Bursting the Filter Bubble: Disincentivizing Echo Chambers in Social Networks

CHRISTIAN BORGS, UC Berkeley, USA
JENNIFER CHAYES, UC Berkeley, USA
CHRISTIAN IKEOKWU, UC Berkeley, USA
ELLEN VITERCIK, Stanford University, USA

In digital social networks, platforms are economically motivated to curate personalized content for users. Users are thus often trapped in "filter bubbles," limiting their exposure to diverse content. In this paper, we develop a mathematical model to explore how platforms can deliver personalized content while mitigating these bubbles. To avoid filter bubbles, we draw on the intuition that if some users are recommended some category of content, then all users should minimally be recommended a small amount of that content. We first analyze a naive formalization of this intuition and show it has unintended consequences: it leads to the "tyranny of the majority" with the burden of diversification borne disproportionately by those with minority interests. We refine our model based on this insight to distribute the burden of diversification more equitably. Based on this mathematical model, we suggest constraints or penalties that can be imposed on the platform to deter the creation of filter bubbles. We provide algorithms a platform can use to optimize content recommendations while adhering to these constraints. Using real-world preference data, we empirically verify that under our model, users share the burden of diversification with minimal detriment to content relevance.

CCS Concepts: • **Theory of computation → Online learning theory**; **Social networks**.

Additional Key Words and Phrases: polarization, filter bubbles, bandit algorithms, social networks, recommender systems, personalization, diversification, polarization tax, exposure diversity, echo chambers

ACM Reference Format:
Christian Borgs, Jennifer Chayes, Christian Ikeokwu, and Ellen Vitercik. 2023. Bursting the Filter Bubble: Disincentivizing Echo Chambers in Social Networks. In *EAAMO'23: ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, October 30 – November 1, 2023, Boston University Boston, MA, USA* . ACM, New York, NY, USA, 38 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Over the past decade, large internet platforms have amassed an unprecedented level of social and political power. Research has shown that the feedback loops generated by algorithmic recommendations increase polarization [21, 23, 30]. Echo chambers created by algorithmic recommendations on these platforms can have a wide range of adverse effects, such as amplifying and creating glass ceilings for minorities [31], as well as limiting exposure and job recommendations [14]. They also lead to disinformation and propaganda being disproportionately spread to minoritized groups [15].

In this paper, we propose an approach to content recommendation that simultaneously preserves the positive aspects of personalization while avoiding the pitfalls of filter bubbles. We do so by introducing a model that ensures that if some users are served a particular category of content, then all users will see at least a small amount of that content. For example, if a network includes individuals across a political spectrum, then every user will be exposed to at least a small amount of news from opposing perspectives. This allows a platform to present diverse content without forcing content

Manuscript submitted to ACM

on its users that no one is interested in. This approach builds upon seminal work by Celis et al. [10] who initiated the study of algorithmic approaches to reducing polarization. However, our approach to avoiding filter bubbles is different and our analysis techniques diverge significantly, as detailed in Section 1.2.

We model a platform recommending content to users with a standard multi-armed bandit formulation. There are k categories of content—such as fashion, sports, left-learning political content, right-leaning political content, and so on—and n users. For each user and content category, the platform receives a stochastic reward from an unknown distribution for showing the user content from that category, measured, for example, in terms of engagement or ad revenue. The platform interacts with the n users over T timesteps, at each timestep choosing a distribution over content categories for each user. The platform's goal is to maximize its cumulative reward. Standard bandit algorithms would eventually learn for each user the category with maximal expected reward and only show them content from that category, at which point the user's content recommendations would be caught in a filter bubble.

1.1 Our contributions

We propose a flexible approach to disincentivizing filter bubbles that adapts to the interests of the individuals on the network. We summarize our contributions along the following two axes.

1.1.1 Modeling contributions. We first analyze an approach that requires that the distribution of content shown to any one user is not far from the distribution shown to the population, so users cannot be siloed into disjoint filter bubbles. However, we show that the optimal recommendations exacerbate *tyranny of the majority*: the burden of diversification is borne by groups with minority interests (as often happens with naive approaches to diversification). A majority group will exclusively see content that they most enjoy while recommendations for minority users become far less relevant.

An equitable approach to preventing filter bubbles. The intuition behind our revised approach is that in order to avoid filter bubbles *and* tyranny of the majority, (1) users should primarily see content that they are most interested in (thus avoiding tyranny of the majority), and (2) if some users are shown a particular type of content, then all users should see at least a small amount of that content (thus avoiding filter bubbles). When both requirements are satisfied, users with majority interests will be exposed to content that interests minority groups and vice versa.

Formally, for each user, we impose the following constraint: we require that the probability she is shown content from a particular category must be at least γ times the average probability the entire population is shown content from that category, where $\gamma \in [0, 1]$ is a tunable parameter. We refer to this model as **Formulation 1.** Setting $\gamma = 0$ corresponds to complete personalization and setting $\gamma = 1$ requires that everyone see the same distribution of content. Moreover, if no one on the network is interested in some type of content, there is no requirement that users be shown that content. When $\gamma \leq \frac{1}{2}$, we show that conditions (1) and (2) are met and thus the burden of diversification is borne more equally among all users. We also provide a second formulation, called **Formulation 2**, where instead of imposing hard constraints, the platform is penalized based on the the extent to which it violates the γ constraint.

Taxation without knowledge of the true content distribution. The penalization described above depends on the true, underlying probabilities that the platform assigns to different types of content at each timestep. To augment the flexibility of our approach, we also analyze a model where an auditor only has access to a dataset describing the types of content that users were actually shown, as opposed to a description of the true distributions. In this model, the platform is penalized at the end of the T timesteps based on the extent to which the *empirical* distribution over content shown to each user violates the γ constraint described above. We refer to this model as **Formulation 3.**

1.1.2 Technical contributions. Since the platform does not know the reward distributions (corresponding to the users' preferences for the different types of content), it must learn a high-reward policy over the course of the T rounds. We analyze the *regret* of the Upper-Confidence-Bound (UCB) algorithm. The key challenges we face are providing nearly-matching lower bounds—which depend on structure exhibited by the specific constraints that we impose—and bounding the regret under Formulation 3, under which the optimal policy may be *history-dependent*.

Regret upper bounds. Under Formulation 1, we measure regret as the difference between (1) the cumulative reward of the optimal distribution over content that satisfies our γ constraint and (2) the cumulative reward of the platform's learning algorithm. Crucially, the optimal distribution (1) is defined by the users' reward distributions, but these are unknown to the learning algorithm. When $\gamma = 1$, a variant of the UCB algorithm achieves a regret of $\tilde{O}(\sqrt{nkT})$ and for $\gamma < 1$, another variant achieves a regret of $\tilde{O}(n\sqrt{kT})$. Under Formulations 2 and 3, we measure regret with respect to the optimal policy that maximizes the cumulative reward minus the penalty. Our regret bounds are $\tilde{O}(n\sqrt{kT})$.

Key challenge. Under Formulation 3, the optimal policy may be *history-dependent*: it may dynamically adjust its recommendations based on the empirical distribution over content thus far, and thus the magnitude of the final penalty. This is in contrast to Formulations 1 and 2, where the optimal policy is a fixed distribution over content.

Regret lower bounds. We provide a nearly-matching lower bound on regret under Formulation 1. As in the upper bound, our lower bound transitions from an $\Omega(n)$ dependence for small γ to an $\Omega(\sqrt{n})$ dependence for large γ. For $k = 2$ arms, we prove a lower bound of $\Omega(n\sqrt{T})$ for $\gamma < \frac{1}{2}$. Meanwhile, for all $k \geq 2$ and all $\gamma \in [0, 1]$, we prove a lower bound of $\Omega(\sqrt{nkT})$. This means that no algorithm has regret better than $\Omega(n\sqrt{T})$ for $\gamma < \frac{1}{2}$ or $\Omega(\sqrt{nkT})$ for any $\gamma \in [0, 1]$.

This transition from a $\Theta(n)$ to $\Theta(\sqrt{n})$ dependence elucidates a tension between the reward of the optimal policy and the ability of the learning algorithm to compete with the optimal policy. As γ grows, the set of distributions that the platform can show the user while still satisfying the γ constraint shrinks. Thus, the optimal policy comes from an increasingly restricted set so the regret benchmark is smaller. Likewise, as γ grows, the learner has to use an increasingly restricted set of policies to compete with the optimal policy. Since regret shrinks as γ grows, we show that the optimal policy's reward diminishes at a faster rate than the learner's handicap in competing with the optimal policy.

Key challenge. Lower bounds for bandit problems typically follow by identifying two worst-case problem instances that are similar enough that any algorithm would not be able to statistically distinguish between them, but are distinct enough to ensure that even if an algorithm has low regret on one instance, it will have high regret on the other. Simply creating n copies (one for each user) of the worst-case problem instances used in standard bandit lower bounds would lead to a large statistical difference between problem instances, thus precluding an $\Omega(n)$ dependence. Our lower bound construction therefore takes advantage of structure specific to our model.

Experiments. We analyze the optimal policies under the formulations from Section 1.1.1 using real user preference data [18]. We empirically verify that when users' preferences are heterogeneous, subgroups share the burden of diversification. We also show that users experience only a minor loss in utility when recommended diversified content.

1.2 Related work

There has been significant interest in understanding the mechanics of how recommender systems affect large-scale opinion dynamics, and if and when they lead to polarization [e.g., 6, 16, 28]. Most of the analysis has focused on how recommender systems impact network structure [32] and how this affects the spread of information and the opinions of members on the network. Recently there have been growing calls to algorithmically increase "exposure diversity" and

combat filter bubbles [7, 13, 19]. Castells et al. [9] discuss methodologies and metrics to assess recommendation diversity, and Halpern et al. [17] analyze the trade-off between diversity and engagement in recommendation algorithms.

The most related research to ours is seminal work by Celis et al. [10], who initiated the study of algorithmic approaches to reducing polarization. There are a variety of differences between our work and theirs, highlighted below.

- *Modeling approach.* Celis et al. [10] suggest that a regulator should place pre-determined, fixed upper and lower bounds on the probability that each arm is played so that no user can exclusively see one type of content. Choosing bounds for each type of content, however, may be challenging. (For example, how should bounds on fashion content and major world events compare?) Moreover, if no user is interested in a type of content, it may not make sense to force all users to see it. The regulator would have to make these differential decisions, which would be a divisive and controversial task. These concerns are largely ameliorated under our model.
- *Stronger assumption on the regulator's knowledge.* Celis et al. [10] assume the regulator can control the exact probabilities that the platform shows different types of content to users. In contrast, in our Formulation 3, we propose a tax based on the content that the platform actually showed the user. As we describe in Section 1.1.2, this introduces technical challenges in providing a no-regret algorithm for the platform.
- *Lower bounds.* Our nearly-matching lower bounds help develop a complete understanding of this problem.

Since the multi-armed bandit problem was proposed [33], many variants have been studied, such as bandits with budgets [1, 5, 29], bandits with constraints [3, 12, 26, 27], and bandits with floors on content [11, 34]. Only a few variants [e.g., 20] study multi-agent settings. However, they usually still involve a common reward like in the classical multi-armed bandit problem. There has also been recent work on fairness in multi-armed bandits [e.g., 20, 22] but none of these focus on the issues of filter bubbles and polarization in social networks.

2 NOTATION AND MODEL

We use $\mathcal{P}^{d-1} = \{\boldsymbol{x} \in [0,1]^d : \|\boldsymbol{x}\|_1 = 1\}$ to denote the probability simplex and $[k]$ to denote the set $[k] = \{1, 2, \ldots, k\}$.

Problem definition. There are n users and k categories of content—for example, fashion, sports, right-leaning news, left-leaning news, and so on—each modeled as an *arm* of a *k-armed bandit*. An instance of our problem, denoted $\nu = \{\mathcal{D}_{i,j} : i \in [n], j \in [k]\}$, is defined by reward distributions $\mathcal{D}_{i,j}$ over $[0,1]$ with density function $f_{i,j} : [0,1] \to \mathbb{R}_{\geq 0}$. This distribution models the platform's reward for showing user i content from category j, measured in terms of engagement or ad revenue, for example. The set of all instances ν is denoted $\mathcal{E}^{n,k}$. The mean of user i's reward distribution for arm j is denoted $\mu_{i,j} \in [0,1]$, with $\boldsymbol{\mu}_i = (\mu_{i,1}, \ldots, \mu_{i,k})$. The instance ν is unknown to the platform.

Interaction between platform and users. This interaction takes place over T timesteps. At each timestep $t \in [T]$:

(1) The platform selects an *action*, which is a distribution over arms for each user. This distribution corresponds to a random variable $\boldsymbol{A}_t \in [k]^n$ over arm choices for each of the n users. We use the notation $\boldsymbol{a}_t \in [k]^n$ to denote the specific set of arms the platform plays on round t, so it is a realization of the random variable $\boldsymbol{A}_t$.
(2) Given the set of arms $\boldsymbol{a}_t = (a_{t,1}, \ldots, a_{t,n}) \in [k]^n$, the platform receives a reward for each user. The reward for user i is drawn from the distribution $\mathcal{D}_{i,a_{t,i}}$. We use the random variable $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,n}) \in [0,1]^n$ to denote the platform's reward on round t. We also use $\boldsymbol{x}_t \in [0,1]^n$ to denote a realization of this random variable.

Platform's learning algorithm. The platform uses a learning algorithm, or *policy*, π to decide the distribution over arms at each timestep. On timestep $t \in [T]$, the (randomized) policy π takes as input the history $\boldsymbol{h}_{t-1} = (\boldsymbol{a}_1, \boldsymbol{x}_1, \ldots, \boldsymbol{a}_{t-1}, \boldsymbol{x}_{t-1}) \in ([k]^n \times [0,1]^n)^{t-1}$ and returns the set of arms $\boldsymbol{a}_t \in [k]^n$ that will be played on round t. The

conditional probability that $\boldsymbol{A}_t = \boldsymbol{a}_t$ given the history $\boldsymbol{A}_1 = \boldsymbol{a}_1, \boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{A}_{t-1} = \boldsymbol{a}_{t-1}, \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1}$ is denoted $\pi(\boldsymbol{a}_t \mid \boldsymbol{a}_1, \boldsymbol{x}_1, \ldots, \boldsymbol{a}_{t-1}, \boldsymbol{x}_{t-1})$, or more compactly as $\pi(\boldsymbol{a}_t \mid \boldsymbol{h}_{t-1})$. The notation $\Pi^{n,k}$ denotes the set of all policies π.

Distribution over outcomes. Since the reward distributions are independent, the conditional distribution of the reward $\boldsymbol{X}_t \in [0,1]^n$ given $\boldsymbol{A}_t = \boldsymbol{a}_t = (a_{t,1}, \ldots, a_{t,n}) \in [k]^n$ has density function

$$f_{\boldsymbol{a}_t}(\boldsymbol{x}_t) = \prod_{i=1}^{n} f_{i,a_{t,i}}(x_{t,i}).$$

The interaction between the policy π and the instance ν induces a distribution $\mathbb{P}_{\pi\nu}$ over outcomes with density function

$$f_{\pi\nu}(\boldsymbol{a}_1, \boldsymbol{x}_1, \ldots, \boldsymbol{a}_T, \boldsymbol{x}_T) = \prod_{t=1}^{T} \pi(\boldsymbol{a}_t \mid \boldsymbol{a}_1, \boldsymbol{x}_1, \ldots, \boldsymbol{a}_{t-1}, \boldsymbol{x}_{t-1}) f_{\boldsymbol{a}_t}(\boldsymbol{x}_t). \tag{1}$$

Platform's goal. The platform's overall goal is to choose a policy π that optimizes its total reward

$$\mathop{\mathbb{E}}_{\pi\nu}\left[\sum_{i=1}^{n}\sum_{t=1}^{T} X_{i,t}\right]. \tag{2}$$

For each user $i \in [n]$, the optimal policy would choose the arm j_i that maximizes expected reward: $j_i = \operatorname{argmax}_{j\in[k]} \{\mu_{i,j}\}$. Classic bandit algorithms will eventually converge to this policy. However, repeatedly showing user i content from category j_i traps the user in a filter bubble. In the next sections, we limit the platform's ability to form filter bubbles.

3 A FIRST ATTEMPT TO DISINCENTIVIZE FILTER BUBBLES

We begin with a naive first attempt at disincentivizing filter bubbles and show that it has the harsh unintended consequence of exacerbating "tyranny of the majority": the burden of diversification is borne by those with minority interests. Interestingly, this issue mirrors real-world attempts at diversification where the labor associated with diversification is put disproportionately on members of the underrepresented groups.

To motivate this first attempt, we observe that in a network with severe filter bubbles, members are partitioned into groups which are exposed to disparate types of content. Thus, our first attempt at avoiding filter bubbles ensures that the content recommendations are not too "spread out." We formalize this intuition by requiring that each user's distribution over content is not too far from the average distribution over content shown to the entire population.

More formally, building on the notation from Section 2, let $\pi_i(j \mid \boldsymbol{h}_{t-1})$ denote the marginal probability that the platform shows user i arm j on timestep t given the history $\boldsymbol{h}_{t-1}$, with $\boldsymbol{\pi}_i(\boldsymbol{h}_{t-1}) = (\pi_i(1 \mid \boldsymbol{h}_{t-1}), \ldots, \pi_i(k \mid \boldsymbol{h}_{t-1}))$. Next, let $\bar{\boldsymbol{\pi}}(\boldsymbol{h}_{t-1}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\pi}_i(\boldsymbol{h}_{t-1})$ denote the average of these marginal distributions. The j^{th} component of $\bar{\boldsymbol{\pi}}(\boldsymbol{h}_{t-1})$, denoted $\bar{\pi}(j \mid \boldsymbol{h}_{t-1})$, measures the average probability that arm j is shown to any user. Under our naive first approach, we require that the distance between the vectors $\boldsymbol{\pi}_i(\boldsymbol{h}_{t-1})$ and $\bar{\boldsymbol{\pi}}(\boldsymbol{h}_{t-1})$ is small under the ℓ_∞-norm:

$$\|\boldsymbol{\pi}_i(\boldsymbol{h}_{t-1}) - \bar{\boldsymbol{\pi}}(\boldsymbol{h}_{t-1})\|_\infty = \max_{j\in[k]} |\pi_i(j \mid \boldsymbol{h}_{t-1}) - \bar{\pi}(j \mid \boldsymbol{h}_{t-1})| \le \Delta \tag{3}$$

for some $\Delta > 0$. (The ℓ_∞-norm could be replaced by any norm, but we use the ℓ_∞-norm for this exposition.)

We now show that the optimal policy $\boldsymbol{p}_1^*, \ldots, \boldsymbol{p}_n^* \in \mathcal{P}^{k-1}$ leads to tyranny of the majority, where

$$\boldsymbol{p}_1^*, \ldots, \boldsymbol{p}_n^* = \operatorname*{argmax}_{\boldsymbol{p}_1,\ldots,\boldsymbol{p}_n} \left\{ \sum_{i=1}^{n} \boldsymbol{\mu}_i \cdot \boldsymbol{p}_i : \left\| \boldsymbol{p}_i - \frac{1}{n}\sum_{i'=1}^{n} \boldsymbol{p}_{i'} \right\|_\infty \le \Delta, \forall i \in [n] \right\}.$$

To illustrate the pitfalls of this approach, we analyze a setting where there are two types of content (e.g., left- and right-leaning political content) and the users can be partitioned into disjoint sets where one set only likes content from the first category (i.e., $\boldsymbol{\mu}_i = (1, 0)$). Meanwhile, the other set only likes content from the second category (i.e., $\boldsymbol{\mu}_i = (0, 1)$). Without loss of generality, we assume that the former set—which we denote as N—is the majority.

When $\Delta \geq \frac{|N|}{n}$, the constraints are meaningless and allow for full personalization: $\boldsymbol{p}_i^* = (1, 0)$ if $i \in [N]$ and $\boldsymbol{p}_i^* = (0, 1)$ if $i \notin [N]$. Therefore, we analyze the case where $\Delta < \frac{|N|}{n}$. We show that under the optimal policy, the majority group will be able to exclusively see the content that they enjoy: $\boldsymbol{p}_i^* = (1, 0)$ if $i \in N$. Meanwhile, the minority group's recommendations take a hit in order to ensure that the constraints are satisfied. In particular, for all $i \notin N$, $\boldsymbol{p}_i^* = \left(1 - \frac{n\Delta}{|N|}, \frac{n\Delta}{|N|}\right)$. The proof of the following lemma is in Appendix A.

Lemma 3.1. *Suppose that there are $k = 2$ arms and for some set $N \subseteq [n]$ with $|N| \geq \frac{n}{2}$, $\boldsymbol{\mu}_i = (1, 0)$ for all $i \in N$ and $\boldsymbol{\mu}_i = (0, 1)$ for all $i \notin N$. If $\Delta < \frac{|N|}{n}$, then $\boldsymbol{p}_i^* = (1, 0)$ if $i \in N$ and $\boldsymbol{p}_i^* = \left(1 - \frac{n\Delta}{|N|}, \frac{n\Delta}{|N|}\right)$ otherwise.*

Lemma 3.1 illustrates that under this approach, tyranny of the majority prevails at the expense of minority interests.

4 EQUITABLE APPROACHES TO DISINCENTIVIZING FILTER BUBBLES

Motivated by Section 3, we propose three different formulations for disincentivizing filter bubbles that avoid tyranny of the majority. The intuition behind these approaches is built upon the following two pillars:

(1) To avoid tyranny of the majority, users should primarily be recommended content they are most interested in,
(2) But to avoid filter bubbles, that content must contain a flavor of the content shown to the entire population.

We show that it is possible to achieve both of these ends. If both conditions are satisfied, then a policy like that of Lemma 3.1 where the majority group sees no minority content is not possible. By the first requirement, groups with minority interests will be recommended content that they are interested in, which means that by the second requirement, the majority group's content recommendations will contain a small amount of that minority content, and vice versa.

