Sequential Decision Problems with Missing Feedback[*]

Filippo Palomba
Princeton University

This version: July 25, 2025.

Abstract

This paper investigates the challenges of optimal online policy learning under missing data. State-of-the-art algorithms implicitly assume that rewards are always observable. I show that when rewards are missing at random, the Upper Confidence Bound (UCB) algorithm maintains optimal regret bounds; however, it selects suboptimal policies with high probability as soon as this assumption is relaxed. To overcome this limitation, I introduce a fully nonparametric algorithm—Doubly-Robust Upper Confidence Bound (DR-UCB)—which explicitly models the form of missingness through observable covariates and achieves a nearly-optimal worst-case regret rate of $\widetilde{O}(\sqrt{T})$. To prove this result, I derive high-probability bounds for a class of doubly-robust estimators that hold under broad dependence structures. Simulation results closely match the theoretical predictions, validating the proposed framework.

Keywords: sequential decision problems, double robustness, missing data

1 Introduction

In recent years, technological advancements and the proliferation of digital platforms have enabled the collection of real-time data describing the interaction between a policy

[*] I thank my advisor, Matias Cattaneo, for invaluable guidance throughout this project. I also thank Thomas Bearpark, Chiara Motta, Andrea Pugnana, and the participants in the Princeton Econometrics Student Seminar for insightful discussions. All errors and omissions remain my own. Corresponding author e-mail address: fpalomba@princeton.edu.

and its targets. Such rich *online* datasets naturally give rise to sequential decision problems in which a decision-maker continuously re-optimizes the implemented policy as more information is gathered. Historically, these decision problems have been modeled as multi-armed bandits (Thompson, 1933; Wald, 1947; Robbins, 1952), in which a decision-maker interactively learns the best policy (*action*) among a set of alternatives by trying out options and observing a signal (*feedback* or *reward*) from the environment.

Naturally, the success of these strategies requires the decision-maker to be able to observe some feedback following each action and, thus, to assess the extent to which the chosen action had the intended impact. However, in many real-world settings, the environment's response to a given policy might not always be observable, making it harder to gauge the true efficacy of an intervention. If this particular form of missing data is correlated with the outcome of interest, it introduces sampling bias (Horvitz and Thompson, 1952) and complicates the identification of optimal decision rules. Despite that, standard bandit algorithms typically rely on the assumption that rewards are always observed upon each action.

To make the problem more concrete, consider the case of a digital platform that experiments with user engagement strategies—such as personalized push notifications or in-app promotions—and gauges satisfaction via voluntary review prompts. While some users may readily submit reviews, a non-trivial fraction will opt out. Suppose the probability of response itself depends on satisfaction. In that case, reviews are observed only for a particular subpopulation, which may differ from the target one in many aspects, thus potentially inducing a suboptimal policy choice. Other real-life examples include a firm implementing different hiring strategies and a graduate admission committee experimenting with alternative types of offers.

In this paper, I first show that the popular UCB algorithm (Auer, Cesa-Bianchi and Fischer, 2002) maintains its optimal (up to logarithmic factor) worst-case regret rate of $\widetilde{O}(\sqrt{T})$ whenever the process that causes missing data is independent from rewards. Intuitively, in this case, rewards are missing completely at random (Rubin, 1976), hence reward-independent missingness only makes the learning process slower, without invalidating it. Nevertheless, this independence assumption is implausible in most practical scenarios where reward observability directly depends on the feedback itself. For example, customer satisfaction is typically surveyed only for extreme types, and different job postings attract different potential employees.

Whenever the process that causes missing data is reward-dependent, I show that the standard UCB algorithm may select suboptimal policies with probability approaching one as the number of trials grows. By explicitly modeling the dependence between rewards and the selection process through observable covariates, I develop a theoretically grounded, fully nonparametric procedure for sequential decision problems with missing feedback. The proposed algorithm achieves a worst-case regret rate of order $\widetilde{O}(\sqrt{T})$, which I show to coincide (up to logarithmic factors) with the rate of a novel lower bound on the minimax regret for the class of bandits with reward-dependent missingness. Additionally, the proposed method is doubly robust in the sense that only one among the conditional expectation of rewards and the conditional probability of missingness needs to be correctly specified for the algorithm to function properly. In this spirit, the algorithm's name is Doubly-Robust Upper Confidence Bound (DR-UCB).

Finally, to the best of my knowledge, this paper provides the first high-probability bounds for a doubly-robust estimator under very general conditions. Specifically, the high-probability bounds derived throughout rely on the theory for martingale difference sequences (Freedman, 1975), and so allow for very general dependence structures in the data. This was necessary in this setting due to the technical challenges brought by the sequential nature of the problem.

1.1 Related Work

Sequential decision-making problems under uncertainty have been extensively studied in economics, statistics, and computer science after the seminal work of Wald (1947) and have mostly focused on the design of optimal strategies and algorithms. Recent advances have particularly emphasized the role of adaptive algorithms that sequentially learn and adjust to uncertain environments, leading to a proliferation of approaches that efficiently balance exploration and exploitation; for a textbook introduction, see Bubeck and Cesa-Bianchi (2012), Lattimore and Szepesvári (2020), and references therein.

This paper builds explicitly on the extensive literature studying the properties of the UCB algorithm. The idea of being optimistic in the face of uncertainty first appeared in Lai and Robbins (1985). Lai (1987) provided the first version of the UCB algorithm, whereas the UCB algorithm analyzed here is closer to the UCB1 analyzed in Auer, Cesa-Bianchi and Fischer (2002). UCB methods have been widely recognized for their effectiveness in handling exploration-exploitation trade-offs, providing provable performance guarantees and optimal regret bounds in multi-armed bandit frameworks.

The literature on multi-armed bandits with delayed feedback is also closely related to this project. Delayed feedback poses significant challenges for standard UCB algorithms, as the decision-maker is forced to take actions before she can receive any signal from the environment she is interacting with. The closest predecessor of this work is probably Lancewicki, Segal, Koren and Mansour (2021), where the authors provide problem-specific regret bounds for the Successive Elimination algorithm when rewards are bounded and delays are occasionally infinite and reward-dependent. Different from them, here I propose distribution-free worst-case regret bounds for a novel version of the UCB and allow rewards to be unbounded.

Finally, this paper intersects with the literature that relies on doubly-robust estimators for different goals, such as handling missing data (Robins, Rotnitzky and Zhao, 1994; Bang and Robins, 2005), estimating the causal impacts of policies (Cattaneo, 2010; Farrell, 2015; Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins, 2018), or to learn optimal policies in offline (Athey and Wager, 2021) and online settings (Kallus, Mao, Wang and Zhou, 2022; Shen, Cai and Song, 2024).

1.2 Organization of the Paper

The paper is organized as follows. Section 2 describes the problem and the algorithms used throughout. Section 3 contains the main results and showcases the worst-case regret properties of the classical UCB algorithm and DR-UCB in environments with reward-independent (Section 3.1) and reward-dependent (Section 3.2) missingness. Section 4 illustrates some simulation evidence. Section 5 concludes. The code replicating the simulation study is available at `https://github.com/filippopalomba/P_2025_banditMissing`.

1.3 Notation

For two positive sequences $\{a_n\}_n, \{b_n\}_n$, I write $a_n = O(b_n)$ if $\exists\, M \in \mathbb{R}_{++} : a_n \leq M b_n$ for all large n, $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n b_n^{-1} = 0$, $a_n = \widetilde{O}(b_n)$ if $\exists\, k \in \mathbb{N}, C \in \mathbb{R}_{++}$: $a_n = O(b_n \ln^k(Cn))$ $a_n \lesssim b_n$ if there exists a constant $C \in \mathbb{R}_{++}$ such that $a_n \leq C b_n$ for all large n, and $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$. For two sequences of random variables $\{A_n\}_n, \{B_n\}_n$, I write $A_n = o_{\mathbb{P}}(B_n)$ if $\forall\, \varepsilon \in \mathbb{R}_{++}, \lim_{n\to\infty} \mathbb{P}[|A_n B_n^{-1}| \geq \varepsilon] = 0$ and $A_n = O_{\mathbb{P}}(B_n)$ if $\forall\, \varepsilon \in \mathbb{R}_{++}, \exists\, M, n_0 \in \mathbb{R}_{++} : \mathbb{P}[|A_n B_n^{-1}| > M] < \varepsilon$, for $n > n_0$. I denote a (possibly multivariate) Gaussian random variable with $\mathsf{N}(\mathbf{a}, \mathbf{B})$, where $\mathbf{a}$ denotes the mean and $\mathbf{B}$ the variance-covariance, with $\mathsf{Be}(p)$ a Bernoulli distribution with $p \in (0, 1]$ denoting the success probability, with $\mathsf{sG}(\sigma)$ a sub-Gaussian random variable with proxy variance at most $\sigma > 0$, and with $\mathcal{SG}(\sigma)$ the space of sub-Gaussian probability distribution with variance proxy at most $\sigma > 0$. A random variable X is sub-Gaussian with parameter $\sigma > 0$ if $\forall\, \lambda \in \mathbb{R}, \mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2\sigma^2/2)$ and $\mathbb{E}[X] = 0$ (Definition 5.1 in Lattimore and Szepesvári, 2020). If $\{X_t\}_{t=1}^{\infty}$ is an $\mathcal{F}$-adapted martingale difference sequence with respect to some filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t=1}^{\infty}$, then it is understood that $X_t \sim \mathsf{sG}(\sigma^2)$ requires $\forall\, \lambda \in \mathbb{R}, \mathbb{E}[\exp(\lambda X_t) \mid \mathcal{F}_t] \leq \exp(\lambda^2\sigma^2/2)$ and $\mathbb{E}[X_t \mid \mathcal{F}_t] = 0$. See also Table SA-1 in the supplemental appendix for a summary of the project-specific notation.

