Learning by Repetition: Stochastic Multi-armed Bandits under Priming Effect

Priyank Agrawal
University of Illinois at Urbana-Champaign
priyank4@illinois.edu

Theja Tulabandhula
University of Illinois at Chicago
theja@uic.edu

Abstract

We study the effect of persistence of engagement on learning in a stochastic multi-armed bandit setting. In advertising and recommendation systems, repetition effect includes a wear-in period, where the user's propensity to reward the platform via a click or purchase depends on how frequently they see the recommendation in the recent past. It also includes a counteracting wear-out period, where the user's propensity to respond positively is dampened if the recommendation was shown too many times recently. Priming effect can be naturally modelled as a temporal constraint on the strategy space, since the reward for the current action depends on historical actions taken by the platform. We provide novel algorithms that achieves sublinear regret in time and the relevant wear-in/wear-out parameters. The effect of priming on the regret upper bound is also additive, and we get back a guarantee that matches popular algorithms such as the UCB1 and Thompson sampling when there is no priming effect. Our work complements recent work on modeling time varying rewards, delays and corruptions in bandits, and extends the usage of rich behavior models in sequential decision making settings.

1 INTRODUCTION

In advertising applications and recommendation systems, there has been a large body of work that models consumer behavior [Hawkins et al., 2009, Solomon et al., 2014]. One such effect that is relatively well studied is the priming/repetition effect. Under the priming effect, an advertiser's payoff (for instance, click through rate) depends on how frequently they have presented the same ad to the same audience in the recent past. If the advertiser presents a specific ad sporadically, then the click through rate is much lower, even if this ad is the best among a collection of ads. Priming can be broken down into two sub-effects; *wear-in* and *wear-out* [Pechmann and Stewart, 1988]. Wear-in effect leads to a user not responding to an ad if it has not been shown enough number of times in the recent past. Whereas, the wear-out effect leads to a user not responding (or becoming insensitive) to an ad if it has been shown too many times in the recent past. Different ads may need different levels of repetition to obtain payoffs, and all the relevant parameters that model the priming effect may not be known a priori to the advertiser [Ma et al., 2016].

This phenomenon also translates to recommendations, such as for products and movies, where repeated display of item(s) can cause positive reinforcement to build over time, culminating in a conversion. It can also lead to fatigue and therefore no conversion. Since these conversion events depend on the past recommendations, they interfere with learning the true underlying (mean) payoffs and demands of different recommendations. Motivated by the above discussion, we define a new class of problems, which we call *bandit learning under priming effect*, to address the agent's need for repetitions. The amount by which the agent (say an ad platform or a recommendation system) needs to replay an arm depends on the degree of priming effect (see Section 2 for a formal treatment). In essence, the platform's current rewards are functions of its previous actions. A diagram illustrating this is shown in Figure 1. Our model and solution is one among a growing literature Kveton et al. [2015], den Boer and Keskin [2017], Shah et al. [2018], Tulabandhula and Wang [2020] that focuses on combining empirically validated behavioral models with sequential decision making.

While our setting can be addressed using Markov Decision Processes/reinforcement learning (RL) techniques, we choose to use the framework of multi-armed ban-
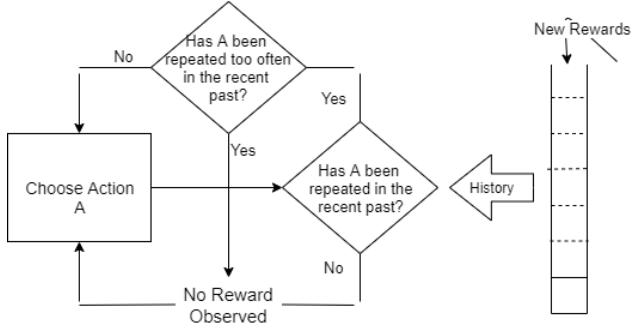
Figure 1: Bandit learning under priming effect.

dits (MAB) for their simplicity, analytical tractability and relatively tighter regret guarantees. The MAB problem, which is a special case of the RL problem, captures exploration-exploitation trade-off in certain sequential decision making settings. In a stochastic MAB set-up, reward distributions are associated with each arm of the set of arms $\mathcal{K}$. When the algorithm plays an arm, it immediately receives a reward with which it can learn, and also suffers a regret, which is the difference between the obtained reward versus the reward that it could have obtained, had it played the best arm in hindsight.

In the presence of priming effect, popular MAB algorithms such as UCB1 [Auer et al., 2002], SE [Even-Dar et al., 2006], MOSS [Audibert and Bubeck, 2009] or Thompson Sampling [Chapelle and Li, 2011] can be ineffective because they cannot directly control the number of times an arm is played in any given time window. For problem instances with non-unique optimal arms, the aforementioned algorithms may in fact switch between these very frequently (see Appendix A.5), potentially causing linear regret. Even when the mean rewards are not close, initial exploration will yield no rewards due to wear-in effects, hampering learning and hence, subsequent exploitation. To address these issues, we develop WI-UCB and WI/WO-UCB, which expand on the phase-based algorithmic template [Auer and Ortner, 2010] to learn in the presence of priming effect.

The consequence of priming effect considered here is closely related to the recent works in corruption [Lykouris et al., 2018] and delay [Pike-Burke et al., 2018] in the reward accrual process. Unlike these settings, priming effect is *endogenous*, and correlates past actions with the current reward. As an unifying view, in all three works one can assume that there is a intermediate function that allows an MAB algorithm to accrue some transformation of the current and past rewards instead of just the current reward. In our case, we accrue rewards that are modulated by the stochastic sequence of actions that our algorithm took previously, which makes learning

more challenging as rewards are now policy dependent.