4.1 Formulation 1: Personalization constraint

In our first formulation, we require that for each user $i \in [n]$, $\boldsymbol{\pi}_i(\boldsymbol{h}_{t-1})$ is at least $\gamma \bar{\boldsymbol{\pi}}(\boldsymbol{h}_{t-1})$ for some $\gamma \in [0, 1]$:

$$\boldsymbol{\pi}_i(\boldsymbol{h}_{t-1}) \geq \gamma \bar{\boldsymbol{\pi}}(\boldsymbol{h}_{t-1}). \tag{4}$$

Each user's recommendations become less personalized as γ grows.

To illustrate the benefit of this approach over that of Section 3, we analyze the same polarized example where there is a majority group N with $\boldsymbol{\mu}_i = (1, 0)$ for all $i \in N$. For the minority group, $\boldsymbol{\mu}_i = (0, 1)$ for all $i \notin N$. For all $\gamma \leq \frac{1}{2}$, we show that under the optimal policy, the majority of each group's content recommendations match their interests, but both groups see some content that appeals to the opposing group. In this case the optimal policy is defined as

$$\boldsymbol{p}_1^*, \ldots, \boldsymbol{p}_n^* = \underset{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \boldsymbol{\mu}_i \cdot \boldsymbol{p}_i : \boldsymbol{p}_i \geq \frac{\gamma}{n} \sum_{i'=1}^{n} \boldsymbol{p}_{i'}, \forall i \in [n] \right\}. \tag{5}$$

The proof of the following lemma is in Appendix B.

Lemma 4.1. *Suppose that there are $k = 2$ arms and for some set $N \subseteq [n]$, $\boldsymbol{\mu}_i = (1, 0)$ for all $i \in N$ and $\boldsymbol{\mu}_i = (0, 1)$ for all $i \notin N$. For $\gamma \leq \frac{1}{2}$, the optimal policy has the following form:*

$$\boldsymbol{p}_i^* = \begin{cases} \left(1 - \frac{\gamma(n-|N|)}{n}, \frac{\gamma(n-|N|)}{n}\right) & \text{if } i \in N \\ \left(\frac{\gamma|N|}{n}, 1 - \frac{\gamma|N|}{n}\right) & \text{if } i \notin N. \end{cases}$$

Since $\gamma \leq \frac{1}{2}$, this policy ensures that users are mostly recommended content that they are interested in: $\boldsymbol{\mu}_i \cdot \boldsymbol{p}_i^* \geq 1 - \gamma \geq \frac{1}{2}$ for all $i \in [n]$. However, they are still shown a small fraction of content that the other set of the population is interested in. We note that when N is the majority group ($|N| \geq \frac{n}{2}$), the minority group $[n] \setminus N$ still sees more content that they are not interested in than the majority group because $\frac{\gamma|N|}{n} \geq \frac{\gamma(n-|N|)}{n}$. However, the burden of diversification is split far more equally among the two groups than in Lemma 3.1. The policy mirrors a typical mode of community forum discussions where members split time between listening to the opinions of each person in the entire group (for a γ-fraction of the time) and breaking into focus groups about specific topics (for a $(1-\gamma)$-fraction of the time).

In Section 5, we provide upper and lower bounds on the platform's *regret* with respect to the optimal policies $\boldsymbol{p}_1^*, \ldots, \boldsymbol{p}_n^*$ defined in Equation (5). Regret measures the difference between the total reward of the optimal policy and that of the platform's policy π. In other words, for any instance ν and policy π, the expected regret is defined as

$$R_{T,1}(\pi, \nu) = T \sum_{i=1}^{n} \boldsymbol{p}_i^* \cdot \boldsymbol{\mu}_i - \mathop{\mathbb{E}}_{\pi\nu}\left[\sum_{i=1}^{n}\sum_{t=1}^{T} X_{i,t}\right]. \tag{6}$$

4.2 Formulation 2: Personalization penalty

We analyze a second formulation where there are no constraints on the platform's policy, but the platform is penalized based on the extent to which Equation (4) is violated. Given a parameter $\eta \geq 0$, this penalty is defined as

$$\eta \sum_{i=1}^{n}\sum_{j=1}^{k} \max\left\{\gamma\bar{\boldsymbol{\pi}}(j \mid \boldsymbol{h}_{t-1}) - \pi_i(j \mid \boldsymbol{h}_{t-1}), 0\right\}.$$

In other words, the platform's goal is to maximize its cumulative reward

$$\begin{aligned} \text{reward}_2(\pi, \nu; \eta, \gamma) &= \mathop{\mathbb{E}}_{\pi\nu}\left[\sum_{i=1}^{n}\left(\sum_{t=1}^{T} X_{i,t} - \eta \sum_{j=1}^{k} \max\left\{\gamma\bar{\boldsymbol{\pi}}(j \mid \boldsymbol{h}_{t-1}) - \pi_i(j \mid \boldsymbol{h}_{t-1}), 0\right\}\right)\right] \\ &= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi\nu}\left[\sum_{i=1}^{n}\left(\boldsymbol{\mu}_i \cdot \boldsymbol{\pi}_i(\boldsymbol{h}_{t-1}) - \eta \sum_{j=1}^{k} \max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n} \pi_{i'}(j \mid \boldsymbol{h}_{t-1}) - \pi_i(j \mid \boldsymbol{h}_{t-1}), 0\right\}\right)\right]. \end{aligned} \tag{7}$$

The policy that maximizes Equation (7) is history independent and can be written as $\boldsymbol{p}^* = (\boldsymbol{p}_1^*, \ldots, \boldsymbol{p}_n^*)$ with $\boldsymbol{p}_i^* \in \mathcal{P}^{k-1}$. The expected regret of a policy π under this formulation is $R_{T,2}(\pi, \nu) = \text{reward}_2(\boldsymbol{p}^*, \nu; \eta, \gamma) - \text{reward}_2(\pi, \nu; \eta, \gamma)$.

4.3 Formulation 3: Personalization penalty on the empirical distribution

Sections 4.1 and 4.2 describe models in which the platform is subject to constraints or penalties based on the *true* distribution over content that it shows users. However, an auditor may only have access to the *realizations* of those distributions—that is, the set of arms $a_{t,i} \in [k]$ shown to each user i at timestep t. Formulation 3 covers a setting in which a regulator penalizes the platform at the end of the T timesteps based on the empirical distribution over content. Specifically, let $\hat{p}_{i,j} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\{A_{t,i}=j\}}$ be the average number of times that the platform pulls arm j for user i. At the

end of the T timesteps, the platform is penalized based on how small $\hat{p}_{i,j}$ is compared to $\frac{\gamma}{n}\sum_{i'=1}^{n}\hat{p}_{i',j}$. In particular, given a normalizing factor η, we define a penalty that is the analogue of Equation (7):

$$\eta \sum_{i=1}^{n}\sum_{j=1}^{k}\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\hat{p}_{i',j}-\hat{p}_{i,j},0\right\}.$$

The platform's goal is therefore to maximize their expected total payoff minus this penalty, which is equal to

$$\begin{aligned}\text{reward}_3(\pi,\nu;\eta,\gamma) &= \mathop{\mathbb{E}}_{\pi\nu}\left[\sum_{i=1}^{n}\left(\sum_{t=1}^{T}X_{i,t}-\eta\sum_{j=1}^{k}\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\hat{p}_{i',j}-\hat{p}_{i,j},0\right\}\right)\right]\\ &= \sum_{i=1}^{n}\left(\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi\nu}\left[\boldsymbol{\mu}_i\cdot\boldsymbol{\pi}_i(\boldsymbol{h}_{t-1})\right]-\eta\sum_{j=1}^{k}\mathop{\mathbb{E}}_{\pi\nu}\left[\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\hat{p}_{i',j}-\hat{p}_{i,j},0\right\}\right]\right). \qquad (8)\end{aligned}$$

Let π^* be the policy that maximizes Equation (8). The regret of π is $R_{T,3}(\pi,\nu)=\text{reward}_3(\pi^*,\nu;\eta,\gamma)-\text{reward}_3(\pi,\nu;\eta,\gamma)$.

A key difference between Equation (7) and Equation (8) is that in Equation (7), the platform is penalized at every timestep whereas in Equation (8), the platform is penalized at the end of the T timesteps. We make this distinction because the empirical distribution over content at a single timestep would be extremely noisy.

5 REGRET ANALYSIS

In this section, we discuss algorithms that the platform can use to minimize regret in the three formulations from Section 4. We also provide a nearly-matching lower bound on regret for Formulation 1 in Section 5.1.3.

5.1 Regret analysis for Formulation 1

We begin with lower bounds on regret under Formulation 1. In Section 5.1.1, we show that a variant of the UCB algorithm has regret $O(n\sqrt{Tk})$ for $\gamma<1$ and in Section 5.1.2, we show that a different variant of UCB has regret $O(\sqrt{nkT})$ for $\gamma=1$. We then prove in Section 5.1.3 that these bounds are nearly optimal: for $\gamma\le\frac{1}{2}$ and $k=2$, no algorithm can achieve regret better than $\Omega(n\sqrt{T})$, and for all $k\ge 2$ and $\gamma\in[0,1]$ (including $\gamma>\frac{1}{2}$) our bound is $\Omega(\sqrt{nkT})$.

The transition from a $\Theta(n)$ to a $\Theta(\sqrt{n})$ dependence illustrates that as γ grows, the platform is better able to compete with the optimal policy subject to the γ constraints. As γ grows, the platform has a smaller set of distributions that it can use to compete with the optimal policy while obeying the γ constraints. However, for the same reason, the cumulative reward of the optimal policy shrinks as γ grows. Intuitively, the transition from a $\Theta(n)$ to a $\Theta(\sqrt{n})$ dependence as γ grows illustrates that the optimal policy's reward degrades faster than the platform's ability to compete with that policy.

5.1.1 Regret upper bound when $\gamma<1$. We analyze a multi-agent variant of the UCB algorithm, which we call n-UCB, and show that it has regret $O(n\sqrt{Tk})$ when $\gamma<1$. The n-UCB algorithm essentially runs a copy of classic UCB for each user, but coordinates amongst these n UCB copies to ensure that they satisfy the global constraints. This requires n-UCB to play distributions over arms from the set of distributions $(\boldsymbol{p}_1,\ldots,\boldsymbol{p}_n)$ that satisfy the constraints: $\boldsymbol{p}_i\ge\frac{\gamma}{n}\sum_{i'=1}^{n}\boldsymbol{p}_{i'}$ for all $i\in[n]$. This is in contrast to the classic case where UCB plays a single arm at each timestep. For completeness, we include a full description of n-UCB (Algorithm 1) and the proof of the following theorem in Appendix C.

Theorem 5.1. *Let π be the policy of n-UCB. Then $R_{T,1}(\pi,\nu)=\tilde{O}(n\sqrt{kT})$.*

5.1.2 Regret upper bound when $\gamma=1$. When $\gamma=1$, all users must be shown the same distribution of content. We can therefore reduce our problem to a single-agent bandit problem with the reward distributions $\mathcal{D}_j=\sum_{i=1}^{n}\mathcal{D}_{i,j}$ for all

arms $j \in [k]$. We adapt the robust-UCB framework by Bubeck et al. [8] with the median-of-means estimator [2], as summarized by Algorithm 2 in Appendix D. The full proof of the following theorem is in Appendix D.

Theorem 5.2. *Let π be the policy of Robust-UCB. Then $R_{T,1}(\pi, \nu) = \tilde{O}(\sqrt{nkT})$.*

5.1.3 Regret lower bound. In this section, we provide nearly-matching regret lower bounds. Our first bound holds when there are $k = 2$ arms, $\gamma \leq \frac{1}{2}$, and n is sufficiently large ($n > 100$). In this case, we prove a regret lower bound of $\Omega(n\sqrt{T})$. Meanwhile, for all $k \geq 2$ and $\gamma \in [0, 1]$ (including $\gamma > \frac{1}{2}$), we provide a bound of $\Omega(\sqrt{nkT})$. We begin with our main result (Theorem 5.3) and show in Corollary 5.4 that it implies a regret bound of $\Omega(n\sqrt{T})$ for $\gamma \leq \frac{1}{2}$ and $n > 100$.

Theorem 5.3. *For all $T \geq 4$, the regret is lower bounded as follows:*

$$\inf_{\pi \in \Pi^{n,2}} \sup_{\nu \in \mathcal{E}^{n,2}} R_{T,1}(\pi, \nu) \geq \max\left\{\sqrt{\frac{T}{8}}\left(\frac{n}{8e} - \gamma\left(\frac{n}{8e} + \sqrt{\frac{n}{2\pi}}\right)\right), \frac{\sqrt{nT}}{16e}\right\}.$$

Proof. This theorem follows directly from Lemmas 5.5 and 5.6. □

Corollary 5.4. *For all $n > 100$, $\gamma \leq \frac{1}{2}$, and $T \geq 4$, the regret is lower bounded as*

$$\inf_{\pi \in \Pi^{n,2}} \sup_{\nu \in \mathcal{E}^{n,2}} R_{T,1}(\pi, \nu) \geq \frac{n\sqrt{T}}{900}.$$

We now provide a proof sketch of the first part of Theorem 5.3. The full proof is in Appendix E.

Lemma 5.5. *For all $T \geq 1$, the regret is lower bounded as follows:*

$$\inf_{\pi \in \Pi^{n,2}} \sup_{\nu \in \mathcal{E}^{n,2}} R_{T,1}(\pi, \nu) \geq \sqrt{\frac{T}{8}}\left(\frac{n}{8e} - \gamma\left(\frac{n}{8e} + \sqrt{\frac{n}{2\pi}}\right)\right). \tag{9}$$

Proof sketch. Our proof is based on worst-case instances $\nu_{\boldsymbol{b}}$ defined for any vector $\boldsymbol{b} \in \{0, 1\}^n$. For each agent $i \in [n]$, their reward distributions for the two arms are Bernoulli with means $\boldsymbol{\mu}_i = (\mu_{i,0}, \mu_{i,1})$ where $\boldsymbol{\mu}_i = \left(\frac{1}{2} + \epsilon, \frac{1}{2}\right)$ if $b_i = 0$, $\boldsymbol{\mu}_i = \left(\frac{1}{2}, \frac{1}{2} + \epsilon\right)$ if $b_i = 1$, and $\epsilon = \sqrt{\frac{1}{8T}}$. We lower bound the expected regret $\mathbb{E}\left[R_{T,1}(\pi, \nu_{\boldsymbol{b}})\right]$ by the right-hand-side of Equation (9), where the expectation is over both the draw of the vector $\boldsymbol{b} \sim \text{Unif}(\{0, 1\}^n)$ and the distribution over outcomes $\mathbb{P}_{\pi \nu_{\boldsymbol{b}}}$. This implies that there exists an instance $\nu_{\boldsymbol{b}}$ such that $R_{T,1}(\pi, \nu_{\boldsymbol{b}}) \geq \sqrt{\frac{T}{8}}\left(\frac{n}{8e} - \gamma\left(\frac{n}{8e} + \sqrt{\frac{n}{2\pi}}\right)\right)$.

Without constraints, the optimal policy would exclusively show arm 0 to each user $i \in [n]$ with $b_i = 0$ since it has higher reward for these users, and similarly it would exclusively show arm 1 to each user $i \in [n]$ with $b_i = 1$. Due to the constraints, both the optimal policy and the policy π must show some users the "wrong" arm on a non-negligible fraction of rounds. In total, the policy π will lose the following reward from showing users the wrong arms:

$$\epsilon \mathop{\mathbb{E}}_{\boldsymbol{b}}\left[\sum_{t=1}^{T}\left(\sum_{i:b_i=0} \mathop{\mathbb{E}}_{\pi \nu_{\boldsymbol{b}}}\left[\pi_i(1 \mid \boldsymbol{h}_{t-1})\right] + \sum_{i:b_i=1} \mathop{\mathbb{E}}_{\pi \nu_{\boldsymbol{b}}}\left[\pi_i(0 \mid \boldsymbol{h}_{t-1})\right]\right)\right].$$

We begin by proving that that optimal policy loses at most

$$\sqrt{\frac{T}{8}} \cdot \frac{\gamma(n-1)}{2} \tag{10}$$

total reward from showing users the wrong arms. Meanwhile, we prove that any policy π will lose at least

$$\sqrt{\frac{T}{8}}\left(\frac{\gamma n}{2} + \frac{n}{8e} - \gamma\left(\frac{n}{8e} + \sqrt{\frac{n}{2\pi}}\right)\right) \tag{11}$$

total reward. We prove this by showing that for any user $i \in [n]$, the distribution over outcomes conditioned on $b_i = 0$ is close to the distribution over outcomes conditioned on $b_i = 1$. Intuitively, this means that any policy π will struggle to distinguish whether $b_i = 0$ or $b_i = 1$. Taking the difference of (11) and (10) implies the lemma. □

We conclude by proving that for all $\gamma \in [0, 1]$ and $k \geq 2$, regret is lower bounded by $\frac{1}{16e}\sqrt{nT(k-1)}$. The proof is similar to that of existing bandit lower bounds [e.g., 24, Theorem 15.2], so we include it for completeness in Appendix E.

Lemma 5.6. *For all $T \geq \frac{7(k-1)}{n}$, the regret is lower bounded as follows:*

$$\inf_{\pi \in \Pi^{n,k}} \sup_{\nu \in \mathcal{E}^{n,k}} R_{T,1}(\pi, \nu) \geq \frac{\sqrt{nT(k-1)}}{16e}.$$

5.2 Regret analysis for Formulation 2

Under Formulation 2, a variation on UCB we call Penalty-UCB (Algorithm 3) achieves regret $\tilde{O}(n\sqrt{kT})$. Penalty-UCB maintains estimates $\hat{\boldsymbol{\mu}}_i^{(t)}$ of each $\boldsymbol{\mu}_i$ and selects the distribution maximizing the estimated reward minus the penalty:

$$\left(\boldsymbol{p}_i^{(t)}\right)_{i \in [n]} = \operatorname{argmax} \left\{ \sum_{i=1}^{n} \boldsymbol{p}_i \cdot \hat{\boldsymbol{\mu}}_i^{(t)} - \eta \sum_{j=1}^{k} \max \left\{ \frac{\gamma}{n} \sum_{i'=1}^{n} p_{i',j} - p_{i,j}, 0 \right\} \right\}.$$

For completeness, we include the proof of the following theorem in Appendix F.

Theorem 5.7. *Let π be the policy of Penalty-UCB. Then $R_{T,2}(\pi, \nu) = \tilde{O}(n\sqrt{kT})$.*

5.3 Regret analysis for Formulation 3

A key challenge under Formulation 3 is that the platform's optimal strategy, given perfect information about the reward distributions $\mathcal{D}_{i,j}$, may be *history dependent*. For example, the platform may choose to increase or decrease personalization dynamically based on the empirical distribution of content chosen thus far. Nonetheless, we show that Penalty-UCB (Algorithm 3) competes with the optimal history-dependent policy by reducing our analysis to that of Section 5.2. We use the notation π^* to denote the optimal policy that maximizes Equation (8).

First, we show that under Formulation 2, the optimal policy obtains a larger reward (measured in terms of reward$_2$) than π^* under Formulation 3 (measured in terms of reward$_3$). The full proof is in Appendix G.

Lemma 5.8. *Let $\boldsymbol{p}^* = (\boldsymbol{p}_1^*, \ldots, \boldsymbol{p}_n^*)$ with $\boldsymbol{p}_i^* \in \mathcal{P}^{k-1}$ be the policy that maximizes reward$_2\left(\boldsymbol{p}, \nu; \frac{\eta}{T}, \gamma\right)$. Then*

$$\text{reward}_2\left(\boldsymbol{p}^*, \nu; \frac{\eta}{T}, \gamma\right) \geq \text{reward}_3(\pi^*, \nu; \eta, \gamma).$$

Proof sketch. For any arm $j \in [k]$, we can exchange the expectation and the maximum in Equation (8) as follows:

$$\mathop{\mathbb{E}}_{\pi^* \nu}\left[\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\hat{p}_{i',j} - \hat{p}_{i,j}, 0\right\}\right] \geq \max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\mathop{\mathbb{E}}_{\pi^* \nu}\left[\hat{p}_{i',j}\right] - \mathop{\mathbb{E}}_{\pi^* \nu}\left[\hat{p}_{i,j}\right], 0\right\}.$$

By definition of $\mathbb{E}_{\pi^* \nu}\left[\hat{p}_{i,j}\right]$, this allows us to show that reward$_3(\pi^*, \nu; \eta, \gamma)$ is upper-bounded by

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\left(\sum_{t=1}^{T}\mu_{i,j}\mathop{\mathbb{E}}_{\pi^* \nu}\left[\pi_i^*(j \mid \boldsymbol{h}_{t-1})\right] - \eta \max\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\gamma}{n}\sum_{i'=1}^{n}\mathop{\mathbb{E}}_{\pi^* \nu}\left[\pi_{i'}^*(j \mid \boldsymbol{h}_{t-1})\right] - \mathop{\mathbb{E}}_{\pi^* \nu}\left[\pi_i^*(j \mid \boldsymbol{h}_{t-1})\right]\right), 0\right\}\right). \tag{12}$$

We next define the history-independent policy $\boldsymbol{p} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n)$ such that

$$p_{i,j} = \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi^* \nu}\left[\pi_i^*(j \mid \boldsymbol{h}_{t-1})\right].$$

We rearrange Equation (12) and use the definition of $\boldsymbol{p}^*$ to get that

$$\begin{aligned}\text{reward}_3(\pi^*, \nu; \eta, \gamma) &\le \sum_{i=1}^{n}\sum_{j=1}^{k}\left(T\mu_{i,j}p_{i,j} - \eta \max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n} p_{i',j} - p_{i,j}, 0\right\}\right) \\ &\le \sum_{i=1}^{n}\sum_{j=1}^{k}\left(T\mu_{i,j}p^*_{i,j} - \eta \max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n} p^*_{i',j} - p^*_{i,j}, 0\right\}\right) = \text{reward}_2\left(\boldsymbol{p}^*, \nu; \frac{\eta}{T}, \gamma\right).\end{aligned}$$

□

Next, we show that for any policy π that deterministically plays each of the k arms once in the first k rounds, the difference between its rewards under Formulations 2 and 3 is bounded. This condition holds for Penalty-UCB (Algorithm 3) and could be removed with a slightly more involved analysis. The full proof is in Appendix G.