2 Problem Setup and Preliminaries

I start by describing a generic instance of a stochastic multi-armed bandit (henceforth, MAB) with (possibly) missing rewards and the decision-maker that interacts with such an environment.

Setting. A decision-maker faces a sequential decision problem over $T \in \mathbb{N}$ rounds in a stochastic environment. At the beginning of each round $t \in [T]$, using all the information available at that point, the decision-maker selects an action $A_t \in \mathcal{A} := \{1, \dots, A\}$. Each action $a \in \mathcal{A}$ is associated with a reward $R_a \in \mathbb{R}, R_a \sim \mathsf{sG}(\sigma_a), \sigma_a > 0$, an indicator for *not* being missing $C_a \in \{0, 1\}$, and some covariates $\mathbf{X}_a \in \mathcal{X} \subseteq \mathbb{R}^k, k \in \mathbb{N}$. All random variables are independent across actions, i.e., $(R_a, C_a, \mathbf{X}_a) \perp\!\!\!\perp (R_{a'}, C_{a'}, \mathbf{X}_{a'})$ for $a \neq a', a, a' \in \mathcal{A}$. A stochastic MAB problem with missing rewards is defined as a collection of random variables $\{(R_{a,\ell}, C_{a,\ell}, \mathbf{X}_{a,\ell})\}_{a\in\mathcal{A},\ell\in[T]}$ where each $(R_{a,\ell}, C_{a,\ell}, \mathbf{X}_{a,\ell})$ is a random draw from $(R_a, C_a, \mathbf{X}_a)$. The reward of action $a \in \mathcal{A}$ in round $t \in [T]$, denoted with $R_{a,t}$, is observed only if $C_{a,t} = 1$. At this level of generality, the class of bandits considered is

$$\mathcal{C} := \left\{ (\nu_a)_{a\in\mathcal{A}} : \nu_a^{[R]} \in \mathcal{SG}(\sigma_a),\ \nu_a^{[C]} = \mathsf{Be}(q_a), q_a \in (0, 1] \right\},$$

where $\nu_a^{[Y]}$ is the marginal distribution of $Y \in \{R, C\}$. Finally, I define $\bar{\sigma} := \sqrt{\max_{a\in\mathcal{A}} \sigma_a^2}$ and $\underline{q} = \min_{a\in\mathcal{A}} q_a$.

Decision-maker. Each decision-maker is characterized by a policy that maps $\{(A_\ell, R_{A_\ell,\ell}, C_{A_\ell,\ell}, \mathbf{X}_{A_\ell,\ell}^\top)\}_{\ell\in[t-1]}$, the history up to the beginning of round t, to the space of probability distributions over actions $\Delta(\mathcal{A})$. Denote the space of policies as

$$\Pi := \left\{ \pi : \pi = \{\pi_t\}_{t\in[T]}, \pi_t : (\mathcal{A} \times \mathbb{R} \times \{0, 1\} \times \mathcal{X})^{t-1} \to \Delta(\mathcal{A}) \right\}.$$

I use interchangeably the words "decision-maker", "algorithm", and "policy" when referring to a generic element $\pi \in \Pi$. Protocol 1 describes the interaction between a decision-maker and a MAB with (possibly) missing feedback.

Protocol 1 Multi-Armed Bandit with Missing Rewards

```
Consider a generic bandit $\nu \in \mathcal{C}$, where $\nu = (\nu_a)_{a \in \mathcal{A}}$
  for $\ell = 1, 2, \ldots, T$ do
      Decision-maker chooses $A_\ell = a$ according to some policy $\pi_t$
      Nature samples $(C_{a,\ell}, R_{a,\ell}, \mathbf{X}_{a,\ell}) \sim \nu_a$
      if $C_{a,\ell} = 1$ then
          Decision-maker observes $R_{a,\ell}$
      else
          Decision-maker receives no feedback
      end if
  end for
```

Regret. The *pseudo-regret* of a decision-maker following a policy π in a MAB with missing rewards $\nu \in \mathcal{C}$ is

$$\mathsf{Regret}_T(\pi; \nu) = \sum_{t=1}^{T} (\max_{a \in \mathcal{A}} \theta_a - \mathbb{E}_\nu[R_{A_t,t}]) = T\overline{\theta} - \sum_{t=1}^{T} \theta_{A_t},$$

which depends on ν via the average rewards and it is a random quantity because the $\{A_t\}_{t \in [T]}$ are random. Note that the latter is true even if the policies considered are deterministic. The reason is that A_t depends on the history, which is random. Furthermore, the regret depends on $R_{A_t,t}$ independently of whether the rewards have been observed. In what follows, I omit the dependence of the regret on ν and simply write $\mathsf{Regret}_T(\pi)$.

Goal. The decision-maker's goal is to find a policy π with good worst-case regret properties. Formally, the decision-maker seeks to find a policy π whose worst-case regret $\overline{\mathsf{Regret}_T}(\pi; \mathcal{C}) := \sup_{\nu \in \mathcal{C}} \mathsf{Regret}_T(\pi; \nu)$ has the nearly optimal rate $\widetilde{O}(\sqrt{T})$. In what follows next, I will not directly optimize the worst case regret over the space of policy Π (see Adusumilli (2024) for an example of such an approach in a standard MAB). Rather, I first derive a lower bound for the minimax regret

$$\mathsf{Regret}^\star_T(\mathcal{C}) := \inf_{\pi \in \Pi} \overline{\mathsf{Regret}_T}(\pi; \mathcal{C}).$$

Then, I derive an upper bound for the worst-case regret $\overline{\mathsf{Regret}_T}(\widetilde{\pi}; \mathcal{C})$ of some specific policy $\widetilde{\pi}$ and, finally, I check that the rates of the two bounds coincide. In what follows, I focus on the popular UCB algorithm (Auer, Cesa-Bianchi and Fischer, 2002) and a novel, doubly-robust modification of it as the algorithms used by the decision-maker to form her policy.

2.1 Algorithms

The UCB algorithm selects the action to be played by solving an exploitation-exploration trade-off. Indeed, the algorithm desires to *explore* new actions, but, since playing a sub-optimal action induces regret, it also wants to *exploit* what is already known about the environment. On the one hand, too little exploration might make a sub-optimal alternative look better than the optimal one because of random fluctuations. On the other hand, too much exploration prevents the algorithm from playing the optimal alternative often enough, which also results in a larger regret.

Let $\hat{\theta}_a(t)$ be an estimator for θ_a after $t-1$ rounds and denote with $b_{a,t}(\delta)$ a bonus term chosen so that $\theta_a \in [\hat{\theta}_a(t) - b_{a,t}(\delta), \hat{\theta}_a(t) + b_{a,t}(\delta)]$ with probability at least $1-\delta$. The UCB algorithm selects the action $a^\star$ at round t that has the highest optimistic mean reward estimate, i.e.

$$a^\star = \arg\max_{a \in \mathcal{A}} \hat{\theta}_a(t) + b_{a,t}(\delta).$$

As such, an action $a \in \mathcal{A}$ can be chosen for two different reasons: because $b_{a,t}(\delta)$ is large, implying that the estimate $\hat{\theta}_a(t)$ is noisy (*explorative* choice); or because $\hat{\theta}_a(t)$ is large (*exploitative* choice). Since the bonus term $b_{a,t}(\delta)$ is constructed to shrink quickly each time alternative a is selected, exploration becomes less frequent over time. When $b_{a,t}(\delta)$ is sufficiently small, the estimated value $\hat{\theta}_a(t)$ closely approximates the true parameter θ_a, assuming that $\hat{\theta}_a(t)$ is a "good" estimator for θ_a. Consequently, UCB naturally balances between exploration and exploitation.

Before describing in great detail the algorithms, let

$$P_a(t) := \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a], \quad \text{and} \quad N_a(t) := \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a] C_{a,\ell}$$

be the number of times an arm $a \in \mathcal{A}$ has been pulled and the number of times the reward R_a has been observed at the beginning of round $t \in [T]$, respectively.

