Active Context Selection Improves Simple Regret in Contextual Bandits

Mohammad Shahverdikondori
College of Management of Technology, EPFL
mohammad.shahverdikondori@epfl.ch

Jalal Etesami
Department of Computer Science, TU Munich
j.etesami@tum.de

Negar Kiyavash
College of Management of Technology, EPFL
negar.kiyavash@epfl.ch

Abstract

We study the contextual multi-armed bandit problem with a finite context space (a.k.a. subpopulations), where the learner recommends a best action for each context and is evaluated by context-weighted simple regret. Our guarantees are worst-case over the reward distributions, while remaining instance-dependent with respect to the context distribution vector $\mathbf{p}$. Akin to experimental design problems where the population of interest is fixed but the sampled subpopulation can be controlled, we allow the learner to actively choose which context to sample from. For a known $\mathbf{p}$, we characterize tight regret rates: passive sampling where contexts are randomly revealed achieves regret of order $\sqrt{n/T\,\|\mathbf{p}\|_{1/2}}$, whereas active sampling with allocation $q_j \propto p_j^{2/3}$ achieves the tight rate $\sqrt{n/T}\,\|\mathbf{p}\|_{2/3}$. The resulting improvement can be as large as $\Theta(k^{1/4})$, where k is the number of contexts. We further extend the analysis to budgeted active sampling, characterize the corresponding tight rate, and identify when a limited active budget suffices to recover the fully active rate. When $\mathbf{p}$ is unknown, we propose the Explore-Explore-Then-Commit (EETC) algorithm, which optimally balances estimating the context distribution and the time to switch to active allocation, such that it matches the known-$\mathbf{p}$ active rate up to constants. Experiments on synthetic and real-world data support our theoretical findings.

1 Introduction

Sequential decision making problems often involve contextual information: the outcome distribution of an action may depend on observable characteristics of the unit under consideration [LS20, BCB$^+$12]. In the finite context regime, each context value naturally defines a subpopulation of the target population, such as an age bracket, geographic region, or disease subtype. This viewpoint is especially common in experimental design applications, where the effect of an action, or treatment, may vary substantially across subpopulations. The learning objective is then not simply to identify a single globally best action, but to recommend the best treatment for each subpopulation [PR13, RKB$^+$21, CHVDS23].

This motivates the context-weighted simple regret objective studied in this paper: after a fixed experimental budget, the learner recommends one treatment per subpopulation. The resulting regret is the average recommendation gap weighted by the population distribution over subpopulations. Even when this distribution is known from historical data, the learner still faces the question of how to allocate the experimental budget across subpopulations. In the standard passive setting, contexts

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

arrive randomly according to the population distribution and the learner only chooses the treatment [KZAB23, JLB22, DSC$^+$18]. However, in many applications, the experimenter may have partial control over which subpopulation to sample. For instance, an online platform may target specific user segments, or a clinical study may recruit participants from selected demographic groups, possibly at different costs. This raises two natural design questions which we study in this paper: when is it worth actively choosing the context, and what is the gain compared to passive sampling?

Besides a fully passive scenario for benchmarking, we study a fully active setting where the learner can choose both the subpopulation and the treatment as well as a budgeted intervention setting, in which only a fraction of the rounds can be active, capturing cases where targeted recruitment or controlled sampling is costly. Our analysis shows how the value of active subpopulation selection depends on the population distribution and when it can lead to meaningful improvements over passive sampling. Additionally, we consider the setting where subpopulation distribution is not a priori known and must be estimated. This setting introduces a trade-off between using passive samples to learn the subpopulation distribution $\mathbf{p}$ and using active samples to improve treatment recommendations. Somewhat surprisingly, we show that for sufficiently large horizons, this uncertainty does not result in worst simple regret rate compared to an active policy that knows the distribution.

Our work is related to several lines of work in literature. Heterogeneous treatment effect and subpopulation-aware decision making problems study settings where treatment effects vary across groups and where ignoring this heterogeneity can lead to suboptimal treatment rules [RKB$^+$21, CHVDS23]. Contextual bandits and policy learning methods also aim to learn context-dependent decisions, including recent work on simple regret in contextual bandits such as Proportional Response [KZAB23], as well as adaptive experimental designs for policy learning [KOIK24, DSC$^+$18, JLB22]. These works typically consider the passive contextual setting, where contexts are revealed by the environment. Our focus is different: we quantify the improvement the learner can obtain by actively selecting which context (or subpopulation) to sample. The closest pure exploration formulation is multi-bandit best arm identification [GGLB11, SBC19], where the goal is to identify the best arm in each of several bandits. This is close to our fully active setting, but it studies best arm identification, where the goal is to minimize the error probability of identifying the best arm, rather than context-weighted simple regret, and does not study active and passive context sampling. Active context sampling has also been recently studied in stochastic contextual linear bandits [BKL26], where, for a known context distribution, choosing context–action pairs can improve the minimax rate by up to a factor of $\sqrt{d}$, with d the linear dimension. In their setting the rewards share a single linear parameter across contexts, whereas we consider unstructured rewards with nothing shared across subpopulations, and we characterize the optimal allocation and the active–passive gap in terms of $\mathbf{p}$, together with budgeted interventions and unknown $\mathbf{p}$. We refer the reader to Appendix A for a more detailed discussion of related work.

Contributions. Our main contributions are as follows:

- For known subpopulation distribution $\mathbf{p}$, we characterize the optimal context-weighted simple regret rate and show that it is attainable, up to constant factors, by a policy that selects which subpopulation to sample from independently of the observed history proportionally to $p_j^{2/3}$, where p_j is the jth element of the subpopulation distribution $\mathbf{p}$. The gap of this active policy's regret compared to passive sampling can be of order $\Theta(k^{1/4})$, where k is the number of contexts.

- We extend the analysis to a budgeted intervention setting, where the learner can actively choose the subpopulation only in α fraction of the rounds. We characterize the optimal allocation in this setting, propose an algorithm that achieves the optimal simple regret rate, and identify the minimum intervention budget required to match the optimal fully active rate.

- We study the setting where the subpopulation distribution $\mathbf{p}$ is unknown to the learner. We show that neither fully passive nor fully active policies are optimal in general, and propose the EETC algorithm, which balances passive estimation with active allocation rounds. We prove that, for large horizons, EETC achieves the same rate as the optimal fully active policy that knows $\mathbf{p}$, up to constant factors.

2 Problem Setup

In this section, we formally define the bandit with subpopulations problem considered in this paper and introduce the relevant notation.

Notation. For any natural number n, let $[n] \coloneqq \{1, \dots, n\}$. Let $\Delta^{n-1} \coloneqq \{\mathbf{x} \in \mathbb{R}^n | \sum_i x_i = 1,\ x_i \geq 0\}$, denotes the probability simplex in $\mathbb{R}^n$ and $\Delta_+^{n-1} \subset \Delta^{n-1}$ the set of vectors with all positive entries. For a vector $\mathbf{v} \in \mathbb{R}^n$ we denote its ℓ_p norm by $\|\mathbf{v}\|_p$.

The bandit with subpopulations problem models a sequential interaction between an agent and an environment. The environment consists of a treatment variable $A \in [n]$, a subpopulation (or context) variable $C \in [k]$, and a reward variable $Y \in [0, 1]$. For each treatment–subpopulation pair $(i, j) \in [n] \times [k]$, the reward distribution is given by $\nu_{i,j}$ with mean $\mu_{i,j}$. Let $\boldsymbol{\nu}$ and $\boldsymbol{\mu} \in [0, 1]^{n \times k}$ denote the matrix of reward distributions and reward means, respectively. We let $\mathbf{p} \in \Delta_+^{k-1}$ denote the population distribution over subpopulations, that is, $\forall j \in [k] : \mathbb{P}(C = j) = p_j > 0$, and define $p_{\min} \coloneqq \min_{j \in [k]} p_j$. The agent interacts with the environment for T rounds. At each round $t \in [T]$, the interaction proceeds as follows.

1. A subpopulation C_t is either sampled according to p and revealed to the agent, or selected by the agent based on the observed history $\mathcal{H}_t = (A_s, C_s, Y_s)_{s=1}^{t-1}$ and, when known, the population distribution vector $\mathbf{p}$.
2. Based on the history $\mathcal{H}_t$ and the current subpopulation C_t, the agent selects a treatment A_t.
3. The agent then observes a random reward Y_t, drawn independently from ν_{A_t, C_t}.

We refer to a round in which C_t is sampled from $\mathbf{p}$ and the agent only chooses A_t as *passive* round. In contrast, when the agent is allowed to choose both C_t and A_t, we call the round *active*.

An instance of the problem is specified by a pair $(\boldsymbol{\nu}, \mathbf{p})$. Let $\mathcal{E}(n, k, \mathbf{p})$ denote the class of all instances with n treatments and k subpopulations distributed according to $\mathbf{p}$, where the reward distributions is supported on $[0, 1]$. Although we focus on a finite discrete set of subpopulations, the same perspective also applies to continuous context spaces, as common in contextual bandits, by introducing a suitable discretization of the context space. Under such a discretization, we obtain the same formulation and its corresponding regret bounds up to the discretization error. With Lipschitz mean rewards, discretizing the context space into cells of diameter at most h adds an error of order $\mathcal{O}(Lh)$, where L is the Lipschitz constant, as in the standard binning approach for bandits with covariates [RZ10, PR13].

After T rounds, the agent outputs, for each subpopulation $j \in [k]$, a recommended best treatment $\hat{a}_T(j)$. The collection of sampling decisions made during the interaction together with the final recommendation rule forms a policy π. For a policy π interacting with an instance $(\boldsymbol{\nu}, \mathbf{p}) \in \mathcal{E}(n, k, \mathbf{p})$ with means matrix $\boldsymbol{\mu}$, we define the simple regret as the expected subpopulation-weighted gap between the optimal and recommended treatments:

$$\mathrm{SR}\left(\pi, T, \boldsymbol{\nu}, \mathbf{p}\right) \coloneqq \mathbb{E}_{j \sim \mathbf{p}}\left[\mu_{a^*(j),j} - \mu_{\hat{a}_T(j),j}\right] = \sum_j p_j \mathbb{E}\left[\Delta_j\right],$$

where $a^*(j) \in \arg\max_{i \in [n]} \mu_{i,j}$ denotes an optimal treatment for subpopulation j, and Δ_j the recommendation gap for subpopulation j. The objective is to minimize the worst-case simple regret (over the class $\mathcal{E}(n, k, \mathbf{p})$) defined as:

$$\mathrm{SR}(\pi, T, \mathbf{p}) \coloneqq \sup_{\boldsymbol{\nu} : (\boldsymbol{\nu}, \mathbf{p}) \in \mathcal{E}(n,k,\mathbf{p})} \mathrm{SR}\left(\pi, T, \boldsymbol{\nu}, \mathbf{p}\right), \tag{1}$$

for a policy π, i.e., find $\arg\inf_\pi \mathrm{SR}(\pi, T, \mathbf{p})$.

This criterion is worst-case with respect to the reward means, but remains parameterized by the subpopulation distribution $\mathbf{p}$. In other words, the performance measure $\mathrm{SR}(\pi, T, \mathbf{p})$ is the supremum over reward configurations for a fixed population profile $\mathbf{p}$, rather than a worst-case $\mathbf{p}$. This is because our objective is not to design algorithms for an adversarial subpopulation distribution, but develop algorithms whose performance *adapts* to the instance-dependent vector $\mathbf{p}$. Thus, the problem is worst-case in terms of the mean rewards but instance-dependent for the subpopulation distribution. This also holds when $\mathbf{p}$ is unknown to the learner (Section 4): the guarantee is for each fixed unknown $\mathbf{p}$, with the worst case taken only over the rewards.

Throughout the paper, we focus on the data-rich regime, in which the horizon $T \gg n, k, 1/p_{\min}$. Concretely, all results of the paper hold once $T \geq C\left(n + k^{1/3}\ln(kT)\right)/p_{\min}$ for a universal constant C;

2.1 Policy Classes

We consider several classes of policies in the paper. A policy is called *passive* if, at each round, the subpopulation is randomly drawn according to the distribution $\mathbf{p}$ and not selected by the agent. In contrast, a policy is called *active* if the agent can pick the subpopulation at every round. We denote the classes of passive and active policies by Π_{pas} and Π_{act}, respectively.

We say that a policy π is *history-free* if its subpopulation selection rule does not depend on the observed history. Formally, C_t is drawn independently of the history $\mathcal{H}_t$, and hence independently across rounds, with a distribution that may depend on t and $\mathbf{p}$. Recall that each policy consists of two components: a subpopulation selection rule and an arm selection rule within each subpopulation. For history-free policies, only the first component is restricted to be independent of the history; the second component can clearly depend on the history. In all the algorithms considered in this paper, this second component is implemented through a standard bandit subroutine, as discussed in the next subsection.

We denote the class of all history-free policies by Π_{hf}. Note that passive policies form a special case of history-free policies (i.e., $\Pi_{\mathrm{pas}} \subseteq \Pi_{\mathrm{hf}}$), since for such policies, the subpopulation is sampled according to $\mathbf{p}$ independently of the history. We denote the class of all policies, without any restriction on how the subpopulation selection rule may depend on observed rewards, by Π.

Finally, we distinguish between policies that know the distribution $\mathbf{p}$ and those that do not by indicating $\mathbf{p}$ in parentheses. For example, $\Pi_{\mathrm{hf}}(\mathbf{p})$ denotes the class of history-free policies that know $\mathbf{p}$, whereas Π_{hf} denotes the class of history-free policies that do not. We use the same convention for the other policy classes.

2.2 Within-Subpopulation Bandit Subroutine

Since we do not impose any structural relation across the reward means of different subpopulations, the learning problem within each subpopulation $j \in [k]$ can be viewed as an independent stochastic bandit problem with n arms. If the number of samples allocated to subpopulation j were fixed, then minimizing $\mathbb{E}[\Delta_j]$ would reduce to the classical simple regret problem in standard bandits, which is well-studied in the literature [BMS09, ZSSJ23, LBA+26]. In our setting, however, these within-subpopulation horizons are induced by the learner's subpopulation selection rule. Thus, the main remaining question is how to allocate the total budget T across subpopulations.

Throughout the paper, we use $\mathcal{A}_{\mathrm{SR}}$ to denote an anytime standard bandit subroutine for simple regret minimization. Each subpopulation maintains an independent copy of $\mathcal{A}_{\mathrm{SR}}$, which is used to select treatments whenever that subpopulation is sampled. This separates the standard within-subpopulation treatment selection problem from the higher level allocation problem.

We rely on the following standard minimax guarantee.

Lemma 2.1 ([LS20], Section 33)**.** *For a stochastic bandit with n arms and T rounds, the minimax simple regret scales as $\Theta\left(\sqrt{n/T}\right)$. Moreover, this rate can be achieved by anytime algorithms, i.e., algorithms that do not require prior knowledge of T.*

For concreteness, uniform sampling over the n arms gives the optimal simple regret rate up to logarithmic factors. Alternatively, an anytime cumulative regret minimization algorithm such as MOSS [AB09], combined with a recommendation rule that outputs arms proportionally to their play counts, achieves the optimal $\sqrt{n/T}$ rate up to constants.

This reduction clarifies the allocation problem. Since the final objective weights the recommendation gap in subpopulation j by p_j, one might expect the natural allocation to be proportional to $\mathbf{p}$, which is exactly the passive allocation. However, because simple regret decreases nonlinearly with the number of samples, this allocation is generally suboptimal. The next sections show that active subpopulation selection leads to a different allocation rule and can largely improve over passive sampling.

3 Known Subpopulation Distribution

In this section, we study the setting where the subpopulation distribution $\mathbf{p}$ is known. We first show that, under the worst-case simple regret criterion in (1), it is sufficient to consider history-free policies: allowing the subpopulation selection rule to depend on observed history does not improve the minimax rate. We then compare passive and fully active policies and characterize the active–passive gap, showing that the optimal active allocation can improve over passive sampling by a factor as large as $\Theta(k^{1/4})$. Finally, we extend the analysis to the budgeted setting, where only an α fraction of rounds can be active, and characterize the optimal allocation under this constraint.

We begin with notation. For a policy π interacting with an instance $(\boldsymbol{\nu}, \mathbf{p})$, let $\mathbf{q}(\pi, (\boldsymbol{\nu}, \mathbf{p})) \in \Delta^{k-1}$ denote the expected proportion of rounds allocated to the different subpopulations under π. More precisely, its j-th coordinate is the expected fraction of rounds t in which $C_t = j$. As a consequence of our definition, for history-free policies, the allocation proportion is independent of $\boldsymbol{\nu}$, and we simply write $\mathbf{q}(\pi, (\boldsymbol{\nu}, \mathbf{p})) = \mathbf{q}(\pi, \mathbf{p})$.

3.1 Optimality of History-Free Policies

We formalize the reduction to history-free policies. The key point is that, for the worst-case simple regret in (1), the optimal rate is determined by the allocation of samples across subpopulations, and history-dependent changes to this allocation do not improve the minimax order.

Lemma 3.1 (Simple Regret Lower Bound)**.** *For any subpopulation distribution $\mathbf{p} \in \Delta_+^{k-1}$ known to the agent and any policy $\pi \in \Pi(\mathbf{p})$,*

$$\mathit{SR}(\pi, T, \mathbf{p}) \in \Omega\Big(\sqrt{\frac{n}{T}}\, \|\mathbf{p}\|_{2/3}\Big).$$

The proof in Appendix B also covers shorter horizons outside the data-rich regime, with no dependence on $p_{\min}$.