Lemma 5.9. *Let π be any policy such that $\pi_i(t \mid \boldsymbol{h}_{t-1}) = 1$ for all $t \le k$ and $i \in [n]$. For any instance ν,*

$$\textit{reward}_2\left(\pi, \nu; \frac{\eta}{T}, \gamma\right) \le \textit{reward}_3(\pi, \nu; \eta, \gamma) + \eta nk(\gamma+1)\sqrt{\frac{10\log T}{T}}.$$

Proof sketch. For any arm $j \in [k]$, we can exchange the expectation and the maximum in Equation (7) as follows:

$$\frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi\nu}\left[\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\pi_{i'}(j \mid \boldsymbol{h}_{t-1}) - \pi_i(j \mid \boldsymbol{h}_{t-1}), 0\right\}\right] \ge \max\left\{\mathop{\mathbb{E}}_{\pi\nu}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\gamma}{n}\sum_{i'=1}^{n}\mathbf{1}_{\{A_{t,i'}=j\}} - \mathbf{1}_{\{A_{t,i}=j\}}\right)\right], 0\right\}. \quad (13)$$

Next, we use a result by Aven [4] to show that the right-hand-side of Equation (13) is lower-bounded by

$$\mathop{\mathbb{E}}_{\pi\nu}\left[\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\hat{p}_{i',j} - \hat{p}_{i,j}, 0\right\}\right] - \sqrt{\frac{1}{2T^2}\cdot\text{Var}\left(\sum_{t=1}^{T}\left(\frac{\gamma}{n}\sum_{i'=1}^{n}\mathbf{1}_{\{A_{t,i'}=j\}} - \mathbf{1}_{\{A_{t,i}=j\}}\right)\right)}. \quad (14)$$

We upper-bound the variance term in Equation (14) by $20T(\gamma+1)^2\log T$, which implies the lemma statement. □

Our regret bound follows from Lemmas 5.8 and 5.9 as well as Theorem 5.7. The proof is in Appendix G.

Theorem 5.10. *Let π be the policy played by Algorithm 3. Then the regret is bounded as*

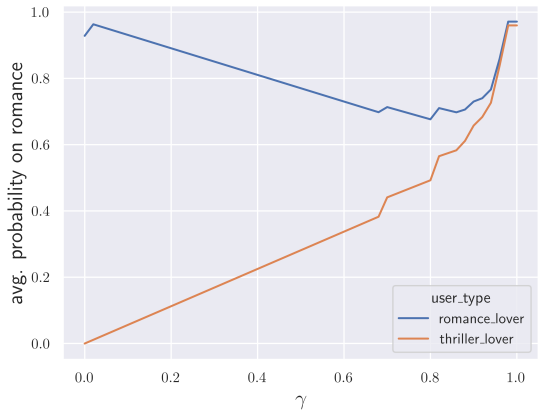
$$\textit{reward}_3(\pi^*, \nu; \eta, \gamma) - \textit{reward}_3(\pi, \nu; \eta, \gamma) = \tilde{O}\left(n\sqrt{kT} + \frac{\eta nk(1+\gamma)}{\sqrt{T}}\right).$$

Even if η grows linearly in T, the regret bound in Theorem 5.10 will only grow with $\sqrt{T}$.

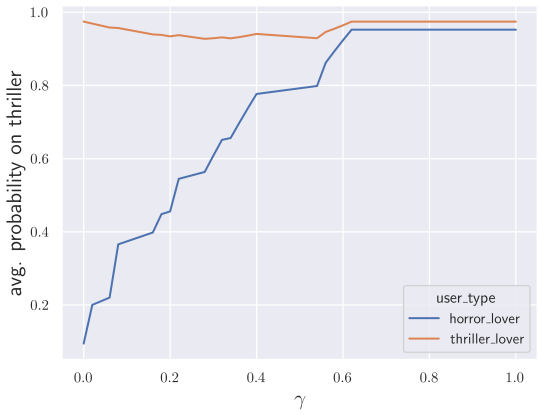
6 EMPIRICAL RESULTS

To explore how our framework impacts exposure diversity in practice, we test it out on real world data: the MovieLens dataset [18] which describes people's expressed preferences for movies[1]. These preferences take the form of <user, item, rating, timestamp> tuples, each the result of a user giving a 0–5 star rating for a movie at a particular time.

[1]We use this dataset in order to analyze our methods on real-world user preferences, recognizing that movie recommendation filter bubbles would likely not be as pernicious as political news filter bubbles, for example.



(a) Average probability placed on romance for romance- and thriller-lovers.



(b) Average probability placed on thriller for horror- and thriller-lovers.

Fig. 1. Polarization cap: Content changes as a function of γ for 2 user groups. We compute the optimal policy for 50 values of γ equally spaced between $[0, 1]$.

6.1 Experimental setup

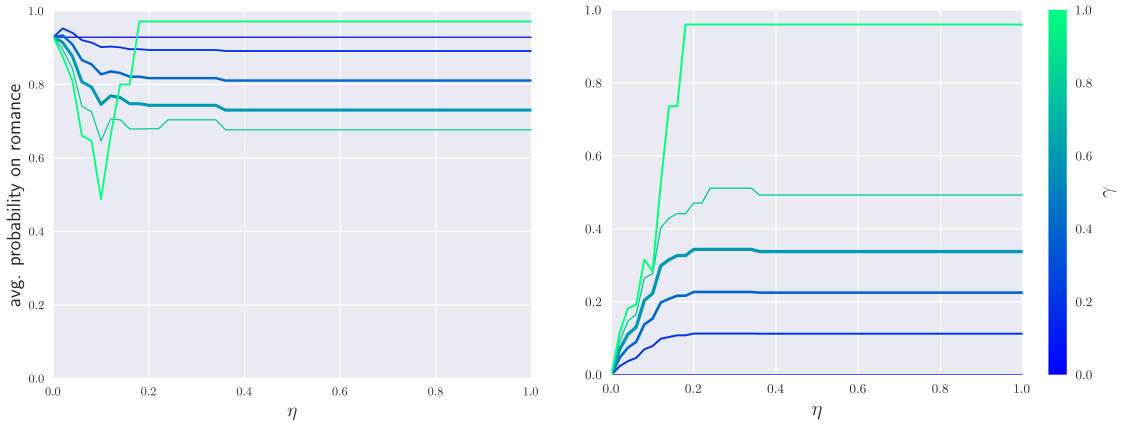
There are $n = 58$ users, randomly selected from the database, and a set $\mathcal{K}$ of $k = 18$ movie genres: $\mathcal{K} = \{$Action, Adventure, Animation, Children, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western$\}$. Each genre is paired with an associated index in $[k]$ determined by alphabetically ordering $\mathcal{K}$. For each movie $m \in \mathcal{M}$, where $\mathcal{M}$ is the set of all movies, there is an associated genre set $m_K \subseteq [k]$ with $|m_K| \geq 1$ (a movie could belong to multiple genres). We use the ratings data to generate preferences for the users. For each movie $m \in \mathcal{M}_i$, where $\mathcal{M}_i$ is the set of movies watched by user $i \in [n]$, the user gives a numeric rating $r_{i,m}$ on a 5-star scale with half-star increments: $r_{i,m} \in \{0.5, 1, 1.5, \ldots, 5.0\}$. We sum these ratings by genre and divide by the number of movies that user i watched from that genre. This results in an average rating $\mu_{i,j} \in [0, 5]$ per genre $j \in [k]$. Finally, we divide $\mu_{i,j}$ by 5 so that $\mu_{i,j} \in [0, 1]$. In the end,

$$\mu_{i,j} = \frac{\sum_{m \in \mathcal{M}_i} r_{i,m} \cdot \mathbb{1}\{j \in m_K\}}{\sum_{m \in \mathcal{M}_i} \mathbb{1}\{j \in m_K\}} \cdot \frac{1}{5}.$$

Using the $\boldsymbol{\mu}_i$s as the mean reward vectors, we use linear programs (LPs) to solve for the optimal policy under no constraints and under both our polarization cap and polarization tax frameworks.

6.2 Effect of the polarization cap and tax on content recommendations

We begin by investigating the effects that our constraints from Formulation 1 (Section 4.1) have on the optimal content distribution. These experiments provide a parallel to Lemma 4.1, which shows that in a polarized population, users share the burden of diversification. To model a polarized society, we restrict our attention to two dissimilar genres: thriller and romance. In this restricted space, $\boldsymbol{\mu}_i \in [0, 1]^2$. We call the users who prefer the thriller genre *thriller-lovers* and those who prefer the romance genre *romance-lovers*. In Figure 1a, we plot the probability placed on romance by the optimal policy (which maximizes $\sum_{i=1}^{n} \boldsymbol{\mu}_i \cdot \boldsymbol{p}_i$ such that $\boldsymbol{p}_i \geq \frac{\gamma}{n} \sum_{i'=1}^{n} \boldsymbol{p}_{i'}$) as a function of γ. For comparison, we run the same experiments for two similar genres: thriller and horror. In Figure 1b, we plot the probability placed on thriller.



(a) Probability placed on romance for romance-lovers.

(b) Probability placed on romance for thriller-lovers.

Fig. 2. Polarization tax: Content changes as function of γ and η for romance- and thriller-lovers. We compute the optimal policy for 6 values of γ equally spaced between $[0, 1]$ and 50 values of η equally spaced between $[0, 1]$.

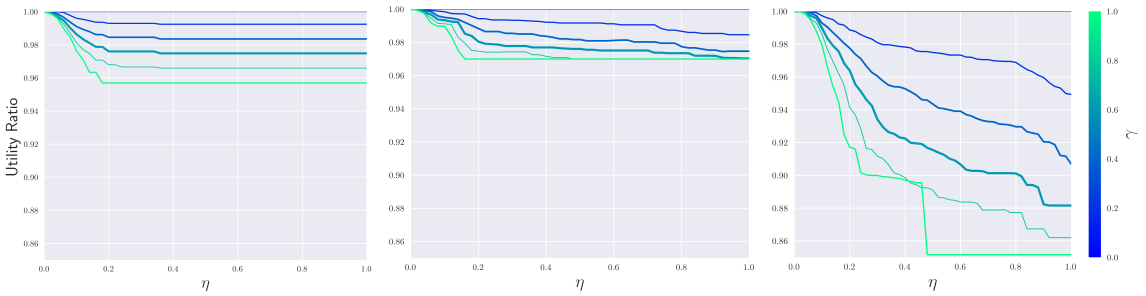
In both Figures 1a and 1b, as γ increases, the content recommendations become more homogeneous. However, the rates at which the recommendations become homogeneous are significantly different. In Figure 1a where the users are polarized, the content recommendations converge slowly. It is not until $\gamma = 0.9$ that the content is completely homogeneous. Meanwhile, in Figure 1b where the groups of users are similar, the recommendations become homogeneous at a faster rate. In this example, they converge at approximately $\gamma = 0.6$.

Under Formulation 2—where the platform is subject to a penalty (Equation (7))—we perform the same experiments for romance- versus thriller-lovers. These experiments are illustrated in Figure 2 where we vary both η and γ. As before, the content distributions converge as γ increases. However, η serves to modulate the impact of γ on content recommendations. When η is small, the platform prefers to pay some tax to show more personalized content than they would under the hard constraint from Formulation 1. In fact, in Figure 2a, we see that even when $\gamma = 1$ (so the platform is penalized for any level of personalization), the platform prefers to pay some tax and personalize its recommendations, but for sufficiently large η (approximately $\eta \gtrsim 0.2$), the platform switches to obeying the γ constraint and paying no tax. For the other values of γ, the content recommendations change more gradually as η grows. However, after a certain point ($\eta \gtrsim 0.4$), only the value of γ leads to differences in the optimal policy.

6.3 Effect of the polarization tax on user utility

We next investigate the impact of the polarization penalty (Formulation 2) on the users' utility. We analyze the same setting from Section 6.2 where there is a polarized society consisting of romance- and thriller-lovers. Letting $(\boldsymbol{p}_i^{\gamma;\eta})_{i\in[n]}$ be the optimal policy under Equation (7) and $(\boldsymbol{p}_i^*)_{i\in[n]}$ policy with no penalty ($\eta = 0$), Figure 3a plots the ratio of the users' cumulative utilities under these two policies: $\sum \boldsymbol{\mu}_i \cdot \boldsymbol{p}_i^{\gamma;\eta} \,/\, \sum \boldsymbol{\mu}_i \cdot \boldsymbol{p}_i^*$. Figure 3b plots the same quantity under the homogenous society from Section 6.2 with only horror- and thriller-lovers.

In Figures 3a and 3b, utility decreases as γ and η grow, but there is a larger utility loss for the polarized group (Figure 3a) compared to the homogeneous group (Figure 3b). Interestingly, in the homogenous group (Figure 3b), utility continuously decreases as η increases, while in the polarized group (Figure 3a), the utility loss eventually flattens out.



(a) Romance- and thriller-lovers (b) Horror- and thriller-lovers (c) All user types

Fig. 3. Multiplicative utility loss as a function of γ and η.

This is because when the population is homogeneous, as η increases the platform only recommends one genre rather than pay the tax, even when γ is small. However, when users are polarized (Figure 2), the platform recommends both genres and pays some tax for most values of γ. It is only when $\gamma = 1$ that the platform recommends only one genre.

Figure 3c plots the same quantity but without restricting the genres ($\boldsymbol{\mu}_i \in [0,1]^{18}$). Since the users' preferences are more diverse, the users' cumulative utility suffers a larger but still minimal loss. This is because each user now sees a larger share of content they might not prefer since there are more groups on the platform. Finally, in Appendix H, we provide plots illustrating the additive utility loss (rather than multiplicative).

7 CONCLUSIONS AND DISCUSSION

Our work proposes a flexible approach to disincentivizing filter bubbles that adapts to the interests of the individuals on the network. Under our model, if some users are shown a particular type of content, then all users see at least a small amount of that content. We show that our model incentivizes diversity in a way that is equitable to users on the platform and discuss algorithms for recommending content under our framework.

There remain many open questions around disincentivizing polarization in social networks. One might want to distinguish between the content of protected minority groups and that of hate-focused or troll groups. Our current formulation does not distinguish between these situations. One could consider a model where the penalties or cap might scale non-linearly with the size of the group, allowing for more effective moderation. In addition, there is more work to be done to understand the precise impacts of our constraints on the utility of the users and platform. Rewards could represent the profit of the platform or the utility of its users, and our current analysis does not address this distinction. However, the difference could be important when there is a wealth disparity between groups and differences in utility of the users might not easily map to differences in the platform's revenue. A related direction could be to extend our model to maximize popular and well-studied notions of fairness like *Nash social welfare*.

REFERENCES

[1] Shipra Agrawal and Nikhil R Devanur. 2014. Bandits with concave rewards and convex knapsacks. In *ACM Conference on Economics and Computation (EC)*. 989–1006.

[2] Noga Alon, Yossi Matias, and Mario Szegedy. 1996. The space complexity of approximating the frequency moments. In *Proceedings of the Annual Symposium on Theory of Computing (STOC)*. 20–29.

[3] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. 2019. Linear stochastic bandits under safety constraints. In *Conference on Neural Information Processing Systems (NeurIPS)*.

[4] Terje Aven. 1985. Upper (lower) bounds on the mean of the maximum (minimum) of a number of random variables. *Journal of applied probability* 22, 3 (1985), 723–728.

[5] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. 2013. Bandits with knapsacks. In *Symposium on Foundations of Computer Science (FOCS)*.

[6] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[7] Engin Bozdag and Jeroen Van Den Hoven. 2015. Breaking the filter bubble: democracy and design. *Ethics and information technology* 17 (2015), 249–265.

[8] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. 2013. Bandits with heavy tail. *IEEE Transactions on Information Theory* 59, 11 (2013), 7711–7717.

[9] Pablo Castells, Neil Hurley, and Saul Vargas. 2021. Novelty and diversity in recommender systems. In *Recommender systems handbook*. Springer, 603–646.

[10] L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. 2019. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the conference on fairness, accountability, and transparency (FAT*)*. 160–169.

[11] Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. 2020. Multi-armed bandits with fairness constraints for distributing resources to human teammates. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 299–308.

[12] Varsha Dani, Thomas P Hayes, and Sham M Kakade. 2008. Stochastic linear optimization under bandit feedback. *Conference on Learning Theory (COLT)*.

[13] Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Erik Knudsen, Helle Sjøvaag, Kristian Tolonen, Øyvind Holmstad, Igor Pipkin, Eivind Throndsen, Agnes Stenbom, et al. 2022. Towards responsible media recommendation. *AI and Ethics* (2022), 1–12.

[14] Francesco Fabbri, Maria Luisa Croci, Francesco Bonchi, and Carlos Castillo. 2022. Exposure inequality in people recommender systems: The long-term effects. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 194–204.

[15] Deen Freelon, Michael Bossetta, Chris Wells, Josephine Lukito, Yiping Xia, and Kirsten Adams. 2022. Black trolls matter: Racial and ideological asymmetries in social media disinformation. *Social Science Computer Review* 40, 3 (2022), 560–578.

[16] Nika Haghtalab, Matthew O Jackson, and Ariel D Procaccia. 2021. Belief polarization in a complex world: A learning theory perspective. *Proceedings of the National Academy of Sciences* 118, 19 (2021), e2010144118.

[17] Daniel Halpern, Ariel D Procaccia, Iyan Rahwan, Itai Shapira, and Manuel Wüthrich. 2023. Optimal Engagement-Diversity Tradeoffs in Social Media. (2023).

[18] F Maxwell Harper and Joseph A Konstan. 2015. The MovieLens Datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TIIS)* 5, 4 (2015), 1–19.

[19] Natali Helberger, Kari Karppinen, and Lucia D'acunto. 2018. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society* 21, 2 (2018), 191–207.

[20] Safwan Hossain, Evi Micha, and Nisarg Shah. 2021. Fair algorithms for multi-agent multi-armed bandits. *Conference on Neural Information Processing Systems (NeurIPS)*, 24005–24017.

[21] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 383–390.

[22] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *Conference on Neural Information Processing Systems (NeurIPS)*.

[23] David Krueger, Tegan Maharaj, and Jan Leike. 2020. Hidden incentives for auto-induced distributional shift. *arXiv preprint arXiv:2009.09153* (2020).

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

[25] mcd (https://math.stackexchange.com/users/1059789/mcd). [n. d.]. $\mathbb{E}[X \mid X \geq n/2]$ when $X \sim \mathrm{Bin}(n, 0.5)$. Mathematics Stack Exchange. arXiv:https://math.stackexchange.com/q/4616304 https://math.stackexchange.com/q/4616304 URL:https://math.stackexchange.com/q/4616304 (version: 2023-01-11).

[26] Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. 2021. Safe linear thompson sampling with side information. *IEEE Transactions on Signal Processing* 69 (2021), 3755–3767.

[27] Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. 2021. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2827–2835.

[28] Agnieszka Rychwalska and Magdalena Roszczyńska-Kurasińska. 2018. Polarization on social media: when group dynamics leads to societal divides. (2018).

[29] Aleksandrs Slivkins. 2019. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272* (2019).

[30] Ana-Andreea Stoica and Augustin Chaintreau. 2019. Hegemony in Social Media and the effect of recommendations. In *Companion Proceedings of The 2019 World Wide Web Conference*. 575–580.

[31] Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. 2018. Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity. In *Proceedings of the International World Wide Web Conference (WWW)*. 923–932.

[32] Jessica Su, Aneesh Sharma, and Sharad Goel. 2016. The effect of recommendations on network structure. In *Proceedings of the International World Wide Web Conference (WWW)*.

[33] William R Thompson. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25, 3-4 (1933), 285–294.

[34] Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. 2016. Conservative bandits. In *International Conference on Machine Learning (ICML)*.

A PROOFS ABOUT FIRST ATTEMPT (SECTION 3)

Lemma 3.1. *Suppose that there are $k = 2$ arms and for some set $N \subseteq [n]$ with $|N| \geq \frac{n}{2}$, $\boldsymbol{\mu}_i = (1, 0)$ for all $i \in N$ and $\boldsymbol{\mu}_i = (0, 1)$ for all $i \notin N$. If $\Delta < \frac{|N|}{n}$, then $\boldsymbol{p}_i^* = (1, 0)$ if $i \in N$ and $\boldsymbol{p}_i^* = \left(1 - \frac{n\Delta}{|N|}, \frac{n\Delta}{|N|}\right)$ otherwise.*

Proof. First, we write the expected reward as

$$\sum_{i=1}^{n} \boldsymbol{p}_i \cdot \boldsymbol{\mu}_i = \sum_{i \in N} p_{i,1} + \sum_{i \notin N} p_{i,2} = n - |N| + \sum_{i \in N} p_{i,1} - \sum_{i \notin N} p_{i,1},$$

so we can write our optimization problem as

$$\begin{aligned} &\text{maximize} && \textstyle\sum_{i \in N} p_{i,1} - \sum_{i \notin N} p_{i,1}, \\ &\text{subject to} && \left| p_{i,1} - \frac{1}{n} \sum_{i'=1}^{n} p_{i',0} \right| \leq \Delta. \end{aligned} \tag{15}$$

Next, we show that there exists an optimal solution such that $p_{i,1} = q_1$ for all $i \in N$ and $p_{i,1} = q_2$ for all $i \notin N$, for some $q_1, q_2 \in [0, 1]$.

Claim A.1. *An optimal solution to Equation (15) has $p_{i,1} = q_1$ for all $i \in N$ and $p_{i,1} = q_2$ for all $i \notin N$, for some $q_1, q_2 \in [0, 1]$.*

Proof of Claim A.1. First, if $N = \emptyset$, then the optimal solution is to set $p_{i,1} = 0$ for all $i \in [n]$, and if $N = [n]$, then the optimal solution is to set $p_{i,1} = 1$ for all $i \in [n]$. In both of these cases, the claim holds.