2.1.1 Classic UCB Algorithm

The (regularized) estimate for $\theta_a, a \in \mathcal{A}$ at the beginning of round $t \in [T]$ is

$$\widehat{R}_a^{\mathsf{UCB}}(t) = \frac{1}{N_a(t) + \lambda} \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a] C_{a,\ell} R_{a,\ell}, \tag{1}$$

where $\lambda > 0$ is a regularization parameter that prevents the estimator from being ill-defined whenever, after initialization, it occurs that $N_a(t) = 0$ for some $a \in \mathcal{A}$ and $t \geq 1$ (see also Remark 1). The optimistic mean reward estimate of action $a \in \mathcal{A}$ after $t \in [T]$ rounds is

$$\widetilde{R}_a^{\mathsf{UCB}}(t, \delta) = \widehat{R}_a^{\mathsf{UCB}}(t) + b_{a,t}^{\mathsf{UCB}}(\delta),$$

where the "bonus" term $b_{a,t}^{\mathsf{UCB}}(\delta)$ is chosen to make sure that the optimistic mean reward estimate $\widetilde{R}_a^{\mathsf{UCB}}(t, \delta)$ upper bounds the true mean reward θ_a with high probability. In this specific case, I define

$$b_{a,t}^{\mathsf{UCB}}(\delta) := \frac{\bar{\sigma}}{\underline{q}_\lambda} \sqrt{\frac{2 \ln(2AT/\delta)}{P_a(t) + \lambda}} + \frac{\lambda \overline{K}}{N_a(t) + \lambda},$$

where $\overline{K}$ is some constant larger than $\overline{\theta}, \underline{q}_\lambda := \inf_{a,t} \frac{N_a(t)+\lambda}{P_a(t)+\lambda}$, and $\delta \in (0,1)$. Intuitively, the first term in $b_{a,t}^{\mathsf{UCB}}(\delta)$ governs the probability with which the optimistic estimate overestimates the mean reward, whereas the second term takes into account the bias induced by the regularization term $\lambda > 0$. Under reward-independent missingness (Assumption 1 below), Lemma SA-6 in the supplemental appendix formally justifies the particular choice of $b_{a,t}^{\mathsf{UCB}}(\delta)$ described above by showing that

$$\forall\, a \in \mathcal{A}, t \in [T], \quad \theta_a \in \left[\widehat{R}_a^{\mathsf{UCB}}(t) - b_{a,t}^{\mathsf{UCB}}(\delta), \widehat{R}_a^{\mathsf{UCB}}(t) + b_{a,t}^{\mathsf{UCB}}(\delta)\right]$$

holds with probability at least $1 - \delta$.

The way the UCB algorithm works is straightforward: at round $t \in [T]$, it selects the arm a that has the highest optimistic mean reward estimate. Algorithm 1 below summarizes all the steps needed by the classic UCB algorithm.

Procedure 1 Update Estimators for UCB

```
for $a \in [A]$ do
    $N_a(t+1) \leftarrow N_a(t) + \mathbb{1}[A_t = a]C_{a,t}$
    $P_a(t+1) \leftarrow P_a(t) + \mathbb{1}[A_t = a]$
    $\widehat{R}_a(t+1) \leftarrow \frac{1}{N_a(t+1)+\lambda} \sum_{\ell=1}^{t} \mathbb{1}[A_\ell = a]C_{a,\ell}R_{a,\ell}$
    $b^{\mathsf{UCB}}_{a,t+1}(\delta) \leftarrow \frac{\bar{\sigma}}{\underline{q}_\lambda}\sqrt{\frac{2\ln(2AT/\delta)}{P_a(t+1)+\lambda}} + \frac{\lambda\overline{K}}{N_a(t+1)+\lambda}$
    $\widetilde{R}_a(t+1,\delta) \leftarrow \widehat{R}_a(t+1) + b^{\mathsf{UCB}}_{a,t+1}(\delta)$
end for
```

Algorithm 1 UCB algorithm

```
Input: $\lambda > 0, \underline{q}_\lambda, \bar{\sigma}, T, \mathcal{A}, \delta, \overline{K}$
Initialization: pull each arm once, get $\widehat{R}^{\mathsf{UCB}}_a(0)$, set $P_a(0) = 1, N_a(0) = C_{a,0}, \forall\, a \in \mathcal{A}$
for $t = 1, 2, \ldots, T$ do
    pull arm $a_t = \arg\max_{a\in\mathcal{A}} \widetilde{R}^{\mathsf{UCB}}_a(t,\delta)$ and set $\pi^{\mathsf{UCB}}_t = a_t$
    call Update Estimators for UCB (Procedure 1)
end for
Output: $\pi^{\mathsf{UCB}} = \{\pi^{\mathsf{UCB}}_t\}_{t\in[T]}$
```

Remark 1. I introduce the regularization parameter $\lambda > 0$ to take care of those cases in which $N_a(0) = 0$, where $t = 0$ denotes the initialization period. In the absence of missing data, i.e., when $N_a(t) = P_a(t)$ almost surely, it is common practice to pull each arm once as initialization. If the action set is finite, the initialization just shifts up the regret of a factor not larger than $A\overline{K}$. In the context studied here, pulling each arm once grants that $P_a(0) = 1$ for all $a \in \mathcal{A}$, but does not guarantee that $N_a(0) = 1$ for all $a \in \mathcal{A}$. Rather, the event $\{N_a(0) = 1, \forall\, a \in \mathcal{A}\}$ realizes with probability $\prod_{a\in\mathcal{A}} q_a \leq 1$, hence the need for regularization. ♣

2.1.2 Doubly-Robust UCB Algorithm

Let the true conditional mean reward and probability of not being missing for arm $a \in \mathcal{A}$ as

$$\theta_a(\mathbf{X}_a) := \mathbb{E}_\nu[R_a \mid \mathbf{X}_a], \qquad q_a(\mathbf{X}_a) = \mathbb{E}_\nu[C_a \mid \mathbf{X}_a] \in [\underline{q}, 1]$$

almost surely, and denote with $\hat{\theta}_a(\cdot)$ and $\hat{q}_a(\cdot)$ their estimated counterparts. Throughout, I use interchangeably the terms "probability of rewards not being missing" and "probability of missingness". The doubly-robust estimator for mean rewards is defined as

$$\widehat{R}^{\mathsf{DR}}_a(t) := \frac{1}{P_a(t)} \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a] \left(\frac{C_{a,\ell}(R_{a,\ell} - \hat{\theta}_a(\mathbf{X}_{a,\ell}))}{\hat{q}_a(\mathbf{X}_{a,\ell})} + \hat{\theta}_a(\mathbf{X}_{a,\ell}) \right).$$

The optimistic mean reward estimator for DR-UCB is defined as

$$\widetilde{R}^{\mathsf{DR}}_a(t,\delta) = \widehat{R}^{\mathsf{DR}}_a(t) + b^{\mathsf{DR}}_{a,t}(\delta), \qquad b^{\mathsf{DR}}_{a,t}(\delta) = K_{\mathsf{ODR}} \sqrt{\frac{2\ln(2AT/\delta)}{P_a(t)}} + b^{\mathsf{res}}_{a,t}(\delta), \tag{2}$$

where $K_{\mathsf{ODR}} := \frac{\bar{\sigma}}{\underline{q}} + \bar{\sigma}$ and $b^{\mathsf{res}}_{a,t}(\delta)$ is defined precisely in Section SA1.3.3 of the supplemental appendix. Under a reward-dependent process that causes missing data (Assumption 2 below), Lemma SA-12 in the supplemental appendix thoroughly justifies the choice of $b^{\mathsf{DR}}_{a,t}(\delta)$ as a bonus term and proves that

$$\forall\, a \in \mathcal{A}, t \in [T], \quad \theta_a \in \left[\widehat{R}^{\mathsf{DR}}_a(t) - b^{\mathsf{DR}}_{a,t}(\delta), \widehat{R}^{\mathsf{DR}}_a(t) + b^{\mathsf{DR}}_{a,t}(\delta) \right]$$

holds with probability at least $1-\delta$.

Finally, Algorithm 2 describes in greater detail how the DR-UCB algorithm works.