Proof sketch. For any policy, we place a prior over hard instances in which, in each subpopulation, one treatment is slightly better than the others and the identity of the best treatment is uncertain. Using the Bretagnolle–Huber inequality and the KL decomposition, we show that a subpopulation j that receives a fraction q_j of the rounds, averaged over the instances, forces regret of order $p_j\sqrt{n/(Tq_j)}$, for adaptive policies as well. Summing over subpopulations gives a lower bound of order $\sqrt{n/T}\sum_{j\in[k]} p_j/\sqrt{q_j}$, and optimizing over the allocation gives $q_j \propto p_j^{2/3}$ and the claimed bound $\sqrt{n/T}\, \|\mathbf{p}\|_{2/3}$. The full proof is deferred to Appendix B. □

We now show that, for history-free policies, the worst-case simple regret depends solely on the expected allocation proportions across subpopulations, namely on the vector $\mathbf{q}(\pi, \mathbf{p})$. The following characterization is tight up to constant factors.

Lemma 3.2 (Simple Regret of History-Free Policies)**.** *For any history-free policy $\pi \in \Pi_{\mathrm{hf}}(\mathbf{p})$ along with the subroutine $\mathcal{A}_{\mathrm{SR}}$ and expected subpopulation proportions $\mathbf{q}(\pi, \mathbf{p})$, if $\forall j \in [k] : Tq(\pi, \mathbf{p})_j > 24\, \ln(2kT)$, then*

$$\mathit{SR}(\pi, T, \mathbf{p}) \in \Theta\Big(\sqrt{\frac{n}{T}} \sum_{j\in[k]} \frac{p_j}{\sqrt{q(\pi, \mathbf{p})_j}}\Big).$$

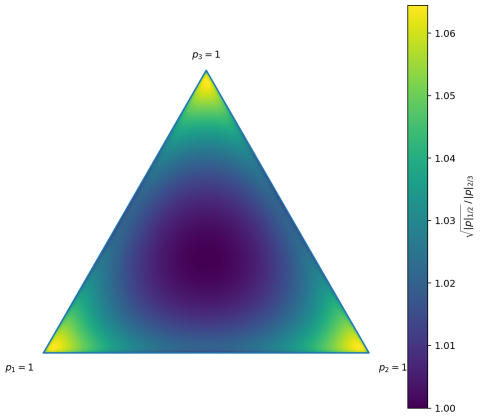
This shows that, for history-free policies, the performance depends solely on the quantity $\sum_{j\in[k]} p_j/\sqrt{q(\pi, \mathbf{p})_j}$. Therefore, for any $\mathbf{q} \in \Delta_+^{k-1}$, we define

$$S_{\mathbf{p}}(\mathbf{q}) \coloneqq \sum_{j\in[k]} \frac{p_j}{\sqrt{q_j}}. \tag{2}$$

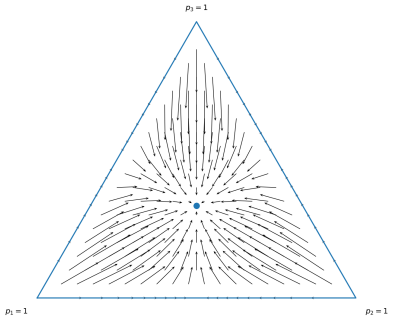
The following lemma identifies the optimal active allocation that minimizes this quantity.

Lemma 3.3 (Optimal Allocation for Active Policies)**.** *For any $\mathbf{p} \in \Delta_+^{k-1}$, the unique minimizer of $S_{\mathbf{p}}(\mathbf{q})$ over $\mathbf{q} \in \Delta_+^{k-1}$ is*

$$q^*(\mathbf{p})_j = \frac{p_j^{2/3}}{\sum_{\ell\in[k]} p_\ell^{2/3}}, \quad j \in [k].$$



(a) Heatmap of $R(\mathbf{p})$ on the 3-simplex.



(b) Flow from $\mathbf{p}$ to the optimal active allocation $\mathbf{q}^*(\mathbf{p})$.

Figure 1: Illustration of the active–passive gap on the probability simplex for $k = 3$. Left: heatmap of $R(\mathbf{p})$ over the simplex, showing that the gap is negligible near the uniform distribution (center) and increases for non-uniform population profiles (corners). Right: for each point $\mathbf{p}$, the arrow points toward the optimal active allocation $\mathbf{q}^*(\mathbf{p})$, illustrating that the optimal active policy shifts the sampling proportions toward a more uniform (center), and hence more exploratory, allocation.

Moreover, $S_{\mathbf{p}}(\mathbf{q}^*(\mathbf{p})) = \|\mathbf{p}\|_{2/3}$.

Combining Lemmas 3.1, 3.2, and 3.3 gives the following optimality result, which states that history-free policies achieve the same regret as general policies that can select the subpopulation allocation based on the history.

Proposition 3.4 (Optimality of History-Free Policies)**.** *For every* $\mathbf{p} \in \Delta_+^{k-1}$,

$$\inf_{\pi \in \Pi_{\mathrm{hf}}(\mathbf{p})} SR(\pi, T, \mathbf{p}) \in \Theta\Big(\inf_{\pi \in \Pi(\mathbf{p})} SR(\pi, T, \mathbf{p}) \Big).$$

That is, there is no optimality penalty for adhering to history-free policies.

3.2 Gap Between Active and Passive Policies

We now compare the best passive and active policies within the class of history-free policies that know $\mathbf{p}$. Under any passive policy, the subpopulation observed at each round is sampled from $\mathbf{p}$, and hence $\mathbf{q}(\pi, \mathbf{p}) = \mathbf{p}$. Lemma 3.2 gives

$$\inf_{\pi \in \Pi_{\mathrm{pas}}(\mathbf{p})} \mathrm{SR}(\pi, T, \mathbf{p}) \in \Theta\Big(\sqrt{\frac{n}{T}} \sum_{j \in [k]} \sqrt{p_j} \Big) = \Theta\Big(\sqrt{\frac{n}{T}}\, \|\mathbf{p}\|_{1/2}^{1/2} \Big). \tag{3}$$

On the other hand, Lemmas 3.2 and 3.3 imply that the best active history-free policy satisfies

$$\inf_{\pi \in \Pi_{\mathrm{act}}(\mathbf{p}) \cap \Pi_{\mathrm{hf}}(\mathbf{p})} \mathrm{SR}(\pi, T, \mathbf{p}) \in \Theta\Big(\sqrt{\frac{n}{T}}\, \|\mathbf{p}\|_{2/3} \Big). \tag{4}$$

This rate is achieved by sampling subpopulation j proportionally to $p_j^{2/3}$ and running $\mathcal{A}_{\mathrm{SR}}$ within each subpopulation. Compared to the passive allocation $\mathbf{p}$, the optimal active allocation is closer to uniform: it downweights large coordinates of $\mathbf{p}$ and upweights small ones, leading to a more exploratory allocation. See Figure 1 for an illustration.

The multiplicative active–passive gap is therefore $R(\mathbf{p}) := \sqrt{\|\mathbf{p}\|_{1/2}} / \|\mathbf{p}\|_{2/3}$ and (3)–(4) imply

$$\frac{\inf_{\pi \in \Pi_{\mathrm{pas}}(\mathbf{p})} \mathrm{SR}(\pi, T, \mathbf{p})}{\inf_{\pi \in \Pi_{\mathrm{act}}(\mathbf{p}) \cap \Pi_{\mathrm{hf}}(\mathbf{p})} \mathrm{SR}(\pi, T, \mathbf{p})} \in \Theta\big(R(\mathbf{p})\big).$$

Algorithm 1 α-Active Simple Regret Algorithm

Input: Standard bandit subroutine $\mathcal{A}_{\mathrm{SR}}$, distribution $\mathbf{p}$, budget α, and horizon T.
Initialize one instance $\mathcal{A}_{\mathrm{SR}}^j$ of $\mathcal{A}_{\mathrm{SR}}$ for each subpopulation $j \in [k]$.
Define $F(c) = \sum_{j\in[k]} \max\big((1-\alpha)p_j,\ cp_j^{2/3}\big)$.
for $t = 1, 2, \ldots, \lfloor(1-\alpha)T\rfloor$ **do**
    Sample $C_t \sim \mathbf{p}$, and choose the treatment A_t recommended by $\mathcal{A}_{\mathrm{SR}}^{C_t}$.
end for
Find the unique value c^* such that $F(c^*) = 1$.
For every $j \in [k]$, set $q_j^* = \max\big((1-\alpha)p_j,\ c^* p_j^{2/3}\big)$, and $r_j = \frac{q_j^* - (1-\alpha)p_j}{\alpha}$.
for $t = \lfloor(1-\alpha)T\rfloor + 1, \ldots, T$ **do**
    Sample subpopulation C_t according to $\mathbf{r}$, and choose the treatment A_t recommended by $\mathcal{A}_{\mathrm{SR}}^{C_t}$.
end for
Output: For each subpopulation j, output the recommended arm $\hat{a}_T(j)$ returned by $\mathcal{A}_{\mathrm{SR}}^j$.

Lemma 3.5 (Active–Passive Gap)**.** *For every $\mathbf{p} \in \Delta_+^{k-1}$, $R(\mathbf{p}) \geq 1$. Moreover, $\sup_{\mathbf{p}\in\Delta_+^{k-1}} R(\mathbf{p}) \in \Theta\big(k^{\frac{1}{4}}\big)$, and this order is tight.*

The inequality $R(\mathbf{p}) \geq 1$ implies that active policies are always at least as good as passive policies. When $\mathbf{p}$ is uniform, $R(\mathbf{p}) = 1$, so the two rates coincide; more generally, the gain is limited near uniform distributions. The largest gap arises for highly non-uniform $\mathbf{p}$, in particular when one subpopulation has a large mass and the remaining coordinates are all equal and of order $k^{-3/2}$.

3.3 Budgeted Interventions

In many practical settings, the agent has a limited budget of active rounds, in which it can choose both the subpopulation and the treatment. We focus on such budgeted setting in this section. More precisely, we consider policies that are allowed to actively select the subpopulation in at most αT of the total T rounds, where $\alpha \in [0, 1]$. We assume the distribution $\mathbf{p}$ is known to the agent, and denote this class of policies by $\Pi_\alpha(\mathbf{p})$. Note that $\Pi_{\alpha=0}(\mathbf{p}) = \Pi_{\mathrm{pas}}(\mathbf{p})$ and $\Pi_{\alpha=1}(\mathbf{p}) = \Pi_{\mathrm{act}}(\mathbf{p})$, so this model interpolates between the passive and fully active settings.

Our algorithm design relies on the result of Lemma 3.2, which states that the regret of a history-free policy π depends on the subpopulation proportions $\mathbf{q}(\pi, \mathbf{p})$ only through the quantity $S_{\mathbf{p}}(\mathbf{q})$ defined in (2). Thus, the goal is to choose the allocation vector $\mathbf{q}$ to minimize this term. Given that the subpopulation cannot be selected by the agent during the $(1-\alpha)T$ rounds but must instead be sampled according to $\mathbf{p}$, any policy in $\Pi_\alpha(\mathbf{p})$ must satisfy $q(\pi, \mathbf{p})_j \geq (1-\alpha)p_j, \forall j \in [k]$. Therefore, the best achievable allocation is given by the solution of the following constrained optimization problem,

$$\min_{\mathbf{q}\in\Delta_+^{k-1}} S_{\mathbf{p}}(\mathbf{q}) = \min_{\mathbf{q}\in\Delta_+^{k-1}} \sum_{j\in[k]} \frac{p_j}{\sqrt{q_j}}, \qquad \text{s.t.} \quad q_j \geq (1-\alpha)p_j, \quad \forall j \in [k]. \tag{5}$$

This optimization problem is convex and its solution admits a simple threshold structure. Let $\mathbf{q}_\alpha^*(\mathbf{p})$ and $v_\alpha^*(\mathbf{p})$ denote the optimizer and the optimal value of the program, respectively. In particular, there exists a constant $c^* \geq 0$ such that $\mathbf{q}_\alpha^*(\mathbf{p})_j = \max\big((1-\alpha)p_j,\ c^* p_j^{2/3}\big), j \in [k]$. Thus, the subpopulations are partitioned into two groups. For those subpopulations whose coordinates satisfy $\mathbf{q}_\alpha^*(\mathbf{p})_j > (1-\alpha)p_j$, the optimal allocation remains proportional to $p_j^{2/3}$, whereas for subpopulations satisfying $\mathbf{q}_\alpha^*(\mathbf{p})_j = (1-\alpha)p_j$, the passive rounds already provide a sufficient number of samples, so there is no need to actively allocate additional pulls to them. Algorithm 1, which is based on this characterization, summarizes the steps of the budgeted intervention setting.

Theorem 3.6 (Simple Regret of Algorithm 1)**.** *For every $\mathbf{p} \in \Delta_+^{k-1}$, Algorithm 1 is a history-free policy in $\Pi_\alpha(\mathbf{p})$ with expected subpopulation proportions equal to $\mathbf{q}_\alpha^*(\mathbf{p})$. Consequently, when T is large, its worst-case simple regret satisfies the optimal rate*

$$SR(\texttt{Alg. 1}, T, \mathbf{p}) \in \mathcal{O}\Big(\sqrt{\frac{n}{T}}\ v_\alpha^*(\mathbf{p})\Big).$$

Algorithm 1 can also be run without knowing T, by spreading the active rounds evenly over time, with the same guarantee up to constants at every round (Appendix C.3).

Another question that arises naturally in this setting is that given a subpopulation distribution $\mathbf{p}$, what is the minimum intervention budget α required to match the fully active setting? Equivalently, how many active rounds are sufficient to attain the optimal fully active rate. We denote this threshold by $\alpha_{\min}(\mathbf{p})$. The next lemma gives a closed-form expression for $\alpha_{\min}(\mathbf{p})$ in terms of $\mathbf{p}$.

Lemma 3.7 (Threshold Budget $\alpha_{\min}(\mathbf{p})$)**.** *For a given $\mathbf{p} \in \Delta_+^{k-1}$, let $\alpha_{min}(\mathbf{p}) := 1 - p_{\max}^{-1/3} / \sum_{j\in[k]} p_j^{2/3}$, where $p_{\max} = \max_j p_j$. Then, for every $\alpha \in [\alpha_{min}(\mathbf{p}), 1]$, we have $v_\alpha^*(\mathbf{p}) = \|\mathbf{p}\|_{2/3}$. Consequently, for every such α, the simple regret bound of Algorithm 1 matches the optimal rate of fully active policies, which is also the optimal rate over all policies.*

4 Unknown Subpopulation Distribution

In this section, we study the setting in which the distribution of subpopulations, $\mathbf{p}$, is unknown to the learner, and investigate how this changes the policy design and the achievable regret rates. We show that neither passive nor fully active policies are optimal in general. Instead, the optimal strategy must combine passive and active rounds. We then propose an algorithm based on this principle and show that, for a sufficiently large horizon T, it achieves the same simple regret bound as a fully active policy that knows $\mathbf{p}$.

The simple regret criterion in (1) depends explicitly on $\mathbf{p}$. Yet a fully active policy agnostic of $\mathbf{p}$ that chooses a subpopulation at every round, receives no information about $\mathbf{p}$. Consequently, if the policy undersamples some subpopulation, the worst-case instance may assign a large mass to that subpopulation, resulting in a large regret. Therefore, without knowledge of $\mathbf{p}$, the minimax fully active policy must allocate samples uniformly across subpopulations, i.e., $\mathbf{q}(\pi, \mathbf{p}) = \frac{1}{k}\mathbf{1}$, where $\mathbf{1}$ denotes the all-ones vector. By Lemma 3.2, the resulting simple regret is

$$\Theta\Big(\sqrt{\frac{n}{T}} \sum_{j\in[k]} \frac{p_j}{\sqrt{1/k}}\Big) = \Theta\Big(\sqrt{\frac{nk}{T}}\Big),$$

which is worse than the passive regret rate $\Theta\big(\sqrt{\frac{n}{T}\,\|\mathbf{p}\|_{\frac{1}{2}}}\big)$ derived in (3), as $\sum_{j\in[k]}\sqrt{p_j} \le \sqrt{k}$, for every $\mathbf{p} \in \Delta_+^{k-1}$.

In the next section, we present an algorithm that outperforms the passive policy. It does so by balancing the exploration–exploitation trade-off in learning the distribution $\mathbf{p}$. Passive rounds provide information about $\mathbf{p}$ and allow the learner to estimate it, while active rounds are used to reduce the simple regret once a sufficiently accurate estimate is available.

4.1 Explore-Explore-Then-Commit (EETC)

For each round $t \in [T]$ and subpopulation $j \in [k]$, let $N(t)_j = \sum_{s=1}^{t} \mathbf{1}\{C_s = j\}$ denote the number of times subpopulation j has been observed up to round t, including round t itself, and let $\hat{\mathbf{p}}(t)$ denote the empirical estimate of $\mathbf{p}$, namely $\hat{\mathbf{p}}(t)_j = N(t)_j/t$.

The algorithm proceeds in three phases: (I) The algorithm acts passively in order to estimate $\mathbf{p}$. The key point is that after $\mathcal{O}(\log(T))$ rounds, the estimate becomes accurate enough so that the allocation $\mathbf{q}^*(\hat{\mathbf{p}}(t))$ computed from $\hat{\mathbf{p}}(t)$ yields a value $S_{\mathbf{p}}(\mathbf{q}^*(\hat{\mathbf{p}}(t)))$ within a constant factor of the optimal value $\|\mathbf{p}\|_{2/3}$. Motivated by this, we define the first stopping time

$$\tau_1 := \inf\Big\{1 \le t \le T \mid \min_{j\in[k]} N(t)_j \ge \log(kT^2)\Big\}. \tag{6}$$

We show that, with high probability, τ_1 is logarithmic in T, and that at time τ_1 the estimate $\hat{\mathbf{p}}(\tau_1)$ is sufficiently accurate for the subsequent allocation steps.