For instance, Lykouris et al. [2018] consider settings with arbitrary exogenous corruptions of rewards, and propose a randomized algorithm that achieves smooth degradation as the corruption level increases. Unlike their setting, priming effect is *endogenous*, and correlates past actions with the current reward. While the amount of corruption due to priming in our setting would be $O(NK \log T)$ (N is a instance dependent parameter in our setting, K is the number of arms, and T is the horizon length) if our algorithm is used, we cannot reuse their analytical techniques because of endogeneity. Impact of delayed rewards on learning has been well-studied recently [Joulani et al., 2013, Perchet et al., 2016, Cesa-Bianchi et al., 2018]. In particular, Joulani et al. [2013] provide a recipe to use any regular MAB algorithm in this setting and show that delay causes an additive regret penalty. In [Cesa-Bianchi et al., 2018] for adversarial bandits and in [Pike-Burke et al., 2018] for stochastic bandits, the authors provide regret guarantees for a much weaker setting where rewards can get mixed up and may partially accrue over time. That is, components of rewards due to multiple previous actions may appear collectively at some future point. In contrast to the delay effect, which is affecting the future accrual of rewards, the priming/repetition effect can be viewed as being caused by the trajectory of past actions. As a result, we obtain very different regret bounds.

There has been recent parallel work on rotting bandits [Levine et al., 2017, Seznec et al., 2019], which can be thought of as capturing the wear-out effect via a sequence of reward random variables with decreasing means. While it does not capture the wear-in effect, the dependence of the reward means on the number of times the ad/arm is played is accounted from the start of the horizon rather than an immediate preceding window, which is critical in our applications. In a closely related subsequent work [Pike-Burke and Grunewalder, 2019], the authors make the mean reward of each arm an unknown function of the time since its last play. While this is an interesting structure, it does not again capture repetition effects (wear-in and wear-out) studied here, and the work relies on a very different modeling setup (using Gaussian processes) to obtain regret bounds.

Some prior works have studied bandit settings under rich temporal user behavior models. For instance, Shah et al. [2018] study a temporal behavioral effect rooted in microeconomic theory, namely self-reinforcement. In addition to delays and corruptions influencing rewards, works such as [Xu and Yun, 2018] and [Gamarnik et al., 2018] also consider the impact of reward accrual in the presence of limited memory, which affects learning and re-

gret. Finally, bandits with switching costs Banks and Sundaram [1994], Dekel et al. [2014] consider penalties for switching arms too often. As we will discuss soon, our algorithms are also candidate solutions to this problem setting by virtue of switching arms rarely. This is because priming effect imposes a hard constraint on switching arms too frequently and too infrequently, which is approximately equivalent to having large switching costs.

Our Results and Techniques: Owing to the nature of the priming effect, algorithms necessarily have to ensure that the arms still under consideration are played frequently, perhaps in batches. Phase based algorithms form a natural algorithmic template for such a mechanism. This family of algorithms date back to [Agrawal et al., 1988], who considered arm switching costs. Our algorithms, WI-UCB and WI/WO-UCB, follow this design pattern, wherein the focus is to eliminate arms between stages (for instance, in WI-UCB each arm is played consecutively for multiple rounds between stages). In particular, both WI-UCB (wear-in effect setting) and WI/WO-UCB (wear-in and wear-out effect setting) are based on algorithms such as UCB-REVISITED [Auer and Ortner, 2010] and UCB2 [Auer et al., 2002], and work under the setting when just the expected priming effect parameters are known. We also introduce a key new idea of *compound-arms* in WI/WO-UCB, which lets us retain the algorithmic structure described above as well as the corresponding analytical machinery to obtain regret bounds.

In our analysis, we design martingales on the sequence of the cumulative sums of the accrued reward deviations from their means. Following the techniques of [Pike-Burke et al., 2018] and [Auer and Ortner, 2010], we use foundational tools, such as the Bernstein inequality for martingales and the Azuma-Hoeffding inequality based Doob's optimal stopping theorem (see Sections 3 and 4), to bound the priming effect under a judicious choice of phase lengths, and guarantee fast convergence of the reward estimates with high probability. Our analysis deviates from these previous works in the following ways: (a) our reward random variables are functions of the past history and *policy* dependent, (b) the use of a phase-based strategy is only possible due to the notion of compound-arms (novel to this work and different from the dueling bandit literature [Yue et al., 2012]), and (c) their regret analysis is not directly applicable to our setting.

The key technical challenge in our setting is due to the opposing wear-in and wear-out effects: while wear-in requires frequent repetition to learn, wear-out hampers learning if arms are played too frequently. Assume that an algorithm can accrue a reward for pulling arm j at time t if it has been tried at least $D_{t,j}$ and at most

Algorithm	Bound
Lower bound [Lai and Robbins, 1985]	$O(\sqrt{KT})$
UCB1 [Auer et al., 2002]	$O(\sqrt{KT \log T})$
MOSS [Audibert and Bubeck, 2009]	$O(\sqrt{KT})$
WI-UCB [*this work*, no priming]	$O(\sqrt{KT \log T})$
WI-UCB [*this work*, wear-in only]	$O\left(\sqrt{KT \log T} + K\sqrt{\log^2 T \mathbb{E}[D]}\right)$
WI/WO-UCB [*this work*, wear-in & wear-out]	$O\left(K\sqrt{T \log T} + K^2\sqrt{\log^2 TN\mathbb{E}[D]}\right)$

Table 1: Results summary.

$Z_{t,j}$ times in the past N rounds, where $D_{t,j}$ and $Z_{t,j}$ are unobserved random variables with means $\mathbb{E}[D_j]$ and $\mathbb{E}[Z_j]$, and N is a known instance-specific fixed positive integer. Then, the regret upper bounds of our algorithms, shown in Table 1, depend *sublinearly* and *additively* on the priming effect parameters (for simplicity assume $\mathbb{E}[D_j] = \mathbb{E}[D]$ for all j). To our knowledge, this is the first work that takes into account both wear-in and wear-out effects for advertising and recommendation systems in an online learning scenario.

2 PROBLEM DEFINITION

There are $K > 1$ arms in the set $\mathcal{K}$ available to the platform (agent/learner), each corresponding to an ad/recommendation. An arm is played by the platform in each of the $T > 1$ rounds of interaction with the user (environment).

