Bandit Learning with Biased Human Feedback

Wei Tang[1] and Chien-Ju Ho[1]

[1]Washington University in St. Louis., Email: w.tang@wustl.edu.
[1]Washington University in St. Louis., Email: chienju.ho@wustl.edu.

Abstract

We study a multi-armed bandit problem with biased human feedback. In our setting, each arm is associated with an unknown reward distribution. When an arm is played, a user receives a realized reward drawn from the distribution of the arm. She then provides feedback, a biased report of the realized reward, that depends on both the realized reward and the feedback history of the arm. The learner can observe only the biased feedback but not the realized rewards. The goal is to design a strategy to sequentially choose arms to maximize the total rewards users receive while only having access to the biased user feedback. We explore two natural feedback models. When user feedback is biased only by the average feedback of the arm (i.e., the ratio of positive feedback), we demonstrate that the evolution of the average feedback over time is mathematically equivalent to users performing online gradient descent for some latent function with a decreasing step size. With this mathematical connection, we show that under some mild conditions, it is possible to design a bandit algorithm achieving *regret* (i.e., the difference between the algorithm performance and the optimal performance of always choosing the best arm) sublinear in the number of rounds. However, in another model when user feedback is biased by both the average feedback and the number of feedback instances, we show that there exist no bandit algorithms that could achieve sublinear regret. Our results demonstrate the importance of understanding human behavior when applying bandit approaches in systems with humans in the loop.

1 Introduction

In a multi-armed bandit problem, a learner sequentially selects from a set of arms. Each arm is associated with some unknown reward distribution. After selecting an arm, the learner observes the realized reward for the selected arm. The goal of the learner is to maximize the total rewards obtained from selected arms over time. The performance of the bandit algorithm is often measured in terms of *regret*, defined as the difference between the algorithm performance and the performance of an oracle which can select the best arm in hindsight. The multi-armed bandit formulation provides a theoretical framework for resolving the classical exploration-exploitation tradeoffs in online decision problems under uncertainty. Therefore, multi-armed bandits have been studied in a wide range of applications in various domains, such as medical trials, online auctions, or web advertisements.

We explore the applications of bandits settings to human-in-the-loop systems. For example, consider user-generated content platforms, such as Youtube, Quora, or Stack Exchange. On these

platforms, content qualities vary across a wide spectrum. Ideally, the platform would like to select the best content to display to users to optimize users' experience. However, the content qualities are often not known in advance, and the platform needs to learn the content qualities through user feedback (e.g., number of likes, upvotes, etc). This naturally leads to a bandit problem, in which the platform needs to balance exploration (display content with fewer feedback instances to users to acquire more information) and exploitation (display content with higher empirical ratings to optimize users' happiness), as studied in the literature [14, 20].

Many challenges arise when humans are involved in the bandit learning process. In recent years, researchers have addressed various strategic issues brought up by humans involved in bandit learning [14, 20, 21, 25]. However, in these works, it is assumed that users' feedback is *unbiased* in representing the reward of selecting an arm (e.g., in user-generated content platforms, users' average ratings are used as the estimates for content qualities). On the other hand, as the empirical studies suggest [22, 28, 29], user feedback is often biased by other users' feedback. For example, users have the tendency to provide feedback that agrees with the majority opinion even if their experience disagrees (i.e., the herding effect). These empirical evidences suggest a different stochastic model in that each observed feedback instance might be biased by the feedback history. Moreover, this biased user feedback introduces additional challenges. Since user feedback only represents biased reports of the realized rewards, suppose the goal of the platform is to maximize the total rewards over time (which may be interpreted as the overall user experience), can a platform achieve sublinear regret from only observing biased feedback?

In this paper, we study a variant of the multi-armed bandit problem with human biased feedback. In our setting, the learner/platform only observes human-generated feedback instead of the realized reward when selecting an arm. The human feedback depends on both the realized reward and other users' feedback for the selected arm. The goal of the learner is to maximize the total realized rewards for the selected arms while only having access to biased human-generated feedback.