(II) This phase commences at time τ_1 and the algorithm still acts passively. While at this point, the estimate is already accurate enough to guide the active allocation, the algorithm can still benefit from additional passive samples as long as doing so does not leave too few active rounds to implement

the optimal active allocation which is suggested by the current estimate. Lemma 3.7 gives exactly this threshold: for an estimate $\hat{\mathbf{p}}(t)$, the fully active rate can still be matched if the remaining active fraction is at least $\alpha_{\min}(\hat{\mathbf{p}}(t))$, or equivalently if the passive fraction is at most $1 - \alpha_{\min}(\hat{\mathbf{p}}(t))$. Accordingly, we define

$$\tau_2 \coloneqq \inf\left\{t \geq \tau_1 + 1 \mid t/T > 1 - \alpha_{\min}(\hat{\mathbf{p}}(t))\right\}. \tag{7}$$

Thus, τ_2 is the first round at which the current estimate no longer allows us to keep playing passively and still match the optimal active rate. The algorithm therefore stops the passive phase at round τ_2 and sets $\hat{\mathbf{p}}^{\mathrm{fr}} \coloneqq \hat{\mathbf{p}}(\tau_2 - 1)$, where $\hat{\mathbf{p}}(\tau_2 - 1)$ is the estimate of the subpopulation distribution vector at the previous round, i.e., $\tau_2 - 1$. For large T, we show that $\tau_2 \leq T$ with high probability. Phases (I) and (II) are both passive and therefore form a single loop in Algorithm 2, during which $\hat{\mathbf{p}}(t)$ is updated at every round.

Algorithm 2 Explore-Explore-Then-Commit (EETC)

Input: standard bandit subroutine $\mathcal{A}_{\mathrm{SR}}$ and horizon T.
Initialize one instance $\mathcal{A}_{\mathrm{SR}}^j$ of $\mathcal{A}_{\mathrm{SR}}$ for each subpopulation $j \in [k]$.
for $t = 1, 2, \ldots$ until $t = \tau_2$ *(phases I and II)* **do**
    Sample $C_t \sim \mathbf{p}$ and choose the treatment A_t recommended by $\mathcal{A}_{\mathrm{SR}}^{C_t}$.
    Update $\hat{\mathbf{p}}(t)$; τ_1 and τ_2 are determined online from (6) and (7).
end for
Compute $\mathbf{q}$ and $\mathbf{r}$ from (8) using $\hat{\mathbf{p}}^{\mathrm{fr}}$.
for $t = \tau_2 + 1, \ldots, T$ **do**
    Sample subpopulation C_t according to $\mathbf{r}$ and choose the treatment A_t recommended by $\mathcal{A}_{\mathrm{SR}}^{C_t}$.
end for
Output: For each subpopulation j, output the recommended arm $\hat{a}_T(j)$ returned by $\mathcal{A}_{\mathrm{SR}}^j$.

(III) The algorithm switches to active play using $\hat{\mathbf{p}}^{\mathrm{fr}}$. The remaining rounds are then allocated exactly as in the budgeted-intervention algorithm of Section 3.3, with active fraction $1 - \frac{\tau_2}{T}$ and estimated distribution $\hat{\mathbf{p}}^{\mathrm{fr}}$. Specifically, for each $c > 0$, define

$$F(c) := \sum_{j \in [k]} \max\{\frac{\tau_2}{T}\hat{p}_j^{\mathrm{fr}},\ (\hat{p}_j^{\mathrm{fr}})^{2/3}c\},$$

let c^* be the unique value such that $F(c^*) = 1$, and define the vectors $\mathbf{q}$ and $\mathbf{r}$ by

$$q_j := \max\left\{\frac{\tau_2}{T}\hat{p}_j^{\mathrm{fr}},\ (\hat{p}_j^{\mathrm{fr}})^{2/3}c^*\right\}, \qquad r_j = \frac{q_j - \frac{\tau_2}{T}\hat{p}_j^{\mathrm{fr}}}{1 - \tau_2/T}. \tag{8}$$

The algorithm then uses the remaining rounds actively, selecting subpopulations according to $\mathbf{r}$. Algorithm 2 presents the pseudocode, and the next theorem states its regret guarantee.

Theorem 4.1 (Simple Regret of EETC)**.** *For sufficiently large horizon T, the worst-case simple regret of EETC is*

$$SR(\texttt{EETC}, T, \mathbf{p}) \in \mathcal{O}\big(\sqrt{\frac{n}{T}}\ \|\mathbf{p}\|_{2/3}\big).$$

Consequently, EETC up to a constant achieves the same simple regret rate as the optimal fully active policy that knows $\mathbf{p}$, and hence also matches the known-$\mathbf{p}$ lower bound up to constant factors. The dependence on $1/p_{\min}$ in the horizon condition can be removed by a thresholded variant of EETC that only estimates the sufficiently large coordinates of $\mathbf{p}$ (Appendix C). EETC can also be made horizon-free by running it in doubling epochs, which gives the same rate up to constants at every sufficiently large round (Appendix C.3).

5 Experiments

This section presents the experimental results on real-world data. Additional dataset details and further experiments are deferred to Appendix D.

Table 1: Simple regret on the MovieLens 1M instance [HK15]. Each entry reports mean $\pm$ 95% confidence interval over 50 independent runs. All values are multiplied by 100 for readability.

Method	$T = 2500$	$T = 5000$	$T = 10000$	$T = 15000$
Passive + Uniform	3.76 ± 0.53	2.05 ± 0.32	1.10 ± 0.23	0.75 ± 0.13
EETC + Uniform	3.52 ± 0.52	1.85 ± 0.32	0.81 ± 0.25	0.59 ± 0.13
Known-$\mathbf{p}$ Active + Uniform	$\mathbf{2.99 \pm 0.51}$	$\mathbf{1.78 \pm 0.37}$	$\mathbf{0.74 \pm 0.17}$	$\mathbf{0.56 \pm 0.14}$
Passive + UCB	2.53 ± 0.37	1.29 ± 0.23	0.71 ± 0.13	0.47 ± 0.11
EETC + UCB	2.57 ± 0.37	0.97 ± 0.17	$\mathbf{0.41 \pm 0.11}$	0.33 ± 0.10
Known-$\mathbf{p}$ Active + UCB	$\mathbf{2.28 \pm 0.46}$	$\mathbf{0.88 \pm 0.21}$	0.49 ± 0.12	$\mathbf{0.31 \pm 0.07}$

MovieLens Experiment. We evaluate the algorithms on a MovieLens 1M instance [HK15]. Treatments correspond to the five most frequently rated movies, and subpopulations correspond to the 14 demographic groups obtained from the Cartesian product of gender and age group. Thus, in this experiment, $n = 5$ and $k = 14$, yielding $nk = 70$ different treatment-subpopulation reward distributions. When an algorithm selects treatment a for subpopulation j, the environment samples uniformly with replacement from the historical ratings matching (a, j) and returns the sampled rating as the reward. The population weights p_j and the mean rewards $\mu_{a,j}$ are computed from the empirical frequencies and cell averages of the filtered dataset. The ratings are used as raw rewards in $\{1, \dots, 5\}$, without normalization to $[0, 1]$, so the reported regret is just a constant rescaling of the normalized regret. For more details see Appendix D.

Within each subpopulation, we instantiate the standard bandit subroutine in two ways: *Uniform*, which samples the available treatments uniformly, and *UCB*, which uses the standard upper-confidence-bound rule. Both are known to achieve the optimal simple regret rate up to logarithmic factors. We compare three policy classes: fully passive, fully active with known $\mathbf{p}$ using the optimal allocation proportional to $p_j^{2/3}$, and EETC. For each policy class, we report results with both subroutines, yielding six methods in total.

Results. Table 1 reports the simple regret after horizons $T \in \{2500, 5000, 10000, 15000\}$, averaged over 50 independent runs, together with confidence intervals. Across both subroutines, the passive policies perform worst, while EETC and the known-$\mathbf{p}$ active policies achieve clearly smaller regret. Moreover, as the horizon increases, the performance of EETC becomes very close to that of the known-p active benchmark. These results consistent with the theory showcase that estimating p from passive observations and then switching to an active allocation is more efficient than remaining fully passive, and that EETC can closely approach the performance of the optimal known-p active policy.

6 Conclusion

We studied the bandit with subpopulations problem with a context-weighted simple regret objective and quantified the benefit of actively choosing which subpopulation to sample. We showed that active subpopulation selection always improves performance over passive sampling, characterized the extent of this improvement, and proved the optimality of a simple history-free allocation rule up to constant factors. Moreover, we extended the analysis to settings with a limited intervention budget and unknown subpopulation distribution. In the latter case, we proposed the EETC algorithm that balances passive estimation with active allocation rounds and for long horizons matches the performance of the optimal fully active policy which knows the population distribution. An interesting direction for future work is to develop gap-dependent guarantees for the same problem. In such a setting, history-free policies are no longer optimal, since optimal algorithms may allocate fewer samples to subpopulations where the best treatment is easy to identify, leading to allocation rules that differ from the worst-case ones studied here. A natural candidate is an elimination-style policy that removes treatments once they are confidently suboptimal and stops sampling a subpopulation once a near-optimal treatment has been identified, with a confidence level that depends on the subpopulation weight p_j. Another possible extension is to consider models where the context is observed only after the treatment is chosen, inspired by related contextual and causal bandit settings [RKB$^+$21, LLR16, SARK25].

References

[AB09] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Colt*, pages 217–226, 2009.

[BBS09] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

[BCB+12] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[BKL26] Emma Brunskill, Ishani Karmarkar, and Zhaoqi Li. Active learning for stochastic contextual linear bandits. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas, editors, *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 577–585. PMLR, 2026.

[BMS09] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

[CHVDS23] Alicia Curth, Alihan Hüyük, and Mihaela Van Der Schaar. Adaptive identification of populations with treatment benefit in clinical trials: machine learning challenges and solutions. In *International Conference on Machine Learning*, pages 6603–6622. PMLR, 2023.

[Coc77] William G. Cochran. *Sampling Techniques*. John Wiley & Sons, 3 edition, 1977.

[DP09] Devdatt P Dubhashi and Alessandro Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[DSC+18] Aniket Anand Deshmukh, Srinagesh Sharma, James W Cutler, Mark Moldwin, and Clayton Scott. Simple regret minimization for contextual bandits. *arXiv preprint arXiv:1810.07371*, 2018.

[FJB+19] Adam Foster, Martin Jankowiak, Elias Bingham, Paul Horsfall, Yee Whye Teh, Thomas Rainforth, and Noah Goodman. Variational bayesian optimal experimental design. *Advances in neural information processing systems*, 32, 2019.

[GGLB11] Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Sébastien Bubeck. Multi-bandit best arm identification. *Advances in Neural Information Processing Systems*, 24, 2011.

[GKK16] Aurélien Garivier, Emilie Kaufmann, and Wouter M Koolen. Maximin action identification: A new bandit framework for games. In *Conference on Learning Theory*, pages 1028–1050. PMLR, 2016.

[HK15] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[HLP52] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 2 edition, 1952.

[JEK24] Fateme Jamshidi, Jalal Etesami, and Negar Kiyavash. Confounded budgeted causal bandits. In *Causal Learning and Reasoning*, pages 423–461. PMLR, 2024.

[JLB22] Matthew Jörke, Jonathan Lee, and Emma Brunskill. Simple regret minimization for contextual bandits using bayesian optimal experimental design. In *ICML2022 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2022.

[KIIK22] Masahiro Kato, Masaaki Imaizumi, Takuya Ishihara, and Toru Kitagawa. Best arm identification with contextual information under a small gap. *arXiv preprint arXiv:2209.07330*, 2022.

[KOIK24] Masahiro Kato, Kyohei Okumura, Takuya Ishihara, and Toru Kitagawa. Adaptive experimental design for policy learning. *arXiv preprint arXiv:2401.03756*, 2024.

[KZAB23] Sanath Kumar Krishnamurthy, Ruohan Zhan, Susan Athey, and Emma Brunskill. Proportional response: Contextual bandits for simple and cumulative regret minimization. *Advances in Neural Information Processing Systems*, 36:30255–30266, 2023.

[LB18] Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? *Advances in neural information processing systems*, 31, 2018.

[LBA+26] Shuai Liu, Alireza Bakhtiari, Alex Ayoub, Botao Hao, and Csaba Szepesvári. Efficient simple regret algorithms for stochastic contextual bandits. *arXiv preprint arXiv:2601.21167*, 2026.

[LLR16] Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in neural information processing systems*, 29, 2016.

[LMTY20] Yangyi Lu, Amirhossein Meisami, Ambuj Tewari, and William Yan. Regret analysis of bandit problems with causal background knowledge. In *Conference on Uncertainty in Artificial Intelligence*, pages 141–150. PMLR, 2020.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[MNS22] Aurghya Maiti, Vineet Nair, and Gaurav Sinha. A causal bandit approach to learning good atomic interventions in presence of unobserved confounders. In *Uncertainty in Artificial Intelligence*, pages 1328–1338. PMLR, 2022.

[Ney34] Jerzy Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934.

[NPS21] Vineet Nair, Vishakha Patil, and Gaurav Sinha. Budgeted and non-budgeted causal bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2017–2025. PMLR, 2021.

[PR13] Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *THE ANNALS of STATISTICS*, pages 693–721, 2013.

[RKB+21] Yoan Russac, Christina Katsimerou, Dennis Bohle, Olivier Cappé, Aurélien Garivier, and Wouter M Koolen. A/b/n testing with control in the presence of subpopulations. *Advances in Neural Information Processing Systems*, 34:25100–25110, 2021.

[RV24] Alessio Russo and Filippo Vannella. Fair best arm identification with fixed confidence. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pages 1173–1180. IEEE, 2024.

[RZ10] Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*, 2010.

[SARK25] Mohammad Shahverdikondori, Amir Mohammad Abouei, Alireza Rezaeimoghadam, and Negar Kiyavash. Optimal best arm identification with post-action context. *arXiv preprint arXiv:2502.03061*, 2025.

[SBC19] Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Overlapping multi-bandit best arm identification. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2544–2548. IEEE, 2019.

[SBK25] Mohammad Shahverdikondori, Mohammad Reza Badri, and Negar Kiyavash. Best group identification in multi-objective bandits. *arXiv preprint arXiv:2505.17869*, 2025.

[SEK25] Mohammad Shahverdikondori, Jalal Etesami, and Negar Kiyavash. Graph learning is suboptimal in causal bandits. *arXiv preprint arXiv:2510.16811*, 2025.

[WS22] Zhenlin Wang and Jonathan Scarlett. Max-min grouped bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8603–8611, 2022.

[WZZ23] Yuhang Wu, Zeyu Zheng, and Tingyu Zhu. Best arm identification with fairness constraints on subpopulations. In *2023 Winter Simulation Conference (WSC)*, pages 540–551. IEEE, 2023.

[ZSSJ23] Yao Zhao, Connor Stephens, Csaba Szepesvári, and Kwang-Sung Jun. Revisiting simple regret: Fast rates for returning a good arm. In *International Conference on Machine Learning*, pages 42110–42158. PMLR, 2023.

Appendix

The appendix is organized as follows. Section A provides a more detailed discussion of related work. Section B contains the proofs omitted from the main text. Section C gives finite-horizon guarantees beyond the data-rich regime and anytime versions of the algorithms. Finally, Section D presents additional experimental details and results.

A Further Discussion on Related Work

In this section, we provide some additional discussion of related work.

Heterogeneous Treatment Effects and Subpopulation-Aware Decision Making. A central line of work studies how treatment heterogeneity across subpopulations fundamentally alters optimal experimental design and decision rules. [RKB$^+$21] shows that ignoring heterogeneity across observed or latent subgroups can lead to systematically suboptimal or even misleading treatment selection, even when each arm is evaluated against a control. When outcomes vary across groups, the problem shifts from selecting a globally best arm to learning *group-conditional policies*, effectively turning A/B/n testing into a policy learning problem. This perspective is further developed in adaptive clinical trial settings by [CHVDS23], who propose methods to sequentially identify subpopulations with positive treatment effects under strict statistical constraints. Their work emphasizes adaptive refinement of target populations via procedures such as AdaGGI and AdaGCPI, focusing not on a single optimal subgroup but on any collection of beneficiaries.

Complementing this line, [WZZ23] and [RV24] study best arm identification under *fairness* constraints across subpopulations. Unlike [RKB$^+$21], which allows fully personalized policies, fairness requirements couple subpopulations by imposing performance constraints across groups. This fundamentally changes the decision problem: the goal is no longer purely optimal identification but balancing optimality with worst-case guarantees across groups, at the cost of increased sample complexity.

Contextual and Policy Learning Bandits. A related but distinct literature incorporates observable covariates into sequential decision making. [PR13] extends the classical bandit setting to contextual information, where the goal is to learn a context-dependent policy rather than a single best arm. Under smoothness assumptions, they show that partition-based methods achieve minimax optimal regret by learning locally optimal actions within regions of the covariate space.

In the small gap regime, [KIIK22] further studies best arm identification with context, showing that contextual structure can significantly reduce sample complexity when arms are hard to distinguish. The key difference is that efficiency gains arise only when contexts meaningfully separate arms; otherwise, learning remains intrinsically difficult. More generally, [KOIK24] shows that even in contextual settings, standard experimental designs are suboptimal when the goal is policy learning rather than estimation, and proposes adaptive designs that concentrate sampling in regions where policy decisions are most uncertain. Similarly, [DSC$^+$18, JLB22] frame simple regret minimization in contextual bandits as a Bayesian experimental design problem where exploration is guided by expected information gain about the optimal policy rather than reward maximization.