Next, suppose $N \neq \emptyset$ and $N \neq [n]$. Let $p_{1,0}, \ldots, p_{n,0}$ be an optimal solution to Equation (15) and let $q_1 = \frac{1}{|N|} \sum_{i \in N} p_{i,1}$ and $q_2 = \frac{1}{n - |N|} \sum_{i \notin N} p_{i,1}$. This is a feasible solution to Equation (15) because

$$\frac{1}{n} \left(|N| q_1 + (n - |N|) q_2 \right) = \frac{1}{n} \sum_{i=1}^{n} p_{i,1},$$

so for all $i \in N$, we have that

$$-\Delta \leq \min_{j \in N} p_{j,0} - \frac{1}{n} \sum_{i'=1}^{n} p_{i',0} \leq q_1 - \frac{1}{n} \left(|N| q_1 + (n - |N|) q_2 \right) \leq \max_{j \in N} p_{j,0} - \frac{1}{n} \sum_{i'=1}^{n} p_{i',0} \leq \Delta.$$

Similarly, for all $i \notin N$,

$$-\Delta \leq \min_{j \notin N} p_{j,0} - \frac{1}{n} \sum_{i'=1}^{n} p_{i',0} \leq q_2 - \frac{1}{n} \left(|N| q_1 + (n - |N|) q_2 \right) \leq \max_{j \notin N} p_{j,0} - \frac{1}{n} \sum_{i'=1}^{n} p_{i',0} \leq \Delta.$$

Moreover, this solution has the same objective function value as $p_{1,0}, \ldots, p_{n,0}$, so it is an optimal solution. □

Using this notation, we simplify the constraints by writing

$$q_1 - \frac{1}{n} \sum_{i=1}^{n} p_{i,1} = q_1 - \frac{1}{n} \left(|N| q_1 + (n - |N|) q_2 \right) = \frac{(n - |N|)(q_1 - q_2)}{n}.$$

Therefore, the constraint (Equation (15)) for any $i \in N$ becomes

$$|q_1 - q_2| \leq \frac{n\Delta}{n - |N|}. \tag{16}$$

Similarly, we may write

$$q_2 - \frac{1}{n}\sum_{i=1}^{n} p_{i,1} = q_2 - \frac{1}{n}\left(|N|q_1 + (n - |N|)q_2\right) = \frac{|N|(q_2 - q_1)}{n}.$$

Therefore, the constraint (Equation (15)) for any $i \notin N$ becomes

$$|q_1 - q_2| \leq \frac{n\Delta}{|N|}. \tag{17}$$

Since $|N| \geq \frac{n}{2}$, Equation (17) is tighter than Equation (16). Therefore, our optimization problem can be written as the LP

$$\begin{aligned}
&\text{maximize} && g(q_1, q_2) = |N|q_1 - (n - |N|)q_2 \\
&\text{subject to} && q_2 \geq q_1 - \frac{n\Delta}{|N|} && (18)\\
& && q_2 \leq q_1 + \frac{n\Delta}{|N|} && (19)\\
& && 0 \leq q_2, q_1 \leq 1
\end{aligned}$$

The vertices (q_1, q_2) of this LP polytope and their objective values $g(q_1, q_2)$ are

Intersection of Equations (18) and (19): Infeasible

Intersection of Equation (18) and $q_2 = 0$: $\left(q_1^{(1)}, q_2^{(1)}\right) = \left(\frac{n\Delta}{|N|}, 0\right)$

$$g\left(q_1^{(1)}, q_2^{(1)}\right) = n\Delta$$

Intersection of Equation (18) and $q_2 = 1$: Infeasible

Intersection of Equation (18) and $q_1 = 0$: Infeasible

Intersection of Equation (18) and $q_1 = 1$: $\left(q_1^{(2)}, q_2^{(2)}\right) = \left(1, 1 - \frac{n\Delta}{|N|}\right)$

$$g\left(q_1^{(2)}, q_2^{(2)}\right) = 2|N| - n + n\Delta\left(\frac{n}{|N|} - 1\right)$$

Intersection of Equation (19) and $q_2 = 0$: Infeasible

Intersection of Equation (19) and $q_2 = 1$: $\left(q_1^{(3)}, q_2^{(3)}\right) = \left(1 - \frac{n\Delta}{|N|}, 1\right)$

$$g\left(q_1^{(3)}, q_2^{(3)}\right) = 2|N| - n + n\Delta$$

Intersection of Equation (19) and $q_1 = 0$: $\left(q_1^{(4)}, q_2^{(4)}\right) = \left(0, \frac{n\Delta}{|N|}\right)$

$$g\left(q_1^{(4)}, q_2^{(4)}\right) = -\frac{n\Delta(n - |N|)}{|N|}$$

Intersection of Equation (19) and $q_1 = 1$: Infeasible.

Finally, we have vertices $\left(q_1^{(5)}, q_2^{(5)}\right) = (0,0)$ with $g\left(q_1^{(5)}, q_2^{(5)}\right) = 0$ and $\left(q_1^{(6)}, q_2^{(6)}\right) = (1,1)$ with $g\left(q_1^{(6)}, q_2^{(6)}\right) = 2|N| - n$. Since $\Delta < \frac{|N|}{n}$, $(q_1, q_2) = (0,1)$ and $(q_1, q_2) = (1,0)$ are infeasible. Since $|N| \geq \frac{n}{2}$, $\left(q_1^{(3)}, q_2^{(3)}\right) = \left(1 - \frac{n\Delta}{|N|}, 1\right)$ maximizes the objective value, which implies the lemma statement. □

B PROOFS ABOUT OUR MORE EQUITABLE APPROACH (SECTION 4)

Lemma 4.1. *Suppose that there are $k = 2$ arms and for some set $N \subseteq [n]$, $\boldsymbol{\mu}_i = (1,0)$ for all $i \in N$ and $\boldsymbol{\mu}_i = (0,1)$ for all $i \notin N$. For $\gamma \leq \frac{1}{2}$, the optimal policy has the following form:*

$$\boldsymbol{p}_i^* = \begin{cases} \left(1 - \frac{\gamma(n-|N|)}{n}, \frac{\gamma(n-|N|)}{n}\right) & \text{if } i \in N \\ \left(\frac{\gamma|N|}{n}, 1 - \frac{\gamma|N|}{n}\right) & \text{if } i \notin N. \end{cases}$$

Proof. Our goal is to find distributions $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n \in \mathcal{P}^1$ to the optimization problem

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^n \boldsymbol{p}_i \cdot \boldsymbol{\mu}_i = \sum_{i \in N} p_{i,1} + \sum_{i \notin N} p_{i,2} \\ \text{such that} & p_{i,1} \geq \frac{\gamma}{n} \sum_{i'=1}^n p_{i',1}, \forall i \in [n] \\ & p_{i,2} \geq \frac{\gamma}{n} \sum_{i'=1}^n p_{i',2}, \forall i \in [n] \end{array} \tag{20}$$

We claim that without loss of generality, we may set $p_{i,1} = q_1$ for all $i \in N$ and $p_{i,2} = q_2$ for all $i \notin N$, for some $q_1, q_2 \in [0,1]$.

Claim B.1. *An optimal solution to Equation (20) has $p_{i,1} = q_1$ for all $i \in N$ and $p_{i,2} = q_2$ for all $i \notin N$, for some $q_1, q_2 \in [0,1]$.*

Proof of Claim B.1. First, if $N = \emptyset$, then the optimal solution is to set $p_{i,2} = 1$ for all $i \in [n]$, and if $N = [n]$, then the optimal solution is to set $p_{i,1} = 1$ for all $i \in [n]$. In both of these cases, the claim holds.

Next, suppose $N \subset [n]$ and $N \neq \emptyset$. Let $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n \in \mathcal{P}^1$ be an optimal solution to Equation (20) and let $q_1 = \frac{1}{|N|} \sum_{i \in N} p_{i,1}$ and $q_2 = \frac{1}{n-|N|} \sum_{i \notin N} p_{i,2}$. This is a feasible solution to the constraints in Equation (20) because

$$\begin{aligned} q_1 &= \frac{1}{|N|} \sum_{i \in N} p_{i,1} \geq \frac{1}{|N|} \sum_{i \in N} \frac{\gamma}{n} \sum_{i'=1}^n p_{i',1} = \frac{\gamma}{n} \sum_{i=1}^n p_{i,1} = \frac{\gamma}{n} \left(\sum_{i \in N} p_{i,1} + \sum_{i \notin N} (1 - p_{i,1}) \right) \\ &= \frac{\gamma}{n} \left(|N| q_1 + (n - |N|)(1 - q_2) \right). \end{aligned}$$

Similarly,

$$\begin{aligned} q_2 &= \frac{1}{|N|} \sum_{i \in N} p_{i,2} \geq \frac{1}{|N|} \sum_{i \in N} \frac{\gamma}{n} \sum_{i'=1}^n p_{i',2} = \frac{\gamma}{n} \sum_{i=1}^n p_{i,2} = \frac{\gamma}{n} \left(\sum_{i \in N} (1 - p_{i,1}) + \sum_{i \notin N} p_{i,1} \right) \\ &= \frac{\gamma}{n} \left(|N| (1 - q_1) + (n - |N|) q_2 \right). \end{aligned}$$

Moreover, the objective functions are the same because

$$|N| q_1 + (n - |N|) q_2 = \sum_{i \in N} p_{i,1} + \sum_{i \notin N} p_{i,2}.$$

Therefore, the claim holds. □

Based on Claim B.1, we may write our optimization problem as

$$\text{maximize} \quad |N| q_1 + (n - |N|) q_2$$

$$\begin{aligned}
\text{such that} \quad & q_1 \geq \frac{\gamma}{n}\left(|N|q_1 + (n-|N|)\,(1-q_2)\right) \\
& 1-q_1 \geq \frac{\gamma}{n}\left(|N|\,(1-q_1) + (n-|N|)q_2\right) \\
& q_2 \geq \frac{\gamma}{n}\left(|N|\,(1-q_1) + (n-|N|)q_2\right) \\
& 1-q_2 \geq \frac{\gamma}{n}\left(|N|q_1 + (n-|N|)\,(1-q_2)\right) \\
& q_1, q_2 \in [0,1].
\end{aligned}$$

Rearranging terms, our optimization problem is

$$\begin{aligned}
\text{maximize} \quad & g(q_1,q_2) = |N|q_1 + (n-|N|)q_2 \\
\text{such that} \quad & q_1 \geq \frac{\gamma(n-|N|)}{n-\gamma|N|}(1-q_2) && (21)\\
& q_1 \leq 1 - \frac{\gamma(n-|N|)}{n-\gamma|N|}\cdot q_2 && (22)\\
& q_1 \geq 1 - \frac{n-\gamma(n-|N|)}{n-\gamma|N|}\cdot q_2 && (23)\\
& q_1 \leq \frac{n-\gamma(n-|N|)}{\gamma|N|}(1-q_2) && (24)\\
& q_1, q_2 \geq 0 \text{ and } q_1, q_2 \leq 1. && (25)
\end{aligned}$$

To analyze the corners of this LP polytope, we identify where the eight hyperplanes in Equations (21)-(25) intersect. Note that Equations (21) and (22) are parallel, so they don't intersect. The same is true of Equations (23) and (24). Moreover, $q_1 = 0$ if and only if $q_2 = 1$ by Equations (21) and (24), so we ignore the intersections with $q_1 = 0$ and $q_2 = 1$. This leads to the following corners (q_1, q_2) with objective values $g(q_1, q_2)$:

$$\begin{aligned}
&\text{Intersection of (21) and (23):} && \left(q_1^{(1)}, q_2^{(1)}\right) = \left(\frac{\gamma(n-|N|)}{n}, \frac{\gamma|N|}{n}\right) \\
& && g\left(q_1^{(1)}, q_2^{(1)}\right) = \frac{2\gamma|N|(n-|N|)}{n} \\
&\text{Intersection of (21) and (24):} && \left(q_1^{(2)}, q_2^{(2)}\right) = (0,1) \\
& && g\left(q_1^{(2)}, q_2^{(2)}\right) = n-|N| \\
&\text{Intersection of (21) and } q_2 = 0\text{:} && \left(q_1^{(3)}, q_2^{(3)}\right) = \left(\frac{\gamma(n-|N|)}{n-\gamma|N|}, 0\right) \\
& && g\left(q_1^{(3)}, q_2^{(3)}\right) = \frac{\gamma(n-|N|)|N|}{n-\gamma|N|} \\
&\text{Intersection of (21) and } q_1 = 1\text{: Not feasible} && \\
&\text{Intersection of (22) and (23):} && \left(q_1^{(4)}, q_2^{(4)}\right) = (1,0) \qquad (26)\\
& && g\left(q_1^{(4)}, q_2^{(4)}\right) = |N| \\
&\text{Intersection of (22) and (24):} && \left(q_1^{(5)}, q_2^{(5)}\right) = \left(1-\gamma\left(1-\frac{|N|}{n}\right), 1-\frac{\gamma|N|}{n}\right)
\end{aligned}$$

Algorithm 1 Multi-agent UCB (defined by parameter δ)

Require: Failure probability $\delta \in (0,1)$

Set $N_{i,j}(0) = 0,\ \forall i \in [n], j \in [k]$; $\hat{\boldsymbol{\mu}}_i^{(0)} = \mathbf{0},\ \forall i \in [n]$
for $t \in \{1, \ldots, T\}$ **do**
  if $t \in \{1, \ldots, k\}$ **then**
    Set $\boldsymbol{p}_i^{(t)} = \boldsymbol{e}_t$
  else
    Set $\left(\boldsymbol{p}_i^{(t)}\right)_{i\in[n]} = \underset{(\boldsymbol{p}_i)_{i\in[n]}\in S}{\operatorname{argmax}} \sum_{i=1}^{n} \boldsymbol{p}_i \cdot \hat{\boldsymbol{\mu}}_i^{(t-1)}$
  end if
  Draw an arm $j_i^{(t)} \sim \boldsymbol{p}_i^{(t)}\ \forall i \in [n]$
  Receive reward $r_i^{(t)} \sim \mathcal{D}_{i,j_i^{(t)}}$
  For all $i \in [n]$, set $N_{i,j_i^{(t)}}(t) = N_{i,j_i^{(t)}}(t-1) + 1$ ▷ Increment the counter for arm $j_i^{(t)}$
  Set $N_{i,j}(t) = N_{i,j}(t-1),\ \forall i \in [n]$ and $j \neq j_i^{(t)}$ ▷ Do not increment the other counters
  Set $\beta_{i,j}^{(t)} = \sqrt{\frac{1}{N_{i,j}(t)} \log \frac{2Tnk}{\delta}},\ \forall i \in [n], j \in [k]$ ▷ Define confidence intervals
  $\hat{\mu}_{i,j}^{(t)} = \frac{1}{N_{i,j}(t)} \sum_{\tau=1}^{t} r_i^{(\tau)} \mathbb{1}\left\{j_i^{(\tau)} = j\right\} + \beta_{i,j}^{(t)},\ \forall i \in [n], j \in [k]$ ▷ Estimate mean rewards
end for

$$g\left(q_1^{(5)}, q_2^{(5)}\right) = n - \frac{2\gamma|N|(n-|N|)}{n}$$

Intersection of (22) and $q_2 = 0$: $(q_1, q_2) = (1, 0)$ as in Equation (26)

Intersection of (22) and $q_1 = 1$: $(q_1, q_2) = (1, 0)$ as in Equation (26)

Intersection of (23) and $q_2 = 0$: $(q_1, q_2) = (1, 0)$ as in Equation (26)

Intersection of (23) and $q_1 = 1$: $(q_1, q_2) = (1, 0)$ as in Equation (26)

Intersection of (24) and $q_2 = 0$: Not feasible.

Intersection of (24) and $q_1 = 1$: Not feasible.

For $\gamma \leq \frac{1}{2}$, the optimum is achieved at $\left(q_1^{(5)}, q_2^{(5)}\right) = \left(1 - \gamma\left(1 - \frac{|N|}{n}\right), 1 - \frac{\gamma|N|}{n}\right)$, so the lemma holds. □

C PROOFS ABOUT THE FORMULATION 1 REGRET UPPER BOUND WHEN $\gamma < 1$ (SECTION 5.1.1)

Claim C.1. *With probability* $1-\delta$, *for all* $i \in [n], t \in [T]$ *and* $j \in [k]$,

$$\hat{\boldsymbol{\mu}}_{i,j}^{(t)} \geq \boldsymbol{\mu}_{i,j} \geq \hat{\boldsymbol{\mu}}_{i,j}^{(t)} - 2\boldsymbol{\beta}_{u,i}^{(t)}$$

Proof. Consider a fixed iteration t. For $i \in [n]$ and $j \in [k]$ and $\ell \in [t]$, let $\tau = \inf_s\{N_{i,j}(s) \geq \ell\}$ and $\hat{v}_{i,j}^{\ell} = \frac{1}{\ell} \sum_{s=1}^{\tau} r_{i,j}^{(s)} \mathbb{1}\left\{j_i^{(s)} = j\right\}$. Since the $\hat{v}_{i,j}^{\ell}$ are independent in ℓ, i, and j applying Hoeffding's inequality with parameter $\delta' = \delta/(tnk)$, with probability greater than $1-\delta'$,

$$|\hat{v}_{i,j}^{\ell} - \mu_{i,j}| \leq \sqrt{\frac{\log(2/\delta')}{\ell}}$$

Applying the union bound for all $\forall \ell \in [t], \forall i \in [n], \forall j \in [k]$ we have that:

$$\mathbb{P}\left(\exists \ell \in [t], \exists i \in [n], \exists j \in [k] : |\hat{v}_{i,j}^{\ell} - \mu_{i,j}| \geq \sqrt{\frac{\log(2/\delta')}{\ell}}\right) \leq \sum_{\ell=1}^{t}\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbb{P}\left(|\hat{v}_{i,j}^{\ell} - \mu_{i,j}| \geq \sqrt{\frac{\log(2/\delta')}{N_{i,j}(t)}}\right)$$
$$\leq tnk\delta' = \delta$$

Thus with probability at least $1-\delta$, $\forall \ell \in [t], \forall i \in [n], \forall j \in [k]$,

$$|\hat{v}_{i,j}^{\ell} - \mu_{i,j}| \leq \sqrt{\frac{\log(2tnk/\delta)}{\ell}}.$$

Since $\hat{\mu}_{i,j}^{t} = \hat{v}_{i,j}^{N_{i,j}(t)} + \beta_{i,j}^{(t)}$ and $N_{i,j}(t) \in [t]$ $\forall i, j$ with probability at least $1-\delta$

$$\hat{\mu}_{i,j}^{t} - 2\beta_{i,j}^{(t)} \leq \mu_{i,j} \leq \hat{\mu}_{i,j}^{t}.$$

□

Theorem 5.1. *Let π be the policy of n-UCB. Then $R_{T,1}(\pi, \nu) = \tilde{O}(n\sqrt{kT})$.*

Proof. Fix a timestep $t \in [T]$

$$\sum_{i\in[n]}(\boldsymbol{p}_i^* \cdot \boldsymbol{\mu}_i - \boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\mu}_i) = \sum_{i\in[n]}(\boldsymbol{p}_i^* \cdot \boldsymbol{\mu}_i - \boldsymbol{p}_i^{(t)} \cdot \hat{\boldsymbol{\mu}}_i^{(t)} + \boldsymbol{p}_i^{(t)} \cdot \hat{\boldsymbol{\mu}}_i^{(t)} - \boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\mu}_i)$$
$$\leq \sum_{i\in[n]}(\boldsymbol{p}_i^* \cdot \boldsymbol{\mu}_i - \boldsymbol{p}_i^* \cdot \hat{\boldsymbol{\mu}}_i^{(t)} + \boldsymbol{p}_i^{(t)} \cdot \hat{\boldsymbol{\mu}}_i^{(t)} - \boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\mu}_i)$$

By Claim C.1, $\boldsymbol{p} \cdot \boldsymbol{\mu}_i \leq \boldsymbol{p} \cdot \hat{\boldsymbol{\mu}}_i^{(t)}$ $\forall i \in [n]$ and all $\boldsymbol{p} \in \mathbb{R}_{\geq 0}^{k}$

$$\sum_{i\in[n]}(\boldsymbol{p}_i^* \cdot \boldsymbol{\mu}_i - \boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\mu}_i) \leq \sum_{i\in[n]}(\boldsymbol{p}_i^* \cdot \hat{\boldsymbol{\mu}}_i^{(t)} - \boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\mu}_i)$$
$$\leq \sum_{i\in[n]}\boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\beta}_i^{(t)}$$

Thus

$$R_T \leq \sum_{t=1}^{T}\sum_{i\in[n]}\boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\beta}_i^{(t)}$$

Let $\mathcal{F}_{i,t-1}$ denote the canonical filtration $\sigma((X_{i,s}, \boldsymbol{p}_i^{(s)}) : 0 \leq s < t)$ on the choice of $\boldsymbol{p}_i^{(t)}$ and let $j_i^{(t)'}$ be a random variable distributed as $\boldsymbol{p}_i^{(t)} \mid \mathcal{F}_{i,t-1}$ and conditionally independent from $j_{i,t}^{(t)}$, i.e. $j_i^{(t)'} \perp j_i^{(t)} \mid \mathcal{F}_{i,t-1}$. Note that by definition the following equality holds:

$$\mathop{\mathbb{E}}_{j_i^{(t)} \sim \boldsymbol{p}_i^{(t)}}[\beta_{i,j_i^{(t)}}^{(t)}] = \mathop{\mathbb{E}}_{j_i^{(t)'} \sim \boldsymbol{p}_i^{(t)}}[\beta_{i,j_i^{(t)'}}^{(t)'} \mid \mathcal{F}_{i,t-1}].$$

Consider the following random variables $A_{i,t} = \mathbb{E}_{j_i^{(t)'} \sim \boldsymbol{p}_i^{(t)}}[\beta_{i,j_i^{(t)'}} \mid \mathcal{F}_{i,t-1}] - \beta_{i,j_i^{(t)}}(t)$. Note that $M_{i,t} = \sum_{s=1}^{t} A_{i,s}$ is a martingale. Since $|A_t| \leq 2\sqrt{2\log(Tnk/\delta)}$, using this as the bound in Azuma-Hoeffding and taking a union bound over $i \in [n]$ and $t \in T$ implies that with probability at least $1-\delta$,

$$R_T = \sum_{t=1}^{T}\sum_{i=1}^{n}\boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\beta}_i^{(t)} \leq \sum_{t=1}^{T}\sum_{i=1}^{n}\boldsymbol{\beta}_{i,j_i^{(t)}}^{(t)} + nT\sqrt{\frac{1}{T}\log\left(\frac{Tnk}{\delta}\right)\log\left(\frac{1}{\delta}\right)}$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{n} \boldsymbol{\beta}^{(t)}_{i,j_i^{(t)}} + n\sqrt{T\log\left(\frac{Tnk}{\delta}\right)\log\left(\frac{1}{\delta}\right)}$$

Now we bound $\sum_{i=1}^{n}\sum_{t=1}^{T} \boldsymbol{\beta}^{(t)}_{i,j_i^{(t)}}$,

$$\sum_{t=1}^{T}\sum_{i=1}^{n} \boldsymbol{\beta}^{(t)}_{i,j_i^{(t)}} = \sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{j=1}^{k} \beta^{(t)}_{i,j}\mathbb{1}\left\{j_i^{(t)} = j\right\}$$

For fixed i, j

$$\sum_{t=1}^{T} \beta^{(t)}_{i,j}\mathbb{1}\left\{j_i^{(t)} = j\right\} = \sqrt{\log(Tnk/\delta)}\sum_{t=1}^{N_{i,j}(T)} 1/\sqrt{t} \le 2\sqrt{N_{i,j}(T)\log(Tnk/\delta)}$$

Therefore

$$\begin{aligned}
\sum_{i\in[n]}\sum_{t=1}^{T} \boldsymbol{\beta}_i^{(t)} &\le 2\sum_{i=1}^{n}\sum_{j=1}^{k}\sqrt{N_{i,j}(T)\log(Tnk/\delta)} \\
&\le 2\sum_{i=1}^{n}\sqrt{k\sum_{j=1}^{k} N_{i,j}(T)\log(Tnk/\delta)} \\
&= 2\sum_{i=1}^{n}\sqrt{kT\log(Tnk/\delta)} \\
&= 2n\sqrt{kT\log(Tnk/\delta)}
\end{aligned}$$

Where the second line follows from the concavity of $\sqrt{\cdot}$ and the penultimate line follows from the fact that $\sum_{j=1}^{k} N_{i,j}(T) = T$.