To address the issues of user biased feedback, we explore two natural user feedback models and study their impacts to the design of bandit algorithms. The first model, avg-herding feedback model, assumes that user feedback for an arm depends on the realized reward and the *average feedback* (i.e., the ratio of positive feedback) of the arm so far. We show that, under this model, the dynamics of user feedback over time is mathematically connected to asymptotic approximation [27]. In particular, the average feedback changes over time as if users are performing online gradient descent on a latent function with a decreasing step size. With this mathematical connection, we characterize the convergence and convergence rates for the average feedback of an arm under some mild conditions. These convergence results enable us to design a bandit algorithm based on UCB (Upper Confidence Bound) algorithm and achieve sublinear regret.

While the results on the first model are promising, our results on another natural model, beta-herding feedback model, paint a very different picture. In this model, user feedback is biased by not only the average feedback in the past, but also the number of feedback instances the arm has received so far. This model captures a natural scenario that users might be biased more heavily if there exists more feedback instances in the history. We show that, under this model, the average feedback of an arm converges to a random variable with non-zero variance. This implies that, even with an infinitely number of feedback instances for the arm, the learner is not able to infer

the expected reward of the arm through observing the average user feedback. We further show that, using arguments from information theory, there exist no bandit algorithms that can achieve sublinear regret when user feedback follows beta-herding feedback model.

We next present a toy example to demonstrate that it is possible to get around the above impossible result by modifying the information structure to break the assumption that users follow beta-herding feedback model. In particular, if the learner is allowed to hide the historical information from a small portion of the users, under some styled user models on how users respond to information structures, it is possible to design an algorithm achieving sublinear regret. This result opens up a potentially interesting line of future research: Can the learner adaptively *design* the information structures to improve the overall utility?

Our results demonstrate the importance of understanding human behavior when learning from human generated feedback. A small deviation on the user behavior model and/or the design of the information structure could have significant impacts on the overall system outcome. Therefore, platforms and decision makers should carefully take these into account when designing learning algorithms in systems with humans in the loop.

2 Related Work

In this section, we review the relevant literature in multi-armed bandit problems, recent studies on human-in-the-loop bandit learning, and the literature on social influences and social learning that share similar motivations of this work.

Multi-armed bandit problems. Our work is a variant of the well-studied multi-armed bandit problem [19]. Bandit problems traditionally assume the rewards generated by each arm at each round are directly observable, and the research focus has been divided into settings in which rewards are either independent and identically distributed (i.i.d.) [4] or adversarial [3, 2]. There exist other works that assume rewards are neither i.i.d. drawn nor adversarial. For example, bandits with Markovian rewards [23, 24] assume the state of each arm evolves according to a Markov process. Other non-stationary bandit problems [6, 12] consider the setting in which the rewards distribution might change over time, independent of previous actions. More recently, researchers have addressed the setting in which the rewards are strategic choices of humans and could be influenced by how the bandit algorithm is designed [14, 20]. Our work differs from the above works in that, in our setting, the "state" (history information) of each arm is correlated with learner's actions and there might be infinitely many states. Moreover, in our setting, the algorithm cannot observe realized rewards but only has access to biased feedback while previous work assume the realized rewards are observable.

Human-in-the-loop bandit learning. Recently, there have been works exploring bandit learning with humans in the loop [18, 10, 21, 25]. In the setting of these works, the learner cannot directly choose which arms to play. Instead, at each time step, a myopic agent, who only aims to maximize her own reward at the single time step she is involved in, chooses which arm to play. Since the agent only cares about her instant payoff, she does not have incentives to explore and tends to always exploit, and this collective arm playing will lead to the convergence to the sub-

optimal arm. Researchers have been attempting to address this problem by considering different ways of *persuading* agents to perform exploration, including offering agents payments to perform exploration [10] or utilizing information asymmetry to lead agents to explore by designing what information to show to each agent [18, 21, 25]. The idea of utilizing information asymmetry to persuade agents is similar to Bayesian Persuasion [17] in economics. The above line of work has focused on settings in which humans are involved in *arm selection*, i.e., which arm is played in each round. In this work, we focus on a parallel aspect of human involvements, in which humans are involved in *feedback generation*.