Beyond bandits, Bayesian optimal experimental design provides a general framework for sequential data acquisition. [FJB$^+$19] develops scalable variational approximations to expected information gain, enabling BOED in complex models where exact computation is infeasible. These methods connect naturally to contextual bandits and policy learning approaches that also select experiments based on information gain rather than immediate reward. Classical stratified sampling is also related: Neyman allocation [Ney34, Coc77] splits a sampling budget across strata to minimize the variance of a single aggregate estimator, $\sum_j p_j^2\sigma_j^2/T_j$, which gives $T_j \propto p_j\sigma_j$. Our objective instead weights the per-subpopulation errors, $\sum_j p_j/\sqrt{T_j}$, which yields $T_j \propto p_j^{2/3}$; the same allocation would arise for any population-weighted estimation error decaying as $1/\sqrt{T_j}$, so it is not specific to bandits. What is specific to the bandit setting is the decision problem of recommending a treatment per subpopulation under worst-case rewards, for which we show that history-free allocations are sufficient and prove matching lower bounds over all adaptive policies.

A complementary perspective arises in supervised learning under covariate shift which could be the result of different subpopulations. [BBS09] shows that distribution mismatch between training and

test covariates biases standard empirical risk minimization, and proposes density-ratio reweighting to correct this issue. While outside the sequential decision-making setting, this line of work shares with contextual bandits and adaptive design the central idea that data collection and inference must be aligned with the target decision environment.

Bandits with Grouped Arms. Our setting is also related to bandit problems with grouped arms. Indeed, each subpopulation can be viewed as a group of n treatment arms, so the learner effectively faces k groups and seeks to identify the best arm within each group. The closest line of work is multi-bandit best arm identification [GGLB11, SBC19], which studies identifying the best arm in each of several bandits with the objective of minimizing the probability of misidentification. This is particularly close to our fully active setting, since both problems require allocating samples across several independent bandit instances. However, those works focus on best arm identification, whereas our objective is to minimize context-weighted simple regret.

Other grouped arm formulations consider different objectives. Maximin action identification [GKK16] and max-min grouped bandits [WS22] study settings in which performance is determined by the worst arm or outcome within a group, while best group identification in multi-objective bandits [SBK25] focuses on identifying a single best group under multi-objective rewards. In contrast, our goal is not to identify one best group or action, but to recommend the best arm separately within every group. In addition, we characterize the gain obtained by actively choosing which group, that is, which subpopulation, to sample.

Causal Bandits. Another related line of work is the general framework of causal bandits [LLR16, LB18, LMTY20, JEK24], where actions correspond to interventions on subsets of variables in a causal graph and the learner observes the reward together with the non-intervened variables; see [LLR16] for details. Our setting can be viewed as a simple causal bandit problem with three nodes: a treatment node, a subpopulation node, and a reward node, where the reward is affected by both the treatment and the subpopulation, while the two parent variables are independent. Under this interpretation, the passive setting corresponds to intervening only on the treatment node, whereas the active setting corresponds to intervening on both the treatment and subpopulation nodes.

A related class of models in this literature is the *parallel graph* setting, in which several independent variables all directly affect the reward [LLR16, NPS21]. In [LLR16] the authors provide algorithms with simple regret guarantees for binary variables, while [NPS21] and [MNS22] study more general no-backdoor and general graphs and settings with intervention costs, and derive both simple and cumulative regret guarantees. Another related work studies best arm identification in a setting where the context is not independent of the treatment but is instead a random function of the chosen treatment [SARK25].

These works are structurally related to ours, but differ substantially in objective and in the type of interventions being compared. In particular, they focus on identifying a single globally best action, whereas our goal is to identify the best treatment separately for each subpopulation and to characterize the gain obtained by actively choosing both the treatment and the subpopulation, rather than only one of them. Another recent work has also studied the effect of the number of intervened nodes at each round, but the results are on cumulative regret in a setting where the causal structure is unknown [SEK25]. Extending our results to settings with multiple variables and richer causal structures would be an interesting future direction.

B Omitted Proofs

This section presents the proofs omitted from the main text due to space constraints. We first start with a Lemma that will be used throughout the proofs.

Lemma B.1 (Concentration of Independent Bernoulli Counts)**.** *Let $X_1, \dots, X_m$ be independent Bernoulli random variables with means $\alpha_1, \dots, \alpha_m \in [0,1]$. Define*

$$S_m \coloneqq \sum_{t=1}^{m} X_t, \qquad M_m \coloneqq \mathbb{E}[S_m] = \sum_{t=1}^{m} \alpha_t.$$

If, for some $\eta \in (0,1)$,

$$M_m \geq 24 \ln\left(\frac{2}{\eta}\right),$$

then

$$\mathbb{P}\left(\frac{M_m}{2} \le S_m \le \frac{3M_m}{2}\right) \ge 1-\eta.$$

Proof. Since $X_1, \dots, X_m$ are independent Bernoulli random variables, possibly with different means, the multiplicative Chernoff bounds apply; see, e.g., [DP09]. In particular,

$$\mathbb{P}\left(S_m < \frac{M_m}{2}\right) \le \exp\left(-\frac{M_m}{8}\right),$$
$$\mathbb{P}\left(S_m > \frac{3M_m}{2}\right) \le \exp\left(-\frac{M_m}{12}\right).$$

Therefore, by a union bound,

$$\begin{aligned}\mathbb{P}\left(\frac{M_m}{2} \le S_m \le \frac{3M_m}{2}\right) &\ge 1 - \exp\left(-\frac{M_m}{8}\right) - \exp\left(-\frac{M_m}{12}\right) \\ &\ge 1 - 2\exp\left(-\frac{M_m}{24}\right).\end{aligned}$$

Since $M_m \ge 24\ln(2/\eta)$, it follows that

$$2\exp\left(-\frac{M_m}{24}\right) \le \eta,$$

and hence

$$\mathbb{P}\left(\frac{M_m}{2} \le S_m \le \frac{3M_m}{2}\right) \ge 1-\eta.$$

□

Next, we introduce inequalities on the norms of vectors in $\mathbb{R}^k$. These inequalities are standard; see, for example, [HLP52, Chapter 2].

Theorem B.2 (Comparison and Interpolation of ℓ_p Norms)**.** *For any vector $x \in \mathbb{R}^k$ and any exponents $0 < p < q < \infty$, we have*

$$\|x\|_q \le \|x\|_p \le k^{\frac{1}{p}-\frac{1}{q}} \|x\|_q.$$

More generally, for any $0 < p \le r \le q < \infty$ and any $\theta \in [0,1]$ such that

$$\frac{1}{r} = \frac{\theta}{p} + \frac{1-\theta}{q},$$

we have

$$\|x\|_r \le \|x\|_p^{\theta} \|x\|_q^{1-\theta}.$$

B.1 Proofs of Section 3

B.1.1 Proof of Lemma 3.1

Lemma 3.1 (Simple Regret Lower Bound)**.** *For any subpopulation distribution $\mathbf{p} \in \Delta_+^{k-1}$ known to the agent and any policy $\pi \in \Pi(\mathbf{p})$,*

$$SR(\pi, T, \mathbf{p}) \in \Omega\Big(\sqrt{\frac{n}{T}}\, \|\mathbf{p}\|_{2/3}\Big).$$

Proof. By Proposition C.1, every policy π satisfies $\mathrm{SR}(\pi, T, \mathbf{p}) \ge cV_T(\mathbf{p})$. Equation (37) gives

$$V_T(\mathbf{p}) \ge \sqrt{\frac{n}{T+nk}}\, \|\mathbf{p}\|_{2/3}.$$

Under the data-rich assumption, $T \ge nk$, so $T + nk \le 2T$. The claimed lower bound follows. □

B.1.2 Proof of Lemma 3.2

Lemma 3.2 (Simple Regret of History-Free Policies)**.** *For any history-free policy $\pi \in \Pi_{\mathrm{hf}}(\mathbf{p})$ along with the subroutine $\mathcal{A}_{\mathrm{SR}}$ and expected subpopulation proportions $\mathbf{q}(\pi, \mathbf{p})$, if $\forall j \in [k] : Tq(\pi, \mathbf{p})_j > 24 \ln(2kT)$, then*

$$\mathit{SR}(\pi, T, \mathbf{p}) \in \Theta\Big(\sqrt{\frac{n}{T}} \sum_{j \in [k]} \frac{p_j}{\sqrt{q(\pi, \mathbf{p})_j}}\Big).$$

Proof. For a history-free policy $\pi \in \Pi_{\mathrm{hf}}(\mathbf{p})$, let

$$T_j \coloneqq \sum_{t=1}^{T} \mathbf{1}\{C_t = j\}$$

denote the number of rounds in which subpopulation j is observed. Since the subpopulation selection rule is history-free, there exist deterministic probabilities $\{q_{t,j}\}_{t \in [T],\, j \in [k]}$, depending only on $\mathbf{p}$, such that

$$\mathbb{P}(C_t = j) = q_{t,j}, \qquad \sum_{j \in [k]} q_{t,j} = 1 \quad \forall t \in [T].$$

Hence T_j is a sum of independent Bernoulli random variables (by independence of C_t across rounds), and

$$\mathbb{E}[T_j] = \sum_{t=1}^{T} q_{t,j} = Tq(\pi, \mathbf{p})_j.$$

We first prove the lower bound. We use the hard-instance prior from the proof of Proposition C.1 in Appendix C. Its derivation of (34) only uses $\delta_j \in (0, 1/8]$, so (34) holds for any such gaps, where $\bar{N}_j$ is the expected number of pulls in subpopulation j under the prior. For a history-free policy, the subpopulation selection rule is independent of the realized rewards and of the instance, so $\bar{N}_j = Tq(\pi, \mathbf{p})_j$. Choose $\delta_j = \min\{1/8, \sqrt{n/(c_1 Tq(\pi, \mathbf{p})_j)}\}$. Then the exponent in (34) is at most 1, and since $Tq(\pi, \mathbf{p})_j \gtrsim n$ (which is part of the data-rich regime), $\delta_j \geq c\sqrt{n/(Tq(\pi, \mathbf{p})_j)}$ for a universal $c > 0$. Hence

$$\mathrm{SR}(\pi, T, \mathbf{p}) \in \Omega\left(\sqrt{\frac{n}{T}} \sum_{j \in [k]} \frac{p_j}{\sqrt{q(\pi, \mathbf{p})_j}}\right). \tag{9}$$

We now prove the upper bound. For each $j \in [k]$, define the event

$$\mathcal{E}_j \coloneqq \left[\frac{Tq(\pi, \mathbf{p})_j}{2} \leq T_j \leq \frac{3Tq(\pi, \mathbf{p})_j}{2}\right],$$

and let

$$\mathcal{E} \coloneqq \bigcap_{j \in [k]} \mathcal{E}_j.$$

Since T_j is a sum of independent Bernoulli random variables with mean $Tq(\pi, \mathbf{p})_j$, Lemma B.1 with $\eta = 1/kT$ implies that

$$\mathbb{P}(\mathcal{E}_j) \geq 1 - \frac{1}{kT},$$

where we use the assumption $Tq(\pi, \mathbf{p})_j > 24 \ln(2kT)$.

Therefore, by a union bound,

$$\mathbb{P}(\mathcal{E}^c) \leq \sum_{j \in [k]} \mathbb{P}(\mathcal{E}_j^c) \leq \frac{1}{T}. \tag{10}$$

Fix any instance $(\boldsymbol{\nu}, \mathbf{p})$. On the event $\mathcal{E}$, each subpopulation j is observed at least $Tq(\pi, \mathbf{p})_j/2$ times. Since the subroutine $\mathcal{A}_{\mathrm{SR}}$ is anytime and achieves the optimal simple regret rate in Lemma 2.1, there

exists a universal constant $c_1 > 0$ such that the expected recommendation gap in subpopulation j is at most

$$c_1\sqrt{\frac{n}{T_j}} \le c_1\sqrt{\frac{2n}{Tq(\pi,\mathbf{p})_j}}.$$

Therefore, on $\mathcal{E}$,

$$\begin{aligned}\mathrm{SR}(\pi, T, (\boldsymbol{\nu}, \mathbf{p})) &= \sum_{j\in[k]} p_j\Delta_j \\ &\le c_1\sqrt{\frac{2n}{T}}\sum_{j\in[k]}\frac{p_j}{\sqrt{q(\pi,\mathbf{p})_j}}.\end{aligned}$$

On the complement event $\mathcal{E}^c$, the simple regret is at most 1, since rewards lie in $[0,1]$ and $\sum_j p_j = 1$. Hence

$$\mathrm{SR}(\pi, T, (\boldsymbol{\nu}, \mathbf{p})) \le c_1\sqrt{\frac{2n}{T}}\sum_{j\in[k]}\frac{p_j}{\sqrt{q(\pi,\mathbf{p})_j}} + \mathbb{P}(\mathcal{E}^c).$$

Taking the supremum over all instances and using (10), we obtain

$$\mathrm{SR}(\pi, T, \mathbf{p}) \le c_1\sqrt{\frac{2n}{T}}\sum_{j\in[k]}\frac{p_j}{\sqrt{q(\pi,\mathbf{p})_j}} + \frac{1}{T}.$$

Finally, since $q(\pi,\mathbf{p})_j \le 1$ for all j, we have

$$\sum_{j\in[k]}\frac{p_j}{\sqrt{q(\pi,\mathbf{p})_j}} \ge \sum_{j\in[k]} p_j = 1,$$

and therefore the term $1/T$ is of smaller order than the main term. Thus,

$$\mathrm{SR}(\pi, T, \mathbf{p}) \in O\left(\sqrt{\frac{n}{T}}\sum_{j\in[k]}\frac{p_j}{\sqrt{q(\pi,\mathbf{p})_j}}\right). \tag{11}$$

Combining (9) and (11) proves the claim. □

B.1.3 Proof of Lemma 3.3

Lemma 3.3 (Optimal Allocation for Active Policies)**.** *For any $\mathbf{p} \in \Delta_+^{k-1}$, the unique minimizer of $S_{\mathbf{p}}(\mathbf{q})$ over $\mathbf{q} \in \Delta_+^{k-1}$ is*

$$q^*(\mathbf{p})_j = \frac{p_j^{2/3}}{\sum_{\ell\in[k]} p_\ell^{2/3}}, \quad j \in [k].$$

Moreover, $S_{\mathbf{p}}(\mathbf{q}^*(\mathbf{p})) = \|\mathbf{p}\|_{2/3}$.

Proof. Recall that

$$S_{\mathbf{p}}(\mathbf{q}) = \sum_{j\in[k]}\frac{p_j}{\sqrt{q_j}}, \qquad \mathbf{q} \in \Delta_+^{k-1}.$$

Since $p_j > 0$ for all $j \in [k]$, each term $p_j q_j^{-1/2}$ is strictly convex on $(0,\infty)$. Therefore, $S_{\mathbf{p}}$ is strictly convex on Δ_+^{k-1}. Hence, if a minimizer exists, it is unique.

We now characterize this minimizer. Since $p_j > 0$ for all j, we have

$$\frac{p_j}{\sqrt{q_j}} \to +\infty \qquad \text{as } q_j \downarrow 0.$$

Therefore, the minimum over the simplex cannot be attained on the boundary, and the minimizer must lie in the interior. Thus, we may use the first order optimality conditions for the equality-constrained problem

$$\begin{aligned} \min_{\mathbf{q}\in\mathbb{R}^k} \quad & \sum_{j\in[k]} \frac{p_j}{\sqrt{q_j}} \\ \text{s.t.} \quad & \sum_{j\in[k]} q_j = 1, \qquad q_j > 0 \ \ \forall j \in [k]. \end{aligned}$$

Consider the Lagrangian

$$\mathcal{L}(\mathbf{q}, \lambda) = \sum_{j\in[k]} p_j q_j^{-1/2} + \lambda \left(\sum_{j\in[k]} q_j - 1 \right).$$

At the optimum, for each $j \in [k]$,

$$\frac{\partial \mathcal{L}}{\partial q_j} = -\frac{1}{2} p_j q_j^{-3/2} + \lambda = 0.$$

Hence

$$q_j^{-3/2} = \frac{2\lambda}{p_j},$$

which implies

$$q_j = \left(\frac{p_j}{2\lambda} \right)^{2/3}.$$

Therefore, there exists a constant $c > 0$ such that

$$q_j = c\, p_j^{2/3} \qquad \forall j \in [k].$$

Using the constraint $\sum_{j\in[k]} q_j = 1$, we obtain

$$1 = \sum_{j\in[k]} q_j = c \sum_{j\in[k]} p_j^{2/3},$$

and thus

$$c = \frac{1}{\sum_{l\in[k]} p_l^{2/3}}.$$

Substituting back gives

$$q_j^*(\mathbf{p}) = \frac{p_j^{2/3}}{\sum_{l\in[k]} p_l^{2/3}}, \qquad j \in [k].$$

Since $S_{\mathbf{p}}$ is strictly convex on Δ_+^{k-1}, this minimizer is unique.

Then, by substituting this back to $S_{\mathbf{p}}(\mathbf{q})$, we obtain

$$S_{\mathbf{p}}(\mathbf{q}^*(\mathbf{p})) = \left(\sum_{j\in[k]} p_j^{2/3} \right)^{3/2} = \|\mathbf{p}\|_{2/3}.$$

□

B.1.4 Proof of Lemma 3.5

Lemma 3.5 (Active–Passive Gap). *For every* $\mathbf{p} \in \Delta_+^{k-1}$, $R(\mathbf{p}) \geq 1$. *Moreover,* $\sup_{\mathbf{p}\in\Delta_+^{k-1}} R(\mathbf{p}) \in \Theta\big(k^{\frac{1}{4}}\big)$, *and this order is tight.*

Proof. Recall that

$$R(\mathbf{p}) = \frac{\|\mathbf{p}\|_{1/2}^{1/2}}{\|\mathbf{p}\|_{2/3}}.$$

We first show that $R(\mathbf{p}) \geq 1$ for every $\mathbf{p} \in \Delta_+^{k-1}$. Let $a_j := \sqrt{p_j}$ for all $j \in [k]$, and let $\mathbf{a} = (a_1, \ldots, a_k)$. Then

$$\|\mathbf{a}\|_1 = \sum_{j\in[k]} \sqrt{p_j} = \|\mathbf{p}\|_{1/2}^{1/2}, \qquad \|\mathbf{a}\|_2^2 = \sum_{j\in[k]} p_j = 1,$$

and

$$\|\mathbf{a}\|_{4/3}^2 = \left(\sum_{j\in[k]} p_j^{2/3}\right)^{3/2} = \|\mathbf{p}\|_{2/3}.$$

Applying the interpolation part of Theorem B.2 with $p = 1$, $r = 4/3$, $q = 2$, and $\theta = 1/2$, we obtain

$$\|\mathbf{a}\|_{4/3} \leq \|\mathbf{a}\|_1^{1/2} \|\mathbf{a}\|_2^{1/2}.$$

Since $\|\mathbf{a}\|_2 = 1$, this yields

$$\|\mathbf{p}\|_{2/3} = \|\mathbf{a}\|_{4/3}^2 \leq \|\mathbf{a}\|_1 = \|\mathbf{p}\|_{1/2}^{1/2},$$

and hence $R(\mathbf{p}) \geq 1$.

