Short-lived High-volume Bandits

Su Jia [*1] **Nishant Oli** [*2] **Ian Anderson** [2] **Paul Duff** [2] **Andrew A. Li** [3] **R. Ravi** [3]

Abstract

We study how to efficiently perform A/B/n testing for a *high-volume* of *short-lived* treatments. We formulate the problem as a multiple-play bandits model. In each round a set of k *actions* arrive. Each action is available for w rounds and has an unknown *reward rate*. In each round, the learner selects a **multiset** of n actions and immediately observes the realized rewards. We aim to minimize the *average loss* under a *random input model* where the instance is randomly drawn from a known prior distribution D. We show that if $k = O(n^\rho)$ for some $\rho > 0$, our policy achieves $\tilde{O}(n^{-\min\{\rho, \frac{1}{2}(1+\frac{1}{w})^{-1}\}})$ average loss on a sufficiently large class of prior distributions. We also complement this result by showing that every policy suffers $\Omega(n^{-\min\{\rho, \frac{1}{2}\}})$ average loss on the same class of distributions. We further validate the effectiveness of our policy through a large-scale field experiment on *Glance*, a content card-serving platform.

1. Introduction

Modern platforms leverage randomized experiments to make informed decisions from a given set of alternatives. As a particularly challenging scenario, these alternatives can potentially have at the same time (i) high *volume*, with thousands of new items being released each hour, and (ii) short *lifetime*, due to the transient nature of the contents. This challenge arises from, for example, recommending short-lived content on a video-sharing platform.

Orthogonal to lifetime, the problem is similarly well understood when there is a *low volume* of contents relative to the number of users - dedicated exploration methods, such as standard *A/B/n testing*, are sufficient for finding the most appealing content for the users; see, e.g., (Kohavi & Longbotham, 2017).

Naturally, then, the most challenging settings are where the contents to be selected are *short-lived* and have *high volume*, which occur in a variety of applications.
(A) **Recommender Systems.** From online advertising to social networks, platforms are sometimes faced with a massive amount of *short-lived* contents to be recommended to users. For example, more than 210 million Snaps were created on Snapchat each day in 2020, most of which expired within just 24 hours (Vuleta, 2021). The brevity of the lifetime can either be caused by the features of the content (e.g. breaking news) itself or by the transient nature of user attentions. The platform needs to decide which contents to appear at the top of the user news feeds to maximize user engagement.
(B) **Website Optimization.** In internet marketing, online platforms perform *multi-variate A/B/n testing* (e.g., (McFarland, 2012; Yang et al., 2017)) to test different designs of their user interface. For example, LinkedIn runs over 400 concurrent experiments per day to compare different website designs, with the goal of encouraging users to establish or refine their personal profiles, or increasing the subscriptions to LinkedIn Premium (Xu et al., 2015). The combinatorial nature of the decision space results in a high volume of items to test. Specifically, the number of possible designs can be exponential in the number of features (e.g., logos, font, background color, etc.).

However, these items can have a short life. For example, considering the non-stationarity of the underlying environment (say, due to seasonality or societal events), any estimation is only reliable for a short amount of time. A simple approach is to partition the time horizon into short segments, in which the conversion rates of the designs are approximately constant, and then view the same design in different segments as separate copies.

Adding to this challenge is *personalization*. A common and naive approach is to cluster the users and solve the problem for each cluster separately. However, when restricting to a cluster, the number of users becomes much smaller while the number of actions remains the same. In other words, personalization would substantially limit the resources available for experimentation, rendering the problem harder.

[*] First and second authors. All other authors are alphabetically ordered. [1] Center of Data Science for Enterprise and Society (CDSES), Cornell University, Ithaca, USA [2] Glance, Bangalore, India [3] Tepper School of Business, Carnegie Mellon University, Pittsburgh, USA. Correspondence to: Su Jia <sj693@cornell.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Thus motivated, we study how to efficiently identify the effective items from a high volume of short-lived candidates. To encapsulate the key features of the problem, we propose the Short-lived High-volume Bandits (SLHVB) problem. We employ a *multiple-play* bandit framework with the following key features:
1) Multiple-play. In each *round*, we can choose each action (or *arm*) multiple times as long as the total number of plays is n, which corresponds to the number of user interactions in this time period;[1]
2) High-Volume of Arrivals. In each round, a set of $k = n^\rho$ actions arrive where $\rho > 0$;
3) Short Lifetime. Each arms is available for w rounds, which is a known, small constant;
4) Random Input Model (RIM). The reward rates are drawn from a fixed known distribution. Our analysis focuses on the **average** performance over all instances, rather than against the worst-case instance.

We present a policy that recursively refines the exploration for up to w times and compares it against what is best achievable for different regimes of $\rho > 0$.

1.1. Our Contributions

This work contributes to the literature of multi-armed bandits and online controlled experiments in the following ways.
1) A Novel and Practical Formulation. Our first contribution is formulating an online learning model that faithfully models a ubiquitous problem faced by online platforms. Our formulation considers a practical metric - the average loss - which better reflects the quality of a policy than worst-case metrics, which are more common in previous literature.
2) Average Regret for Batched Bandits. As a subroutine for our showing main result, we show that the Batched Successive Elimination algorithm (Gao et al., 2019) for the Batched Bandits (BB) problem achieves $\tilde{O}((k/n)^{\frac{\ell}{\ell+2}})$ **average** regret using ℓ rounds of adaptivity whenever $\rho \geq \frac{\ell-1}{2\ell+1}$.[2] In particular, for $\ell \geq 2$, this bound is asymptotically better than the optimal $\tilde{O}((k/n)^{1/2} \cdot n^{2^{-\ell}})$ **worst-case** regret. This contrast is strongest when $\rho = 1/2$ - in this case our policy has average regret $n^{-1/2+O(1/\ell)}$ whereas the optimal worst-case regret is $\tilde{O}(n^{-1/4+2^{-\ell}})$.
3) Policy for SLHVB. We show that any policy for BB with $R(n,k)$ average regret can be converted into a policy for SLHVB with $\tilde{O}\left(n^{-\rho} + R(n,k)\right)$ average loss. Our average regret bound for BB then implies an $\tilde{O}(n^{-\min\{\rho, \frac{w}{2(w+1)}\}})$ average loss bound for SLHVB, by choosing a suitable ℓ depending on ρ and w.
4) Nearly Matching Lower Bound. We show that any policy for SLHVB suffers an $\Omega(n^{-\min\{\rho,\frac{1}{2}\}})$ average loss. Further, we juxtapose this result with a lower bound on the **worst-case** loss which is higher, and hence highlights the value of our RIM.
5) Large-Scale Field Experiment. Finally and most importantly, we validated the effectiveness of our policy in a field experiment via collaboration with *Glance*, a leading lock-screen content platform in India. This firm faces exactly the aforementioned challenge: they generate hundreds of *content cards* on an hourly basis, which are available for at most 48 hours. Their current recommender is based on a state-of-the-art Deep Neural Network (DNN), which is time-consuming to re-train and hence unable to utilize user feedback in a timely manner. In a field experiment, we observed that our 1-Layered Sieve policy outperforms their DNN-based recommender by $4-7\%$ in user engagement.

1.2. Related Work

Our problem is a variant of the *Multi-Armed Bandits* (MAB) problem (Lai et al., 1985). Three lines of work are most related to ours: *multiple-play bandits*, *mortal bandits* and *high-volume Bandits*.

Multiple-play Bandits. In this variant, several arms are selected in each round. Many results in single-play bandits can be generalized to the multi-play variant, for example, (Komiyama et al., 2015) showed that the instance dependent regret bound for Thompson sampling can be generalized to the multi-play setting. One motivation of the multi-play variant is *online ranking*, see e.g., (Radlinski et al., 2008; Lagrée et al., 2016; Gauthier et al., 2022), where the learner presents an *ordered* list of items to each user, viewed sequentially under certain click model. Unlike in our formulation, the arms have infinite lifetime. Further, there is no arrival of new arms, so the learner does not need to take into consideration the ages of the arms.

Mortality of Arms. A quintessential motivation for the mortality of arms is online advertising. In the classical *pay-by-click model*, the ad broker matches each ad from a large corpus to contents, and is paid by the advertiser (i.e. who created the ad) only when an ad is clicked. As the key feature, an ad becomes unavailable when its advertiser's budget is run out. Thus motivated, (Chakrabarti et al., 2008) introduced the *mortal bandits* problem and considered two death models. In the deterministic model, each arm dies after being selected for a certain number of times, which corresponds to the advertisers' budget in the advertising example. In the stochastic lifetime model, an arm dies with a fixed probability every time it is selected. Relatedly, in *rotting bandits* (Levine et al., 2017), each arm's reward rate decays in the number of times it has been selected. In particular, if the reward function is an indicator function, then effectively each arm has a finite lifetime.

[1] Do not confuse this "n" with the "n" in the term "A/B/n" testing - the latte is actually our k, i.e., the number of treatments.

[2] All big-O's are with respect to $n \to \infty$, with ρ fixed.

Motivated by demand learning in assortment planning, (Farias & Madan, 2011) considered the *irrevocable bandits* problem that bears both the multi-play and mortality features: arms are selected in batches and discarded immediately once selected. Unlike in this work, however, none of these models considered arrivals, and hence the learner does not need to take into consideration the age of the actions.

High Volume of Arms. Most existing work concerning large volume of arms considered the *worst-case* regret of a policy, i.e., the regret on the worst input in a given family (e.g. (Berry et al., 1997; Zhang & Frazier, 2021)). As a distinctive feature, in this work we consider a random input model where the reward rates follow a known distribution, and perform an *average case* analysis. As we will soon see, our formulation leads to theoretical results that would be otherwise impossible.

(Wang et al., 2008) also assumed that the reward rates of the arms are independently drawn from a common distribution such that the probability of being ε-optimal is $O(\varepsilon^\beta)$ where $\beta \in [0,1]$ is a known constant. However, unlike in our problem, there are no arrivals and hence the policy does not need to balance the exploration for items with different ages.

Finally, we are aware of another variant of MAB closely related to the multi-play bandits (and hence to this work).

Low-Adaptive Bandits Algorithms. In the *batched bandits* problem (Perchet et al., 2016; Agarwal et al., 2017) we aim to achieve low regret using low *adaptivity*. Unlike in the multi-play setting, here the learner can partition the time horizon $[T]$ into w *batches* where w is a given constant, and choose a batch (i.e., a multiset) of arms based on the realizations in the previous batches. Alternatively, w can be interpreted as a constraint on the adaptivity of the policy. In the classical setting, the learner has unlimited adaptivity, i.e., $w = T$. (Gao et al., 2019) showed that for any k arms, we can achieve $\tilde{O}(\sqrt{kT})$ regret whenever $w = \Omega(\log\log T)$, which is optimal among all policies with **arbitrary** adaptivity.

2. Formulation

Suppose at the start of each round $t = 1, 2, \cdots$, a set A_t of k *actions* (or *arms*) arrives, available in rounds $t, \ldots, t+w$. We call w the *lifetime* and viewed it as a small constant. In each round, the learner selects a *multiset* of n available arms - each arm can be chosen multiple times, as long as the total number of plays is n. If an arm a is selected m times, the learner immediately observes i.i.d. rewards $X_1, \ldots, X_m$ with mean μ_a.[3] For simplicity, we consider only Bernoulli random variables, although the analysis can be generalized to subgaussian reward distributions in a straightforward manner.

For concreteness, let us relate the above formulation to the recommendation problem. Most online platforms retrieve user interaction data and update the predictions periodically. A round corresponds to such a period. Further, n corresponds to the number of user impressions in a round, and is assumed to be independent of the quality of the recommendation. Selecting a multiset of arms corresponds to recommending exactly **one** item to each impression.

A Random Input Model. Unlike most work in MAB, here we consider a *Random Input Model* (RIM) where μ_a's are assumed be drawn i.i.d. from a known distribution D. This is realistic in practice as D can be approximated using past data. In contrast to the minimax framework, this formulation better reflects the reality and more importantly, enables us to obtain theoretical guarantees that would be otherwise impossible; see Section 3.

As a standard assumption in statistical learning (Ghosal, 2001; Petrone & Wasserman, 2002; Audibert & Tsybakov, 2007), we assume the density of D is bounded from above and below away from 0, although our analysis extends to more general distributions (with possibly weaker guarantees).

Assumption 2.1 (Bounded Density Assumption)**.** The distribution D admits a density function f with a compact support $\mathcal{C}$, and there exist constants $C_1, C_2 > 0$ such that $C_1 \le f(x) \le C_2$ for all $x \in \mathcal{C}$. W.l.o.g.[4] we assume $\mathcal{C} = [0,1]$.

The Average Loss. The procedure of selecting arms can be encoded by a *policy* $\pi = (\pi_t)$ satisfying $\sum_{a \in A^t_{t-w}} \pi_t(a) = n$ for each t, where $A^{t'}_t := \bigcup_{s=t}^{t'} A_s$ for $0 < t < t'$. The expected reward in round t is then $\sum_a \pi_t(a) \cdot \mu(a)$. The problem is easy if μ_a's were known. In fact, let $a^*_t = \arg\max\{\mu_a : a \in A^t_{t-w}\}$, then the optimal policy is given by $\pi^*_t(a) = n \cdot \mathbb{1}[a = a^*_t]$.

When reward rates are unknown, the performance of a policy is measured by the following notion of *average loss*.[5] We first define the **finite-time** loss. For any integer $T \ge w$ and reward rates $\mu = \{\mu_a\}$, consider

$$\mathrm{Loss}_n(\pi; T, \mu) := \frac{1}{nT} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a \in A^t_{t-w}} \pi_t(a) \cdot (\mu^*_t - \mu_a)\right],$$

where the expectation is only over the reward realizations but **not** μ. To characterize the long-run performance of the system, let $T \to \infty$ and define the *average loss* (or simply,

[3] "i.i.d." stands for "identically independently distributed"

[4] "w.l.o.g." stands for "without loss of generality"

[5] To avoid confusions, we use the term "loss" for SLHVB and "regret" for Batched Bandits in Section 4 to prevent confusions.

loss) as

$$\mathrm{Loss}_n(\pi) := \varlimsup_{T\to\infty} \mathbb{E}_{\mu\sim D}\left[\mathrm{Loss}_n(\pi; T, \mu)\right].$$

We are interested in how rapidly $\mathrm{Loss}_n(\pi)$ vanishes given a fixed $\rho > 0$.

3. Lower Bounds

We first show that every policy suffers $\Omega(n^{-\min\{\rho,\frac{1}{2}\}})$ average loss. We defer the details to Appendix A and only explain the high level ideas here. Consider the average loss L_t given by

$$L_t = \frac{1}{n} \cdot \mathbb{E}\left[\sum_{a\in A_{t-w}^t} \pi_t(a) \cdot (\mu_t^* - \mu_a)\right],$$

where $\mu_t^* = \mu_{\max}(A_{t-w}^t)$. By linearity of expectation, it suffices to lower bound L_t. We will show that the regret is large in each of the following two cases. First consider the case where $\pi_t(A_t) \geq \frac{n}{2}$, i.e., the policy sufficiently explores the arms arriving at t. Due to the RIM and by Assumption 2.1, the reward rates μ_a's are approximately evenly spaced. Since there are wk arms available at any time, the gap between the best and second best arms a, a' is approximately $\frac{1}{wk}$, formally, $\mu_a - \mu_{a'} \gtrsim \frac{1}{wk}$. Consequently, if the arriving batch A_t does not contain the best available arm, i.e., $a_t^* \notin A_t$, then $\mu(a_t^*) - \mu_{\max}(A_t) \gtrsim \frac{1}{wk}$. Moreover, by symmetry a_t^* appears in each of the w available batches equally likely, so the above event occurs with probability $1 - \frac{1}{w} \geq \frac{1}{2}$, we have $L_t \gtrsim \frac{1}{2} \cdot \frac{1}{wk} \sim \frac{1}{w} \cdot n^{-\rho}$.

Now consider the other case, $\pi_t(A_t) < \frac{n}{2}$, i.e., A_t is under-explored in round t. Consider the event $a_t^* \in A_t$. Then, by an argument similar to the first case, we have $\mu(a_t^*) - \mu_{\max}\left(A_{t-1}^{t-w}\right) \sim \frac{1}{wk}$. Further, this event occurs with probability $\frac{1}{w}$, so in this case $L_t \gtrsim \frac{1}{w} \cdot \frac{1}{wk} = \frac{1}{w^2} n^{-\rho}$. Thus in both cases, we have $L_t \geq \frac{1}{w^2} n^{-\rho}$, as formally stated below. We defer the proof to Appendix A.1.

Proposition 3.1 (Lower Bound)**.** *For any policy π and $\rho \geq 0$, we have* $\mathrm{Loss}_n(\pi) \geq \frac{1}{12w^2} \cdot n^{-\rho}$.

However, this bound becomes very weak when ρ is large. We next present a lower bound specific to $\rho \geq \frac{1}{2}$. We argue that to avoid an $\Omega(n^{-1/2})$ regret, a policy has to identify an $n^{-1/2}$-optimal arm a, i.e., $\mu_a \geq \mu_t^* - O(n^{-1/2})$. To this goal, by Assumption 2.1, the learner needs to explore $\gtrsim n^{1/2}$ distinct arms, leading to an $\Omega(n^{1/2})$ regret. We formally state the second lower bound below. We defer the proof to Appendix A.2.