Priming Effect: Each arm $j \in \mathcal{K}$ is associated with a reward distribution ξ_j, which has support in $[0, 1]$. The mean reward for arm j is μ_j. μ^* is the maximum of all μ_j and corresponds to the arm j^*. Let $N \in \mathbb{N}$ denote the number of historical rounds, and let ξ_j^D and ξ_j^Z for each $j \in \mathcal{K}$, with supports in $\{0, ..., a\}$ and $\{b, ..., N\}$ respectively, parameterize the priming effect (wear-in and wear-out). In particular, the distributions ξ_j^D and ξ_j^Z (where $0 \le a < b \le N$ are fixed non-negative integers) are associated with the wear-in and wear-out effects respectively, and characterize the stochastic user who is unknown to the platform a priori. At each round t of interaction between the platform and the stochastic user, the following happens:

- The platform selects an arm (e.g., shows an ad) denoted by J_t (say, $J_t = j$).

- The user/environment generates a sample $R_{t,j} \sim \xi_j$, as well as samples from the wear-in and wear-out distributions: $D_{t,j} \sim \xi_j^D$ and $Z_{t,j} \sim \xi_j^Z$.

- The user returns the following derived sample reward $X_{t,j}$ via a click/purchase given by:

$$X_{t,j} = R_{t,j}\mathbb{I}\left[Z_{t,j} \ge f_{t,j}(N) \ge D_{t,j}\right], \quad (1)$$

which is the only quantity observed by the platform. Here $\mathbb{I}[]$ is the indicator function, and $f_{t,j}(N)$ is a *history function* that encapsulates contribution of prior user interactions, i.e., events and outcomes of rounds $\{t - N, ...t - 1\}$ that capture the priming effect.

We assume that the value of $f_{t,j}(N)$ only depends on the number of times arm j was played by the platform in the past N rounds and is independent of the other choices made, and is non-decreasing in N. We also assume that each play of arm j in the relevant history contributes equally to the value of $f_{t,j}(N)$. With these assumptions, we focus our analysis on the following reward accrual model:

$$X_{t,j} = R_{t,j}\mathbb{I}\left[Z_{t,j} \geq \left(\sum_{k=\max(t-N,0)}^{t} \mathbb{I}[J_k = j]\right) \geq D_{t,j}\right]. \quad (2)$$

Justifying the reward model: The reward model, as a product of $R_{t,j}$ and an indicator function of the history of arms played along with the wear-in/wear-out effects, is quite practical while being amenable to analysis, and can be viewed as a stepping stone for more realistic models in the future. In fact, when ξ_j are Bernoulli, as is the case with clicks and checkouts in applications such as e-commerce, there is very little loss in expressivity when using an indicator function as a multiplier versus any other continuous unimodal function of the history. One could imagine applications where a more sensitive function of history[1] (for example, $X_{t,j}$ gradually increases due to wear-in and then decreases due to wear-out) could be relevant, where such an indicator function (even as an approximation) may be oversimplified. In these cases, a different algorithmic approach and analysis will be needed, potentially relying on RL techniques (we are already capturing the impact of past actions on the current reward without relying on RL methodology here). Further note that the priming effect in Equation 1 entails the necessity of having non-overlapping support for the distributions $\{\xi_j^D\}$ and $\{\xi_j^D\}$ (i.e., $a < b$). If not, there will be problem instances where no rewards would be accrued for any policy.

The stochastic nature of user behavior: We are in a (non-contextual) stochastic bandit setting. As noted earlier, every time an arm j is presented, the user generates the three random variables (and using N and $f_{t,j}(N)$) responds by giving back a transformed reward $X_{t,j}$. In doing so, they are agnostic to the strategy of the platform. Making the priming effect stochastic in each round captures a natural time-varying behavior of users on such

---

[1]It is possible to extend our analysis to case when different intervals of the past N rounds have a weighted contribution in the definition of $f_{t,j}(N)$.
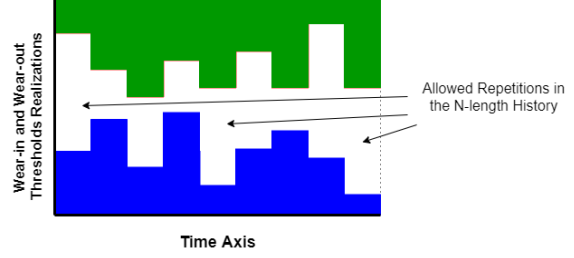

Figure 2: Model Illustration: At any round (x axis), if the cumulative number of times a chosen arm was played is in the white region, then a reward is obtained by the platform. If it is lower (wear-in, blue region) or higher (wear-out, green region), the platform does not get the reward. Each pair of blue and green bars is for the arm that was pulled in that time step.

advertising/recommendation platforms. It is important to distinguish between our setup and a setting where $D_{t,j}$ and $Z_{t,j}$ do not depend on time. The latter is a restricted stochastic setting where the priming effect is static across time ($R_{t,j}$ is the only randomness in the environment) and can result in potentially simpler learning strategies (such as playing each arm for $O(\mathbb{E}[D_j])$ rounds, and then estimating μ_is), although it is unclear if there will be any improvements in terms of the regret guarantees over ours.

Knowledge of $\mathbb{E}[D_j]$: We assume that the platform knows the first moment $\mathbb{E}[D_j]$, as well as N and T. This may seem limiting at first, but we argue that it is fairly benign: the platform needs no additional distributional knowledge to be able to achieve sub-linear regret matching the performance of stochastic MABs (modulo additive factors, see Section 3). Further knowing (just) this first moment is readily possible for a platform using observational data or from previous interactions with its users. For simplicity of exposition, we assume that $\xi_j^D = \xi^D$ and $\xi_j^Z = \xi^Z$ (and thus work with $\mathbb{E}[D]$ and $\mathbb{E}[Z]$ moving forward).

Goal: We want to design an online algorithm for the platform that plays a sequence of arms $\{J_t\}$ such that the expected (pseudo-)regret R_T of the algorithm when compared to a benchmark policy is sublinear in the time horizon T:

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} X_{t,\pi_t}\right] - \mathbb{E}\left[\sum_{t=1}^{T} X_{t,J_t}\right], \quad (3)$$

where $\pi = (\pi_1, ..., \pi_T)$ lies in the class of benchmark policies Π_B, and each π_t is a function that maps the history of arms played in the past and their rewards to arm choice in round t. Note that the rewards for the benchmark are censored according to that policy and not the policy the learner is playing. In the standard bandit set-

ting, it is standard to assume that π is a set of constant functions (i.e., $\pi_1 = ... = \pi_T$) that are determined by the mean rewards μ_j for $j \in \mathcal{K}$. Similarly, in our setting, we consider benchmark policies that are designed to take into account the priming effect. Intuitively, the benchmark policy of playing the best arm in hindsight may not be suitable anymore, especially when both wear-in and wear-out effects are present (Section 4). We defer further discussion on the choice of the benchmark policy to Sections 3 and 4.