The result then follows by setting $\delta = \frac{1}{nT}$. □

D PROOFS ABOUT THE FORMULATION 1 REGRET UPPER BOUND WHEN $\gamma = 1$ (SECTION 5.1.2)

In this section, the distribution $\mathcal{D}_j = \sum_{i=1}^{n} \mathcal{D}_{i,j}$ is supported on $[0, n]$ instead of $[0, 1]$ and we denote the average reward for arm $j \in [k]$ as $\mu_j = \sum_{i=1}^{n} \mu_{i,j}$.

Definition D.1 (median-of-means estimator [2]). Let $\delta \in (0, 1)$ and $X_1, \ldots, X_T$ be i.i.d random variables with mean $\mathbb{E}[X] = \mu$ and variance $\mathbb{E}|X - \mu|^2 = \sigma^2$. Let $m = \lfloor 8\log(1/\delta) \wedge T/2 \rfloor$ and $t = \lfloor T/m \rfloor$. Let $\bar{\mu}^1, \ldots, \bar{\mu}^m$ be m empirical mean estimates, each one calculated on t data points as follows:

$$\bar{\mu}^1 = \frac{1}{t}\sum_{s=1}^{t} X_s\ ,\ \bar{\mu}^2 = \frac{1}{t}\sum_{s=t+1}^{2t} X_s\ ,\ldots,\ \bar{\mu}^m = \frac{1}{t}\sum_{s=(m-1)t+1}^{mt} X_s.$$

The *median-of-means estimator* $\bar{\boldsymbol{\mu}}(T, \delta)$ is the median of these m empirical means.

Theorem 5.2. *Let π be the policy of Robust-UCB. Then $R_{T,1}(\pi, \nu) = \tilde{O}(\sqrt{nkT})$.*

Proof. For all $t \in [T]$, $j \in [k]$ the median-of-mean estimate at time t, $\hat{\mu}_j^{(t)}$ with probability at least $1 - \delta$ we have

$$|\hat{\mu}_j^{(t)} - \mu_j| \le \sqrt{\frac{24n\log(kT/\delta)}{T}}$$

Algorithm 2 Robust-UCB (defined by parameter δ)

Require: Failure probability $\delta \in (0,1)$, median-of-means estimator $\bar{\boldsymbol{\mu}}(t,\delta)$

Set $N_j(0) = 0$, $\hat{\mu}_j^{(0)} = 0 \ \ \forall j \in [k]$
for $t \in \{1,\dots,T\}$ **do**
  if $t \in \{1,\dots,k\}$ **then**
    Set $\boldsymbol{p}^{(t)} = \boldsymbol{e}_t$
  else
    Set $\boldsymbol{p}^{(t)} = \underset{\boldsymbol{p}\in\mathcal{P}^{k-1}}{\operatorname{argmax}}\, \boldsymbol{p}\cdot\hat{\boldsymbol{\mu}}^{(t-1)}$
  end if
  Draw an arm $j^{(t)} \sim \boldsymbol{p}^{(t)}$
  Receive reward $r^{(t)} \sim \mathcal{D}_{j^{(t)}}$
  Set $N_{j^{(t)}}(t) = N_{j^{(t)}}(t-1) + 1$ ▷ Increment the counter for arm $j^{(t)}$
  Set $N_j(t) = N_j(t-1),\ \forall j \neq j^{(t)}$ ▷ Do not increment the other counters
  Set $\beta_j^{(t)} = \sqrt{\frac{24n}{N_j(t)}\log\frac{Tk}{\delta}},\ \forall j \in [k]$ ▷ Define confidence intervals
  $\hat{\mu}_j^{(t)} = \bar{\boldsymbol{\mu}}_j(N_j(t),\delta) + \beta_j^{(t)},\ \forall j \in [k]$ ▷ Get mean rewards estimates
end for

By applying Lemma D.2 with $\sigma_j^2 = n/4$ which is justified by ClaimD.3 and then taking a union bound over all arms $j \in [k]$.

This event and Proposition 1 in Bubeck et al. [8] imply the desired regret upper bounds for the Robust-UCB with median-of-means estimator. The result then follows by setting $\delta = \frac{1}{nT}$. □

Lemma D.2. *Let $\delta \in (0,1)$. Let $X_{j,1},\dots,X_{j,T}$ be i.i.d random variables with mean $\mathbb{E}[X_j] = \mu_j$ and $\mathbb{E}\,|X_j - \mu_j|^2 = \sigma_j^2$. Let $m = \lfloor 8\log(1/\delta) \wedge T/2 \rfloor$ and $t = \lfloor T/m \rfloor$. Let*

$$\bar{\mu}_j^1 = \frac{1}{t}\sum_{s=1}^{t} X_{j,s}\ ,\ \bar{\mu}_j^2 = \frac{1}{t}\sum_{s=t+1}^{2t} X_{j,s}\ ,\dots,\ \bar{\mu}_j^m = \frac{1}{t}\sum_{s=(m-1)t+1}^{kt} X_{j,s},$$

be m empirical mean estimates, each one computed on t data points. Let $\hat{\mu}_j$ be the median of these m empirical means. Then with probability at least $1-\delta$

$$|\hat{\mu}_j - \mu| \le \sigma\sqrt{\frac{96\log(1/\delta)}{T}}.$$

Proof. By Chebyshev's inequality $\forall \ell \in [m]$

$$\mathbb{P}[|\bar{\mu}_j^\ell - \mu_j| \le \sigma\sqrt{12/t}] \ge 3/4$$

Let $\epsilon > 0$ and $Y_\ell = \mathbb{1}\left\{|\hat{\mu}_j^\ell - \mu_j| > \epsilon\right\}$ for $\ell \in [k]$, For $\epsilon = \sigma\sqrt{12/t}$, Y_ℓ is stochastically dominated by a Bernoulli distribution with parameter $p = 1/4$. Thus using Hoeffding's inequality for the tail of a binomial distribution we get

$$\mathbb{P}(|\hat{\mu}_j - \mu_j| > \epsilon) = \mathbb{P}\left(\sum_{\ell=1}^{m} Y_\ell \ge m/2\right) \le \mathbb{P}(Bin(m,1/4) \ge m/2) \le \exp(-2m(p-1/2)^2) = \exp(-m/8) = \delta$$

□

Claim D.3. $\sigma_j^2 = \mathbb{E}[|X_j - \mu_j|^2] \le n/4\ \forall j \in [k]$

Proof.

$$\begin{aligned}\mathbb{E}[|X_j - \mu_j|^2] &= \mathbb{E}\left[\left|\sum_{i=1}^{n} X_{i,j} - \mu_{i,j}\right|^2\right] \\ &\leq \sum_{i=1}^{n} \mathbb{E}[|X_{i,j} - \mu_{i,j}|^2] \\ &\leq \sum_{i=1}^{n} 1/4 = n/4\end{aligned}$$

where the second line follows from the independence across users i and triangle inequality. While the last line follows from Popoviciu's variance inequality. □

E PROOFS ABOUT THE FORMULATION 1 REGRET LOWER BOUNDS (SECTION 5.1.3)

Lemma 5.5. *For all $T \geq 1$, the regret is lower bounded as follows:*

$$\inf_{\pi \in \Pi^{n,2}} \sup_{\nu \in \mathcal{E}^{n,2}} R_{T,1}(\pi, \nu) \geq \sqrt{\frac{T}{8}}\left(\frac{n}{8e} - \gamma\left(\frac{n}{8e} + \sqrt{\frac{n}{2\pi}}\right)\right). \tag{9}$$

Proof. Our proof is based on worst-case instances $\nu_{\boldsymbol{b}}$ defined for any vector $\boldsymbol{b} \in \{0,1\}^n$. For each user $i \in [n]$, their reward distributions for the two arms are Bernoulli with means $\boldsymbol{\mu}_i = (\mu_{i,0}, \mu_{i,1})$ where

$$\boldsymbol{\mu}_i = \begin{cases} \left(\frac{1}{2} + \epsilon, \frac{1}{2}\right) & \text{if } b_i = 0 \\ \left(\frac{1}{2}, \frac{1}{2} + \epsilon\right) & \text{if } b_i = 1 \end{cases} \tag{27}$$

where $\epsilon = \sqrt{\frac{1}{8T}}$. We will lower bound the expected regret $\mathbb{E}_{\boldsymbol{b}}\left[R_T\left(\pi, \nu_{\boldsymbol{b}}\right)\right]$ over both the randomness of the draw of the vector $\boldsymbol{b} \sim \text{Unif}\left(\{0,1\}^n\right)$ and the distribution over outcomes $\mathbb{P}_{\pi\nu_{\boldsymbol{b}}}$. This will imply that for any policy π, there exists an instance $\nu_{\boldsymbol{b}}$ such that

$$R_T\left(\pi, \nu_{\boldsymbol{b}}\right) \geq \sqrt{\frac{T}{8}}\left(\frac{n}{8e} - \gamma\left(\frac{n}{8e} + \sqrt{\frac{n}{2\pi}}\right)\right).$$

Given an instance $\nu_{\boldsymbol{b}}$, the following distributions $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ with

$$\boldsymbol{p}_i = \begin{cases} \left(1 - \frac{\gamma\|\boldsymbol{b}\|_1}{n}, \frac{\gamma\|\boldsymbol{b}\|_1}{n}\right) & \text{if } b_i = 0 \\ \left(\frac{\gamma(n-\|\boldsymbol{b}\|_1)}{n}, 1 - \frac{\gamma(n-\|\boldsymbol{b}\|_1)}{n}\right) & \text{if } b_i = 1. \end{cases}$$

are feasible policies. This is because $n - \|\boldsymbol{b}\|_1$ is the number of 0's in $\boldsymbol{b}$ and $\|\boldsymbol{b}\|_1$ is the number of 1's in $\boldsymbol{b}$, so for any i such that $b_i = 0$,

$$1 - \frac{\gamma\|\boldsymbol{b}\|_1}{n} \geq \frac{\gamma}{n}\left(\sum_{i:b_i=0}\left(1 - \frac{\gamma\|\boldsymbol{b}\|_1}{n}\right) + \sum_{i:b_i=1}\frac{\gamma\left(n - \|\boldsymbol{b}\|_1\right)}{n}\right) = \gamma - \frac{\gamma\|\boldsymbol{b}\|_1}{n}$$

and

$$\frac{\gamma\|\boldsymbol{b}\|_1}{n} = \frac{\gamma}{n}\left(\sum_{i:b_i=0}\frac{\gamma\|\boldsymbol{b}\|_1}{n} + \sum_{i:b_i=1}\left(1 - \frac{\gamma\left(n - \|\boldsymbol{b}\|_1\right)}{n}\right)\right).$$

Similarly, for any i such that $b_i = 1$,

$$\frac{\gamma(n - \|\boldsymbol{b}\|_1)}{n} = \frac{\gamma}{n}\left(\sum_{i:b_i=0}\left(1 - \frac{\gamma\,\|\boldsymbol{b}\|_1}{n}\right) + \sum_{i:b_i=1}\frac{\gamma\,(n - \|\boldsymbol{b}\|_1)}{n}\right)$$

and

$$1 - \frac{\gamma(n - \|\boldsymbol{b}\|_1)}{n} = 1 - \gamma + \frac{\gamma\,\|\boldsymbol{b}\|_1}{n} \geq \frac{\gamma}{n}\left(\sum_{i:b_i=0}\frac{\gamma\,\|\boldsymbol{b}\|_1}{n} + \sum_{i:b_i=1}\left(1 - \frac{\gamma\,(n - \|\boldsymbol{b}\|_1)}{n}\right)\right) = \frac{\gamma\,\|\boldsymbol{b}\|_1}{n}.$$

After simplifying, this policy has an objective value of

$$\begin{aligned}
&\sum_{i:b_i=0}\left(\frac{1}{2} + \left(1 - \frac{\gamma\,\|\boldsymbol{b}\|_1}{n}\right)\epsilon\right) + \sum_{i:b_i=1}\left(\frac{1}{2} + \left(1 - \frac{\gamma(n - \|\boldsymbol{b}\|_1)}{n}\right)\epsilon\right)\\
&= \frac{n}{2} + (n - \|\boldsymbol{b}\|_1)\left(1 - \frac{\gamma\,\|\boldsymbol{b}\|_1}{n}\right)\epsilon + \|\boldsymbol{b}\|_1\left(1 - \frac{\gamma(n - \|\boldsymbol{b}\|_1)}{n}\right)\epsilon\\
&= \frac{n}{2} + \epsilon\left(n - \frac{1}{n}\cdot 2\gamma\,\|\boldsymbol{b}\|_1\,(n - \|\boldsymbol{b}\|_1)\right).
\end{aligned}$$

Thus, the optimal policy's expected cumulative reward is at least

$$\frac{nT}{2} + \epsilon\left(nT - \frac{1}{n}\cdot 2T\gamma\,\|\boldsymbol{b}\|_1\,(n - \|\boldsymbol{b}\|_1)\right). \tag{28}$$

Meanwhile, for any policy π, let $\pi_{i,0}^{(t)} = \pi_i\,(0 \mid \boldsymbol{h}_{t-1})$ denote the probability that the policy chooses arm 0 for user i on round t given the history $\boldsymbol{h}_{t-1}$. The value $\pi_{i,0}^{(t)}$ is therefore a random variable that depends on the history $\boldsymbol{h}_{t-1}$. Similarly, let $\pi_{i,1}^{(t)} = \pi_i\,(1 \mid \boldsymbol{h}_{t-1})$. The expected cumulative reward of policy π is

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\pi\nu_{\boldsymbol{b}}}\left[\sum_{t=1}^{T}\left(\sum_{i:b_i=0}\left(\left(\frac{1}{2}+\epsilon\right)\pi_{i,0}^{(t)} + \frac{1}{2}\pi_{i,1}^{(t)}\right) + \sum_{i:b_i=1}\left(\frac{1}{2}\pi_{i,0}^{(t)} + \left(\frac{1}{2}+\epsilon\right)\pi_{i,1}^{(t)}\right)\right)\right]\\
&= \frac{nT}{2} + \epsilon\left(\sum_{t=1}^{T}\left(\sum_{i:b_i=0}\mathop{\mathbb{E}}_{\pi\nu_{\boldsymbol{b}}}\left[\pi_{i,0}^{(t)}\right] + \sum_{i:b_i=1}\mathop{\mathbb{E}}_{\pi\nu_{\boldsymbol{b}}}\left[\pi_{i,1}^{(t)}\right]\right)\right)\\
&= \frac{nT}{2} + \epsilon\left(nT - \sum_{t=1}^{T}\left(\sum_{i:b_i=0}\mathop{\mathbb{E}}_{\pi\nu_{\boldsymbol{b}}}\left[\pi_{i,1}^{(t)}\right] + \sum_{i:b_i=1}\mathop{\mathbb{E}}_{\pi\nu_{\boldsymbol{b}}}\left[\pi_{i,0}^{(t)}\right]\right)\right).
\end{aligned} \tag{29}$$

Therefore, the expected regret of π is at least Equation (28) minus Equation (29), which is

$$\epsilon\left(\sum_{t=1}^{T}\left(\sum_{i:b_i=0}\mathop{\mathbb{E}}_{\pi\nu_{\boldsymbol{b}}}\left[\pi_{i,1}^{(t)}\right] + \sum_{i:b_i=1}\mathop{\mathbb{E}}_{\pi\nu_{\boldsymbol{b}}}\left[\pi_{i,0}^{(t)}\right]\right) - \frac{1}{n}\cdot 2T\gamma\,\|\boldsymbol{b}\|_1\,(n - \|\boldsymbol{b}\|_1)\right). \tag{30}$$

We will begin incorporating the constraints by rewriting the first part of Equation (30) as

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\sum_{i:b_i=0}\mathop{\mathbb{E}}_{\pi\nu_{\boldsymbol{b}}}\left[\pi_{i,1}^{(t)}\right] + \sum_{i:b_i=1}\mathop{\mathbb{E}}_{\pi\nu_{\boldsymbol{b}}}\left[\pi_{i,0}^{(t)}\right]\right)\\
&= \sum_{t=1}^{T}\left(\sum_{i:b_i=0}\mathop{\mathbb{E}}_{\pi\nu_{\boldsymbol{b}}}\left[\pi_{i,1}^{(t)} - \frac{\gamma}{n}\sum_{j=1}^{n}\pi_{j,1}^{(t)}\right] + \sum_{i:b_i=1}\mathop{\mathbb{E}}_{\pi\nu_{\boldsymbol{b}}}\left[\pi_{i,0}^{(t)} - \frac{\gamma}{n}\sum_{j=1}^{n}\pi_{j,0}^{(t)}\right]\right)
\end{aligned} \tag{31}$$

$$+\sum_{t=1}^{T}\left(\frac{\gamma\left(n-\|\boldsymbol{b}\|_1\right)}{n}\sum_{j=1}^{n}\mathop{\mathbb{E}}_{\pi\sim\nu_{\boldsymbol{b}}}\left[\pi_{j,1}^{(t)}\right]+\frac{\gamma\|\boldsymbol{b}\|_1}{n}\sum_{j=1}^{n}\mathop{\mathbb{E}}_{\pi\sim\nu_{\boldsymbol{b}}}\left[\pi_{j,0}^{(t)}\right]\right).$$

If $\|\boldsymbol{b}\|_1 < \frac{n}{2}$, then $\frac{\gamma(n-\|\boldsymbol{b}\|_1)}{n} > \frac{\gamma\|\boldsymbol{b}\|_1}{n}$, so

$$\sum_{t=1}^{T}\left(\frac{\gamma\left(n-\|\boldsymbol{b}\|_1\right)}{n}\sum_{j=1}^{n}\mathop{\mathbb{E}}_{\pi\sim\nu_{\boldsymbol{b}}}\left[\pi_{j,1}^{(t)}\right]+\frac{\gamma\|\boldsymbol{b}\|_1}{n}\sum_{j=1}^{n}\mathop{\mathbb{E}}_{\pi\sim\nu_{\boldsymbol{b}}}\left[\pi_{j,0}^{(t)}\right]\right)\geq\frac{\gamma\|\boldsymbol{b}\|_1}{n}\mathop{\mathbb{E}}_{\pi\sim\nu_{\boldsymbol{b}}}\left[\sum_{t=1}^{T}\sum_{j=1}^{n}\left(\pi_{j,1}^{(t)}+\pi_{j,0}^{(t)}\right)\right]=\gamma\|\boldsymbol{b}\|_1 T.$$

Similarly, if $\|\boldsymbol{b}\|_1 > \frac{n}{2}$,

$$\begin{aligned}\sum_{t=1}^{T}\left(\frac{\gamma\left(n-\|\boldsymbol{b}\|_1\right)}{n}\sum_{j=1}^{n}\mathop{\mathbb{E}}_{\pi\sim\nu_{\boldsymbol{b}}}\left[\pi_{j,1}^{(t)}\right]+\frac{\gamma\|\boldsymbol{b}\|_1}{n}\sum_{j=1}^{n}\mathop{\mathbb{E}}_{\pi\sim\nu_{\boldsymbol{b}}}\left[\pi_{j,0}^{(t)}\right]\right)&\geq\frac{\gamma\left(n-\|\boldsymbol{b}\|_1\right)}{n}\mathop{\mathbb{E}}_{\pi\sim\nu_{\boldsymbol{b}}}\left[\sum_{t=1}^{T}\sum_{j=1}^{n}\left(\pi_{j,1}^{(t)}+\pi_{j,0}^{(t)}\right)\right]\\&=\gamma\left(n-\|\boldsymbol{b}\|_1\right)T.\end{aligned}$$