Proposition 3.2 (Lower Bound, $\rho > \frac{1}{2}$)**.** *If $\rho \geq \frac{1}{2}$, then for any policy π, we have* $\mathrm{Loss}_n(\pi) \geq \frac{1}{96w^2} \cdot n^{-1/2}$.

Note that when $\rho = 1/2$, the above two lower bounds have the same asymptotic order. By combining the above two lower bounds, we immediately obtain the following.

Theorem 3.3 (Lower Bound)**.** *For any $\rho > 0$ and policy π, it holds that* $\mathrm{Loss}_n(\pi) \geq \frac{1}{96w^2} \cdot n^{-\min\{\rho,1/2\}}$.

Finally, we present a lower bound on the *worst-case* regret that highlights the advantage of the RIM. We show that no policy achieves $o(1)$ worst-case regret, defined as

$$\mathrm{Loss}'_n(\pi) := \max_{\mu} \varlimsup_{T\to\infty} \mathrm{Loss}(\pi; T, \mu),$$

where $\max$ is over all instances $0 \leq \mu \leq 1$.

Proposition 3.4 (Worst-case Regret)**.** *For any policy π and lifetime $w > 0$, we have* $\mathrm{Loss}'_n(\pi) \geq \frac{1}{2w^2}$.

This bound can be proved by constructing an instance with binary rewards, such that in each round there is exactly one arm a with $\mu_a = 1$.

4. Upper Bounds

In this section, we establish the connections to the Batched Bandits (BB) problem. We first explain how a *semi-adaptive* algorithm for BB can be converted into a policy for SLHVB, and show how the guarantees translate from one problem to another. Then, we consider a variant of the *Batched Successive Elimination* (BSE) algorithm (Gao et al., 2019) and analyze its regret. Finally, using this result, we obtain a policy for the SLHVB problem with $\tilde{O}(n^{-\min\{\rho,\frac{1}{2}\cdot(1+\frac{1}{w})^{-1}\}})$ average loss. We defer the details to Appendix C.

4.1. Batched Bandits

In the BB problem (Perchet et al., 2016), the learner is given k arms, n *slots*, and an *adaptivity level* $\ell \geq 1$. In each *phase* $i = 0, 1, \ldots, \ell$, the learner selects a multiset[6] M_i of arms such that $\sum_{i=0}^{\ell} |M_i| = n$. Each time an arm a is selected, a reward is randomly drawn from a Bernoulli distribution with unknown *reward rate* μ_a, and is immediately observed. If an arm is selected multiple times in one phase, the realized rewards may be different. The goal is to maximize the expected total reward.

Given an instance $(\mu_a)_{a\in[k]}$, the *regret*[7] of an algorithm $\mathbb{A}$ for BB is defined as

$$\mathrm{Reg}_n(\mathbb{A}; \mu) := \frac{1}{n} \cdot \mathbb{E}\left[\sum_{a\in[k]} (\mu^* - \mu_a) \cdot N_a\right]$$

where $\mu^* = \arg\max_{a\in[k]}\{\mu_a\}$ and N_a is the number of times an arm a is selected. Most existing work for MAB

[6] an easy way to remember the indexing rule: M_i is the batch of arms selected when using the i-th chance of being adaptive.

[7] To avoid confusions, we use synonymous terms *"algorithm"* for BB and *"policy"* for SLHVB; for the objective we use *"regret"* for BB and *"loss"* for SLHVB.

considered the *worst-case regret*

$$\mathrm{Reg}_n^{\mathrm{wc}}(\mathbb{A}) := \max_{\mu\in[0,1]^k} \mathrm{Reg}_n(\mathbb{A};\mu).$$

Analogous to the notion of average loss for the SLHVB problem, for BB we define the *average regret* over all instances drawn from a distribution D as

$$\mathrm{Reg}_n^{\mathrm{avg}}(\mathbb{A}) = \mathbb{E}_{\mu\sim D}\left[\mathrm{Reg}_n(\mathbb{A};\mu)\right].$$

An appealing class of algorithms is the *semi-adaptive* algorithms, where the cardinality of each batch M_i of arms selected are decided in advance **non-adaptively** and the selection of arms in each batch depends on the reward realizations **adaptively**.

Definition 4.1 (Semi-adaptive Algorithm)**.** Given an *adaptivity level* $\ell > 0$, a *semi-adaptive* algorithm is specified by (i) *grid sizes* $\varepsilon_0,\ldots,\varepsilon_{\ell-1} \in (0,1)$ with $\sum_{i=0}^{\ell-1}\varepsilon_i < 1$, and (ii) a family of decision rules

$$\mathbb{A}_j : ([k]\times\mathbb{R})^{n_{j-1}} \to [k]^{n_j - n_{j-1}}$$

for $j = 0,\ldots,\ell$ where[8]

$$n_j := \begin{cases} \sum_{i=0}^{j}\varepsilon_i n, & \text{if} \quad j = 0,\ldots,\ell-1,\\ n, & \text{if} \quad j = \ell,\\ 0, & \text{if} \quad j = -1.\end{cases}$$

(Gao et al., 2019) proposed the following *Batched Successive Elimination* (BSE) algorithm. In each phase the algorithm keeps track of a subset $S_i \subseteq [k]$ of *surviving arms* (or simply *survivors*) that serve as the candidates for the optimal arm, computed in the following manner. Initially $S_0 = [k]$. In each phase $i = 0,\ldots,\ell-1$, the algorithm explores arms in S_i uniformly, i.e., selects each arm $\varepsilon_i n/|S_i|$ times. Finally in the last phase, i.e., phase ℓ, we arbitrarily choose an arm from S_ℓ. For completeness, we formally state the algorithm in Algorithm 1.

(Gao et al., 2019) analyzed the performance of the BSE algorithm under two types of grids. The first is called the *minimax* grid: let $c = n^{\frac{1}{1-2^{-\ell}}} = \tilde{O}(\sqrt{n}\cdot n^{2^{-\ell}})$, we recursive define $\varepsilon_{i+1} = c\sqrt{\varepsilon_i}$ where $\varepsilon_0 = c$. The second is called *geometric* grid, as the batch sizes form a geometric progression. Specifically, let $c = n^{1/\ell}$, recursively we define $\varepsilon_{i+1} = c\cdot\varepsilon_i$ for $\ell = 1,\ldots,\ell-1$.

As the name suggests, the BSE algorithm under the minimax grid achieves the minimax regret. The geometric grid, on the other hand, is shown to achieve nearly-optimal instance-dependent regret. We rephrase (Gao et al., 2019)'s main result as follows.

Theorem 4.2 (Theorem 1 of (Gao et al., 2019))**.** *Given any adaptivity level ℓ, denote by $\mathrm{BSE}_\ell^{\mathrm{minimax}}$ and $\mathrm{BSE}_\ell^{\mathrm{geo}}$ the BSE algorithms under the minimax and geometric grids. Then, for any k-armed instance, it holds that $\mathrm{Reg}_n^{\mathrm{wc}}(\mathrm{BSE}_\ell^{\mathrm{minimax}}) = \tilde{O}(\sqrt{k/n}\cdot n^{2^{-\ell}})$. Further, if the highest and second highest reward rates differ by Δ, then $\mathrm{Reg}_n^{\mathrm{wc}}(\mathrm{BSE}_\ell^{\mathrm{geo}}) = \tilde{O}(n^{\frac{1}{\ell}-1}\cdot k/\Delta)$.*

Algorithm 1 Batched Successive Elimination Policy $\mathrm{BSE}(\varepsilon_0,\ldots,\varepsilon_{\ell-1};k')$ for Batched Bandits.

```
Input:
    $\ell$: adaptivity level
    $\varepsilon_0,\ldots,\varepsilon_{\ell-1} \in (0,1)$: exploration intensities
    $n$: number of arms to be selected in total
    $A$: a set of $k$ arms
Let $\tilde{A}$ be a uniformly random subset of $A$ of size $k'$
for $i = 0,1,...,\ell-1$ do
    if $|S_i| = 1$ then
        Set $S_\ell = S_i$; Break
    end if
    if $|S_i| \geq 2$ then
        $n_i \leftarrow \left\lfloor \frac{\varepsilon_i n}{|S_i|} \right\rfloor$
        for $a \in S_i$ do
            Select arm $a$ for $n_i$ times and observe rewards $X_{a,1},...,X_{a,n_i}$
            $\overline{X}_a \leftarrow \frac{1}{n_i}\sum_{j=1}^{n_i} X_{a,j}$
        end for
        $\overline{X}_{\max} \leftarrow \max\left\{\overline{X}_a : a \in S_i\right\}$
        $S_{i+1} = \{a \in S_i : |\overline{X}_a - \overline{X}_{\max}| \leq 3n_i^{-1/2}\log^{1/2} n\}$
    end if
end for
Select any arm in $S_\ell$ for $n - \sum_{i=0}^{\ell-1}\lfloor\varepsilon_i n\rfloor$ times
```

Apparently, the worst-case regret bound trivially holds for average regret. On the other hand, the worst-case regret of $\mathrm{BSE}_\ell^{\mathrm{geo}}$ depends on the parameter Δ, which is random under the RIM. By Assumption 2.1, we have $\mathbb{E}_\mu[\Delta^{-1}] = \tilde{\Theta}(k)$, immediately implying an $\tilde{O}(k^2 n^{\frac{1}{\ell}-1})$ bound on the average regret, as summarized below.

Corollary 4.3 (Average Regret of BSE)**.** *For any adaptivity level $\ell \geq 1$, we have $\mathrm{Reg}_n^{\mathrm{avg}}(\mathrm{BSE}_\ell^{\mathrm{minimax}}) = \tilde{O}(\sqrt{k/n}\cdot n^{2^{-\ell}})$ and $\mathrm{Reg}_n^{\mathrm{avg}}(\mathrm{BSE}_\ell^{\mathrm{geo}}) = \tilde{O}(k^2 n^{-(1-\frac{1}{\ell})})$.*

In particular, when $\ell = O(\log\log n)$, we have $n^{2^{-\ell}} = \tilde{O}(1)$ and hence the bound in Theorem 4.2 becomes $\tilde{O}(\sqrt{k/n})$. Surprisingly, this matches the $\Omega(\sqrt{k/n})$ lower bound for **unlimited** adaptivity level; see, e.g., (Auer et al., 2002).

4.2. Breaking the $\sqrt{k/n}$ Bound

Using a different geometric grid, we obtain a stronger bound on the average regret for the BSE algorithm, compared to

[8] To avoid integrality issues, we assume $\varepsilon_i n/|S_i|$ is an integer. This does not affect the asymptotic order of our bounds.

the $\tilde{O}(\sqrt{k/n})$ bound in Corollary 4.7. More precisely, the common ratio in this geometric progression is k/n, whereas in (Gao et al., 2019) the dependence only depends on n. With some foresight, let's consider the following choice of exploration intensities.

Definition 4.4 (Revised Geometric Grid)**.** For any adaptivity level ℓ and $i \le \ell$, we define $\varepsilon^\star_{i,\ell} = (k/n)^{(\ell-i)/(\ell+2)}$ for integers $i = 0, \dots, \ell-1$. For each fixed ℓ, denote by $\mathrm{BSE}^\star_\ell$ the algorithm BSE_ℓ with grid size $(\varepsilon^\star_{0,\ell}, \dots, \varepsilon^\star_{\ell-1,\ell})$.

Our bound for a fixed ℓ requires ρ to be larger than the following threshold, otherwise there is no need for having ℓ phases.

Definition 4.5 (Threshold Exponent)**.** For each integer $\ell \ge 1$, we define the *threshold exponent* as $\theta_\ell := \frac{\ell-1}{2\ell+1}$.

To see why we need ρ to be large, observe that by the RIM, μ_a's are $\sim n^{-\rho}$ distance apart. On the other hand, if ρ is small, the confidence intervals are expected to be narrower than $n^{-\rho}$ with strictly less than ℓ phases, and hence there is no need for extra layers.

More concretely, suppose $\ell = 2$ and consider $\rho = 0.04 < \theta_2 = 1/5$. To see why the second phase is **redundant**, note that by Definition 4.4, $\varepsilon^\star_0 = \tilde{O}((k/n)^{1/2})$, so each arm is selected $\sim \varepsilon^\star_0 n/k = (n/k)^{1/2} = n^{0.48}$ times upon arrival. Thus, the confidence interval of each arm after selecting the first batch of arms has width $\sim (n^{0.48})^{-1/2} = n^{-0.24}$. On the other hand, the reward rates are spaced at $n^{-\rho} = n^{-0.04}$ distance apart on average. Therefore, the optimal arm in A_t is likely to be identified after just one phase.

Theorem 4.6 (Average Regret of $\mathrm{BSE}^\star_\ell$)**.** *For any adaptivity level ℓ with $\rho \ge \theta_\ell$, the average regret of the BSE algorithm under the revised geometric grids can be bounded as* $\mathrm{Reg}^{\mathrm{avg}}_n(\mathrm{BSE}^\star_\ell) = \tilde{O}((k/n)^{\ell/(\ell+2)})$.

The proof of this result crucially relies on the RIM. We argue that due to the RIM, the number of surviving arms is bounded by a function in n, ρ, j w.h.p.[9] after a number j of phases. Consequently, each surviving arm then is guaranteed at least a certain amount of slots in phase $j+1$.

As a caveat, the above argument breaks when there remains only one surviving arm after a number of phases. This event, however, occurs with low probability, since ρ is greater than the *threshold exponent*.

Note that average regret becomes lower as ℓ increases. But ℓ can not be arbitrarily high since we required $\theta_\ell < \rho$. To find the maximal feasible ℓ, consider the following meta-algorithm: If $\rho < 1/2$, choose the maximum ℓ with $\rho \ge \theta_\ell$. Equivalently, choose $\ell = \ell^*(\rho) := \lfloor \frac{1+\rho}{1-2\rho} \rfloor$. If $\rho \ge \frac{1}{2}$, then the condition $\rho \ge \theta_\ell$ holds for any ℓ, and so we can choose arbitrarily large ℓ. We formally state this result as follows.

Corollary 4.7 (Explicit Form of the Average Regret)**.** *The average regret of the BSE algorithm under the revised geometric grids satisfies the following:*

(i) *If $0 < \rho < \frac{1}{2}$, then for adaptivity level $\ell^*(\rho) = \left\lfloor \frac{1+\rho}{1-2\rho} \right\rfloor$ the average regret can be bounded as*

$$\mathrm{Reg}^{\mathrm{avg}}_n\left(\mathrm{BSE}^\star_{\ell^*(\rho)}\right) = \tilde{O}\left(\left(\frac{k}{n}\right)^{\frac{\ell^*(\rho)}{\ell^*(\rho)+2}}\right).$$

(ii) *If $\rho \ge \frac{1}{2}$, then for any adaptivity level $\ell \ge 1$, the average regret can be bounded as*

$$\mathrm{Reg}^{\mathrm{avg}}_n\left(\mathrm{BSE}^\star_\ell\right) = \tilde{O}\left(\left(\frac{k}{n}\right)^{\frac{\ell}{\ell+2}}\right).$$

In particular, $\mathrm{Reg}^{\mathrm{avg}}_n(\mathrm{BSE}^\star_\ell) = (k/n)^{1-O(1/\ell)}$ *as* $\ell \to \infty$.

Note that $n^{2^{-\ell}} > 1$ for any ℓ, so the $\tilde{O}(\sqrt{k/n} \cdot n^{2^{-\ell}})$ bound in Theorem 4.2 is no better than $\tilde{O}((k/n)^{1/2})$. In contrast, note that whenever $\rho \ge 1/5$, we have $\ell^*(\rho) \ge 2$, so the bound in Corollary 4.7 is stronger.

This contrast becomes sharper as ρ increases. For example, with $\rho = 1/3$, we have $\ell^*(\rho) = 4$, and hence $\mathrm{Reg}^\star_{\ell^*(\rho)} = \tilde{O}((k/n)^{2/3})$, which is better than $\tilde{O}((k/n)^{1/2})$. In the extreme case, for $\varepsilon > 0$ and $\rho = \frac{1}{2} - \varepsilon$ we have $\ell^*(\rho) = \Omega(1/\varepsilon)$ and hence $(k/n)^{\Omega(1/\varepsilon)}$, which is much stronger a guarantee than $\tilde{O}((k/n)^{1/2})$.

4.3. From Regret to Loss

A semi-adaptive algorithm for BB induces the following policy for SLHVB. Let ε_i be the grid sizes, which is usually chosen such that $\sum_{j=0}^{\ell-1} \varepsilon_i = o(n)$, e.g., as in Definition 4.4. At each time t, the induced policy uses $\varepsilon_i n$ slots to perform the i-th phase of the BB algorithm on A_{t-i}, i.e., the arms arriving at time $t-i$. Meanwhile, we use the remaining $(1 - \sum_{i=0}^{\ell-1} \varepsilon_i)n$ slots for "exploitation"; see Algorithm 2 for a formal statement.