3 WEAR-IN EFFECT

We start with those instances where wear-out effect is non-existent, i.e., $Z_{t,j} = N$ for all t and $j \in \mathcal{K}$. This simplifies the treatment, while allowing for the same analytical tools to be extended to the general setting in Section 4.

Methods such as UCB1 [Auer et al., 2002], MOSS [Audibert and Bubeck, 2009] or Thompson Sampling [Chapelle and Li, 2011] will not succeed in our problem setting for any reasonable values of $\mathbb{E}[D]$ (and $\mathbb{E}[Z]$). For instance, the UCB1 algorithm, which is based on the optimism principle, adaptively decreases its optimism over mean rewards such that arms are easily distinguishable from each other. But while doing this, it does not allow for a direct control on how arms switch between rounds. Even when only wear-in effect is present, is impact is minimal only when $\mathbb{E}[D] << N$, which is quite limiting. To control the switching between arms, one could employ algorithms such as SE [Even-Dar et al., 2006], UCB-REVISITED [Auer and Ortner, 2010] or UCB2 [Auer et al., 2002], which either play arms in a predictable round robin fashion or play the same arm consecutively. And this is precisely what we attempt to do here.

Our proposed strategy, WI-UCB (Algorithm 1) plays arms in phases (indexed by m). These phases are distinct and non-overlapping. The algorithm maintains a set of active arms, and in every phase, each arm from the active set is played repeatedly and consecutively. The algorithm also maintains a confidence bound $\tilde{\Delta}_m$ on the estimates of mean rewards in each phase. At the end of each phase, arms are eliminated based on confidence gaps computed using this arm-agnostic bound. We refer to the number of rounds that each active arm is played in a given phase as the *incremental phase length*. The total length of the phase is the sum of such individual incremental phase lengths. In particular, a sequence of (cumulative) phase lengths $\{n_m | m = 0, 1, 2, ...\}$ determine the number of rounds each active arm has been played by phase m (thus, the incremental phase length is $n_m - n_{m-1}$). And we denote the set of active arms in

---

Input: A set of arms $\mathcal{K}$, time horizon T, and phase length parameters $\{n_m | m = 0, 1, 2, ...\}$.
Initialization: Phase index $m = 1$, $\mathcal{K}_m = \mathcal{K}$, $\tilde{\Delta}_1 = 1$, $T_j(0) = \phi \;\; \forall j \in \mathcal{K}$, where ϕ is the empty set, and time index $t = 1$.
while $t \leq T$ **do**
   if $|\mathcal{K}_m| > 1$ **then**
      Play Arms:
      for *each active arm j in $\mathcal{K}_m$* **do**
         Set $T_j(m) = T_j(m-1)$.
         Play j for $n_m - n_{m-1}$ consecutive rounds and update $T_j(m)$.
         Accrue rewards according to the environment model Equation (2).
      end
      Eliminate Sub-optimal Arms:
      for *each active arm j in $\mathcal{K}_m$* **do**
         $\overline{X}_{m,j} = \frac{1}{|T_j(m)|} \sum_{s \in T_j(m)} X_{s,j}$.
      end
      Construct $\mathcal{K}_{m+1}$ by eliminating arms j in $\mathcal{K}_m$ for which:
      $\overline{X}_{m,j} + \tilde{\Delta}_m / 2 < \max_{j' \in \mathcal{K}_m} \overline{X}_{m,j'} - \tilde{\Delta}_m / 2$.
      Update the Confidence Bound:
      Set $\tilde{\Delta}_{m+1} = \frac{\tilde{\Delta}_m}{2}$.
      Increment phase index m by 1 and update t based on $\{n_m\}$ values up to the current phase.
   end
   Play the single arm in $\mathcal{K}_m$ and update t.
end

Algorithm 1: WI-UCB

---

phase m using the set $\mathcal{K}_m$.

Intuitively, longer incremental phase lengths help in negating the wear-in effect, whereas shorter incremental phase lengths help in negating the wear-out effect. These opposing consequences necessitate a different algorithm design when both the priming effect are present (see Section 4). In the current setting, switching the arms too often reduces the rewards accumulated due to the wear-in effect and impedes the algorithm from exploring as well as exploiting what it has learned so far. Thus, a careful design of phase length is necessary, which we discuss below. In Algorithm 1, we use $T_j(m)$ to refer to the collection of times when the jth arm is played up to phase m. Further, the estimated mean reward for arm j at the end of phase m is denoted by $\overline{X}_{m,j}$.

Benchmark Policy: To bound the regret defined in Equation (3) for this algorithm, we consider the benchmark policy π to be one which plays the arm with the highest mean reward at all rounds. It can be shown that

such a policy is optimal even with the wear-in effect (See Appendix A.4 for a proof).

There are two key aspects to bounding regret with respect to the aforementioned benchmark for Algorithm 1: (a) identifying an appropriate n_m that depends on the wear-in effect parameter $\mathbb{E}[D]$, and (b) showing that this n_m swiftly eliminates the sub-optimal arms. With these two steps addressed, we can get a regret guarantee such as below.

Theorem 1. *For any $\lambda > 0$, the expected (pseudo-)regret of* WI-UCB *(Algorithm 1) is bounded as:*

$$R_T \leq \sum_{i \in \mathcal{K}_1} \left(\Delta_i + \frac{64 \log(T)}{\Delta_i} + \frac{64 \log(T)}{3} \right.$$
$$\left. + 32 \sqrt{\log\left(\frac{4}{\Delta_i}\right) \mathbb{E}[D] \log(T)} \right)$$
$$+ \sum_{i \in \mathcal{K}_1} \frac{4 \Delta_i}{T} + \sum_{i \in \mathcal{K}_2} \frac{32}{T} + \max_{\{i \in \mathcal{K}_2 : \Delta_i < \lambda\}} \Delta_i T,$$

where $\mathcal{K}_1 = \{i \in \mathcal{K} | \Delta_i > \lambda\}$, $\mathcal{K}_2 = \{i \in \mathcal{K} | \Delta_i > 0\}$, and $\Delta_i = \mu^* - \mu_i$.