Therefore,

$$\mathop{\mathbb{E}}_{\boldsymbol{b}\sim\{0,1\}^n}\left[\sum_{t=1}^{T}\left(\frac{\gamma\left(n-\|\boldsymbol{b}\|_1\right)}{n}\sum_{j=1}^{n}\mathop{\mathbb{E}}_{\pi\sim\nu_{\boldsymbol{b}}}\left[\pi_{j,1}^{(t)}\right]+\frac{\gamma\|\boldsymbol{b}\|_1}{n}\sum_{j=1}^{n}\mathop{\mathbb{E}}_{\pi\sim\nu_{\boldsymbol{b}}}\left[\pi_{j,0}^{(t)}\right]\right)\right] \tag{32}$$

$$\begin{aligned}&\geq\gamma T\mathop{\mathbb{E}}_{\boldsymbol{b}}\left[\|\boldsymbol{b}\|_1\mathbf{1}_{\{\|\boldsymbol{b}\|_1<\frac{n}{2}\}}+\left(n-\|\boldsymbol{b}\|_1\right)\mathbf{1}_{\{\|\boldsymbol{b}\|_1>\frac{n}{2}\}}\right]\\&=\gamma T\mathop{\mathbb{E}}_{\boldsymbol{b}}\left[\|\boldsymbol{b}\|_1\left(1-\mathbf{1}_{\{\|\boldsymbol{b}\|_1>\frac{n}{2}\}}\right)+\left(n-\|\boldsymbol{b}\|_1\right)\mathbf{1}_{\{\|\boldsymbol{b}\|_1>\frac{n}{2}\}}\right]\\&=\gamma T\left(\mathop{\mathbb{E}}_{\boldsymbol{b}}\left[\|\boldsymbol{b}\|_1\right]+n\,\mathbb{E}\left[\mathbf{1}_{\{\|\boldsymbol{b}\|_1>\frac{n}{2}\}}\right]-2\,\mathbb{E}\left[\|\boldsymbol{b}\|_1\mathbf{1}_{\{\|\boldsymbol{b}\|_1>\frac{n}{2}\}}\right]\right).\end{aligned} \tag{33}$$

When $\boldsymbol{b}\sim\mathrm{Unif}\left(\{0,1\}^n\right)$, $\|\boldsymbol{b}\|_1\sim\mathrm{Bin}\left(n,\frac{1}{2}\right)$. Therefore, Equation (33) is equal to

$$\gamma T\left(n-2\,\mathbb{E}\left[\|\boldsymbol{b}\|_1\,\middle|\,\|\boldsymbol{b}\|_1>\frac{n}{2}\right]\Pr\left[\|\boldsymbol{b}\|_1>\frac{n}{2}\right]\right)=\gamma T\left(n-\mathbb{E}\left[\|\boldsymbol{b}\|_1\,\middle|\,\|\boldsymbol{b}\|_1>\frac{n}{2}\right]\right).$$

Since $\|\boldsymbol{b}\|_1\sim\mathrm{Bin}\left(n,\frac{1}{2}\right)$,

$$\mathbb{E}\left[\|\boldsymbol{b}\|_1\,\middle|\,\|\boldsymbol{b}\|_1>\frac{n}{2}\right]=\frac{n}{2^n}\left(2^{n-1}+\binom{n-1}{\frac{n-1}{2}}\right)$$

[25] and by Stirling's approximation,

$$\mathbb{E}\left[\|\boldsymbol{b}\|_1\,\middle|\,\|\boldsymbol{b}\|_1>\frac{n}{2}\right]<\frac{n}{2}+\sqrt{\frac{n}{2\pi}}.$$

We can use these facts to bound Equation (32) as follows:

$$\mathop{\mathbb{E}}_{\boldsymbol{b}\sim\{0,1\}^n}\left[\sum_{t=1}^{T}\left(\frac{\gamma\left(n-\|\boldsymbol{b}\|_1\right)}{n}\sum_{j=1}^{n}\mathop{\mathbb{E}}_{\pi\sim\nu_{\boldsymbol{b}}}\left[\pi_{j,1}^{(t)}\right]+\frac{\gamma\|\boldsymbol{b}\|_1}{n}\sum_{j=1}^{n}\mathop{\mathbb{E}}_{\pi\sim\nu_{\boldsymbol{b}}}\left[\pi_{j,0}^{(t)}\right]\right)\right]\geq\gamma T\left(\frac{n}{2}-\sqrt{\frac{n}{2\pi}}\right). \tag{34}$$

Returning to Equation (31), we will next bound

$$\mathop{\mathbb{E}}_{\boldsymbol{b}}\left[\sum_{t=1}^{T}\left(\sum_{i:b_i=0}\mathop{\mathbb{E}}_{\pi\sim\nu_{\boldsymbol{b}}}\left[\pi_{i,1}^{(t)}-\frac{\gamma}{n}\sum_{j=1}^{n}\pi_{j,1}^{(t)}\right]+\sum_{i:b_i=1}\mathop{\mathbb{E}}_{\pi\sim\nu_{\boldsymbol{b}}}\left[\pi_{i,0}^{(t)}-\frac{\gamma}{n}\sum_{j=1}^{n}\pi_{j,0}^{(t)}\right]\right)\right]$$

$$= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\boldsymbol{b}} \left[\sum_{t=1}^{T} \left(\mathop{\mathbb{E}}_{\pi \nu_{\boldsymbol{b}}} \left[\pi_{i,1}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,1}^{(t)} \right] \mathbf{1}_{\{b_i=0\}} + \mathop{\mathbb{E}}_{\pi \nu_{\boldsymbol{b}}} \left[\pi_{i,0}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,0}^{(t)} \right] \mathbf{1}_{\{b_i=1\}} \right) \right].$$

For each user $i \in [n]$, we will therefore lower bound

$$\begin{aligned} &\mathop{\mathbb{E}}_{\boldsymbol{b}} \left[\sum_{t=1}^{T} \left(\mathop{\mathbb{E}}_{\pi \nu_{\boldsymbol{b}}} \left[\pi_{i,1}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,1}^{(t)} \right] \mathbf{1}_{\{b_i=0\}} + \mathop{\mathbb{E}}_{\pi \nu_{\boldsymbol{b}}} \left[\pi_{i,0}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,0}^{(t)} \right] \mathbf{1}_{\{b_i=1\}} \right) \right] \\ &= \frac{1}{2} \left(\mathop{\mathbb{E}}_{\boldsymbol{b}} \left[\mathop{\mathbb{E}}_{\pi \nu_{\boldsymbol{b}}} \left[\sum_{t=1}^{T} \left(\pi_{i,1}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,1}^{(t)} \right) \right] \middle| \, b_i = 0 \right] + \mathop{\mathbb{E}}_{\boldsymbol{b}} \left[\mathop{\mathbb{E}}_{\pi \nu_{\boldsymbol{b}}} \left[\sum_{t=1}^{T} \left(\pi_{i,0}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,0}^{(t)} \right) \right] \middle| \, b_i = 1 \right] \right). \end{aligned} \tag{35}$$

Let $\boldsymbol{b}_{-i} \in \{0,1\}^{n-1}$ denote the vector $\boldsymbol{b}$ with all components except the i^{th} component. Moreover, to simplify notation, let $\mathbb{P}_{i,0}$ denote the distribution over outcomes $(\boldsymbol{A}_1, \boldsymbol{X}_1, \ldots, \boldsymbol{A}_T, \boldsymbol{X}_T) \in (\{0,1\}^n \times \{0,1\}^n)^T$ defined by first drawing $\boldsymbol{b}_{-i} \sim \text{Unif}\left(\{0,1\}^{n-1}\right)$ and then running the policy π on the instance $\nu_{(0,\boldsymbol{b}_{-i})}$. Similarly, let $\mathbb{P}_{i,1}$ denote the distribution over outcomes $(\boldsymbol{A}_1, \boldsymbol{X}_1, \ldots, \boldsymbol{A}_T, \boldsymbol{X}_T) \in (\{0,1\}^n \times \{0,1\}^n)^T$ defined by first drawing $\boldsymbol{b}_{-i} \sim \text{Unif}\left(\{0,1\}^{n-1}\right)$ and then running the policy π on the instance $\nu_{(1,\boldsymbol{b}_{-i})}$. We can then rewrite Equation (35) as

$$\begin{aligned} &\frac{1}{2} \left(\mathop{\mathbb{E}}_{i,0} \left[\sum_{t=1}^{T} \left(\pi_{i,1}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,1}^{(t)} \right) \right] + \mathop{\mathbb{E}}_{i,1} \left[\sum_{t=1}^{T} \left(\pi_{i,0}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,0}^{(t)} \right) \right] \right) \\ &= \frac{1}{2} \left(\mathop{\mathbb{E}}_{i,0} \left[\sum_{t=1}^{T} \left(\pi_{i,1}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,1}^{(t)} \right) \right] + \mathop{\mathbb{E}}_{i,1} \left[(1-\gamma)T - \sum_{t=1}^{T} \left(\pi_{i,1}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,1}^{(t)} \right) \right] \right). \end{aligned} \tag{36}$$

Based on the constraints, we know that $\pi_{i,1}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,1}^{(t)} \geq 0$ and $\pi_{i,0}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,0}^{(t)} \geq 0$ with probability 1. Therefore, by Markov's inequality and the Bretagnolle–Huber inequality,

$$\begin{aligned} &\frac{1}{2} \left(\mathop{\mathbb{E}}_{i,0} \left[\sum_{t=1}^{T} \left(\pi_{i,1}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,1}^{(t)} \right) \right] + \mathop{\mathbb{E}}_{i,1} \left[(1-\gamma)T - \sum_{t=1}^{T} \left(\pi_{i,1}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,1}^{(t)} \right) \right] \right) \\ &\geq \frac{T(1-\gamma)}{4} \left(\mathbb{P}_{i,0} \left[\sum_{t=1}^{T} \left(\pi_{i,1}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,1}^{(t)} \right) \geq \frac{T(1-\gamma)}{2} \right] + \mathbb{P}_{i,1} \left[\sum_{t=1}^{T} \left(\pi_{i,1}^{(t)} - \frac{\gamma}{n} \sum_{j=1}^{n} \pi_{j,1}^{(t)} \right) < \frac{T(1-\gamma)}{2} \right] \right) \\ &\geq \frac{T(1-\gamma)}{8} \exp\left(-D\left(\mathbb{P}_{i,0}, \mathbb{P}_{i,1} \right) \right). \end{aligned}$$

In the following claim, we bound $D\left(\mathbb{P}_{i,0}, \mathbb{P}_{i,1}\right)$.

Claim E.1. $D\left(\mathbb{P}_{i,0}, \mathbb{P}_{i,1}\right) \leq 8\epsilon^2 T$.

Proof of Claim E.1. In this proof, we will use the following notation to distinguish the reward distributions for each instance $\nu_{\boldsymbol{b}}$. For any vector of rewards $\boldsymbol{x_t} = (x_{t,1}, \ldots, x_{t,n}) \in \{0,1\}^n$ and any choice of arms $\boldsymbol{a_t} = (a_{t,1}, \ldots, a_{t,n}) \in \{0,1\}^n$, we will use the notation $f_{\boldsymbol{a_t}}^{\boldsymbol{b}}(\boldsymbol{x_t})$ to denote the probability that the platform receives rewards $\boldsymbol{x_t}$ under instance $\nu_{\boldsymbol{b}}$ after choosing arms $\boldsymbol{a_t}$. We also use $f_{i,0}^{(b_i)} : \{0,1\} \to [0,1]$ to denote the PMF of arm 0 for user i and $f_{i,1}^{(b_i)} : \{0,1\} \to [0,1]$ to denote the PMF of arm 1 for user i. In other words, $f_{i,0}^{(0)}$ is the $\text{Bern}\left(\frac{1}{2} + \epsilon\right)$ PMF, $f_{i,1}^{(0)}$ is the $\text{Bern}\left(\frac{1}{2}\right)$ PMF, $f_{i,0}^{(1)}$ is

the Bern$\left(\frac{1}{2}\right)$ PMF, and $f_{i,1}^{(1)}$ is the Bern$\left(\frac{1}{2}+\epsilon\right)$ PMF. With this notation,

$$f_{\boldsymbol{a}_t}^{\boldsymbol{b}}(\boldsymbol{x}_t) = \prod_{i=1}^{n} f_{i,a_{t,i}}^{(b_i)}\left(x_{t,i}\right). \tag{37}$$

Moving now to KL divergence between $\mathbb{P}_{i,0}$ and $\mathbb{P}_{i,1}$, let $f_{i,0} : (\{0,1\}^n \times \{0,1\}^n)^T \to [0,1]$ be the probability mass function of the distribution $\mathbb{P}_{i,0}$, and define $f_{i,1}$ similarly. By definition,

$$D\left(\mathbb{P}_{i,0}, \mathbb{P}_{i,1}\right) = \sum_{(\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T} f_{i,0}\left((\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T\right) \log \frac{f_{i,0}\left((\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T\right)}{f_{i,1}\left((\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T\right)}. \tag{38}$$

We will begin by simplifying the logarithm in Equation (38). Beginning with the numerator of the logarithm, we have that

$$f_{i,0}\left((\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T\right) = \mathbb{P}_{i,0}\left[(\boldsymbol{A}_t,\boldsymbol{X}_t)_{t=1}^T = (\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T\right] = \frac{1}{2^{n-1}} \sum_{\boldsymbol{b}_{-i}\in\{0,1\}^{n-1}} \mathbb{P}_{\pi_{\nu_{(0,\boldsymbol{b}_{-i})}}}\left[(\boldsymbol{A}_t,\boldsymbol{X}_t)_{t=1}^T = (\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T\right].$$

Using the notation defined in Section 2 (Equation (1)), we have that

$$f_{i,0}\left((\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T\right) = \frac{1}{n-1}\sum_{\boldsymbol{b}_{-i}} f_{\pi_{\nu_{(0,\boldsymbol{b}_{-i})}}}\left((\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T\right) = \frac{1}{2^{n-1}}\sum_{\boldsymbol{b}_{-i}}\prod_{t=1}^{T}\pi(\boldsymbol{a}_t \mid \boldsymbol{a}_1,\boldsymbol{x}_1,\ldots,\boldsymbol{a}_{t-1},\boldsymbol{x}_{t-1}) f_{\boldsymbol{a}_t}^{(0,\boldsymbol{b}_{-i})}(\boldsymbol{x}_t).$$

Applying Equation (37), we have that

$$f_{i,0}\left((\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T\right) = \frac{1}{2^{n-1}}\sum_{\boldsymbol{b}_{-i}}\prod_{t=1}^{T}\left(\pi(\boldsymbol{a}_t \mid \boldsymbol{a}_1,\boldsymbol{x}_1,\ldots,\boldsymbol{a}_{t-1},\boldsymbol{x}_{t-1}) f_{i,a_{t,i}}^{(0)}\left(x_{t,i}\right)\prod_{j\neq i} f_{j,a_{t,j}}^{(b_j)}(x_{t,j})\right)$$

where b_j indicates the j^{th} component of the vector $\boldsymbol{b}_{-i}$. Rearranging the product within the summation, we have that $f_{i,0}\left((\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T\right)$ is equal to

$$\frac{1}{2^{n-1}}\sum_{\boldsymbol{b}_{-i}}\left(\prod_{t=1}^{T}\left(\pi(\boldsymbol{a}_t \mid \boldsymbol{a}_1,\boldsymbol{x}_1,\ldots,\boldsymbol{a}_{t-1},\boldsymbol{x}_{t-1}) f_{i,a_{t,i}}^{(0)}\left(x_{t,i}\right)\right)\prod_{t=1}^{T}\left(\prod_{j\neq i} f_{j,a_{t,j}}^{(b_j)}(x_{t,j})\right)\right).$$

Since $\prod_{t=1}^{T}\pi(\boldsymbol{a}_t \mid \boldsymbol{a}_1,\boldsymbol{x}_1,\ldots,\boldsymbol{a}_{t-1},\boldsymbol{x}_{t-1}) f_{i,a_{t,i}}^{(0)}\left(x_{t,i}\right)$ does not depend on $\boldsymbol{b}_{-i}$, we rearrange the summation over $\boldsymbol{b}_{-i}$ as

$$f_{i,0}\left((\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T\right) = \frac{1}{2^{n-1}}\prod_{t=1}^{T}\left(\pi(\boldsymbol{a}_t \mid \boldsymbol{a}_1,\boldsymbol{x}_1,\ldots,\boldsymbol{a}_{t-1},\boldsymbol{x}_{t-1}) f_{i,a_{t,i}}^{(0)}\left(x_{t,i}\right)\right)\sum_{\boldsymbol{b}_{-i}}\prod_{t=1}^{T}\prod_{j\neq i} f_{j,a_{t,j}}^{(b_j)}(x_{t,j}). \tag{39}$$

Similarly,

$$f_{i,1}\left((\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T\right) = \frac{1}{2^{n-1}}\prod_{t=1}^{T}\left(\pi(\boldsymbol{a}_t \mid \boldsymbol{a}_1,\boldsymbol{x}_1,\ldots,\boldsymbol{a}_{t-1},\boldsymbol{x}_{t-1}) f_{i,a_{t,i}}^{(1)}\left(x_{t,i}\right)\right)\sum_{\boldsymbol{b}_{-i}}\prod_{t=1}^{T}\prod_{j\neq i} f_{j,a_{t,j}}^{(b_j)}(x_{t,j}). \tag{40}$$

We now return to the logarithm in Equation (38). Based on Equations (39) and (40), much of the numerator and denominator cancel out, leaving us with

$$\log\frac{f_{i,0}\left((\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T\right)}{f_{i,1}\left((\boldsymbol{a}_t,\boldsymbol{x}_t)_{t=1}^T\right)} = \log\prod_{t=1}^{T}\frac{f_{i,a_{t,i}}^{(0)}\left(x_{t,i}\right)}{f_{i,a_{t,i}}^{(1)}\left(x_{t,i}\right)} = \sum_{t=1}^{T}\log\frac{f_{i,a_{t,i}}^{(0)}\left(x_{t,i}\right)}{f_{i,a_{t,i}}^{(1)}\left(x_{t,i}\right)}.$$

We can therefore can write the KL divergence as

$$D\left(\mathbb{P}_{i,0}, \mathbb{P}_{i,1}\right) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{i,0}\left[\log \frac{f^{(0)}_{i,A_{t,i}}\left(X_{t,i}\right)}{f^{(1)}_{i,A_{t,i}}\left(X_{t,i}\right)}\right].$$

Moreover, by the law of total expectation, $D\left(\mathbb{P}_{i,0}, \mathbb{P}_{i,1}\right)$ is equal to

$$\sum_{t=1}^{T}\left(\mathop{\mathbb{E}}_{i,0}\left[\log \frac{f^{(0)}_{i,0}\left(X_{t,i}\right)}{f^{(1)}_{i,0}\left(X_{t,i}\right)} \,\middle|\, A_{t,i}=0\right]\mathbb{P}\left[A_{t,i}=0\right] + \mathop{\mathbb{E}}_{i,0}\left[\log \frac{f^{(0)}_{i,1}\left(X_{t,i}\right)}{f^{(1)}_{i,1}\left(X_{t,i}\right)} \,\middle|\, A_{t,i}=1\right]\mathbb{P}\left[A_{t,i}=1\right]\right). \tag{41}$$

Inspecting each conditional expectation in this sum, we have that

$$\mathop{\mathbb{E}}_{i,0}\left[\log \frac{f^{(0)}_{i,0}\left(X_{t,i}\right)}{f^{(1)}_{i,0}\left(X_{t,i}\right)} \,\middle|\, A_{t,i}=0\right] = \mathbb{P}_{i,0}\left[X_{t,i}=0 \mid A_{t,i}=0\right]\cdot\log\frac{f^{(0)}_{i,0}(0)}{f^{(1)}_{i,0}(0)} + \mathbb{P}_{i,0}\left[X_{t,i}=1 \mid A_{t,i}=0\right]\cdot\log\frac{f^{(0)}_{i,0}(1)}{f^{(1)}_{i,0}(1)}.$$

By Equation (27), for any instance $v_{\boldsymbol{b}}$ such that $b_i = 0$, the probability that $X_{t,i} = 0$ given that $A_{t,i} = 0$ is $\frac{1}{2} - \epsilon = f^{(0)}_{i,0}(0)$. Similarly, the probability that $X_{t,i} = 1$ given that $A_{t,i} = 0$ is $\frac{1}{2} + \epsilon = f^{(0)}_{i,0}(1)$. Therefore,

$$\mathop{\mathbb{E}}_{i,0}\left[\log \frac{f^{(0)}_{i,0}\left(X_{t,i}\right)}{f^{(1)}_{i,0}\left(X_{t,i}\right)} \,\middle|\, A_{t,i}=0\right] = f^{(0)}_{i,0}(0)\log\frac{f^{(0)}_{i,0}(0)}{f^{(1)}_{i,0}(0)} + f^{(0)}_{i,0}(1)\cdot\log\frac{f^{(0)}_{i,0}(1)}{f^{(1)}_{i,0}(1)} = D\left(f^{(0)}_{i,0}, f^{(1)}_{i,0}\right). \tag{42}$$

Similarly,

$$\mathop{\mathbb{E}}_{i,0}\left[\log \frac{f^{(0)}_{i,1}\left(X_{t,i}\right)}{f^{(1)}_{i,1}\left(X_{t,i}\right)} \,\middle|\, A_{t,i}=1\right] = D\left(f^{(0)}_{i,1}, f^{(1)}_{i,1}\right). \tag{43}$$

