Online Learning for Repeated Nudging

Anand Kalvit
Stanford University, akalvit@stanford.edu

Divya Singhvi
NYU Leonard N. Stern School of Business, divya.singhvi@stern.nyu.edu

We consider the problem of optimal repeated user nudging on online platforms when nudge effectiveness is unknown and repeated use of the same nudge over time reduces its effectiveness. We model the optimal nudging problem as an online learning problem with $\mathcal{K}$-types (corresponding to different nudge-types), bandit feedback and non-stationary rewards. Furthermore, our model also incorporates costs of designing new nudges which are essential to ensure that they remain effective over time. We show that in the full information setting (when all the model parameters are known), a cyclic policy which regenerates arms of a single type after a fixed interval is optimal for maximizing the long-run-average-reward. Somewhat surprisingly, we find that this cyclic policy incurs constant regret (independent of time) even in the finite time setting. Leveraging ideas from this analysis, we reduce the online learning problem of optimizing repeated nudges to learning the optimal nudge-type and the corresponding cycle-length and construct a Upper Confidence Bound (UCB) based algorithm that incurs sublinear regret ($\tilde{\mathcal{O}}(\sqrt{T})$) which is rate-optimal in this setting. Numerical experiments based on both synthetic data as well as a model calibrated with real-world data in an EdTech setting show considerable improvement over benchmark methods and demonstrate the applicability of the proposed framework.

Key words: $\mathcal{K}$-Type Bandits, Optimal Nudging, Social-good

1. Introduction

Ensuring consistent user engagement is one of the foremost challenges that platforms face today. With the average American spending about 7 hours per day interacting with digital media[1], competition for users' attention is fierce, even for platforms designed to improve lives. For example, on EdTech platforms, student retention in online courses can be as low as 5-10% (Reich and Ruipérez-Valiente 2019). Similarly, in digital health, a recent study found that while 80% of adults were willing to try digital health tools, only 10% had actually started using them (Rock Health 2020).

To address these engagement issues, platforms employ various techniques, from providing personalized content recommendations to monetary discounts. Another commonly used approach is to utilize various forms of digital nudges - subtle design elements that guide user behavior without restricting choice. Common nudges include personalized notifications, progress trackers, social comparisons, and gamification elements (Sunstein et al. 2022). Digital nudges have shown to be highly

[1] `https://datareportal.com/reports/digital-2022-global-overview-report`

effective in various settings (see, e.g., Kelders et al. (2012) for studies in healthcare and Goyal et al. (2024) and references therein for recent studies in EdTech). Increased engagement in these sectors can lead to better learning outcomes and improved health behaviors. Furthermore, digital nudges are relatively cheap in comparison to other forms of user engagement related interventions (e.g., monetary compensation). Hence, an efficient digital nudging strategy could be critical to the success of a platform.

However, estimating efficient user nudging strategies can be challenging due to various reasons. Nudges can highly vary in their effectiveness. For example, recent studies in health-care and education (Chang et al. 2023, Nazaret and Sapiro 2023, Milkman et al. 2024, Agrawal et al. 2023, Goyal et al. 2024) have shown that nudges have highly differential effect with some nudges being highly effective, while others showing very limited effect. Hence, platforms need to estimate nudge preferences and design nudging strategies accordingly. To estimate nudge preferences, platforms can either leverage existing data when available, or dynamically learn preferences through efficient data-collection. This preference-learning and efficient nudging problem can be framed as a Bandit problem (Chu et al. 2011). The bandit framework has a long history and researchers have leveraged this framework in various contexts including pricing, donations, recommendations and health-care amongst others (e.g., Li et al. 2010, Besbes and Zeevi 2015, Bastani et al. 2021, Singhvi and Singhvi 2022). The core idea of these algorithms is to randomize over available options to learn preferences, and then efficiently manage the learning and earning tradeoff between randomizing and maximizing engagement, or other outcomes (Lattimore and Szepesvári 2020). However, the nudging context provides unique challenges that makes the direct application of existing algorithms particularly challenging.

First and foremost is the non-stationary nature of nudge effectiveness on users. For instance many empirical studies have shown that nudges lead to initial increase in user performance or engagement but this effect tends to diminish over time as users become accustomed to it (Rogers and Frey 2015, Asensio and Delmas 2019, Lichand and Christen 2020). This effect, often referred to as the *novelty effect* Chen et al. (2020a), presents a significant challenge since, if not correctly accounted for, it can lead to misleading interpretations of nudge effectiveness and complicate the optimization process. Second, is the complexity of modeling and designing nudge interventions. In particular, digital nudges encompass a diverse array of intervention types, each targeting distinct behavioral mechanisms to influence user engagement.[2] While these broad categories define the underlying psychological principles at work, the digital nature of these nudges allows for virtually

[2] These nudge categories include, but are not limited to, gamification elements that leverage intrinsic motivation and competition, self-comparison nudges that promote personal growth and goal-setting, and peer-comparison nudges that harness social influence and normative behavior.

infinite variations in their presentation and implementation by dynamically adjusting visual elements, timing, wording, and interactive features. Consequently, a single nudge-type, targeting a single behavioral mechanism can manifest in countless forms, each of which can be designed by incurring additional design costs (Dhar et al. 2017, Harrison and Patel 2020, Tor 2023). Platforms have to simultaneously minimize nudge *design costs* while maximizing user engagement. Hence, to accurately model the online nudge preference learning and optimization problem one has to carefully model the non-stationary effects of nudging while incorporating nudge design costs and different nudge types.

1.1. Contributions

We highlight our main contributions below:

- *Modeling of the learning with repeated nudging problem:* We model the problem of repeated nudging of users on online platforms where user preferences for different nudges are unknown, and repeated usage of the same nudge reduces its effectiveness. Our framework also incorporates distinct characteristics of the nudging problem. Notably, we posit that nudges belong to different types (for e.g., reminder nudges, peer comparison nudges, self-comparison nudges etc.) and new nudges of any type can be generated at some fixed cost. We also model the effect of nudge re-generation on rewards. Hence, in our learning problem the platform has to not only decide what nudge to send, but also decide when to generate a new nudge. To the best of our knowledge, our paper is the first paper that studies these tradeoffs in the learning and nudging context.
- *Characterizing a near-optimal policy under full information:* To develop insights into the structure of the optimal policy and achievable performance in our problem, we begin by analyzing its full-information variant. In Theorem 1, we establish that a simple heuristic, which sends only a single nudge type that *regenerates* after a fixed number of periods, remains within a constant additive factor of the optimal reward at all times (and is therefore long-run-average optimal as well). This result is significant, as the optimal policy—and consequently, the optimal reward—may be computationally intractable, yet the proposed heuristic serves as a simple and effective proxy. Interestingly, we show that this heuristic directly corresponds to the celebrated Whittle Index policy, which is widely studied in the context of restless bandits (Theorem 2).
- *Rate-optimal online learning with finite-time guarantees:* Following the intuition from the full-information setting, we focus on efficiently learning the Whittle indices of each nudge type in the finite-horizon learning setting. Analytically, we establish that our learning problem is statistically as hard as the K-armed bandit problem by proving that a regret of $\Omega(\sqrt{KT})$ is

unavoidable under non-anticipating policies (Theorem 3). We then show that our proposed UCB algorithms, built on the widely used principle of *Optimism under Uncertainty*, achieve rate-optimal regret (up to logarithmic terms) of order $\tilde{\mathcal{O}}\left(\sqrt{|\mathcal{L}|T}\right)$, where $|\mathcal{L}|$ denotes the number of unknown parameters (Theorems 4 and 5). Our algorithms update estimates of the unknown parameters in an episodic manner, with each episode converging, in the limit, to a regeneration cycle of the underlying Whittle Index Policy. Our regret analysis leverages a careful and novel adaptation of the well-known *Elliptical Potential Lemma* from the linear bandits literature (see, e.g., Carpentier et al. (2020), Hamidi and Bayati (2023)) to derive tight bounds in our setting, which may be of independent interest. Finally, to capture salient aspects of learning in the presence of auxiliary information, we also extend our analysis to linear models, where each arm type is associated with a static feature vector. We establish similar performance bounds in this setting (Theorems 6 and 7), demonstrating the robustness of our approach.

- *Extensive numerical experiments to demonstrate the empirical effectiveness of the algorithm:* Finally, we perform extensive numerical experiments on synthetic data, and in parameter settings inspired from a nudging intervention in the Early Childhood Education setting. We benchmark the proposed algorithm against benchmark learning policies such as Upper Confidence Bound and Thompson Sampling (TS) (Lattimore and Szepesvári (2020)), and find that our algorithm considerably outperforms benchmark algorithms and is also robust to parameter misspecification. Overall, the numerical results further enforce the practical applicability of the proposed algorithm.

2. Literature Review

Our work relates to two primary streams of literature. On one hand, we model and analyze our recommendation problem within the well-established bandit learning framework and develop a novel algorithm with broad applicability across various domains. On the other hand, from an applied perspective, our focus is on data-driven nudging in platforms designed for social good.

Bandits: The bandit framework aims to efficiently balance the tradeoff between learning and earning in settings where user preferences are unknown and the platform's objective is to learn user preferences while maximizing rewards. For a comprehensive overview of the vast literature in this area, we refer the interested readers to Bubeck and Cesa-Bianchi (2012), Slivkins et al. (2019), Lattimore and Szepesvári (2020). Since the seminal work of Robbins (1952), numerous extensions and modifications have been analyzed, including the case of non-stationary rewards and arms of $\mathcal{K}$-types, extensions most closely related to ours that we discuss in more detail next.

Non-stationary bandits generalize the classical multi-armed bandit problem to settings where reward distributions evolve over time. Unlike stationary bandits, which assume fixed expected

rewards, non-stationary models accommodate dynamic environments, making them particularly relevant to applications such as dynamic pricing, adaptive resource allocation, and online recommendation systems. To address challenges arising from non-stationarity, various strategies have been proposed, including sliding window algorithms (Garivier and Moulines 2008), discounted-UCB approaches (Kocsis and Szepesvári 2006), and meta-learning frameworks (Chen et al. 2020b), each adapting to evolving reward distributions by prioritizing recent observations or detecting distribution shifts. A fundamental challenge in non-stationary bandits lies in balancing the exploitation of currently high-reward arms with exploration to detect change points (Besbes et al. 2014). In comparison, our work introduces a novel formulation of non-stationarity in rewards in the context of nudges. Departing from prior work on non-stationary bandits, we consider a $\mathcal{K}$-typed bandit setting (Pandey et al. 2007, Baransi et al. 2014, Kalvit and Zeevi 2020) with an infinite number of arms, each belonging to one of K types and exhibiting non-stationary decaying rewards. Crucially, new arms can be generated at an additional cost, requiring the platform to jointly optimize arm selection and arm regeneration. This structure not only captures reward decay but also enables strategic intervention, allowing the platform to regulate the timing of arm regeneration and manage shifts in reward dynamics.

From a modeling perspective, our work is closely related to studies on mortal bandits and rotting bandits. Mortal bandit problems consider settings in which available arms can "die" or become inactive over time (Chakrabarti et al. 2008, Tracà et al. 2020). These models are motivated by applications in health care and marketing, where certain treatments or advertisements may become unavailable as a study progresses. Research in this area typically focuses on developing strategies that rapidly detect and adapt to the loss of arms, ensuring that decision-making remains efficient and effective. Similarly, rotting bandit problems involve a potentially infinite number of arms, each of which may degrade—or "rot"—over time as a function of its usage Levine et al. (2017), Seznec et al. (2019). Our proposed model draws on ideas from both frameworks yet is substantially different from either. In our model, much like in the mortal bandit framework, arms have a fixed lifetime after which they become irrelevant. Moreover, we allow an arm's rewards to "rot" as a function of its age rather than the number of times it is pulled. Although this aspect differs from the traditional rotting bandit model, we show that our model subsumes the standard model of reward decay found in rotting bandits. This insight follows from the observation that the rotting rewards model can be cast as a resting bandit problem, whereas our proposed problem is instead a restless bandit problem where rewards of different arms can change, regardless of whether that particular arm is pulled or not (Whittle 1980, 1988). In fact, we leverage the problem structure, and an independent novel sample path based approach to show the optimality of a cyclic policy, which also coincides with the Whittle Index Policy. Nevertheless, unlike most prior studies on

Whittle Index, which primarily focus on the infinite-horizon average cost setting, we demonstrate the near-optimality of our proposed policy in the finite horizon setting as well.

Digital Nudging: From an application perspective, our work addresses the problem of digital nudging—particularly on platforms aimed at social good. Subtle digital interventions, such as personalized notifications, progress trackers, and social comparisons, have been shown to effectively guide user behavior while preserving choice (Sunstein et al. 2022). Empirical studies in healthcare (Kelders et al. 2012) and education (Goyal et al. 2024) further illustrate that well-designed nudges can significantly boost engagement and improve outcomes. However, these prior works generally assume a static setting and do not explicitly account for the diminishing returns (the so-called *novelty effect*) that occur when the same nudge is repeatedly employed (Rogers and Frey 2015, Asensio and Delmas 2019, Chen et al. 2020a, Lichand and Christen 2020). In contrast, our work models the learning problem of repeated nudging in which user preferences for different nudge types are unknown and the effectiveness of a nudge decays with repetition. The study most closely related to our work is Chen et al. (2020a), in which the authors employ a bandit learning model to optimize digital nudge delivery on the Duolingo mobile app. That study also utilizes an exponential decay model to account for the reduction in engagement resulting from repeated nudges. Nevertheless, our work substantially differs from Chen et al. (2020a): while they assume that the decay model parameters can be estimated separately, our approach jointly learns both the nudge effects and their decay rates, and crucially incorporates the regeneration cost for designing a new nudge. This comprehensive framework enables us not only to determine when a nudge's effectiveness wanes, but also to identify the optimal point at which it is beneficial for the platform to invest in generating a new nudge rather than reusing an old one. Consequently, both our analytical approach and the main insights diverge significantly from those of Chen et al. (2020a).

Outline of the Paper

§3 introduces our model and formally defines the problem. In §4, we characterize a near-optimal policy for the full-information baseline setting, assuming perfect prior knowledge of all problem primitives. §5 addresses the challenges posed by parameter uncertainty: §5.1 examines the fundamental statistical complexity of the problem, while §5.2.1 and §5.2.2 introduce two variants of our rate-optimal learning algorithm, each tailored to different configurations of unknown parameters. Numerical experiments comparing the performance of our algorithms against several baseline policies are presented in §6. For brevity, discussions on model extensions incorporating auxiliary covariate information are deferred to §H and §I in the supplementary material. Finally, all technical developments, including proofs and ancillary results, are relegated to the appendices.

3. Model and Problem Formulation

Notation. Before formally introducing our model, we review the key mathematical notation used throughout this paper. We adopt the standard Landau convention, where $f(t) = \mathcal{O}(g(t))$ (equivalently, $g(t) = \Omega(f(t))$) if there exists a positive constant C independent of t such that $f(t) \leqslant Cg(t)$. Moreover, if C is independent of all problem primitives, we refer to it as an absolute constant. The notation $\tilde{\mathcal{O}}(\cdot)$ is used to suppress polylogarithmic factors in $\mathcal{O}(\cdot)$. For any $N \in \mathbb{N}$, we denote $[N] = \{1, \ldots, N\}$. For any $w \in \mathbb{R}$, $\lfloor w \rfloor$ and $\lceil w \rceil$ represent the greatest integer less than or equal to w and the smallest integer greater than or equal to w, respectively. For any $a, b \in \mathbb{N}$, the modulo operator is defined as $a \% b := a - \lfloor a/b \rfloor b$. The indicator random variable for an event E is denoted by $\mathbb{1}\{E\}$. The scalar inner product of vectors $x, y \in \mathbb{R}^d$ is denoted by $\langle x, y \rangle$, while the outer product of x with itself is denoted by $x \otimes x \in \mathbb{R}^{d \times d}$. The *elliptical norm* of $x \in \mathbb{R}^d$ with respect to a symmetric positive definite matrix $A \in \mathbb{R}^{d \times d}$ is defined as $\|x\|_A := \sqrt{\langle x, Ax \rangle}$.

In this section, we present a stylized model for optimizing nudge interventions. We consider a universe of countably many arms indexed by $\mathbb{N}$, each belonging to one of K possible arm types in $[K]$. The type of an arm is an observable attribute specified by the mapping $\texttt{type} : \mathbb{N} \to [K]$. For each type $\kappa \in [K]$, there is an associated mean reward $\mu_\kappa \geq 0$, which is unknown a priori. For example, in the nudging context, arms correspond to different nudges, and the K arm types represent different nudge categories (e.g., reminder nudges might constitute one type, while peer-comparison nudges form another). Similarly, mean reward μ_κ could then denote the average engagement due to sending a *new* nudge of type κ to the user. The user interacts with the platform sequentially over T rounds. In each round $t \in [T]$, the decision maker may either pull a *new* arm of type κ (i.e., one not yet explored) by incurring a known cost $c_\kappa \in [0, \mu_\kappa]$, or pull an *old* arm (i.e., one previously pulled) at no additional cost. Pulling a new arm corresponds to generating a new nudge, whereas pulling an old arm reuses an existing nudge, yielding a diminished reward.

More precisely, let $\mathcal{A}_{t-1} \subset \mathbb{N}$ denote the set of arms pulled up to round $t-1$ (with $\mathcal{A}_0 := \varnothing$). Then, the net expected reward from playing arm i in round t, conditional on $\texttt{type}(i) = \kappa$, is given by:

$$\begin{aligned} &\texttt{Rew}(i,t) \,|\, \texttt{type}(i) = \kappa \\ &= \begin{cases} \mu_\kappa - c_\kappa & \text{if } i \notin \mathcal{A}_{t-1}, \\ \mu_\kappa \, g_\kappa\big(\texttt{age}_t(i)\big) & \text{if } i \in \mathcal{A}_{t-1} \text{ and } \texttt{age}_t(i) \leqslant M-1, \\ -\infty & \text{otherwise,} \end{cases} \end{aligned}$$

where $g_\kappa : \{0\} \cup [M-1] \to \mathbb{R}_+$ is a non-increasing mapping (possibly unknown) with $g_\kappa(0) = 1$ for each $\kappa \in [K]$, and $\texttt{age}_t(i) := t - t_0(i)$ with $t_0(i) \in [t]$ denoting the first round in which arm i

was pulled.[3] Simply put, $g_\kappa\big(\texttt{age}_t(i)\big)$ captures the decay in the efficacy of arm i since its initial pull, and M represents the maximum life of the arm after which it is discarded. Figure 1 plots a representative reward from different arm-types, as a function of the age of the arm pulled, with and without the cost of regeneration. The underlying decay function follows a Logit specification (more details on the parameter specification are provided in §6.1). Notice that different arm-types have different decay patterns, and have heterogeneous regeneration costs. This heterogeneity allows for the model to capture fairly general reward and cost structures.

In the sequential setting, we let π_t denote the arm pulled in round $t \in [T]$. Then the net realized reward r_t is generated according to the following model:

$$r_t = \xi_t + \begin{cases} \mu_{\texttt{type}(\pi_t)} - c_{\texttt{type}(\pi_t)} & \text{if } \pi_t \notin \mathcal{A}_{t-1} \\ \mu_{\texttt{type}(\pi_t)} g_{\texttt{type}(\pi_t)}(\texttt{age}_t(\pi_t)) & \text{if } \pi_t \in \mathcal{A}_{t-1} \text{ and } \texttt{age}_t(\pi_t) \leqslant M-1 \\ -\infty & \text{otherwise} \end{cases}, \tag{1}$$

where ξ_t is a mean-zero noise term satisfying Assumption 1 which is standard in the study of Bandit learning algorithms (Lattimore and Szepesvári 2020).

Assumption 1 (Independent σ-Sub-Gaussian Noise) *In each round $t \in [T]$, $\xi_t \sim \texttt{subG}(\sigma^2)$, generated independently of $\mathcal{H}_{t-1} \cup \{\pi_t\}$, where $\mathcal{H}_{t-1}$ denotes the past history defined as $\mathcal{H}_{t-1} := \texttt{SigmaAlgebra}\{(\pi_s, r_s) : s \in [t-1]\}$.*[4]

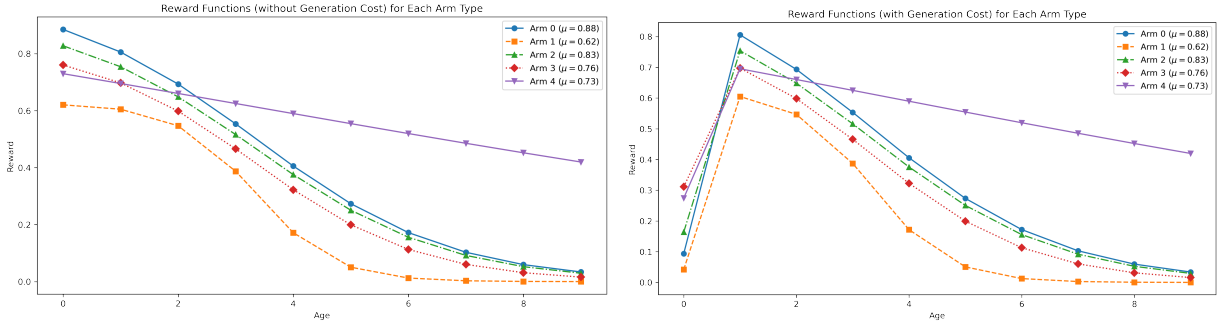

Figure 1 **Representative expected reward outcomes for different arms and corresponding reward decay rates. On the left, we plot the expected reward without accounting for cost of regeneration and on the right we plot the expected reward after accounting for the cost of regeneration.**

The decision maker's objective is to maximize cumulative reward over T time periods. To formally define this objective, we first define the set of feasible non-anticipating policies in our setting. The

[3] A natural alternative is to model $\texttt{age}_t(i)$ as the number of times arm i has been pulled up to round t. Such a model would be similar to the rotting bandits framework where reward decays as a function of the arm usage (Levine et al. 2017, Seznec et al. 2019). This would result in a "simpler" resting model in which our algorithms and results would still apply; we omit the discussion for brevity.