Given an appropriate choice for phase lengths n_m, the proof of the above theorem follows that in Auer and Ortner [2010], where a similar phased-based algorithm was suggested for the vanilla stochastic MAB. Following Theorem 1, we can also obtain a corresponding instance independent bound, as shown in the following corollary.

Corollary 1. *For all $T \geq K$, choosing $\lambda = \sqrt{\frac{K \log(T)}{T}}$ and using $\log(1/\tilde{\Delta}_m) \leq \log(T)$, the expected (pseudo-)regret of* WI-UCB *is* $O\left(\sqrt{KT \log T} + K\sqrt{\log^2 T \mathbb{E}[D]} \right)$.

A key point to note is that the *wear-in* effect parameter $\mathbb{E}[D]$ appears as an additive penalty. The leading term, $\sqrt{KT \log T}$ is only a logarithmic factor away from the best known bounds for the vanilla stochastic MAB [Audibert and Bubeck, 2009]. Thus our regret upper bound behaves gracefully with the level of the wear-in priming effect.

To define suitable phase length parameters $\{n_m\}$, we design martingales on the sequences of bias adjusted rewards sequences and eventually bound the growth of such martingales under our model. Appendix A.1 contains an overview of martingales, stopping times, and key concentration bounds required for the analysis. For a detailed discussion on martingale properties, one can refer Mitzenmacher and Upfal [2005]. For any active arm j and phase m, let $S_{m,j}$ denote the time in this phase when the algorithm starts playing this arm. Similarly let

$U_{m,j}$ denote the time in this phase when the algorithm stops playing this arm. Also, let $T_t(j, N)$ be a random variable that denotes the number of times that the arm j was played in the rounds $\{t - N, ...t - 1\}$. We define a filtration $\{\mathcal{G}_s\}_{s=0}^{\infty}$ by setting $\{\mathcal{G}_0\} = \{\Omega, \phi\}$ with Ω suitably defined, and letting $\{\mathcal{G}_t\}$ to be the σ-algebra over $(X_1....X_t, J_1....J_t, D_{1,J_1}....D_{t,J_t}, R_{1,J_1}...R_{t,J_t})$. In Lemma 1 below, we give a constructive proof for the choice of n_m such that the estimated mean reward for an arm j gets closer to its true mean at the end of phase, and we can use this property to eliminate sub-optimal arms quickly.

Lemma 1. *There exists a positive n_m for which the estimate $\overline{X}_{m,j}$ calculated by Algorithm 1 for an active arm j ($j \in \mathcal{K}_m$) and phase m, satisfies $\overline{X}_{m,j} - \mu_j \leq \tilde{\Delta}_m/2$ with probability at least $1 - \frac{2}{T^2}$.*

Below, we show how unlikely it is to grossly overestimate the mean value, assuming j is a sub-optimal arm.

Outline of the proof: We build on the observation that the cumulative sums of bias adjusted rewards ($\overline{X}_{m,j} - \mu_j$) can be decomposed into a couple of martingale sequences (see the first two terms in Equation 5). As the algorithm progresses, we show via Lemmas 3 and 5 that the growth of both these martingales can be bounded with high probability in terms of the phase length n_m and the wear-in effect parameter $\mathbb{E}[D]$. To start, it follows that for each arm j:

$$\sum_{i=1}^{m} \sum_{t=S_{i,j}}^{U_{i,j}} (X_{t,j} - \mu_j) \leq \sum_{i=1}^{m} \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,J_t} - \mu_j)$$
$$- \sum_{i=1}^{m} \sum_{t=S_{i,j}}^{U_{i,j}} R_{t,J_t} \mathbb{I}\{T_t(J_t, N) \leq D_{t,J_t}\}. \quad (4)$$

Since only one arm is played at a time, therefore, $J_t = j$ and $T_t(J_t, N) \leq t - S_{i,j}$ within a phase. Define $A_{i,t} := R_{t,J_t} \mathbb{I}\{t \leq S_{i,j} + d_{t,J_t}\}$ and $M_t := \sum_{i=0}^{m} A_{i,t} \mathbb{I}\{S_{i,j} \leq t \leq U_{i,j}\}$. We can upper bound Equation 4 and write it in terms of M_t as:

$$\sum_{i=1}^{m} \sum_{t=S_{i,j}}^{U_{i,j}} (X_t - \mu_j) \leq \sum_{i=1}^{m} \sum_{t=S_{i,j}}^{U_{i,j}} (R_{t,j} - \mu_j)$$
$$+ \sum_{t=1}^{U_{m,j}} (\mathbb{E}[M_t | G_{t-1}] - M_t) - \sum_{t=1}^{U_{m,j}} \mathbb{E}[M_t | G_{t-1}]. \quad (5)$$

Due to the above construction, we are able to succinctly separate the loss/modulation in rewards due to the *wear-in* effect. Next, we bound each term in Equation 5 individually. The first term is the deviations of the un-modulated rewards from their true means, and a reasonable upper bound is desired to get the right dependence

on T in the Corollary 1. Next, the second term captures the impact of the wear-in effect, and can be upper bounded using Lemmas 3 and 5. Finally, by taking trivial non-negative upper bound on the last term in Equation 5 above, and applying a simple union bound, we obtain the following expression for n_m that guarantees the claim made in the statement of Lemma 1 with probability at least $1 - \frac{2}{T^2}$:

$$n_m \leq 1 + \frac{4\log(T)}{\tilde{\Delta}_m^2} + \frac{16\log(T)}{3\tilde{\Delta}_m} + \frac{8\sqrt{m\mathbb{E}[D]\log(T)}}{\tilde{\Delta}_m}. \quad (6)$$

Below, we now discuss the supporting lemmas needed for Lemma 1. First, we start with the second term of Equation 5. The following lemma shows that $Y_s = \sum_{t=1}^{s}(\mathbb{E}[M_t|G_{t-1}] - M_t)$ forms a martingale. This fact is a prerequisite for Lemma 5.