Procedure 2 Update Estimators for DR-UCB

```
for $a \in [A]$ do
    $N_a(t+1) \leftarrow N_a(t) + \mathbb{1}[A_t = a] C_{a,\ell}$
    $P_a(t+1) \leftarrow P_a(t) + \mathbb{1}[A_t = a]$
    Update $\hat{q}_a$ and $\hat{\theta}_a$ if required
    $\widehat{R}^{\mathsf{DR}}_a(t+1) := \frac{1}{P_a(t+1)} \sum_{\ell=1}^{t} \mathbb{1}[A_\ell = a] \left( \frac{C_{a,\ell}(R_{a,\ell} - \hat{\theta}_a(\mathbf{X}_{a,\ell}))}{\hat{q}_a(\mathbf{X}_{a,\ell})} + \hat{\theta}_a(\mathbf{X}_{a,\ell}) \right)$
    $\widetilde{R}^{\mathsf{DR}}_a(t+1,\delta) \leftarrow \widehat{R}^{\mathsf{DR}}_a(t+1) + b^{\mathsf{DR}}_{a,t}(\delta)$
end for
```

Algorithm 2 DR-UCB algorithm

```
Input: $\lambda > 0, T, \mathcal{A}, \delta, \{\hat{q}_a(\cdot), \hat{\theta}_a(\cdot)\}_{a \in \mathcal{A}}$
Initialization: pull each arm once, get $\widehat{R}^{\mathsf{DR}}_a(0)$ and set $P_a(0) = 1, N_a(0) = C_{a,0}, \forall\, a \in \mathcal{A}$
Nuisances: get estimates $\{\hat{q}_a(\mathbf{X}_{a,0}), \hat{\theta}_a(\mathbf{X}_{a,0})\}_{a \in \mathcal{A}}$ according to Assumption 3(iii)
for $t = 1, 2, \ldots, T$ do
    pull arm $a_t = \arg\max_{a \in \mathcal{A}} \widetilde{R}^{\mathsf{DR}}_a(t,\delta)$ and set $\pi^{\mathsf{DR}}_t = a_t$
    call Update Estimators for DR-UCB (Procedure 2)
end for
Output: $\pi^{\mathsf{DR}} = \{\pi^{\mathsf{DR}}_t\}_{t \in [T]}$
```

3 Multi-armed Bandits with Missing Data

In Section 3.1, I begin by analyzing the performance of the UCB algorithm under the assumption that the process that causes missing data does not depend on rewards (Assumption 1). In Section 3.2, I then show that, when the previous assumption is relaxed to allow for reward-dependent missingness (Assumption 2), the UCB algorithm can suffer from linear regret, whereas DR-UCB achieves sub-linear worst-case regret. In Section 3.3, I present a lower bound for the minimax regret. Finally, in Section 3.4, I conclude by giving some practical advice on how to implement DR-UCB.

3.1 Reward-independent Missingness

I begin by considering the scenario where rewards are missing at random, as formalized in Assumption 1. Additionally, rewards are assumed to be sub-Gaussian, a common condition that constrains their tail behavior and enables the use of standard concentration inequalities; see Vershynin (2018) and Wainwright (2019) for an introduction.

Assumption 1. *For each action $a \in \mathcal{A}, C_a \perp\!\!\!\perp R_a$.*

Under Assumption 1, the class of bandits of interest is restricted to

$$\mathcal{C}_1 := \left\{(\nu_a)_{a\in\mathcal{A}} : \nu_a^{[R]} \in \mathcal{SG}(\sigma_a),\ \ \nu_a^{[C]} = \mathsf{Be}(q_a), q_a \in (0,1],\ \ \nu_a = \nu_a^{[R]} \cdot \nu_a^{[C]}\right\} \subset \mathcal{C}.$$

The paper's first main result establishes that the UCB algorithm achieves a near-optimal worst-case regret rate over the class of bandits $\mathcal{C}_1$. The proof of such a result follows using standard arguments. Namely, I consider a "good" event and show that it occurs with high probability in the setting considered throughout. In this spirit, define such an event as

$$\mathcal{G}(\delta_1, \delta_2) = \overline{\mathcal{F}^{\mathsf{UCB}}(\delta_1)} \cap \overline{\mathcal{F}^{\mathsf{MIS}}(\delta_2)},$$

for some $\delta_1, \delta_2 \in (0,1)$, where

$$\mathcal{F}^{\mathsf{UCB}}(\delta) := \{\exists\, a \in \mathcal{A}, t \in [T], \left|\widehat{R}_a^{\mathsf{UCB}}(t) - \theta_a\right| \geq b_{a,t}^{\mathsf{UCB}}(\delta)\},$$

$$\mathcal{F}^{\mathsf{MIS}}(\delta) :=\{\exists\, a \in \mathcal{A}, t \in [T] : N_a(t) \leq (1-\delta)q_a P_a(t), P_a(t) \geq \underline{T}_a\},$$

where $\underline{T}_a := 1 + \frac{24\ln(T)}{q_a}$. Under the good event, which realizes with probability at least $1-\delta_1-\delta_2$, it holds that: (i) for each action, the optimistic reward estimator always covers the true mean; (ii) after a minimum amount of pulls $\underline{T}_a$, the missing data mechanism is not too extreme in terms of percentage deviation from its mean.

In the supplemental appendix, Lemma SA-6 and Lemma SA-7 show that $\mathcal{F}^{\mathsf{UCB}}(\delta_1)$ and $\mathcal{F}^{\mathsf{MIS}}(\delta_2)$ occur with arbitrarily small probability. Then, under the good event $\mathcal{G}(\delta_1, \delta_2)$, it is possible to bound the worst case regret of UCB uniformly over bandits in $\mathcal{C}_1$ and horizons $T \in \mathbb{N}$. This result is presented formally in the next theorem and proven in the supplemental appendix.

Theorem 1. *Let Assumption 1 hold,* $\lambda = o(T^{1/2}), \delta_1 \in (0,1), \delta_2 = \sqrt{\frac{1+\kappa}{12}}$, *and* $\kappa > 0$. *Then, for any* $T \in \mathbb{N}$ *and bandit* $\nu \in \mathcal{C}_1$

$$\mathsf{Regret}_T(\pi^{\mathsf{UCB}}) \lesssim \frac{4\bar{\sigma}}{\underline{q}_\lambda}\sqrt{2AT\ln(2AT/\delta_1)},$$

with probability at least $1-\delta_1 - O(T^{-\kappa})$.

3.2 Reward-dependent Missingness

The previous section showed that, under the assumption that the process that causes missing data does not depend on rewards, the UCB algorithm has nearly-optimal regret. However, this independence assumption is hard to defend in practical applications. Borrowing from the program evaluation literature, I relax Assumption 1 and, instead, rely on a *conditional ignorability assumption* (CIA), which states that independence holds only conditional on the vector of covariates $\mathbf{X}_a$ (Rosenbaum and Rubin, 1983).[1] Put differently, the CIA imposes that knowledge of $\mathbf{X}_a$ is sufficient to break the dependence between rewards and the missing data mechanism in each arm.

The CIA can be expressed in two ways: by imposing some structure on the conditional

[1] This is a standard assumption in the causal inference literature. I refer the interested reader to Imbens (2004) and Chapter 21 in Wooldridge (2010) for thorough discussions of the plausibility of such an assumption in various contexts.

expectation of C_a (so-called *design-based* approach); by imposing some structure on the conditional expectation of R_a (so-called *model-based* approach). Depending on the specific application, either version of the CIA might be more appealing. The next assumption formalizes this idea.

Assumption 2 (Model- and Design-based Ignorability)**.** *For each $a \in \mathcal{A}$, either*

$$\mathbb{E}_\nu[R_a \mid \mathbf{X}_a, C_a] = \mathbb{E}_\nu[R_a \mid \mathbf{X}_a] =: \theta_a(\mathbf{X}_a) \qquad a.s. \tag{\textbf{MB}}$$

or

$$\mathbb{E}_\nu[C_a \mid \mathbf{X}_a, R_a] = \mathbb{E}_\nu[C_a \mid \mathbf{X}_a] =: q_a(\mathbf{X}_a) \qquad a.s. \tag{\textbf{DB}}$$

holds. Moreover, $R_a \mid \mathbf{X}_a, \sim \mathsf{sG}(\sigma_a)$ and $q_a(\mathbf{x}) \in [\underline{q}, 1]$, for some constants $0 \leq \overline{K}_\theta < \infty$ and $\underline{q} \in (0, 1]$.

On top of the CIA, Assumption 2 requires sub-Gaussianity of rewards only conditional on $\mathbf{X}_a$ and has two other mild requirements: (i) the conditional expectation of each R_a is uniformly bounded over $\mathcal{X}$; (ii) the probability of observing rewards is non-zero ($q_a > 0$). Hence, the class of bandits considered throughout is

$$\mathcal{C}_2 := \left\{\nu = (\nu_a)_{a\in\mathcal{A}} : \nu_a^{[C]} = \mathsf{Be}(q_a), q_a \in (0,1], \text{ and Assumption 2 holds}\right\} \subset \mathcal{C}.$$

In this framework, it is well-known that the mean-reward estimator defined in (1) is not consistent anymore for θ_a (Horvitz and Thompson, 1952). Even more worryingly, the estimators $\widehat{R}_a^{\mathsf{UCB}}(t), a \in \mathcal{A}$ might have probability limits $\widetilde{\theta}_a$ such that

$$\arg\max_{a\in\mathcal{A}} \widetilde{\theta}_a \neq \arg\max_{a\in\mathcal{A}} \theta_a,$$

so that even after a large number of rounds T, the UCB algorithm will not learn the best arm. Section SA2.2.1 of the supplemental appendix shows an example of a bandit in $\mathcal{C}_2$ such that this realizes.