[4] See Definition 2 for details on sub-Gaussian random variables.

sequence $\pi := (\pi_t : t \in [T])$ is called a feasible policy if each π_t is a non-anticipating mapping from the history to actions, $\pi_t : \mathcal{H}_{t-1} \mapsto \mathbb{N}$ (possibly randomized). We will refer to $\pi_t(\mathcal{H}_{t-1})$ simply as π_t. The exhaustive set of feasible policies is denoted by Π. The decision maker's problem of maximizing cumulative expected rewards over T rounds is given by

$$\sup_{\pi \in \Pi} \mathbb{E}_\pi \left[\sum_{t \in [T]} r_t \right] , \tag{2}$$

where the expectation is w.r.t. all possible sources of randomness in the problem (rewards as well as policy). We denote the optimal value of (2) when all problem primitives are known a priori by $\texttt{OPT}_T$. The *regret*[5] incurred by policy π over the T rounds is then given by

$$\mathfrak{R}(T) := \texttt{OPT}_T - \sum_{(\kappa,m) \in [K] \times [M]} (\mu_\kappa g_\kappa(m-1) - \mathbb{1}\{m = 1\} c_\kappa) N_{\kappa,m}(T) , \tag{3}$$

where $N_{\kappa,m}(T)$ denotes the number of rounds in $[T]$ in which an arm of type κ and age $m-1$ is pulled by π. Using the Tower property of expectations, it can be shown that $\mathbb{E}_\pi[\mathfrak{R}(T)] = \mathbb{E}_\pi\left[\texttt{OPT}_T - \sum_{t \in [T]} r_t\right]$, and therefore, (2) is equivalent to the problem of minimizing the widely studied metric of expected regret, given by

$$\inf_{\pi \in \Pi} \mathbb{E}_\pi[\mathfrak{R}(T)] . \tag{4}$$

4. A Near-Optimal Static Index Policy Under Complete Information

This section aims to outline key properties of $\texttt{OPT}_T$, the optimal value of (2) when all primitives, including $(\mu_\kappa, g_\kappa(\cdot) : \kappa \in [K])$, are known a priori. To this end, we first impose a natural restriction on the policy space Π to prune it to a tractable class Π_0. We further show that this restriction incurs no loss of optimality and preserves the achievability of $\texttt{OPT}_T$ within the pruned policy class. Finally, we characterize a static index-based, age-triggered policy from Π_0 that solves (2) near-optimally and shed light on its connections to prior work on related models in the literature.

Definition 1 (No-Recall Policy) *A policy $\pi \in \Pi$ is referred to as a "no-recall policy" iff the following implications hold for all rounds $t, s \in [T]$ satisfying $t + 2 \leqslant s \leqslant T$:*

1. $\pi_t \neq \pi_{t+1} \implies \pi_t \neq \pi_s$.
2. $\texttt{type}(\pi_t) \neq \texttt{type}(\pi_{t+1}) \implies \texttt{type}(\pi_t) \neq \texttt{type}(\pi_s)$.

We denote the exhaustive collection of such policies by Π_0.

[5] Note that this definition is occasionally used for *pseudo-regret* in the bandit literature; see, e.g., Slivkins et al. (2019).

Informally, a no-recall policy is a non-anticipating policy that batches arms and types together, where each arm is pulled only during consecutive rounds (referred to as an *epoch* henceforth) and then discarded. In what follows, we show that, without loss of optimality, one can focus exclusively on no-recall policies to solve the regret minimization problem.

Lemma 1 (Restricting to No-Recall Policies is Sufficient) *For every $\pi \in \Pi$, there exists some $\pi' \in \Pi_0$ such that $\mathbb{E}_{\pi'}\left[\sum_{t\in[T]} r_t\right] \geqslant \mathbb{E}_{\pi}\left[\sum_{t\in[T]} r_t\right]$.*

Details of the proof are deferred to §A.1 but follow from the monotonicity of the decay functions. Specifically, in the presence of multiple arms of the same type at different ages, one can, without loss of optimality, focus solely on the "youngest" arm of that type and discard all others. Consequently, all pulls of an arm can be batched together. Moreover, batching arm types together is naturally without loss of generality. We next leverage Lemma 1 to characterize a near-optimal solution to (2) from the policy class Π_0.

Theorem 1 (A Near-Optimal Age-Triggered Policy) *For each $(\kappa, m) \in [K] \times [M]$, define the index*

$$\bar{\mu}_{\kappa,m} := \frac{\mu_\kappa \sum_{t=0}^{m-1} g_\kappa(t) - c_\kappa}{m} \ , \tag{5}$$

and let $(\kappa^*, m^*) \in \arg\max_{(\kappa,m)\in[K]\times[M]} \bar{\mu}_{\kappa,m}$. Then, the no-recall policy π^* that keeps pulling an arm of type κ^* while its age is less than m^* rounds, switches to a new arm of the same type thereafter and repeats the process, satisfies for any $T \in \mathbb{N}$:

$$\texttt{OPT}_T - c_{\kappa^*} \mathbb{1}\left\{T\%m^* \neq 0\right\} \leqslant \mathbb{E}_{\pi^*}\left[\sum_{t\in[T]} r_t\right] \leqslant \texttt{OPT}_T \leqslant \bar{\mu}_{\kappa^*,m^*} T \ . \tag{6}$$

Further, when $T\%m^* = 0$ or $c_{\kappa^*} = 0$, all inequalities in (6) are tight, i.e., $\texttt{OPT}_T = \mathbb{E}_{\pi^*}\left[\sum_{t\in[T]} r_t\right] = \bar{\mu}_{\kappa^*,m^*} T$.

Details of the analysis are deferred to §A.2, but in essence, the proof leverages Lemma 1 to optimize over the pruned policy class Π_0, identifying the structure of π^* from the optimal objective value. Theorem 1 shows that, without any significant loss of optimality, it suffices to focus exclusively on arms of type κ^*. Furthermore, the "smallest age" serves as a sufficient state descriptor—tracking only the age of the most recently pulled type κ^* arm suffices for near-optimal performance[6]. Figure 2 demonstrates the structure of the optimal policy for the same setting as

[6] A version of Theorem 1 can be shown to hold in the more general setting where the maximum lifetime M is unbounded. However, we do not pursue this result at this level of generality, as practical scenarios typically satisfy $M < \infty$.

that of Figure 1. In this case, the optimal policy generates a new arm with arm-type 4 after pulling the arm for 7 consecutive periods and then discards it. Finally, a stationary, Markov, and deterministic policy that prescribes pulling a fixed arm of type κ^* repeatedly while its age is less than m^* rounds, then discarding it and replacing it with a new arm of the same type, remains within c_{κ^*} of $\texttt{OPT}_T$ in value at all times.

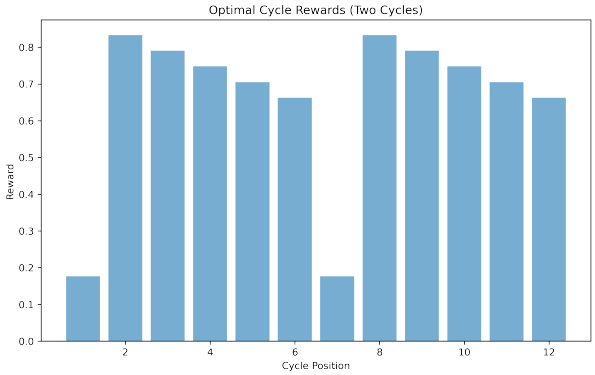

Figure 2 **Optimal policy structure of the age-triggered policy. The reward and decay parameters are selected as in Figure 1. Notice that arm-type 0 yields the highest reward when a new arm of this type is generated. Nevertheless, since rewards decay over time, the optimal policy in fact chooses arm-type 4 and regenerates in after 6 consecutive periods .**

Before concluding this section, we show that the no-recall policy π^* described in Theorem 1 is, in fact, closely related to a well-known heuristic for a special class of Markov Decision Processes known as *Restless Bandits* (Whittle 1988).

Theorem 2 (Connection to the Whittle Index Policy for Restless Bandits) *Under complete knowledge of all primitives,* (2) *is a special case of a restless K-armed bandit problem that satisfies the Whittle-indexability condition. Moreover, the no-recall policy π^* described in Theorem 1 matches the Whittle Index Policy, with static Whittle indices given by* $\left(\max_{m\in[M]} \bar{\mu}_{\kappa,m} : \kappa \in [K]\right)$.

The details of this connection are fleshed out in §A.3. Several remarks are in order: (i) It is important to note that Theorem 1 does not follow from Theorem 2 or any prior result in the restless bandit literature. Instead, our proof of optimality of the no-recall policy π^* (which, incidentally, happens to match the Whittle Index Policy) relies on an independent, novel sample-path-based analysis that exploits our problem structure. (ii) Prior literature on the restless bandit problem, beginning with the seminal works of Whittle (1980, 1988), primarily focuses on the infinite-horizon average-cost formulation under indexability and additional technical conditions that are generally hard to verify (only a few papers study finite-horizon formulations; see Hu and Frazier (2017),

Zayas-Caban et al. (2019), Brown and Smith (2020), Zhang and Frazier (2021)). (iii) Moreover, while the indexability condition ensures that the Whittle Index Policy is well-defined, it alone may not be sufficient to guarantee the policy's optimality (see, e.g., Weber and Weiss (1990)). (iv) Most importantly, the notion of optimality adopted in the aforementioned references and in the broader restless bandit literature differs from ours: there, optimality is defined w.r.t. K in an asymptotic regime where the *proportion* of arms pulled in each round is held fixed (budget/activation constraint). In contrast, we show that the no-recall policy π^* is near-optimal for any $K \in \mathbb{N}$ in our setting, where exactly one arm is pulled in each round.

To conclude, we reiterate that the above connection to the Whittle Index Policy and the restless bandit model is established solely for completeness of theoretical exposition and does not affect our analyses or main results.

5. Regret Minimization Under Parameter Uncertainty

We now focus on the incomplete information setting where the mean reward vector $(\mu_\kappa : \kappa \in [K])$, and potentially the decay kernels $(g_\kappa(\cdot) : \kappa \in [K])$ as well, are unknown a priori, but can be estimated from noisy observations. In §5.1, we establish a fundamental hardness result for the regret minimization problem in (4), demonstrating that in the absence of prior knowledge of these parameters, no algorithm can achieve regret smaller than $\Omega\Big(\sqrt{T}\Big)$ uniformly across all parameter configurations. Subsequently, in §5.2, we propose an adaptive learning algorithm that achieves a near-optimal $\mathcal{O}\big(\sqrt{T \log T}\big)$ regret.

5.1. Fundamental Statistical Complexity

In what follows, an "instance" of the problem refers to a specific parameter configuration, $\{(\mu_\kappa, \sigma, c_\kappa, g_\kappa(m)) : (\kappa, m) \in [K] \times [M-1]\}$, along with the distribution of $(\xi_t : t \in [T])$. For clarity, we will slightly overload notation to emphasize the dependence of achievable regret on the given problem instance.

Theorem 3 (Lower Bound on Achievable Regret) *Fix $T \geqslant K$. Then, for every policy $\pi \in \Pi$, there exists a problem instance ν_π such that $\mathbb{E}_\pi[\mathfrak{R}(T; \nu_\pi)] = \Omega\Big(\sigma\sqrt{KT}\Big)$, where the $\Omega(\cdot)$ only hides absolute constants.*

The proof relies on a reduction to a K-armed bandit problem, leveraging Theorem 1 (see §C for details). Informally, this implies that our problem is statistically at least as hard as a K-armed bandit problem. In §5.2, we will see that this bound is, in fact, tight with respect to its leading order in T and, depending on the dimension of the estimand (the mean reward vector, potentially with the decay kernels as well), can also capture the correct order of dependence on K (up to logarithmic factors in other primitives).

5.2. An Optimism-based Approach for Learning the Whittle Index Policy

In this section, we propose an approach based on the celebrated *optimism under uncertainty* principle from the multi-armed bandits literature (see, e.g., Slivkins et al. (2019)) to learn the static Whittle indices $\left(\max_{m\in[M]} \bar{\mu}_{\kappa,m} : \kappa \in [K]\right)$ for each arm type in an online, data-driven manner.

Recall that the mean reward vector $(\mu_i : i \in [K])$, and potentially also the decay kernels $(g_i(\cdot) : i \in [K])$, are initially unknown. We first address the case of known decay kernels in §5.2.1, before investigating unknown decay kernels in §5.2.2.

5.2.1 Learning the Mean Reward Vector

A key challenge in learning the unknown mean reward vector lies in determining how to appropriately use samples of varying "ages" for estimation. To address this, we propose in (7) a stratified estimator, $\hat{\mu}_\kappa^{(t)}$, for estimating the true mean reward μ_κ associated with arm-type κ, using observations collected from arms of type κ up to round t.

In what follows, for $(\kappa, m, t) \in [K] \times [M] \times [T]$, let $N_{\kappa,j}(t)$ denote the total number of samples collected by the end of round t from type κ arms of age $j-1$. We denote the empirical mean of these samples by $\hat{\nu}_{\kappa,j}^{(N_{\kappa,j}(t))}$, with $\hat{\nu}_{\kappa,j}^{(0)} := 0$. Note that $\hat{\nu}_{\kappa,j}^{\left(N_{\kappa,j}(t)\right)}$ serves as an empirical estimate of $\mu_\kappa g_\kappa(j-1)$. Our stratified estimator, defined for $(\kappa, t) \in [K] \times [T]$, is given by

$$\hat{\mu}_\kappa^{(t)} := \frac{\sum_{j\in[M]} \left(\frac{\rho_{\kappa,j} \hat{\nu}_{\kappa,j}^{\left(N_{\kappa,j}(t)\right)} \left(1+N_{\kappa,j}(t)\right)}{g_\kappa(j-1)} \right)}{1 + \sum_{j\in[M]} \rho_{\kappa,j} N_{\kappa,j}(t)} \,, \tag{7}$$

where $(\rho_{\kappa,j} : (\kappa, j) \in [K] \times [M])$ are hyper-parameters satisfying $\rho_{\kappa,j} \geqslant 0 \ \forall \ (\kappa, j) \in [K] \times [M]$ and $\sum_{j\in[M]} \rho_{\kappa,j} = 1 \ \forall \ \kappa \in [K]$.

Remark 1 (Possible Stratifications) *For any $\kappa \in [K]$, the vector $(\rho_{\kappa,j} : j \in [M])$ determines how samples from type κ arms of different ages contribute to the estimation of μ_κ in (7). For example, setting $\rho_{\kappa,1} = 1$ would completely discard samples of age at least 1, while setting $\rho_{\kappa,j} = 1/M \ \forall \ j \in [M]$ would assign equal weight to samples of all ages.*

Our algorithm also relies on computing the following key estimation metrics, defined for each $(\kappa, m, t) \in [K] \times [M] \times [T]$:

$$\hat{\mu}_{\kappa,m}^{(t)} := \frac{\hat{\mu}_\kappa^{(t)} \sum_{j\in[m]} g_\kappa(j-1) - c_\kappa}{m} \,, \tag{8}$$

$$\texttt{rad}_{\kappa,m,t} := 2\sigma \left(\frac{\sum_{j\in[M]} \sqrt{\alpha \left(\frac{\rho_{\kappa,j}}{g_\kappa(j-1)} \right)^2 (1+N_{\kappa,j}(t)) \log\left(\frac{1}{\delta}\right)}}{1 + \sum_{j\in[M]} \rho_{\kappa,j} N_{\kappa,j}(t)} \right) \left(\frac{\sum_{j\in[m]} g_\kappa(j-1)}{m} \right) \,, \tag{9}$$

where $\hat{\mu}_{\kappa}^{(t)}$ is as defined in (7), σ is the sub-Gaussian parameter from Assumption 1, and $\alpha \geqslant 0$ is a hyper-parameter that controls the *confidence radius* $\texttt{rad}_{\kappa,m,t}$ of the estimate $\hat{\mu}_{\kappa,m}^{(t)}$, henceforth referred to as the *exploration coefficient*, and δ is the *failure probability*. An appropriately tuned exploration coefficient guarantees that the estimate $\hat{\mu}_{\kappa,m}^{(t)}$ lies within $\pm\texttt{rad}_{\kappa,m,t}$ of the underlying estimand $\bar{\mu}_{\kappa,m}$ (see (5)), *with high probability*.

We are now ready to present our algorithm for the case of known decay kernels. In what follows, a "new" arm refers to one that has never been pulled before.

Algorithm 1 Decay-Rate Aware Optimistic Nudging (DRAWON)

- **Input:** Types $[K]$, Max life M, sub-Gaussian param σ, Exploration coefficient α, Horizon T, Decay coefficients $\{g_\kappa(j-1) : (\kappa, j) \in [K] \times [M]\}$, Weights $\{\rho_{\kappa,j} : (\kappa, j) \in [K] \times [M]\}$, Costs $\{c_\kappa : \kappa \in [K]\}$, Failure probability δ.
- **Initialize:** Round $t = 1$.
- **While** $t \leqslant T$

  Fix $(\kappa_t, m_t) \in \arg\max_{(\kappa,m)\in[K]\times[M]} \left\{\hat{\mu}_{\kappa,m}^{(t-1)} + \texttt{rad}_{\kappa,m,t-1}\right\}$. (`UCB rule`)

  Select a new arm of type κ_t; pull it for m_t consecutive rounds.

  $t \leftarrow t + m_t$.

Dynamics. At the beginning of each episode, Algorithm 1 commits to pulling a new arm of a fixed type for a pre-specified number of rounds, after which the arm is discarded. Based on the observed feedback, the parameter estimates of the unknown mean reward vector are updated. The type and commitment time in each episode are chosen based on the UCB parameter estimates. The underlying premise is that the algorithm tracks the best $\bar{\mu}_{\kappa,m}$, eventually converging to the Whittle Index Policy described in Theorem 1. The result below provides bounds on the regret incurred during this process.

Theorem 4 (High Probability Regret Bounds for Algorithm 1) *Fix exploration coefficient α and failure probability δ such that $1 - 2KMT\delta^\alpha > 0$. Then, the regret of Algorithm 1 satisfies the following w.p. at least $1 - 2KMT\delta^\alpha$:*

$$\mathfrak{R}(t) \leqslant \max_{\kappa\in[K]} c_\kappa + 4\sigma\sqrt{2\alpha K \mathcal{C}_t \log\left(1 + \frac{t}{K}\right)\log\left(\frac{1}{\delta}\right)} \quad \forall\ t \in [T]\ , \tag{10}$$

where

$$\mathcal{C}_t := \sum_{s\in\mathcal{S}_{1,t}} \left\{ \left(\sum_{j\in[M]} \left(\frac{\rho_{\kappa_s,j}/\rho_{\kappa_s,1}}{g_{\kappa_s}^2(j-1)} \right) \right) \left(\sum_{j\in[m_s]} g_{\kappa_s}(j-1) \right)^2 \right\}\ , \tag{11}$$

with $\mathcal{S}_{1,t} \subseteq [t]$ denoting the subset of rounds until t in which the UCB rule applies.

In essence, the result follows from a decomposition of regret over the episodes of the algorithm. A tight upper bound on the sum of the corresponding UCB terms is then derived via a novel and careful application of the Elliptical Potential Lemma (Carpentier et al. 2020), adapted to our setting; we note that this result is typically used in the analysis of bandits with covariates and is uncommon in the study of "simple" (covariate-free) bandits. Nonetheless, its necessity in analyzing our model arises due to the complex interdependencies introduced by the decay kernels, for which standard stochastic bandit proof techniques appear to be inadequate, yielding bounds that are weaker by a factor of $\sqrt{M}$. Full details are deferred to §D. In what follows, we derive a bound from (10) for the special case where $\rho_{\kappa,1} = 1$ for all $\kappa \in [K]$.

Corollary 1 (Simple Mean Estimation Using Age-0 Samples) *In the setting of Theorem 4, if $\rho_{\kappa,1} = 1$ for all $\kappa \in [K]$, then the following holds w.p. at least $1 - 2KMT\delta^{\alpha}$:*

$$\mathfrak{R}(t) = \max_{\kappa \in [K]} c_\kappa + \tilde{\mathcal{O}}\left(\sqrt{K(t+M)}\ \min\left(\sqrt{M}\ ,\ \max_{\kappa \in [K]}\left(\sum_{j \in [M]} g_\kappa(j-1)\right)\right)\right) \quad \forall\ t \in [T]\ . \tag{12}$$

To see this, note that when $\rho_{\kappa,1} = 1$ for all $\kappa \in [K]$, we obtain $\mathcal{C}_t = \sum_{s \in \mathcal{S}_{1,t}} \left(\sum_{j \in [m_s]} g_{\kappa_s}(j-1)\right)^2 \leqslant (t+M) \min\left(M,\ \max_{\kappa \in [K]} \left(\sum_{j \in [M]} g_\kappa(j-1)\right)^2\right)$, leading to the bound in (12).

Effect of mean estimator selection: Corollary 1 suggests that relying solely on age-0 samples for estimation is generally sufficient to achieve good regret performance when decays are "fast." Specifically, if $\sum_{j \in [M]} g_\kappa(j-1)$ is bounded above by an absolute constant C for each $\kappa \in [K]$, then the upper bound in (12) matches the $\Omega\left(\sqrt{Kt}\right)$ lower bound from Theorem 3, up to logarithmic factors, for $t \geqslant M$. This suggests that stratification offers no advantage in such cases. On the other hand, when decays are "slow," the upper bounds do not indicate which estimator is preferable. However, we hypothesize that stratified estimation often results in improved numerical performance in this regime.

5.2.2 Learning Decay Kernels Simultaneously

While the previous section assumed that the decay rate functions g were known, in this section we analyze the case when neither the mean reward vector $(\mu_\kappa : \kappa \in [K])$ nor the decay coefficients $\{g_\kappa(j-1) : (\kappa, j) \in [K] \times [M]\}$ are known a priori. Recall that $\hat{\nu}_{\kappa,j}^{\left(N_{\kappa,j}(t)\right)}$ is an estimator for $\mu_\kappa g_\kappa(j-1)$, computed as the empirical mean of the $N_{\kappa,j}(t)$ samples collected from type κ arms of age $j-1$ by the end of round t. Our algorithm estimates the $\mu_\kappa g_\kappa(j-1)$ values directly. To this end, we redefine estimation metrics for $(\kappa, m, t) \in [K] \times [M] \times [T]$ as follows:

$$\hat{\mu}_{\kappa,m}^{(t)} := \frac{1}{m}\left(\sum_{j \in [m]} \hat{\nu}_{\kappa,j}^{\left(N_{\kappa,j}(t)\right)} - c_\kappa\right)\ . \tag{13}$$

$$\texttt{rad}_{\kappa,m,t} := \frac{2\sigma}{m} \sum_{j\in[m]} \sqrt{\frac{\alpha \log\left(\frac{1}{\delta}\right)}{1+N_{\kappa,j}(t)}} \ . \tag{14}$$

These are defined similarly to their counterparts in §5.2.1, except that they no longer rely on prior knowledge of the decay kernels. We now present our algorithm for the case of unknown decay kernels. In what follows, like before, a "new" arm refers to one that has never been pulled before.

Algorithm 2 Decay-Rate Agnostic Optimistic Nudging (DRAGON)

- **Input:** Types $[K]$, Max life M, sub-Gaussian param σ, Exploration coefficient α, Horizon T, Failure probability δ, Costs $\{c_\kappa : \kappa \in [K]\}$.
- **Initialize:** Round $t = 1$.
- **While** $t \leqslant T$

  Fix $(\kappa_t, m_t) \in \arg\max_{(\kappa,m)\in[K]\times[M]} \left\{\hat{\mu}^{(t-1)}_{\kappa,m} + \texttt{rad}_{\kappa,m,t-1}\right\}$. (`UCB rule`)

  Select a new arm of type κ_t; pull it for m_t consecutive rounds.

  $t \leftarrow t + m_t$.

Dynamics. The algorithm operates similarly to Algorithm 1, except that it uses the redefined estimator $\hat{\mu}^{(t-1)}_{\kappa,m}$ and confidence radius $\texttt{rad}_{\kappa,m,t-1}$ (as per (13) and (14) respectively), in the UCB rule. The result below provides an upper bound on the regret incurred by Algorithm 2.