As a nice property, the induced policy of any semi-adaptive algorithm selects **exactly** n arms in each round and is hence valid for SLHVB. In fact, suppose the semi-adaptive policy for BB has grid sizes (ε_i). Using the notations in Algorithm 2, for each round t and each phase $j = 0, \dots, \ell$, an arm $a \in A_{t-j}$ is selected $N^t_{j,a}$ times. By definition of ε_j, we have $\sum_{a \in A_{t-j}} N^t_{j,a} = \varepsilon_j n$. Summing over all j's, the total number of arms selected is

$$\sum_{j=0}^{\ell} \sum_{a \in A_{t-j}} N^t_{j,a} = \sum_{j=0}^{\ell-1} \varepsilon_j n + \left(n - \sum_{j=0}^{\ell-1} \varepsilon_j n\right) = n.$$

[9] We say an event occurs "w.h.p." (which stands for "with high probability") if it has probability $1 - n^{\Omega(1)}$.

Algorithm 2 Induced Policy $\pi[\mathbb{A}; k']$ for SLHV Bandits

```
Input:
    $\mathbb{A}$: a semi-adaptive algorithm for BB
    $k'$: cardinality of the random subset of arms
for $t = 1, 2, \dots$ do
    Sample a random subset $\tilde{A}_t$ of $A_t$ of size $k'$
    $A_t \leftarrow \tilde{A}_t$
    for $j = 0, \dots, \ell$ do
        for $a \in A_{t-j}$ do
            Query $\mathbb{A}$ for the number $N^t_{j,a}$ of times to select arm $a$ in the $j$-th batch
            Select $a$ for $N^t_{j,a}$ times, and return the rewards to $\mathbb{A}$
        end for
    end for
end for
```

Note that in Algorithm 2, we only use age-ℓ arms for "exploitation". This is not practical – an obviously better policy would exploit the empirically best arm in $A_{t-w}^{t-\ell} = \bigcup_{j=\ell}^{w} A_{t-j}$ rather than just in $A_{t-\ell}$. We choose to state the policy in this way for simplicity of analysis. By doing this, the loss only increases by $\tilde{O}(n^{-\rho})$, which is on the same order as the $\tilde{\Omega}(n^{-\rho})$ lower bound in Theorem 3.1, and is hence not essential.

We can convert the **regret** of any semi-adaptive BB algorithm to the **loss** of the induced SLHVB policy as follows.

Proposition 4.8 (Regret-to-Loss Conversion). *Suppose $\mathbb{A}$ is a semi-adaptive algorithm with average regret $R(n,k)$ on any BB instance with k arms and n slots. Then for any k', the loss of the induced policy $\pi[\mathbb{A}, k']$ for SLHVB satisfies* $\mathrm{Loss}_n(\pi[\mathbb{A}, k']) = \tilde{O}(1/k' + \rho^{-1} n^{-\rho} + R(n, k'))$.

We explain the high level idea and defer the details to Appendix B. The term $1/k'$ captures the gap between the optimal reward rate in the resampled subset and the original set of arms. The second term, $n^{-\rho} = 1/k$, captures the loss due to the reward uncertainty of new arrivals – we have to decide how many times to choose them without any knowledge of their qualities apart from the RIM assumption. The final term, $R(n, k')$, bounds the average regret on the resampled subset, which contains k' arms.

4.4. Loss of the Induced Policy

The average regret bound in Proposition 4.6 and the regret-to-less conversion formula in Proposition 4.8 immediately lead to a loss bound for the induced SLHVB policy. To formalize this, we first recapitulate some notations. Given a semi-adaptive algorithm $\mathbb{A}$ for BB, we defined $\pi[\mathbb{A}, k']$ as the induced policy for SLHVB with resampling size k'. Let $\mathrm{BSE}^\star(\ell, k') := \pi[\mathrm{BSE}^\star_\ell, k']$ be the SLHVB policy induced by the BSE algorithm with the revised geometric grid $(\varepsilon^\star_{j;\ell})$ specified in Definition 4.4.

Proposition 4.9 (Loss of the Induced Policy). *If $\ell \le w$ and $\rho \ge \theta_\ell$, then*

$$\mathrm{Loss}_n\left(\mathrm{BSE}^\star(\ell, k)\right) = \tilde{O}\left(n^{-\rho} + n^{(\rho-1)\cdot\frac{\ell}{\ell+2}}\right) = \begin{cases} \tilde{O}\left(n^{-\rho}\right), & \textit{if } \rho < \frac{\ell}{2(\ell+1)}, \\ \tilde{O}\left(n^{(\rho-1)\cdot\frac{\ell}{\ell+2}}\right), & \textit{otherwise.} \end{cases} \tag{1}$$

Different from the previous subsections, we now have an extra lifetime constraint that $\ell \le w$. Our meta-policy, called the Hybrid policy, chooses suitable ℓ for any given ρ; see Algorithm 3).

Algorithm 3 Hybrid Policy $\mathbb{H}(\rho; w)$

```
If $\rho < \frac{1}{5}$ then set $\ell = 1$ and $k' = k$.
If $\frac{1}{5} \le \rho < \frac{w}{2w+2}$, then set $k' = k$ and choose any $\ell \le w$ such that $\theta_\ell = \frac{\ell-1}{2\ell+1} \le \rho < \frac{\ell}{2\ell+2}$.
If $\rho \ge \frac{w}{2w+2}$, then set $\ell = w$ and $k' = n^{\frac{w}{2w+2}}$.
Invoke $\mathrm{BSE}^\star(\ell, k')$.
```

We prove the following in Appendix D.

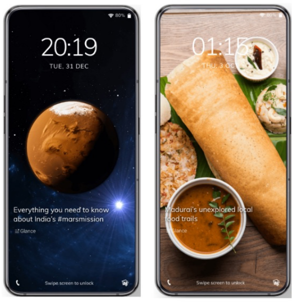
Theorem 4.10 (Loss of the Hybrid Policy). *For any SLHVB instance with n slots per round, volume exponent $\rho > 0$ and lifetime $w \ge 1$, the average loss of the Hybrid policy satisfies* $\mathrm{Loss}_n(\mathbb{H}(\rho; w)) = \tilde{O}(\rho^{-1} \cdot n^{-\min\{\rho, \frac{1}{2}(1+\frac{1}{w})^{-1}\}})$.

To better understand the impact of the lifetime w, observe that for $w = 1, 2$, the regret bounds are asymptotically $r_1(n) := n^{-\min\{\rho, 1/4\}}$ and $r_2(n) := n^{-\min\{\rho, 1/3\}}$. In particular, when $\rho > 1/4$, we have $r_2 = o(r_1)$ as $n \to \infty$.

5. Field Experiment

We further validated the effectiveness of our policy in a field experiment, via collaboration with Glance, a leading lock-screen content platform that faces the aforementioned challenge. Specifically, their marketing team curates around 200 *content cards* (or simply, *cards*) per hour, and around 70% of them expire in 48 hours. Each card consists of a link to some external content (e.g., video, news or article), along with a short text description; see Figure 1. The firm needs to recommend cards to users with the goal of maximizing the total user engagement, measured by the total *duration* of the interactions and the number of click-throughs.

This problem can be cast as an SLHVB problem, assuming that the total impressions is independent of the recommendations. Two key quantities are of particular interest for each card: (i) the *click-through rate* (CTR) and (ii) the *average duration* per impression. Both metrics are unknown when a card is released. We mix the above two metrics by considering the *conversions*. A conversion occurs if either a



Everything you need to know about India's #marsmission

Madurai's unexplored local food trails

Figure 1. Glance's content cards. The image and text summarize the content, and by clicking on the link the user is shown the details.

click-through occurs or the duration reaches a threshold of $\theta = 5$ seconds. For simplicity, we assume that the rewards across different impressions are i.i.d. random variables.

The platform sends cards to the users on an hourly basis. For each user, the platform replaces the cards that the user has viewed in the past with new content cards, provided the device is connected to the internet. The users can swipe through the cards stored in the platform's app. When a user is interested in the content, they can click on the provided link and be redirected to an external source for further engagement, and then be redirected back to the app when finished. To decide which cards to send to each user, the firm deployed a recommender system based on a state-of-the-art Deep Neural Network (DNN). This DNN predicts, for each pair of user and card, the expected conversion probability, using (i) the user interaction history and (ii) the card's text, image features.

Although the current recommender system works reasonably well, there is a considerable potential for improvement. Most notably, the current system only uses the user feedback to update the users' behavioral signature for future prediction, and does not directly leverage the feedback in making recommendations for similar users. In particular, they do not use the feedback to adjust the predictions on the conversion rates **directly**. This may have caused substantial loss in user engagement.

It is thus vital for the platform to find a recommendation policy that (i) can learn the true conversion rates of new cards quickly using user interaction data and (ii) is computationally simple to deploy. Our policy is well-suited for this task. We defer the implementation details to Appendix E.

We perform a detailed analysis on the field experiment result and show that the simplest version of our policy outperforms the DNN-based recommender by about 4% in total duration (see Figure 2) and nearly 7% in the total number of click-throughs per-user-per-day; see Figure 3. We defer the detailed analysis in Appendix F.

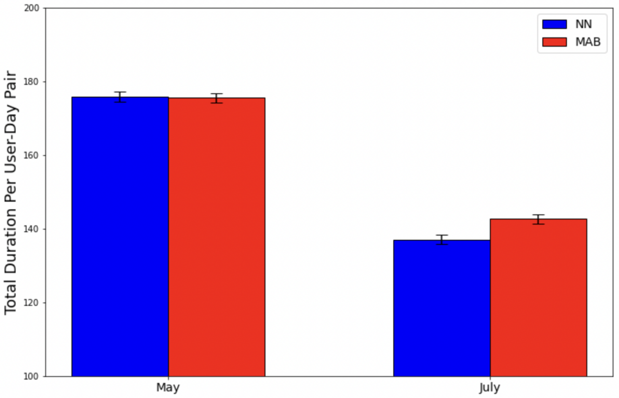


Figure 2. Duration per user-day pair.

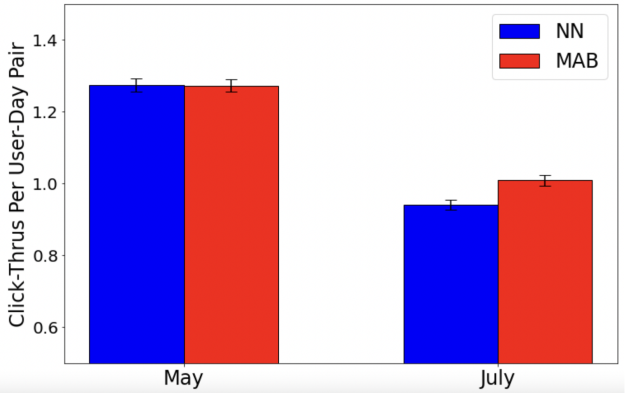


Figure 3. Click-through per impression.

References

Agarwal, A., Agarwal, S., Assadi, S., and Khanna, S. Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Conference on Learning Theory*, pp. 39–75. PMLR, 2017.

Audibert, J.-Y. and Tsybakov, A. B. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2): 608–633, 2007.

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

Berry, D. A., Chen, R. W., Zame, A., Heath, D. C., and Shepp, L. A. Bandit problems with infinitely many arms. *The Annals of Statistics*, 25(5):2103–2116, 1997.

Chakrabarti, D., Kumar, R., Radlinski, F., and Upfal, E. Mortal multi-armed bandits. *Advances in neural information processing systems*, 21:273–280, 2008.

Farias, V. F. and Madan, R. The irrevocable multiarmed bandit problem. *Operations Research*, 59(2):383–399, 2011.

Gao, Z., Han, Y., Ren, Z., and Zhou, Z. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32, 2019.

Gauthier, C.-S., Gaudel, R., and Fromont, E. Unirank: Unimodal bandit algorithms for online ranking. In *International Conference on Machine Learning*, pp. 7279–7309. PMLR, 2022.

Ghosal, S. Convergence rates for density estimation with bernstein polynomials. *The Annals of Statistics*, 29(5): 1264–1280, 2001.

Kohavi, R. and Longbotham, R. Online controlled experiments and a/b testing. *Encyclopedia of machine learning and data mining*, 7(8):922–929, 2017.

Komiyama, J., Honda, J., and Nakagawa, H. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *International Conference on Machine Learning*, pp. 1152–1161. PMLR, 2015.

Lagrée, P., Vernade, C., and Cappe, O. Multiple-play bandits in the position-based model. *Advances in Neural Information Processing Systems*, 29, 2016.

Lai, T. L., Robbins, H., et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6 (1):4–22, 1985.

Levine, N., Crammer, K., and Mannor, S. Rotting bandits. *Advances in neural information processing systems*, 30, 2017.

McFarland, C. *Experiment!: Website conversion rate optimization with A/B and multivariate testing*. New Riders, 2012.

Perchet, V., Rigollet, P., Chassang, S., Snowberg, E., et al. Batched bandit problems. *Annals of Statistics*, 44(2): 660–681, 2016.

Petrone, S. and Wasserman, L. Consistency of bernstein polynomial posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(1):79–100, 2002.

Radlinski, F., Kleinberg, R., and Joachims, T. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, pp. 784–791, 2008.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Vuleta, B. How much data is created every day? 2021. URL `https://seedscientific.com/how-much-data-is-created-every-day/`.

Wang, Y., Audibert, J.-Y., and Munos, R. Algorithms for infinitely many-armed bandits. *Advances in Neural Information Processing Systems*, 21, 2008.

Wasserman, L. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

Xu, Y., Chen, N., Fernandez, A., Sinno, O., and Bhasin, A. From infrastructure to culture: A/b testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2227–2236, 2015.

Yang, F., Ramdas, A., Jamieson, K. G., and Wainwright, M. J. A framework for multi-a (rmed)/b (andit) testing with online fdr control. *Advances in Neural Information Processing Systems*, 30, 2017.

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

A. Details of the Lower Bounds

A.1. Proof of Proposition 3.1

Consider the event

$$W_{t-1} = \left\{ 1 - \frac{2}{(w-1)k} \le \mu_{\max}(A_{t-w}^{t-1}) \le 1 - \frac{1}{(w-1)k} \right\}.$$

We first show that this event occurs with large probability.

Lemma A.1. *If $K \ge 10$, then $\mathbb{P}(W_{t-1}) \ge \frac{1}{8}$.*

Proof. Denote $\mu_{\max} = \mu_{max}(A_{t-w}^{t-1})$. Let $H = \left\{ \mu_{\max} \ge 1 - \frac{1}{(w-1)k} \right\}$ and $H' = \left\{ \mu_{\max} < 1 - \frac{2}{(w-1)k} \right\}$. Then for any $k \ge 10$, we have

$$\mathbb{P}\left[\overline{H}\right] = \mathbb{P}\left[\mu_{\max} < 1 - \frac{1}{(w-1)k} \right] = \left(1 - \frac{1}{(w-1)k} \right)^{(w-1)k} \ge \frac{3}{4} \cdot e^{-1},$$

and thus $\mathbb{P}[H] \le 1 - \frac{3}{4e}$. Moreover, since $\mathbb{P}[H'] = \left(1 - \frac{2}{(w-1)k} \right)^{\frac{(w-1)k}{2} \cdot 2} \le e^{-2}$, we conclude that

$$\mathbb{P}[W_{t-1}] \ge 1 - \mathbb{P}[H'] - \mathbb{P}[H] \ge 1 - \left(1 - \frac{3}{4e} \right) - e^{-2} > \frac{1}{8}.$$

□

Thus, by giving up a constant factor in the loss, we may restrict our attention on the event W_{t-1}. For each time period t, define probability measure $\mathbb{P}_t(\cdot) = \mathbb{P}[\cdot | W_{t-1}]$ and $\mathbb{E}_t(\cdot) = \mathbb{E}[\cdot | W_{t-1}]$. Consider the event that at time t, the policy selects arms from A_t at least $n/2$ times, i.e., $\mathcal{E}_t = \left\{ \sum_{a \in A_t} \pi_t(a) \ge \frac{n}{2} \right\}$. Further, define

$$\mathcal{B}_t^- = \left\{ \mu_{\max}(A_t) \le \mu_{\max}(A_{t-w}^{t-1}) - \frac{1}{k} \right\} \text{ and } \mathcal{B}_t^+ = \left\{ \mu_{\max}(A_t) \ge \mu_{\max}(A_{t-w}^{t-1}) + \frac{1}{wk} \right\}.$$

Observe that

$$\begin{aligned} \mathbb{E}_t(L_t) &= \mathbb{E}_t(L_t | \mathcal{E}_t \cap \mathcal{B}_t^-) \cdot \mathbb{P}_t(\mathcal{E}_t \cap \mathcal{B}_t^-) + \mathbb{E}_t(L_t | \mathcal{E}_t \cap \mathcal{B}_t^+) \cdot \mathbb{P}_t(\mathcal{E}_t \cap \mathcal{B}_t^+) \\ &\quad + \mathbb{E}_t(L_t | \overline{\mathcal{E}}_t \cap \mathcal{B}_t^-) \cdot \mathbb{P}_t(\overline{\mathcal{E}}_t \cap \mathcal{B}_t^-) + \mathbb{E}_t(L_t | \overline{\mathcal{E}}_t \cap \mathcal{B}_t^+) \cdot \mathbb{P}_t(\overline{\mathcal{E}}_t \cap \mathcal{B}_t^+) \\ &\ge \mathbb{E}_t[L_t | \mathcal{E}_t \cap \mathcal{B}_t^-] \cdot \mathbb{P}_t(\mathcal{E}_t \cap \mathcal{B}_t^-) + \mathbb{E}_t(L_t | \overline{\mathcal{E}}_t \cap \mathcal{B}_t^+) \cdot \mathbb{P}_t(\overline{\mathcal{E}}_t \cap \mathcal{B}_t^+). \end{aligned} \tag{2}$$

We next lower bound the above two terms. Note that the event in the first term, i.e., $\mathcal{E}_t \cap \mathcal{B}_t^-$, says that the policy selects arms from A_t for at least $n/2$ times, but the best arm in A_t is suboptimal (w.r.t. $\mu_{\max}(A_{t-w}^t)$) by $1/k$. Intuitively, when this event occurs, the policy suffers at least $\frac{n}{2} \cdot \frac{1}{k} = \frac{n}{2k}$ loss in this round. On the other hand, the event in the other term, i.e., $\overline{\mathcal{E}}_t \cap \mathcal{B}_t^+$, says that the policy selects A_t for less than $n/2$ times, but the best arm in $\mu(A_{t-w}^{t-1})$ is suboptimal by $n/(wk)$. By a similar argument, we can show that the loss in this round is at least $\frac{1}{2wk}$ on this event. At a high level, both these two lower bounds are caused by not knowing quality of the new batch of arms. We formalize the above ideas in the following lemma.