Assumption 2 ensures that θ_a can be identified from the data had the nuisance functions $\{\theta_a(\cdot), q_a(\cdot), a \in \mathcal{A}\}$ been known. However, in practice, these nuisances need to be estimated, and extra care is required in doing so. Before elucidating how nuisance estimation should be conducted, define the following ℓ_2-estimation errors for each

$a \in \mathcal{A}$

$$\mathsf{Err}_t(\hat{\theta}_a) := \sqrt{\frac{1}{P_a(t)} \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a](\hat{\theta}_a(\mathbf{X}_{a,\ell}) - \widetilde{\theta}_a(\mathbf{X}_{a,\ell}))^2}\,,$$

and

$$\mathsf{Err}_t(\hat{q}_a) := \sqrt{\frac{1}{P_a(t)} \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a](\hat{q}_a(\mathbf{X}_{a,\ell}) - \widetilde{q}_a(\mathbf{X}_{a,\ell}))^2}$$

for some known $\widetilde{q}_a(\mathbf{x})$ and $\widetilde{\theta}_a(\mathbf{x})$.

Assumption 3 precisely states all the requirements the nuisance estimators have to satisfy; see also Section 3.4 for two examples of practical procedures that satisfy Assumption 3.

Assumption 3 (Nuisance Estimation)**.** *For each $a \in \mathcal{A}$, the following are true:*

(a) *(double robustness) either $\widetilde{q}_a(\mathbf{x}) = q_a(\mathbf{x})$ or $\widetilde{\theta}_a(\mathbf{x}) = \theta_a(\mathbf{x})$;*

(b) *(truncation) $\forall\, \mathbf{x} \in \mathcal{X}, \widehat{q}_a(\mathbf{x}) \in [\underline{q}, 1], \underline{q} \in (0, 1]$;*

(c) *(independence) $(\hat{q}_a(\mathbf{X}_a), \hat{\theta}_a(\mathbf{X}_a)) \perp\!\!\!\perp (R_a, C_a) \mid \mathbf{X}_a$;*

(d) *(ℓ_2-error rate) there exist rates $\alpha > 1/2, \alpha_q > 0$, and $\alpha_\theta > 0$ such that*

$$\mathsf{Err}_t(\hat{q}_a) \lesssim \frac{1}{P_a(t)^{\alpha_q}}, \qquad \mathsf{Err}_t(\hat{\theta}_a) \lesssim \frac{1}{P_a(t)^{\alpha_\theta}}, \qquad \mathsf{Err}_t(\hat{q}_a)\mathsf{Err}_t(\hat{\theta}_a) \lesssim \frac{1}{P_a(t)^{\alpha}}$$

with probability $1 - \delta_\mathfrak{c}, \delta_\mathfrak{c} \in (0, 1)$.

First, Assumption 3(a) requires one between the true conditional probability of missingness, $q_a(\cdot)$, and the true conditional expectation of rewards, $\theta_a(\cdot)$, to be the probability limit of one of the nuisance estimators, $\hat{\theta}_a(\cdot)$ and $\hat{q}_a(\cdot)$. In other words, it suffices to have well-specified only one of the two conditional expectations. Second, Assumption 3(b) bounds the estimated probability of (not) being missing away from 0, a typical regularity condition in such problems. Third, to avoid over-fitting bias, Assumption 3(c) asks the nuisance functions to be estimated in an independent (conditional on $\mathbf{X}_a$) sample. Fourth, 3(d) controls the estimation error of the nuisance estimators in two ways: (i) the estimation error of each nuisance need to be shrinking in $P_a(t)$; (ii)

the product of the estimation errors must decay faster than $1/\sqrt{P_a(t)}$. These conditions make the sampling error dominate the estimation error induced by the fact that $\{(\theta_a(\cdot), q_a(\cdot)), a \in \mathcal{A}\}$ are estimated.

Assumption 3 is fundamental to make the term $b^{\mathsf{res}}_{a,t}(\delta)$ of higher order in the bonus term $b^{\mathsf{DR}}_{a,t}(\delta)$ defined in (2). Formally, under such an assumption, it follows that $b^{\mathsf{res}}_{a,t}(\delta) = o_{\mathbb{P}}(1/\sqrt{P_a(t)})$ and the next lemma follows.

Lemma 1. *Let Assumptions 2 and 3 hold, $\delta \in (0,1), a \in \mathcal{A}$, and $t \in [T]$. Then,*

$$\left|\widehat{R}^{\mathsf{DR}}_a(t) - \theta_a\right| \geq b^{\mathsf{DR}}_{a,t}(\delta A T)$$

with probability at most δ.

The above result shows that the bonus term for the DR-UCB algorithm has been chosen appropriately in the sense that it controls the probability with which $\widehat{R}^{\mathsf{DR}}_a(t)$ deviates from θ_a. This result is of independent interest as it is the first one that provides high-probability bounds for a doubly-robust estimator under mild assumptions; see Lemma SA-12 in the supplemental appendix for a formal statement, a proof of the result, and some heuristics on the logic behind the strategy used in the proof.

To provide a bound on the worst-case regret over the class $\mathcal{C}_2$, a similar strategy to the one used in Section 3.1 is adopted. Define the failure event

$$\mathcal{F}^{\mathsf{DR}}(\delta) := \left\{\exists\, a \in \mathcal{A}, t \in [T], \left|\widehat{R}^{\mathsf{DR}}_a(t) - \theta_a\right| \geq b^{\mathsf{DR}}_{a,t}(\delta)\right\}.$$

When $\mathcal{F}^{\mathsf{DR}}(\delta)$ occurs the optimistic doubly-robust reward estimator $\widetilde{R}^{\mathsf{DR}}_a(t) = \widehat{R}^{\mathsf{DR}}_a(t) + b^{\mathsf{DR}}_{a,t}(t)$ does not cover the true mean reward θ_a. Using Lemma 1, it is immediate to see that $\mathbb{P}[\mathcal{F}^{\mathsf{DR}}(\delta)] \leq \delta$ for some $\delta \in (0,1)$. The next theorem shows that the regret of the DR-UCB algorithm is nearly optimal (up to logarithmic factors) and provides an upper bound that holds with high probability uniformly over the horizon T and the class of bandits $\mathcal{C}_2$.

Theorem 2. *Let Assumptions 2 and 3 hold with $\delta \in (0,1)$ and $\delta_{\mathfrak{c}} \in (0,1)$. Then, for any* *horizon $T \in \mathbb{N}$ and bandit $\nu \in \mathcal{C}_2$*

$$\mathsf{Regret}_T(\pi^{\mathsf{DR}}) \lesssim \frac{4\bar{\sigma}}{\underline{q}} \sqrt{AT \ln(2AT/\delta)}$$

with probability $1 - \delta - \delta_{\mathfrak{c}}$.

Comparing Theorem 2 with Theorem 1, it is possible to see that the leading order terms of the worst case regret bound are identical. Indeed, Assumption 3(d) is crucial in granting that the uncertainty due to the estimation of the nuisance functions is dominated by the one due to sampling error; see the supplemental appendix for a formal proof and additional heuristics to foster intuition for this result.

3.3 Lower Bound

In this section, I show that the minimax regret

$$\mathsf{Regret}^{\star}_T(\mathcal{C}_j) := \inf_{\pi \in \Pi} \sup_{\nu \in \mathcal{C}_j} \mathsf{Regret}_T(\pi; \nu), \; j \in \{1, 2\}$$

is lower bounded by a constant times a factor of $\sqrt{T}$.

The logic of the proof is similar to that behind classic lower bound results obtained via Le Cam's two-point lemma. In particular, a generic policy $\tilde{\pi} \in \Pi$ is considered and its regret on two particular instances $\nu, \nu' \in \mathcal{C}$ is lower-bounded, where $\mathcal{C}$ is some class of bandits. Then, it follows that

$$\sup_{\tilde{\nu} \in \mathcal{C}} \mathsf{Regret}_T(\tilde{\pi}; \tilde{\nu}) \geq \max\{\mathsf{Regret}_T(\tilde{\pi}, \nu), \mathsf{Regret}_T(\tilde{\pi}, \nu')\} \geq f(T),$$

for some function $f(\cdot)$. Because the policy $\tilde{\pi}$ was generic, then the bound above holds for any policy in Π, hence

$$\mathsf{Regret}^{\star}_T(\mathcal{C}) = \inf_{\pi \in \Pi} \sup_{\nu \in \mathcal{C}} \mathsf{Regret}_T(\pi; \nu) \geq f(T).$$

In particular, to get a lower bound for the minimax regret over the classes of bandits $\mathcal{C}_1$ and $\mathcal{C}_2$, I derive a lower bound for the minimax regret for the classes of Gaussian bandits

$$\mathcal{C}_1^{\mathsf{gau}} := \{(\nu_a)_{a \in \mathcal{A}} : \nu_a^{[R]} = \mathsf{N}(\theta_a, 1), \nu_a^{[C]} = \mathsf{Be}(q_a), q_a \in (0, 1]\} \subset \mathcal{C}_1$$

and

$$\mathcal{C}_2^{\mathsf{gau}} := \big\{\nu = (\nu_a)_{a\in\mathcal{A}} : \nu_a^{[R|X]} = \mathsf{N}(\theta_a(X), 1), \nu_a^{[C]} = \mathsf{Be}(q_a), q_a \in (0,1],$$
$$\text{and Assumption 2 holds}\big\} \subset \mathcal{C}_2.$$

Because the "sup" of the minimax is taken over a smaller subset, the minimax bound for Gaussian bandits extends immediately to the classes of sub-Gaussian bandits $\mathcal{C}_1$ and $\mathcal{C}_2$. A textbook example can be found in Lattimore and Szepesvári (2020), Chapter 15, and the proof of Theorem 3 is almost identical to that of Theorem 15.2 in LS and, indeed, the bound is identical.