For the upper bound, applying the comparison part of Theorem B.2 with $p = 1/2$ and $q = 2/3$ gives

$$\|\mathbf{p}\|_{1/2} \leq k^{1/2} \|\mathbf{p}\|_{2/3}.$$

Taking square roots and using $\|\mathbf{p}\|_{2/3} \geq \|\mathbf{p}\|_1 = 1$, we obtain

$$\|\mathbf{p}\|_{1/2}^{1/2} \leq k^{1/4} \|\mathbf{p}\|_{2/3},$$

which implies $R(\mathbf{p}) \leq k^{1/4}$.

It remains to prove the matching lower bound. For $k = 1$, the claim is trivial, so assume $k \geq 2$. Consider the family of probability vectors

$$\mathbf{p}^{(\varepsilon)} = \left(1 - \varepsilon, \frac{\varepsilon}{k-1}, \ldots, \frac{\varepsilon}{k-1}\right), \qquad \varepsilon \in (0, 1).$$

Choose

$$\varepsilon = \frac{1}{\sqrt{k-1}}.$$

Then

$$\left\|\mathbf{p}^{(\varepsilon)}\right\|_{1/2}^{1/2} = \sqrt{1-\varepsilon} + (k-1)\sqrt{\frac{\varepsilon}{k-1}} = \sqrt{1-\varepsilon} + \sqrt{\varepsilon(k-1)} \geq (k-1)^{1/4}.$$

On the other hand,

$$\begin{aligned}\left\|\mathbf{p}^{(\varepsilon)}\right\|_{2/3} &= \left((1-\varepsilon)^{2/3} + (k-1)\left(\frac{\varepsilon}{k-1}\right)^{2/3}\right)^{3/2} \\ &= \left((1-\varepsilon)^{2/3} + (k-1)^{1/3}\varepsilon^{2/3}\right)^{3/2}.\end{aligned}$$

With $\varepsilon = (k-1)^{-1/2}$, we have

$$(k-1)^{1/3}\varepsilon^{2/3} = 1,$$

and therefore

$$\left\|\mathbf{p}^{(\varepsilon)}\right\|_{2/3} = \left((1-\varepsilon)^{2/3} + 1\right)^{3/2} \leq 2^{3/2} = \sqrt{8}.$$

Combining the two bounds yields

$$R\left(\mathbf{p}^{(\varepsilon)}\right) = \frac{\left\|\mathbf{p}^{(\varepsilon)}\right\|_{1/2}^{1/2}}{\left\|\mathbf{p}^{(\varepsilon)}\right\|_{2/3}} \geq \frac{(k-1)^{1/4}}{\sqrt{8}}.$$

Hence

$$\sup_{\mathbf{p}\in\Delta_+^{k-1}} R(\mathbf{p}) \in \Omega\left(k^{1/4}\right).$$

Combining the upper and lower bounds, we conclude that

$$\sup_{\mathbf{p}\in\Delta_+^{k-1}} R(\mathbf{p}) \in \Theta\left(k^{1/4}\right),$$

and this order is tight. □

B.1.5 Proof of Theorem 3.6

Theorem 3.6 (Simple Regret of Algorithm 1)**.** *For every $\mathbf{p} \in \Delta_+^{k-1}$, Algorithm 1 is a history-free policy in $\Pi_\alpha(\mathbf{p})$ with expected subpopulation proportions equal to $\mathbf{q}_\alpha^*(\mathbf{p})$. Consequently, when T is large, its worst-case simple regret satisfies the optimal rate*

$$SR(\texttt{Alg. 1}, T, \mathbf{p}) \in \mathcal{O}\big(\sqrt{\frac{n}{T}}\, v_\alpha^*(\mathbf{p})\big).$$

Proof. For simplicity, assume that αT is an integer. Replacing αT by $\lfloor \alpha T \rfloor$ changes the number of passive and active rounds by at most one and therefore does not affect the regret rate.

We first characterize the optimizer of

$$\begin{aligned} \min_{\mathbf{q}\in\Delta_+^{k-1}} \quad & S_{\mathbf{p}}(\mathbf{q}) = \sum_{j\in[k]} \frac{p_j}{\sqrt{q_j}} \\ \text{s.t.} \quad & q_j \geq (1-\alpha)p_j, \qquad \forall j \in [k]. \end{aligned} \tag{12}$$

When $\alpha = 0$, the constraints reduce to $q_j \geq p_j$ for all j, and since $\sum_j q_j = \sum_j p_j = 1$, the feasible set contains only the point $\mathbf{q} = \mathbf{p}$. Hence $\mathbf{q}_\alpha^*(\mathbf{p}) = \mathbf{p}$, and the claim is immediate. We therefore assume in the remainder of the proof that $\alpha \in (0, 1]$.

Since Δ_+^{k-1} is not compact, we instead consider the same problem over the closed simplex Δ^{k-1}. The feasible set

$$\mathcal{Q}_\alpha \coloneqq \left[\mathbf{q} \in \Delta^{k-1} : q_j \geq (1-\alpha)p_j,\ \forall j \in [k]\right]$$

is compact and nonempty, so the minimum is attained on $\mathcal{Q}_\alpha$. Moreover, because each $p_j > 0$,

$$\frac{p_j}{\sqrt{q_j}} \to +\infty \qquad \text{as } q_j \downarrow 0.$$

Hence the minimizer cannot lie on the boundary $\{q_j = 0\}$, and therefore lies in Δ_+^{k-1}. Thus the minimizer over $\mathcal{Q}_\alpha$ is also the minimizer of (12).

The objective is strictly convex on Δ_+^{k-1}, since each map $q_j \mapsto p_j q_j^{-1/2}$ is strictly convex on $(0, \infty)$. Hence the minimizer is unique.

We now derive its form. Let

$$b_j \coloneqq (1-\alpha)p_j, \qquad j \in [k].$$

By the KKT conditions, there exist $\lambda \in \mathbb{R}$ and multipliers $\nu_j \geq 0$ such that

$$-\frac{1}{2}p_j q_j^{-3/2} + \lambda - \nu_j = 0, \qquad \nu_j(b_j - q_j) = 0, \qquad q_j \geq b_j.$$

Therefore, if $q_j > b_j$, then $\nu_j = 0$ and

$$q_j = \left(\frac{p_j}{2\lambda}\right)^{2/3} = c\, p_j^{2/3}$$

for some constant $c > 0$. If $q_j = b_j$, then the lower bound constraint is active. Hence the unique optimizer must be of the form

$$q^*_\alpha(\mathbf{p})_j = \max\big((1-\alpha)p_j,\ c^* p_j^{2/3}\big), \qquad j \in [k],$$

for some constant $c^* \geq 0$.

To determine c^*, define

$$F(c) \coloneqq \sum_{j\in[k]} \max\big((1-\alpha)p_j,\ cp_j^{2/3}\big).$$

Then F is continuous and nondecreasing on $[0,\infty)$. Moreover, since $\alpha > 0$,

$$F(0) = \sum_{j\in[k]} (1-\alpha)p_j = 1-\alpha < 1,$$

and clearly $F(c) \to \infty$ as $c \to \infty$. Hence there exists at least one $c^* \geq 0$ such that $F(c^*) = 1$.

We now show uniqueness. Let $p_{\min} \coloneqq \min_j p_j$. If $c < p_{\min}^{1/3}$, then for every $j \in [k]$,

$$cp_j^{2/3} < p_{\min}^{1/3} p_j^{2/3} \leq p_j, \qquad (1-\alpha)p_j < p_j,$$

and therefore

$$\max\big((1-\alpha)p_j,\ cp_j^{2/3}\big) < p_j.$$

Summing over j gives $F(c) < 1$ for all $c < p_{\min}^{1/3}$. On the other hand, for $c \geq p_{\min}^{1/3}$, the coordinate corresponding to $p_{\min}$ is already on the linear branch, so F has strictly positive slope and is therefore strictly increasing on $[p_{\min}^{1/3}, \infty)$. Since $F(c) < 1$ for all $c < p_{\min}^{1/3}$ and F is strictly increasing afterwards, the equation $F(c) = 1$ has a unique solution c^*.

Thus,

$$q^*_\alpha(\mathbf{p})_j = \max\big((1-\alpha)p_j,\ c^* p_j^{2/3}\big), \qquad j \in [k],$$

is indeed the unique optimizer of (12).

We now show that Algorithm 1 has expected subpopulation proportions equal to $\mathbf{q}^*_\alpha(\mathbf{p})$. By construction, the algorithm computes the same constant c^* and sets

$$q_j^* = \max\big((1-\alpha)p_j,\ c^* p_j^{2/3}\big) = q^*_\alpha(\mathbf{p})_j, \qquad r_j = \frac{q_j^* - (1-\alpha)p_j}{\alpha}.$$

Since $q_j^* \geq (1-\alpha)p_j$, we have $r_j \geq 0$, and

$$\sum_{j\in[k]} r_j = \frac{\sum_{j\in[k]} q_j^* - (1-\alpha)\sum_{j\in[k]} p_j}{\alpha} = \frac{1-(1-\alpha)}{\alpha} = 1,$$

so $\mathbf{r} \in \Delta_+^{k-1}$.

During the first $(1-\alpha)T$ rounds, the algorithm samples subpopulations according to $\mathbf{p}$, hence the expected number of observations of subpopulation j in this phase is

$$(1-\alpha)Tp_j.$$

During the remaining αT rounds, the algorithm samples according to $\mathbf{r}$, so the expected number of observations of subpopulation j in the active phase is

$$\alpha T r_j = \alpha T \cdot \frac{q_j^* - (1-\alpha)p_j}{\alpha} = T\big(q_j^* - (1-\alpha)p_j\big).$$

Therefore the total expected number of observations of subpopulation j is

$$(1-\alpha)Tp_j + T\big(q_j^* - (1-\alpha)p_j\big) = Tq_j^*.$$

Hence the expected subpopulation proportions of Algorithm 1 are exactly

$$\mathbf{q}(\texttt{Alg.1}, \mathbf{p}) = \mathbf{q}^* = \mathbf{q}^*_\alpha(\mathbf{p}).$$

The subpopulation selection rule of the algorithm depends only on $\mathbf{p}$, α, and the deterministic vector $\mathbf{r}$. It is therefore independent of the observed history, so the algorithm is history-free. Moreover, it is passive for exactly $(1-\alpha)T$ rounds and active for exactly αT rounds, hence it belongs to $\Pi_\alpha(\mathbf{p})$.

We next verify the condition of Lemma 3.2. Since $c^* \geq p_{\min}^{1/3}$ by the argument above, for every $j \in [k]$,

$$q_\alpha^*(\mathbf{p})_j = \max\left((1-\alpha)p_j,\ c^* p_j^{2/3}\right) \geq c^* p_j^{2/3} \geq p_{\min}^{1/3} p_j^{2/3} \geq p_{\min}.$$

Thus

$$\min_{j\in[k]} q_\alpha^*(\mathbf{p})_j \geq p_{\min}.$$

Since we work in the data rich regime and assume $T \gg 1/p_{\min}$, it follows that,

$$T q_\alpha^*(\mathbf{p})_j > 24\ln(2kT), \qquad \forall j \in [k].$$

We then apply Lemma 3.2 with $\mathbf{q}(\pi, \mathbf{p}) = \mathbf{q}_\alpha^*(\mathbf{p})$ to obtain

$$\begin{aligned}\mathrm{SR}(\texttt{Alg.1}, T, \mathbf{p}) &\in \mathcal{O}\Big(\sqrt{\frac{n}{T}} \sum_{j\in[k]} \frac{p_j}{\sqrt{q_\alpha^*(\mathbf{p})_j}}\Big) \\ &= \mathcal{O}\Big(\sqrt{\frac{n}{T}}\, S_{\mathbf{p}}(\mathbf{q}_\alpha^*(\mathbf{p}))\Big).\end{aligned}$$

By definition of $v_\alpha^*(\mathbf{p})$ as the optimal value of (12),

$$S_{\mathbf{p}}(\mathbf{q}_\alpha^*(\mathbf{p})) = v_\alpha^*(\mathbf{p}).$$

Hence

$$\mathrm{SR}(\texttt{Alg.1}, T, \mathbf{p}) \in \mathcal{O}\Big(\sqrt{\frac{n}{T}}\, v_\alpha^*(\mathbf{p})\Big).$$

This proves the theorem. □

B.1.6 Proof of Lemma 3.7

Lemma 3.7 (Threshold Budget $\alpha_{\min}(\mathbf{p})$)**.** *For a given $\mathbf{p} \in \Delta_+^{k-1}$, let $\alpha_{min}(\mathbf{p}) := 1 - p_{\max}^{-1/3} / \sum_{j\in[k]} p_j^{2/3}$, where $p_{\max} = \max_j p_j$. Then, for every $\alpha \in [\alpha_{min}(\mathbf{p}), 1]$, we have $v_\alpha^*(\mathbf{p}) = \|\mathbf{p}\|_{2/3}$. Consequently, for every such α, the simple regret bound of Algorithm 1 matches the optimal rate of fully active policies, which is also the optimal rate over all policies.*

Proof. Recall that the optimal value $v_\alpha^*(\mathbf{p})$ is defined as the minimum of

$$\begin{aligned}\min_{\mathbf{q}\in\Delta_+^{k-1}} \quad & S_{\mathbf{p}}(\mathbf{q}) = \sum_{j\in[k]} \frac{p_j}{\sqrt{q_j}} \\ \text{s.t.} \quad & q_j \geq (1-\alpha)p_j, \qquad \forall j \in [k].\end{aligned} \tag{13}$$

By Lemma 3.3, the unique minimizer of the unconstrained problem over Δ_+^{k-1} is

$$q_j^{\mathrm{act}} := \frac{p_j^{2/3}}{\sum_{\ell\in[k]} p_\ell^{2/3}}, \qquad j \in [k],$$

and its objective value is

$$S_{\mathbf{p}}(\mathbf{q}^{\mathrm{act}}) = \|\mathbf{p}\|_{2/3}.$$

We claim that $\mathbf{q}^{\mathrm{act}}$ is feasible for (13) whenever $\alpha \geq \alpha_{\min}(\mathbf{p})$. Indeed, the feasibility condition is

$$q_j^{\mathrm{act}} \geq (1-\alpha)p_j, \qquad \forall j \in [k],$$

which is equivalent to

$$\frac{p_j^{2/3}}{\sum_{\ell\in[k]} p_\ell^{2/3}} \geq (1-\alpha)p_j, \qquad \forall j \in [k].$$

Since $p_j > 0$, dividing by p_j gives

$$\frac{p_j^{-1/3}}{\sum_{\ell\in[k]} p_\ell^{2/3}} \geq 1-\alpha, \qquad \forall j \in [k].$$

Equivalently,

$$1-\alpha \leq \min_{j\in[k]} \frac{p_j^{-1/3}}{\sum_{\ell\in[k]} p_\ell^{2/3}}.$$

Now the function $x \mapsto x^{-1/3}$ is decreasing on $(0,\infty)$, so

$$\min_{j\in[k]} p_j^{-1/3} = p_{\max}^{-1/3},$$

where $p_{\max} = \max_j p_j$. Therefore,

$$\min_{j\in[k]} \frac{p_j^{-1/3}}{\sum_{\ell\in[k]} p_\ell^{2/3}} = \frac{p_{\max}^{-1/3}}{\sum_{\ell\in[k]} p_\ell^{2/3}}.$$

Hence the feasibility condition becomes

$$1-\alpha \leq \frac{p_{\max}^{-1/3}}{\sum_{\ell\in[k]} p_\ell^{2/3}},$$

or equivalently

$$\alpha \geq 1 - \frac{p_{\max}^{-1/3}}{\sum_{\ell\in[k]} p_\ell^{2/3}} = \alpha_{\min}(\mathbf{p}).$$

Thus, for every $\alpha \in [\alpha_{\min}(\mathbf{p}), 1]$, the fully active optimizer $\mathbf{q}^{\mathrm{act}}$ is feasible for the budgeted problem (13).

Since (13) minimizes the same objective $S_{\mathbf{p}}$ as the fully active problem, but over a smaller feasible set, its optimal value is at least the unconstrained optimum:

$$v_\alpha^*(\mathbf{p}) \geq \|\mathbf{p}\|_{2/3}.$$

On the other hand, when $\alpha \geq \alpha_{\min}(\mathbf{p})$, the vector $\mathbf{q}^{\mathrm{act}}$ is feasible, and therefore

$$v_\alpha^*(\mathbf{p}) \leq S_{\mathbf{p}}(\mathbf{q}^{\mathrm{act}}) = \|\mathbf{p}\|_{2/3}.$$

Combining the two inequalities yields

$$v_\alpha^*(\mathbf{p}) = \|\mathbf{p}\|_{2/3}, \qquad \forall \alpha \in [\alpha_{\min}(\mathbf{p}), 1].$$

The final statement follows immediately from Theorem 3.6: for every $\alpha \in [\alpha_{\min}(\mathbf{p}), 1]$, Algorithm 1 satisfies

$$\mathrm{SR}(\texttt{Alg.1}, T, \mathbf{p}) \in \mathcal{O}\Big(\sqrt{\frac{n}{T}}\, v_\alpha^*(\mathbf{p})\Big) = \mathcal{O}\Big(\sqrt{\frac{n}{T}}\, \|\mathbf{p}\|_{2/3}\Big).$$

This is exactly the optimal rate of fully active policies, which, by Lemma 3.1, is also the optimal rate over all policies up to constant factors. □

B.2 Proofs of Section 4

This section presents the omitted proofs of Section 4 of the main text.