Theorem 5 (High Probability Regret Bound for Algorithm 2) *Fix exploration coefficient α and failure probability δ such that $1 - 2KMT\delta^\alpha > 0$. Then, the regret of Algorithm 2 satisfies the following w.p. at least $1 - 2KMT\delta^\alpha$:*

$$\mathfrak{R}(t) \leqslant \max_{\kappa\in[K]} c_\kappa + 4\sigma\sqrt{\alpha \log\left(\frac{1}{\delta}\right)}\sqrt{2KM(t+M)\log\left(1+\frac{t+M}{K}\right)} \quad \forall\ t \in [T] \ .$$

Details of the analysis are deferred to §E, but follow a similar approach as in the proof of Theorem 4. It is noteworthy that this $\tilde{\mathcal{O}}\left(\sqrt{KMt}\right)$ bound is relatively benign compared to the complexity terms in general-purpose reinforcement learning approaches for finite-state, finite-horizon MDPs. For instance, a direct implementation of the `UCRL2` algorithm (Auer et al. 2008) yields a regret bound of order $\tilde{\mathcal{O}}\left(D|\mathcal{S}|\sqrt{|\mathcal{A}|t}\right)$, where the nominal size of the state space is $|\mathcal{S}| = \mathcal{O}(M^K)$, the action space is of size $|\mathcal{A}| = \mathcal{O}(K)$, and D denotes the MDP diameter, which is bounded above by MK. The exponential improvement achieved in our setting arises from an explicit incorporation of problem structure into the learning algorithm. Specifically, the structure of the near-optimal policy π^* characterized in Theorem 1 enables a state-space reduction from M^K to KM, corresponding to the actual dimensionality of the unknown parameter space.

Calibrating the maximum life parameter: Thus far, we have treated the maximum life parameter M as an independent model primitive. However, in practice, M could potentially be an endogenous parameter, as the viable life of any arm is intrinsically linked to its decay characteristics. In the following, we present a method to calibrate M using benign knowledge of other model primitives. To this end, we impose a mild assumption on the mean rewards and costs associated with each type.

Assumption 2 (Cost-to-Reward Ratio) *$c_\kappa \leqslant \beta\mu_\kappa$ holds for each $\kappa \in [K]$, where $\beta < 1$ is a known constant.*

The assumption posits that the decision maker is a priori aware of a meaningful upper bound on the cost-to-reward ratio for each arm type. Based on this information and prior knowledge of the decay kernels, we now present a formal result for calibrating the max life parameter M.

Proposition 1 (Sufficient Max Life of an Arm) *Suppose that Assumption 2 holds, and that $f_\kappa(m)$ is an upper bound on $g_\kappa(m)$ for each $(\kappa, m) \in [K] \times \mathbb{N}$. If M is an upper bound on $\max_{\kappa\in[K]} \mathcal{M}_\kappa$, where $\mathcal{M}_\kappa := \inf\{m \in \mathbb{N} : f_\kappa(m) \leqslant 1-\beta\}$, then $\inf(\arg\sup_{m\in\mathbb{N}} \bar{\mu}_{\kappa,m}) \leqslant M \ \forall\ \kappa \in [K]$.*

We defer the proof to §F. A direct implication of Proposition 1 is that, under the proposed calibration of M, the value $\sup_{m\in\mathbb{N}} \bar{\mu}_{\kappa,m}$ is attained at some $m^*_\kappa \in [M]$ for each type $\kappa \in [K]$. Consequently, it suffices to discard any arm once it reaches M rounds of age. In what follows, we provide examples of M that satisfy Proposition 1, derived from standard choices for decay kernels.

Example 1 (Light-Tailed Decays) *If $f_\kappa(m) = a_\kappa \exp(-b_\kappa m)$ for some $a_\kappa, b_\kappa > 0$, then it suffices to set $M = \max_{\kappa\in[K]} \mathcal{M}_\kappa$, where $\mathcal{M}_\kappa = \left\lceil \frac{1}{b_\kappa} \log\left(\frac{a_\kappa}{1-\beta}\right)\right\rceil$.*

Example 2 (Heavy-Tailed Decays) *If $f_\kappa(m) = a_\kappa m^{-b_k}$ for some $a_\kappa, b_\kappa > 0$, then it suffices to set $M = \max_{\kappa\in[K]} \mathcal{M}_\kappa$, where $\mathcal{M}_\kappa = \left\lceil \left(\frac{a_\kappa}{1-\beta}\right)^{\frac{1}{b_\kappa}} \right\rceil$.*

Finally, in more practical settings when a meaningful upper-bound on the cost-to-reward ratio is not readily available, we recommend over-estimating the max-life parameter based on numerical results in §6 that show the robustness of the algorithm to the max-life parameter.

6. Numerical Experiments

We now compare the empirical performance of the proposed algorithms with state-of-the-art benchmark algorithms using both synthetic (§6.1) as well as real-world data (§6.2).

Benchmark Algorithms: We implement benchmark algorithms based on Upper Confidence Bounds and Thompson Sampling, the two most widely used algorithms in academic research for online learning problems.

1. Decay-Aware and Decay-Agnostic Benchmark UCB Algorithms: We consider two-versions of the classical UCB algorithm for multi-armed bandit problems as proposed by Auer (2002), (one for the decay aware and the other for the decay agnostic case). The algorithms rely on a direct cost-to-reward comparison for arm selection and regeneration decisions instead of the Whittle index rule that the proposed DRAWON (Algorithm 1) and DRAGON (Algorithm 2) algorithms rely on. We present the exact pseudo-code for the benchmark algorithms in §J.
2. Decay-Aware and Decay-Agnostic Benchmark TS Algorithms: We also implement two different versions of the Thompson Sampling algorithm (Lattimore and Szepesvári (2020)) that uses a sampling based approach for exploration and is widely used due to its ease of implementation. As before, the algorithms use a direct cost-to-reward comparison to decide which arms to pull and when to regenerate them. We present the pseudo-code for the algorithm and additional details on the sampling strategy in §J.

6.1. Synthetic Data Experiments

Data generation: Recall that in our setting K denotes distinct arm types where each arm-type $i \in [K]$ is characterized by a base mean reward μ_i, a decay function $g_i(\cdot)$, and a regeneration cost c_i. We generate these parameters as follows: (i) Each $\mu_i \sim \mathrm{U}[0.5, 1]$ where $\mathrm{U}[a, b]$ denotes a uniform random variable between a and b; (ii) Given μ_i, we let the cost parameter $c_i \sim \mathrm{U}[\frac{\mu_i}{2}, \mu_i]$; (iii) For each arm type i, we specify a maximum life M, so that $g_i(\texttt{age})$ is only meaningful for $\texttt{age} \in \{0, \ldots, M-1\}$. We experiment with different decay functions that we discuss in-more detail below; (iv) Finally, we let $T = 10,000$ and whenever an arm of type κ is pulled in period t, rewards are generated according to the reward model (1) specified in §3 with $\varepsilon_t \sim \mathcal{N}(0, 1)$.

Decay functions and optimal policy structure: Recall that the decay function $g_i(.)$ plays a crucial role in our setting since it determines the structure of the optimal policy. We implement two natural parametric forms of the decay functions discussed in more detail below.

Exponential decay function: The exponential decay function captures an initial drop in rewards, followed by a more gradual tapering off from subsequent usage. The reward from pulling an arm of type κ and age j is given by:

$$g_\kappa(j) = \mu_\kappa \exp\left(-\lambda_\kappa j\right), \tag{15}$$

where μ_κ is the baseline effect of pulling arm κ of age 0 sampled as discussed before and λ_κ[7] denotes the decay rate associated with using an arm of type κ. In Figure 3, we plot representative

[7] The exact functional form of $\lambda_k = c_2 + c_3 U[0, 2]$ where c_2 and c_3 are independent constants. This functional form allows for substantial variability in decay rates across different arm-types.

$\mu_\kappa g_\kappa(.)$, the realized reward (after accounting for the cost of regeneration, and finally the rewards for the clairvoyant Whittles Index based optimal policy.

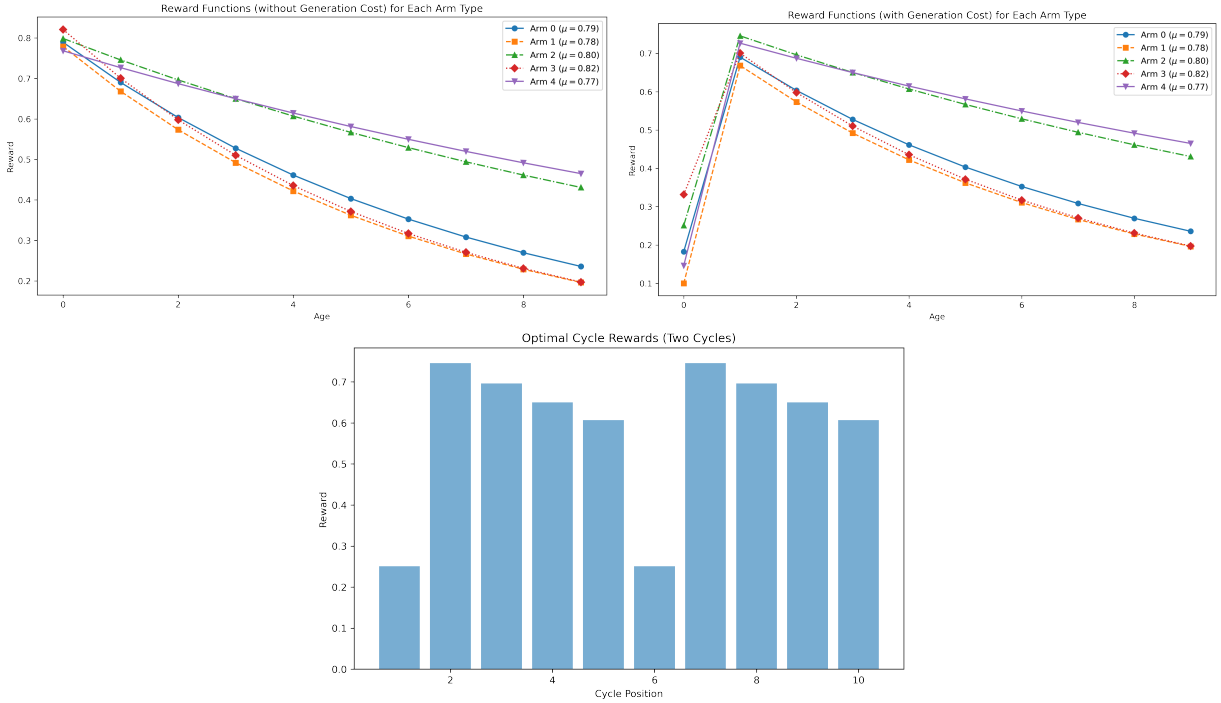

Figure 3 **Reward structure and optimal policy structure. Note that here the cost of regeneration is constant across arms for ease of exposition. Notice that while Arm-type 4 yields the highest reward when a new arm of this type is generated. Nevertheless, since rewards decay over time, the optimal policy in fact chooses arm-type 2 with a cycle-length of 5 before regenerating the arm again.**

Logit decay function: The Logit functional form allows for rewards to remain relatively stable when the arm's age is small. Nevertheless, as the arm-age goes up, the reward experiences a sharp drop and stabilizes again. In this case, the reward from pulling an arm of type k and age j is defined as:

$$g_k(j) = \frac{\mu_k}{1 + \exp(\beta_k j)}, \tag{16}$$

μ_k and β_k are arm-specific parameters selected to ensure that substantial variability in the reward decay rates. As in the exponential decay case, In Figure 1, we plot representative $\mu_\kappa g_\kappa(.)$, the realized reward (after accounting for the cost of regeneration, and finally the rewards for the clairvoyant Whittles Index based optimal policy in the Logit case (Figure 2).

Regret comparison: In Figure 4, we plot the regret of the proposed DRAWON (Algorithm 1) algorithm and compare it to the benchmark algorithms in the decay-aware case. Note that in this case all algorithms are provided with complete information on the decay function g but

nevertheless have to learn the initial reward μ_κ. On the left, we plot regret results from the case when the underlying decay model is exponential, and on the right we plot regret results from the case when the underlying decay model is Logit. We find that the proposed algorithm considerably outperforms the benchmark algorithms. This difference can be attributed to sub-optimal action selection under the benchmark algorithms since they do not track the right cycle-averaged reward metric to optimize. For instance, notice that the representative optimal policies in the exponential and Logit case (Figure 3 and Figure 2) neither pulls the arm-type that maximizes the initial reward (μ_κ), nor the one that maximizes the cost adjusted initial reward ($\mu_\kappa - c_\kappa$). Since the benchmark algorithms use a direct cost-to-reward comparison, they end up taking suboptimal actions in this setting.

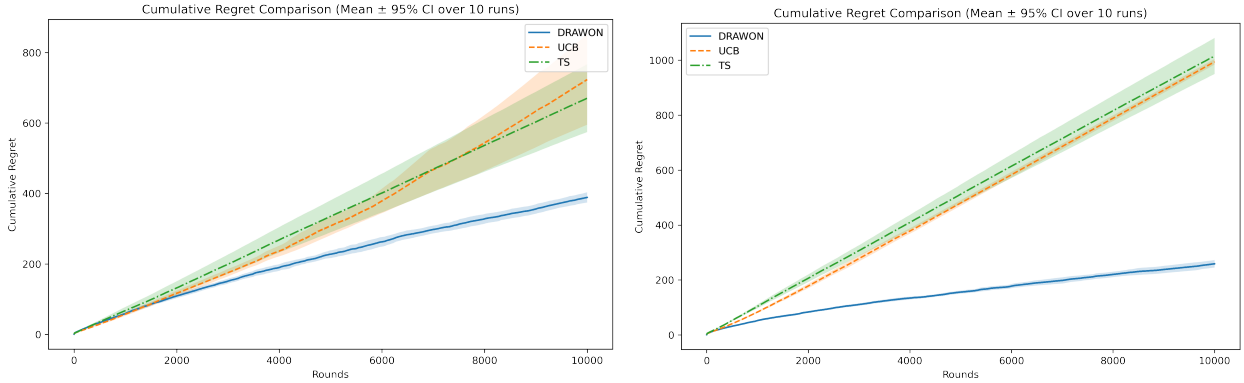

Figure 4 **Cumulative regret comparison between the proposed DRAWON Algorithm and the UCB and TS Algorithms when all algorithms have access to the decay rate function g. On the left, we plot cumulative regret in the case when g follows an exponential decay. On the right, we plot cumulative regret in the case when g follows the logistic or s-shaped decay. In both cases the proposed DRAWON algorithm considerably outperforms both the benchmark algorithms.**

Similarly, in Figure 5, we plot the regret of different benchmark algorithms in the decay-agnostic case. That is, no algorithm is provided with the information of the decay function g. As before, the proposed DRAGON Algorithm (Algorithm 2) substantially outperforms other benchmark algorithms in this setting as well. In fact, the performance gap between benchmark algorithms and DRAGON Algorithm considerably increases in compare to the decay-aware case. This is because in this case the decay parameters are also unknown and hence all algorithms need to learn additional parameters from the adaptively collected data.

Robustness to max-life parameter M***:*** Recall that our proposed model and algorithms rely on the max-life parameter M which signifies the maximum time the platform is willing to use a nudge since its generating. Since it practice, the exact value of this parameter might be hard to

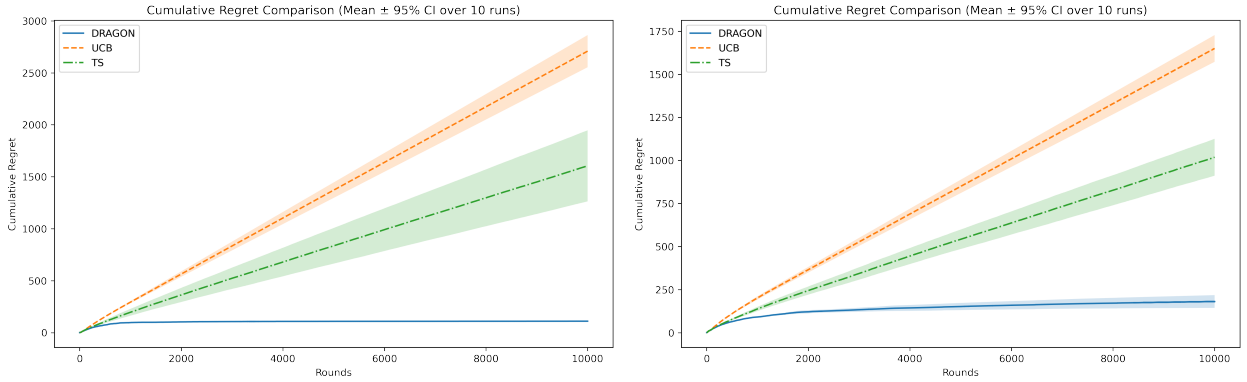

Figure 5 **Cumulative regret comparison between the proposed DRAGON Algorithm and the UCB and TS Algorithms when all algorithms have access to the decay rate function g. On the left, we plot cumulative regret in the case when g follows an exponential decay. On the right, we plot cumulative regret in the case when g follows the logistic or s-shaped decay. In both cases the proposed DRAGON algorithm considerably outperforms both the benchmark algorithms.**

estimate a-priori, in this section we test the robustness of the proposed algorithm to this parameter. In Figure 6, we plot the cumulative regret of the proposed DRAGON algorithm under different values of the max-life parameter. As expected, the cumulative regret goes up as M increases but nevertheless, the increase is not substantial (in comparison to the performance of the benchmark algorithms) and hence we conclude that the algorithm is robust to parameter miss-specification in M.

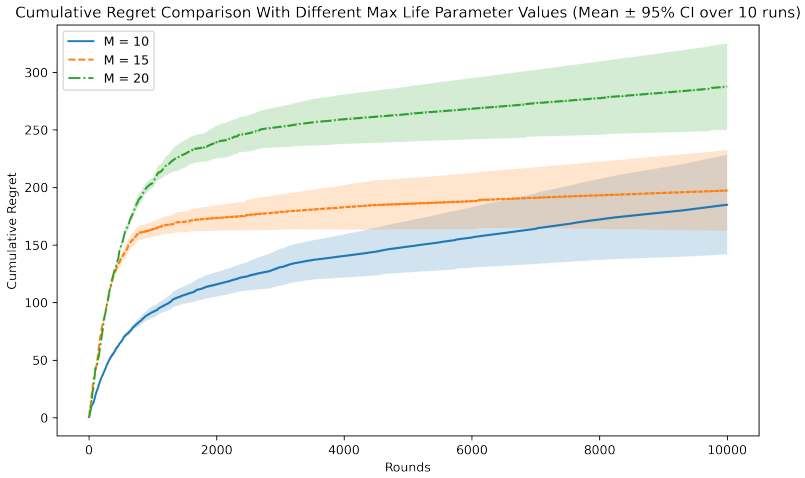

Figure 6 **Cumulative regret comparison of the proposed DRAGON Algorithm (Algorithm 2) for different values of the max life parameter M. Notice that the algorithm is robust to the selection of M and converges to the optimal policy in a comparable time across all the values of M.**

6.2. Case Study: EdTech Platforms

We now present simulation results from a model calibrated based on recent empirical results on the effect of repeated nudging of users on Rocket Learning, one of India's largest nonprofit EdTech startups dedicated to enhancing early childhood education (Goyal et al. 2024). We start this section by briefly discussing the setting; how we calibrated the ground truth model and simplifying assumptions; and finally present simulation results.

Rocket Learning and Digital Nudging: Rocket Learning (RL) works alongside state and national government bodies to connect parents from public schools and day-care centers (referred to in Hindi as "Anganwadi") through digital communities, primarily using WhatsApp groups. Every physical classroom corresponds to an online group that includes both parents and the class teacher. Through these virtual communities, Rocket Learning provides age-appropriate, play-based materials to support learning beyond the classroom (Goyal et al. 2024).

One of the central challenges that RL as well as other digital platforms face is that of consistent user engagement. Hence, RL uses various interventions, including digital nudging, to keep users engaged. The current paper leverages results from a large-scale Randomized Control Trial involving nearly 150,000 parents to test the effects of one-on-one user nudging on RL's platform (Goyal et al. (2024)). The experiment tested the effectiveness of two types of behavioral nudges: peer-comparison and self-comparison nudges. WhatsApp groups were randomly assigned to one of the two treatment arms (or the control arm) and parents in these groups were nudged with the corresponding nudge over a four week-period, once every week. Both the nudges were found to be highly effective, leading up to a 8-13.5% average improvement in daily engagement on the WhatsApp groups. Most importantly, the authors also estimate long-term temporal effects of behavioral nudges. Figure 7 (Figure 5 of Goyal et al. 2024) demonstrates the estimated temporal effect of both the nudges. Note that Week 0 denotes the start of the experiment and the experiment ran for 4 weeks. Hence, weeks -3, -2 and -1 denote pre-treatment weeks and weeks 4, 5 6 and 7 denote post-treatment weeks. Interestingly, both nudges show differential trends. While peer-comparison nudges show very limited decrease in effectiveness due to repeated usage, self-comparison nudges show more pronounced decrease in effectiveness over time. In what follows, we discuss how we leverage these estimates to generate a ground-truth model for our experimental evaluations.