Social influences and social learning. Our feedback models are motivated by the empirical evidences that users' decisions are influenced by not only their own experience but also other users' decisions [28, 22, 29]. For example, Muchnik et al. [22] empirically show that, a post on a forum is more likely to receive positive feedback (i.e., *upvotes*) if the platform insert an upvote right after the post is made. Similar discussion also appears in the social learning literature in economics [5, 7, 30]. They discuss the setting in which users' decisions might be influenced by other users' decisions. Therefore, under certain conditions, users might collectively make the bad decision since they might follow what other users do regardless of what they privately know. In prior work, there is not much discussion on either the convergence rate of users' aggregate behavior or the impacts on the system designer's perspective. In this work, we focus on deriving the dynamics of user feedback over time and explore the impacts on the design of bandit algorithms.

3 Problem Setting

Let K be the number of arms. Each arm $k \in [K] = \{1, ..., K\}$ is associated with an unknown quality $\theta_k \in [0, 1]$. Let $I^* = \text{argmax}_k \theta_k$ and $\theta^* = \theta_{I^*}$ be the index of the best arm and the associated highest expected quality. At each round t, a user randomly drawn from the population arrives, the learner selects an arm $I_t \in \{1, ..., K\}$ for the arriving user. The user then gets a binary reward Z_t (positive or negative experience) with mean θ_{I_t}.

$$Z_t \sim \texttt{Bernoulli}[\theta_{I_t}]$$

The reward is not observable to the learner. However, after receiving the reward, each user provides a binary feedback $X_t \in \{0, 1\}$ about this arm. The goal of the learner is to maximize the total rewards users receive while observing only the (potentially biased) feedback. Note that when the feedback is the same as the realized reward, i.e., $X_t = Z_t$ for all t, this problem reduces to standard bandit setting. Below we describe the user feedback models, i.e., how X_t is generated.

User feedback models

Users' feedback depends on both the realized rewards and the feedback history of the arms. The feedback history of arm k up to time t can be summarized by $n_{k,t}$ and $\rho_{k,t}$, which represent the number of feedback instances and the ratio of positive feedback for arm k up to round t. We assume $n_{k,0} = \rho_{k,0} = 0$ to simplify the presentation, however, our results can be easily extended to settings with non-zero $n_{k,0}$ and $\rho_{k,0}$, which can be used to represent the users' *prior* of the arm quality. Again, if users provide unbiased feedback, we should have $X_t = Z_t$ for all t.

In this paper, we model the feedback generation as a stochastic process. We define a feedback function to model the probability of obtaining positive feedback for an arm from a user randomly drawn from the population. Note that a feedback function describes the characteristics of the *user population* the platform is interacting with instead of a single specific user. In particular, we introduce $\mathsf{Feedback}(\theta, \rho, n)$ to model the probability of obtaining positive feedback from a user given that the arm quality is θ and the history information of the arm is summarized by its average feedback ρ and the number of feedback instances n.

As a special case, when $\mathsf{Feedback}(\theta, \rho, n) = \theta$, user feedback represents unbiased samples of the arm quality.

In this paper, we explore two natural feedback models.

- Avg-herding feedback model:

  In this feedback model, user feedback is biased by the average feedback of the arm. In particular, the feedback function has the form

  $$\mathsf{Feedback}(\theta, \rho, n) = F(\theta, \rho).$$

  In Section 4, we study the stochastic process of user feedback specified by a general set of feedback functions F. We then discuss the impacts of this stochastic feedback generation on the design of bandit algorithms.

- Beta-herding feedback model:

  In this feedback model, user feedback is biased by the average feedback and the number of feedback instances. In particular, we consider a natural setting and assume users update their beliefs about the arm quality in a Bayesian manner. Users treat the historical ratings as the prior signals of the arm quality and update the posterior based on their own experience. They then provide feedback according to their posterior.

  We introduce a factor $m \geq 0$,which can be interpreted as the weights users put on their own experience. When the arm quality is θ and the arm history is (n, ρ), the expected number of *positive signals* the user will obtain is $m\theta + n\rho$, where the first term is the expected positive signals the users receive from their own experience (i.e., arm quality multiplied by the weight) and the second term is the number of positive signals from other users. The total number of signals is $m + n$.

  Therefore, the probability of obtaining positive feedback for arm k at round t can be written as

  $$\mathsf{Feedback}(\theta, \rho, n) = \frac{m\theta + n\rho}{m + n}. \tag{1}$$

  Note that when $m \to \infty$, user feedback provides unbiased samples of the arm quality.