Lemma A.2 (Loss for Uncertainty in the New Batch)**.** *The loss in any round t satisfies $\mathbb{E}_t[L_t | \mathcal{E}_t \cap \mathcal{B}_t^-] \ge \frac{n}{2k}$ and $\mathbb{E}_t[L_t | \overline{\mathcal{E}}_t \cap \mathcal{B}_t^+] \ge \frac{n}{2wk}$.*

Proof. Write $\mu_t^* = \mu_{\max}(A_t)$. Recall from the definition that when $\mathcal{B}_t^-$ occurs, we have $\mu_t^* - \mu_{\max}(A_t) \ge 1/k$. Thus, if π

selects arms from A_t for $\Omega(n)$ times, an $\Omega\left(n/k\right)$ loss is incurred. Formally,

$$\begin{aligned}
\mathbb{E}_t\left[L_t \middle| \, \mathcal{E}_t \cap \mathcal{B}_t^-\right] &= \mathbb{E}_t\left[\sum_{a\in A_{t-w}^t} \pi_t(a)\cdot(\mu_t^* - \mu_a) \,\middle|\, \mathcal{E}_t \cap \mathcal{B}_t^-\right] \\
&\geq \mathbb{E}_t\left[\sum_{a\in A_t} \pi_t(a)\cdot(\mu_t^* - \mu_a) \,\middle|\, \mathcal{E}_t \cap \mathcal{B}_t^-\right] \\
&\geq \mathbb{E}_t\left[\sum_{a\in A_t} \pi_t(a)\cdot(\mu_t^* - \mu_{\max}(A_t)) \,\middle|\, \mathcal{E}_t \cap \mathcal{B}_t^-\right]. \qquad (3)
\end{aligned}$$

Note that conditional on $\mathcal{E}_t, \mathcal{B}_t^-$ and W_{t-1} (recall that $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|W_{t-1}]$), we have $\sum_{a\in A_t} \pi_t(a) \geq \frac{n}{2}$ and $\mu_t^* - \mu_{\max}(A_t) \geq 1/k$. Thus,

$$(3) \geq \frac{n}{2}\cdot\frac{1}{k} = \frac{n}{2k}.$$

Now consider the second term in (2). When $\mathcal{B}_t^+$ occurs, we have $\mu_{\max}(A_t) - \mu_{\max}(A_{t-w}^{t-1}) \geq \frac{1}{wk}$, so if the π selects arms in A_{t-w}^{t-1} for $\Omega(n)$ times, an $\Omega(n/k)$ loss is incurred. Formally,

$$\begin{aligned}
\mathbb{E}_t\left[L_t \middle| \, \overline{\mathcal{E}_t} \cap \mathcal{B}_t^+\right] &= \mathbb{E}_t\left[\sum_{a\in A_{t-w}^t} \pi_t(a)\cdot(\mu_t^* - \mu_a) \,\middle|\, \overline{\mathcal{E}_t} \cap \mathcal{B}_t^+\right] \\
&\geq \mathbb{E}_t\left[\sum_{a\in A_{t-w}^{t-1}} \pi_t(a)\cdot\left(\mu_{\max}\left(A_t\right) - \mu_{\max}(A_{t-w}^{t-1})\right) \,\middle|\, \overline{\mathcal{E}_t} \cap \mathcal{B}_t^+\right]. \qquad (4)
\end{aligned}$$

Note that conditional on $\overline{\mathcal{E}_t}, \mathcal{B}_t^-$ and W_{t-1}, we have $\sum_{a\in A_{t-w}^{t-1}} \pi_t(a) \geq \frac{n}{2}$ and $A_t - \mu_{\max}(A_{t-w}^{t-1}) \geq \frac{1}{wk}$. Therefore,

$$(4) \geq \frac{n}{2}\cdot\frac{1}{wk} = \frac{n}{2wk}.$$

□

A.2. Proof of Proposition 3.2

Consider the event $G_t = \left\{\mu_{\max}(A_{t-w}^t) \geq 1 - n^{-1/2}\right\}$ where $t \geq w$.

Lemma A.3 ($\mu_{\max}$ is Close to 1)**.** *For $k > \sqrt{n}$, we have $\mathbb{P}[G_t] \geq \frac{1}{2}$.*

Proof. Since $|A_{t-w}^t| = wk$ and the reward rate of each arm is drawn i.i.d. uniformly, we have $\mathbb{P}[\overline{G_t}] = (1-\delta)^{kw} = (1-\delta)^{\frac{1}{\delta}\cdot kw\delta} \leq e^{-kw\delta}$. Since $k > \sqrt{n}$, we have $k\delta > n^{\frac{1}{2}} \cdot n^{-\frac{1}{2}} \geq 1$, so $\mathbb{P}[\overline{G_t}] \leq e^{-kw\delta} \leq e^{-w} \leq \frac{1}{2}$, i.e. $\mathbb{P}[G_t \geq \frac{1}{2}]$. □

An arm is said to be *unexplored* at time t if it has never been selected by the policy before. Our proof considers the number of unexplored arms selected in each round, as formalized below.

Definition A.4 (Unexplored Arms)**.** We say an arm a is *unexplored* at round t if it has never been selected by (the start of) round t; formally, this means $\sum_{\tau=1}^{t-1} \pi_\tau(a) = 0$. For each round t, define M_t as the subset of unexplored arms selected in this round; formally, $M_t = \{a \in A_{t-w}^t : \pi_t(a) > 0 \text{ and } \sum_{s=1}^{t-1} \pi_s(a) = 0\}$. Moreover, as for any t, t', we write $M_t^{t'} = \bigcup_{\tau=t}^{t'} S_\tau$.

From a myopic perspective, selecting too many unexplored arms in a round leads to high loss since, by Assumption 2.1, each unexplored arm is suboptimal by $\Omega(1)$ on average. We formalize this in the following lemma.

Lemma A.5 (Cost of Selecting Many unexplored Arms)**.** *For any $t \geq 1$, consider the event $V_t = \left\{|M_t| \geq \frac{\sqrt{n}}{4w}\right\}$. The expected loss in round t conditional on V_t satisfies $\mathbb{E}\left[L_t|V_t\right] \geq \frac{\sqrt{n}}{24w}$.*

Proof. Write $\mu_t^* = \mu_{\max}(A_{t-w}^t)$ and $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|G_t]$ for any t. We start with lower bounding the conditional loss. Observe that

$$\begin{aligned}\mathbb{E}_t\left[L_t|V_t\right] &= \mathbb{E}_t\left[\sum_{a\in A}\pi_t(a)\cdot(\mu_t^*-\mu_a)\Big|\ V_t\right]\\ &\geq \sum_{a\in A}\mathbb{E}_t\left[\pi_t(a)\cdot\mathbb{1}(a\in M_t)\cdot(\mu_t^*-\mu_a)\Big|\ V_t\right]\\ &\geq \sum_{a\in A}\mathbb{E}_t[\mathbb{1}(a\in M_t)\cdot\Big(1-n^{-1/2}-\mu_a\Big)\mid V_t], \end{aligned} \tag{5}$$

where the last inequality follows by definition of $\mathbb{E}_t$ and that $a \in M_t$ implies $\pi_t(a) \geq 1$. To further simplify, note that the reward rate of any unexplored arm a is an independent draw and is hence independent with the set M_t of unexplored arms selected in this round, so

$$(5) = \sum_{a\in A}\mathbb{E}_t\left[\mathbb{1}(a\in M_t)|\ V_t\right]\cdot\mathbb{E}_t\left[1-n^{-1/2}-\mu_a|\ V_t\right]. \tag{6}$$

Note that $\mathbb{E}_t[\mu_a|V_t] = \mathbb{E}[\mu_a] = \frac{1}{2}$, we have $\mathbb{E}_t\left[1-n^{-1/2}-\mu_a|\ V_t\right] = 1-n^{-1/2}-\frac{1}{2} \geq \frac{1}{3}$ for any $n \geq 1$. Therefore,

$$(6) \geq \mathbb{E}_t\left[\sum_{a\in A}\mathbb{1}(a\in M_t)\Big|V_\tau\right]\cdot\frac{1}{3} = \mathbb{E}_t\left[|M_t|\Big|V_\tau\right]\cdot\frac{1}{3} \geq \frac{\sqrt{n}}{12w}.$$

By Lemma A.5, we have $\mathbb{P}[G_t] \geq \frac{1}{2}$, and thus

$$\mathbb{E}[L_t|V_t] \geq \mathbb{E}_t[L_t|V_t]\cdot\mathbb{P}[G_t] \geq \frac{1}{2}\cdot\frac{\sqrt{n}}{12w} = \frac{\sqrt{n}}{24w}.$$

□

The above says that selecting too many unexplored arms leads to high loss. On the flip side, selecting too few unexplored arms also leads to high loss. To show this, consider the *cold-start* event where no δ-good arms that π ever selected is still available at the start of round t.

Definition A.6 (Cold-start Event). For each $t \geq 1$, the *cold-start event* is defined as $B_t = \left\{\mu_{\max}(S_{t-w}^t) \leq 1 - n^{-1/2}\right\}$.

As the name suggests, when this event occurs, the policy has little information about the available arms, and hence it behaves as if the time horizon has **restarted**. This leads to high loss. In fact, a policy has to identify a δ-good arm to have low loss from time $t-w$ to $t+w$. This forces the policy to explore $\Omega(1/\delta) = \Omega(n^{1/2})$ unexplored arms, which leads to an $\Omega(n^{1/2})$ loss. This result is implied by the lower bound for *infinite armed bandits* (Wang et al., 2008).

Lemma A.7 (Cold-start Incurs High Loss). *For any t, we have* $\mathbb{E}\left[\sum_{\tau=t}^{t+w} L_\tau\Big|B_t\right] \geq \frac{\sqrt{n}}{6w}$.

To apply the above, we next characterize when B_t would occur. Consider the number N_t of new arms selected in $[t-w, t]$. If $\mathbb{E}N_t \geq 1/\delta$, then on average the policy suffers $\Omega(1/\delta) = \Omega(\sqrt{n})$ loss. If $\mathbb{E}N_t < 1/\delta$, then with $\Omega(1)$ probability, **none** of those N_t arms are δ-good, which leads to the cold-start event B_t. We formalize this idea below.

Lemma A.8 (Under-Exploration Leads to Cold-start Event). *Suppose* $\mathbb{E}\left[\sum_{s=t-w}^{t} L_s\right] \leq \frac{\sqrt{n}}{96w^2}$, *then* $\mathbb{P}[B_t] \geq \frac{1}{2}$.

Proof. Consider the event $E_s = \left\{\mu_{\max}(S_s) \geq 1 - \frac{1}{\sqrt{n}}\right\}$ that the for any $s = 1, 2\ldots$, then our goal reduces to showing that $\mathbb{P}[E_s] \leq \frac{1}{2w}$ for every $s = t-w, \ldots, t$. In fact, if none of the events $E_{t-w}, \ldots, E_t$ occurs, i.e. no unexplored arm selected in the past w rounds is $n^{-1/2}$-good, then B_t occurs. It then follows from the union bound that

$$\mathbb{P}\left[B_t\right] \geq \mathbb{P}\left[\bigcap_{\tau=t-w}^{t}\overline{E_\tau}\right] = 1 - \mathbb{P}\left[\bigcup_{\tau=t-w}^{t}E_\tau\right] \geq 1 - w\cdot\frac{1}{2w} = \frac{1}{2},$$

and the proof will be complete.

To show $\mathbb{P}[E_s] \le \frac{1}{2w}$ for $s \in \{t-w, ..., t\}$, observe that for any fixed s,

$$\mathbb{P}[E_s] = \mathbb{P}[E_s|V_s] \cdot \mathbb{P}[V_s] + \mathbb{P}[E_s|\overline{V_s}] \cdot \mathbb{P}[\overline{V_s}] \le \mathbb{P}[V_s] + \mathbb{P}[E_s|\overline{V_s}], \tag{7}$$

where we recall that $V_t = \left\{|M_t| \ge \frac{\sqrt{n}}{4w}\right\}$. we will bound each of the two terms in (7) by $\frac{1}{4w}$ separately. To see $\mathbb{P}[V_s] \le \frac{1}{4w}$, note that

$$\frac{\sqrt{n}}{96w^2} \ge \sum_{t'=t-w}^{t} \mathbb{E}[L_{t'}] \ge \mathbb{E}[L_s] \ge \mathbb{E}[L_s|V_s] \cdot \mathbb{P}[V_s].$$

Note that by Lemma A.5, we have $\mathbb{E}[L_s|V_s] \ge \frac{\sqrt{n}}{24w}$, so $P[V_s] \le \frac{1}{4w}$.

To bound the second term in (7), note that by definition of E_s, for any $C, \delta > 0$ we have

$$\mathbb{P}\left[\overline{E}_s \middle| S_s \le C\right] \ge (1-\delta)^C \ge (1-\delta)^{\frac{1}{\delta}\cdot\delta C} \ge e^{-\delta C} \ge 1 - \delta C,$$

where the last inequality follows since $e^{-x} \ge 1 - x$ for any $x \in \mathbb{R}$. In particular, for $C = \frac{\sqrt{n}}{4w}$ and $\delta = \frac{1}{\sqrt{n}}$, we have

$$\mathbb{P}\left[\overline{E_s}\middle|\overline{V_s}\right] = \mathbb{P}\left[\overline{E_s}\middle| S_s \le \frac{\sqrt{n}}{4w}\right] \ge 1 - \frac{1}{\sqrt{n}} \cdot \frac{\sqrt{n}}{4w} = 1 - \frac{1}{4w},$$

i.e.,

$$\mathbb{P}\left[E_s\middle|\overline{V_s}\right] = 1 - \mathbb{P}\left[\overline{E}_s\middle|\overline{V_s}\right] \le \frac{1}{4w}.$$

□

We now combine the above to establish the $\sqrt{n}$ lower bound.

Proof of Proposition 3.2. Fix any round t. Decompose L_{t-w}^{t+w} into the loss before and after round t as

$$L_{t-w}^{t+w} = \sum_{\tau=t-w}^{t+w} \mathbb{E}L_\tau = \sum_{\tau=t-w}^{t-1} \mathbb{E}L_\tau + \sum_{\tau=t}^{t+w} \mathbb{E}L_\tau.$$

If the first term, i.e., the loss before time t, is greater than $\frac{\sqrt{n}}{96w^2}$, then the claim holds trivially. Otherwise, by Lemma A.8, the cold-start event B_t occurs w.p. $\mathbb{P}[B_t] \ge \frac{1}{2}$. Thus, the loss after time t can be bounded using Lemma A.7 as

$$\sum_{\tau=t}^{t+w} \mathbb{E}L_\tau \ge \mathbb{E}\left[\sum_{\tau=t}^{t+w} L_\tau \middle| B_t\right] \cdot \mathbb{P}[B_t] \ge \frac{\sqrt{n}}{6w} \cdot \frac{1}{2} = \frac{\sqrt{n}}{12w} > \frac{\sqrt{n}}{96w^2}.$$

□

A.3. Proof of Theorem 3.1

Now we are ready to show the $\Omega(1/k)$ lower bound. By (2) and Lemma A.2, we have

$$\begin{aligned}\mathbb{E}_t(L_t) &\ge \mathbb{E}_t[L_t|\mathcal{E}_t \cap \mathcal{B}_t^-] \cdot \mathbb{P}_t(\mathcal{E}_t \cap \mathcal{B}_t^-) + \mathbb{E}_t(L_t|\overline{\mathcal{E}}_t \cap \mathcal{B}_t^+) \cdot \mathbb{P}_t(\overline{\mathcal{E}}_t \cap \mathcal{B}_t^+) \\ &\ge \frac{n}{2k} \cdot \mathbb{P}_t(\mathcal{E}_t \cap \mathcal{B}_t^-) + \frac{n}{2wk} \cdot \mathbb{P}_t(\overline{\mathcal{E}_t} \cap \mathcal{B}_t^+).\end{aligned} \tag{8}$$

Note that the events $\mathcal{E}_t$ and $\mathcal{B}_t^+$ ($\overline{\mathcal{E}_t}$ and $\mathcal{B}_t^-$ resp.) are independent conditional on G_{t-1}, so

$$(8) \ge \frac{n}{2wk} \cdot \left(\frac{1}{2} \cdot \mathbb{P}_t(\mathcal{E}_t) + \frac{1}{w} \cdot \mathbb{P}_t(\overline{\mathcal{E}_t})\right) \ge \frac{n}{2w^2k}.$$

Therefore,

$$\mathbb{E}[L_t] \geq \mathbb{E}[L_t \cdot \mathbb{1}(G_{t-1})] = \mathbb{E}[L_t | G_{t-1}] \cdot \mathbb{P}[G_{t-1}] \geq \frac{n}{2w^2 k} \cdot \frac{1}{8} = \frac{n}{16w^2 k}.$$

Summing over t and taking the limit, we conclude that

$$\mathrm{Loss}_n(\pi) = \overline{\lim_{T\to\infty}} \frac{1}{nT} \sum_{t=1}^{T} \mathbb{E}[L_t] \geq \frac{1}{16w^2 k}.$$

□

B. Proof of Proposition 4.8: Regret-to-Loss Conversion

As we recall, given a policy $\pi = (\pi_t)$ for SLHV bandits, the loss in round t is $L_t = \frac{1}{n} \cdot \mathbb{E}\left[\sum_{a \in A^t_{t-w}} \pi_t(a) \cdot (\mu^*_t - \mu_a)\right]$. We first decompose L_t into the following *internal* and *external* loss.