Let $N_{i,0}(T)$ be the number of rounds that user i is shown arm 0 and let $N_{i,1}(T)$ be the number of rounds that user i is shown arm 0, so $N_{i,0}(T) + N_{i,1}(T) = T$. Combining Equations (41), (42), and (43), we have that

$$\begin{aligned} D\left(\mathbb{P}_{i,0}, \mathbb{P}_{i,1}\right) &= \sum_{t=1}^{T}\left(D\left(f^{(0)}_{i,0}, f^{(1)}_{i,0}\right)\mathbb{P}_{i,0}\left[A_{t,i}=0\right] + D\left(f^{(0)}_{i,1}, f^{(1)}_{i,1}\right)\mathbb{P}_{i,0}\left[A_{t,i}=1\right]\right) \\ &= D\left(f^{(0)}_{i,0}, f^{(1)}_{i,0}\right)\mathop{\mathbb{E}}_{i,0}\left[N_{i,0}(T)\right] + D\left(f^{(0)}_{i,1}, f^{(1)}_{i,1}\right)\mathop{\mathbb{E}}_{i,0}\left[N_{i,1}(T)\right]. \end{aligned}$$

Since $f^{(0)}_{i,0}$ is Bern$\left(\frac{1}{2}+\epsilon\right)$, $f^{(0)}_{i,1}$ is Bern$\left(\frac{1}{2}\right)$, $f^{(1)}_{i,0}$ is Bern$\left(\frac{1}{2}\right)$, and $f^{(1)}_{i,1}$ is Bern$\left(\frac{1}{2}+\epsilon\right)$,

$$D\left(\mathbb{P}_{i,0}, \mathbb{P}_{i,1}\right) \le 8\epsilon^2\left(\mathop{\mathbb{E}}_{i,0}\left[N_{i,0}(T)\right] + \mathop{\mathbb{E}}_{i,0}\left[N_{i,1}(T)\right]\right) = 8\epsilon^2 T.$$

□

By Claim E.1, $D\left(\mathbb{P}_{i,0}, \mathbb{P}_{i,1}\right) \le 8\epsilon^2 T$, so if we set $\epsilon = \sqrt{\frac{1}{8T}}$, we have that

$$\frac{1}{2}\left(\mathop{\mathbb{E}}_{i,0}\left[\sum_{t=1}^{T}\left(\pi^{(t)}_{i,1} - \frac{\gamma}{n}\sum_{j=1}^{n}\pi^{(t)}_{j,1}\right)\right] + \mathop{\mathbb{E}}_{i,1}\left[(1-\gamma)T - \sum_{t=1}^{T}\left(\pi^{(t)}_{i,1} - \frac{\gamma}{n}\sum_{j=1}^{n}\pi^{(t)}_{j,1}\right)\right]\right) \ge \frac{T(1-\gamma)}{8e}. \tag{44}$$

Combining Equations (30), (31), and (44), we have that the regret is lower bounded by

$$\sqrt{\frac{1}{8T}}\left(\frac{nT(1-\gamma)}{8e} + \gamma T\left(\frac{n}{2} - \sqrt{\frac{n}{2\pi}}\right) - \frac{2T\gamma}{n}\mathop{\mathbb{E}}_{\boldsymbol{b}\sim\{0,1\}^n}\left[\|\boldsymbol{b}\|_1\left(n - \|\boldsymbol{b}\|_1\right)\right]\right).$$

Since $\mathbb{E}_{\boldsymbol{b}\sim\{0,1\}^n}\left[\|\boldsymbol{b}\|_1\left(n-\|\boldsymbol{b}\|_1\right)\right]=\frac{n}{4}(n-1)$, we have that the expected regret is lower bounded by

$$\sqrt{\frac{T}{8}}\left(\frac{n(1-\gamma)}{8e}+\gamma\left(\frac{n}{2}-\sqrt{\frac{n}{2\pi}}\right)-\frac{\gamma(n-1)}{2}\right)\geq\sqrt{\frac{T}{8}}\left(\frac{n}{8e}-\gamma\left(\frac{n}{8e}+\sqrt{\frac{n}{2\pi}}\right)\right).$$

□

Lemma 5.6. *For all $T\geq\frac{7(k-1)}{n}$, the regret is lower bounded as follows:*

$$\inf_{\pi\in\Pi^{n,k}}\sup_{\nu\in\mathcal{E}^{n,k}}R_{T,1}(\pi,\nu)\geq\frac{\sqrt{nT(k-1)}}{16e}.$$

Proof. We begin by defining the worst-case instance ν where for each user $i\in[n]$, their reward distributions for the k arms are Bernoulli with means

$$\boldsymbol{\mu}_1=\cdots=\boldsymbol{\mu}_n=\left(\frac{1}{2}+\epsilon,\frac{1}{2},\frac{1}{2},\ldots,\frac{1}{2}\right)$$

where $\epsilon=\sqrt{\frac{k-1}{8nT}}$. We will use the notation $N_{i,j}(T)$ to denote the number of rounds that user i is shown arm j and $N_j(T)=\sum_{i=1}^n N_{i,j}(T)$ to denote the total number of rounds that all users are shown arm j. This means that $\sum_{j=1}^k N_j(T)=nT$. Under instance ν, the optimal policy obtains a reward of $nT\left(\frac{1}{2}+\epsilon\right)$. Meanwhile, an arbitrary policy π will obtain a reward of

$$\left(\frac{1}{2}+\epsilon\right)\mathop{\mathbb{E}}_{\pi\nu}\left[N_1(T)\right]+\frac{1}{2}\sum_{j=2}^n\mathop{\mathbb{E}}_{\pi\nu}\left[N_j(T)\right]=\frac{nT}{2}+\epsilon\mathop{\mathbb{E}}_{\pi\nu}\left[N_1(T)\right].$$

Therefore, the regret of policy π on instance ν is

$$R_T(\pi,\nu)=\epsilon\left(nT-\mathop{\mathbb{E}}_{\pi\nu}\left[N_1(T)\right]\right).$$

Fix a policy π and let $j^*=\operatorname{argmin}_{j>2}\mathbb{E}_{\pi\nu}[N_j(T)]$. Since $\sum_{j=2}^k N_j(T)\leq nT$, we have that $\mathbb{E}_{\pi\nu}\left[N_{j^*}(T)\right]\leq\frac{nT}{k-1}$. We now use j^* to construct a second worst-case instance ν' where for each user $i\in[n]$,

$$\mu_{i,j}=\begin{cases}\frac{1}{2}+\epsilon & \text{if } j=1\\ \frac{1}{2}+2\epsilon & \text{if } j=j^*\\ \frac{1}{2} & \text{else.}\end{cases}$$

Under instance ν', the optimal policy obtains a reward of $nT\left(\frac{1}{2}+2\epsilon\right)$. Meanwhile, policy π will obtain a reward of

$$\left(\frac{1}{2}+\epsilon\right)\mathop{\mathbb{E}}_{\pi\nu'}\left[N_1(T)\right]+\left(\frac{1}{2}+2\epsilon\right)\mathop{\mathbb{E}}_{\pi\nu'}\left[N_{j^*}(T)\right]+\frac{1}{2}\sum_{j\notin\{1,j^*\}}\mathop{\mathbb{E}}_{\pi\nu'}\left[N_j(T)\right]=\frac{nT}{2}+\epsilon\mathop{\mathbb{E}}_{\pi\nu'}\left[N_1(T)\right]+2\epsilon\mathop{\mathbb{E}}_{\pi\nu'}\left[N_{j^*}(T)\right].$$

Therefore,

$$\begin{aligned}R_T(\pi,\nu')&=\epsilon\left(2nT-\mathop{\mathbb{E}}_{\pi\nu'}\left[N_1(T)\right]-2\mathop{\mathbb{E}}_{\pi\nu'}\left[N_{j^*}(T)\right]\right)=\epsilon\left(2\sum_{j=1}^k\mathop{\mathbb{E}}_{\pi\nu'}\left[N_j(T)\right]-\mathop{\mathbb{E}}_{\pi\nu'}\left[N_1(T)\right]-2\mathop{\mathbb{E}}_{\pi\nu'}\left[N_{j^*}(T)\right]\right)\\&\geq\epsilon\mathop{\mathbb{E}}_{\pi\nu'}\left[N_1(T)\right].\end{aligned}$$

By Markov's inequality,

$$R_T(\pi, \nu) + R_T(\pi, \nu') \geq \frac{\epsilon n T}{2} \left(\mathbb{P}_{\pi\nu} \left[N_1(T) \leq \frac{nT}{2} \right] + \mathbb{P}_{\pi\nu'} \left[N_1(T) > \frac{nT}{2} \right] \right),$$

so by the Bretagnolle–Huber inequality,

$$R_T(\pi, \nu) + R_T(\pi, \nu) \geq \frac{\epsilon n T}{4} \exp\left(-D\left(\mathbb{P}_{\pi\nu}, \mathbb{P}_{\pi\nu'}\right)\right). \tag{45}$$

Claim E.2. *For $\epsilon < \frac{1}{5}$, $D\left(\mathbb{P}_{\pi\nu}, \mathbb{P}_{\pi\nu'}\right) \leq \frac{4\epsilon^2 nT}{k-1}$.*

Proof of Claim E.2. In this proof, we will use the following notation to distinguish the reward distributions for the instances ν and ν'. For any vector of rewards $\boldsymbol{x}_t = \left(x_{t,1}, \ldots, x_{t,n}\right) \in \{0,1\}^n$ and any choice of arms $\boldsymbol{a}_t = \left(a_{t,1}, \ldots, a_{t,n}\right) \in [k]^n$, we use the notation $f_{\boldsymbol{a}_t}(\boldsymbol{x}_t)$ (respectively, $f'_{\boldsymbol{a}_t}(\boldsymbol{x}_t)$) to denote the probability that the platform receives rewards $\boldsymbol{x}_t$ after choosing arms $\boldsymbol{a}_t$ under the instance ν (respectively, ν'). We also use $f_{i,j} : \{0,1\} \to [0,1]$ (respectively, $f'_{i,j} : \{0,1\} \to [0,1]$) to denote the PMF of arm j for user i. With this notation,

$$f_{\boldsymbol{a}_t}(\boldsymbol{x}_t) = \prod_{i=1}^{n} f_{i,a_{t,i}}\left(x_{t,i}\right) \tag{46}$$

and

$$f'_{\boldsymbol{a}_t}(\boldsymbol{x}_t) = \prod_{i=1}^{n} f'_{i,a_{t,i}}\left(x_{t,i}\right). \tag{47}$$

By definition,

$$D\left(\mathbb{P}_{\pi\nu}, \mathbb{P}_{\pi\nu'}\right) = \sum_{(\boldsymbol{a}_t, \boldsymbol{x}_t)_{t=1}^T} f_{\pi\nu}\left((\boldsymbol{a}_t, \boldsymbol{x}_t)_{t=1}^T\right) \log \frac{f_{\pi\nu}\left((\boldsymbol{a}_t, \boldsymbol{x}_t)_{t=1}^T\right)}{f_{\pi\nu'}\left((\boldsymbol{a}_t, \boldsymbol{x}_t)_{t=1}^T\right)}. \tag{48}$$

We will begin by simplifying the logarithm in Equation (48). Beginning with the numerator of the logarithm and using the notation defined in Section 2 (Equation (1)), we have that

$$f_{\pi\nu}\left((\boldsymbol{a}_t, \boldsymbol{x}_t)_{t=1}^T\right) = \mathbb{P}_{\pi\nu}\left[(\boldsymbol{A}_t, \boldsymbol{X}_t)_{t=1}^T = (\boldsymbol{a}_t, \boldsymbol{x}_t)_{t=1}^T\right] = \prod_{t=1}^{T} \pi(\boldsymbol{a}_t \mid \boldsymbol{a}_1, \boldsymbol{x}_1, \ldots, \boldsymbol{a}_{t-1}, \boldsymbol{x}_{t-1}) f_{\boldsymbol{a}_t}(\boldsymbol{x}_t).$$

By Equation (46), we have that

$$f_{\pi\nu}\left((\boldsymbol{a}_t, \boldsymbol{x}_t)_{t=1}^T\right) = \prod_{t=1}^{T} \pi(\boldsymbol{a}_t \mid \boldsymbol{a}_1, \boldsymbol{x}_1, \ldots, \boldsymbol{a}_{t-1}, \boldsymbol{x}_{t-1}) \prod_{i=1}^{n} f_{i,a_{t,i}}\left(x_{t,i}\right). \tag{49}$$

Similarly,

$$f_{\pi\nu'}\left((\boldsymbol{a}_t, \boldsymbol{x}_t)_{t=1}^T\right) = \prod_{t=1}^{T} \pi(\boldsymbol{a}_t \mid \boldsymbol{a}_1, \boldsymbol{x}_1, \ldots, \boldsymbol{a}_{t-1}, \boldsymbol{x}_{t-1}) \prod_{i=1}^{n} f'_{i,a_{t,i}}\left(x_{t,i}\right). \tag{50}$$

We now return to the logarithm in Equation (48). Based on Equations (49) and (50), much of the numerator and denominator cancel out, leaving us with

$$\log \frac{f_{\pi\nu}\left((\boldsymbol{a}_t, \boldsymbol{x}_t)_{t=1}^T\right)}{f_{\pi\nu'}\left((\boldsymbol{a}_t, \boldsymbol{x}_t)_{t=1}^T\right)} = \log \prod_{t=1}^{T} \prod_{i=1}^{n} \frac{f_{i,a_{t,i}}\left(x_{t,i}\right)}{f'_{i,a_{t,i}}\left(x_{t,i}\right)} = \sum_{t=1}^{T} \sum_{i=1}^{n} \log \frac{f_{i,a_{t,i}}\left(x_{t,i}\right)}{f'_{i,a_{t,i}}\left(x_{t,i}\right)}.$$

We can therefore can write the KL divergence as

$$D\left(\mathbb{P}_{\pi\nu}, \mathbb{P}_{\pi\nu'}\right) = \sum_{i=1}^{n}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi\nu}\left[\log \frac{f_{i,A_{t,i}}\left(X_{t,i}\right)}{f'_{i,A_{t,i}}\left(X_{t,i}\right)}\right].$$

Moreover, by the law of total expectation,

$$D\left(\mathbb{P}_{\pi\nu}, \mathbb{P}_{\pi\nu'}\right) = \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{k} \mathop{\mathbb{E}}_{\pi\nu}\left[\log \frac{f_{i,j}\left(X_{t,i}\right)}{f'_{i,j}\left(X_{t,i}\right)} \,\middle|\, A_{t,i} = j\right] \mathbb{P}_{\pi\nu}\left[A_{t,i} = j\right].$$

We know that for all $j \neq j^*$, $f_{i,j} = f'_{i,j}$, which means that

$$D\left(\mathbb{P}_{\pi\nu}, \mathbb{P}_{\pi\nu'}\right) = \sum_{i=1}^{n}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi\nu}\left[\log \frac{f_{i,j^*}\left(X_{t,i}\right)}{f'_{i,j^*}\left(X_{t,i}\right)} \,\middle|\, A_{t,i} = j^*\right] \mathbb{P}_{\pi\nu}\left[A_{t,i} = j^*\right]. \tag{51}$$

By further conditioning,

$$\mathop{\mathbb{E}}_{\pi\nu}\left[\log \frac{f_{i,j^*}\left(X_{t,i}\right)}{f'_{i,j^*}\left(X_{t,i}\right)} \,\middle|\, A_{t,i} = j^*\right] = \mathbb{P}_{\pi\nu}\left[X_{t,i} = 0 \mid A_{t,i} = j^*\right] \cdot \log \frac{f_{i,j^*}(0)}{f'_{i,j^*}(0)} + \mathbb{P}_{\pi\nu}\left[X_{t,i} = 1 \mid A_{t,i} = j^*\right] \cdot \log \frac{f_{i,j^*}(1)}{f'_{i,j^*}(1)}.$$

Under instance ν, the probability that $X_{t,i} = 0$ given that $A_{t,i} = j^*$ is $f_{i,j^*}(0)$. Similarly, the probability that $X_{t,i} = 1$ given that $A_{t,i} = j^*$ is $f_{i,j^*}(1)$. Therefore,

$$\mathop{\mathbb{E}}_{\pi\nu_0}\left[\log \frac{f_{i,j^*}\left(X_{t,i}\right)}{f'_{i,j^*}\left(X_{t,i}\right)} \,\middle|\, A_{t,i} = j^*\right] = f_{i,j^*}(0) \log \frac{f_{i,j^*}(0)}{f'_{i,j^*}(0)} + f_{i,j^*}(1) \cdot \log \frac{f_{i,j^*}(1)}{f'_{i,j^*}(1)} = D\left(f_{i,j^*}, f'_{i,j^*}\right). \tag{52}$$

Moreover, since f_{i,j^*} is the $\mathrm{Bin}\left(\frac{1}{2} + \epsilon\right)$ PMF and f'_{i,j^*} is the $\mathrm{Bin}\left(\frac{1}{2} + 2\epsilon\right)$ PMF, we have that $D\left(f_{i,j^*}, f'_{i,j^*}\right) \leq 4\epsilon^2$ for $\epsilon < \frac{1}{5}$.

Combining Equations (51) and (52), we have that

$$D\left(\mathbb{P}_{\pi\nu}, \mathbb{P}_{\pi\nu'}\right) = \sum_{i=1}^{n}\sum_{t=1}^{T} D\left(f_{i,j^*}, f'_{i,j^*}\right) \mathbb{P}_{\pi\nu}\left[A_{t,i} = j^*\right] \leq 4\epsilon^2 \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi\nu}\left[N_{i,j^*}(T)\right] = 4\epsilon^2 \mathop{\mathbb{E}}_{\pi\nu}\left[N_{j^*}(T)\right] \leq \frac{4\epsilon^2 nT}{k-1}.$$

□

Combining Equation (45) and Claim E.2, and setting $\epsilon = \sqrt{\frac{k-1}{4nT}}$ (in which case $\epsilon < \frac{1}{5}$ for $nT > 7(k-1)$), we have that $R_T(\pi, \nu_0) + R_T(\pi, \nu_1) \geq \frac{1}{8e}\sqrt{nT(k-1)}$, so $\max\left\{R_T(\pi, \nu_0), R_T(\pi, \nu_1)\right\} \geq \frac{1}{16e}\sqrt{nT(k-1)}$. □

F PROOFS ABOUT THE FORMULATION 2 REGRET BOUNDS

In this section, we will use the following notation. For any distributions $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n \in \mathcal{P}^{k-1}$, denote the penalty attributed to user i as

$$P_i((\boldsymbol{p}_i)_{i\in[n]}; \gamma, \eta) = \eta \sum_{j=1}^{k} \max\left\{0, \frac{\gamma}{n}\sum_{i'=1}^{n} p_{i',j} - p_{i,j}\right\}.$$

The total penalty across all n users is

$$P((\boldsymbol{p}_i)_{i\in[n]}; \gamma, \eta) = \sum_{i=1}^{n} P_i((\boldsymbol{p}_i)_{i\in[n]}; \gamma, \eta).$$

Algorithm 3 Penalty-UCB (defined by parameter δ)

Require: Failure probability $\delta > 0$

Set $N_{i,j}(0) = 0,\ \forall i \in [n], j \in [k]$; $\hat{\boldsymbol{\mu}}_i^{(0)} = \mathbf{0},\ \forall i \in [n]$
for $t \in \{1, \dots, T\}$ **do**
  if $t \in \{1, \dots, k\}$ **then**
    Set $\boldsymbol{p}_i^{(t)} = \boldsymbol{e}_t$
  else
    Set $\left(\boldsymbol{p}_i^{(t)}\right)_{i\in[n]} = \operatorname{argmax}\left\{\sum_{i=1}^{n} \boldsymbol{p}_i \cdot \hat{\boldsymbol{\mu}}_i^{(t-1)} - \eta \sum_{j=1}^{k} \max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n} p_{i',j} - p_{i,j}, 0\right\}\right\}$
  end if
  Draw $j_i^{(t)} \sim \boldsymbol{p}_i^{(t)}\ \forall i \in [n]$
  Receive reward $r_i^{(t)} \sim X_{i,j_i^{(t)}}$
  $N_{i,j_i^t}(t) = N_{i,j_i^t}(t-1) + 1,\ \forall i \in [n]$
  $N_{i,j}(t) = N_{i,j}(t-1),\ \forall i \in [n]$ and $j \neq j_i^t$
  $\beta_{i,j}^{(t)} = \sqrt{\frac{\log(2Tnk/\delta)}{N_{i,j}(t)}},\ \forall i \in [n], j \in [k]$
  $\hat{\boldsymbol{\mu}}_{i,j}^t = \frac{1}{N_{i,j}(t)} \sum_{\tau=1}^{t} r_i^{(\tau)} \mathbb{1}\left\{j_i^{(\tau)} = j\right\} + \beta_{i,j}^{(t)},\ \forall i \in [n], j \in [k]$
end for

Next, let $\boldsymbol{p}_1^*, \dots, \boldsymbol{p}_n^* \in \mathcal{P}^{k-1}$ be distributions that maximize the expected reward minus the penalties. Then the expected regret of a policy π under this formulation is

$$T\sum_{i=1}^{n}\left(\boldsymbol{p}_i^* \cdot \boldsymbol{\mu}_i - P_i((\boldsymbol{p}_i^*)_{i\in[n]}; \gamma, \eta)\right) - \mathbb{E}\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\left(X_{i,t} - P_i((\boldsymbol{\pi}_i(\boldsymbol{h}_{t-1}))_{i\in[n]}; \gamma, \eta)\right)\right].$$

Theorem 5.7. *Let π be the policy of Penalty-UCB. Then $R_{T,2}(\pi, \nu) = \tilde{O}(n\sqrt{kT})$.*