Lemma 2. $Y_s := \sum_{t=1}^{s}(\mathbb{E}[M_t|G_{t-1}] - M_t)$ *for all* $s \geq 1$ *with* $Y_0 = 0$ *is a martingale with respect to the filtration* $\{G_s\}_{s=0}^{\infty}$ *with increments* $C_s = \mathbb{E}[M_s|G_{s-1}] - M_s$ *satisfying* $\mathbb{E}[C_s|G_{s-1}] = 0$ *and* $C_s \leq 1$ *for all* $s \geq 1$.

The next lemma is used to bound the sum of deviations of the received rewards from their mean values at the end of a phase (the first term in Equation 5). In effect, we show that the sum of accrued rewards will be close to their sum of means with high probability. This is a high probability guarantee that reasonably long sequences of rewards can give information about the mean reward parameters of the arms.

Lemma 3. *With probability at least* $1 - \frac{1}{T^2}$, $\sum_{i=1}^{m}\sum_{t=S_{i,j}}^{U_{m,j}}(R_{t,j} - \mu_j) \leq \sqrt{n_m\log(T)}$.

Outline of the proof: For an arm j that was played at time t, the quantity $R_{t,J_t} - \mu_j$ can be either positive or negative. And we are interested in bounding the maximum cumulative positive growth. One can interpret this bound to be the maximum expected value of a finite 1-dimensional random walk, where an agent is taking steps at n_m randomly chosen time instances in the given horizon T. The cumulative displacements of the agent because of the random walk form a martingale with an appropriately defined filtration. We use a modified version of the Azuma-Hoeffding inequality (see Lemma 9 in Appendix A.1) to upper bound this displacement with high probability.

Next, through the following lemma, we show that the wear-in effect martingale sequence (the second term in Equation 5 does not have any *sudden jumps* by upper bounding the corresponding quadratic variation process.

Lemma 4. *For any* t, *let* $P_t = \mathbb{E}[M_t|G_{t-1}] - M_t$. *It follows that:* $\sum_{t=1}^{U_{m,j}}\mathbb{E}[P_t^2|G_{t-1}] \leq m\mathbb{E}[D]$.

Outline of the proof: From the definition of P_t, it is easy to see that $\sum_{t=1}^{U_{m,j}}\mathbb{E}[P_t^2|G_{t-1}] \leq \sum_{t=1}^{U_{m,j}}\mathbb{E}[M_t^2|G_{t-1}]$.

Recall that M_t denotes the sum of m random variables of which only one is non-negative while others are zero, and this depends on the phase number the time index t belongs to. Intuitively, this contributes the factor m to bound claimed. In a phase i, for arm j, the reward is lost if $\mathbb{I}\{t < S_{i,j} + D_{t,j}\} = 1$. We are interested in sum of these indicator values when t varies from 1 to T. In expectation, each indicator is given by $\mathbb{P}(t < S_{i,j} + D_{t,j})$, the summation of which can be upper bounded by $\mathbb{E}[D]$.

Finally, we can now bound the second term in Equation 5, using the following lemma.

Lemma 5. *With probability at least* $1 - \frac{1}{T^2}$, $\sum_{t=1}^{U_{m,j}}(\mathbb{E}[M_t|G_{t-1}] - M_t) \leq \frac{2}{3}\log(T) + \sqrt{\frac{4\log^2(T)}{9} + 4m\mathbb{E}[D]\log(T)}$.

Outline of the proof: It follows from Lemma 2 that $Y_{U_{m,j}} = \sum_{t=1}^{U_{m,j}}(\mathbb{E}[M_t|G_{t-1}] - M_t)$ forms a martingale, and represents an upper bound on the missed rewards due the wear-in effect. Next, note that each reward is bounded. Further, we had claimed in Lemma 4 that there are no sudden jumps in the growth of the corresponding martingale sequences. Using Lemma 7(see Appendix A.1 for the statement) implies that $Y_{U_{m,j}}$ is bounded as well, which can then be used to show that the above inequality is true with high probability. Detailed proofs of Lemmas 1-5 and Theorem 1 are provided in Appendix A.2.

4 WEAR-IN AND WEAR-OUT EFFECTS

The general setting involving both wear-in and wear-out effects is significantly harder to tackle, primarily because of the opposing nature of these two. Qualitatively, wear-in effect necessitates continued exploitation (repetition of actions), while on the other hand, wear-out penalized continued exploitation. In WI-UCB, arms are consecutively repeated $n_m - n_{m-1}$ times (see Equation 6) in each phase m. The amount of repetition quickly supersedes N, which leads to zero reward accrual due to the wear-out effect for a majority of the rounds (see Equation 2). Hence, an appropriate algorithm for the general priming situation should be able to track the number of plays of each arm like WI-UCB, potentially in a phased manner, but also also have sufficient *local exploration* to discourage any detrimental wear-out effect.

Recall (from Section 2) that the distributions ξ_j^D and ξ_j^Z have supports on $\{0, ..., a\}$ and $\{b, ..., N\}$, with $N \in \mathbb{N}$ and $0 \leq a < b \leq N$ being unknown fixed constants. Clearly, $D_t = 0$ or $Z_t = N$ would imply there are no loss of rewards due to wear-in or wear-out respectively (drop-

ping the dependence on arm j here for clarity). Additionally, $D_t > N/2$ would make the wear-in effect *too strong* as it necessitates the following: any algorithm will need to repeat the same arm for a majority portion of any contiguous N rounds, and further the algorithm would fail to accrue rewards from other arms that are played in the remaining portion of this set of N rounds. Hence, we assume $a \leq N/2 < b$ (refer to Appendix A.3 for a more detailed discussion about this assumption).