Definition B.1 (External and Internal Loss)**.** Let $\mathbb{A}$ be a semi-adaptive algorithm for the BB problem with adaptivity level ℓ and $\pi = (\pi_t)$ be the induced policy for the SLHVB problem. For any round t, integer $j \leq w$, let $\Delta_{t,j} = \mu_{\max}(A^t_{t-w}) - \mu_{\max}(A_t)$. Define the *external* and *internal loss* as

$$L^{\mathrm{ext}}_t = \mathbb{E}\left[\max_{1\leq j\leq w} \Delta_{t,j}\right] \quad \text{and} \quad L^{\mathrm{int}}_t = \mathbb{E}\left[\sum_{j=0}^{\ell} \sum_{a \in A_{t-j}} \frac{\pi_t(a)}{n} \cdot (\mu_{\max}(A_{t-j}) - \mu(a))\right]. \tag{9}$$

Here, the term $\Delta_{t,j}$ is *external* in the sense that it does **not** depend on the policy, but only on the randomness in the RIM. On the other hand, the term L^{int}_t is internal – it measures the reward gap between the selected arms and the best arms among the respective age groups. It is straightforward to show the following.

Lemma B.2 (Loss Decomposition)**.** *In any round t, the loss satisfies* $L_t \leq L^{\mathrm{int}}_t + L^{\mathrm{ext}}_t$.

Proof. By the definition of internal and external loss, we have

$$\begin{aligned}
L_t &= \frac{1}{n} \cdot \mathbb{E}\left[\sum_{a \in A^t_{t-w}} \pi_t(a) \cdot (\mu^*_t - \mu_a)\right] \\
&= \mathbb{E}\left[\sum_{j=0}^{\ell} \sum_{a \in A_{t-j}} \frac{\pi_t(a)}{n} \cdot \left(\mu_{\max}(A^t_{t-w}) - \mu_a\right)\right] && (\text{since } \pi_t(a) = 0 \text{ if } a \in A^t_{t-w} \backslash A^t_{t-\ell}) \\
&= \mathbb{E}\left[\sum_{j=0}^{\ell} \sum_{a \in A_{t-j}} \frac{\pi_t(a)}{n} \cdot (\Delta_{t,j} + (\mu_{\max}(A_{t-j}) - \mu_a))\right] && (\text{by definition of } \Delta_{i,j}) \\
&\leq \mathbb{E}\left[\max_{1\leq j\leq w} \Delta_{t,j}\right] + \mathbb{E}\left[\sum_{j=0}^{\ell} \sum_{a \in A_{t-j}} \frac{\pi_t(a)}{n} \cdot (\mu_{\max}(A_{t-j}) - \mu_a)\right]. && \left(\text{since } \sum_{j=0}^{\ell} \sum_{a \in A_{t-j}} \frac{\pi_t(a)}{n} = 1\right)
\end{aligned}$$

□

We next bound the external and internal losses separately. We start by showing the external regret is $\tilde{O}(1/k)$. Recall from the definition of the RIM that $\mu_a \sim D$ and from Assumption 2.1 that the density of D is bounded by $C_2, C_1 > 0$ from above and below.

Proposition B.3 (External Loss)**.** *For any round t, the external loss can be bounded as* $L^{\mathrm{ext}}_t \leq \frac{3\rho \log n}{C_1 k}$ *for any sufficiently large n.*[10]

[10] We say that a property $\mathcal{P}_n$ (e.g., an inequality) holds for "any sufficiently large n" if there exists a constant $n_0 > 0$ such that $\mathcal{P}_n$ holds whenever $n \geq n_0$.

Proof. Consider any $i \in [w]$ and $Y_i := \mu_{\max}(A_{t-i})$. For any $\varepsilon < 1$ and arm $a \in A$, by the RIM and Assumption 2.1, we have $\mathbb{P}[\mu_a > 1-\varepsilon] \geq C_1\varepsilon$, or equivalently, $\mathbb{P}[\mu_a \leq 1-\varepsilon] \leq 1 - C_1\varepsilon$. In particular, for $\varepsilon = \frac{2\rho \log n}{C_1 k}$, we have

$$\mathbb{P}\left[Y_i < 1 - \frac{2\rho \log n}{C_1 k}\right] \leq \left(1 - C_1 \cdot \frac{2\rho \log n}{C_1 k}\right)^k = \left(1 - \frac{2\rho \log n}{k}\right)^{\frac{k}{2\rho \log n}\cdot 2\rho \log n} \leq e^{-2\rho \cdot \log n} = n^{-2\rho}.$$

Thus, the event $B := \bigcup_{i\in[w]} \left\{Y_i < 1 - \frac{2\rho \log n}{C_1 k}\right\}$ has probability $\mathbb{P}[B] \leq \sum_{i\in[w]} n^{-2\rho} \leq wn^{-2\rho}$. It follows that

$$\mathbb{E}\left[\min_{i\in[w]} Y_i\right] = \mathbb{E}\left[\min_{i\in[w]} Y_i \mid \bar{B}\right] \cdot \mathbb{P}\left[\bar{B}\right] \geq \left(1 - \frac{2\rho \log n}{C_1 k}\right) \cdot \left(1 - wn^{-2\rho}\right) \geq 1 - \frac{3\rho \log n}{C_1 k},$$

where the last inequality follows since $wn^{-2\rho} \leq \frac{2\rho \log n}{C_1 k}$ for any sufficiently large n. Therefore, $\mathbb{E}\left[\max_{i\in[w]} \Delta_{t,i}\right] \leq 1 - \mathbb{E}\left[\min_{i\in[w]} Y_i\right] \leq \frac{3\rho \log n}{C_1 k}$. □

Now we are ready to prove the regret-to-loss conversion formula.

Proof of Proposition 4.8. Recall that k' is the size of the resampled subset of each set A_t of arriving arms. For any fixed $T \geq w$, by re-arranging the terms in (9), we obtain

$$\begin{aligned}
\sum_{t=1}^{T} L_t^{\text{int}} &= \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=0}^{\ell} \sum_{a\in A_{t-j}} \frac{\pi_t(a)}{n} \cdot \left(\mu_{\max}\left(A_{t-j}\right) - \mu_a\right)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=0}^{\ell} \sum_{a\in A_{t-j}} \frac{\pi_t(a)}{n} \cdot \left[\left(\mu_{\max}\left(A_{t-j}\right) - \mu_{\max}\left(A'_{t-j}\right)\right) + \left(\mu_{\max}\left(A'_{t-j}\right) - \mu_a\right)\right]\right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[\max_{0\leq j\leq \ell} \left|\mu_{\max}\left(A'_{t-j}\right) - \mu_{\max}\left(A_{t-j}\right)\right| + \frac{1}{n} \cdot \sum_{j=0}^{\ell} \sum_{a\in A_{t-j}} \pi_t(a) \cdot \left(\mu_{\max}\left(A'_{t-j}\right) - \mu_a\right)\right] \qquad (10)
\end{aligned}$$

where the inequality follows since for any t, the total number of arms the policy selects satisfies $\sum_{j=0}^{\ell} \sum_{a\in A_{t-j}} \pi_t(a) = n$. Note that $\mathbb{E}\left[\max_{0\leq j\leq \ell} \left|\mu_{\max}\left(A'_{t-j}\right) - \mu_{\max}\left(A_{t-j}\right)\right|\right] \leq \frac{1}{k'}$, so

$$(10) \leq \frac{T}{k'} + \sum_{t=\ell}^{T-\ell} \mathbb{E}\left[\sum_{a\in A_t} \frac{1}{n} \cdot \sum_{j=0}^{\ell} \pi_{t+j}(a) \cdot \left(\mu_{\max}(A_t) - \mu_a\right)\right] + 2\ell \leq \frac{T}{k'} + (T - 2\ell) \cdot R(n, k') + 2\ell,$$

where the 2ℓ term in the first inequality is because we are summing from ℓ to $T - \ell$. It follows that

$$\begin{aligned}
\text{Loss}_n(\pi) &\leq \varlimsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left(L_t^{\text{int}} + L_t^{\text{ext}}\right) \\
&\leq \frac{1}{k'} + \frac{3 \log k}{C_1 \rho k} + \varlimsup_{T\to\infty} \frac{(T - 2\ell) \cdot R(n, k') + 2\ell}{T} \\
&= \frac{1}{k'} + \frac{3 \log k}{C_1 \rho k} + R(n, k').
\end{aligned}$$

where the last identity follows since $R(n, k')$ does not depend on T, and $\ell = O(1)$ as $T \to \infty$. □

C. Proof of Theorem 4.6: Average Regret of BSE

In this section we prove Theorem 4.10, which says that the average regret is $\tilde{O}((k/n)^{\ell/(\ell+2)})$ for the BSE algorithm with the revised geometric grid, specified in Definition 4.4. We illustrate the key ideas by considering adaptivity levels $\ell = 1$ and 2 as warm-up first in Section C.2 and C.2 respectively, and then in Section C.4 we present the proof for general $\ell \geq 1$.

C.1. Preliminaries

We introduce some tools for our analysis. For any arm a, consider i.i.d. Bernoulli rewards $(Z^a_{i,j})_{i\in[w],j\in[n]}$ with mean μ_a. We first state a standard concentration bound for independent random variables, see e.g., (Vershynin, 2018).

Lemma C.1 (Concentration Bounds)**.** *Let $Z_1, ..., Z_m$ be independent random variables supported on* $[0, 1]$*, and $\bar{Z} = \frac{1}{m}\sum_{i=1}^{m} Z_i$, then for any $\delta > 0$, it holds that*

$$\mathbb{P}(|\bar{Z} - \mathbb{E}(\bar{Z})| > \delta) \le \exp\left(-\frac{\delta^2 m}{2}\right) \qquad \text{(Hoeffding's inequality)},$$

and

$$\mathbb{P}\left(|\bar{Z} - \mathbb{E}(\bar{Z})| > \delta \cdot \mathbb{E}(\bar{Z})\right) \le \exp\left(-\frac{\mathbb{E}\left(\bar{Z}\right)}{2}\delta^2 m\right) \qquad \text{(Chernoff's inequality)}.$$

Consider the event that the empirical mean rewards of all available arms available at time t satisfy Hoeffding's inequality.

Definition C.2 (Clean Event)**.** For any integers i, m and arm a, define

$$\mathcal{C}_a^{i,m} = \left\{ \Big| \sum_{j=1}^{m} Z^a_{i,j} - m\mu_a \Big| \le \sqrt{2m \cdot (\rho + 3) \cdot \log n} \right\}.$$

We define the *clean event* as $\mathcal{C}_t = \bigcap \mathcal{C}_a^{i,m}$ where the intersection is over $m \in [n], i \in [w]$ and $a \in A^t_{t-w}$.

We use Lemma C.1 to show that $\mathcal{C}_t$ occurs with high probability in each round t.

Lemma C.3 (Clean Event is Likely)**.** *For any time t, the clean event satisfies $\mathbb{P}\left[\overline{\mathcal{C}}_t\right] \le n^{-1}$.*

Proof. Fix an arbitrary $i \in [w]$ and arm $a \in A^t_{t-w}$. Then for any integer m, by Hoeffding's inequality,

$$\mathbb{P}\left[\overline{\mathcal{C}_a^{i,m}}\right] = \mathbb{P}\left[\Big|\sum_{j=1}^{m} Z^a_{i,j} - m\mu_a\Big| > \sqrt{2m(\rho+3)\log n}\right] \le \exp\left(-\frac{1}{2m} \cdot 2m(\rho+3)\log n\right) = n^{-3}.$$

By the union bound over all $a \in [k]$, $i \in [w]$ and $m \in [n]$, we have

$$\begin{aligned}
\mathbb{P}\left[\bigcup_{\substack{m\in[n],i\in[w],\\ a\in A^t_{t-w}}} \overline{\mathcal{C}_a^{i,m}}\right] &\le \sum_{\substack{m\in[n],i\in[w],\\ a\in A^t_{t-w}}} \mathbb{P}\left[\overline{\mathcal{C}_a^{i,m}}\right] \\
&= \sum_{\substack{m\in[n],i\in[w],\\ a\in A^t_{t-w}}} \mathbb{P}\left[\Big|\sum_{j=1}^{m} Z^a_{i,j} - m\mu_a\Big| > \sqrt{2m(\rho+3)\log n}\right] \\
&\le nwk \cdot n^{-\rho-3} = w \cdot n^{-2} \le n^{-1}.
\end{aligned}$$

□

We will also repeatedly apply the following simple fact.

Lemma C.4 (3Δ-Inequality)**.** *Let $\Delta > 0$ and $X = \{x_j\}_{j\in[k]}, X' = \{x'_j\}_{j\in[k]}$ be two sets of real numbers such that $|x_j - x'_j| \le \Delta$ for all $j \in [k]$. Suppose for some i we have $x'_i \ge \max X' - \Delta$. Then, $x_i \ge \max X - 3\Delta$.*

Proof. Since $|x_j - x'_j| \le \Delta$ for all $j \in [k]$, we have $|\max X - \max X'| \le \Delta$. Hence, $x'_i \ge \max X' - \Delta \ge \max X - 2\Delta$. Therefore, $x_i \ge x'_i - \Delta \ge \max X - 2\Delta - \Delta = \max X - 3\Delta$. □

C.2. Warm-up: $\ell = 1$

In this subsection we show that the average regret is $\tilde{O}\left((k/n)^{1/3}\right)$ for the BSE algorithm with grid $\varepsilon^{\star}_{0,1}$. We first prove that w.h.p. all surviving arms are close to optimal. Recall from Algorithm 1 that S_1 denotes the surviving arms at the end of the first phase.

Lemma C.5 (Surviving Arms Have High Rewards)**.** *If the clean event $\mathcal{C}$ occurs, then $\mu_{\max}(A) - \mu_{\min}(S_1) \le \delta_1$ where $\delta_1 = 3\left(\frac{\varepsilon_0 n}{k}\right)^{-1/2} \log^{1/2} n$.*

Proof. By Hoeffding's inequality (see Lemma C.1), the deviation of the empirical mean reward of each arm satisfies $|\hat{\mu}_a - \mu_a| \le \frac{1}{3}\delta_1$ for all arms $a \in A$. Consider $\hat{a} = \arg\max_a \hat{\mu}_a$ and $a^* = \arg\max_a \mu_a$. By definition of S_1 (see Algorithm 1), an arm a survives *only if* $|\hat{\mu}_a - \hat{\mu}_{\hat{a}}| \le \frac{1}{3}\delta_1$. Thus, by the 3Δ-inequality (Lemma C.4) we have $\mu_a \ge \mu_{\max}(A) - \delta_1$. □

Combining the above with the regret decomposition (Lemma B.2), we obtain the following.