Generating a ground-truth model: Recall from §3 that our modeling framework considers each nudge-type κ to be associated with a mean-reward μ_κ, a decay function $g_\kappa(.)$ and a cost of arm-regeneration c_κ. The temporal effects estimated from the RCT can be readily used to estimate all the three parameters. In particular, estimates from Week 0 and Week 1 can be used to estimate both μ_κ and c_κ. Recall that c_κ denotes the cost of creating new nudges. While designing new nudges of any type can be relatively cheap (in comparison to the relative gains from engaging additional

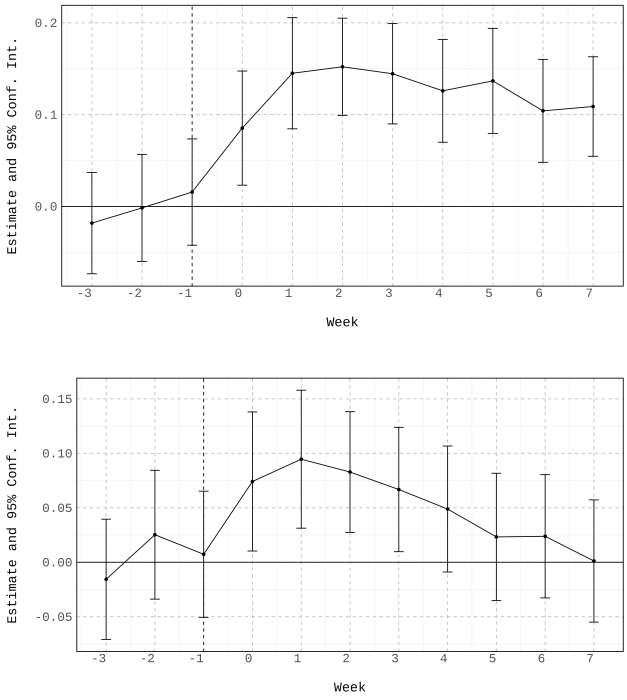

Figure 7 **Long-Term Temporal Effects Estimated in Goyal et al. (2024). Using a large-scale RCT, the authors estimate long-term temporal effects of digital nudging on user engagement. The RCT experimented with peer and self comparison nudges where users were sent identical nudges over the course of 4 weeks (week 0,1,2 and 3). Weeks 4 and onward constitutes post-treatment data. The plot at the top estimates treatment effect of the peer-comparison nudge while the plot on the bottom constitutes treatment effect for the self-comparison nudge. We use these treatment effects estimates to caliberate our ground-truth model.**

users, especially in the EdTech context), there is an inherent *adoption* cost that platforms pay when introducing new interventions or nudges. This cost reflects the initial warm-up phase when users get accustomed to the new intervention. For example, notice that the estimated effect at the start of the experiment (Week 0) is in fact smaller than the effect in Week 1 for both the nudge-types. This difference in fact reflects the "cost" of nudging users with new nudges. Letting Ψ_k^t denote the point estimate of the effect of nudge k in period t, then we have that

$$\Psi_k^0 = \mu_\kappa - c_\kappa \;\&\; \Psi_k^1 = \mu_\kappa \implies c_\kappa = \Psi_k^1 - \Psi_k^0 \,, \forall k \in \{peer, self\}\,.$$

Hence, we have that $c_{peer} = 0.6$ and $\mu_{peer} = 0.15$. Similarly, $c_{self} = 0.015$ and $\mu_{self} = 0.09$. Notice that as expected, the cost of nudging is higher for the peer-comparison nudge which is the more complex intervention amongst the two interventions. Finally, since the same nudge was sent from

weeks 0 to 3, we can leverage these values to estimate g_κ. In fact, since we do not put parametric assumptions on $g(.)$, we simply let $g_\kappa(0) = 1$ and

$$g_\kappa(j) = \max\left\{1, \frac{\Psi_\kappa^j}{\Psi_\kappa^1}\right\},\ \forall j \in [1,2,3],\ \kappa \in \{self, peer\}\,. \tag{17}$$

Since the experiment ran for four weeks, to estimate $g_\kappa(j), \forall j > 3$, we consider two different heuristics. The first heuristic simply uses (17) from the post-treatment periods (weeks 4-7) to estimate $g_\kappa(j), \forall j \in [4,5,6,7]$. Letting the max-life parameter be $M = 8$ completes the ground-truth specification in this case. Table 1 presents the estimated $g_\kappa(.)$ from this approach.

Table 1 Non-Parametric Estimates of the Decay Function

j	0	1	2	3	4	5	6	7
g_{peer}	1.00	1.00	0.99	0.93	0.83	0.83	0.67	0.67
g_{self}	1.00	1.00	0.89	0.78	0.56	0.28	0.28	0.00

Alternatively, our second heuristic extrapolates data from the experiment using different curve fitting methods to estimate the decay parameter for later periods. Figure 8 plots the result from these extrapolations. Note that we estimate three different parametric decay functions (linear, exponential and S-shaped Logit functions). Since the data from the experimental periods is limited, we also re-estimate the decay functions using both in-treatment as well as post-treatment data (Figure 8, on the right). Finally, we let the max life parameter $M = 15$ since in the RL's context, this would mean using the same nudge for 4 months at most. Nevertheless, as discussed before, the performance of the proposed algorithm is robust to the selection of this parameter. Finally, as a note of caution, we note that our proposed model also assumes that new nudges of the same type leads to reward regeneration. In practice, platforms run continuous experimentation with many interventions and hence this modeling assumption might be easy to verify. Nevertheless, since RL's nudging experiment did not use multiple types of the same nudge, we cannot verify this assumption from the data, and consider it a limitation of the experimental results presented in the following section.

Regret Results: Each ground truth model can be used to generate a corresponding clairvoyant policy, and in-turn compare different benchmark algorithms in terms of their cumulative regret. Since we estimate one non-parametric ground truth model and six parametric ground-truth models, this leads to a total of seven different potential settings to benchmark different algorithms. Since in practice, the decay functions are hardly known in advance, we consider the decay agnostic setting. In Figure 9, we plot cumulative regret from three different ground-truth settings: (i) the non-parametric estimation; (ii) the parametric Logit based estimation using only in-treatment

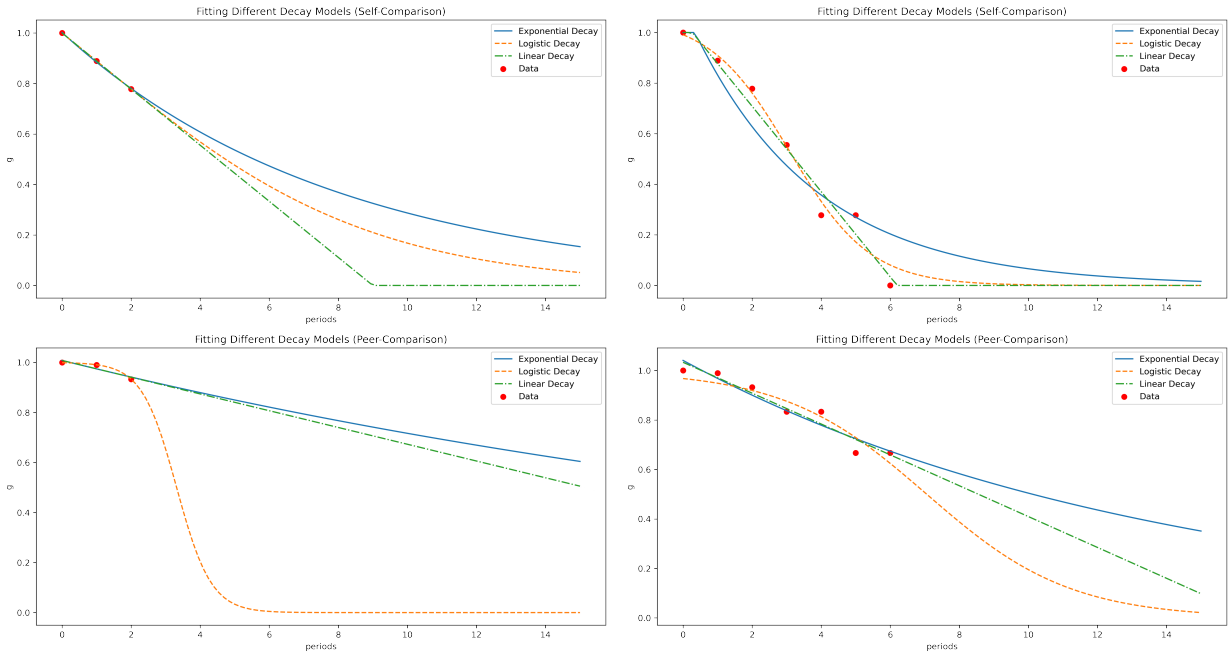

Figure 8 **Parametric Estimates of the decay function g_{self} and g_{peer}. We leverage the long term temporal effects estimated from the RCT to estimate the decay functions. The top (bottom) row plots results from estimating g_{self} (g_{peer}). Each plot contains outputs from fitting three different parametric functions: linear, exponential and S-shaped Logit functions. Plots on the left only use data from during-treatment weeks while plots on the right use data from all the seven weeks.**

data (4 weeks); and (iii) the parametric Logit based estimation using both in-treatment and post-treatment data (7 weeks). In each case, the proposed algorithm considerably outperforms the benchmark algorithms in terms of cumulative regret. In fact, the cumulative regret of DRAGON is one sixth of the regret of the benchmark algorithms. This improvement directly translates to increased engagement of users on the platform. We find similar insights in the other four settings but relegate these results to §J for the sake of brevity.

7. Concluding Remarks

In this work, we examine the challenge of repeatedly nudging users on online platforms, taking into account initially unknown nudge effectiveness and the diminished impact of reusing the same nudge over time. We model the problem using an online learning framework with multiple nudge types, bandit feedback, and non-stationary rewards, while also incorporating the costs of creating new nudges to maintain their effectiveness. Under full information (when all parameters are known), we show that a cyclic policy, which periodically regenerates a single nudge type, maximizes the long-run average reward and, somewhat surprisingly, achieves constant regret in finite horizons. Building on these insights, we reduce the broader online learning problem to determining both the optimal nudge type and its cycle length, and we propose a UCB-based algorithm

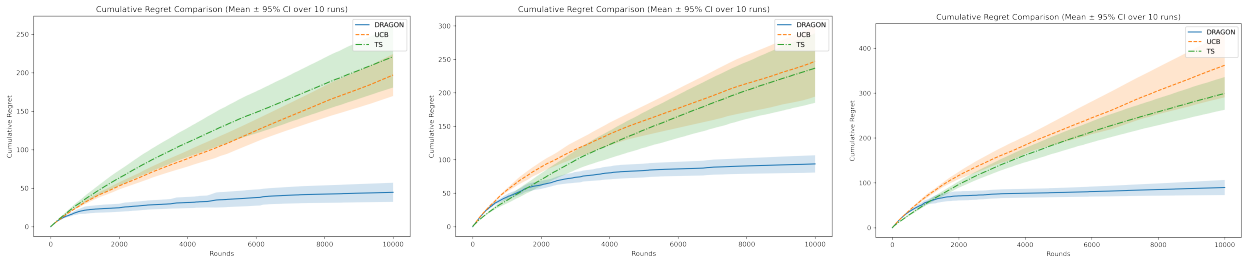

Figure 9 **Cumulative Regret comparison between the proposed DRAGON and the benchmark algorithms on the calibrated model. On the left, we consider results from considering the ground truth model using non-parametric estimation. In the middle, we plot regret results from the case when the ground truth model considers the logit parametric model but only considers the data from the 4 weeks of treatment. Finally, the plot of the right also considers the logit function to estimate the ground truth model but uses data from seven weeks to estimate the ground truth.**

that achieves sublinear near-optimal regret guarantees. Numerical experiments using both synthetic data and real-world data from an EdTech setting highlight considerable improvements over benchmark methods, illustrating the practical relevance and efficacy of the proposed framework.

There are several potential directions for future work. We highlight two that are particularly relevant and closely related to our current model. First, an important extension would be to study scenarios where rewards diminish across episodes, i.e., regeneration is "lossy." This would correspond not only to a nudge becoming stale over time but also to its type losing effectiveness, making the model more reflective of real-world intervention dynamics. Second, a more realistic setting would involve the platform operating under a budget constraint for regenerating interventions, a characteristic feature of many social impact operations. However, both these extensions introduce a fundamentally more complex setting that requires a first-principles analysis to characterize optimal performance. In such cases, existing algorithms and results would no longer directly apply. We leave these investigations to future work.

References

Abbasi-Yadkori, Yasin, Dávid Pál, Csaba Szepesvári. 2011. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* **24**.

Agrawal, Keshav, Susan Athey, Ayush Kanodia, Emil Palikot. 2023. Digital interventions and habit formation in educational technology. *arXiv preprint arXiv:2310.10850* .

Asensio, Omar I, Magali A Delmas. 2019. The dynamics of consumer behavior: Novelty and framing effects. *Journal of Economic Behavior and Organization* .

Auer, Peter. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* **3**(Nov) 397–422.

Auer, Peter, Thomas Jaksch, Ronald Ortner. 2008. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems* **21**.

Baransi, Amos, Odalric-Ambrym Maillard, Shie Mannor. 2014. Subgroup identification in multi-armed bandits. *International Conference on Machine Learning (ICML)*. 409–417.

Bastani, Hamsa, Kimon Drakopoulos, Vishal Gupta, Ioannis Vlachogiannis, Christos Hadjichristodoulou, Pagona Lagiou, Gkikas Magiorkinis, Dimitrios Paraskevis, Sotirios Tsiodras. 2021. Efficient and targeted covid-19 border testing via reinforcement learning. *Nature* **599**(7883) 108–113.

Bertsekas, Dimitri. 2012. *Dynamic programming and optimal control: Volume I*, vol. 4. Athena scientific.

Besbes, Omar, Yonatan Gur, Assaf Zeevi. 2014. Stochastic multi-armed-bandit problem with non-stationary rewards. *Operations Research* **62**(5) 1234–1250.

Besbes, Omar, Assaf Zeevi. 2015. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science* **61**(4) 723–739.

Brown, David B, James E Smith. 2020. Index policies and performance bounds for dynamic selection problems. *Management Science* **66**(7) 3029–3050.

Bubeck, Sébastien, Nicolò Cesa-Bianchi. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *CoRR* **abs/1204.5721**. URL `http://arxiv.org/abs/1204.5721`.

Carpentier, Alexandra, Claire Vernade, Yasin Abbasi-Yadkori. 2020. The elliptical potential lemma revisited. *arXiv preprint arXiv:2010.10182* .

Chakrabarti, Deepayan, Ravi Kumar, Filip Radlinski, Eli Upfal. 2008. Mortal multi-armed bandits. *Advances in neural information processing systems* **21**.

Chang, Tom Y, Mireille Jacobson, Manisha Shah, Matthew Kopetsky, Rajiv Pramanik, Samir B Shah. 2023. Reminders, but not monetary incentives, increase covid-19 booster uptake. *Proceedings of the National Academy of Sciences* **120**(31) e2302725120.

Chen, Jane, et al. 2020a. A sleeping, recovering bandit algorithm for optimizing recurring notifications. *Proceedings of the 2020 International Conference on Machine Learning for Human Behavior*. 123–134.

Chen, Xinyi, Wei Xu, Zheng Zhou. 2020b. Adaptive meta-learning for non-stationary bandits. *Advances in Neural Information Processing Systems*.

Chu, Wei, Lihong Li, Lev Reyzin, Robert Schapire. 2011. Contextual bandits with linear payoff functions. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 208–214.

Dani, Varsha, Thomas P Hayes, Sham M Kakade. 2008. Stochastic linear optimization under bandit feedback. *COLT*, vol. 2. 3.

Dhar, Julia, Allison Bailey, Stéphanie Mingardon, Jennifer Tankersley. 2017. The persuasive power of the digital nudge. `https://www.bcg.com/publications/2017/`

people-organization-operations-persuasive-power-digital-nudge. Accessed: 2025-02-02; Boston Consulting Group.

Garivier, Aurélien, Éric Moulines. 2008. On upper-confidence bound policies for switching bandit problems. *Algorithmic Learning Theory*. Springer, 174–188.

Goyal, Nishit, Alex Akira Okuno, Divya Singhvi, Somya Singhvi. 2024. Increasing parents' engagement on ed-tech platforms: Evidence from the field. *Available at SSRN 4918175* .

Hamidi, Nima, Mohsen Bayati. 2023. The elliptical potential lemma for general distributions with an application to linear thompson sampling. *Operations Research* **71**(4) 1434–1439.

Harrison, Joseph D, Mitesh S Patel. 2020. Designing nudges for success in health care. *AMA Journal of Ethics* **22**(9) 796–801.

Hu, Weici, Peter Frazier. 2017. An asymptotically optimal index policy for finite-horizon restless bandits. *arXiv preprint arXiv:1707.00205* .

Kalvit, Anand, Assaf Zeevi. 2020. From finite to countable-armed bandits. *Advances in Neural Information Processing Systems* **33** 8259–8269.

Kelders, Saskia M, Robin N Kok, Hans C Ossebaard, Julia EWC Van Gemert-Pijnen. 2012. Persuasive system design does matter: a systematic review of adherence to web-based interventions. *Journal of medical Internet research* **14**(6) e152.

Kocsis, Levente, Csaba Szepesvári. 2006. Discounted ucb. *Proceedings of the International Conference on Machine Learning*.

Lattimore, T., Cs. Szepesvári. 2020. *Bandit Algorithms*. Cambridge University Press. URL https://tor-lattimore.com/downloads/book/book.pdf.

Levine, Nir, Koby Crammer, Shie Mannor. 2017. Rotting bandits. *Advances in neural information processing systems* **30**.

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

Lichand, Guilherme, Julien Christen. 2020. Using nudges to prevent student dropouts in the pandemic. *arXiv preprint arXiv:2009.04767* .

McCullagh, Peter, John A. Nelder. 1989. *Generalized Linear Models*, *Monographs on Statistics and Applied Probability*, vol. 37. 2nd ed. Chapman and Hall/CRC. doi:10.1007/978-1-4899-3242-6.

Milkman, Katherine L, Sean F Ellis, Dena M Gromet, Youngwoo Jung, Alex S Luscher, Rayyan S Mobarak, Madeline K Paxson, Ramon A Silvera Zumaran, Robert Kuan, Ron Berman, et al. 2024. Megastudy shows that reminders boost vaccination but adding free rides does not. *Nature* **631**(8019) 179–188.

Nazaret, Achille, Guillermo Sapiro. 2023. A large-scale observational study of the causal effects of a behavioral health nudge. *Science Advances* **9**(38) eadi1752.

Pandey, Shilpa, Deepayan Chakrabarti, Deepak Agarwal. 2007. Multi-armed bandit problems with dependent arms. *International Conference on Machine Learning (ICML)*. 721–728.

Reich, Justin, José A Ruipérez-Valiente. 2019. The mooc pivot. *Science* **363**(6423) 130–131.

Robbins, Herbert. 1952. Some aspects of the sequential design of experiments .

Rock Health. 2020. Digital health consumer adoption report 2020. Tech. rep., Rock Health. URL `https://rockhealth.com/insights/digital-health-consumer-adoption-report-2020/`.

Rogers, Todd, Erin Frey. 2015. Changing behavior beyond the here and now. *The Wiley Blackwell handbook of judgment and decision making* **2** 723–748.

Rusmevichientong, Paat, John N Tsitsiklis. 2010. Linearly parameterized bandits. *Mathematics of Operations Research* **35**(2) 395–411.

Seznec, Julien, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, Michal Valko. 2019. Rotting bandits are no harder than stochastic ones. *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2564–2572.

Singhvi, Divya, Somya Singhvi. 2022. Online learning with sample selection bias. *Available at SSRN 4045779* .

Slivkins, Aleksandrs, et al. 2019. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning* **12**(1-2) 1–286.

Sunstein, Cass R, Cass Sunstein, Richard H Thaler. 2022. *Nudge*. Vuibert.

Tor, Avishalom. 2023. Digital nudges: Contours and challenges. *International Law and Economics Conference*. Springer, 3–18.

Tracà, Stefano, Cynthia Rudin, Weiyu Yan. 2020. Reducing exploration of dying arms in mortal bandits. *Uncertainty in Artificial Intelligence*. PMLR, 156–163.

Vershynin, Roman. 2018. *High-Dimensional Probability: An Introduction with Applications in Data Science*, *Cambridge Series in Statistical and Probabilistic Mathematics*, vol. 47. Cambridge University Press. doi:10.1017/9781108231596.

Weber, Richard R, Gideon Weiss. 1990. On an index policy for restless bandits. *Journal of applied probability* **27**(3) 637–648.

Whittle, Peter. 1980. Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society: Series B (Methodological)* **42**(2) 143–149.

Whittle, Peter. 1988. Restless bandits: Activity allocation in a changing world. *Journal of applied probability* **25**(A) 287–298.

Zayas-Caban, Gabriel, Stefanus Jasin, Guihua Wang. 2019. An asymptotically optimal heuristic for general nonstationary finite-horizon restless multi-armed, multi-action bandits. *Advances in Applied Probability* **51**(3) 745–772.

Zhang, Xiangyu, Peter I Frazier. 2021. Restless bandits with many arms: Beating the central limit theorem. *arXiv preprint arXiv:2107.11911* .

A. Analysis of the Complete Information Setting

A.1. Proof of Lemma 1

If $\pi \in \Pi_0$, the claim trivially holds for $\pi' = \pi$ (note that $\Pi_0 \subset \Pi$). Therefore, without loss of generality, suppose that $\pi \in \Pi \backslash \Pi_0$. Then, we know that at least one of the implications in Definition 1 is false for some $t, s \in [T]$ s.t. $t+2 \leqslant s \leqslant T$. There are two exhaustive and mutually exclusive possibilities:

1. Implication (1) is false for some $t, s \in [T]$ s.t. $t+2 \leqslant s \leqslant T$.
2. Implication (1) is always true, but implication (2) is false for some $t, s \in [T]$ s.t. $t+2 \leqslant s \leqslant T$.

We address the two cases separately below, presenting an algorithmic approach to prove the stated assertion.

Case I: Implication (1) is false for some t, s.

In this case, define

$$\tau_1 := \min \{t \in [T] : \pi_t \neq \pi_{t+1} \text{ and } \exists\ s \in \{t+2, \cdots, T\} \text{ s.t. } \pi_t = \pi_s\}.$$
$$\tau_2 := \min \{s \in \{t+2, \cdots, T\} : \pi_{\tau_1} = \pi_s\}.$$

Then, note that $\{\pi_1, \cdots, \pi_{\tau_1}\} \cap \{\pi_{\tau_1+1}, \cdots, \pi_{\tau_2-1}\} = \varnothing$. Consider now an alternative policy π' whose arm-pull sequence is given by

$$\pi'_t := \begin{cases} \pi_t & \text{for } 1 \leqslant t \leqslant \tau_1 \\ \pi_{\tau_1} & \text{for } t = \tau_1 + 1 \\ \pi_{t-1} & \text{for } \tau_1 + 2 \leqslant t \leqslant \tau_2 \\ \pi_t & \text{for } t \geqslant \tau_2 + 1. \end{cases}$$

It follows that the cumulative T-round payoff of π' is weakly greater than that of π, owing to the non-increasing property of the decay kernels. If $\pi' \in \Pi_0$, the claim is proved. If not, the same argument can be applied iteratively until a policy π'' satisfying implication (1) is constructed. If π'' violates implication (2), one can construct a new policy by simply rearranging the epochs to "stack" identical arm types together (this operation preserves implication (1) as well as the cumulative T-round payoff). The resulting policy, denoted by π''', will then satisfy both implications and have a T-round cumulative payoff that is necessarily weakly greater than that of π.

Case II: Implication (1) is always true, but implication (2) is false for some t, s.

By the same "stacking" argument as discussed above, we can construct a policy π' from π that satisfies implication (2) while preserving implication (1) as well as the T-round cumulative payoff.

Combining the two cases proves the assertion. □

A.2. Proof of Theorem 1

By Lemma 1, we know that there exists a no-recall policy that solves (2) optimally and achieves $\texttt{OPT}_T$. We will therefore, without loss of generality, focus only on policies in Π_0.