Regret notions

The goal of the learner is to maximize the sum of rewards users receive over time. Let $\mathcal{A}$ be the algorithm the learner deploys and $\{I_t\}$ are the arms selected by $\mathcal{A}$. We define the *regret* as $R_{\mathcal{A}}(T)$.

$$\mathbb{E}[R_{\mathcal{A}}(T)] = T\theta^* - \mathbb{E}_{\mathcal{A}}\left[\sum_{t=1}^{T} \theta_{I_t}\right],$$

where the expectation is taken over the randomness of the reward realization and the algorithm. In particular, we are interested in the region of $T \to \infty$ and aim to understand under what conditions we can achieve asymptotic sublinear regret, i.e., $\mathbb{E}[R(T)] = o(T)$, when user feedback is biased by historical feedback.

4 Bandits with Avg-Herding Feedback Model

In this section, we explore the bandit learning problem when user feedback follows avg-herding feedback model. We first derive the stochastic process of the feedback generation for a single arm and characterize the convergence and convergence rate of users' average feedback over time. We then discuss how this user feedback model impacts the design and analysis of bandit algorithms.

4.1 Stochastic process of feedback generation

In the following discussion, we explore the feedback dynamics of a single arm, i.e., the stochastic process of feedback generation. We omit the arm's index k in the subscript when it is clear from the context. Also, since user feedback is biased by the history of only the selected arm, to simplify the presentation, we consider the case that the same arm is repeatedly selected and therefore $n_t \equiv t$ when studying the stochastic process for a single arm.

4.1.1 Connection to stochastic approximation

Recall that in avg-herding feedback model, when the quality of the arm is θ and the average feedback of the arm is ρ, the probability for a user to provide a positive feedback is $F(\theta, \rho)$. The stochastic process of the feedback dynamics can be expressed as follows: at the $(t+1)$–th round, the feedback X_{t+1} provided by the user is drawn randomly from a Bernoulli distribution: $\texttt{Bernoulli}[F(\theta, \rho_t)]$. The history information of the arm (n_{t+1}, ρ_{t+1}) are updated based on the realized feedback.

As mentioned, we simplify the presentation by setting $n_t \equiv t$. Therefore, we focus on how ρ_t evolves over time. By simple weighted averaging, we have

$$\rho_{t+1} = \frac{t}{t+1}\rho_t + \frac{1}{t+1}X_{t+1} = \rho_t - \frac{1}{t+1}(\rho_t - X_{t+1}).$$

Define the noise term $\xi_t = \mathbb{E}[X_t|\mathcal{F}_{t-1}] - X_t = F(\theta, \rho_{t-1}) - X_t$, where $\mathcal{F}_t = \sigma(\{X_t\}_{t\geq 1})$ is the filtration of the stochastic process. It is easy to see that $\mathbb{E}[\xi_t|\mathcal{F}_{t-1}] = 0$. Also let $\eta_t = 1/t$ be the step size (learning rate). We can rewrite the above recursive definition as an update rule in stochastic approximation [27, 11].

$$\rho_{t+1} = \rho_t - \eta_{t+1}(\rho_t - F(\theta, \rho_t) + \xi_{t+1}) \tag{2}$$

In particular, suppose there exists a latent function $G(\theta, \rho)$, such that $\partial G/\partial \rho = \rho - F(\theta, \rho)$, then Equation (2) is equivalent to the update rule for stochastic gradient descent with step size η_{t+1}:

$$\rho_{t+1} = \rho_t - \eta_{t+1}(\nabla_\rho G(\theta, \rho_t) + \xi_{t+1})$$

With this observation, the stochastic process of the average feedback updates is equivalent to users collectively performing stochastic gradient descent for a latent function G with a decreasing step size. Below we utilize this mathematical connection and discuss conditions on the

convergence and convergence rates of the average feedback ρ. We then discuss the impacts of this stochastic process on the design and analysis of bandit algorithms.

4.1.2 On the convergence and convergence rate of $\lim_{t\to\infty}\rho_t$.

We first specify the assumptions needed to establish the asymptotic behavior of the limit of average feedback.