Proposition C.6 (Average Regret of $\mathrm{BSE}^{\star}_1$)**.** *Suppose $\rho \le 1$. Then the average regret of $\mathrm{BSE}^{\star}_1$, the BSE algorithm with exploration intensities $\varepsilon^{\star}_{0,1}$, satisfies*

$$\mathrm{Reg}^{\mathrm{avg}}_n(\mathrm{BSE}^{\star}_1) \le 5\left(\frac{k}{n}\right)^{1/3} \log^{1/3} n.$$

Proof. To suppress notations, write $\varepsilon_0 := \varepsilon^{\star}_{0;1}$. By Lemma C.5, we have $\mu_{\max}(A) - \mu_{\min}(S_1) \le \delta_1$, and thus

$$\begin{aligned}\mathrm{Reg}^{\mathrm{avg}}_n(\mathrm{BSE}^{\star}_1) &\le \varepsilon_0 + (1-\varepsilon_0)\cdot \mathbb{E}\left[\mu_{\max}(A) - \mu_{\min}(S_1)\right] \\ &\le \varepsilon_0 + \delta_1 \cdot \mathbb{P}\left[\mathcal{C}\right] + \mathbb{P}\left[\overline{\mathcal{C}}\right] \\ &\le \varepsilon_0 + \delta_1 + n^{-1},\end{aligned}$$

where the last inequality follows from Lemma C.3. Expanding δ_1 using Lemma C.5 and recalling that $\varepsilon^{\star}_{0,1} = \left(\frac{k}{n}\right)^{1/3} \log^{1/3} n$, we have

$$\mathrm{Reg}^{\mathrm{avg}}_n(\mathrm{BSE}^{\star}_1) \le 4\left(\frac{k}{n}\right)^{\frac{1}{3}} \log^{\frac{1}{3}} n + n^{-1} \le 5\left(\frac{k}{n}\right)^{\frac{1}{3}} \log^{\frac{1}{3}} n,$$

where the last inequality follows since $\frac{1}{n} \le \frac{k}{n} \le \left(\frac{k}{n}\right)^{1/3}$. □

C.3. Bounding the Number of Survivors: Analysis for $\ell = 2$

Recall that for adaptivity level $\ell = 2$, we first select a batch of arms and compute a subset S_1 of surviving arms whose confidence intervals are not dominated by any other arm. Then we select another batch of arms, and compute a further subset $S_2 \subseteq S_1$ in a similar manner. Finally, in the exploitation phase, we choose an arbitrary arm from S_2.

The key step is bounding the number of survivors after the first phase - if we can upper-bound S_1's cardinality (w.h.p.), then we can lower-bound the number of times each arm in S_1 is selected in phase 2, leading to a guarantee on the width of the confidence interval. To this goal, consider the following *uniform event* that μ_a's are distributed approximately uniformly.

Definition C.7 (Uniform Event)**.** Consider any constant $\delta \in (0,1)$, and the number N_δ of arms in A that lie in a δ-neighborhood of the optimal arm, formally, $N_\delta = \left|\{a \in A : \mu_a \ge \mu_{\max}(A) - \delta\}\right|$. We define the *uniform event* as

$$U_\delta = \left\{\frac{1}{2}C_1\delta k \le N_\delta \le \frac{3}{2}C_2\delta k\right\}.$$

We show that the uniform event is likely to occur when k is large. This is a direct consequence of the RIM and Assumption 2.1, and its proof is similar to that of Proposition B.3.

Lemma C.8 (Uniform Event is Likely)**.** *Suppose $\delta k \ge \frac{8}{C_1}\log n$. Then for any $\delta \in (0,1)$ and round t, the uniform event satisfies $\mathbb{P}\left[\overline{U_\delta}\right] \le 2n^{-1}$.*

Proof. Index the arms from 1 to k, and denote by f the density of D, the distribution from which the reward rates μ_i's are drawn. We first upper bound $\mathbb{P}[\bar{G} \mid \mu_{\max} = 1-\gamma]$ for any fixed $\gamma \in (0,1)$.

We subsequently fix an arbitrary $i \in [k]$. Denote by $Z_i = \mathbb{1}[\mu_i \geq \mu_{max} - \delta]$ and consider the largest reward rate $\mu_{-i}^{\max}$ in $[k]\backslash\{i\}$, i.e.,

$$\mu_{-i}^{\max} = \max\{\mu_1, \dots, \mu_{i-1}, \mu_{i+1}, \dots, \mu_k\}.$$

Observe that

$$\mathbb{P}[\, Z_i = 1 \mid \mu_{\max} = 1-\gamma\,] = \mathbb{P}\left[\mu_i \in [1-\gamma-\delta,\ 1-\gamma] \mid \mu_{-i}^{\max} - \delta\right] = \int_{1-\gamma-\delta}^{1-\gamma} f(z)\, dz,$$

where the last identity follows since μ_i's are independent, in particular, $\mu_{-i}^{\max}$ and μ_i are independent. Recall by Assumption 2.1 that the density satisfies $C_1 \leq f(z) \leq C_2$, so

$$C_1\delta \ \leq\ \mathbb{P}[\, Z_i = 1 \mid \mu_{\max} = 1-\gamma\,] \ \leq\ C_2\delta.$$

Note that conditional on the event $\{\mu_{\max} = 1-\gamma\}$, the random variables Z_i are still i.i.d., so by Chernoff's inequality (see Lemma C.1) we obtain

$$\begin{aligned}
\mathbb{P}\left[\bar{G} \,\middle|\, \mu_{\max} = 1-\gamma\right] &\leq \mathbb{P}\left[\sum_{i\in[k]} Z_i > \frac{3}{2}\cdot C_2\delta k \,\middle|\, \mu_{\max} = 1-\gamma\right] + \mathbb{P}\left[\sum_{i\in[k]} Z_i < \frac{1}{2}\cdot C_1\delta k \,\middle|\, \mu_{\max} = 1-\gamma\right]\\
&\leq 2\exp\left(-\frac{1}{2}\cdot\left(\frac{1}{2}\right)^2\cdot C_1\delta k\right)\\
&\leq 2\exp\left(-\frac{C_1}{8}\cdot\frac{8}{C_1}\log n\right)\\
&= 2k^{-\frac{C_1}{8}\cdot\frac{8}{C_1\rho}} = 2n^{-1}.
\end{aligned}$$

Therefore,

$$\mathbb{P}[\bar{G}] = \mathbb{E}_\gamma\left[\mathbb{P}[\bar{G} \mid \mu_{\max} = 1-\gamma]\right] \leq 2n^{-1}.$$

□

Assuming that U_{δ_1} and $\mathcal{C}$ both occur, we have $|S_1| \sim \delta_1 k$. Thus, in the second phase, each surviving arm is selected $\gtrsim \frac{\varepsilon_1 n}{\delta_1 k}$ times, and hence the confidence intervals have widths $\lesssim \left(\frac{\varepsilon_1 n}{\delta_1 k}\right)^{-1/2}$. We next make these ideas precise.

Definition C.9 (Widths of Confidence Intervals)**.** For any $\varepsilon_0, \varepsilon_1 \in (0,1)$, define

$$\delta_1 = \delta_1(\varepsilon_0) = \frac{8}{C_1}\varepsilon_0^{-\frac{1}{2}}\left(\frac{k}{n}\right)^{\frac{1}{2}}\log n \quad\text{and}\quad \delta_2 = \delta_2(\varepsilon_0,\varepsilon_1) = 6^{\frac{3}{2}}\cdot\sqrt{(\rho+3)\,\frac{C_2}{C_1}}\cdot\varepsilon_0^{-\frac{1}{4}}\varepsilon_1^{-\frac{1}{2}}\left(\frac{k}{n}\right)^{\frac{3}{4}}\log^{\frac{5}{4}} n.$$

For any adaptivity level ℓ, denote by $\delta_{j,\ell}^\star = \delta_{j,\ell}(\varepsilon_0^\star, \varepsilon_1^\star)$ for $j = 1, \dots, \ell$.

Under the above notations, we can bound the suboptimality of S_2, the surviving arms after the second phase as follows.

Lemma C.10 (Suboptimality of S_2)**.** *Consider the algorithm* $\mathrm{BSE}_2(\varepsilon_0, \varepsilon_1)$ *such that* $\delta_1 = \delta_1(\varepsilon_0,\varepsilon_1)$ *satisfies* $\delta_1 k > \frac{8}{C_1}\log n$*. If the clean event* $\mathcal{C}$ *and the uniform event* U_{δ_1} *both occur, then*

$$\mu_{\max}(A) - \mu_{\min}(S_2) \leq \delta_2.$$

Proof. We first upper bound the cardinality of S_1. By Lemma C.5, since $\mathcal{C}$ occurs, we have $\mu_{\max}(A) - \mu_{\min}(S_1) \leq \delta_1$. In other words, for an arm $a \in A$ to survive the first phase, its mean reward needs to be δ_1-close to $\mu_{\max}(A)$. Since the uniform event U_{δ_1} occurs, by Lemma C.8 the number of such arms can be bounded as

$$|S_1| \leq \frac{3}{2}C_2\delta_1 k. \tag{11}$$

Note that in phase 2, each arm in S_1 is selected $m_1 := \frac{\varepsilon_1 n}{|S_1|}$ times. By definition of the clean event $\mathcal{C}$, the empirical mean reward of every arm deviates from the mean by at most $\sqrt{2(\rho+3)} \cdot m_1^{-1/2} \log^{1/2} n$. Thus by the 3Δ-inequality (Lemma C.4), we can bound the suboptimality of arms in S_2 as

$$\mu_{\max}(A) - \mu_{\min}(S_2) \leq 3\sqrt{2(\rho+3)} \cdot m_1^{-1/2} \log^{1/2} n.$$

To further bound the above, we use (11) to lower bound $m_1 = \frac{\varepsilon_1 n}{|S_1|}$. Specifically,

$$\begin{aligned} &\mu_{\max}(A) - \mu_{\min}(S_2) \\ &\leq 3\sqrt{2(\rho+3)} \cdot \left(\frac{\varepsilon_1 n}{\frac{3}{2} C_2 \delta_1 k}\right)^{-\frac{1}{2}} \log^{\frac{1}{2}} n \\ &\leq 3^{\frac{3}{2}} \sqrt{(\rho+3) C_2} \cdot \varepsilon_1^{-\frac{1}{2}} \delta_1^{\frac{1}{2}} \left(\frac{k}{n}\right)^{\frac{1}{2}} \log^{\frac{1}{4}} k \cdot \log^{\frac{1}{2}} n. \end{aligned} \tag{12}$$

Recall from Definition C.9 that

$$\delta_1 = \frac{8}{C_1} \varepsilon_0^{-\frac{1}{2}} \left(\frac{k}{n}\right)^{\frac{1}{2}} \log n \quad \text{and} \quad \delta_2 = \delta_2(\varepsilon_0, \varepsilon_1) = 6^{\frac{3}{2}} \cdot \sqrt{(\rho+3)\frac{C_2}{C_1}} \cdot \varepsilon_0^{-\frac{1}{4}} \varepsilon_1^{-\frac{1}{2}} \left(\frac{k}{n}\right)^{\frac{3}{4}} \log^{\frac{5}{4}} n,$$

we can simplify the above inequality as

$$(12) \leq 6^{\frac{3}{2}} \cdot \sqrt{(\rho+3)\frac{C_2}{C_1}} \cdot \varepsilon_1^{-\frac{1}{2}} \varepsilon_0^{-\frac{1}{4}} \left(\frac{k}{n}\right)^{\frac{3}{4}} \log^{\frac{1}{4}} k \cdot \log n = \delta_2.$$

□

We can now bound the average regret of the BSE algorithm for any grid as follows.

Proposition C.11 (Average Regret of BSE with Arbitrary Grid, $\ell = 2$)**.** *Suppose* $0 < \varepsilon_0 < \varepsilon_1 < 1$ *and* $\delta_1 k > \frac{8 \log n}{C_1}$*. Then, the average regret of the BSE algorithm with grid size* $\varepsilon_0, \varepsilon_1$ *satisfies*

$$\mathrm{Reg}_n^{\mathrm{avg}}(\mathrm{BSE}_{\ell=2}(\varepsilon_0, \varepsilon_1)) \leq \varepsilon_0 + \varepsilon_1 \delta_1 + \delta_2 + O\left(n^{-1}\right).$$

Proof. By definition of regret, we have

$$\mathrm{Reg}_n^{\mathrm{avg}}(\mathrm{BSE}_{\ell}(\varepsilon_0, \varepsilon_1)) = \varepsilon_0 + \varepsilon_1 \cdot \mathbb{E}[\mu_{\max}(A) - \mu_{\min}(S_1)] + (1 - \varepsilon_0 - \varepsilon_1) \cdot \mathbb{E}[\mu_{\max}(A) - \mu_{\min}(S_2)].$$

We bound each term separatly. By Lemma C.3 and Lemma C.5, the second term can be bounded as

$$\begin{aligned} &\mathbb{E}[\mu_{\max}(A) - \mu_{\min}(S_1)] \\ &= \mathbb{E}[\mu_{\max}(A) - \mu_{\min}(S_1) \mid \mathcal{C}] \cdot \mathbb{P}[\mathcal{C}] + \mathbb{E}\left[\mu_{\max}(A) - \mu_{\min}(S_1) \mid \overline{\mathcal{C}}\right] \cdot \mathbb{P}[\overline{\mathcal{C}}] \\ &\leq \delta_1 \cdot 1 + 2n^{-1}. \end{aligned}$$

By Lemma C.10, if the events U_{δ_1} and $\mathcal{C}$ both occur, then $\mu_{\max}(A) - \mu_{\min}(S_2) \leq \delta_2$. Further, by Lemma C.8, we have $\mathbb{P}[U_{\delta_1}] \leq 2n^{-1}$ whenever $\delta_1 k \geq \frac{8 \log n}{C_1}$. Combining the above facts, we obtain

$$\mathrm{Reg}_n^{\mathrm{avg}}(\mathrm{BSE}_{\ell}(\varepsilon_0, \varepsilon_1)) \leq \varepsilon_0 + \varepsilon_1 \cdot \left(\delta_1 + 2n^{-1}\right) + \delta_2 + n^{-1},$$

and the proof is complete. □

The revised geometric grids in Definition 4.4 are motivated by minimizing the above bound. Since we are focusing on $\ell = 2$, we will suppress ℓ and write $\varepsilon_i^\star = \varepsilon_{i,\ell}^\star$ for $i \leq \ell = 2$. Choose ε_0 so that

$$\varepsilon_0 = \varepsilon_1 \cdot \delta_1(\varepsilon_0) = \delta_2(\varepsilon_0, \varepsilon_1),$$

then we have

$$\varepsilon_0 \sim \left(\frac{k}{n}\right)^{\frac{1}{2}} \quad \text{and} \quad \varepsilon_1 \sim \left(\frac{k}{n}\right)^{\frac{1}{4}},$$

as specified in Definition 4.4. We obtain the following guarantee for the revised geometric grid by choosing $\varepsilon_i = \varepsilon_i^\star$ in Proposition C.11. The proof is straightforward – we only need to verify that $\delta_1^\star k \geq \frac{8 \log n}{C_1}$.

Corollary C.12 (Average Regret of $\mathrm{BSE}_2^\star$)**.** *If $\rho \geq \theta_2$, then* $\mathrm{Reg}_n\left(\mathrm{BSE}_{\ell=2}^\star\right) \leq \tilde{O}\left(\left(\frac{k}{n}\right)^{1/2}\right)$.

Proof. Recall from Definition 4.4 that $\varepsilon_0^\star = \left(\frac{k}{n}\right)^{\frac{1}{2}}$, and from Definition C.9 that for any ε_0 we defined

$$\delta_1 = \delta_1(\varepsilon_0) = \frac{8}{C_1}\varepsilon_0^{-\frac{1}{2}}\left(\frac{k}{n}\right)^{\frac{1}{2}}\log n.$$

Expanding the expressions for $\varepsilon_0^\star$ and $\delta_1^\star = \delta_1(\varepsilon_0^\star)$, we have

$$\delta_1^\star k \geq \frac{8}{C_1}\varepsilon_0^{-\frac{1}{2}}\left(\frac{k}{n}\right)^{\frac{1}{2}}\log n \cdot k = \frac{8}{C_1}\cdot k^{\frac{5}{4}}\cdot n^{-\frac{1}{4}}\cdot \log n \geq \frac{8}{C_1}\log n,$$

where the last inequality follows since $\rho \geq \theta_2 = \frac{1}{5}$. By Proposition C.11, we conclude that

$$\mathrm{Reg}_n^{\mathrm{avg}}\left(\mathrm{BSE}_{\ell=2}^\star\right) \leq \varepsilon_0^\star + \varepsilon_1^\star\delta_1^\star + \delta_2^\star + O\left(n^{-1}\right) = O\left(\left(\frac{k}{n}\right)^{\frac{1}{2}}\log^{\frac{5}{4}} n\right).$$

□

C.4. Proof of Theorem 4.6

We now extend the analysis to the general ℓ case. For each available arm, an ℓ-Layer Sieve policy will recursively explore its reward rate and derive a series of confidence intervals of the following widths.

Definition C.13 (Width of Confidence Interval)**.** Given $\varepsilon_0, \ldots, \varepsilon_{\ell-1}$, for each $i = 1, 2, \ldots, \ell$, we define

$$\delta_i = \delta_i(\varepsilon_0, \ldots, \varepsilon_{i-1}) = 3\cdot 15^{i-1}\cdot C_2^{\frac{i-1}{2}}\cdot \varepsilon_0^{-2^{-i}}\cdot \varepsilon_1^{-2^{-(i-1)}}\cdot \ldots \cdot \varepsilon_{i-1}^{-\frac{1}{2}}\cdot \left(\frac{k}{n}\right)^{1-2^{-i}}\cdot \log^{1+\frac{i-1}{4}} n.$$

It is straightforward to verify that the above δ_i's satisfy the following recursion. As in the analysis for $\ell = 2$ in Section C.2, we will show that the loss on the j-th layer is bounded as follows.