Fix $\pi \in \Pi_0$. Let M_κ denote the number of distinct type κ arms pulled by π, and let $\ell_{\kappa,m}$ denote the length of the m^{th} epoch of type κ (defined for $m \in [M_\kappa]$). Then, observe that

$$\mathbb{E}_\pi\left[\sum_{t\in[T]} r_t\right] = \sum_{\kappa\in[K]}\sum_{m\in[M_\kappa]}\left(\mu_\kappa \sum_{n=0}^{\ell_{\kappa,m}-1} g_\kappa(n) - c_\kappa\right)\mathbb{1}\left\{\ell_{\kappa,m}\geqslant 1\right\} \underset{\dagger}{\leqslant} \sum_{\kappa\in[K]}\sum_{m\in[M_\kappa]} \bar{\mu}_{\kappa^*,m^*}\ell_{\kappa,m} = \bar{\mu}_{\kappa^*,m^*}T \ , \tag{18}$$

where (†) follows from the definition of $\bar{\mu}_{\kappa^*,m^*}$ in (5). Since π is arbitrary, this implies

$$\texttt{OPT}_T = \sup_{\pi\in\Pi}\mathbb{E}_\pi\left[\sum_{t\in[T]} r_t\right] \underset{\texttt{a}}{=} \sup_{\pi\in\Pi_0}\mathbb{E}_\pi\left[\sum_{t\in[T]} r_t\right] \underset{\texttt{b}}{\leqslant} \bar{\mu}_{\kappa^*,m^*}T \ , \tag{19}$$

where (a) is due to Lemma 1, and (b) follows from (18). Now consider the no-recall policy π^* described in Theorem 1, and observe that

$$\begin{aligned}
\mathbb{E}_{\pi^*}\left[\sum_{t\in[T]} r_t\right] &= \bar{\mu}_{\kappa^*,m^*}m^*\left\lceil\frac{T}{m^*}\right\rceil - \left(\mu_{\kappa^*}\sum_{n=T\%m^*}^{m^*-1} g_{\kappa^*}(n)\right)\mathbb{1}\left\{T\%m^*\geqslant 1\right\} \\
&\underset{\dagger}{\geqslant} \bar{\mu}_{\kappa^*,m^*}m^*\left\lceil\frac{T}{m^*}\right\rceil - \left(\mu_{\kappa^*}\sum_{n=0}^{m^*-T\%m^*-1} g_{\kappa^*}(n)\right)\mathbb{1}\left\{T\%m^*\geqslant 1\right\} \\
&= \bar{\mu}_{\kappa^*,m^*}m^*\left\lceil\frac{T}{m^*}\right\rceil - c_{\kappa^*}\mathbb{1}\left\{T\%m^*\geqslant 1\right\} - \left(\mu_{\kappa^*}\sum_{n=0}^{m^*-T\%m^*-1} g_{\kappa^*}(n) - c_{\kappa^*}\right)\mathbb{1}\left\{T\%m^*\geqslant 1\right\} \\
&\underset{\ddagger}{\geqslant} \bar{\mu}_{\kappa^*,m^*}m^*\left\lceil\frac{T}{m^*}\right\rceil - c_{\kappa^*}\mathbb{1}\left\{T\%m^*\geqslant 1\right\} - \bar{\mu}_{\kappa^*,m^*}(m^*-T\%m^*)\mathbb{1}\left\{T\%m^*\geqslant 1\right\} \\
&= \bar{\mu}_{\kappa^*,m^*}\left(m^*\left\lceil\frac{T}{m^*}\right\rceil - (m^*-T\%m^*)\mathbb{1}\left\{T\%m^*\geqslant 1\right\}\right) - c_{\kappa^*}\mathbb{1}\left\{T\%m^*\geqslant 1\right\} \\
&= \bar{\mu}_{\kappa^*,m^*}T - c_{\kappa^*}\mathbb{1}\left\{T\%m^*\geqslant 1\right\} \\
&\underset{\star}{\geqslant} \texttt{OPT}_T - c_{\kappa^*}\mathbb{1}\left\{T\%m^*\geqslant 1\right\} \ ,
\end{aligned} \tag{20}$$

where (†) follows from the monotonicity of the decay kernels, (‡) follows from the definition of $\bar{\mu}_{\kappa^*,m^*}$ in (5), and ($\star$) follows from (19). Finally, note that π^* satisfies (18). Combining everything establishes the stated assertion. □

A.3. Proof of Theorem 2

The problem in (2) can be seen as a special case of a restless K-armed bandit problem. To show this, we reformulate the problem as a Markov Decision Process (MDP).

To this end, first note that, for each arm type, one can, without loss of optimality, focus exclusively on the "youngest arm." This significantly reduces the dimensionality of the problem and leads to the following MDP formulation:

There are K Markov bandit processes, each with a common state space of $\{0,\dots,M-1\}$, where the state $s\in[M-1]$ of any process denotes the age of the youngest arm of the corresponding type. State 0 indicates the absence of any viable arm to pull (which occurs when the maximum life M is exceeded for all living arms), and the process can only be continued by generating a "new" arm.

In what follows, we will refer to process κ simply as "arm κ," and we use $S_\kappa(t)$ to denote its state at the beginning of round $t\in\mathbb{N}$.

Action space. We will use $a^1_\kappa(t)=1$ if arm κ is played in round t and $a^1_\kappa(t)=0$ otherwise. Similarly, $a^2_\kappa(t)=1$ if arm κ is regenerated at time t and $a^2_\kappa(t)=0$ otherwise. If an arm is regenerated, it must also be played, i.e., $a^2_\kappa(t)\leqslant a^1_\kappa(t)\ \forall\ \kappa\in[K],\ \forall\ t$. Thus, the feasible ordered set of actions (a^1_κ,a^2_κ) in any state is $\{(0,0),(1,0),(1,1)\}$.

Rewards. For each $\kappa\in[K]$, the reward function $R_\kappa(S_\kappa,a^1_\kappa,a^2_\kappa)$ is defined as:

$$R_\kappa(S_\kappa,a^1_\kappa,a^2_\kappa)=\begin{cases}0 & \text{if } a^1_\kappa=0 \text{ and } a^2_\kappa=0\\ \mu_\kappa g_\kappa(S_\kappa)-c_\kappa\mathbb{1}\{S_\kappa=0\} & \text{if } a^1_\kappa=1 \text{ and } a^2_\kappa=0\\ \mu_\kappa-c_\kappa & \text{if } a^1_\kappa=1 \text{ and } a^2_\kappa=1\ ,\end{cases}$$

where the state S_κ takes values in $\{0,\cdots,M-1\}$.

Transition Kernels. For each $(\kappa,t)\in[K]\times\mathbb{N}$, state transitions are specified by the rules:

1. If $S_\kappa(t)=0$, then
$$S_\kappa(t+1)=\begin{cases}0 & \text{if } a^1_\kappa(t)=0 \text{ and } a^2_\kappa(t)=0\\ 1 & \text{if } a^1_\kappa(t)=1 \text{ and } a^2_\kappa(t)=0\\ 1 & \text{if } a^1_\kappa(t)=1 \text{ and } a^2_\kappa(t)=1\ .\end{cases}$$
2. If $S_\kappa(t)\in[M-2]$, then
$$S_\kappa(t+1)=\begin{cases}S_\kappa(t)+1 & \text{if } a^1_\kappa(t)=0 \text{ and } a^2_\kappa(t)=0\\ S_\kappa(t)+1 & \text{if } a^1_\kappa(t)=1 \text{ and } a^2_\kappa(t)=0\\ 1 & \text{if } a^1_\kappa(t)=1 \text{ and } a^2_\kappa(t)=1\ .\end{cases}$$
3. If $S_\kappa(t)=M-1$, then
$$S_\kappa(t+1)=\begin{cases}0 & \text{if } a^1_\kappa(t)=0 \text{ and } a^2_\kappa(t)=0\\ 0 & \text{if } a^1_\kappa(t)=1 \text{ and } a^2_\kappa(t)=0\\ 1 & \text{if } a^1_\kappa(t)=1 \text{ and } a^2_\kappa(t)=1\ .\end{cases}$$

The infinite-horizon average-cost formulation of the problem in (2) is then given by:

$$\begin{aligned}&\sup_\pi \liminf_{T\to\infty}\mathbb{E}_\pi\left[\frac{1}{T}\sum_{t\in[T]}\sum_{\kappa\in[K]}R_\kappa\big(S_\kappa,a^1_\kappa(t),a^2_\kappa(t)\big)\right]\\ \text{s.t. }&\sum_{\kappa\in[K]}a^1_\kappa(t)=1\quad\forall\ t\\ &a^2_\kappa(t)\leqslant a^1_\kappa(t)\quad\forall\ \kappa\in[K],\ \forall\ t\\ &a^1_\kappa(t),a^2_\kappa(t)\in\{0,1\}\quad\forall\ \kappa\in[K],\ \forall\ t\ .\end{aligned}$$

Lagrangian Problem:

$$\sup_{a^1(\cdot),a^2(\cdot)} L(\lambda) = \liminf_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]}\left(\sum_{\kappa\in[K]} R_\kappa\big(S_\kappa(t),\ a^1_\kappa(t),\ a^2_\kappa(t)\big)\ -\ \lambda\left(\sum_{\kappa\in[K]} a^1_\kappa(t)\ -\ 1\right)\right)\right]$$
$$\text{s.t. } a^2_\kappa(t) \leqslant a^1_\kappa(t) \quad \forall\ \kappa\in[K],\ \forall\ t$$
$$a^1_\kappa(t), a^2_\kappa(t) \in \{0,1\} \quad \forall\ \kappa\in[K],\ \forall\ t\ .$$

The Lagrangian objective can be simplified to

$$\begin{aligned} L(\lambda) &= \liminf_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]}\left(\sum_{\kappa\in[K]} R_\kappa\big(S_\kappa(t),\ a^1_\kappa(t),\ a^2_\kappa(t)\big) - \lambda\sum_{\kappa\in[K]} a^1_\kappa(t) + \lambda\right)\right] \\ &= \left(\liminf_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]}\sum_{\kappa\in[K]}\left(R_\kappa\big(S_\kappa(t),\ a^1_\kappa(t),\ a^2_\kappa(t)\big) - \lambda a^1_\kappa(t)\right)\right]\right) + \lambda\ . \end{aligned}$$

Define

$$\tilde{R}_\kappa\big(S_\kappa(t),\ a^1_\kappa(t),\ a^2_\kappa(t)\big) := R_\kappa\big(S_\kappa(t),\ a^1_\kappa(t),\ a^2_\kappa(t)\big) - \lambda a^1_\kappa(t)\ .$$

Then, we have that

$$L(\lambda) = \left(\liminf_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]}\sum_{\kappa\in[K]}\tilde{R}_\kappa\big(S_\kappa(t),\ a^1_\kappa(t),\ a^2_\kappa(t)\big)\right]\right) + \lambda\ ,$$

which is separable in the arms. Letting

$$\eta_\kappa(\lambda) := \liminf_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t\in[T]}\tilde{R}_\kappa\big(S_\kappa(t),\ a^1_\kappa(t),\ a^2_\kappa(t)\big)\right]\ ,$$

we can write the decoupled problem for arm κ as

$$\begin{aligned} \mathcal{P}_\kappa(\lambda) := \sup_{a^1_\kappa(\cdot),a^2_\kappa(\cdot)} \quad & \eta_\kappa(\lambda) \\ \text{s.t. } & a^2_\kappa(t) \leqslant a^1_\kappa(t) \quad \forall\ t \\ & a^1_\kappa(t), a^2_\kappa(t) \in \{0,1\} \quad \forall\ t\ . \end{aligned} \tag{21}$$

Lemma 2 *The optimal value of* (21)*, starting from any initial state* $S_\kappa(1) \in \{0,\cdots,M-1\}$*, is given by*

$$\mathcal{P}_\kappa(\lambda) = \left(\max_{m\in[M]}\left(\frac{\mu_\kappa\sum_{t=0}^{m-1} g_\kappa(t) - c_\kappa}{m}\right) - \lambda\right)^+\ .$$

Proof of Lemma 2.

Define

$$\bar{\mu}_{\kappa,m} := \frac{\mu_\kappa \sum_{t=0}^{m-1} g_\kappa(t) - c_\kappa}{m} \quad \text{for } m \in [M].$$

$$m^* := \inf\left(\arg\max_{m\in[M]} \bar{\mu}_{\kappa,m}\right).$$

$$V^*(s) := \begin{cases} 0 & \text{if } s = 0 \\ (\bar{\mu}_{\kappa,m^*} - \bar{\mu}_{\kappa,s})s & \text{if } 1 \leqslant s \leqslant m^* \\ 0 & \text{if } m^* < s \leqslant M-1\ . \end{cases} \tag{22}$$

Without loss of generality, assume that $\lambda \leqslant \bar{\mu}_{\kappa,m^*}$. We will show that the following holds:

$$\bar{\mu}_{\kappa,m^*} - \lambda + V^*(s) = \begin{cases} \max\{V^*(0)\ ,\ \mu_\kappa - c - \lambda + V^*(1)\} & \text{if } s = 0 \\ \max\{V^*(s+1)\ ,\ \mu_\kappa g_\kappa(s) - \lambda + V^*(s+1)\ ,\ \mu_\kappa - c - \lambda + V^*(1)\} & \text{if } s \in [M-2] \\ \max\{V^*(0)\ ,\ \mu_\kappa g_\kappa(M-1) - \lambda + V^*(0)\ ,\ \mu_\kappa - c - \lambda + V^*(1)\} & \text{if } s = M-1 \end{cases} \tag{23}$$

Note that (23) is Bellman's equation for the infinite-horizon average-cost dynamic program in (21). We will now show that the vector $(V^*(s) : s = 0, \ldots, M-1)$ defined in (22) satisfies (23).

Note that $\mu_\kappa - c - \lambda + V^*(1) = \bar{\mu}_{\kappa,m^*} - \lambda$. The $s = 0$ case is easily verified. Now consider $s \in [m^* - 1]$. Note that

$$\begin{aligned} \mu_\kappa g_\kappa(s) - \lambda + V^*(s+1) - (\mu_\kappa - c - \lambda + V^*(1)) &= \mu_\kappa g_\kappa(s) - \lambda + (\bar{\mu}_{\kappa,m^*} - \bar{\mu}_{\kappa,s+1})(s+1) - (\bar{\mu}_{\kappa,m^*} - \lambda) \\ &= (\bar{\mu}_{\kappa,m^*} - \bar{\mu}_{\kappa,s})s \\ &= V^*(s) \\ &\geqslant 0\ . \end{aligned}$$

Also note that $\bar{\mu}_{\kappa,m^*}$ is the average of $\{\mu_\kappa - c_\kappa, \mu_\kappa g_\kappa(1), \ldots, \mu_\kappa g_\kappa(m^* - 1)\}$. By the definition of m^*, it follows that $\mu_\kappa g_\kappa(m^* - 1) \geqslant \bar{\mu}_{\kappa,m^*}$ must necessarily hold (else it would imply $\bar{\mu}_{\kappa,m^*-1} > \bar{\mu}_{\kappa,m^*}$, leading to a contradiction). Therefore, we have $\mu_\kappa g_\kappa(s) \geqslant \bar{\mu}_{\kappa,m^*}$ for all $s \in \{0, \ldots, m^* - 1\}$. Finally, note that

$$\mu_\kappa g_\kappa(s) - \lambda + V^*(s+1) = \bar{\mu}_{\kappa,m^*} - \lambda + V^*(s).$$

Thus, the case of $s \in [m^* - 1]$ is verified as well. Lastly, note that for $m^* \leqslant s \leqslant M-1$, we have

$$\mu_\kappa - c - \lambda + V^*(1) - (\mu_\kappa g_\kappa(s) - \lambda + V^*(s+1)) = \bar{\mu}_{\kappa,m^*} - \mu_\kappa g_\kappa(s) \geqslant \bar{\mu}_{\kappa,m^*} - \mu_\kappa g_\kappa(m^*) \underset{\dagger}{\geqslant} 0\ ,$$

where (†) again follows from the definition of m^*. Thus, we have verified that (23) holds when $\lambda \leqslant \bar{\mu}_{\kappa,m^*}$. From the theory of infinite-horizon average-cost dynamic programming (see, e.g., Proposition 5.5.1 in Bertsekas (2012)), it then follows that $\mathcal{P}_\kappa(\lambda) = \bar{\mu}_{\kappa,m^*} - \lambda$ whenever $\lambda \leqslant \bar{\mu}_{\kappa,m^*}$.

Similarly, when $\lambda > \bar{\mu}_{\kappa,m^*}$, one can show that the "always idle" policy solves a similar Bellman equation with $\tilde{V}^*(s) = 0$ for all $s \in \{0, \ldots, M-1\}$, achieving $\mathcal{P}_\kappa(\lambda) = 0$. We skip the details for brevity. Combining the two results proves the assertion. □

Lemma 3 *Consider the decoupled problem for any arm κ, $\mathcal{P}_\kappa(\lambda)$. Then $\mathcal{P}_\kappa(\lambda)$ is indexable. That is, let $\mathcal{S}_\kappa(\lambda)$ denote the set of all states for which it is optimal to idle (i.e., $a^1_\kappa(t) = 0$). Then $\mathcal{S}_\kappa(\lambda)$ increases monotonically as λ increases from 0 to $+\infty$. Furthermore, let $W_\kappa(s)$ denote the Whittle Index in state s which is defined as the infimum charge λ that makes the passive and the best of the active decisions equally desirable. Then, we have that*

$$W_\kappa(s) := \max_{m\in[M]} \left(\frac{\mu_\kappa \sum_{t=0}^{m-1} g_\kappa(t) - c_\kappa}{m} \right).$$

Before we prove this result, we make some observations about the Whittle index computed in Lemma 3. (i) The Whittle index for any arm is independent of s which implies that unlike conventional restless bandit problems, the Whittle index computation in our case is not expensive since it needs to be performed only once. (ii) The optimal arm to pull is stationary and does not change over time. Hence, this implies that we continue to pull the same arm repeatedly instead of switching between different arms over time.

Proof of Lemma 3: Recall that the Whittle index of a state is the *infimum* λ that makes the no-pull decision equivalent to the pull decision in that state. We know from (the proof of) Lemma 2 that for $\lambda < \bar{\mu}_{\kappa,m^*}$, the optimal policy, starting from initial state $s \in \{0,\ldots,M-1\}$, is to continue pulling the arm (without regeneration) until state 0 is reached and then switch to the periodic policy characterized by cycle length m^*.

Next, note that for any initial state s, when $\lambda > \bar{\mu}_{\kappa,m^*}$, the optimal policy is to make the no-pull decision at all times (also follows from Lemma 2). Hence, the smallest charge λ that makes the pull and no-pull decisions equivalent in state s is

$$W(s) = \bar{\mu}_{\kappa,m^*} \ .$$

Note that this result also shows the indexability of the decoupled problem since, if $\lambda \leqslant W(s)$, then $\mathcal{S}(\lambda)$ is an empty set. Otherwise, when $\lambda > W(s)$, we have $\mathcal{S}(\lambda) = \{0,\ldots,M-1\}$. That is, it includes all states. Hence, this proves the final result. □

B. Online Learning: Technical Preliminaries

Below, we present the key technical definitions and facts used in the regret analysis, adapted from Vershynin (2018).

Definition 2 (sub-Gaussianity) *A real-valued random variable W is said to be sub-Gaussian with variance proxy σ^2 (aka σ-sub-Gaussian, where $\sigma \geqslant 0$), denoted by $W \sim \texttt{subG}(\sigma^2)$, iff*

$$\mathbb{E}[W] = 0 \qquad \textit{and} \qquad \mathbb{E}[\exp(\theta W)] \leqslant \exp\left(\frac{\sigma^2\theta^2}{2}\right) \ \forall\ \theta \in \mathbb{R}.$$

Fact 1 (sub-Gaussianity with scale) *If* $W \sim \texttt{subG}(\sigma^2)$*, then* $aW \sim \texttt{subG}(a^2\sigma^2)$ *for any* $a \in \mathbb{R}$*.*

Fact 2 (sub-Gaussian tail bound) *Fix* $c \geqslant 0$ *and* $\delta \in (0,1]$*. If* $W \sim \texttt{subG}(\sigma^2)$*, then* $\max\left\{\mathbb{P}\left(W \geqslant \sigma\sqrt{2c\log(1/\delta)}\right),\ \mathbb{P}\left(W \leqslant -\sigma\sqrt{2c\log(1/\delta)}\right)\right\} \leqslant \delta^c$*.*

Fact 3 (Sum of independent sub-Gaussian random variables) *Suppose* $(W_i : i \in [n])$ *is a collection of* n *independent random variables satisfying* $W_i \sim \texttt{subG}(\sigma_i^2)$ *for each* $i \in [n]$*. Then,* $\sum_{i\in[n]} W_i \sim \texttt{subG}\left(\sum_{i\in[n]} \sigma_i^2\right)$*.*

C. Lower Bound: Proof of Theorem 3

Fix horizon T, policy $\pi \in \Pi$, costs $c_\kappa = 0\ \forall\ \kappa \in [K]$, and suppose that the noise ξ_t is i.i.d. Gaussian with variance σ^2 in each round $t \in [T]$. Then,

$$\mathbb{E}_\pi[\mathfrak{R}(T)] = \mathbb{E}_\pi\left[\texttt{OPT}_T - \sum_{t\in[T]} r_t\right] \underset{\dagger}{=} \mathbb{E}_\pi\left[\mu_{\kappa^*}T - \sum_{t\in[T]} r_t\right], \tag{24}$$

where (†) follows using Theorem 1.

Now consider a K-armed bandit problem endowed with i.i.d. Gaussian rewards with unknown means $(\mu_\kappa : \kappa \in [K])$ and variance σ^2. We will use $\bar{r}_t$ to denote the reward realized in round t of this problem, and $\bar{\Pi}$ to denote the natural class of history-dependent, non-anticipating policies for the problem.

Claim 1 *For every* $\pi \in \Pi$*, there exists some* $\bar{\pi} \in \bar{\Pi}$ *such that* $\mathbb{E}_{\bar{\pi}}\left[\sum_{t\in[T]} \bar{r}_t\right] \geqslant \mathbb{E}_\pi\left[\sum_{t\in[T]} r_t\right]$*.*

Proof of Claim 1.

Consider an arbitrary policy $\pi \in \Pi$. Construct a policy $\bar{\pi} \in \bar{\Pi}$ as follows: whenever π prescribes pulling an arm of type κ in the original problem (either new or old), $\bar{\pi}$ prescribes pulling the arm with mean μ_κ in the K-armed bandit problem. The claim then follows from the monotonicity property $g_\kappa(m-1) \leqslant g_\kappa(m)$ for all $(\kappa, m) \in [K] \times [M-1]$, together with the assumption that the costs satisfy $c_\kappa = 0$ for all $\kappa \in [K]$. □

Invoking Claim 1 in (24), we obtain $\mathbb{E}_\pi[\mathfrak{R}(T)] \geqslant \mathbb{E}_{\bar{\pi}}\left[\mu_{\kappa^*}T - \sum_{t\in[T]} \bar{r}_t\right]$, where the policy $\bar{\pi} \in \bar{\Pi}$ satisfies the condition in Claim 1. Note that $\mathbb{E}_{\bar{\pi}}\left[\mu_{\kappa^*}T - \sum_{t\in[T]} \bar{r}_t\right]$ represents the (expected cumulative) regret of $\bar{\pi}$ in the aforementioned K-armed bandit problem.

From standard multi-armed bandit theory (see, e.g., Theorem 15.2 in Lattimore and Szepesvári (2020)), we know that for every policy $\bar{\pi} \in \bar{\Pi}$, there exists a mean reward vector $(\mu_\kappa : \kappa \in [K])$ such that $\mathbb{E}_{\bar{\pi}}\left[\mu_{\kappa^*}T - \sum_{t\in[T]} \bar{r}_t\right] = \Omega\left(\sigma\sqrt{KT}\right)$. Thus, it follows that for every policy $\pi \in \Pi$, there exists an instance ν of the original problem such that $\mathbb{E}_\pi[\mathfrak{R}(T;\nu)] = \Omega\left(\sigma\sqrt{KT}\right)$. □

D. Regret Analysis of Algorithm 1

First, we adapt the following key lemma from Slivkins et al. (2019) to our setting.