Proof. Fix a timestep $t \in [T]$

$$\begin{aligned}
&\sum_{i=1}^{n}(\boldsymbol{p}_i^* \cdot \boldsymbol{\mu}_i - \boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\mu}_i) - (P_i((\boldsymbol{p}_i^*)_{i\in[n]}; \gamma, \eta) - P_i((\boldsymbol{p}_i^{(t)})_{i\in[n]}; \gamma, \eta)) \\
&\quad= \sum_{i=1}^{n}(\boldsymbol{p}_i^* \cdot \boldsymbol{\mu}_i - P_i((\boldsymbol{p}_i^*)_{i\in[n]}; \gamma, \eta)) - (\boldsymbol{p}_i^{(t)} \cdot \hat{\boldsymbol{\mu}}_i^{(t)} - P_i((\boldsymbol{p}_i^{(t)})_{i\in[n]}; \gamma, \eta)) - \boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\mu}_i + \boldsymbol{p}_i^{(t)} \cdot \hat{\boldsymbol{\mu}}_i^{(t)} \\
&\quad\leq \sum_{i=1}^{n}(\boldsymbol{p}_i^* \cdot \boldsymbol{\mu}_i - P_i((\boldsymbol{p}_i^*)_{i\in[n]}; \gamma, \eta)) - (\boldsymbol{p}_i^* \cdot \hat{\boldsymbol{\mu}}_i^{(t)} - P_i((\boldsymbol{p}_i^*)_{i\in[n]}; \gamma, \eta)) - \boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\mu}_i + \boldsymbol{p}_i^{(t)} \cdot \hat{\boldsymbol{\mu}}_i^{(t)} \\
&\quad= \sum_{i=1}^{n}(\boldsymbol{p}_i^* \cdot \boldsymbol{\mu}_i - \boldsymbol{p}_i^* \cdot \hat{\boldsymbol{\mu}}_i^{(t)} - \boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\mu}_i + \boldsymbol{p}_i^{(t)} \cdot \hat{\boldsymbol{\mu}}_i^{(t)})
\end{aligned}$$

By Claim C.1, $\boldsymbol{p} \cdot \boldsymbol{\mu}_i \leq \boldsymbol{p} \cdot \hat{\boldsymbol{\mu}}_i^{(t)}\ \forall i \in [n]$ and all $\boldsymbol{p} \in \mathbb{R}^k_{\geq 0}$

$$\sum_{i=1}^{n}(\boldsymbol{p}_i^* \cdot \boldsymbol{\mu}_i - \boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\mu}_i) - (P_i((\boldsymbol{p}_i^*)_{i\in[n]}; \gamma, \eta) - P_i((\boldsymbol{p}_i^{(t)})_{i\in[n]}; \gamma, \eta)) \leq \sum_{i=1}^{n}(\boldsymbol{p}_i^{(t)} \cdot \hat{\boldsymbol{\mu}}_i^{(t)} - \boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\mu}_i)$$

$$\leq \sum_{i=1}^{n} \boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\beta}_i^{(t)}$$

Thus

$$R_T \leq \sum_{t=1}^{T} \sum_{i\in[n]} \boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\beta}_i^{(t)}$$

Let $\mathcal{F}_{i,t-1}$ be the sigma algebra defined up to the choice of $\boldsymbol{p}_i^{(t)}$ and $j_i^{(t)'}$ be a random variable distributed as $\boldsymbol{p}_i^{(t)} \mid \mathcal{F}_{i,t-1}$ and conditionally independent from $j_{i,t}^{(t)}$, i.e. $j_i^{(t)'} \perp j_i^{(t)} \mid \mathcal{F}_{i,t-1}$. Note that by definition the following equality holds:

$$\mathop{\mathbb{E}}_{j_i^{(t)}\sim \boldsymbol{p}_i^{(t)}}[\beta_{i,j_i^{(t)}}] = \mathop{\mathbb{E}}_{j_i^{(t)'}\sim \boldsymbol{p}_i^{(t)}}[\beta_{i,j_i^{(t)'}} \mid \mathcal{F}_{i,t-1}].$$

Consider the following random variables $A_{i,t} = \mathbb{E}_{j_i^{(t)'}\sim \boldsymbol{p}_i^{(t)}}[\beta_{i,j_i^{(t)'}} \mid \mathcal{F}_{i,t-1}] - \beta_{i,j_i^{(t)}}(t)$. Note that $M_{i,t} = \sum_{s=1}^{t} A_{i,s}$ is a martingale. Since $|A_t| \leq 2\sqrt{2\log(nkT/\delta)}$, a simple application of Azuma-Hoeffding implies that with probability at least $1-\delta$,

$$R_T = \sum_{t=1}^{T} \sum_{i\in[n]} \boldsymbol{p}_i^{(t)} \cdot \boldsymbol{\beta}_i^{(t)} \leq \sum_{t=1}^{T} \sum_{i\in[n]} \boldsymbol{\beta}_{i,j_i^{(t)}}^{(t)} + n\sqrt{T\log\left(\frac{nkT}{\delta}\right)\log\left(\frac{1}{\delta}\right)}$$

Now let us bound $\sum_{i\in[n]} \sum_{t=i}^{T} \boldsymbol{\beta}_i^{(t)}$,

$$\sum_{i\in[n]} \sum_{t=i}^{T} \boldsymbol{\beta}_i^{(t)} = \sum_{i\in[n]} \sum_{j\in[k]} \sum_{t=i}^{T} \beta_{i,j}^{(t)} \mathbb{1}\left\{j_i^{(t)} = j\right\}$$

For fixed i, j

$$\sum_{t=i}^{T} \beta_{i,j}^{(t)} \mathbb{1}\left\{j_i^{(t)} = j\right\} = \sqrt{\log(Tnk/\delta)} \sum_{t=1}^{N_{i,j}(T)} 1/\sqrt{t} \leq 2\sqrt{N_{i,j}(T)\log(Tnk/\delta)}$$

Therefore

$$\begin{aligned}
\sum_{i\in[n]} \sum_{t=i}^{T} \boldsymbol{\beta}_i^{(t)} &\leq 2 \sum_{i\in[n]} \sum_{j\in[k]} \sqrt{N_{i,j}(T)\log(Tnk/\delta)} \\
&\leq 2 \sum_{i\in[n]} \sqrt{k \sum_{j\in[k]} N_{i,j}(T)\log(Tnk/\delta)} \\
&= 2 \sum_{i\in[n]} \sqrt{kT\log(Tnk/\delta)} \\
&= 2n\sqrt{kT\log(Tnk/\delta)}
\end{aligned}$$

Where the second line follows from the concavity of $\sqrt{\cdot}$ and the penultimate line follows from the fact that $\sum_{j\in[k]} N_{i,j}(T) = T$. □

G PROOFS ABOUT THE FORMULATION 3 REGRET BOUNDS

Lemma 5.8. *Let $\boldsymbol{p}^* = (\boldsymbol{p}_1^*, \ldots, \boldsymbol{p}_n^*)$ with $\boldsymbol{p}_i^* \in \mathcal{P}^{k-1}$ be the policy that maximizes reward$_2\left(\boldsymbol{p}, v; \frac{\eta}{T}, \gamma\right)$. Then*

$$\text{reward}_2\left(\boldsymbol{p}^*, v; \frac{\eta}{T}, \gamma\right) \geq \text{reward}_3(\pi^*, v; \eta, \gamma).$$

Proof. First, for any arm $j \in [k]$, we can exchange the expectation and the maximum in Equation (8) as follows:

$$\mathop{\mathbb{E}}_{\pi^* v}\left[\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\hat{p}_{i',j} - \hat{p}_{i,j}, 0\right\}\right] \geq \max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\mathop{\mathbb{E}}_{\pi^* v}\left[\hat{p}_{i',j}\right] - \mathop{\mathbb{E}}_{\pi^* v}\left[\hat{p}_{i,j}\right], 0\right\}. \tag{53}$$

Moreover, we can rewrite the expected empirical distribution as follows:

$$\mathop{\mathbb{E}}_{\pi^* v}\left[\hat{p}_{i,j}\right] = \frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi^* v}\left[\mathbf{1}_{\{a_{t,i}=j\}}\right] = \frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi^* v}\left[\pi_i^*(j \mid \boldsymbol{h}_{t-1})\right].$$

Therefore, by Equation (53), we have that

$$\mathop{\mathbb{E}}_{\pi^* v}\left[\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\hat{p}_{i',j} - \hat{p}_{i,j}, 0\right\}\right] \geq \max\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\gamma}{n}\sum_{i'=1}^{n}\mathop{\mathbb{E}}_{\pi^* v}\left[\pi_{i'}^*(j \mid \boldsymbol{h}_{t-1})\right] - \mathop{\mathbb{E}}_{\pi^* v}\left[\pi_i^*(j \mid \boldsymbol{h}_{t-1})\right]\right), 0\right\}.$$

We can therefore bound reward$_3(\pi^*, v; \eta, \gamma)$ as follows:

$$\begin{aligned}
&\text{reward}_3(\pi^*, v; \eta, \gamma)\\
&= \sum_{i=1}^{n}\left(\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi^* v}\left[\boldsymbol{\mu}_i \cdot \boldsymbol{\pi}_i^*(\boldsymbol{h}_{t-1})\right] - \eta\sum_{j=1}^{k}\mathop{\mathbb{E}}_{\pi^* v}\left[\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\hat{p}_{i',j} - \hat{p}_{i,j}, 0\right\}\right]\right)\\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{k}\left(\sum_{t=1}^{T}\mu_{i,j}\mathop{\mathbb{E}}_{\pi^* v}\left[\pi_i^*(j \mid \boldsymbol{h}_{t-1})\right] - \eta\max\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\gamma}{n}\sum_{i'=1}^{n}\mathop{\mathbb{E}}_{\pi^* v}\left[\pi_{i'}^*(j \mid \boldsymbol{h}_{t-1})\right] - \mathop{\mathbb{E}}_{\pi^* v}\left[\pi_i^*(j \mid \boldsymbol{h}_{t-1})\right]\right), 0\right\}\right).
\end{aligned} \tag{54}$$

Define the history-independent policy $\boldsymbol{p} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n)$ such that

$$p_{i,j} = \frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi^* v}\left[\pi_i^*(j \mid \boldsymbol{h}_{t-1})\right].$$

This is a distribution because for any user $i \in [n]$,

$$\sum_{j=1}^{k} p_{i,j} = \frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi^* v}\left[\sum_{j=1}^{k}\pi_i^*(j \mid \boldsymbol{h}_{t-1})\right] = \frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi^* v}[1] = 1.$$

We rearrange Equation (54) to get that

$$\begin{aligned}
&\text{reward}_3(\pi^*, v; \eta, \gamma)\\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{k}\left(\mu_{i,j}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi^* v}\left[\pi_i^*(j \mid \boldsymbol{h}_{t-1})\right] - \eta\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi^* v}\left[\pi_{i'}^*(j \mid \boldsymbol{h}_{t-1})\right] - \frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi^* v}\left[\pi_i^*(j \mid \boldsymbol{h}_{t-1})\right]\right), 0\right\}\\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\left(T\mu_{i,j}p_{i,j} - \eta\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}p_{i',j} - p_{i,j}, 0\right\}\right).
\end{aligned}$$

By definition of $\boldsymbol{p}^*$, this means that

$$\text{reward}_3(\pi^*, v; \eta, \gamma) \le \sum_{i=1}^{n}\sum_{j=1}^{k}\left(T\mu_{i,j}p^*_{i,j} - \eta\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}p^*_{i',j} - p^*_{i,j}, 0\right\}\right) = \text{reward}_2\left(\boldsymbol{p}^*, v; \frac{\eta}{T}, \gamma\right).$$

□

Lemma 5.9. *Let π be any policy such that $\pi_i(t \mid \boldsymbol{h}_{t-1}) = 1$ for all $t \le k$ and $i \in [n]$. For any instance v,*

$$reward_2\left(\pi, v; \frac{\eta}{T}, \gamma\right) \le reward_3(\pi, v; \eta, \gamma) + \eta nk(\gamma+1)\sqrt{\frac{10\log T}{T}}.$$

Proof. First, for any arm $j \in [k]$, we can exchange the expectation and the maximum in Equation (7) as follows:

$$\frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi v}\left[\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\pi_{i'}(j \mid \boldsymbol{h}_{t-1}) - \pi_i(j \mid \boldsymbol{h}_{t-1}), 0\right\}\right] \ge \mathop{\mathbb{E}}_{\pi v}\left[\max\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\gamma}{n}\sum_{i'=1}^{n}\pi_{i'}(j \mid \boldsymbol{h}_{t-1}) - \pi_i(j \mid \boldsymbol{h}_{t-1})\right), 0\right\}\right]$$
$$\ge \max\left\{\mathop{\mathbb{E}}_{\pi v}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\gamma}{n}\sum_{i'=1}^{n}\pi_{i'}(j \mid \boldsymbol{h}_{t-1}) - \pi_i(j \mid \boldsymbol{h}_{t-1})\right)\right], 0\right\}.$$

Using the fact that $\mathbb{E}_{\pi v}[\pi_i(j \mid \boldsymbol{h}_{t-1})] = \mathbb{E}_{\pi v}[\mathbf{1}_{\{A_{t,i}=j\}}]$, we have that

$$\frac{1}{T}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi v}\left[\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\pi_{i'}(j \mid \boldsymbol{h}_{t-1}) - \pi_i(j \mid \boldsymbol{h}_{t-1}), 0\right\}\right] \ge \max\left\{\mathop{\mathbb{E}}_{\pi v}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\gamma}{n}\sum_{i'=1}^{n}\mathbf{1}_{\{A_{t,i'}=j\}} - \mathbf{1}_{\{A_{t,i}=j\}}\right)\right], 0\right\}. \quad (55)$$

Next, we use the fact [4] that

$$\begin{aligned}
&\max\left\{\mathop{\mathbb{E}}_{\pi v}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\gamma}{n}\sum_{i'=1}^{n}\mathbf{1}_{\{A_{t,i'}=j\}} - \mathbf{1}_{\{A_{t,i}=j\}}\right)\right], 0\right\} \\
&\ge \mathop{\mathbb{E}}_{\pi v}\left[\max\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\gamma}{n}\sum_{i'=1}^{n}\mathbf{1}_{\{A_{t,i'}=j\}} - \mathbf{1}_{\{A_{t,i}=j\}}\right), 0\right\}\right] - \sqrt{\frac{1}{2}\cdot\text{Var}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\gamma}{n}\sum_{i'=1}^{n}\mathbf{1}_{\{A_{t,i'}=j\}} - \mathbf{1}_{\{A_{t,i}=j\}}\right)\right)} \\
&= \mathop{\mathbb{E}}_{\pi v}\left[\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\hat{p}_{i',j} - \hat{p}_{i,j}, 0\right\}\right] - \sqrt{\frac{1}{2T^2}\cdot\text{Var}\left(\sum_{t=1}^{T}\left(\frac{\gamma}{n}\sum_{i'=1}^{n}\mathbf{1}_{\{A_{t,i'}=j\}} - \mathbf{1}_{\{A_{t,i}=j\}}\right)\right)}. \quad (56)
\end{aligned}$$

Let $Y_t = \frac{\gamma}{n}\sum_{i'=1}^{n}\mathbf{1}_{\{A_{t,i'}=j\}} - \mathbf{1}_{\{A_{t,i}=j\}}$ and define the martingale difference sequence $D_t := \sum_{\tau=1}^{t}(Y_\tau - \mathbb{E}[Y_\tau])$. Then $\text{Var}[D_T] = \text{Var}\left[\sum_{t=1}^{T}Y_t\right]$ and D_t is martingale with bounded increments $|D_t - D_{t-1}| \le 2(\gamma+1)$. By assumption $\pi_i(t \mid \boldsymbol{h}_{t-1}) = 1$ for all $t \le k$ and $i \in [n]$ so $D_0 = 0$ deterministically. Let B be the event that $|D_T| \le (\gamma+1)\sqrt{8T\log T}$. Applying Azuma-Hoeffding for martingales we know that $\Pr[B^c] \le \frac{1}{T}$. Moreover, with probability 1, $|D_T| \le 2T(\gamma+1)$. Therefore, by the law of total variance and Popoviciu's inequality,

$$\begin{aligned}
&\text{Var}\left[\sum_{t=1}^{T}Y_t\right] \\
&= \text{Var}[D_T] \\
&= \text{Var}[D_T \mid B]\Pr[B] + \text{Var}[D_T \mid B^c]\Pr[B^c] + \left(\mathbb{E}[D_T \mid B]^2 + \mathbb{E}[D_T \mid B^c]^2 - 2\,\mathbb{E}[D_T \mid B]\,\mathbb{E}[D_T \mid B^c]\right)\Pr[B]\Pr[B^c] \\
&\le \text{Var}[D_T \mid B] + \frac{1}{T}\left(\text{Var}[D_T \mid B^c] + \mathbb{E}[D_T \mid B]^2 + \mathbb{E}[D_T \mid B^c]^2 - 2\,\mathbb{E}[D_T \mid B]\,\mathbb{E}[D_T \mid B^c]\right) \\
&\le 2T(\gamma+1)^2\log T + 17T(\gamma+1)^2
\end{aligned}$$

$$\leq 19T(\gamma+1)^2 \log T.$$

Combining this fact with Equations (55) and (56), we have that

$$\frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi\nu}\left[\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\pi_{i'}(j \mid \boldsymbol{h}_{t-1}) - \pi_i(j \mid \boldsymbol{h}_{t-1}), 0\right\}\right] \geq \mathop{\mathbb{E}}_{\pi\nu}\left[\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\hat{p}_{i',j} - \hat{p}_{i,j}, 0\right\}\right] - \sqrt{\frac{10(\gamma+1)^2\log T}{T}}. \quad (57)$$

As a result,

$$\begin{aligned}
\text{reward}_2\left(\pi, \nu; \frac{\eta}{T}, \gamma\right) &= \mathop{\mathbb{E}}_{\pi\nu}\left[\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\boldsymbol{\mu}_i \cdot \boldsymbol{\pi}_i(\boldsymbol{h}_{t-1}) - \frac{\eta}{T}\sum_{j=1}^{k}\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\pi_{i'}(j \mid \boldsymbol{h}_{t-1}) - \pi_i(j \mid \boldsymbol{h}_{t-1}), 0\right\}\right)\right] \\
&\leq \mathop{\mathbb{E}}_{\pi\nu}\left[\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\boldsymbol{\mu}_i \cdot \boldsymbol{\pi}_i(\boldsymbol{h}_{t-1}) - \eta\sum_{j=1}^{k}\max\left\{\frac{\gamma}{n}\sum_{i'=1}^{n}\hat{p}_{i',j} - \hat{p}_{i,j}, 0\right\}\right)\right] + \eta nk(\gamma+1)\sqrt{\frac{10\log T}{T}} \\
&= \text{reward}_3(\pi, \nu; \eta, \gamma) + \eta nk(\gamma+1)\sqrt{\frac{10\log T}{T}}.
\end{aligned}$$

□

Theorem 5.10. *Let π be the policy played by Algorithm 3. Then the regret is bounded as*

$$\textit{reward}_3(\pi^*, \nu; \eta, \gamma) - \textit{reward}_3(\pi, \nu; \eta, \gamma) = \tilde{O}\left(n\sqrt{kT} + \frac{\eta nk(1+\gamma)}{\sqrt{T}}\right).$$

Proof. Let $\boldsymbol{p}^*$ be the policy that maximizes $\text{reward}_2\left(\boldsymbol{p}, \nu; \frac{\eta}{T}, \gamma\right)$. We expand the regret as

$$\begin{aligned}
&\text{reward}_3(\pi^*, \nu; \eta, \gamma) - \text{reward}_3(\pi, \nu; \eta, \gamma) \\
&= \text{reward}_3(\pi^*, \nu; \eta, \gamma) - \text{reward}_2\left(\boldsymbol{p}^*, \nu; \frac{\eta}{T}, \gamma\right) + \text{reward}_2\left(\boldsymbol{p}^*, \nu; \frac{\eta}{T}, \gamma\right) - \text{reward}_3(\pi, \nu; \eta, \gamma) \\
&\leq \text{reward}_2\left(\boldsymbol{p}^*, \nu; \frac{\eta}{T}, \gamma\right) - \text{reward}_3(\pi, \nu; \eta, \gamma) && \text{(Lemma 5.8)} \\
&\leq \text{reward}_2\left(\boldsymbol{p}^*, \nu; \frac{\eta}{T}, \gamma\right) - \text{reward}_2\left(\pi, \nu; \frac{\eta}{T}, \gamma\right) + \eta nk(\gamma+1)\sqrt{\frac{10\log T}{T}} && \text{(Lemma 5.9)} \\
&= O\left(n\sqrt{kT\log(Tnk)} + \sqrt{T\log^2(Tnk)} + \eta nk(\gamma+1)\sqrt{\frac{10\log T}{T}}\right). && \text{(Theorem 5.7)}
\end{aligned}$$

□

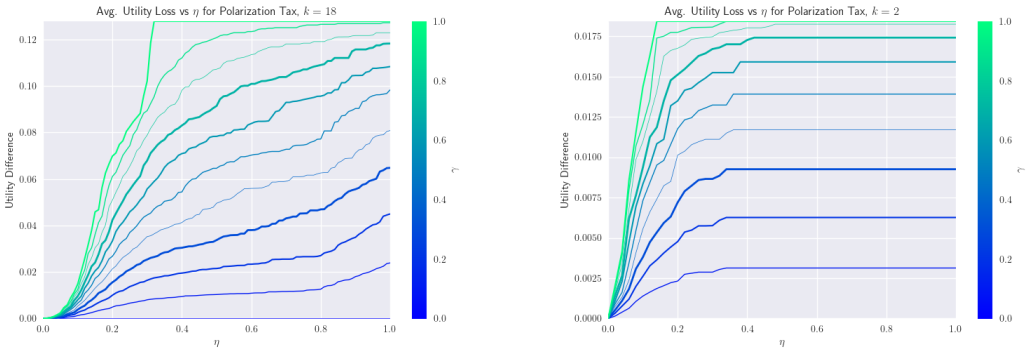
H ADDITIONAL INFORMATION ABOUT THE EXPERIMENTS

Figure 4 plots the change in the additive utility loss

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\mu}_i \cdot \boldsymbol{p}_i^* - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\mu}_i \cdot \boldsymbol{p}_i^{\gamma;\eta}$$

for all genres and the entire population of users described in Section 6.1.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009



(a) All User Types (b) Action and Crime Content Lovers

Fig. 4. Polarization Tax: Utility Difference as function γ and η