Theorem 3. *Let $T \in \mathbb{N}, T \geq A - 1$ and consider the classes of bandits $\mathcal{C}_1$ and $\mathcal{C}_2$. Then,*

$$\mathsf{Regret}_T^\star(\mathcal{C}_j) = \inf_{\pi\in\Pi} \sup_{\nu\in\mathcal{C}_j} \mathsf{Regret}_T(\pi;\nu) \geq \frac{\sqrt{T(A-1)}}{16\sqrt{e}}.$$

3.4 Nuisance Estimation

The near-optimality of DR-UCB hinges on having nuisance estimators $\hat{\theta}_a(\cdot)$ and $\hat{q}_a(\cdot)$ that satisfy Assumption 3. In what follows, I give two examples of classes of estimators that satisfy the conditions in Assumption 3 and give practical advice for implementation.

First, I start by highlighting a classical trade-off between Assumptions 3(a) and 3(d): relying on flexible nonparametric estimators makes Assumption 3(a) more likely to be satisfied than when parametric estimators are used; however, nonparametric estimators typically have slower convergence rates than their parametric competitors, thus making 3(d) harder to be satisfied. As an instance, machine learning methods –such as lasso, ridge, random forests, neural networks– can be used to estimate the nuisances $\{(\theta_a(\cdot), q_a(\cdot)), a \in \mathcal{A}\}$, as long as their convergence rate (or ℓ_2 error bounds) decay at a faster rate than $1/\sqrt{P_a(T)}$. For some real-life applications and discussions of the feasibility of these methods, see Ahrens, Chernozhukov, Hansen, Kozbur, Schaffer and Wiemann (2025).

I now turn to Assumption 3(c), but before discussing it, it is necessary to introduce some notation. Define the main estimation dataset as

$$\mathcal{D}^T_{\texttt{main}} := \left\{\mathbf{Z}_\ell, \ell \in [T]\right\}, \qquad \mathbf{Z}_j := (R_{A_j,j}, C_{A_j,j}, \mathbf{X}^\top_{A_j,j})^\top.$$

Appropriate nuisance estimators $\hat{\theta}_a$ and $\hat{q}_a$ that satisfy Assumption 3(c) can be constructed in (at least) two ways:

M1 *Different batch.* Use samples from a dataset $\mathcal{D}_{\texttt{M1}}$ such that

$$\mathcal{D}_{\texttt{M1}} := \left\{\mathbf{Z}_\ell, \ell \notin [T]\right\}.$$

For example, suppose two similar decision-makers are interacting with two copies of the same bandit with missing data. Let the datasets generated by their interactions with the MAB be denoted with $\mathcal{D}^{T_j}_{\texttt{main},j}, j = 1, 2$. Then, $\hat{\theta}_a$ and $\hat{q}_a$ can be estimated for the first decision-maker using $\mathcal{D}^{T_2}_{\texttt{main},2}$ and for the second decision-maker using $\mathcal{D}^{T_1}_{\texttt{main},1}$.

M2 *Leave-one-out.* When updating $\widehat{R}^{\mathsf{DR}}_a(t)$ at the end of round $t-1$, one can use estimators $\hat{\theta}_a$ and $\hat{q}_a$ that use the dataset $\mathcal{D}^{t-2}_{\texttt{main}}$, i.e.,

$$\mathcal{D}^{t-2}_{\texttt{main}} := \{\mathbf{Z}_\ell, \ell = 1, \dots, t-2\}.$$

In words, $\hat{\theta}_a$ and $\hat{q}_a$ are constructed online using the "leave-one-out" principle, where the sample left out corresponds to the last observed sample of data.

As a final remark, I stress that in virtue of approaches such as M2, I should be writing $\hat{\theta}^{(\ell)}_a$ and $\hat{q}^{(\ell)}_a$ for each round $\ell \in [T]$. I avoid doing so to save notation, but it is maintained that the nuisance estimators are allowed to be updated with the rounds.

Third, the estimated conditional probability of missingness needs to be bounded between $\underline{q}$ and 1 to avoid dealing with "small denominators", i.e., when some of the $\hat{q}_a$'s are close to 0. Various alternatives exist in the literature: Crump, Hotz, Imbens and Mitnik (2009) proposes a data-driven trimming procedure that minimizes the variance of the estimator, and it is optimal under homoskedasticity (see Khan and Ugander (2025) for a recent extension to the heteroskedastic case); (Ma and Wang, 2020) warn against ad hoc trimming and propose a data-driven procedure that minimizes the asymptotic mean squared error of the resulting estimator. I stress that, while these procedures have been extensively validated via simulations, they rely on asymptotic guarantees and should therefore be applied with caution in finite-sample settings such as the one analyzed here.

4 Simulation Evidence

In this section, I present simulation results for the performance of the algorithms discussed. In the simulation, for each $a \in \mathcal{A}$, I model $X_{a,j} \overset{\text{iid}}{\sim} \mathsf{N}(0,1), j = 1, \dots, d, u_{a,j} \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma_j^2), j \in \{C, R\}$, and define

$$C_a = \mathbb{1}\left[\sum_{\ell=1}^{d} X_{a,\ell}\beta_\ell + u_{a,C} > \tau(q_a)\right], \quad R_a = \theta_a + \sum_{\ell=1}^{d} X_{a,\ell}\beta_\ell + u_{a,R},$$

where $\boldsymbol{\beta} := (\beta_1, \dots, \beta_d)^\top \in \mathbb{R}_{++}$ and $\tau(q_a) : \mathbb{P}[\sum_{\ell=1}^{d} X_{a,\ell}\beta_\ell + u_{a,C} > \tau(q_a)] = q_a$. Note that

$$R_a \sim \mathsf{N}(\theta, d + \sigma_R^2), \qquad C_a \sim \mathsf{Be}(q_a), \qquad C_a \perp\!\!\!\perp R_a \mid \mathbf{X}_a.$$

I then set $A = 2, T = 5000, d = 1$, and consider three different scenarios: no missing data; reward-independent missingness; and reward-dependent missingness. Table 1 reports the most important characteristics of these scenarios, which are the true mean rewards (θ_1, θ_2), the probability limit of the mean reward estimator that uses only observed rewards $(\widetilde{\theta}_1, \widetilde{\theta}_2)$, and the probability of missingness. More details about the simulation, estimation of nuisance functions, and parametrization can be found in the supplemental appendix Section SA3.

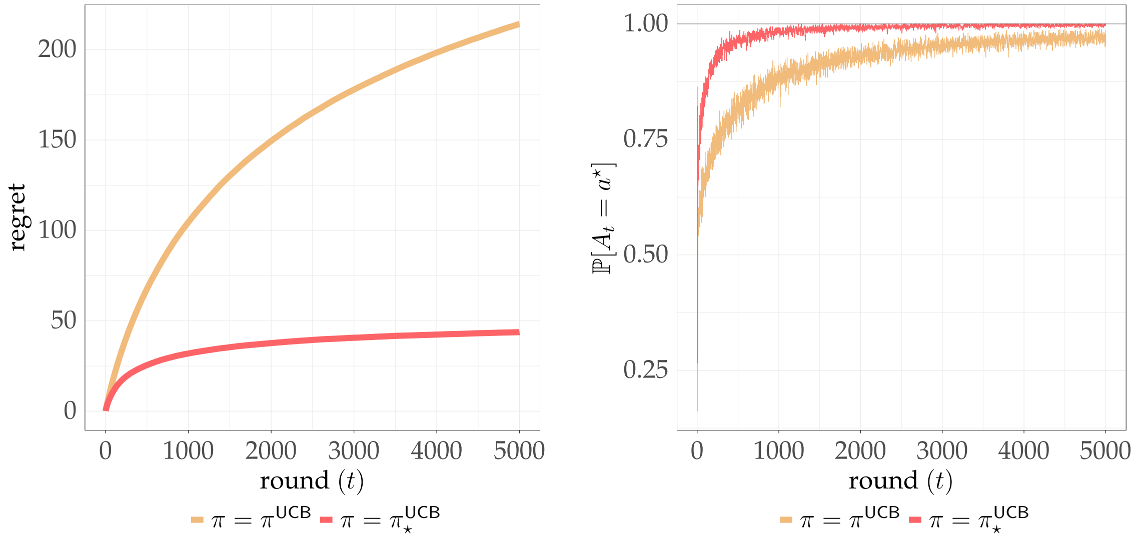
Table 1: Details of the three simulated scenarios.