B.2.1 Proof of Theorem 4.1

Theorem 4.1 (Simple Regret of EETC)**.** *For sufficiently large horizon T, the worst-case simple regret of EETC is*

$$SR(\texttt{EETC}, T, \mathbf{p}) \in \mathcal{O}\big(\sqrt{\frac{n}{T}}\,\|\mathbf{p}\|_{2/3}\big).$$

Proof. We analyze a slightly modified version of EETC in which the passive phase ends at time $\tau_2 - 1$, the frozen estimate is

$$\hat{\mathbf{p}}^{\mathrm{fr}} \coloneqq \hat{\mathbf{p}}(\tau_2 - 1),$$

and the active phase starts at round τ_2. Compared to the algorithm as stated, this removes the single wasted passive round at time τ_2. This changes the number of samples collected from each subpopulation by at most one and therefore does not affect the regret rate. We will therefore ignore this one round discrepancy in the remainder of the proof.

To analyze the stopping times, consider an auxiliary passive sequence $(\widetilde{C}_t)_{t\in[T]}$ of i.i.d. draws from $\mathbf{p}$. Up to time $\tau_2 - 1$, EETC observes exactly the prefix of this sequence, so both stopping times τ_1, τ_2 are measurable functions of $(\widetilde{C}_t)_{t\in[T]}$. For $t \in [T]$ and $j \in [k]$, define

$$\widetilde{N}_{t,j} \coloneqq \sum_{s=1}^{t} \mathbf{1}\{\widetilde{C}_s = j\}, \qquad \hat{p}_{t,j} \coloneqq \frac{\widetilde{N}_{t,j}}{t}.$$

For $t \le \tau_2 - 1$, $\widetilde{N}_{t,j}$ and $\hat{p}_{t,j}$ are exactly the counts $N(t)_j$ and empirical frequencies $\hat{p}(t)_j$ of Section 4, and $\hat{\mathbf{p}}_t$ below denotes the vector $\hat{\mathbf{p}}(t)$.

We first show that after $O(\log T)$ passive rounds, the empirical proportions are uniformly accurate. Let

$$m_T \coloneqq \left\lceil \frac{24\ln(2kT^2)}{p_{\min}} \right\rceil, \qquad p_{\min} \coloneqq \min_{j\in[k]} p_j.$$

Since $p_{\min} > 0$ is fixed, $m_T \in \mathcal{O}(\log T)$.

For each fixed $t \ge m_T$ and $j \in [k]$, the count $\widetilde{N}_{t,j}$ is binomial with mean $tp_j \ge m_T p_{\min} \ge 24\ln(2kT^2)$. Hence, by Lemma B.1,

$$\mathbb{P}\left(\frac{tp_j}{2} \le \widetilde{N}_{t,j} \le \frac{3tp_j}{2}\right) \ge 1 - \frac{1}{kT^2}.$$

Define the event

$$\mathcal{G}_1 \coloneqq \bigcap_{t=m_T}^{T} \bigcap_{j\in[k]} \left[\frac{tp_j}{2} \le \widetilde{N}_{t,j} \le \frac{3tp_j}{2}\right].$$

By a union bound over at most kT pairs (t, j),

$$\mathbb{P}(\mathcal{G}_1^c) \le \sum_{t=m_T}^{T} \sum_{j=1}^{k} \frac{1}{kT^2} \le \frac{1}{T}. \tag{14}$$

We next record the consequences of $\mathcal{G}_1$.

First, on $\mathcal{G}_1$, for every $j \in [k]$,

$$\widetilde{N}_{m_T,j} \ge \frac{m_T p_j}{2} \ge \frac{m_T p_{\min}}{2} \ge 12\ln(2kT^2) > \ln(kT^2),$$

and therefore, by the definition of τ_1,

$$\tau_1 \le m_T \in \mathcal{O}(\log T). \tag{15}$$

We now show that the frozen estimate $\hat{\mathbf{p}}^{\mathrm{fr}}$ is accurate. Define

$$\beta(\mathbf{u}) \coloneqq 1 - \alpha_{\min}(\mathbf{u}) = \frac{u_{\max}^{-1/3}}{\sum_{j\in[k]} u_j^{2/3}}, \qquad u_{\max} \coloneqq \max_j u_j.$$

We first note that $\beta(\mathbf{u})$ admits a uniform positive lower bound over all probability vectors $\mathbf{u} \in \Delta^{k-1}$. Indeed, since $u_{\max} \le 1$, we have

$$u_{\max}^{-1/3} \ge 1.$$

Moreover, since $x \mapsto x^{2/3}$ is concave on $[0, \infty)$, Jensen's inequality gives

$$\frac{1}{k} \sum_{j \in [k]} u_j^{2/3} \le \left(\frac{1}{k} \sum_{j \in [k]} u_j \right)^{2/3} = \left(\frac{1}{k} \right)^{2/3}.$$

Multiplying both sides by k yields

$$\sum_{j \in [k]} u_j^{2/3} \le k^{1/3}.$$

Hence, for every $\mathbf{u} \in \Delta^{k-1}$,

$$\beta(\mathbf{u}) \ge k^{-1/3}. \tag{16}$$

Since $m_T = \mathcal{O}(\log T)$, for sufficiently large T we have

$$\frac{m_T}{T} < k^{-1/3}.$$

Combining this with (16), we obtain

$$\frac{t}{T} \le \frac{m_T}{T} < k^{-1/3} \le \beta(\hat{\mathbf{p}}_t), \qquad \forall t \le m_T.$$

Therefore, the stopping condition defining τ_2 cannot be satisfied at any time $t \le m_T$, and thus

$$\tau_2 > m_T. \tag{17}$$

On the other hand, on the event $\mathcal{G}_1$, we have for every $t \in [m_T, T]$ and every $j \in [k]$,

$$\frac{p_j}{2} \le \hat{p}_{t,j} \le \frac{3p_j}{2}. \tag{18}$$

Since (17) implies $\tau_2 - 1 \ge m_T$, it follows from (18) that, on $\mathcal{G}_1$,

$$\frac{p_j}{2} \le \hat{p}_j^{\mathrm{fr}} \le \frac{3p_j}{2}, \qquad \forall j \in [k]. \tag{19}$$

This implies

$$2^{-2/3} p_j^{2/3} \le (\hat{p}_j^{\mathrm{fr}})^{2/3} \le \left(\frac{3}{2} \right)^{2/3} p_j^{2/3}.$$

Summing over j and raising to the power $3/2$ proves

$$\frac{1}{2} \|\mathbf{p}\|_{2/3} \le \left\| \hat{\mathbf{p}}^{\mathrm{fr}} \right\|_{2/3} \le \frac{3}{2} \|\mathbf{p}\|_{2/3}. \tag{20}$$

We now analyze phase (III). Let

$$s \coloneqq \frac{\tau_2 - 1}{T}.$$

By the definition of τ_2 and the modified algorithm, we have

$$s \le \beta(\hat{\mathbf{p}}^{\mathrm{fr}}) = 1 - \alpha_{\min}(\hat{\mathbf{p}}^{\mathrm{fr}}).$$

Therefore, by Lemma 3.7, the optimizer of the budgeted problem with passive fraction s and distribution $\hat{\mathbf{p}}^{\mathrm{fr}}$ coincides with the fully active optimizer for $\hat{\mathbf{p}}^{\mathrm{fr}}$, namely

$$q_j = \frac{(\hat{p}_j^{\mathrm{fr}})^{2/3}}{\sum_{\ell \in [k]} (\hat{p}_\ell^{\mathrm{fr}})^{2/3}}, \qquad j \in [k]. \tag{21}$$

This is exactly the vector used by the algorithm in the active phase, and the active phase sampling probabilities are then

$$r_j = \frac{q_j - s\hat{p}_j^{\mathrm{fr}}}{1-s}.$$

We next show that the final number of observations of each subpopulation is of order Tq_j with high probability. For $j \in [k], t \in [T]$, let

$$N_{t,j} \coloneqq \sum_{u=1}^{t} \mathbf{1}\{C_u = j\}$$

denote the number of times EETC observes subpopulation j by time t. Conditional on $\hat{\mathbf{p}}^{\mathrm{fr}}$ and τ_2, the passive counts are fixed and equal to

$$N_{\tau_2 - 1, j} = sT\hat{p}_j^{\mathrm{fr}}.$$

Let A_j denote the number of times subpopulation j is selected during the active phase. Conditional on $\hat{\mathbf{p}}^{\mathrm{fr}}$ and τ_2, the variable A_j is a sum of independent Bernoulli random variables with mean

$$M_j \coloneqq T(q_j - s\hat{p}_j^{\mathrm{fr}}).$$

Define

$$\mathcal{J}_{\mathrm{big}} \coloneqq [j \in [k] : M_j \geq 24\ln(2kT)], \qquad \mathcal{J}_{\mathrm{small}} \coloneqq [k] \setminus \mathcal{J}_{\mathrm{big}}.$$

For each $j \in \mathcal{J}_{\mathrm{big}}$, Lemma B.1 with $\eta = 1/(kT)$ gives

$$\mathbb{P}\left(A_j \geq \frac{M_j}{2} \,\middle|\, \hat{\mathbf{p}}^{\mathrm{fr}}, \tau_2\right) \geq 1 - \frac{1}{kT}.$$

Let

$$\mathcal{G}_2 \coloneqq \bigcap_{j \in \mathcal{J}_{\mathrm{big}}} \left[A_j \geq \frac{M_j}{2}\right].$$

Then, conditional on $(\hat{\mathbf{p}}^{\mathrm{fr}}, \tau_2)$, a union bound gives

$$\mathbb{P}(\mathcal{G}_2^c \mid \hat{\mathbf{p}}^{\mathrm{fr}}, \tau_2) \leq \frac{1}{T}.$$

Taking expectation with respect to $(\hat{\mathbf{p}}^{\mathrm{fr}}, \tau_2)$ yields

$$\mathbb{P}(\mathcal{G}_2^c) = \mathbb{E}\left[\mathbb{P}(\mathcal{G}_2^c \mid \hat{\mathbf{p}}^{\mathrm{fr}}, \tau_2)\right] \leq \frac{1}{T}. \tag{22}$$

We claim that on $\mathcal{G}_1 \cap \mathcal{G}_2$,

$$N_{T,j} \geq \frac{Tq_j}{2}, \qquad \forall j \in [k]. \tag{23}$$

Indeed, if $j \in \mathcal{J}_{\mathrm{big}}$, then on $\mathcal{G}_2$,

$$\begin{aligned} N_{T,j} &= sT\hat{p}_j^{\mathrm{fr}} + A_j \geq sT\hat{p}_j^{\mathrm{fr}} + \frac{M_j}{2} \\ &= sT\hat{p}_j^{\mathrm{fr}} + \frac{T(q_j - s\hat{p}_j^{\mathrm{fr}})}{2} = \frac{T(q_j + s\hat{p}_j^{\mathrm{fr}})}{2} \geq \frac{Tq_j}{2}. \end{aligned}$$

If $j \in \mathcal{J}_{\mathrm{small}}$, then $M_j < 24\ln(2kT)$. We also have

$$N_{T,j} \geq sT\hat{p}_j^{\mathrm{fr}} = Tq_j - M_j.$$

Moreover, from (21),

$$q_j = \frac{(\hat{p}_j^{\mathrm{fr}})^{2/3}}{\sum_{\ell \in [k]} (\hat{p}_\ell^{\mathrm{fr}})^{2/3}} \geq \hat{p}_{\min}^{\mathrm{fr}},$$

because

$$\sum_{\ell \in [k]} (\hat{p}_\ell^{\mathrm{fr}})^{2/3} \leq (\hat{p}_{\min}^{\mathrm{fr}})^{-1/3} \sum_{\ell \in [k]} \hat{p}_\ell^{\mathrm{fr}} = (\hat{p}_{\min}^{\mathrm{fr}})^{-1/3}.$$

Together with (19), this implies

$$q_j \geq \hat{p}^{\mathrm{fr}}_{\min} \geq \frac{p_{\min}}{2}.$$

Since $T \gg 1/p_{\min}, k$, we have

$$Tq_j \geq \frac{Tp_{\min}}{2} \geq 48\ln(2kT),$$

and therefore, for $j \in \mathcal{J}_{\mathrm{small}}$,

$$N_{T,j} \geq Tq_j - M_j \geq Tq_j - 24\ln(2kT) \geq \frac{Tq_j}{2}.$$

This proves (23).

Now fix an arbitrary instance. Let Δ_j denote the recommendation gap in subpopulation j at the end of the horizon. Since the bandit subroutine $\mathcal{A}_{\mathrm{SR}}$ is anytime and achieves the optimal standard bandit simple regret rate, there exists a universal constant $c_3 > 0$ such that, conditional on $N_{T,j}$,

$$\mathbb{E}[\Delta_j \mid N_{T,j}] \leq c_3\sqrt{\frac{n}{N_{T,j}}}.$$

Therefore,

$$\begin{aligned}
\mathrm{SR}(\texttt{EETC}, T, \mathbf{p}) &= \sum_{j\in[k]} p_j \mathbb{E}[\Delta_j] \\
&= \sum_{j\in[k]} p_j \mathbb{E}[\Delta_j \mathbf{1}_{\mathcal{G}_1\cap\mathcal{G}_2}] + \sum_{j\in[k]} p_j \mathbb{E}[\Delta_j \mathbf{1}_{(\mathcal{G}_1\cap\mathcal{G}_2)^c}] \\
&\leq c_3 \sum_{j\in[k]} p_j \mathbb{E}\left[\sqrt{\frac{n}{N_{T,j}}}\mathbf{1}_{\mathcal{G}_1\cap\mathcal{G}_2}\right] + \mathbb{P}((\mathcal{G}_1\cap\mathcal{G}_2)^c).
\end{aligned}$$

Using (23), this yields

$$\mathrm{SR}(\texttt{EETC}, T, \mathbf{p}) \leq c_3\sqrt{\frac{2n}{T}} \sum_{j\in[k]} \frac{p_j}{\sqrt{q_j}} + \mathbb{P}((\mathcal{G}_1\cap\mathcal{G}_2)^c). \tag{24}$$

It remains to bound the sum. By (21),

$$\sum_{j\in[k]} \frac{p_j}{\sqrt{q_j}} = \left(\sum_{\ell\in[k]} (\hat{p}^{\mathrm{fr}}_\ell)^{2/3}\right)^{1/2} \sum_{j\in[k]} p_j (\hat{p}^{\mathrm{fr}}_j)^{-1/3}.$$

From (19),

$$\sum_{\ell\in[k]} (\hat{p}^{\mathrm{fr}}_\ell)^{2/3} \leq \left(\frac{3}{2}\right)^{2/3} \sum_{\ell\in[k]} p_\ell^{2/3}, \qquad (\hat{p}^{\mathrm{fr}}_j)^{-1/3} \leq 2^{1/3} p_j^{-1/3}.$$

Therefore,

$$\begin{aligned}
\sum_{j\in[k]} \frac{p_j}{\sqrt{q_j}} &\leq 3^{1/3} \left(\sum_{\ell\in[k]} p_\ell^{2/3}\right)^{1/2} \sum_{j\in[k]} p_j^{2/3} \\
&= 3^{1/3} \left\|\mathbf{p}\right\|_{2/3}.
\end{aligned} \tag{25}$$

Finally, by (14) and (22),

$$\mathbb{P}((\mathcal{G}_1\cap\mathcal{G}_2)^c) \leq \mathbb{P}(\mathcal{G}_1^c) + \mathbb{P}(\mathcal{G}_2^c) \leq \frac{2}{T}.$$

Substituting this and (25) into (24), we obtain

$$\mathrm{SR}(\texttt{EETC}, T, \mathbf{p}) \leq c_4\sqrt{\frac{n}{T}}\ \left\|\mathbf{p}\right\|_{2/3} + \frac{2}{T}$$

for a universal constant $c_4 > 0$. Hence, for sufficiently large T,

$$\mathrm{SR}(\texttt{EETC}, T, \mathbf{p}) \in \mathcal{O}\Big(\sqrt{\frac{n}{T}}\,\|\mathbf{p}\|_{2/3}\Big).$$

The final claim follows from the known-$\mathbf{p}$ analysis: the optimal fully active policy that knows $\mathbf{p}$ achieves rate

$$\Theta\Big(\sqrt{\frac{n}{T}}\,\|\mathbf{p}\|_{2/3}\Big),$$

and this also matches the known-$\mathbf{p}$ lower bound up to constant factors. Therefore EETC achieves the same simple regret rate, up to constants, as the optimal fully active known-$\mathbf{p}$ policy. □

C Beyond the Data-Rich Regime

This section gives finite-horizon guarantees when some subpopulations are too rare to be learned accurately along with anytime versions of the algorithms in the main text. We assume $n \geq 2$ arms and use an anytime, n-armed simple-regret subroutine whose expected recommendation gap after $s \geq 1$ pulls is at most $C_{\mathrm{sr}} \min\{1, \sqrt{n/s}\}$, uniformly over reward distributions; with no pulls we recommend an arbitrary arm and use the bound 1.