Lemma 4 (Clean Execution of Algorithm 1) *Define "clean event"*

$$E := \bigcap_{\substack{\kappa\in[K]\\ m\in[M]\\ t\in[T]}} \left\{ \left|\hat{\mu}^{(t)}_{\kappa,m} - \bar{\mu}_{\kappa,m}\right| \leqslant \texttt{rad}_{\kappa,m,t} \right\} ,$$

where $\bar{\mu}_{\kappa,m}$, $\hat{\mu}^{(t)}_{\kappa,m}$ *and* $\texttt{rad}_{\kappa,m,t}$ *are as defined in* (5), (8) *and* (9) *respectively. Then, one has that* $\mathbb{P}(E) \geqslant 1 - 2KMT\delta^{\alpha}$.

Proof of Lemma 4. Recall that $\hat{\nu}^{(n)}_{\kappa,j}$ denotes the empirical mean of n samples from type κ arms of age $j-1$. Then, note that

$$F := \bigcap_{\substack{\kappa\in[K]\\ j\in[M]\\ n\in[T]}} \left\{ \left|\hat{\nu}^{(n)}_{\kappa,j} - \mu_\kappa g_\kappa(j-1)\right| \leqslant 2\sigma\sqrt{\frac{\alpha\log\left(\frac{1}{\delta}\right)}{1+n}} \right\} \subseteq E. \tag{25}$$

We know that $\left(\hat{\nu}^{(n)}_{\kappa,j} - \mu_\kappa g_\kappa(j-1)\right) \sim \texttt{subG}(\sigma^2/n)$. Then, it follows from Fact 2 that $\mathbb{P}(F^{\mathsf{c}}) \leqslant 2KMT\delta^{\frac{2\alpha n}{1+n}} \leqslant 2KMT\delta^{\alpha}$. The claim now follows using (25). □

The next result is adapted from Lemma 11 of Abbasi-Yadkori et al. (2011).

Lemma 5 (Elliptical Potential Lemma) *Let* $\mathcal{S}_{j,t} \subseteq [t]$ *denote the subset of rounds until* t *in which an arm of age* $j-1$ *is played by Algorithm 1. Then, the following holds for any* $t \in \mathbb{N}$*:*

$$\sum_{s\in\mathcal{S}_{1,t}} \left(\frac{\rho_{\kappa_s,1}}{1+\sum_{j\in[M]}\rho_{\kappa_s,j}N_{\kappa_s,j}(s-1)} \right) \leqslant 2K\log\left(1+\frac{t}{K}\right) .$$

Proof of Lemma 5.

Let $\{x_k : k\in[K]\}$ be the collection of standard basis vectors of $\mathbb{R}^K$, with each x_k being 1 in its k^{th} coordinate, and 0 everywhere else. Let $\mathcal{I}$ be the identity matrix in $\mathbb{R}^{K\times K}$. Recall that $N_{k,j}(t)$ denotes the number of rounds in $[t]$ in which a type k arm of age $j-1$ is played by the algorithm. Define for $t\in\mathbb{N}$ the following:

$$V_t := \mathcal{I} + \sum_{j\in[M]}\sum_{u\in\mathcal{S}_{j,t}} \rho_{\kappa_u,j}(x_{\kappa_u}\otimes x_{\kappa_u}) ,$$

Note that $\|x_{\kappa_s}\|^2_{V_{s-1}^{-1}} = \frac{1}{1+\sum_{j\in[M]}\rho_{\kappa_s,j}N_{\kappa_s,j}(s-1)}$. Observe that V_t can also be expressed as:

$$V_t = \mathcal{I} + \sum_{u\in[t]} \rho_{\kappa_u,\texttt{age}_u(\pi_u)+1}(x_{\kappa_u}\otimes x_{\kappa_u}) .$$

Finally, note that

$$\begin{aligned}\sum_{s\in\mathcal{S}_{1,t}}\left(\frac{\rho_{\kappa_s,1}}{1+\sum_{j\in[M]}\rho_{\kappa_s,j}N_{\kappa_s,j}(s-1)}\right)=\sum_{s\in\mathcal{S}_{1,t}}\left\|\sqrt{\rho_{\kappa_s,1}}x_{\kappa_s}\right\|^2_{V_{s-1}^{-1}}&=\sum_{s\in\mathcal{S}_{1,t}}\left\|\sqrt{\rho_{\kappa_s,\mathtt{age}_s(\pi_s)+1}}x_{\kappa_s}\right\|^2_{V_{s-1}^{-1}}\\&\leqslant\sum_{s\in[t]}\left\|\sqrt{\rho_{\kappa_s,\mathtt{age}_s(\pi_s)+1}}x_{\kappa_s}\right\|^2_{V_{s-1}^{-1}}\\&\underset{\dagger}{\leqslant}2\log\left(\mathtt{det}(V_t)\right)\\&\underset{\ddagger}{\leqslant}2K\log\left(1+\frac{t}{K}\right),\end{aligned}$$

where (†) follows using Lemma 11 of Abbasi-Yadkori et al. (2011), and (‡) follows using the Determinant-Trace inequality (see, e.g., Lemma 10 of Abbasi-Yadkori et al. (2011)). □

D.1. Proof of Theorem 4

Recall that $\mathcal{S}_{1,t}\subseteq[t]$ is the subset of rounds until t in which the UCB rule applies. Let $\bar{s}_t$ be the largest element of $\mathcal{S}_{1,t}$, and define $\tau_t:=t-\bar{s}_t+1$. Note that $\tau_t\leqslant m_{\bar{s}_t}\leqslant M$. Then, the regret of Algorithm 1 after t rounds is given by

$$\begin{aligned}\mathfrak{R}(t)&=\mathtt{OPT}_t-\left(\sum_{s\in\mathcal{S}_{1,t}\setminus\{\bar{s}_t\}}\bar{\mu}_{\kappa_s,m_s}m_s+\mu_{\kappa_{\bar{s}_t}}\sum_{u=0}^{\tau_t-1}g_{\kappa_{\bar{s}_t}}(u)-c_{\kappa_{\bar{s}_t}}\right)\\&\underset{\dagger}{\leqslant}\bar{\mu}_{\kappa^*,m^*}t-\left(\sum_{s\in\mathcal{S}_{1,t}\setminus\{\bar{s}_t\}}\bar{\mu}_{\kappa_s,m_s}m_s+\mu_{\kappa_{\bar{s}_t}}\sum_{u=0}^{\tau_t-1}g_{\kappa_{\bar{s}_t}}(u)-c_{\kappa_{\bar{s}_t}}\right)\\&=\sum_{s\in\mathcal{S}_{1,t}\setminus\{\bar{s}_t\}}(\bar{\mu}_{\kappa^*,m^*}-\bar{\mu}_{\kappa_s,m_s})m_s+\bar{\mu}_{\kappa^*,m^*}\tau_t-\left(\mu_{\kappa_{\bar{s}_t}}\sum_{u=0}^{\tau_t-1}g_{\kappa_{\bar{s}_t}}(u)-c_{\kappa_{\bar{s}_t}}\right)\\&=\sum_{s\in\mathcal{S}_{1,t}\setminus\{\bar{s}_t\}}(\bar{\mu}_{\kappa^*,m^*}-\bar{\mu}_{\kappa_s,m_s})m_s+\bar{\mu}_{\kappa^*,m^*}\tau_t-\bar{\mu}_{\kappa_{\bar{s}_t},m_{\bar{s}_t}}m_{\bar{s}_t}+\mu_{\kappa_{\bar{s}_t}}\sum_{u=\tau_t}^{m_{\bar{s}_t}-1}g_{\kappa_{\bar{s}_t}}(u)\\&=\sum_{s\in\mathcal{S}_{1,t}}(\bar{\mu}_{\kappa^*,m^*}-\bar{\mu}_{\kappa_s,m_s})m_s+\mu_{\kappa_{\bar{s}_t}}\sum_{u=\tau_t}^{m_{\bar{s}_t}-1}g_{\kappa_{\bar{s}_t}}(u)-\bar{\mu}_{\kappa^*,m^*}(m_{\bar{s}_t}-\tau_t)\\&\underset{\ddagger}{\leqslant}\sum_{s\in\mathcal{S}_{1,t}}(\bar{\mu}_{\kappa^*,m^*}-\bar{\mu}_{\kappa_s,m_s})m_s+\mu_{\kappa_{\bar{s}_t}}\sum_{u=0}^{m_{\bar{s}_t}-\tau_t-1}g_{\kappa_{\bar{s}_t}}(u)-\bar{\mu}_{\kappa^*,m^*}(m_{\bar{s}_t}-\tau_t)\\&\leqslant\sum_{s\in\mathcal{S}_{1,t}}(\bar{\mu}_{\kappa^*,m^*}-\bar{\mu}_{\kappa_s,m_s})m_s-\left(\bar{\mu}_{\kappa^*,m^*}-\bar{\mu}_{\kappa_{\bar{s}_t},m_{\bar{s}_t}-\tau_t}\right)(m_{\bar{s}_t}-\tau_t)+c_{\kappa_{\bar{s}_t}}\\&\underset{*}{\leqslant}c_{\kappa_{\bar{s}_t}}+\sum_{s\in\mathcal{S}_{1,t}}(\bar{\mu}_{\kappa^*,m^*}-\bar{\mu}_{\kappa_s,m_s})m_s\\&\leqslant\max_{\kappa\in[K]}c_\kappa+\sum_{s\in\mathcal{S}_{1,t}}(\bar{\mu}_{\kappa^*,m^*}-\bar{\mu}_{\kappa_s,m_s})m_s\ ,\end{aligned}\tag{26}$$

where (†) follows from Theorem 1, (‡) follows since $\mu_\kappa\geqslant 0$ and $g_\kappa(u)$ is non-increasing in u for any $\kappa\in[K]$, $(*)$ follows since $(\kappa^*,m^*)\in\arg\max_{(\kappa,m)\in[K]\times[M]}\bar{\mu}_{\kappa,m}$ and $m_{\bar{s}_t}\geqslant\tau_t$.

Clean Execution.

Suppose the "clean event" E defined in Lemma 4 is true. Then, one has

$$\bar{\mu}_{\kappa_s,m_s} + 2\texttt{rad}_{\kappa_s,m_s,s-1} \underset{\dagger}{\geqslant} \hat{\mu}^{(s-1)}_{\kappa_s,m_s} + \texttt{rad}_{\kappa_s,m_s,s-1} \underset{*}{\geqslant} \hat{\mu}^{(s-1)}_{\kappa^*,m^*} + \texttt{rad}_{\kappa^*,m^*,s-1} \underset{\ddagger}{\geqslant} \bar{\mu}_{\kappa^*,m^*},$$

where (†) and (‡) follow since E is true and $(*)$ holds since the UCB rule applies in round s. We therefore have that

$$\begin{aligned}
\bar{\mu}_{\kappa^*,m^*} - \bar{\mu}_{\kappa_s,m_s} &\leqslant 2\texttt{rad}_{\kappa_s,m_s,s-1} \\
&= 4\sigma \left(\frac{\sum_{j\in[M]} \sqrt{\alpha \left(\frac{\rho_{\kappa_s,j}}{g_{\kappa_s}(j-1)} \right)^2 (1 + N_{\kappa_s,j}(s-1)) \log\left(\frac{1}{\delta}\right)}}{1 + \sum_{j\in[M]} \rho_{\kappa_s,j} N_{\kappa_s,j}(s-1)} \right) \left(\frac{\sum_{j\in[m_s]} g_{\kappa_s}(j-1)}{m_s} \right) \\
&\underset{\dagger}{\leqslant} 4\sigma \sqrt{\frac{\alpha \left(\sum_{j\in[M]} \left(\frac{\rho_{\kappa_s,j}}{g^2_{\kappa_s}(j-1)} \right) \right) \log\left(\frac{1}{\delta}\right)}{1 + \sum_{j\in[M]} \rho_{\kappa_s,j} N_{\kappa_s,j}(s-1)}} \left(\frac{\sum_{j\in[m_s]} g_{\kappa_s}(j-1)}{m_s} \right). \qquad (27)
\end{aligned}$$

where (†) follows using the Cauchy-Schwarz inequality, together with $\sum_{j\in[M]} \rho_{\kappa_s,j} = 1$. Combining (26) and (27), we have that the following holds under the "clean event" E:

$$\begin{aligned}
&\mathfrak{R}(t) - \max_{\kappa\in[K]} c_\kappa \\
&\leqslant 4\sigma \sqrt{\alpha \log\left(\frac{1}{\delta}\right)} \sum_{s\in\mathcal{S}_{1,t}} \sqrt{\frac{\left(\sum_{j\in[M]} \left(\frac{\rho_{\kappa_s,j}}{g^2_{\kappa_s}(j-1)} \right) \right) \left(\sum_{j\in[m_s]} g_{\kappa_s}(j-1) \right)^2}{1 + \sum_{j\in[M]} \rho_{\kappa_s,j} N_{\kappa_s,j}(s-1)}} \\
&\underset{\dagger}{\leqslant} 4\sigma \sqrt{\alpha \log\left(\frac{1}{\delta}\right) \left(\sum_{s\in\mathcal{S}_{1,t}} \left\{ \left(\sum_{j\in[M]} \left(\frac{\rho_{\kappa_s,j}/\rho_{\kappa_s,1}}{g^2_{\kappa_s}(j-1)} \right) \right) \left(\sum_{j\in[m_s]} g_{\kappa_s}(j-1) \right)^2 \right\} \right) \sum_{s\in\mathcal{S}_{1,t}} \left\{ \frac{\rho_{\kappa_s,1}}{1 + \sum_{j\in[M]} \rho_{\kappa_s,j} N_{\kappa_s,j}(s-1)} \right\}} \\
&\underset{\ddagger}{\leqslant} 4\sigma \sqrt{2\alpha K \log\left(1 + \frac{t}{K}\right) \log\left(\frac{1}{\delta}\right) \sum_{s\in\mathcal{S}_{1,t}} \left\{ \left(\sum_{j\in[M]} \left(\frac{\rho_{\kappa_s,j}/\rho_{\kappa_s,1}}{g^2_{\kappa_s}(j-1)} \right) \right) \left(\sum_{j\in[m_s]} g_{\kappa_s}(j-1) \right)^2 \right\}},
\end{aligned}$$

where (†) follows using the Cauchy-Schwarz inequality, and (‡) follows from Lemma 5. □

E. Regret Analysis of Algorithm 2

We proceed by establishing a "clean execution" result for Algorithm 2, analogous to Lemma 4.

Lemma 6 (Clean Execution of Algorithm 2) *Define "clean event"*

$$E := \bigcap_{\substack{\kappa\in[K] \\ m\in[M] \\ t\in[T]}} \left\{ \left| \hat{\mu}^{(t)}_{\kappa,m} - \bar{\mu}_{\kappa,m} \right| \leqslant \texttt{rad}_{\kappa,m,t} \right\},$$

where $\bar{\mu}_{\kappa,m}$, $\hat{\mu}^{(t)}_{\kappa,m}$ *and* $\texttt{rad}_{\kappa,m,t}$ *are as defined in* (5), (13) *and* (14) *respectively. Then, one has that* $\mathbb{P}(E) \geqslant 1 - 2KMT\delta^\alpha$.

Proof of Lemma 6. Recall that $\hat{\nu}^{(n)}_{\kappa,j}$ denotes the empirical mean of n samples from type κ arms of age $j-1$. Then, note that

$$F := \bigcap_{\substack{\kappa\in[K]\\ j\in[M]\\ n\in[T]}} \left\{ \left| \hat{\nu}^{(n)}_{\kappa,j} - \mu_\kappa g_\kappa(j-1) \right| \leqslant 2\sigma \sqrt{\frac{\alpha \log\left(\frac{1}{\delta}\right)}{1+n}} \right\} \subseteq E. \tag{28}$$

We know that $\left(\hat{\nu}^{(n)}_{\kappa,j} - \mu_\kappa g_\kappa(j-1)\right) \sim \texttt{subG}(\sigma^2/n)$. Then, it follows from Fact 2 that $\mathbb{P}(F^{\mathtt{c}}) \leqslant 2KMT\delta^{\frac{2\alpha n}{1+n}} \leqslant 2KMT\delta^{\alpha}$. The claim now follows using (28). □

The next result is adapted from Lemma 11 of Abbasi-Yadkori et al. (2011).

Lemma 7 (Elliptical Potential Lemma) *Let $\mathcal{S}_{m,t} \subseteq [t]$ denote the subset of rounds until t in which an arm of age $m-1$ is played by Algorithm 2. Then, the following holds for any $m \in [M]$ and $t \in \mathbb{N}$:*

$$\sum_{s\in\mathcal{S}_{m,t}} \left(\frac{1}{1+N_{\kappa_s,m}(s-1)} \right) \leqslant 2K \log\left(1 + \frac{|\mathcal{S}_{m,t}|}{K}\right) .$$

Proof of Lemma 7.

Let $\{x_k : k \in [K]\}$ be the collection of standard basis vectors of $\mathbb{R}^K$, with each x_k being 1 in its k^{th} coordinate, and 0 everywhere else. Let $\mathcal{I}$ be the identity matrix in $\mathbb{R}^{K\times K}$. Recall that $N_{k,m}(t)$ denotes the number of rounds in $[t]$ in which a type k arm of age $m-1$ is played by the algorithm. Define for $m \in [M]$ and $t \in \mathbb{N}$ the following:

$$V_{m,t} := \mathcal{I} + \sum_{s\in\mathcal{S}_{m,t}} x_{\kappa_s} \otimes x_{\kappa_s} \ ,$$

with the convention that $V_{m,t} := \mathcal{I}$ if $\mathcal{S}_{m,t} = \varnothing$. Then, note that $\|x_{\kappa_s}\|^2_{V^{-1}_{m,s-1}} = \frac{1}{1+N_{\kappa_s,m}(s-1)}$. We know from Lemma 11 of Abbasi-Yadkori et al. (2011) that $\sum_{s\in\mathcal{S}_{m,t}} \|x_{\kappa_s}\|^2_{V^{-1}_{m,s-1}} \leqslant 2\log\left(\texttt{det}(V_{m,t})\right)$. By the Determinant-Trace inequality (Lemma 10 of Abbasi-Yadkori et al. (2011)), $\texttt{det}(V_{m,t}) \leqslant \left(1 + |\mathcal{S}_{m,t}|/K\right)^K$. Combining everything establishes the stated assertion. □

E.1. Proof of Theorem 5

Let $\mathcal{S}_{m,t} \subseteq [t]$ denote the subset of rounds until t in which an arm of age $m-1$ is played by Algorithm 2. Then, similarly to the derivation of (26) in the analysis of Algorithm 1, one can show that the regret of Algorithm 2 after any t rounds is bounded as

$$\mathfrak{R}(t) \leqslant \max_{\kappa\in[K]} c_\kappa + \sum_{s\in\mathcal{S}_{1,t}} (\bar{\mu}_{\kappa^*,m^*} - \bar{\mu}_{\kappa_s,m_s}) m_s \ . \tag{29}$$

Clean Execution.

Suppose the "clean event" E is true. Consider round $s \in \mathcal{S}_{1,t}$. Then, one has

$$\bar{\mu}_{\kappa_s,m_s} + 2\texttt{rad}_{\kappa_s,m_s,s-1} \underset{\dagger}{\geqslant} \hat{\mu}^{(s-1)}_{\kappa_s,m_s} + \texttt{rad}_{\kappa_s,m_s,s-1} \underset{*}{\geqslant} \hat{\mu}^{(s-1)}_{\kappa^*,m^*} + \texttt{rad}_{\kappa^*,m^*,s-1} \underset{\ddagger}{\geqslant} \bar{\mu}_{\kappa^*,m^*},$$

where (†) and (‡) follow since E is true and $(*)$ holds since the UCB rule applies in round $s \in \mathcal{S}_{1,t}$. We therefore have that

$$\bar{\mu}_{\kappa^*,m^*} - \bar{\mu}_{\kappa_s,m_s} \leqslant 2\texttt{rad}_{\kappa_s,m_s,s-1} = \frac{4\sigma}{m_s} \sum_{j\in[m_s]} \sqrt{\frac{\alpha \log\left(\frac{1}{\delta}\right)}{1+N_{\kappa_s,j}(s-1)}}\ . \tag{30}$$

Combining (29) and (30),

$$\begin{aligned}
\mathfrak{R}(t) &\leqslant \max_{\kappa\in[K]} c_\kappa + 4\sigma\sqrt{\alpha\log\left(\frac{1}{\delta}\right)} \sum_{s\in\mathcal{S}_{1,t}} \sum_{j\in[m_s]} \frac{1}{\sqrt{1+N_{\kappa_s,j}(s-1)}} \\
&\underset{\dagger}{\leqslant} \max_{\kappa\in[K]} c_\kappa + 4\sigma\sqrt{\alpha\log\left(\frac{1}{\delta}\right)} \sqrt{t^* \sum_{s\in\mathcal{S}_{1,t}} \sum_{j\in[m_s]} \left(\frac{1}{1+N_{\kappa_s,j}(s-1)}\right)} \\
&= \max_{\kappa\in[K]} c_\kappa + 4\sigma\sqrt{\alpha\log\left(\frac{1}{\delta}\right)} \sqrt{t^* \sum_{m\in[M]} \sum_{s\in\mathcal{S}_{m,t^*}} \left(\frac{1}{1+N_{\kappa_s,m}(s-1)}\right)} \\
&\underset{\ddagger}{\leqslant} \max_{\kappa\in[K]} c_\kappa + 4\sigma\sqrt{\alpha\log\left(\frac{1}{\delta}\right)} \sqrt{2Kt^* \sum_{m\in[M]} \log\left(1+\frac{|\mathcal{S}_{m,t^*}|}{K}\right)} \\
&\leqslant \max_{\kappa\in[K]} c_\kappa + 4\sigma\sqrt{\alpha\log\left(\frac{1}{\delta}\right)} \sqrt{2KMt^* \log\left(1+\frac{t^*}{K}\right)} \\
&\underset{\star}{\leqslant} \max_{\kappa\in[K]} c_\kappa + 4\sigma\sqrt{\alpha\log\left(\frac{1}{\delta}\right)} \sqrt{2KM(t+M) \log\left(1+\frac{t+M}{K}\right)}\ ,
\end{aligned}$$

where (†) follows using Jensen's inequality with $t^* := \sum_{s\in\mathcal{S}_{1,t}} m_s$, (‡) follows using Lemma 7, and $(\star)$ follows since $t^* < t+M$. □

F. Proof of Proposition 1

Fix $\kappa \in [K]$. Consider an arbitrary $m \geqslant M$. Then, note that

$$\begin{aligned}
\bar{\mu}_{\kappa,m} - \bar{\mu}_{\kappa,m+1} &= \left(\frac{\mu_\kappa \sum_{u=0}^{m-1} g_\kappa(u) - c_\kappa}{m}\right) - \left(\frac{\mu_\kappa \sum_{u=0}^{m} g_\kappa(u) - c_\kappa}{m+1}\right) \\
&= \frac{1}{m+1}\left(\frac{\mu_\kappa \sum_{u=0}^{m-1} g_\kappa(u) - c_\kappa}{m} - \mu_\kappa g_\kappa(m)\right) \\
&\underset{\dagger}{\geqslant} \frac{1}{m+1}\left(\frac{\mu_\kappa - c_\kappa + \mu_\kappa(m-1)g_\kappa(m)}{m} - \mu_\kappa g_\kappa(m)\right) \\
&= \frac{\mu_\kappa}{m(m+1)}\left(1 - g_\kappa(m) - \frac{c_\kappa}{\mu_\kappa}\right) \\
&\underset{\ddagger}{\geqslant} \frac{\beta\mu_\kappa - c_\kappa}{m(m+1)} \\
&\underset{\star}{\geqslant} 0\ ,
\end{aligned}$$

where (†) follows using the monotonicity of $g_\kappa(\cdot)$ and non-negativity of μ_κ, (‡) follows since $g_\kappa(m) \leqslant g_\kappa(M) \leqslant f_\kappa(M) \leqslant 1-\beta$, and finally ($\star$) follows from Assumption 2. Since $m \geqslant M$ is arbitrary, it follows that $\bar{\mu}_{\kappa,M} \geqslant \bar{\mu}_{\kappa,m}\ \forall\ m \geqslant M$. Therefore, one has $\inf\left(\arg\sup_{m\in\mathbb{N}} \bar{\mu}_{\kappa,m}\right) \leqslant M\ \forall\ \kappa \in [K]$, as asserted. □

G. Regret Analysis of Algorithm 3

First, we adapt the following key lemmas from Abbasi-Yadkori et al. (2011) to our setting.

