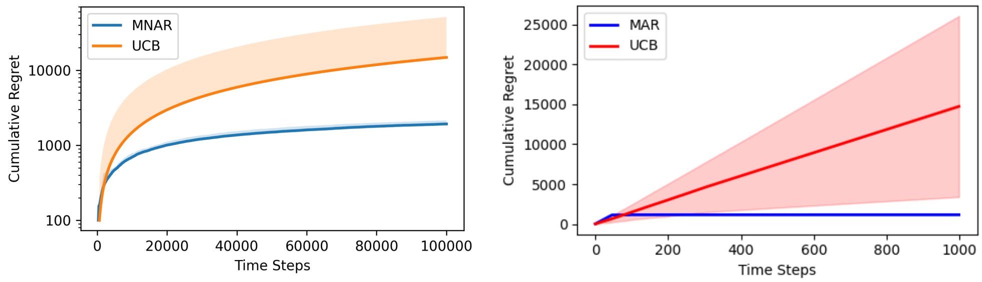
$$
\begin{aligned}
|\mu_a - \hat{\mu_a}| &\leq 2\epsilon \left(8 \frac{C_a}{\|\hat{\Theta}_a\|} \frac{K}{\hat{\lambda}_a} + \frac{1}{\hat{\lambda}_a} \right) + \sqrt{\frac{\alpha \log(T)}{2} \sum_{m \in \mathbb{M}} \frac{32 p_{m,a}^2}{T_a p_{m,a} \lambda_{m,a} \gamma_{m,a}}} \\
&= 2\epsilon \left(8 \frac{C_a}{\|\hat{\Theta}_a\|} \frac{K}{\hat{\lambda}_a} + \frac{1}{\hat{\lambda}_a} \right) + \sqrt{\sum_{m \in \mathbb{M}} \frac{32 p_{m,a}}{\lambda_{m,a} \gamma_{m,a}}} \\
&\leq 2\epsilon \max \left(\left(32 \frac{C_a}{\|\Theta\|_a} \frac{K}{\lambda_a} + \frac{2}{\lambda_a} \right), \sqrt{\sum_{m \in \mathbb{M}} \frac{32 p_{m,a}}{\lambda_{m,a} \gamma_{m,a}}} \right).
\end{aligned}
$$

Following the same reasoning as in previous proofs, we conclude:

$$
\mathbb{E}[R_T] = O\left(\sqrt{\alpha T \log(T) \sum_a S_a^2} \right) = O\left(\sqrt{\alpha T \log(T) n S} \right).
$$

$\square$

D ADDITIONAL EMPIRICAL EVALUATION



(a) MNAR and UCB algorithms in the MNAR bandit environment.

(b) MAR and UCB algorithms in a real-world MAR bandit environment.

Figure 5: Complementary evaluation results for our proposed algorithms.

In Figure 5a, we compare the performance of the UCB and MNAR algorithms in the MNAR bandit environment. The results clearly demonstrate that the cumulative regret of the UCB algorithm is consistently higher than that of the MNAR algorithm. Additionally, the y-axis is displayed on a logarithmic scale, further highlighting the considerable difference in the performance of our algorithm compared to the UCB algorithm. The environment is generated as before, with a horizon of $T = 100{,}000$, and the experiment is repeated 10 times.

D.1 REAL-WORLD SIMULATION

The dataset used in this study is the Primary Biliary Cirrhosis (PBC) dataset from the Mayo Clinic [6], containing 418 observations and 19 variables. Collected over a 10-year span (1974–1984), it focuses on a randomized, placebo-controlled trial of D-penicillamine for treating PBC, and includes both trial participants and observational data from non-participants.

To simulate a real-world setting, we structured the data as follows: the **Z1** variable (1 for D-penicillamine, 2 for placebo) was treated as the *arms of the bandit*, representing treatment groups. The **X** variable, denoting the time in days from registration to death, liver transplantation, or censoring, was used as the outcome. The **D** variable, indicating whether **X** measures time until death (1) or censoring (0), served as the *mediator*.

The **D** mediator captures whether the time interval **X** is associated with death or censoring, offering key insights into the progression of the disease and the effect of treatment. This setup allows us to model the pathways from treatment to outcome, where **Z1** represents the action taken, **X** is the reward (days survived), and **D** explains the intermediate state between treatment and survival or death.

Applying the MAR algorithm to this MAR bandit environment yielded results consistent with those seen in synthetic data, as shown in Figure 5b.

[6] https://www.openml.org/d/200