C.1 Known subpopulation distribution

For $\mathbf{t} = (t_1, \ldots, t_k) \in \mathbb{R}_+^k$ with $\sum_j t_j = T$, define

$$V_T(\mathbf{p}) := \min_{\substack{t_j \geq 0 \\ \sum_j t_j = T}} \sum_{j=1}^{k} p_j \sqrt{\frac{n}{n + t_j}}. \tag{26}$$

The summand is of order p_j when $t_j \lesssim n$ and of order $p_j\sqrt{n/t_j}$ when $t_j \gtrsim n$. Thus the objective itself accounts for subpopulations on which learning is not worthwhile at horizon T.

Proposition C.1 (Finite-horizon active regret)**.** *There are universal constants $0 < c < C < \infty$ such that, for every $T \geq 1$ and every fixed $\mathbf{p} \in \Delta_+^{k-1}$,*

$$cV_T(\mathbf{p}) \leq \inf_{\pi \in \Pi(\mathbf{p})} \mathrm{SR}(\pi, T, \mathbf{p}) \leq CV_T(\mathbf{p}). \tag{27}$$

The upper bound is attained by a deterministic, reward-independent allocation followed by the within-subpopulation simple-regret subroutine.

Proof. Let $\mathbf{t}^*$ minimize (26). Allocate $N_j = \lfloor t_j^* \rfloor$ active rounds to subpopulation j and distribute the remaining rounds arbitrarily. The subroutine can ignore these additional rounds. For every $t \geq 0$, $n \geq 2$, and $N = \lfloor t \rfloor$, its regret bound implies

$$\mathbb{E}[\Delta_j] \leq C' \sqrt{\frac{n}{n + t}}. \tag{28}$$

Indeed, if $t \leq n$, the right-hand square root is at least $1/\sqrt{2}$, while the regret is at most 1. If $t > n$, then $N \geq t/2$ and $\sqrt{n/N} \leq 2\sqrt{n/(n+t)}$. Multiplying (28) by p_j and summing proves the upper bound.

For the lower bound, write $f(t) = \sqrt{n/(n+t)}$, and fix a small universal $\varepsilon \in (0, 1/8]$. Set $\delta_j = \varepsilon f(t_j^*)$. Consider the following product prior on reward instances. Independently for each j, draw $\omega_j \sim \mathrm{Bernoulli}(1/2)$ and b_j uniformly from $\{2, \ldots, n\}$. Arm 1 in subpopulation j has Bernoulli mean $1/2 + \delta_j$; arm b_j has mean $1/2 + 2\delta_j\omega_j$; all other arms have mean $1/2$. For $\omega_j = 0$ the best arm is 1, whereas for $\omega_j = 1$ it is b_j. An incorrect recommendation incurs a gap of at least δ_j.

Let $N_{a,j}$ be the number of times the policy pulls arm a in subpopulation j, and let $N_j = \sum_a N_{a,j}$. Expectations in the remainder of the proof average over both the random instance drawn from the

prior and the policy's interaction with it. Write $\bar{N}_j = \mathbb{E}[N_j]$. Since $\sum_j N_j = T$ on every interaction,

$$\sum_{j=1}^{k} \bar{N}_j = T. \tag{29}$$

Fix j, and let $U_j = (\omega_{-j}, \mathbf{b}_{-j})$ contain the random bits and rival-arm identities in all subpopulations other than j. A value $u = (\widetilde{\omega}_{-j}, \widetilde{\mathbf{b}}_{-j})$ specifies one realization of these variables. We also fix $b_j = b \in \{2, \dots, n\}$. Let $P_0^{u,b}$ and $P_1^{u,b}$ be the laws of the full interaction, including the final recommendation, with $\omega_j = 0$ and $\omega_j = 1$, respectively. Thus u and b fix the reward instance outside this single bit; the policy's randomization and observed rewards remain random under both laws. Put $A_j = \{\hat{a}_j = 1\}$. Under $P_0^{u,b}$, the error event is A_j^c; under $P_1^{u,b}$, A_j is contained in the error event because arm b is optimal. Thus the sum of the two error probabilities is at least

$$P_0^{u,b}(A_j^c) + P_1^{u,b}(A_j) \geq \frac{1}{2} \exp\Big\{ -D(P_0^{u,b} \| P_1^{u,b}) \Big\}, \tag{30}$$

where the inequality is Bretagnolle–Huber.

These two environments differ only in the reward distribution of arm b in subpopulation j. The KL chain rule for adaptive experiments therefore gives

$$D(P_0^{u,b} \| P_1^{u,b}) = \mathbb{E}_0^{u,b}[N_{b,j}] \, \mathrm{kl}\left(\frac{1}{2}, \frac{1}{2} + 2\delta_j \right) \leq c_0 \delta_j^2 \mathbb{E}_0^{u,b}[N_{b,j}], \tag{31}$$

with a universal c_0 (for example, $c_0 = 16$). Indeed, $\mathrm{kl}(1/2, 1/2 + 2\delta_j) = -\frac{1}{2}\log(1 - 16\delta_j^2) \leq 16\delta_j^2$ for $\delta_j \leq 1/8$. The learner's context and arm selection kernels are the same under both laws conditional on the past, so they contribute no additional KL terms.

Let e_j be the probability, averaged over the prior and the interaction, of recommending a non-optimal arm in subpopulation j. Since ω_j is an independent fair bit, (30) and (31) imply the following. To shorten the display, write $K_j(u, b) = \mathbb{E}_0^{u,b}[N_{b,j}]$ and $\bar{K}_j = \mathbb{E}_{U_j, b_j}[K_j(U_j, b_j)]$. Then

$$\begin{aligned} e_j &\geq \frac{1}{4} \mathbb{E}_{U_j, b_j}\left[e^{-c_0 \delta_j^2 K_j(U_j, b_j)} \right] \\ &\geq \frac{1}{4} e^{-c_0 \delta_j^2 \bar{K}_j}. \end{aligned} \tag{32}$$

For the second inequality, take the random variable $X = K_j(U_j, b_j)$ and the function $\phi(x) = e^{-c_0 \delta_j^2 x}$. Since $\phi''(x) = (c_0 \delta_j^2)^2 e^{-c_0 \delta_j^2 x} > 0$, Jensen's inequality gives $\mathbb{E}[\phi(X)] \geq \phi(\mathbb{E}[X])$; here $\mathbb{E}[X] = \bar{K}_j$. The function is decreasing, but its convexity determines the direction of Jensen's inequality.

The random choice of b_j now supplies the factor $1/(n-1)$. When $\omega_j = 0$, arm b_j has mean $1/2$, just like every arm other than arm 1. Thus, conditional on $U_j = u$, the interaction law under $\omega_j = 0$ is independent of the hidden choice b_j. Denote this common law by P_0^u. Averaging over the uniform b_j gives

$$\begin{aligned} \bar{K}_j &= \mathbb{E}_{U_j} \frac{1}{n-1} \sum_{b=2}^{n} \mathbb{E}_0^{U_j}[N_{b,j}] \\ &\leq \frac{\mathbb{E}[N_j \mid \omega_j = 0]}{n-1} \leq \frac{2\bar{N}_j}{n-1}. \end{aligned} \tag{33}$$

The last inequality follows from $\bar{N}_j = \frac{1}{2}\mathbb{E}[N_j \mid \omega_j = 0] + \frac{1}{2}\mathbb{E}[N_j \mid \omega_j = 1]$. Because every incorrect recommendation costs at least δ_j, $\mathrm{SR}(\pi, T, \mathbf{p}) \geq \sum_j p_j \delta_j e_j$. Combining this with (32) and (33), and using $1/(n-1) \leq 2/n$ for $n \geq 2$, yields

$$\mathrm{SR}(\pi, T, \mathbf{p}) \geq \frac{1}{4} \sum_j p_j \delta_j \exp\left\{ -c_1 \frac{\delta_j^2 \bar{N}_j}{n} \right\}. \tag{34}$$

Here $c_1 = 4c_0$ is a universal constant.

It remains to show that the exponent in (34) is bounded after averaging across contexts, regardless of the policy's expected allocation $\bar{\mathbf{N}}$. The function $f(t) = \sqrt{n}\,(n+t)^{-1/2}$ is convex and has derivative $f'(t) = -\sqrt{n}/[2(n+t)^{3/2}]$. Let $\lambda > 0$ be the multiplier for $\sum_j t_j = T$ at $\mathbf{t}^*$, and let $\eta_j \geq 0$ be the multiplier for $t_j \geq 0$. Stationarity and complementary slackness give $-p_j\sqrt{n}/[2(n+t_j^*)^{3/2}] + \lambda - \eta_j = 0$ and $\eta_j t_j^* = 0$. Equivalently,

$$\frac{p_j\sqrt{n}}{2(n+t_j^*)^{3/2}} \begin{cases} = \lambda, & t_j^* > 0, \\ \leq \lambda, & t_j^* = 0. \end{cases} \tag{35}$$

By the choice $\delta_j = \varepsilon\sqrt{n/(n+t_j^*)}$,

$$\frac{p_j\delta_j^3}{n} = \varepsilon^3 \frac{p_j\sqrt{n}}{(n+t_j^*)^{3/2}} \leq 2\varepsilon^3\lambda \qquad \text{for every } j. \tag{36}$$

For coordinates with $t_j^* > 0$, equality holds in (35). Since $\sum_j t_j^* = T$, we also have

$$\lambda T = \sum_{j:t_j^*>0} \lambda t_j^* = \sum_{j:t_j^*>0} \frac{p_j f(t_j^*) t_j^*}{2(n+t_j^*)} \leq \frac{1}{2} V_T(\mathbf{p}).$$

To apply Jensen across contexts, set $W = \sum_j p_j\delta_j = \varepsilon V_T(\mathbf{p})$ and $w_j = p_j\delta_j/W$, so $\sum_j w_j = 1$. Consider a discrete random variable X that equals $c_1\delta_j^2\bar{N}_j/n$ with probability w_j. Jensen gives $\sum_j w_j e^{-c_1\delta_j^2\bar{N}_j/n} \geq e^{-\sum_j w_j c_1\delta_j^2\bar{N}_j/n}$. Multiplying by W and substituting the definition of w_j yields

$$\begin{aligned} \sum_j p_j\delta_j \exp\left\{-c_1\frac{\delta_j^2\bar{N}_j}{n}\right\} &= W\sum_j w_j \exp\left\{-c_1\frac{\delta_j^2\bar{N}_j}{n}\right\} \\ &\geq W\exp\left\{-\frac{c_1}{W}\sum_j \frac{p_j\delta_j^3}{n}\bar{N}_j\right\}. \end{aligned}$$

By (36), (29), and $\lambda T \leq V_T(\mathbf{p})/2$, the quantity inside this last exponent is at most

$$\frac{c_1}{W}\sum_j \frac{p_j\delta_j^3}{n}\bar{N}_j \leq \frac{2c_1\varepsilon^3\lambda T}{\varepsilon V_T(\mathbf{p})} \leq c_1\varepsilon^2.$$

Substituting this bound into (34) proves

$$\mathrm{SR}(\pi, T, \mathbf{p}) \geq \frac{\varepsilon V_T(\mathbf{p})}{4}\exp\{-c_1\varepsilon^2\}.$$

This holds for every policy, including reward-adaptive allocation rules. □

For comparison with the data-rich statement in the main text, the value in (26) satisfies

$$\sqrt{\frac{n}{T+nk}}\,\|\mathbf{p}\|_{2/3} \leq V_T(\mathbf{p}) \leq \min\left\{1, \sqrt{\frac{n}{T}}\,\|\mathbf{p}\|_{2/3}\right\}. \tag{37}$$

For the left inequality, write $q_j = (t_j+n)/(T+nk)$ and minimize $\sum_j p_j/\sqrt{q_j}$ over the probability simplex. For the right inequality, use $t_j = Tp_j^{2/3}/\sum_\ell p_\ell^{2/3}$, and also note that every summand in (26) is at most p_j. Thus $T \geq nk$ is sufficient for the finite-horizon minimax value to match the uncapped $\sqrt{n/T}\,\|\mathbf{p}\|_{2/3}$ rate up to constants.

A simple explicit alternative to solving (26) is to truncate the tail. Order the masses as $p_{(1)} \geq \cdots \geq p_{(k)}$ and, for $m \geq 1$, set $Z_m = \sum_{i=1}^m p_{(i)}^{2/3}$. Sample only the largest m subpopulations, with expected or deterministically rounded allocation proportional to $p_{(i)}^{2/3}$, and recommend arbitrary arms elsewhere. The same rounding argument gives

$$\mathrm{SR} \leq C\left\{\sum_{i=m+1}^k p_{(i)} + \sqrt{\frac{n}{T}}\,Z_m^{3/2}\right\}. \tag{38}$$

For $m = 0$, the policy samples no designated subpopulation and has regret at most $1 = \sum_i p_{(i)}$. One may minimize the right-hand side of (38) over $m \in \{0, \dots, k\}$ using the known $\mathbf{p}$. This is an upper bound for a convenient family of truncation policies; the exact finite-horizon minimax value is (26). In particular, taking $m = k$ recovers $C\sqrt{n/T}\,\|\mathbf{p}\|_{2/3}$ whenever that bound is informative.

C.2 Unknown p: thresholded EETC

The original EETC analysis estimates every coordinate of $\mathbf{p}$ multiplicatively. For rare subpopulations this requires a passive phase that depends on $1/p_{\min}$. The following fixed-horizon variant instead estimates only masses large enough to affect the desired regret scale. It retains the passive estimation and active allocation structure of EETC; we call it thresholded EETC.

Fix $T \geq n$, and define

$$\gamma_T := \frac{1}{8k}\sqrt{\frac{n}{T}}, \qquad L_T := \log(4kT^2), \qquad M_T := \left\lceil \frac{48 L_T}{\gamma_T} \right\rceil. \tag{39}$$

If $M_T > T/2$, the algorithm can use any passive policy; the theorem below concerns $M_T \leq T/2$. Otherwise, in the first M_T rounds draw contexts passively from $\mathbf{p}$ and form $\widehat{p}_j = M_T^{-1} \sum_{t=1}^{M_T} \mathbf{1}\{C_t = j\}$. Retain

$$\widehat{S} = \{j : \widehat{p}_j \geq 2\gamma_T\}, \qquad \widehat{q}_j = \frac{\widehat{p}_j^{\,2/3}}{\sum_{\ell \in \widehat{S}} \widehat{p}_\ell^{\,2/3}} \quad (j \in \widehat{S}).$$

For each of the remaining $T - M_T$ rounds, draw a context independently from $\widehat{\mathbf{q}}$ and use its simple-regret subroutine to choose an arm. The subroutines may start afresh at this transition, discarding passive-phase rewards; this convention makes the conditional analysis immediate. At the end, output their recommendations for retained contexts and arbitrary arms for the others. If $\widehat{S} = \varnothing$, sample any contexts in the remaining rounds and use arbitrary recommendations for all contexts.

Theorem C.2 (Thresholded EETC without a $p_{\min}$ condition)**.** *There is a universal constant C such that, for every fixed $\mathbf{p} \in \Delta_+^{k-1}$, if $T \geq n$ and $M_T \leq T/2$, then*

$$\mathrm{SR}(\texttt{thresholded EETC}, T, \mathbf{p}) \leq C\sqrt{\frac{n}{T}}\,\|\mathbf{p}\|_{2/3}. \tag{40}$$

The condition on M_T is implied, up to a universal numerical constant, by $T \gtrsim \max\{n, k^2 \log^2(kT)/n\}$ and has no dependence on $p_{\min}$.

Proof. We first establish a simultaneous event for the passive estimate. For each j with $p_j \geq \gamma_T$, the multiplicative Chernoff bound gives

$$\mathbb{P}(\widehat{p}_j \notin [p_j/2, 3p_j/2]) \leq 2\exp(-M_T p_j/12) \leq 2\exp(-4L_T).$$

For $p_j < \gamma_T$, stochastic domination by a binomial variable with mean $M_T\gamma_T$ and the upper-tail Chernoff bound give

$$\mathbb{P}(\widehat{p}_j \geq 2\gamma_T) \leq \exp(-M_T\gamma_T/3) \leq \exp(-16L_T).$$

A union bound thus gives an event G of probability at least $1 - T^{-2}$ on which both conclusions hold for all j.

On G, every retained context has $p_j \geq \gamma_T$ and $p_j/2 \leq \widehat{p}_j \leq 3p_j/2$. Every context with $p_j \geq 4\gamma_T$ is retained, because then $\widehat{p}_j \geq p_j/2 \geq 2\gamma_T$. Hence

$$\sum_{j \notin \widehat{S}} p_j \leq 4k\gamma_T = \frac{1}{2}\sqrt{\frac{n}{T}}. \tag{41}$$

Also $\widehat{S}$ is nonempty on G: some $p_j \geq 1/k$, while $4\gamma_T \leq 1/(2k)$ since $T \geq n$.

Conditional on the passive history, each retained context receives a binomial number of active samples with mean $(T - M_T)\widehat{q}_j$. The within-context subroutine and a binomial lower-tail bound imply

$$\mathbb{E}[\Delta_j \mid \text{passive history}] \leq C_1\sqrt{\frac{n}{(T - M_T)\widehat{q}_j}}. \tag{42}$$

For completeness, if the binomial mean $u < 1$, the right-hand side is at least a constant and the gap is at most 1. If $u \geq 1$, on the event that the count is at least $u/2$ the subroutine bound is $C_{\mathrm{sr}}\sqrt{2n/u}$, while the complementary event has probability at most $e^{-u/8} \leq C'/\sqrt{u}$. This proves (42), including the case of zero active pulls.