Lemma 8 (High Probability Confidence Ellipsoids) *Fix failure probability $\delta \in (0,1]$ and for $t = 1, 2, \cdots$, consider the sequence of confidence ellipsoids generated by Algorithm 3:*

$$\Theta_t := \left\{ \theta \in \mathbb{R}^d : \left\| \hat{\theta}_t - \theta \right\|_{V_t} \leqslant \alpha_t \right\}, \tag{31}$$

where V_t, $\hat{\theta}_t$, and α_t are as defined in (38), (39), and (42) respectively. Define "clean event"

$$E := \{\theta^* \in \Theta_t\ \forall\ t = 0, 1, 2, \cdots\}. \tag{32}$$

Then, one has that $\mathbb{P}(E) \geqslant 1-\delta$.

Proof of Lemma 8. The result follows instructively from Theorem 2 of Abbasi-Yadkori et al. (2011). We skip the steps for brevity. □

Lemma 9 (Elliptical Potential Lemma) *In the setting of Lemma 8, if $\lambda \geqslant \max(1, L^2)$, then the following holds for all $t = 1, 2, \cdots$:*

$$\sum_{s\in\mathcal{S}_t} \|x_{\kappa_s}\|^2_{V_{s-1}^{-1}} \leqslant 2d \log\left(1 + \frac{L^2|\mathcal{S}_t|}{\lambda d}\right).$$

Proof of Lemma 9. The result follows instructively by combining Lemma 10 and 11 of Abbasi-Yadkori et al. (2011). We skip the steps for brevity. □

G.1. Proof of Theorem 6

Suppose the "clean event" E in (32) is true. Let $(\kappa^*, m^*) \in \arg\max_{(\kappa,m)\in[K]\times[M]} \bar{\mu}_{\kappa,m}(\theta^*)$. Recall that $\mathcal{S}_t \subseteq [t]$ denotes the subset of rounds until t in which the UCB rule applies. Then, similarly to the derivation of (26) in the analysis of Algorithm 1, one can show that the regret of Algorithm 3 after any t rounds is bounded as

$$\begin{aligned}
\mathfrak{R}(t) &\leqslant \max_{\kappa\in[K]} c_\kappa + \sum_{s\in\mathcal{S}_t} \left(\bar{\mu}_{\kappa^*,m^*}(\theta^*) - \bar{\mu}_{\kappa_s,m_s}(\theta^*)\right) m_s \\
&= \max_{\kappa\in[K]} c_\kappa + \sum_{s\in\mathcal{S}_t} \left(\bar{\mu}_{\kappa^*,m^*}\left(\hat{\theta}_{s-1}\right) - \bar{\mu}_{\kappa_s,m_s}\left(\hat{\theta}_{s-1}\right)\right) m_s + \sum_{s\in\mathcal{S}_t} \left(\bar{\mu}_{\kappa^*,m^*}(\theta^*) - \bar{\mu}_{\kappa^*,m^*}\left(\hat{\theta}_{s-1}\right)\right) m_s
\end{aligned}$$

$$+\sum_{s\in\mathcal{S}_t}\left(\bar{\mu}_{\kappa_s,m_s}\left(\hat{\theta}_{s-1}\right)-\bar{\mu}_{\kappa_s,m_s}(\theta^*)\right)m_s$$

$$\leqslant_{\dagger}\max_{\kappa\in[K]}c_\kappa+\sum_{s\in\mathcal{S}_t}(\texttt{rad}_{\kappa_s,m_s,s-1}-\texttt{rad}_{\kappa^*,m^*,s-1})m_s+\left(\frac{\sum_{j\in[m^*]}g_{\kappa^*}(j-1)}{m^*}\right)\sum_{s\in\mathcal{S}_t}\left\langle x_{\kappa^*},\theta^*-\hat{\theta}_{s-1}\right\rangle m_s$$

$$+\sum_{s\in\mathcal{S}_t}\left(\left\langle x_{\kappa_s},\hat{\theta}_{s-1}-\theta^*\right\rangle\sum_{j\in[m_s]}g_{\kappa_s}(j-1)\right), \tag{33}$$

where (†) follows since (κ_s,m_s) is chosen by the UCB rule in round $s\in\mathcal{S}_t$. Now observe that for any round $s\in\mathcal{S}_t$ and any $\kappa\in[K]$,

$$\left|\left\langle x_\kappa,\hat{\theta}_{s-1}-\theta^*\right\rangle\right|\leqslant_{\ddagger}\|x_\kappa\|_{V_{s-1}^{-1}}\left\|\hat{\theta}_{s-1}-\theta^*\right\|_{V_{s-1}}\leqslant_{*}\|x_\kappa\|_{V_{s-1}^{-1}}\left(\sigma\sqrt{d\log\left(\frac{1+(s-1)L^2/\lambda}{\delta}\right)}+\sqrt{\lambda}S\right)$$

$$=_{\$}\alpha_{s-1}\|x_\kappa\|_{V_{s-1}^{-1}}, \tag{34}$$

where (‡) follows from the Cauchy-Schwarz inequality, and (∗) follows from (31) since $\theta^*\in\Theta_{s-1}$ (the clean event E is true, by assumption), and ($) follows from (42). Combining (33) and (34),

$$\mathfrak{R}(t)\leqslant\max_{\kappa\in[K]}c_\kappa+\sum_{s\in\mathcal{S}_t}(\texttt{rad}_{\kappa_s,m_s,s-1}-\texttt{rad}_{\kappa^*,m^*,s-1})m_s+\left(\frac{\sum_{j\in[m^*]}g_{\kappa^*}(j-1)}{m^*}\right)\sum_{s\in\mathcal{S}_t}\alpha_{s-1}\|x_{\kappa^*}\|_{V_{s-1}^{-1}}m_s$$

$$+\sum_{s\in\mathcal{S}_t}\left(\alpha_{s-1}\|x_{\kappa_s}\|_{V_{s-1}^{-1}}\sum_{j\in[m_s]}g_{\kappa_s}(j-1)\right)$$

$$=_{\star}\max_{\kappa\in[K]}c_\kappa+2\sum_{s\in\mathcal{S}_t}\left(\alpha_{s-1}\|x_{\kappa_s}\|_{V_{s-1}^{-1}}\sum_{j\in[m_s]}g_{\kappa_s}(j-1)\right)$$

$$\leqslant\max_{\kappa\in[K]}c_\kappa+2\alpha_t\sum_{s\in\mathcal{S}_t}\left(\|x_{\kappa_s}\|_{V_{s-1}^{-1}}\sum_{j\in[m_s]}g_{\kappa_s}(j-1)\right)$$

$$\leqslant_{\dagger}\max_{\kappa\in[K]}c_\kappa+2\alpha_t\sqrt{\left(\sum_{s\in\mathcal{S}_t}\|x_{\kappa_s}\|_{V_{s-1}^{-1}}^2\right)\left(\sum_{s\in\mathcal{S}_t}\left(\sum_{j\in[m_s]}g_{\kappa_s}(j-1)\right)^2\right)}$$

$$\leqslant_{\ddagger}\max_{\kappa\in[K]}c_\kappa+2\alpha_t\sqrt{2d\log\left(1+\frac{L^2t}{\lambda d}\right)\left(\sum_{s\in\mathcal{S}_t}\left(\sum_{j\in[m_s]}g_{\kappa_s}(j-1)\right)^2\right)}$$

$$\leqslant\max_{\kappa\in[K]}c_\kappa+2\max_{\kappa\in[K]}\left(\sum_{j\in[M]}g_\kappa(j-1)\right)\alpha_t\sqrt{2dt\log\left(1+\frac{L^2t}{\lambda d}\right)},$$

where (⋆) follows from (41), (†) using the Cauchy-Schwarz inequality, and (‡) using Lemma 9. We have thus proved both (43) and (44). The bound in (45) follows from (43) by observing that

$$\sum_{s\in\mathcal{S}_t}\left(\sum_{j\in[m_s]}g_{\kappa_s}(j-1)\right)^2\leqslant\sum_{s\in\mathcal{S}_t}m_s^2\leqslant M\sum_{s\in\mathcal{S}_t}m_s<M(t+M).$$

□

H. Incorporating Feature-Based Learning

In practical scenarios, auxiliary information is often available and can be leveraged to reformulate the learning problem within a structured feature space that is potentially low-dimensional. This transformation may enable faster learning rates that scale with the ambient feature dimension d rather than K, which is particularly beneficial when $d \ll K$. Below, we introduce a feature-based variant of our model and establish connections between its primitives and those of the original (feature-free) model.

Each type $\kappa \in [K]$ is associated with a (static) feature vector $x_\kappa \in \mathbb{R}^d$, satisfying $\|x_\kappa\|_2 \leqslant L$. The mean reward μ_κ for type κ is given by

$$\mu_\kappa := \langle x_\kappa, \theta^* \rangle, \tag{35}$$

where $\theta^* \in \Theta_0 \subseteq \mathbb{R}^d$ is a *latent* parameter satisfying $\|\theta^*\|_2 \leqslant S$. For ease of technical exposition, we assume a linear model; however, our analytical approach and techniques can be extended to more complex models, such as generalized linear models (McCullagh and Nelder 1989).

For clarity, we revisit the sequential learning setting under the feature-rich model below. However, this follows directly from our original model (1) under the mapping (35).

Generative Model. The reward r_t in round t is realized according to the model:

$$r_t = \xi_t + \begin{cases} \left\langle x_{\texttt{type}(\pi_t)}, \theta^* \right\rangle - c_{\texttt{type}(\pi_t)} & \text{if } \pi_t \notin \mathcal{A}_{t-1} \\ \left\langle x_{\texttt{type}(\pi_t)}, \theta^* \right\rangle g_{\texttt{type}(\pi_t)}(\texttt{age}_t(\pi_t)) & \text{if } \pi_t \in \mathcal{A}_{t-1} \text{ and } \texttt{age}_t(\pi_t) \leqslant M-1 \\ -\infty & \text{otherwise.} \end{cases}, \tag{36}$$

where ξ_t is a mean-zero noise term satisfying Assumption 1, $\mathcal{A}_{t-1} \subset \mathbb{N}$ denotes the set of arms pulled up to and including round $t-1$ (with $\mathcal{A}_0 := \varnothing$), and π_t denotes the arm pulled in round t.

Objective. The goal is to maximize the cumulative expected reward:

$$\sup_{\pi \in \Pi} \mathbb{E}_\pi \left[\sum_{t \in [T]} r_t \right], \tag{37}$$

where the expectation is taken over all sources of randomness in the problem, including the rewards and the policy. This is equivalent to minimizing the expected regret:

$$\inf_{\pi \in \Pi} \left(\texttt{OPT}_T - \mathbb{E}_\pi \left[\sum_{t \in [T]} r_t \right] \right),$$

where $\texttt{OPT}_T$ denotes the optimal value of (37) under ex ante knowledge of θ^*.

Remark 2 (Near-Optimal Policy) *It is instructive to note that the value* $\texttt{OPT}_T$ *remains unchanged in both the feature-rich and feature-less models and is attained (up to an* $\mathcal{O}(1)$ *additive loss) by the age-triggered policy* π^* *described in Theorem 1. Consequently, the online learning algorithm we propose next for the feature-rich model also aims to learn* π^* *from data.*

Learning Framework. We adapt the standard framework for learning in linear bandits (see, e.g., Auer (2002), Dani et al. (2008), Rusmevichientong and Tsitsiklis (2010), Chu et al. (2011), Abbasi-Yadkori et al. (2011)) to our setting, which we describe next.

In what follows, $\mathcal{S}_t \subseteq [t]$ denotes the subset of rounds up to t in which the UCB rule applies (see Algorithm 3 below), $\kappa_s := \texttt{type}(\pi_s)$ denotes the type of the arm pulled in round $s \in \mathcal{S}_t$, and $\mathcal{I}$ denotes the identity matrix in $\mathbb{R}^{d\times d}$. We define for $(\kappa, m, t, \theta) \in [K] \times [M] \times [T] \times \Theta_0$ the following:

$$V_t := \lambda \mathcal{I} + \sum_{s\in[\mathcal{S}_t]} x_{\kappa_s} \otimes x_{\kappa_s} \ . \tag{38}$$

$$\hat{\theta}_t := V_t^{-1} \sum_{s\in[\mathcal{S}_t]} r_s x_{\kappa_s} \ . \tag{39}$$

$$\bar{\mu}_{\kappa,m}(\theta) := \left(\frac{\langle x_\kappa, \theta \rangle \sum_{j\in[m]} g_\kappa(j-1) - c_\kappa}{m} \right) . \tag{40}$$

$$\texttt{rad}_{\kappa,m,t} := \alpha_t \|x_\kappa\|_{V_t^{-1}} \left(\frac{\sum_{j\in[m]} g_\kappa(j-1)}{m} \right) , \tag{41}$$

with the convention that $V_0 := \lambda\mathcal{I}$, $\texttt{rad}_{\kappa,m,0} := \alpha_0 \|x_\kappa\|_{V_0^{-1}} \left(\frac{\sum_{j\in[m]} g_\kappa(j-1)}{m} \right)$, and $\hat{\theta}_0 := \theta_0$, where θ_0 is some known parameter in the parameter set Θ_0. Note that $\bar{\mu}_{\kappa,m}(\theta)$ represents the long-run-average value of an age-triggered policy parameterized by type κ and cycle length m when the underlying model is θ. Algorithm 3 aims to optimize this value over the space of types $[K]$ and cycle lengths $[M]$ using an ℓ^2-regularized least-squares estimator (RLSE), $\hat{\theta}_t$, for the true latent model θ^*, defined with respect to a λ-regularized covariance matrix V_t. Lastly, $\texttt{rad}_{\kappa,m,t}$ denotes the confidence radius of $\hat{\theta}_t$ along the direction of x_κ in the elliptical norm induced by V_t^{-1}, further modulated by the exploration coefficient α_t.

Algorithm 3 Decay-Aware Feature-Based Optimistic Nudging

- **Input:** Types $[K]$, Max life M, Feature vectors $\{x_\kappa : \kappa \in [K]\}$, Parameter set Θ_0,
  Decay coefficients $\{g_\kappa(j-1) : (\kappa, j) \in [K] \times [M]\}$, Costs $\{c_\kappa : \kappa \in [K]\}$, Horizon T,
  Regularization param λ, Exploration sequence $(\alpha_0, \alpha_1, \cdots)$, Initial estimate $\theta_0 \in \Theta_0$.
- **Initialize:** Round $t = 1$, RLSE $\hat{\theta}_0 = \theta_0$.
- **While** $t \leqslant T$
  Fix $(\kappa_t, m_t) \in \arg\max_{(\kappa,m)\in[K]\times[M]} \left\{ \bar{\mu}_{\kappa,m}\left(\hat{\theta}_{t-1}\right) + \texttt{rad}_{\kappa,m,t-1} \right\}$. (`UCB rule`)
  Select a new arm of type κ_t; pull it for m_t consecutive rounds.
  Observe feedback and update RLSE $\hat{\theta}_{t+m_t-1}$.
  $t \leftarrow t + m_t$.

Dynamics. Algorithm 3 is an adaptation of the celebrated LinUCB algorithm (Chu et al. 2011) to our setting (additionally incorporating regularization to address the initial cold-start problem).

The algorithm operates episodically: at the end of each episode, the RLSE of θ^* is updated based on the observed feedback. It then commits to an arm type and duration for the next episode based on the UCB rule, which is computed using optimistic estimates of the target long-run-average value. This ensures progressively improving performance over time as the RLSE converges and the algorithm approximates π^* more closely. The precise regret rates achieved depend on the specifics of the estimation metrics described in (38), (39), and (41).

Remark 3 (Data Requirements for RLSE Computation) *Note that only the feedback from samples of age* 0 *is used in computing the RLSE. While incorporating samples of higher age is possible, it complicates the theoretical analysis and generally does not lead to meaningful improvements in the regret upper bound, as noted in §5.2.1.*

Remark 4 (Unknown Decay Kernels) *A version of Algorithm 3 for the case of unknown decay kernels (à la §5.2.2) can also be designed and analyzed in a principled manner. For brevity, the details of this extension are deferred to §I.*

Theorem 6 (High Probability Regret Bounds for Algorithm 3) *Fix failure probability* $\delta \in (0,1]$ *and set* $\lambda \geqslant \max(1, L^2)$. *For* $t = 0, 1, \cdots$, *specify the exploration sequence as follows:*

$$\alpha_t = \sigma \sqrt{d \log\left(\frac{1 + tL^2/\lambda}{\delta}\right)} + \sqrt{\lambda} S \ . \tag{42}$$

For any t, *let* $\mathcal{S}_t \subseteq [t]$ *denote the subset of rounds until* t *in which the UCB rule applies. Then, the regret of Algorithm 3 satisfies the following w.p. at least* $1 - \delta$*:*

$$\mathfrak{R}(t) \leqslant \max_{\kappa \in [K]} c_\kappa + 2\alpha_t \sqrt{2d \log\left(1 + \frac{L^2 t}{\lambda d}\right) \left(\sum_{s \in \mathcal{S}_t} \left(\sum_{j \in [m_s]} g_{\kappa_s}(j-1) \right)^2 \right)} \quad \forall\ t = 0, 1, \cdots \tag{43}$$

Furthermore, both (44) *and* (45) *also hold w.p. at least* $1 - \delta$*:*

$$\mathfrak{R}(t) \leqslant \max_{\kappa \in [K]} c_\kappa + 2 \max_{\kappa \in [K]} \left(\sum_{j \in [M]} g_\kappa(j-1) \right) \alpha_t \sqrt{2dt \log\left(1 + \frac{L^2 t}{\lambda d}\right)} \quad \forall\ t = 0, 1, \cdots \tag{44}$$

$$\mathfrak{R}(t) \leqslant \max_{\kappa \in [K]} c_\kappa + 2\alpha_t \sqrt{2dM(t+M) \log\left(1 + \frac{L^2 t}{\lambda d}\right)} \quad \forall\ t = 0, 1, \cdots \tag{45}$$

Details of the analysis are deferred to §G. For a horizon of length T, one can set $\delta = 1/T$ to obtain $\mathbb{E}[\mathfrak{R}(t)] = \tilde{\mathcal{O}}\big(d\sqrt{t}\big)$ for all times $t \in [T]$. Note that the upper bounds scale as $\tilde{\mathcal{O}}\big(d\sqrt{t}\big)$, compared to $\tilde{\mathcal{O}}\Big(\sqrt{Kt}\Big)$ in the feature-less setting. Thus, Algorithm 3 can offer significant performance benefits when $d \ll K$. However, this potentially comes at the expense of computational complexity, as the

implementation of the UCB rule now involves a rank-d matrix inversion, which can be computationally prohibitive. To mitigate this, one can compute the inverse efficiently using rank-1 updates at the end of each episode without sacrificing performance.

Remark 5 (Comparison of Upper Bounds) *Note that both* (44) *and* (45) *follow from* (43)*. Depending on the specifics of the decay kernels, one or the other may be tighter and therefore preferred. The bound in* (44) *is tighter when the decays are "sharp," and may potentially be independent of M altogether.*

Remark 6 (Improved Rates) *Note that our regret upper bounds are a factor of $\sqrt{d}$ away from the optimal $\tilde{\mathcal{O}}(\sqrt{dt})$ rate for the linear K-armed bandit problem (Chu et al. 2011). This gap can be closed by employing more sophisticated learning algorithms, such as the one proposed in the cited reference. However, we do not pursue these extensions in this work for brevity.*

I. Incorporating Feature-Based Learning: General Model

As before, each type $\kappa \in [K]$ is associated with a (static) feature vector $x_\kappa \in \mathbb{R}^d$, satisfying $\|x_\kappa\|_2 \leqslant L$. For $m \in [M]$, the mean reward obtained by pulling a type κ arm of age $m-1$ is denoted by $\mu_{\kappa,m}$ and given by

$$\mu_{\kappa,m} := \langle x_\kappa, \theta^*_m \rangle \ , \tag{46}$$

where $\theta^*_m \in \Theta_0 \subseteq \mathbb{R}^d$ is a *latent* parameter satisfying $\|\theta^*_m\|_2 \leqslant S$. For each $\kappa \in [K]$, the M mean reward parameters satisfy $\mu_{\kappa,1} \geqslant \cdots \geqslant \mu_{\kappa,M}$. The M model parameters $\{\theta^*_1, \cdots, \theta^*_M\} \subset \Theta_0$ are ex ante unknown and will be estimated from online data. We will use $\bar{\theta}^*$ to denote the true parameter configuration $(\theta^*_1, \cdots, \theta^*_M)$, $\bar{\theta}$ to denote a generic parameter configuration $(\theta_1, \cdots, \theta_M) \in \Theta_0^M$, and $\hat{\bar{\theta}}_t$ to denote the configuration specified by the M parameter estimates $\left(\hat{\theta}_{1,t}, \cdots, \hat{\theta}_{M,t}\right)$ computed at the end of round $t \in [T]$, with the convention that $\hat{\bar{\theta}}_0 := \left(\hat{\theta}_{1,0}, \cdots, \hat{\theta}_{M,0}\right) := (\theta_{1,0}, \cdots, \theta_{M,0})$, where $\{\theta_{1,0}, \cdots, \theta_{M,0}\}$ are some known parameters in the parameter set Θ_0. Define for $\left(\kappa, m, \bar{\theta}\right) \in [K] \times [M] \times \Theta_0^M$ the following:

$$\bar{\mu}_{\kappa,m}\left(\bar{\theta}\right) := \frac{\sum_{j\in[m]} \mu_{\kappa,j} - c_\kappa}{m} = \frac{\sum_{j\in[m]} \left\langle x_\kappa, \theta^*_j \right\rangle - c_\kappa}{m} \ . \tag{47}$$

For clarity, we revisit the sequential learning setting under the feature-rich model below. However, this follows directly from our original model (1) under the mapping (46).