MAB	(θ_1, θ_2)	$(\widetilde{\theta}_1, \widetilde{\theta}_2)$	(q_1, q_2)
Standard	$(0.5, 1)$	$(0.5, 1)$	$(1, 1)$
w/ independent missingness (Section 3.1)	$(0.5, 1)$	$(0.5, 1)$	$(0.25, 0.9)$
w/ dependent missingness (Section 3.2)	$(0.5, 1)$	$(1.16, 1.08)$	$(0.25, 0.9)$

In the first two scenarios, there is no wedge between the probability limits $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ and the true mean reward, suggesting that the UCB algorithm would work fine. This is exactly what can be deduced from Figure 1 and Figure 2.

The left panel of those figures shows the instance-specific regret of the UCB algorithm, which exhibits the classical logarithmic shape in the number of rounds. As a benchmark, the performance of UCB is compared with an oracle version of the algorithm, which does not rely on the optimistic mean reward estimator $\widetilde{R}_a^{\mathsf{UCB}}(t;\delta)$, but rather on a confidence interval constructed as if the data-generating process was known. The right panel of Figure 1 and Figure 2 portrays the probability with which each algorithm selects the optimal arm $a^\star = 2$ and shows that it approaches one as the number of rounds grows large. Notably, there is not much of a difference between scenario 1 and scenario 2. The only exception is that UCB needs more rounds to discover the optimal arm, consistent with the fact that rewards are not observed at all rounds.

Fig. 1: Regret and optimal arm selection - No missingness

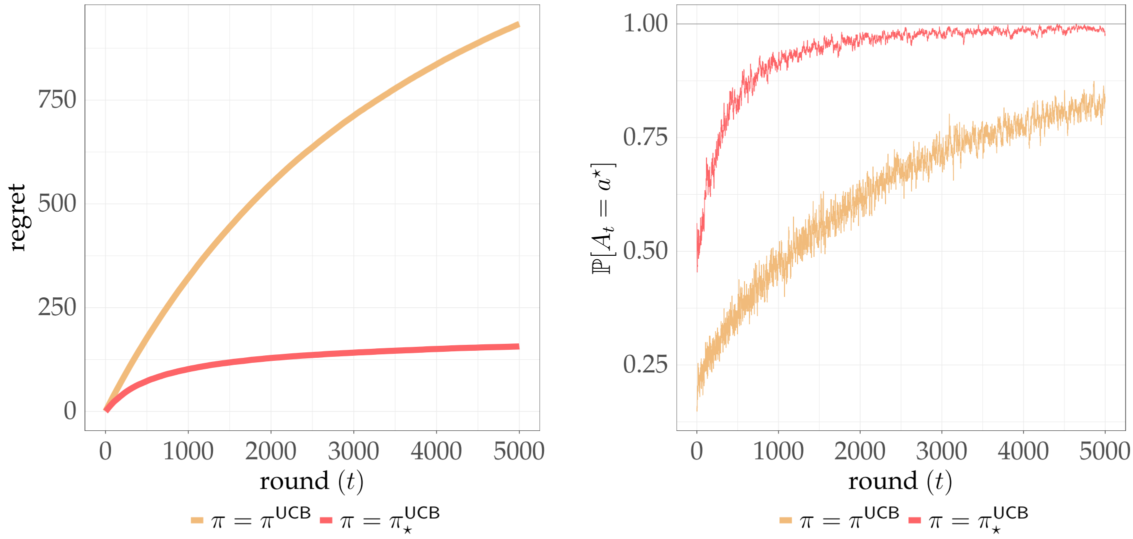


Notes: the left panel shows the cumulative regret averaged over $S = 500$ draws of a bandit parametrized according to the specifics of scenario 1. A similar exercise has been conducted in the right panel to plot the probability with which each policy selects the optimal arm. The algorithm behind the policy π^{UCB} is described in Algorithm 1, whereas the one behind $\pi_\star^{\mathsf{UCB}}$ is described in Section SA3.2 of the supplemental appendix.

Sensible differences emerge under the last scenario. Indeed, the parametrization under the third scenario has been chosen so that

$$\theta_1 < \theta_2 \quad \text{but} \quad \widetilde{\theta}_1 > \widetilde{\theta}_2,$$

Fig. 2: Regret and optimal arm selection - reward-independent missingness



Notes: the left panel shows the cumulative regret averaged over $S = 500$ draws of a bandit parametrized according to the specifics of scenario 2. A similar exercise has been conducted in the right panel to plot the probability with which each policy selects the optimal arm. The algorithm behind the policy π^{UCB} is described in Algorithm 1, whereas the one behind $\pi_{\star}^{\mathsf{UCB}}$ is described in Section SA3.2 of the supplemental appendix.

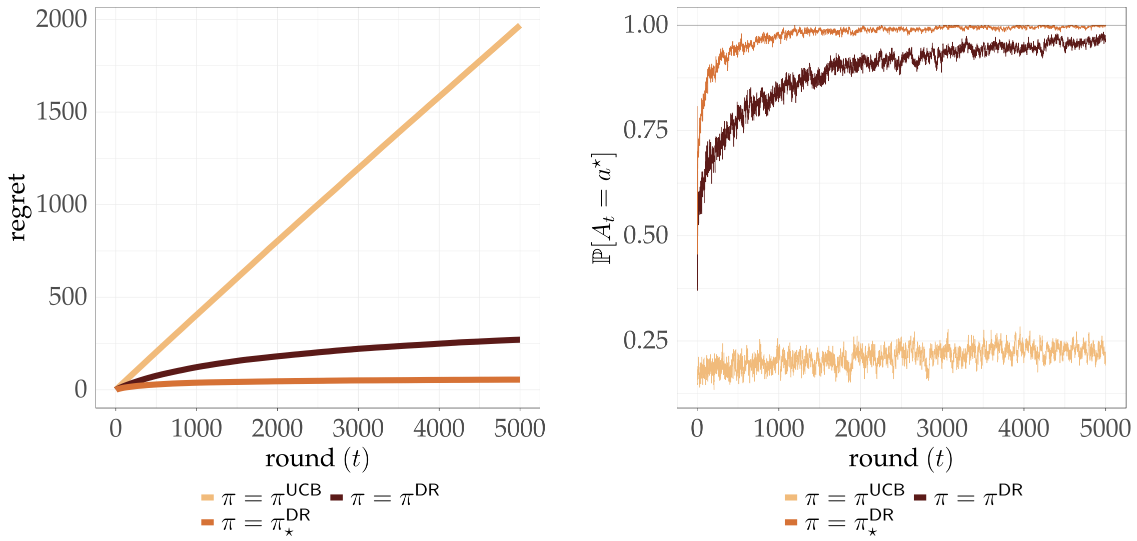
which implies that UCB will select with high probability $a = 1$, which is not the best action. On the contrary, DR-UCB should still be able to pick the best arm $a^{\star} = 2$. Figure 3 demonstrates that this intuition is indeed correct, together with the theoretical results presented in Section 3.

Most importantly, the one under scenario 3 is an instance of a bandit in $\mathcal{C}_2$ that makes the regret of π^{UCB} grow linearly with the number of rounds. Indeed, the vanilla UCB algorithm selects the correct action only a quarter of the time after 5,000 rounds. If the number of rounds were to increase, this probability would slowly approach 0. Intuitively, it takes time for the UCB algorithm to stick to the suboptimal arm $a^{\star} = 1$ because the probability limits $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ are very close to each other.

On the other hand, the DR-UCB shows logarithmic regret, and its probability of selecting the optimal arm quickly approaches one. Even in Figure 3, I also display the performance of an oracle algorithm that does not use the bonus term $b_{a,t}^{\mathsf{DR}}(\delta)$, but instead leverages knowledge of the underlying data-generating process.

Finally, Figure 4 illustrates why DR-UCB works, whereas UCB does not under reward-