Let $Z = \sum_{j\in\widehat{S}} \widehat{p}_j^{\,2/3}$. On G,

$$\begin{aligned}\sum_{j\in\widehat{S}} \frac{p_j}{\sqrt{\widehat{q}_j}} &= \sqrt{Z}\sum_{j\in\widehat{S}} \frac{p_j}{\widehat{p}_j^{\,1/3}} \\ &\leq (3/2)^{1/3}2^{1/3}\left(\sum_{j\in\widehat{S}} p_j^{2/3}\right)^{3/2} \\ &\leq 3^{1/3}\|\mathbf{p}\|_{2/3}.\end{aligned}$$

Combining this inequality, (41), (42), and $T - M_T \geq T/2$ bounds the conditional regret on G by $C_2\sqrt{n/T}\,\|\mathbf{p}\|_{2/3}$; we used $\|\mathbf{p}\|_{2/3} \geq 1$. On G^c, weighted simple regret is at most 1, so its contribution is at most $T^{-2} \leq \sqrt{n/T}$. Taking expectations proves (40). □

The thresholded rule gives the uncapped active rate without estimating the smallest coordinate of $\mathbf{p}$. For horizons where the right-hand side of (40) exceeds 1, the regret bound is of course vacuous; the known-$\mathbf{p}$ benchmark in Proposition C.1 retains its capped form at all horizons.

C.3 Anytime versions

All policies in the paper can be made anytime, i.e., independent of the horizon T, at the cost of a constant factor in regret; below, t denotes the current round at which the recommendation is requested.

Passive and fully active policies ($\mathbf{p}$ known). Their sampling distributions, $\mathbf{p}$ and $\mathbf{q}^*(\mathbf{p})$, do not depend on T, and $\mathcal{A}_{\mathrm{SR}}$ is anytime, so no change is needed.

Budgeted policy. Algorithm 1 uses T only to arrange the passive and active rounds. Instead, call round t active if $\lfloor \alpha t\rfloor > \lfloor \alpha(t-1)\rfloor$ and passive otherwise; passive rounds sample $C_t \sim \mathbf{p}$ and active rounds sample $C_t \sim \mathbf{r}$ as in Algorithm 1. Exactly $\lfloor \alpha t\rfloor \leq \alpha t$ of the first t rounds are active, so the policy is history-free and respects the budget at every t. Since $\mathbf{q}^*_\alpha(\mathbf{p}) = (1-\alpha)\mathbf{p} + \alpha\mathbf{r}$, the expected number of rounds in subpopulation j among the first t is

$$(t - \lfloor \alpha t\rfloor)p_j + \lfloor \alpha t\rfloor r_j = t\,\mathbf{q}^*_\alpha(\mathbf{p})_j + (\lfloor \alpha t\rfloor - \alpha t)(r_j - p_j),$$

which differs from $t\,\mathbf{q}^*_\alpha(\mathbf{p})_j$ by at most 1. Hence the expected allocation fractions $\mathbf{q}^{(t)}$ satisfy $q_j^{(t)} \geq \mathbf{q}^*_\alpha(\mathbf{p})_j/2$ whenever $t\,\mathbf{q}^*_\alpha(\mathbf{p})_j \geq 2$, and therefore $S_{\mathbf{p}}(\mathbf{q}^{(t)}) \leq \sqrt{2}\,v^*_\alpha(\mathbf{p})$. Applying Lemma 3.2 at horizon t (its proof allows round-dependent selection probabilities), as in the proof of Theorem 3.6, gives regret $\mathcal{O}\big(\sqrt{n/t}\,v^*_\alpha(\mathbf{p})\big)$ at every t with $t\,p_{\min} \geq C\ln(2kt)$ for a large universal constant C.

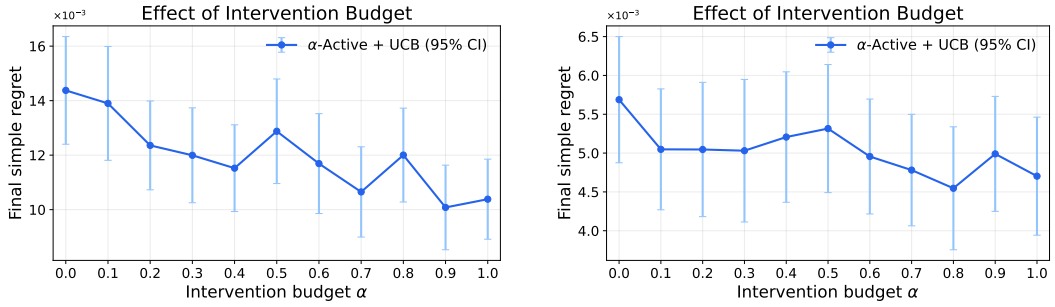
EETC. Split time into epochs $r = 0, 1, 2, \ldots$ of lengths $H_r = 2^r$. In epoch r, run a fresh copy of EETC (new instances of $\mathcal{A}_{\mathrm{SR}}$ and a new estimate $\hat{\mathbf{p}}$) with horizon H_r; if H_r is too short for the estimation phase to complete, stay passive during the epoch. At time t, output the recommendations of the most recently completed epoch. This epoch r satisfies $2^{r+1} - 1 \leq t < 2^{r+2} - 1$, so $H_r = 2^r \geq t/4$. Let T_0 be such that Theorem 4.1 holds for every horizon $T \geq T_0$ (its horizon conditions are monotone in T). For $t \geq 4T_0$, the last completed epoch has horizon at least T_0, and since its recommendation depends only on the data of that epoch, Theorem 4.1 gives

$$\mathrm{SR} \leq \mathcal{O}\Big(\sqrt{\tfrac{n}{H_r}}\;\|\mathbf{p}\|_{2/3}\Big) = \mathcal{O}\Big(\sqrt{\tfrac{n}{t}}\;\|\mathbf{p}\|_{2/3}\Big),$$

with no additional logarithmic factor, because only one epoch is used for the output.

D Additional Experiments

This section provides additional details for the real-world experiment on the MovieLens dataset described in the main text, together with further experimental results.



(a) $T = 5000$ (b) $T = 10000$

Figure 2: Simple regret of the budgeted active algorithm on the MovieLens instance as a function of the intervention budget α. As predicted by the theory, the regret decreases as the active budget increases, interpolating between the passive and fully active regimes.

D.1 Details of the MovieLens Dataset

Our real-world experiment is based on the MovieLens 1M dataset, which contains 1,000,209 ratings from 6,040 users on roughly 3,900 movies. We construct a bandit instance by restricting attention to the five most frequently rated movies in the dataset, namely *American Beauty (1999)*, *Star Wars: Episode IV - A New Hope (1977)*, *Star Wars: Episode V - The Empire Strikes Back (1980)*, *Star Wars: Episode VI - Return of the Jedi (1983)*, and *Jurassic Park (1993)*. These five movies define the treatment set, so that $n = 5$.

The subpopulations are defined by the Cartesian product of gender and age group, yielding $2 \times 7 = 14$ demographic groups and therefore $k = 14$ subpopulations. After filtering the data to the selected movies and joining with the demographic attributes, the resulting dataset contains 14,964 rating records. Each record consists of a treatment–subpopulation pair together with the corresponding MovieLens rating, which we use directly as the reward. Thus, rewards take values in $\{1, 2, 3, 4, 5\}$. (No normalization to $[0, 1]$ is applied.)

For each subpopulation j, the population weight p_j is estimated by the empirical fraction of filtered records belonging to that group. The resulting subpopulation distribution is

$$\mathbf{p} = (0.0084,\ 0.0444,\ 0.0863,\ 0.0456,\ 0.0216,\ 0.0165,\ 0.0088,\\ 0.0233,\ 0.1549,\ 0.2994,\ 0.1474,\ 0.0558,\ 0.0539,\ 0.0336).$$

The coordinates are ordered by gender and age group, with the seven female groups first and the seven male groups second:

$(F,$ Under 18$)$, $(F,$ 18–24$)$, $(F,$ 25–34$)$, $(F,$ 35–44$)$, $(F,$ 45–49$)$, $(F,$ 50–55$)$, $(F,$ 56+$)$,
$(M,$ Under 18$)$, $(M,$ 18–24$)$, $(M,$ 25–34$)$, $(M,$ 35–44$)$, $(M,$ 45–49$)$, $(M,$ 50–55$)$, $(M,$ 56+$)$.

Similarly, for each treatment–subpopulation pair (a, j), the mean reward $\mu_{a,j}$ is computed as the empirical average of the corresponding ratings in the filtered dataset. During interaction, when an algorithm selects treatment a for subpopulation j, the environment samples uniformly with replacement from the historical ratings associated with (a, j) and returns the sampled rating as the reward.

At the end of the horizon, each algorithm recommends one treatment for every subpopulation, and performance is evaluated using the context-weighted simple regret defined in the main text, with the empirical MovieLens frequencies and empirical cell means used as the ground-truth values.

D.2 Budgeted Interventions on MovieLens

We next evaluate the budgeted active algorithm on the MovieLens instance [HK15]. In this experiment, we vary the intervention budget α from 0 to 1 in increments of 0.1, where $\alpha = 0$ corresponds to the fully passive policy and $\alpha = 1$ to the fully active policy. We report results for two horizons, $T = 5000$ and $T = 10000$, and for UCB as the standard bandit subroutine.

Figure 2 shows the average simple regret over 100 independent runs, together with 95% confidence intervals. Although the curves are somewhat noisy due to the variance of the real data environment, both figures clearly exhibit the expected decreasing trend: as the intervention budget α increases, the simple regret decreases. This is consistent with the theory, which predicts that allowing more active control over the sampled subpopulations improves the achievable regret by moving the allocation closer to the optimal active design.

Overall, these results support the budgeted intervention analysis and illustrate the gradual interpolation between passive and fully active performance on real data.

D.3 Synthetic Experiments

We next report synthetic experiments designed to highlight the gains predicted by the theory. In all synthetic experiments, for each treatment–subpopulation pair (a, j), the reward distribution is Bernoulli and independent across pairs. For each subpopulation j, one treatment is selected uniformly at random to be the optimal one, and its mean is set to

$$\frac{1}{2} + \delta_j, \qquad \delta_j \coloneqq \sqrt{\frac{n}{T}}\, p_j^{-1/3},$$

while the remaining $n - 1$ treatments have mean $1/2$. This construction follows the worst-case instances appearing in our theoretical analysis.

To induce a non-uniform subpopulation distribution with a pronounced active-passive gap, we choose $\mathbf{p}$ as

$$\left(1 - \epsilon,\ \frac{\epsilon}{k-1},\ \ldots,\ \frac{\epsilon}{k-1}\right),$$

where $\epsilon = \frac{1}{\sqrt{k-1}}$. That is, one subpopulation has a larger weight, while the remaining $k - 1$ subpopulations have equal smaller weights. This choice is aligned with the worst-case regime discussed in the proof of Lemma 3.5, in which the gap factor $R(\mathbf{p})$ becomes large. All results below are averaged over 100 independent runs and are reported with 95% confidence intervals.

We compare three algorithms: the passive baseline, the known-$\mathbf{p}$ fully active algorithm, and EETC. In each case, the standard bandit subroutine is UCB, since it has a better performance compared to the uniform subroutine.

Varying the number of subpopulations. In the first synthetic experiment, we fix $n = 5$ and vary the number of subpopulations over $k \in \{5, 10, 20, 30, 40\}$. For each value of k, we set $T = 100nk$, so that the horizon scales linearly with the total number of treatment–subpopulation pairs and the algorithms have sufficient budget to explore all pairs.

Figure 3 summarizes the results. The left panel reports the average simple regret, while the right panel reports the regret ratio between the passive baseline and each algorithm, namely

$$\frac{\text{Regret of passive baseline}}{\text{Regret of algorithm}},$$

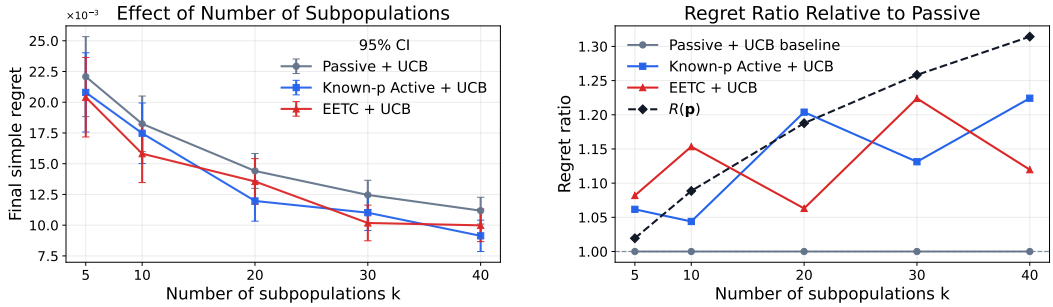
so values larger than 1 indicate improvement over the passive policy. The right panel also includes the theoretical quantity $R(\mathbf{p})$.

The figure shows that both proposed algorithms uniformly outperform the passive baseline over all tested values of k. Moreover, the gap becomes larger as k increases, which is consistent with the theory, since the value of $R(\mathbf{p})$ also increases with k in this construction. The known-$\mathbf{p}$ fully active policy and EETC are not clearly separated from one another, which is expected: they have the same asymptotic rate, and the finite sample variance is relatively high.

Varying the number of treatments. In the second synthetic experiment, we fix the number of subpopulations to $k = 20$ and vary the number of treatments over

$$n \in \{5, 10, 15, 20, 25, 30\}.$$

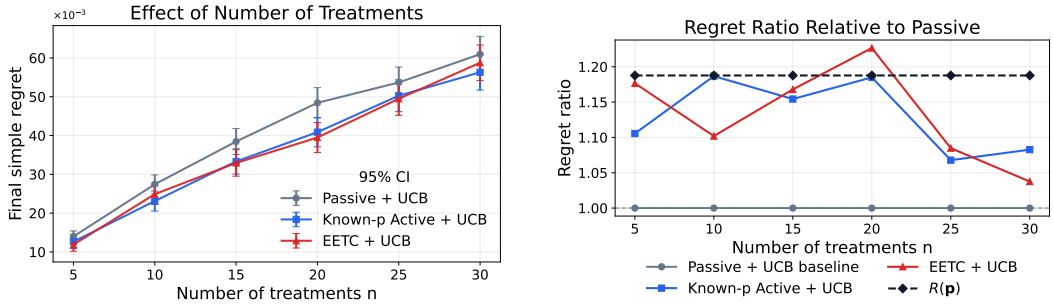
Unlike the previous experiment, we keep the horizon fixed at $T = 10000$ for all values of n. For each value of n, we fix one synthetic instance and average the final simple regret over 100 independent



(a) Average simple regret.

(b) Improvement ratio over passive baseline.

Figure 3: Synthetic experiment with varying number of subpopulations k and fixed $n = 5$, with horizon $T = 100nk$. Left: average simple regret over 100 runs with 95% confidence intervals. Right: ratio between the average regret of the passive baseline and the regret of each algorithm. Values above 1 indicate improvement over the passive policy. The theoretical factor $R(\mathbf{p})$ is also shown for reference.



(a) Average simple regret.

(b) Improvement ratio over passive baseline.

Figure 4: Synthetic experiment with varying number of treatments n and fixed $k = 20$, with horizon $T = 10000$. Left: average simple regret over 100 runs with 95% confidence intervals. Right: ratio between the average regret of the passive baseline and the regret of each algorithm. Values above 1 indicate improvement over passive sampling, while the passive baseline appears as the constant line at value 1.

runs, varying only the reward samples and the algorithmic randomness. All algorithms use UCB as the within-subpopulation subroutine.

Figure 4 reports the same two summaries as in the previous experiment. The left panel shows the average simple regret with 95% confidence intervals, and the right panel shows the regret ratio relative to the passive baseline,

$$\frac{\text{Regret of passive baseline}}{\text{Regret of algorithm}}.$$

Thus, values above 1 indicate improvement over passive sampling, while the passive baseline itself appears as the constant line at value 1.

The active algorithms outperform the passive baseline for all tested values of n. The relative gain becomes smaller as n increases, which is expected in this fixed-horizon experiment: larger n creates more treatment–subpopulation reward distributions to explore, while the total budget remains unchanged. As before, the known-$\mathbf{p}$ active algorithm and EETC are close to each other and are not uniformly ordered, consistent with their matching theoretical rates and the finite-sample variability of the experiment.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the claims in abstract and introduction have been discussed in Sections 3, 4, 5, and also appendices.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 2 states all the assumptions (finite contexts, guarantees that are worst-case over rewards for a fixed $\mathbf{p}$, and a data-rich regime with an explicit sufficient condition on T). Appendix C treats shorter horizons, and the Conclusion discusses limitations and future work, such as gap-dependent guarantees.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theoretical results are stated clearly with the assumptions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The implementation of our algorithms is straightforward and is discussed in related sections. Also, Appendix D contains the necessary information.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We do not release the code of this paper. The experiments use the public MovieLens 1M dataset, and the information needed to reproduce them is described in Section 5 and Appendix D.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: In Section 5 and Appendix D, we discuss about experiment setups.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Refer to Figures and the Table of results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our experiments are simple to run and do not require any specific computational resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper adheres fully to the NeurIPS Code of

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: There is no societal impact as our work primarily focuses on theoretical analysis of a problem in online decision-making (bandits).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The focus of our paper is the theoretical understanding of a specific problem in the area of Bandit algorithms. Therefore, this question is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The only existing asset is the MovieLens 1M dataset [HK15] released by GroupLens, which we cite and describe in Section 5 and Appendix D. We use it only for research purposes, in line with its terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The focus of our paper is the theoretical understanding of a specific problem in the area of Bandit algorithms. Therefore, this question is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The focus of our paper is the theoretical understanding of a specific problem in the area of Bandit algorithms. Therefore, this question is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The focus of our paper is the theoretical understanding of a specific problem in the area of Bandit algorithms. Therefore, this question is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The focus of our paper is the theoretical understanding of a specific problem in the area of Bandit algorithms. Therefore, this question is not applicable.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.