The way we tackle both the wear-in and wear-out effects is through a *key observation*: that playing a pair of arms (or more) with equal probability may provide sufficient *local exploration* which could nullify wear-out effect, while ensuring that arms are repeated often enough for sufficient *global exploitation* to also counter wear-in effect simultaneously. This is exactly what we do in the algorithm WI/WO-UCB (see Algorithm 2). Similar to the algorithm WI-UCB (of Section 3), WI/WO-UCB plays arms in phases. However, instead of repeating the same arm continuously for $n_m - n_{m-1}$ times, the algorithm plays a pair of arms with equal probability for a collective $n_m - n_{m-1}$ times in the m-th phase (for some new optimally chosen phase length parameters $\{n_m\}$). Intuitively it is similar to the following hypothetical setting: construct $^{|\mathcal{K}|}C_2$ pairs of arms from the original set $\mathcal{K}$, appropriately calculate the mean rewards for each pair and run an instance of WI-UCB with $^{|\mathcal{K}|}C_2$ arms. To be able to reuse the techniques and analysis from Section 3, we formalize the above idea of arm-pair or *compound arm* next. Notation wise, let (i, j) denote the compound arm constructed by playing the arms i and j with equal probability and $\mathcal{K}^2$ denote the set of all possible pairs composed of arms in $\mathcal{K}$. Clearly $\mu_{(i,j)} = (\mu_i + \mu_j)/2$.

Benchmark Policy: Again, recall the definition of expected regret from Section 2 (Equation 3). Since any benchmark also endures the priming effect, the benchmark policy of Section 3, which plays the best arm consistently for all rounds is not optimal (in fact, it will have linear regret). So for the setting in this Section, we define π to be the policy that knows the mean rewards of all arms and plays the top two arms (in terms of mean rewards, denoted by $\mu_{(1)}^*$ and $\mu_{(2)}^*$) with equal probability in each round. The policy π may not necessarily be optimal, however it provides a natural performance measure to contrast against for cases where a learning algorithm plays *compound arms*. For a more detailed discussion on the above benchmark and the optimal benchmark, see Appendix A.4.

As in the previous section, we assume the knowledge of the wear-in parameter $\mathbb{E}[D]$. The key design challenge is to calculate the $\{n_m\}$ sequence such that Algorithm 2 quickly eliminates sub-optimal *compound arms*, which

---

Input: Compound arms composed of pairs from $\mathcal{K}$: $\mathcal{K}^2$, time horizon T, and phase length parameters $\{n_m | m = 0, 1, 2, ...\}$.
Initialization: Phase index $m = 1$, $\mathcal{K}_m^2 = \mathcal{K}^2$, $\tilde{\Delta}_1 = 1$, $T_{(i,j)}(0) = \phi \ \forall (i,j) \in \mathcal{K}^2$, where ϕ is the empty set, and time index $t = 1$.
while $t \leq T$ **do**
  if $|\mathcal{K}_m^2| > 1$ **then**
    Play Compound Arms:
    for *each compound arm (i, j) in $\mathcal{K}_m^2$* **do**
      Set $T_{(i,j)}(m) = T_{(i,j)}(m - 1)$.
      Play either i or j with equal probability for $(n_m - n_{m-1})$ consecutive rounds and update $T_{(i,j)}(m)$.
      Accrue rewards according to the environment model Equation 2.
    end
    Eliminate Sub-optimal Pairs:
    for *each active pair (i, j) in $\mathcal{K}_m^2$* **do**
      $\overline{X}_{m,(i,j)} = \frac{1}{|T_{(i,j)}(m)|} \sum_{s \in T_{(i,j)}(m)} X_{s,(i,j)}$.
    end
    Construct $\mathcal{K}_{m+1}^2$ by eliminating all pairs (i, j) in $\mathcal{K}_m^2$ for which:
    $\overline{X}_{m,(i,j)} + \tilde{\Delta}_m/2 < \max_{(i,j)' \in \mathcal{K}_m^2} \overline{X}_{m,(i,j)'} - \tilde{\Delta}_m/2$.
    Update the Confidence Bound:
    Set $\tilde{\Delta}_{m+1} = \frac{\tilde{\Delta}_m}{2}$.
    Increment phase index m by 1 and update t based on $\{n_m\}$ values up to the current phase.
  end
  Play the single compound arm in $\mathcal{K}_m^2$ and update t.
end

Algorithm 2: WI/WO-UCB

---

we do in Lemma 6. Using this choice for WI/WO-UCB, we obtain an expected regret upper bound against the above benchmark as follows.

Theorem 2. *For any $\lambda > 0$, the expected (pseudo-)regret of* WI/WO-UCB *is bounded as:*

$$R_T \leq \sum_{i \in \mathcal{K}_1^2} \left(\Delta_i + \frac{64 \log(T)}{\Delta_i} + \frac{64 \log(T)}{3} \right.$$

$$\left. + 32 \sqrt{\log\left(\frac{4}{\Delta_i}\right) N \mathbb{E}[D] \log(T)} \right)$$

$$+ \sum_{i \in \mathcal{K}_1^2} \frac{4\Delta_i}{T} + \sum_{i \in \mathcal{K}_2^2} \frac{32}{T} + \max_{\{i \in \mathcal{K}_2^2 : \Delta_i < \lambda\}} \Delta_i T,$$

where $\mathcal{K}_1^2 = \{(i,j) \in \mathcal{K}^2 | \Delta_{(i,j)} > \lambda\}$, $\mathcal{K}_2^2 = \{(i,j) \in \mathcal{K}^2 | \Delta_{(i,j)} > 0\}$, *and* $\Delta_{(i,j)} = \frac{1}{2}(\mu_{(1)}^* + \mu_{(2)}^* - \mu_i - \mu_j)$.

The proof of Theorem 2 follows a similar proof strategy as that of Theorem 1, and has been provided in detail in Appendix A.2. It relies on an appropriate choice for $\{n_m\}$, as given by Equation 7 (see Appendix A.2). As before, we can also get an instance independent bound as shown below.

Corollary 2. *For all* $T \geq K^2$, *choosing* $\lambda = \sqrt{\frac{K^2 \log(T)}{T}}$, *using the inequality* $\log(1/\tilde{\Delta}_m) \leq \log(T)$, *and the observation that both* $|\mathcal{K}_1^2|$ *and* $|\mathcal{K}_2^2|$ *are* $O(K^2)$, *gives the following upper bound on the expected (pseudo-)regret of* WI/WO-UCB *is* $R_T \leq O\left(K\sqrt{T \log T} + K^2\sqrt{\log^2 T N \mathbb{E}[D]}\right)$.