Fig. 3: Regret and optimal arm selection - Reward-dependent missingness



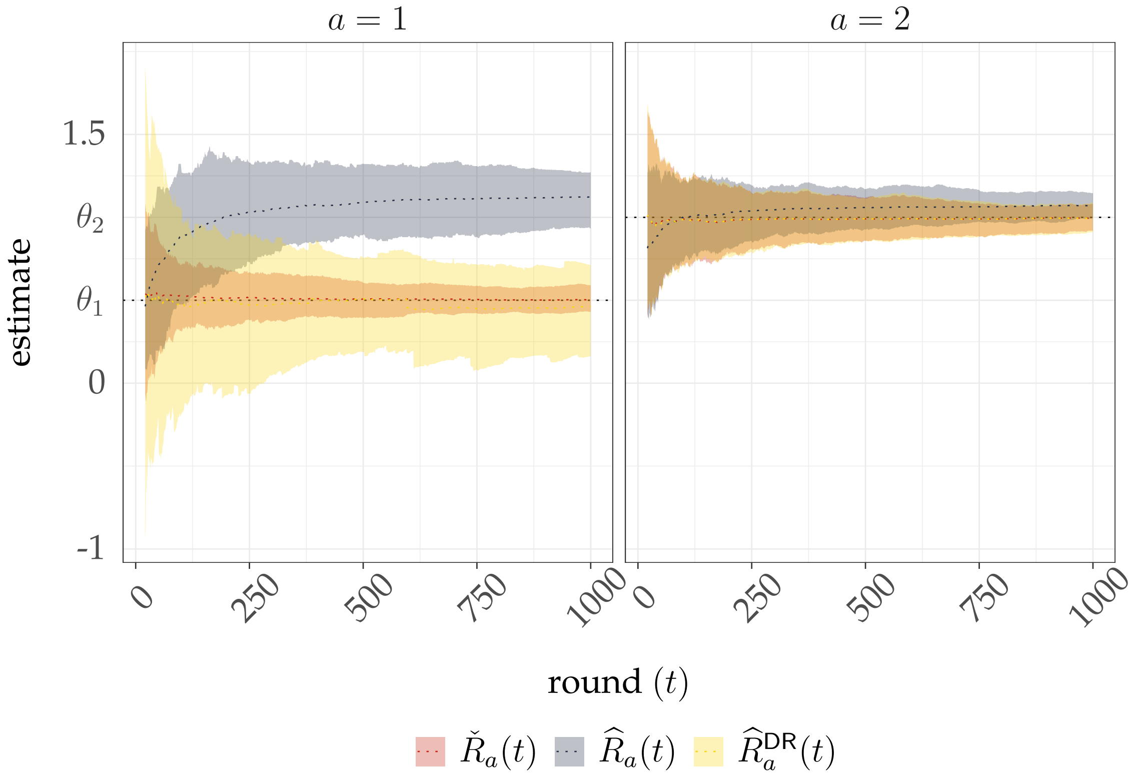
Notes: the left panel shows the cumulative regret averaged over $S = 500$ draws of a bandit parametrized according to the specifics of scenario 3. A similar exercise has been conducted in the right panel to plot the probability with which each policy selects the optimal arm. The algorithm behind the policy π^{UCB} is described in Algorithm 1, the one to implement the policy π^{DR} in Algorithm 2, whereas the one behind $\pi^{\mathsf{DR}}_{\star}$ is described in Section SA3 of the supplemental appendix.

dependent missingness. The graph plots the mean value, along with the 2.5th and 97.5th percentiles of the values obtained by three estimators across 1,000 draws of a bandit parametrized as in scenario 3. The three estimators are: the estimator that uses only observed rewards, $\widehat{R}_a(t)$; the doubly-robust estimator, $\widehat{R}^{\mathsf{DR}}_a(t)$; and an oracle estimator that always observes rewards, $\check{R}_a(t)$.

The right panel showcases the results for action $a = 2$. Under this arm, the difference between the true mean reward $\theta_2 = 1$ and the probability limit of $\widehat{R}_2(t)$, $\widetilde{\theta}_2 = 1.08$ is minimal, but still detectable from the graph. As expected, the doubly-robust and oracle estimators both quickly converge to the true mean reward θ_2. The left panel portrays the differences between the estimators in a neater way. The naïve estimator $\widehat{R}_1(t)$ rapidly approaches $\widetilde{\theta}_1 = 1.16$. The other two estimators converge to $\theta_1 = 0.5$. The larger uncertainty of the doubly-robust estimator, when compared to the oracle, comes from the fact that it estimates the nuisance functions, and because the incidence of missing data is strong under this arm, $q_1 = 0.2$.

Comparing the two panels also shows that the UCB algorithm flips the true ordering of mean rewards. Put differently, UCB picks with increasing probability $a = 1$, because $\widetilde{\theta}_1 > \widetilde{\theta}_2$. This is nothing more than a standard sample selection problem, where the selection occurs on dimensions that are (possibly directly) related to the outcomes of interest.

Fig. 4: Mean reward estimators behavior



Notes: dotted lines are the average value taken by an estimator across 500 draws of a bandit parametrized as under scenario 3; shaded areas are bounded between the 2.5th and 97.5th percentiles. Horizontal black lines indicate the values of the true mean rewards θ_a.

5 Conclusion

This paper examined a sequential decision-making problem in which feedback may be missing when the decision-maker interacts with the environment. I showed that standard methods—most notably the popular UCB algorithm—incur linear minimax regret across a wide range of such problems. In contrast, the proposed DR-UCB algorithm matches the optimal minimax regret rate (up to logarithmic factors), as established by a new lower bound for this problem class. I also provide practical guidance for implementing DR-UCB. Extending this framework to more general settings—such as contextual bandits or models with time-dependent feedback—constitutes a promising avenue for future work.

References

Adusumilli, Karun, “Risk and Optimal Policies in Bandit Experiments,” January 2024.

Ahrens, Achim, Victor Chernozhukov, Christian Hansen, Damian Kozbur, Mark Schaffer, and Thomas Wiemann, “An Introduction to Double/Debiased Machine Learning,” April 2025.

Athey, Susan and Stefan Wager, “Policy Learning With Observational Data,” *Econometrica*, 2021, *89* (1), 133–161.

Auer, Peter, Nicolò Cesa-Bianchi, and Paul Fischer, “Finite-Time Analysis of the Multiarmed Bandit Problem,” *Machine Learning*, May 2002, *47* (2), 235–256.

Bang, Heejung and James M. Robins, “Doubly Robust Estimation in Missing Data and Causal Inference Models,” *Biometrics*, December 2005, *61* (4), 962–973.

Bubeck, Sébastien and Nicolò Cesa-Bianchi, “Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems,” *Foundations and Trends in Machine Learning*, 2012, *5* (1), 1–122.

Cattaneo, Matias D., “Efficient Semiparametric Estimation of Multi-Valued Treatment Effects under Ignorability,” *Journal of Econometrics*, April 2010, *155* (2), 138–154.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, “Double/Debiased Machine Learning for Treatment and Structural Parameters,” *The Econometrics Journal*, February 2018, *21* (1), C1–C68.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik, “Dealing with Limited Overlap in Estimation of Average Treatment Effects,” *Biometrika*, January 2009, *96* (1), 187–199.

Farrell, Max H., “Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations,” *Journal of Econometrics*, November 2015, *189* (1), 1–23.

Freedman, David A., “On Tail Probabilities for Martingales,” *The Annals of Probability*, February 1975, *3* (1).

Horvitz, D. G. and D. J. Thompson, “A Generalization of Sampling Without Replacement from a Finite Universe,” *Journal of the American Statistical Association*, December 1952, *47* (260), 663–685.

Imbens, Guido W., “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,” *The Review of Economics and Statistics*, February 2004, *86* (1), 4–29.

Kallus, Nathan, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou, “Doubly Robust Distributionally Robust Off-Policy Evaluation and Learning,” July 2022.

Khan, S and J Ugander, “Doubly Robust and Heteroscedasticity-Aware Sample Trimming for Causal Inference,” *Biometrika*, March 2025, *112* (2), asae053.

Lai, T.L and Herbert Robbins, “Asymptotically Efficient Adaptive Allocation Rules,” *Advances in Applied Mathematics*, March 1985, *6* (1), 4–22.

Lai, Tze Leung, “Adaptive Treatment Allocation and the Multi-Armed Bandit Problem,” *The Annals of Statistics*, 1987, *15* (3), 1091–1114.

Lancewicki, Tal, Shahar Segal, Tomer Koren, and Yishay Mansour, "Stochastic Multi-Armed Bandits with Unrestricted Delay Distributions," in "Proceedings of the 38th International Conference on Machine Learning" PMLR July 2021, pp. 5969–5978.

Lattimore, Tor and Csaba Szepesvári, *Bandit Algorithms*, 1 ed., Cambridge University Press, July 2020.

Ma, Xinwei and Jingshen Wang, "Robust Inference Using Inverse Probability Weighting," *Journal of the American Statistical Association*, October 2020, *115* (532), 1851–1860.

Robbins, Herbert, "Some Aspects of the Sequential Design of Experiments," *Bulletin of the American Mathematical Society*, 1952, *58* (5), 527–535.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao, "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 1994, *89* (427), 846–866.

Rosenbaum, Paul R. and Donald B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 1983, *70* (1), 41–55.

Rubin, Donald B., "Inference and Missing Data," *Biometrika*, 1976, *63* (3), 581–592.

Shen, Ye, Hengrui Cai, and Rui Song, "Doubly Robust Interval Estimation for Optimal Policy Evaluation in Online Learning," *Journal of the American Statistical Association*, October 2024, *119* (548), 2811–2821.

Thompson, William R., "On the Likelihood That One Unknown Probability Exceeds Another in View of the Evidence of Two Samples," *Biometrika*, December 1933, *25* (3/4), 285.

Vershynin, Roman, *High-Dimensional Probability: An Introduction with Applications in Data Science*, 1 ed., Cambridge University Press, September 2018.

Wainwright, Martin J., *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, 1 ed., Cambridge University Press, February 2019.

Wald, Abraham, *Sequential Analysis*, Wiley, New York, 1947.

Wooldridge, Jeffrey M., *Econometric Analysis of Cross Section and Panel Data*, 2nd ed ed., Cambridge, Mass: MIT Press, 2010.