Generative Model. The reward r_t in round t is realized according to the model:

$$r_t = \xi_t + \begin{cases} \left\langle x_{\mathtt{type}(\pi_t)}, \theta^*_{\mathsf{age}_t(\pi_t)+1} \right\rangle - c_{\mathtt{type}(\pi_t)} & \text{if } \pi_t \notin \mathcal{A}_{t-1} \\ \left\langle x_{\mathtt{type}(\pi_t)}, \theta^*_{\mathsf{age}_t(\pi_t)+1} \right\rangle g_{\mathtt{type}(\pi_t)}(\mathsf{age}_t(\pi_t)) & \text{if } \pi_t \in \mathcal{A}_{t-1} \text{ and } \mathsf{age}_t(\pi_t) \leqslant M-1 \\ -\infty & \text{otherwise.} \end{cases} \ , \tag{48}$$

where ξ_t is a mean-zero noise term satisfying Assumption 1, $\mathcal{A}_{t-1} \subset \mathbb{N}$ denotes the set of arms pulled up to and including round $t-1$ (with $\mathcal{A}_0 := \varnothing$), and π_t denotes the arm pulled in round t.

Objective. The goal is to maximize the cumulative expected reward:

$$\sup_{\pi \in \Pi} \mathbb{E}_\pi \left[\sum_{t \in [T]} r_t \right], \tag{49}$$

where the expectation is taken over all sources of randomness in the problem, including the rewards and the policy. This is equivalent to minimizing the expected regret:

$$\inf_{\pi \in \Pi} \left(\texttt{OPT}_T - \mathbb{E}_\pi \left[\sum_{t \in [T]} r_t \right] \right),$$

where $\texttt{OPT}_T$ denotes the optimal value of (49) under ex ante knowledge of $\bar{\theta}^*$.

Remark 7 (Near-Optimal Policy) *It is instructive to note that the value $\texttt{OPT}_T$ remains unchanged in both the feature-rich and feature-less models and is attained (up to an $\mathcal{O}(1)$ additive loss) by the age-triggered policy π^* described in Theorem 1. Consequently, the online learning algorithm we propose next for the feature-rich model also aims to learn π^* from data.*

The parameters (κ^*, m^*) of the age-triggered policy π^* satisfy

$$(\kappa^*, m^*) \in \underset{(\kappa,m) \in [K] \times [M]}{\arg\max} \bar{\mu}_{\kappa,m}\big(\bar{\theta}^*\big) \ . \tag{50}$$

Learning Framework. In what follows, $\mathcal{S}_{m,t} \subseteq [t]$ denotes the subset of rounds up to t in which an arm of age $m-1$ is pulled, $\kappa_s := \texttt{type}(\pi_s)$ denotes the type of the arm pulled in round s, and $\mathcal{I}$ denotes the identity matrix in $\mathbb{R}^{d \times d}$. We define for $(\kappa, m, t) \in [K] \times [M] \times [T]$ the following:

$$V_{m,t} := \lambda \mathcal{I} + \sum_{s \in \left[\mathcal{S}_{m,t}\right]} x_{\kappa_s} \otimes x_{\kappa_s} \ . \tag{51}$$

$$\hat{\theta}_{m,t} := V_{m,t}^{-1} \sum_{s \in \left[\mathcal{S}_{m,t}\right]} r_s x_{\kappa_s} \ . \tag{52}$$

$$\texttt{rad}_{\kappa,m,t} := \alpha_t \left(\frac{\sum_{j \in [m]} \left\| x_\kappa \right\|_{V_{j,t}^{-1}}}{m} \right) \ , \tag{53}$$

with the convention that $V_{m,0} := \lambda \mathcal{I}$, $\texttt{rad}_{\kappa,m,0} := \alpha_0 \left(\frac{\sum_{j \in [m]} \left\| x_\kappa \right\|_{V_{j,0}^{-1}}}{m} \right)$, and $\hat{\theta}_{m,0} := \theta_{m,0}$, where $\theta_{m,0}$ is some known parameter in the parameter set Θ_0.

Algorithm 4 Decay-Agnostic Feature-Based Optimistic Nudging

- **Input:** Types $[K]$, Max life M, Feature vectors $\{x_\kappa : \kappa \in [K]\}$, Costs $\{c_\kappa : \kappa \in [K]\}$, Horizon T
  Regularization param λ, Exploration sequence $(\alpha_0, \alpha_1, \cdots)$,
  Parameter set Θ_0, Initial estimate $(\theta_{1,0}, \cdots, \theta_{M,0}) \in \Theta_0^M$.
- **Initialize:** Round $t = 1$, RLSE $\hat{\bar{\theta}}_0 = (\theta_{1,0}, \cdots, \theta_{M,0})$.
- **While** $t \leqslant T$
  Fix $(\kappa_t, m_t) \in \arg\max_{(\kappa,m)\in[K]\times[M]} \left\{ \bar{\mu}_{\kappa,m}\left(\hat{\bar{\theta}}_{t-1}\right) + \texttt{rad}_{\kappa,m,t-1} \right\}$. (`UCB rule`)
  Select a new arm of type κ_t; pull it for m_t consecutive rounds.
  Observe feedback and update RLSE $\hat{\bar{\theta}}_{t+m_t-1}$.
  $t \leftarrow t + m_t$.

Theorem 7 (High Probability Regret Bound for Algorithm 4) *Fix failure probability $\delta \in (0, 1/M]$ and set $\lambda \geqslant \max(1, L^2)$. For $t = 0, 1, \cdots$, specify the exploration sequence as follows:*

$$\alpha_t = \sigma \sqrt{d \log\left(\frac{1 + tL^2/\lambda}{\delta}\right)} + \sqrt{\lambda} S \ . \tag{54}$$

Then, the regret of Algorithm 4 satisfies the following w.p. at least $1 - M\delta$:

$$\mathfrak{R}(t) \leqslant \max_{\kappa\in[K]} c_\kappa + 2\alpha_t \sqrt{2d(t+M)M \log\left(1 + \frac{L^2(t+M)}{\lambda d}\right)} \quad \forall\ t = 0, 1, \cdots$$

We observe that the upper bound is $\tilde{\mathcal{O}}\left(d\sqrt{Mt}\right)$. A factor of $\sqrt{d}$ can be further reduced using more sophisticated learning algorithms, such as those following the approach of Chu et al. (2011). This could potentially yield a bound of $\tilde{\mathcal{O}}\left(\sqrt{dMt}\right)$, exhibiting a tight dependence on the actual number of unknown scalar parameters (dM). However, we do not pursue such extensions in the present paper for brevity.

I.1. Regret Analysis of Algorithm 4

First, we adapt the following key lemmas from Abbasi-Yadkori et al. (2011) to our setting.

Lemma 10 (High Probability Confidence Ellipsoids) *Fix failure probability $\delta \in (0, 1/M]$, and for $m \in [M]$ and $t = 1, 2, \cdots$, consider the sequence of confidence ellipsoids generated by Algorithm 3:*

$$\Theta_{m,t} := \left\{ \theta \in \mathbb{R}^d : \left\| \hat{\theta}_{m,t} - \theta \right\|_{V_{m,t}} \leqslant \alpha_t \right\} , \tag{55}$$

where $V_{m,t}$, $\hat{\theta}_{m,t}$, and α_t are as defined in (51), (52), and (54) respectively. Define "clean event"

$$E := \{\theta^*_m \in \Theta_{m,t} \ \forall\ m \in [M] \ \forall\ t = 0, 1, 2, \cdots\}. \tag{56}$$

Then, one has that $\mathbb{P}(E) \geqslant 1 - M\delta$.

Proof of Lemma 10. The result follows instructively from Theorem 2 of Abbasi-Yadkori et al. (2011). We skip the steps for brevity. □

Lemma 11 (Elliptical Potential Lemma) *In the setting of Lemma 10, if* $\lambda \geqslant \max(1, L^2)$*, then the following holds for all* $m \in [M]$ *and* $t = 1, 2, \cdots$*:*

$$\sum_{s \in \mathcal{S}_{m,t}} \|x_{\kappa_s}\|^2_{V^{-1}_{m,s-1}} \leqslant 2d \log\left(1 + \frac{L^2 |\mathcal{S}_{m,t}|}{\lambda d}\right).$$

Proof of Lemma 11. The result follows instructively by combining Lemma 10 and 11 of Abbasi-Yadkori et al. (2011). We skip the steps for brevity. □

I.1.1 Proof of Theorem 7

Suppose the "clean event" E in (56) is true. Recall that $(\kappa^*, m^*) \in \arg\max_{(\kappa,m)\in[K]\times[M]} \bar{\mu}_{\kappa,m}(\bar{\theta}^*)$, and that $\mathcal{S}_{1,t} \subseteq [t]$ denotes the subset of rounds until t in which the UCB rule applies (see Algorithm 4). Then, similarly to the derivation of (26) in the analysis of Algorithm 1, one can show that the regret of Algorithm 4 after any t rounds is bounded as

$$\begin{aligned}
\mathfrak{R}(t) &\leqslant \max_{\kappa\in[K]} c_\kappa + \sum_{s\in\mathcal{S}_{1,t}} \left(\bar{\mu}_{\kappa^*,m^*}(\bar{\theta}^*) - \bar{\mu}_{\kappa_s,m_s}(\bar{\theta}^*)\right) m_s \\
&= \max_{\kappa\in[K]} c_\kappa + \sum_{s\in\mathcal{S}_{1,t}} \left(\bar{\mu}_{\kappa^*,m^*}\left(\hat{\bar{\theta}}_{s-1}\right) - \bar{\mu}_{\kappa_s,m_s}\left(\hat{\bar{\theta}}_{s-1}\right)\right) m_s + \sum_{s\in\mathcal{S}_{1,t}} \left(\bar{\mu}_{\kappa^*,m^*}(\bar{\theta}^*) - \bar{\mu}_{\kappa^*,m^*}\left(\hat{\bar{\theta}}_{s-1}\right)\right) m_s \\
&\quad + \sum_{s\in\mathcal{S}_{1,t}} \left(\bar{\mu}_{\kappa_s,m_s}\left(\hat{\bar{\theta}}_{s-1}\right) - \bar{\mu}_{\kappa_s,m_s}(\bar{\theta}^*)\right) m_s \\
&\underset{\dagger}{\leqslant} \max_{\kappa\in[K]} c_\kappa + \sum_{s\in\mathcal{S}_{1,t}} (\mathtt{rad}_{\kappa_s,m_s,s-1} - \mathtt{rad}_{\kappa^*,m^*,s-1}) m_s + \frac{1}{m^*} \sum_{s\in\mathcal{S}_{1,t}} \sum_{j\in[m^*]} \left\langle x_{\kappa^*}, \theta^*_j - \hat{\theta}_{j,s-1} \right\rangle m_s \\
&\quad + \sum_{s\in\mathcal{S}_{1,t}} \sum_{j\in[m_s]} \left\langle x_{\kappa_s}, \hat{\theta}_{j,s-1} - \theta^*_j \right\rangle, \qquad (57)
\end{aligned}$$

where (†) follows since (κ_s, m_s) is chosen by the UCB rule in round $s \in \mathcal{S}_{1,t}$. Now observe that for any round $s \in \mathcal{S}_{1,t}$, $j \in [m_s]$, and any $\kappa \in [K]$,

$$\begin{aligned}
\left|\left\langle x_\kappa, \hat{\theta}_{j,s-1} - \theta^*_j \right\rangle\right| &\underset{\ddagger}{\leqslant} \|x_\kappa\|_{V^{-1}_{j,s-1}} \left\|\hat{\theta}_{j,s-1} - \theta^*_j\right\|_{V_{j,s-1}} \underset{*}{\leqslant} \|x_\kappa\|_{V^{-1}_{j,s-1}} \left(\sigma\sqrt{d\log\left(\frac{1+(s-1)L^2/\lambda}{\delta}\right)} + \sqrt{\lambda}S\right) \\
&\underset{\$}{=} \alpha_{s-1}\|x_\kappa\|_{V^{-1}_{j,s-1}}, \qquad (58)
\end{aligned}$$

where (‡) follows from the Cauchy-Schwarz inequality, and (∗) follows from (55) since $\theta^*_j \in \Theta_{j,s-1}$ (the clean event E is true, by assumption), and ($) follows from (54). Combining (57) and (58),

$$\mathfrak{R}(t) \leqslant \max_{\kappa\in[K]} c_\kappa + \sum_{s\in\mathcal{S}_{1,t}} (\mathtt{rad}_{\kappa_s,m_s,s-1} - \mathtt{rad}_{\kappa^*,m^*,s-1}) m_s + \frac{1}{m^*} \sum_{s\in\mathcal{S}_{1,t}} \sum_{j\in[m^*]} \alpha_{s-1}\|x_{\kappa^*}\|_{V^{-1}_{j,s-1}} m_s$$

$$
\begin{aligned}
&\quad + \sum_{s\in\mathcal{S}_{1,t}} \sum_{j\in[m_s]} \alpha_{s-1} \|x_{\kappa_s}\|_{V_{j,s-1}^{-1}} \\
&\underset{\star}{=} \max_{\kappa\in[K]} c_\kappa + 2 \sum_{s\in\mathcal{S}_{1,t}} \sum_{j\in[m_s]} \alpha_{s-1} \|x_{\kappa_s}\|_{V_{j,s-1}^{-1}} \\
&\leqslant \max_{\kappa\in[K]} c_\kappa + 2\alpha_t \sum_{s\in\mathcal{S}_{1,t}} \sum_{j\in[m_s]} \|x_{\kappa_s}\|_{V_{j,s-1}^{-1}} \\
&\underset{\dagger}{\leqslant} \max_{\kappa\in[K]} c_\kappa + 2\alpha_t \sqrt{t^* \sum_{s\in\mathcal{S}_{1,t}} \sum_{j\in[m_s]} \|x_{\kappa_s}\|^2_{V_{j,s-1}^{-1}}} \\
&= \max_{\kappa\in[K]} c_\kappa + 2\alpha_t \sqrt{t^* \sum_{m\in[M]} \sum_{s\in\mathcal{S}_{m,t^*}} \|x_{\kappa_s}\|^2_{V_{m,s-1}^{-1}}} \\
&\underset{\ddagger}{\leqslant} \max_{\kappa\in[K]} c_\kappa + 2\alpha_t \sqrt{2dt^* \sum_{m\in[M]} \log\left(1 + \frac{L^2|\mathcal{S}_{m,t^*}|}{\lambda d}\right)} \\
&\leqslant \max_{\kappa\in[K]} c_\kappa + 2\alpha_t \sqrt{2dt^* M \log\left(1 + \frac{L^2 t^*}{\lambda d}\right)} \\
&\underset{\$}{\leqslant} \max_{\kappa\in[K]} c_\kappa + 2\alpha_t \sqrt{2d(t+M)M \log\left(1 + \frac{L^2(t+M)}{\lambda d}\right)}\,,
\end{aligned}
$$

where $(\star)$ follows using (53), $(\dagger)$ follows with $t^* := \sum_{s\in\mathcal{S}_{1,t}} m_s$ using Jensen's inequality, $(\ddagger)$ follows using Lemma 11, and finally $(\$)$ follows since $t^* < t + M$. □

J. Additional Details on Benchmark Algorithms and Additional Numerical Results

In this appendix, we first provide additional details on the benchmark UCB (Algorithm 5) and TS (Algorithm 6) algorithms. Then, we also present results from additional numerical experiments from §6 of the paper. Note that we only present benchmark algorithms for the decay aware setting and skip the decay agnostic case for the sake of brevity.

Additional details on the benchmark algorithms: In comparison to the proposed DRAWON algorithm (Algorithm 1), the benchmark algorithms only differ in their arm selection/regeneration rule. Instead of estimating the Whittles Index, both the benchmark algorithms compute an estimate of the current reward of each arm, given its age and type. Then, a regeneration decision is made by a direct cost-to-reward comparison (see the **If** statement in both the algorithms).

Algorithm 5 Decay-Aware Benchmark UCB

- **Input:** Types $[K]$, Max life M, sub-Gaussian param σ, Exploration coefficient α, Horizon T, Decay coefficients $\{g_\kappa(j-1) : (\kappa, j) \in [K] \times [M]\}$.
- **Initialize:** Round $t = 1$.
- **While** $1 \leqslant t \leqslant 2K$
  - Pull a *new* arm of type $1 + t\%K$.
  - $t \leftarrow t + 1$.
- Let $m_\kappa = 0,\ \forall \kappa \in [K]$.
- **While** $2K + 1 \leqslant t \leqslant T$
  - For each $k \in [K]$, compute $\hat{\mu}_\kappa^t$ according to the stratified estimator 7, with $\rho_{k,j} = 1/M$.
  - Let $\text{UCB}_k^t = \hat{\mu}_\kappa^t + 2\sigma\sqrt{\frac{\log T}{N_\kappa(t)}}$ where $N_\kappa(t)$ denotes the number of arm pulls of type κ.
  - Let $\kappa_t = \arg\max_{k \in [K]} \text{UCB}_k^t g_k(m_k)$.
  - **If**, $\text{UCB}_{\kappa_t}^t g_{\kappa_t}(m_{\kappa_t}) \leqslant \text{UCB}_{\kappa_t}^t g_{\kappa_t}(0) - c_{\kappa_t}$
    - Let $\kappa_t = \arg\max_{k \in [K]} \text{UCB}_k^t g_k(0) - c_k$.
    - Regenerate and pull arm κ_t; Also let $m_{\kappa_t} = 0$.
  - **Otherwise** pull arm κ_t without regeneration.
  - For each $k \in [K],\ m_k \to m_k + 1,\ t \to t + 1$.

Additional results from the numerical experiments: In Figure 10 we present additional numerical results from the model calibrated with real-world data from an EdTech platform. We find identical insights: the proposed DRAGON algorithm substantially outperforms benchmark algorithms regardless of how we generate the ground-truth model.

Algorithm 6 Decay-Aware Benchmark Thompson Sampling

- **Input:** Types $[K]$, max life M, sub-Gaussian param σ, horizon T, generation costs $\{c_\kappa\}_{\kappa\in[K]}$, decay coefficients $\{g_\kappa(j-1)\}$.
- **While** $1 \le t \le 2K$:

  Pull a *new* arm of type $(\,1+(t-1) \bmod K\,)$.

  $t \leftarrow t+1$.
- Let $m_\kappa = 0,\ \forall \kappa \in [K]$.
- **While** $2K+1 \le t \le T$:

  For each $\kappa \in [K]$, compute $\hat{\mu}_\kappa^t$ according to the stratified estimator 7, with $\rho_{k,j} = 1/M$.

  Sample $\theta_\kappa \sim \mathcal{N}\big(\widehat{\mu}_\kappa,\, \sigma^2/N_\kappa\big)$ where $N_\kappa(t)$ denotes the number of arm pulls of type κ.

  Compute $R_\kappa = \theta_\kappa \,\times\, g_\kappa\big(m_\kappa\big)$.

  $\kappa_t \leftarrow \arg\max_{\kappa\in[K]}\, R_\kappa$.

  If, $R_{\kappa_t} \leqslant R_{\kappa_t} g_{\kappa_t}(0) - c_{\kappa_t}$

  Let $\kappa_t = \arg\max_{k\in[K]} R_k g_k(0) - c_k$.

  Regenerate and pull arm κ_t; Also let $m_{\kappa_t} = 0$.

  Otherwise pull arm κ_t without regeneration.

  For each $k \in [K],\ m_k \to m_k + 1,\ t \to t+1$.

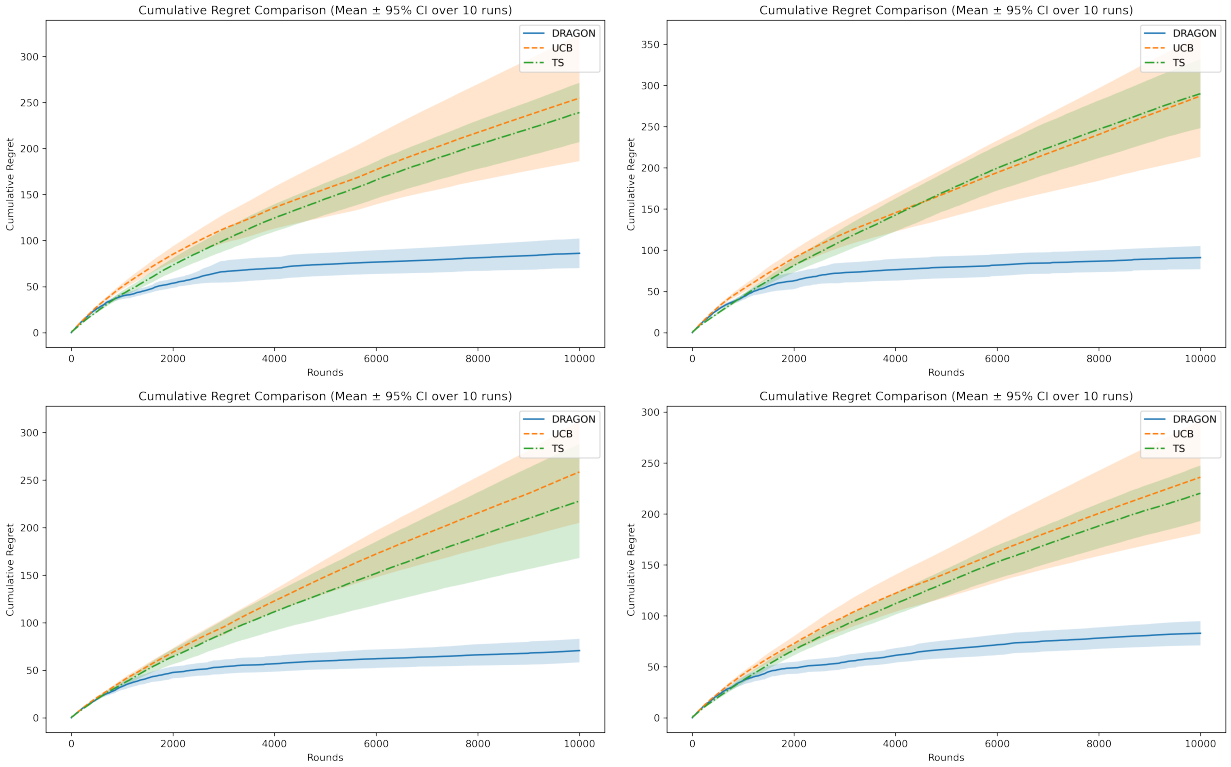

Figure 10 **Cumulative Regret comparison between the proposed DRAGON and the benchmark algorithms on the calibrated model. The top left (top right) plot demonstrates cumulative regret when the underlying ground-truth model is estimated using the exponential function (linear function) and data from both in-treatment and post-treatment weeks. Similarly, the bottom left (bottom right) plot demonstrates cumulative regret when the underlying ground-truth model is estimated using the exponential function (linear function) and data from only in-treatment weeks.**