Lemma C.14 (Regret on the j-th Layer)**.** *Fix integers t, j and suppose the events $G_{t-1}^{\delta_1}, \ldots, G_{t-j}^{\delta_j}$ and $\mathcal{C}_t$ occur. Then,*

$$\mu_{\max}(A_{t-j}) - \mu_{\min}(S_{t-j}^j) \leq \delta_j \quad \textit{and} \quad |S_{t-j}^j| \leq 6C_2\delta_j k\cdot \log^{1/2} k.$$

Proof. Proof. We show this inductively on j. To show the base case $j = 1$, note that by Lemma C.5, we have $\mu_{\max}(A_{t-1}) - \mu_{\min}(S_{t-1}^1) \leq \delta_1$. Moreover, we showed in the first paragraph in the proof of Lemma **??** that $|S_{t-1}^1| \leq 6C_2\delta_1 k\log^{1/2} k$, and thus the claim holds for $j = 1$.

Now consider $j \geq 2$. As the *induction hypothesis*, assume the claim holds for $1, \ldots, j-1$. Then, $|S_{t-(j-1)}^{j-1}| \leq 6C_2\delta_{j-1}k\log^{1/2} k$. Consequently, each arm in $S_{t-(j-1)}^{j-1}$ is selected

$$m_{j-1} = \frac{\varepsilon_j n}{|S_{t-(j-1)}^{j-1}|} \geq \frac{\varepsilon_j n}{6C_2\delta_{j-1}k\cdot\log^{1/2} k} \tag{13}$$

times. Since the clean event $\mathcal{C}_t$ occurs, the empirical mean of each arm from $S_{t-(j-1)}^{j-1}$ deviates from the mean by $3m_{j-1}^{1/2}\log^{1/2} n$. Thus, if an arm $a \in A_{t-j}$ survives, we have

$$\mu_{\max}\left(A_{t-(j-1)}\right) - \mu_{\min}\left(S_{t-(j-1)}^{j-1}\right) \leq 9m_{j-1}^{-\frac{1}{2}}\log^{\frac{1}{2}} m_{j-1}. \tag{14}$$

Since the good event $G_{t-j}^{\delta_j}$ occurs, we have

$$|S_{t-j}^j| \leq 6C_2\cdot\left(3m_{j-1}^{-\frac{1}{2}}\log^{\frac{1}{2}} m_{j-1}\right)k\log^{1/2} k.$$

To simplify, note that by (13),

$$3m_{j-1}^{-\frac{1}{2}}\log^{\frac{1}{2}} m_{j-1} \le 3\cdot\left(\frac{\varepsilon_j n}{6C_2\delta_{j-1}k\cdot\log^{1/2}k}\right)^{-\frac{1}{2}}\log^{\frac{1}{2}}\left(\frac{\varepsilon_j n}{6C_2\delta_{j-1}k\cdot\log^{1/2}k}\right)\le\delta_j,$$

and hence $|S_{t-j}^j| \le 6C_2\delta_j k\log^{1/2}k$.

To simplify (14), we expand m_{j-1} again and obtain

$$\begin{aligned}
&\mu_{\max}\left(A_{t-(j-1)}\right)-\mu_{\min}\left(S_{t-(j-1)}^{j-1}\right)\\
&\le 3m_{j-1}^{-\frac{1}{2}}\log^{\frac{1}{2}} m_{j-1}\\
&\le 3\left(\frac{\varepsilon_j n}{6C_2\delta_{j-1}k\cdot\log^{1/2}k}\right)^{-\frac{1}{2}}\log^{\frac{1}{2}} n\\
&\le 9\sqrt{C_2}\cdot\varepsilon_j^{-\frac{1}{2}}\delta_{j-1}^{\frac{1}{2}}\cdot\left(\frac{k}{n}\right)^{\frac{1}{2}}\cdot\log^{\frac{1}{4}}k\cdot\log^{\frac{1}{2}}n, \qquad (15)
\end{aligned}$$

where the last inequality relies on $\delta_{j-1}k > 1$. By Definition C.13, we have

$$\delta_{j-1} = 3\cdot 5^{j-2}\cdot\varepsilon_0^{-2^{-(j-1)}}\cdot\varepsilon_1^{-2^{-(j-2)}}\cdots\varepsilon_{j-1}^{-2^{-1}}\cdot\left(\frac{k}{n}\right)^{1-2^{-(j-1)}}\log^{1+\frac{j-2}{4}}n.$$

Substituting into (15), we conclude that

$$\mu_{\max}\left(A_{t-(j-1)}\right)-\mu_{\min}\left(S_{t-(j-1)}^{j-1}\right)\le 3\cdot 5^{j-1}C_2^{\frac{j-1}{2}}\cdot\varepsilon_0^{-2^{-j}}\cdot\varepsilon_1^{-2^{-(j-1)}}\cdots\varepsilon_j^{-2^{-1}}\left(\frac{k}{n}\right)^{1-2^{-j}}\log^{1+\frac{j-1}{4}}n\le\delta_2. \qquad \square$$

We now have the following regret bound for BSE algorithm with adaptivity ℓ.

Proposition C.15 (Average Regret of BSE, Arbitrary Grid)**.** *Consider arbitrary adaptive level ℓ and grid sizes $0<\varepsilon_0<\cdots<\varepsilon_{\ell-1}<1$ and $\delta_i k>1$ for any $i\le\ell-1$. Then for any round t,*

$$\mathrm{Reg}_n^{\mathrm{avg}}\left(\mathrm{BSE}_\ell\left(\varepsilon_0,\cdots,\varepsilon_{\ell-1}\right)\right)=\tilde{O}\left(\varepsilon_0+\varepsilon_1\delta_1+\varepsilon_2\delta_2+\cdots+\varepsilon_{\ell-1}\delta_{\ell-1}+\delta_\ell+n^{-\rho}\right).$$

Proof. Proof. By definition of internal regret,

$$\mathrm{Reg}_n^{\mathrm{avg}}\left(\mathrm{BSE}_\ell\left(\varepsilon_0,\cdots,\varepsilon_{\ell-1}\right)\right)\le\sum_{j=0}^{\ell-1}\varepsilon_j\cdot\mathbb{E}\left[\mu_{\max}(A_{t-j})-\mu_{\min}(S_{t-j}^j)\right]+\mathbb{E}\left[\max_{j\in[w]}\left\{\mu_{\max}\left(A_{t-j}\right)-\mu_{\min}\left(S_{t-j}^2\right)\right\}\right]. \qquad (16)$$

Note that if the events $\mathcal{C}_t$ and $G_{t-j}^{\delta_j}$ occur, $j=0,\ldots,\ell-1$, then by Lemma C.14, we have

$$\mu_{\max}(A_{t-j})-\mu_{\min}(S_{t-j}^j)\le\delta_j \quad \text{for } j=0,\ldots,\ell-1,$$

and

$$\max_{\ell\le j\le w}\left\{\mu_{\max}(A_{t-j})-\mu_{\min}(S_{t-j}^2)\right\}\le\delta_2.$$

Further, by Lemma C.8 that for each j we have $\mathbb{P}\left[G_{t-j}^{\delta_j}\right]\le k^{-2}$ provided $\delta_i k\ge 1$ for all $j=0,\ldots,\ell-1$, so

$$\begin{aligned}
&\mathbb{E}\left[\mu_{\max}(A_{t-j})-\mu_{\min}(S_{t-j}^j)\right]\\
&\le\delta_j\cdot\mathbb{P}\left[\mathcal{C}_t\cap\bigcap_{\ell\le j\le w}G_{t-j}^{\delta_j}\right]+\sum_{\ell\le j\le w}\mathbb{P}\left[\overline{G_{t-j}^{\delta_j}}\right]+\mathbb{P}\left[\overline{\mathcal{C}_t}\right]\\
&\le\delta_j\cdot 1+n^{-2+\rho}\cdot w.
\end{aligned}$$

Similarly,

$$\begin{aligned}&\mathbb{E}\left[\max_{\ell\le j\le w}\left\{\mu_{\max}(A_{t-j})-\mu_{\min}(S^2_{t-j})\right\}\right]\\ &\le \delta_\ell\cdot\mathbb{P}\left[\mathcal{C}_t\cap\bigcap_{\ell\le j\le w}G^{3\delta_1}_{t-j}\right]+\sum_{\ell\le j\le w}\mathbb{P}\left[\overline{G^{\delta_1}_{t-j}}\right]+\mathbb{P}\left[\overline{\mathcal{C}_t}\right]\\ &\le \delta_\ell+k^{-2}w+n^{-2+\rho}w.\end{aligned}$$

When $\rho\le 1$, the above is bounded by $\delta_2+n^{-2\rho}w+o\left(\frac{1}{n}\right)$. Combining, we conclude that

$$\begin{aligned}\operatorname{Reg}^{\text{avg}}_n\left(\text{BSE}_\ell\left(\varepsilon_0,\cdots,\varepsilon_{\ell-1}\right)\right)&\le\varepsilon_0+\sum_{j=1}^{\ell}\varepsilon_j\left(\delta_j+n^{-3}\right)+n^{-2\rho}w+o\left(\frac{1}{n}\right)\\ &\le\varepsilon_0+\sum_{j=1}^{\ell}\varepsilon_j\delta_j+n^{-\rho}+o\left(\frac{1}{n}\right).\end{aligned}$$

□

D. Proof of Proposition 4.9

To derive the upper bound for SLHVB, we need to compare the asymptotic orders of the two terms in (1). Consider the following three cases for ρ.

Case 1. Suppose $\rho\le\frac{1}{5}$. Consider $\ell=1$. By Proposition 4.9, we have

$$\operatorname{Loss}_n\left(\text{BSE}^\star_1,k\right)=\tilde{O}\left(n^{-\rho}+n^{(\rho-1)\cdot\frac{1}{3}}\right).$$

When $\rho\le\frac{1}{5}$, it holds that $n^{-\rho}\ge n^{(\rho-1)\cdot\frac{1}{3}}$, so the above becomes $\tilde{O}\left(n^{-\rho}\right)$.

Case 2. Suppose $\frac{1}{5}<\rho\le\frac{w}{2w+2}$. The claimed ℓ in Step 2 exists[11] due to the following elementary fact: for any integer $w\ge 2$,

$$\left[\frac{1}{5},\ \frac{w}{2w+2}\right]\subseteq\bigcup_{2\le\ell\le w}\left[\frac{\ell-1}{2\ell+1},\ \frac{\ell}{2\ell+2}\right]. \tag{17}$$

It then follows by Proposition 4.9 that

$$\operatorname{Loss}_n\left(\text{BSE}^\star\left(\ell,k\right)\right)=\tilde{O}\left(n^{-\rho}+n^{(\rho-1)\cdot\frac{\ell}{\ell+2}}\right).$$

Further, note that when $\rho<\frac{\ell}{2\ell+2}$, we have $n^{-\rho}>n^{(\rho-1)\cdot\frac{\ell}{\ell+2}}$, so the above becomes $\tilde{O}\left(n^{-\rho}\right)$.

Case 3. Suppose $\rho\ge\frac{w}{2w+2}=\theta_w$. Note that the threshold exponent θ_ℓ increases in ℓ, in particular, for any $\ell\le w$, we have $\theta_\ell\le\theta_w\le\rho$. It then follows from Proposition 4.9 that

$$\operatorname{Loss}_n\left(\text{BSE}^\star\left(w,k\right)\right)=\tilde{O}\left(n^{-\rho}+n^{(\rho-1)\cdot\frac{w}{w+2}}\right). \tag{18}$$

Although this is already sublinear in n, we observe that the two terms in (18) are, in general, not equal, which suggests a potential improvement. Consider the resampling size $k'=n^{\rho'}$ that renders those two terms equal, that is, $\rho'=\frac{w}{2w+2}\le\rho$. Then, due to Proposition 4.8 we obtain

$$\begin{aligned}\operatorname{Loss}_n\left(\text{BSE}^\star(w,k')\right)&=\tilde{O}\left(n^{-\rho'}+\rho^{-1}n^{-\rho}+n^{(\rho'-1)\cdot\frac{w}{w+2}}\right)\\ &\le\tilde{O}\left(n^{-\frac{w}{2w+2}}+\rho^{-1}\cdot n^{-\frac{w}{2w+2}}+n^{-\frac{w}{2w+2}}\right)\\ &=\tilde{O}\left(\rho^{-1}\cdot n^{-\frac{w}{2w+2}}\right).\end{aligned}$$

where the inequality follows since $\rho'\le\rho$. We summarize the above discussion in the following theorem.

[11] Alternatively, one can show this constructively by showing that $\ell^*=\ell^*(\rho)=\left\lfloor\frac{\rho+1}{1-2\rho}\right\rfloor$ satisfies $\theta_{\ell^*}\le\rho\le\frac{\ell^*}{2\ell^*+2}$. However, the proof - which is essentially arithmetic manipulation - is slightly tedious, so we choose not to specify this explicit form.

E. Field Experiment: Implementation Details

In this section we provide more details about the implementation of the field experiment.

E.1. Randomized BSE: A Thompson Sampling Variant

We implement a variant of semi-adaptive policy for SLHVB induced by the BSE policy (Algorithm 1) with $\ell = 1$. We modify the policy due to the following practical concerns. The first issue is the lack of knowledge of n. In our implementation, each round is set to be an hour, so n corresponds to the number of impressions per hour. But in practice, n is unknown. This can be easily fixed via randomization: for each card slot, we assign a newly-arriving card (for *exploration*) with probability ε_0 and an old card (for *exploitation*) otherwise.

Algorithm 4 SetPrior(g).

```
Input: a card $g$ and $m$ users
Output: $\hat{\alpha}, \hat{\beta}$
Randomly sample $m$ users $u_1, \ldots, u_m$
for $i = 1, \ldots, m$ do
    $\mu(u_i, g) \leftarrow$ DNN-predicted reward on $(u_i, g)$
end for
Compute the sample mean $\bar{\mu}$ and variance $\bar{v}$:
    $$\bar{\mu} \leftarrow \frac{1}{m}\sum_{i=1}^{m} \mu(u_i, g), \quad \bar{v} \leftarrow \frac{1}{m-1}\sum_{i=1}^{m} (\mu(u_i, g) - \bar{\mu})^2$$
Return
    $$\hat{\alpha} \leftarrow \bar{\mu}\left(\frac{\bar{\mu}(1-\bar{\mu})}{\bar{v}} - 1\right) \quad \text{and} \quad \hat{\beta} = \frac{1-\bar{\mu}}{\bar{\mu}}\hat{\alpha}$$
```

The second issue is the *prior information* for the rewards rates of the newly arriving cards. Although the DNN predictions are sometimes inaccurate, they do provide useful information. To utilize such information, we fit a Beta prior distribution for each card's reward rate using the *method of moments* (see e.g., (Wasserman, 2006)) based on the DNN predictions (see Algorithm 4). Specifically, recall that the DNN returns a predicted conversion rate for each **pair** of user and card. For a fixed card, denote by $\bar{\mu}, \bar{v}$ the mean and variance of the predicted conversion rates on $m = 500$ randomly sampled users. The fitted Beta prior $\mathrm{B}(\hat{\alpha}, \hat{\beta})$ is then given by

$$\hat{\alpha} = \bar{\mu}\left(\frac{\bar{\mu}(1-\bar{\mu})}{\bar{v}} - 1\right) \quad \text{and} \quad \hat{\beta} = \frac{1-\bar{\mu}}{\bar{\mu}}\hat{\alpha}.$$

The final issue is the recommended contents' *diversity*. Recall that the basic version of the BSE policy selects the empirically best arm to "exploit". However, this is not reasonable in a practical setting where such an extreme promotion of a single card is undesirable. We thus consider the Thompson Sampling version of the Sieve policy under a Beta-Bernoulli reward model; see Algorithm 5. Specifically, for each slot we draw a score for each card from its posterior. Then we assign to this slot the card with the highest score. Note that the posterior can be efficiently updated using the Bayesian rule, since the Beta distribution is a conjugate prior for the Bernoulli distribution.

E.2. Implementation Details

The firm has maintained a partition of the users into several hundreds of *buckets* for various online experiments. This partition is randomly re-generated every six months. We chose three buckets as the treatment group, involving over $600,000$ users and accounting for around 1% of the total traffic. We implemented the randomized BSE policy with $\ell = 1$ on their real system in the first 14 days of July 2021. For comparison, we also analyzed the interaction data from the first 14 days of May in the same year.