When compared to Corollary 1, we immediately notice the following. The first term has an additional factor of $\sqrt{K}$ and the second term has an additional factor of $K\sqrt{N}$. The increased dependence on K can easily be attributed to the choice of using compound-arms by the algorithm. The dependence on $\sqrt{N}$ comes from the wear-out effect. Because $\mathbb{E}[D]$ can be $O(N)$ itself, overall, one can conclude that the priming effect (wear-in and wear-out) lead to an additive linear term in N (which also parameterizes these effects).

The following lemma provides a recipe to calculate n_m which is an input to Algorithm 2.

Lemma 6. *There exists a positive* n_m *for which the estimate* $\overline{X}_{m,(i,j)}$ *calculated by Algorithm 2 for an active pair* (i, j) $((i, j) \in \mathcal{K}_m^2)$ *in phase* m *satisfies the following inequality with probability at least* $1 - \frac{2}{T^2}$:

$$\overline{X}_{m,(i,j)} - \mu_{(i,j)} \leq \tilde{\Delta}_m/2,$$

where $\mu_{(i,j)} = \frac{1}{2}(\mu_i + \mu_j)$.

Outline of the proof: The phases are defined with respect to the compound arms and the reward R_{t,J_t} is due to arm played by Algorithm 2 at time t. We continue to build up on the same observation as used in the proof of Lemma 1: the cumulative sums of bias adjusted rewards $(\overline{X}_{m,(i,j)} - \mu_{(i,j)})$ can be decomposed into three sequences, where the additional third is due to the loss of rewards due to wear-out effect see Appendix A.2 for details. The value of n_m for which the lemma holds is given by:

$$n_m \leq 1 + \frac{4\log(T)}{\tilde{\Delta}_m^2} + \frac{16\log(T)}{3\tilde{\Delta}_m} + \frac{8\sqrt{Nm\mathbb{E}[D]\log(T)}}{\tilde{\Delta}_m}. \quad (7)$$

5 CONCLUSION

We considered the problem of showing recommendations and ads when the user behavior is influenced by priming effect: wear-in, where the user responds positively when shown the same recommendation multiple times in the recent past, and wear-out, where the user response dampens when the same recommendation is shown too frequently in the recent past. Modeling this in a bandit framework, we develop a new algorithm and show how its performance in terms of regret can be bounded.

An open problem is to develop a theory for general temporal dependencies of current rewards on past actions and rewards. Any departure from the functional form as described in the Section 2 may need additional analytical tools beyond what was presented here. One possible way is to model the temporal dependency structures via Markov decision processes. Regret analysis may not be straightforward in such settings or may produce loose bounds.

Further, the current regret bounds contain additive terms for priming, and it might be possible to achieve a better dependence on the priming effect parameters while having a worse dependence on other parameters. In this regard, characterizing the benchmarks thoroughly and finding tight lower bounds under the current model is a potential starting point. Extensions to adversarial behavior models, inclusion of contexts, handling non-stationarity, and most importantly, developing regret minimizing algorithms when multiple user behavioral effects including priming occur simultaneously, are some additional future research directions.

References

Rajeev Agrawal, MV Hedge, and Demosthenis Teneketzis. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. *IEEE Transactions on Automatic Control*, 33 (10):899–906, 1988.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory*, pages 217–226, 2009.

Peter Auer and Ronald Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2): 55–65, 2010.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Jeffrey S Banks and Rangarajan K Sundaram. Switching costs and the Gittins index. *Econometrica*, pages 687–694, 1994.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Nonstochastic bandits with composite anony-

mous feedback. In *Conference on Learning Theory*, pages 750–773, 2018.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Neural Information Processing Systems*, pages 2249–2257, 2011.

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: pow(T,2/3) regret. In *ACM Symposium on Theory of Computing*, pages 459–467. ACM, 2014.

Arnoud den Boer and N Bora Keskin. Dynamic pricing with demand learning and reference effects. *Available at SSRN 3092745*, 2017.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(1):1079–1105, 2006.

David A Freedman. On tail probabilities for martingales. *Annals of Probability*, pages 100–118, 1975.

David Gamarnik, John N Tsitsiklis, and Martin Zubeldia. Delay, memory, and messaging tradeoffs in distributed service systems. *Stochastic Systems*, 8(1):45–74, 2018.

Delbert Hawkins, Roger J Best, and Kenneth A Coney. *Consumer behavior*. McGraw-Hill Publishing, 2009.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461, 2013.

Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pages 767–776, 2015.

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

Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. In *Advances in neural information processing systems*, pages 3074–3083, 2017.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *ACM SIGACT Symposium on Theory of Computing*, pages 114–122. ACM, 2018.

Hao Ma, Xueqing Liu, and Zhihong Shen. User fatigue in online news recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1363–1372, 2016.

Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press, 2005.

Cornelia Pechmann and David W Stewart. Advertising repetition: A critical review of wearin and wearout. *Current issues and research in advertising*, 11(1-2): 285–329, 1988.

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, Erik Snowberg, et al. Batched bandit problems. *Annals of Statistics*, 44(2):660–681, 2016.

Ciara Pike-Burke and Steffen Grunewalder. Recovering bandits. In *Advances in Neural Information Processing Systems*, pages 14122–14131, 2019.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4102–4110, 2018.

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

Virag Shah, Jose Blanchet, and Ramesh Johari. Bandit learning with positive externalities. In *Advances in Neural Information Processing Systems*, pages 4918–4928, 2018.

Michael R Solomon, Dahren William Dahl, Katherine White, Judith L Zaichkowsky, and Rosemary Polegato. *Consumer behavior: Buying, having, and being*, volume 10. Pearson London, 2014.

István Szita and Csaba Szepesvári. Agnostic KWIK learning and efficient approximate reinforcement learning. In *Conference on Learning Theory*, pages 739–772, 2011.

Theja Tulabandhula and Yunjuan Wang. Thompson sampling for a fatigue-aware online recommendation system. In *International Symposium on AI and Mathematics*, 2020.

Kuang Xu and Se-Young Yun. Reinforcement with fading memories. In *ACM Conference on Measurement and Modeling of Computer Systems*, pages 90–92. ACM, 2018.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Shengjia Zhao, Enze Zhou, Ashish Sabharwal, and Stefano Ermon. Adaptive concentration inequalities for sequential decision problems. In *Advances in Neural Information Processing Systems*, pages 1343–1351, 2016.