Using an offline semi-synthetic simulation, we determined the empirically optimal parameter to be around $\varepsilon_0 = 0.2$, which we used in the field experiment. This choice is also consistent with our theoretical analysis. In fact, as we recall from

Algorithm 5 Randomized One-Layer Sieve Policy

```
Input: $\varepsilon \in [0,1], \theta \geq 0$
for each hour $t = 1, 2, ...$ do
    Receive a set $A_{\text{new}}$ of new cards
    Update the set $A$ of available cards
    for each card $g \in A$ do
        if $g \in A_{\text{new}}$ then
            $(\alpha_g, \beta_g) \leftarrow \text{SetPrior}(g)$.
        else
            $n_g \leftarrow$ number of interactions of $g$ in the last hour
            $h_g \leftarrow$ number of conversions of $g$ in the last hour
            $\alpha_g \leftarrow \alpha_g + h_g,\ \beta_g \leftarrow \beta_g + n_g - h_g$
        end if
    end for
    $A_{\text{well}} \leftarrow \{g : \alpha_g + \beta_g > \theta\}$
    for each user $u$ do
        Receive the number $r_u$ of cards requested by $u$
        $A_u \leftarrow$ cards that have never been assigned to $u$
        for each card $g \in A_u$ do
            Draw $X_{u,g} \sim \text{Beta}(\alpha_g, \beta_g)$
        end for
        Sort $A_u \bigcap A_{\text{well}}$ by $X_{u,g}$ in non-increasing order as $g_1, g_2, ...$
        Sort $A_u \backslash A_{\text{well}}$ by $X_{u,g}$ in non-increasing order as $g'_1, g'_2, ...$
        $i, i' \leftarrow 1$
        for $j = 1, ...r_u$ do
            $Z_j \leftarrow \text{Ber}(\varepsilon)$
            if $Z_j = 1$ then
                $S_j \leftarrow g_i$
                $i \leftarrow i + 1$
            else
                $S_j \leftarrow g'_{i'}$
                $i' \leftarrow i' + 1$
            end if
            Send to $u$ the cards $\{S_j : j = 1, ..., r_u\}$
        end for
    end for
end for
```

Proposition C.6, the optimal parameter is $\varepsilon^\star_{0;1} \sim (k/n)^{1/3}$. In our scenario, we observed from past data that there were on average around 11 million impressions per 14 days, so the number n of impressions per **hour** is $\approx 3.2 \times 10^4$. Moreover, there are on average $k = 150$ cards released per hour, and thus $\varepsilon^\star_{0,1} \approx 0.14$.

F. Field Experiment: Analysis

We now present a detailed statistical analysis of the field experiment, including a bootstrapping hypothesis testing and a Difference-in-Differences (DID) analysis. Our analysis shows that our policy outperforms the DNN-based recommender by about 4% in total duration and nearly 7% in the total number of click-throughs per-user-per-day.

We first explain how to eliminate outliers. An outlier is typically introduced in two ways. In practice, users may accidentally swipe through two cards in a row, without even looking at the first one. We thus remove any impression with duration less than 0.2 seconds. On the other hand, users may sometimes leave their devices unattended for minutes, generating an abnormally high duration. Since most cards' content can be fully consumed within 300 seconds, we remove any impression with duration over $u = 300$ seconds.

Table 1. Types of Data In the Analysis

	CT	Duration
Per User-Day	integral	numeric
Per Impression	binary	numeric

Table 2. Overall Statistics

			May		July	
			NN	MAB	NN	MAB
Per-User-Per-Day	Duration	Mean	175.910	175.548	137.059	142.618
		SE Mean	0.699	0.659	0.6081	0.597
		Median	44.250	44.279	32.973	34.430
	#CT	Mean	1.275	1.273	0.941	1.010
		SE Mean	9.251e-03	8.814e-03	7.276e-03	7.549e-03
Per Impression	Duration	Mean	3.9697	4.0195	4.1183	4.2391
		SE Mean	4.529e-03	4.402e-03	5.738e-03	5.599e-03
		Median	0.693	0.697	0.702	0.703
	CTR	Mean	2.887e-02	2.915e-02	2.827e-02	3.001e-02
		SE Mean	4.698e-05	4.568e-05	5.804e-05	5.671e-05

F.1. Analysis of $2^2 = 4$ Metrics

The firm is interested in the analysis at the *per-user-per-day* and *per-impression* levels, and two measures for user engagement: duration and click-throughs. So altogether we have **four** metrics in total, as shown in Table 1. In the per-impression analysis, we treat each impression as an independent sample. For example, the row for per-impression duration in Table 2 is the ratio between the total duration and total number of impressions.

However, the firm's objective is the **total** user engagement rather than the *per-impression* engagement, which motivated our analysis on the per-user-per-day level. At first sight, it seems reasonable to consider the total engagement of a fixed user over **all** 14 days during the experiment. However, this metric is flawed since the frequency at which the users enter the app is affected by many external factors, such as holidays and weekends, which introduces extra noise. We will thus only consider the days when a user entered the app (i.e., has at least one impression). Formally, for each user u and day d where this user has at least one impression, we define a tuple (u, d, D_{ud}) where D_{ud} is the total duration. Thus, the number of tuples associated with each user is between 1 and 14.

Under this definition of total engagement, we summarize[12] the experimental results in Table 2 and visualize the per-user-per-day user engagement in Figure 4 and Figure 5. We observe that in May the user engagement of the two groups are approximately identical, but in July the MAB group has a significantly higher mean user engagement. Moreover, such improvement also appeared in *median* duration, indicating that this improvement is unlikely to be caused by a heavier tail in the distribution.

It is also worth noting that the user engagement per-user-per-day decreased from May to July. This is because May 2021 was when the Covid-19 pandemic reached its peak in the country where most users were located. During the lockdown, the users may have had more time to spend on the app, resulting in a higher total engagement. In Section F.2, we perform a DID analysis which incorporates the underlying change of environment across time.

Finally we emphasize that in our implementation, our policy is **not** personalized. Despite this disadvantage, our Sieve policy still outperforms their personalized DNN recommender in **all** metrics for user-engagement, both at the per-impression and per-user levels; see Table 2.

F.2. Significance Tests

From Figure 4 and Figure 5 we observe that our policy outperforms the control policy. We now test whether this improvement is statistically significant.

[12] The unit of duration in all tables is *second*

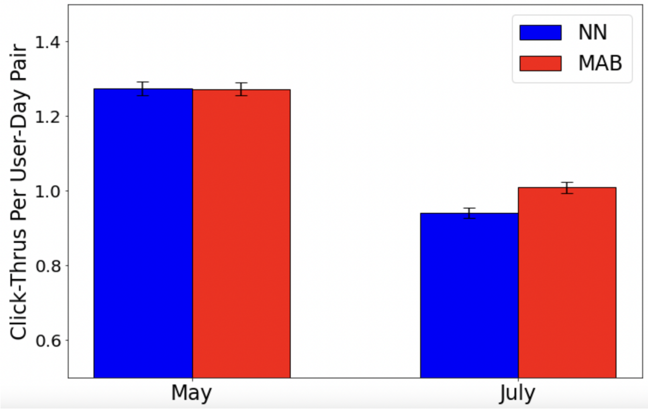


Figure 4. Number of click-throughs per-user-per-day

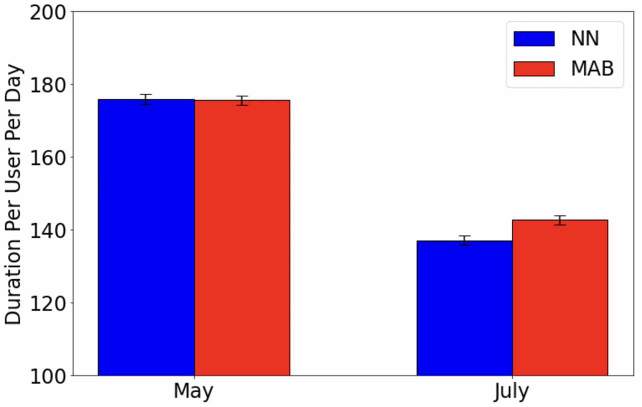


Figure 5. Duration per-user-per-day

Table 3. Significance Testing

		Basic		Bootstrap	
		Z-score	p-value	Z-score	p-value
Per-User-Per-Day	Duration	4.610	2.018e-06	4.6197	1.921e-06
	CT	4.259	1.027e-05	4.2556	1.042e-05
Per Impression	Duration	6.963	1.665e-12	6.972	1.556e-12
	CT	12.999	6.127e-39	12.933	1.469e-38

For each month $m \in \{\text{May, July}\}$, denote by X^m, Y^m the user engagement (either duration or number of click-throughs) in the NN and MAB group respectively. Similarly, denote by $\overline{X}^m, \overline{Y}^m$ be the sample means of X^m, Y^m in month m. We are interested in the *difference-in-differences* in user engagement before and after our policy was deployed, i.e.,

$$\Delta = (Y^{\text{July}} - X^{\text{July}}) - (Y^{\text{May}} - X^{\text{May}}).$$

Consider the hypotheses

$$H_0 : \mathbb{E}[\Delta] \leq 0 \quad \text{vs.} \quad H_1 : \mathbb{E}[\Delta] > 0.$$

First consider the basic Z-score given by

$$Z = \frac{\left(\overline{Y}^{\text{July}}\overline{X}^{\text{July}}\right) - \left(\overline{Y}^{\text{May}} - \overline{X}^{\text{May}}\right)}{\hat{S}} \tag{19}$$

where

$$\hat{S} = \text{SE}\left(\left(\overline{Y}^{\text{July}} - \overline{X}^{\text{July}}\right) - \left(\overline{Y}^{\text{May}} - \overline{X}^{\text{May}}\right)\right) = \sqrt{\text{Var}\left(\left(\overline{Y}^{\text{July}} - \overline{X}^{\text{July}}\right) - \left(\overline{Y}^{\text{May}} - \overline{X}^{\text{May}}\right)\right)}$$

is the estimated standard deviation. For $Z \in \{X^{\text{May}}, X^{\text{July}}, Y^{\text{May}}, Y^{\text{July}}\}$ we denote by N_Z the number of i.i.d. samples of Z, and let S_Z^2 be the sample variance. Assuming the samples are independent, we may approximate the above as

$$\hat{S} \approx \sqrt{\frac{1}{N_{X^{\text{May}}}} S^2_{X^{\text{May}}} + \frac{1}{N_{X^{\text{July}}}} S^2_{X^{\text{July}}} + \frac{1}{N_{Y^{\text{May}}}} S^2_{Y^{\text{May}}} + \frac{1}{N_{Y^{\text{July}}}} S^2_{Y^{\text{July}}}}.$$

As shown in the “Basic” column of Table 3, the p-values for each of the four metrics are very low. We therefore reject the null hypothesis H_0 and conclude that the treatment effect is statistically significant.

However, in reality the samples are **not** independent, since (1) each user may appear in both months, (2) each user has multiple data points in a month, and (3) the same set of cards are shown to both the treatment and control group. We remove the dependence by bootstrapping as follows. From each of these four pools of data points, we randomly draw 10^6 samples with replacement and redefine each $\bar{Z}$ as the bootstrap sample mean where $Z = X^{\text{May}}, X^{\text{July}}, Y^{\text{May}}, Y^{\text{July}}$; see the “Bootstrap” column in Table 2. We still observe very low p-values, which further validates our conclusion.

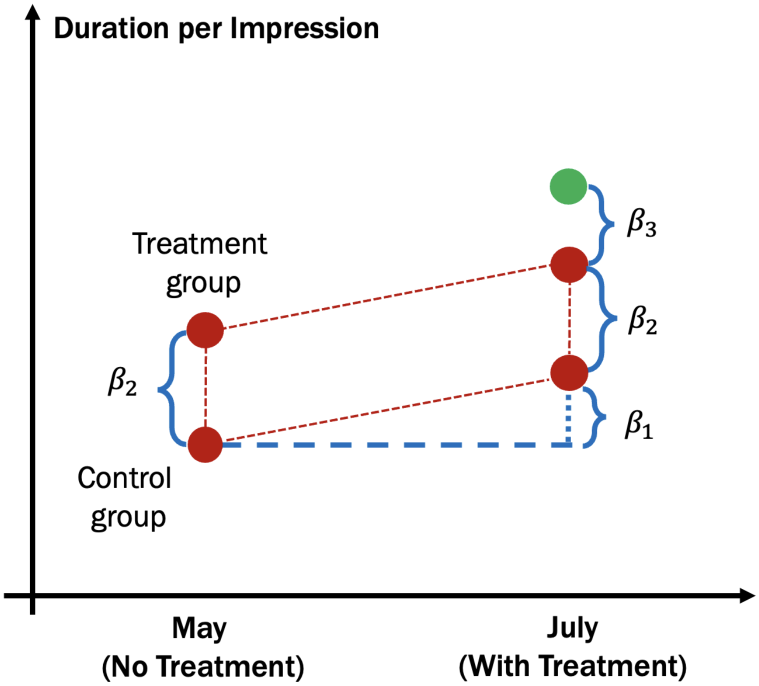


Figure 6. Illustration of DID regression for duration per impression

F.3. DID Regression

Now we consider a DID regression analysis. We first illustrate the idea of DID regression in Figure 6 using per-impression duration as an example. We first vectorize each tuple (u, d, Y_{ud}) into a four-dimensional vector $(t_{ud}, i_{ud}, t_{ud} \cdot i_{ud}, Y_{ud})$ where

$$t_{ud} = \mathbb{1}[\text{day } d \text{ is in July}] \quad \text{and} \quad i_{ud} = \mathbb{1}[\text{user } u \text{ is in MAB group}]$$

are the time and intervention indicators, and $Y_{ud} \in \{C_{ud}, D_{ud}\}$ is the metric under consideration (i.e., click-throughs or duration of user u on day d). The basic assumption in a DID analysis is that the outcome follows the linear model

$$Y_{ud} = \beta_0 + \beta_1 t_{ud} + \beta_2 i_{ud} + \beta_3 t_{ud} \cdot i_{ud} + \varepsilon_{ud} \tag{20}$$

where $\varepsilon_{ud} \sim N(0, \sigma^2)$ with unknown variance σ^2. Intuitively, β_1 measures the effect of being assigned to the treatment group and β_2 captures the overall trend over time. Thus, if there is no treatment effect, the differences between the two groups should remain unchanged across May and July, and therefore the means of the samples from the four pools (*shown as the red dots* in Figure 6) will form a perfect parallelogram.

Now suppose there is indeed a positive treatment effect. Then, the top-right corner of this quadrilateral will be raised; see the green dot in Figure 6. This lift is measured by the variable β_3. To see this, by setting $t_{ud} = i_{ud} = 0$, we observe that β_0 is the mean engagement of control group users in May. Further, note that the top-right corner of the parallelogram is $\beta_1 + \beta_2 + \beta_0$. In contrast, for day d in July and user u in MAB group, if $i_{ud} = t_{ud} = 1$, then the mean outcome satisfies $\mathbb{E}[Y_{ud}] = \beta_0 + \beta_1 + \beta_2 + \beta_3$, which is higher than the top-right red dot by β_3.

Assuming the Gaussian noise, we are able to compute confidence intervals and p-values for the coefficients β_i's; see Tables 4. For both duration and CT, the coefficients β_3 are positive and have very low p-values. Therefore, it is is indeed significant whether a user is assigned to the MAB group. Moreover, note that β_2 has high p-values, so we conclude that the partition of users is sufficiently random, at least on the per-user-per-day level.

As shown in the second half of Table 4, we also consider per-impression user engagement. Similar to the above analysis, we convert each impression j into a three-dimensional binary vector (t_j, i_j, Y_j) where Y_j is either the duration or click-through indicator for impression j. Note that the duration per impression is numerical so we can analyze it using linear regression

Table 4. Difference-In-Differences Regression

			Coef.	Std. Dev.	t	p-value	0.025Q	0.975Q
Per User-Day	Duration	β_0	175.9103	0.640	274.941	0.000	174.656	177.164
		β_1	-38.8514	0.942	-41.263	0.000	-40.697	-37.006
		β_2	-0.3622	0.887	-0.409	0.683	-2.100	1.375
		β_3	**5.9208**	1.303	4.544	**2.759e-06**	3.367	8.475
	#CT	β_0	1.2750	0.008	153.851	0.000	1.259	1.291
		β_1	-0.3341	0.012	-27.394	1.616e-165	-0.358	-0.310
		β_2	-0.0016	0.011	-0.141	0.888	-0.024	0.021
		β_3	**0.0704**	0.017	4.171	**1.516e-05**	0.037	0.103
Per Impression	Duration	β_0	3.9697	0.005	863.796	0.000	3.961	3.979
		β_1	0.1486	0.007	20.234	2.753e-89	0.134	0.163
		β_2	0.0497	0.006	7.781	3.597e-15	0.037	0.062
		β_3	**0.0711**	0.010	6.998	**1.298e-12**	0.051	0.091
	CTR	β_0	-3.5198	0.002	-2092.794	0.000	-3.523	-3.517
		β_1	-0.0161	0.003	-5.947	1.365e-09	-0.021	-0.011
		β_2	0.0133	0.002	5.712	5.582e-09	0.009	0.018
		β_3	**0.0474**	0.004	12.819	**6.417e-38**	0.040	0.055

Note: All regression are linear regression except for per impression CT, where we applied logistic regression due to binary labels.

(20). In contrast, the per impression click-throughs (Y_j) are *binary*, so we instead apply logistic regression: we assume $Y_j \sim \mathrm{Ber}\left(\frac{e^z}{1+e^z}\right)$ where

$$z = \beta_0 + \beta_1 t_j + \beta_2 i_j + \beta_3 t_j i_j.$$

As opposed to the per-user-per-day regression, in this case **all** coefficients have tiny p-values for both CT and duration. In particular, the coefficient β_2 for intervention has low p-value, indicating that the initial user partition may not be truly random in terms of per impression engagement. Nonetheless, this difference is interpretable: our experiment was performed on random user-groups that Glance has been using for months prior to our field test, on which some previous experiments have been performed, causing this discrepancy in user behavior.

We next quantify the improvement. For the per-user-day conversion, we observe that the total duration and the number of click-throughs improved by $\beta_3/(\beta_0+\beta_1) \approx 4.319\%$ and 7.482% respectively. For the per impression conversion, the duration improved by 1.726%. Finally, note that the β's for CTR are based on *logistic* regression. The improvement in the odds ratio is $e^{\beta_3} - 1 \approx 4.854\%$.