A1. $F(\theta,\rho)$ is strictly increasing in θ and non-decreasing in ρ;

A2. $F(\theta,\rho)$ is differentiable and L_F^ρ–Lipschitz continuous with respect to ρ.

A1 implies that, conditional on the same quality (average feedback), an arm with better average feedback (quality) receives more positive feedback in expectation. *A2* assumes the improvement is smooth with respect to ρ. While the differentiable property of F can be satisfied if the population is large and smooth, we note that the differentiable property is only for analytical convenience. Our results still hold even if F is only continuously differentiable in some local neighbourhood of equilibrium points.

We would also like to note that these two assumptions are relatively mild. As an example, below we give a general set of feedback functions F that satisfy the above assumptions.

Example 1. *Consider the following set of feedback functions: $F(\theta,\rho)=w_1\theta+w_2\rho$, for any $w_1,w_2\geq 0$ and $w_1+w_2=1$. This set of feedback functions satisfies both of the assumptions. It also has very natural interpretations. In particular, it specifies that, the probability of receiving a positive feedback from a random user (drawn from the population) $F(\theta,\rho)$ is the weighted average of the arm quality θ and other users' average feedback ρ.*

Armed with the above assumptions, we can formally characterize the convergence of ρ_t.

Lemma 4.1. *Let $\mathcal{S}_\theta:=\{\rho:\rho-F(\theta,\rho)=0\}$. We have $\mathbb{P}(\lim_{t\to\infty}\rho_t\in\mathcal{S}_\theta)=1$.*

The above lemma demonstrates that ρ_t converges to one of the points in a set $\mathcal{S}_\theta$ and characterizes the points in $\mathcal{S}_\theta$. Recall that the latent function $G(\theta,\rho)$ satisfies $\partial G/\partial\rho=\rho-F(\theta,\rho)$. Therefore, the lemma illustrates that the average feedback will converge to one of the points in $\mathcal{S}_\theta$, the set of the local optimal points for the latent function G. This intuition suggests that, when the latent function G is strongly convex, since there exists only one local optimal point (which is the global optimal), we should be able to show that ρ_t will almost surely converge to the global optimal.

Moreover, the convexity of G is correlated with the value of the Lipschitz constant L_F^ρ. In particular, when $L_F^\rho<1$, by definition, we have $\nabla_\rho F(\theta,\rho)<1$ for all θ and ρ. Since $\nabla_\rho^2 G(\theta,\rho)=1-\nabla_\rho F(\theta,\rho)$, when $L_F^\rho<1$, $\nabla_\rho^2 G(\theta,\rho)>0$ for all θ and ρ. Therefore, G is strongly convex when $L_F^\rho<1$. Below we formally characterize the convergence of ρ_t when G is strongly convex.

Corollary 4.2. *Given $L_F^\rho<1$, i.e., G is strongly convex, there exists a unique ρ^* that satisfies $\rho^*-F(\theta,\rho^*)=0$, such that $\mathbb{P}(\lim_{t\to\infty}\rho_t=\rho^*)=1$.*

Next we provide the results on the convergence rate of ρ_t and focus on the case when G is strongly convex. In particular, we introduce $\bar{\lambda}>0$, such that $\nabla_\rho^2 G\geq\bar{\lambda}>0$.

Theorem 4.3. *Given $L_F^\rho < 1$, i.e., G is strongly convex. $\forall \epsilon > 0$, we have,*

$$\mathbb{P}(|\rho_t - \rho^*| \geq \epsilon) \leq \exp\left(-\frac{(\epsilon - \epsilon_t)^2}{2\sum_{i=1}^t L_i}\right),$$

where $L_i = \eta_i^2(\prod_{j=i}^{t-1}(\eta_{j+1}^2(L_F^\rho - 1)^2) - 2\bar{\lambda}\eta_{j+1} + 1)$, $\epsilon_t = \exp(-\bar{\lambda}S_t)|\rho_0 - \rho^*| + \sqrt{\sum_{i=0}^{t-1}\eta_{i+1}^2\exp(-2\bar{\lambda}(S_t - S_{i+1}))}$, *and* $S_t = \sum_{i=1}^t \eta_i$.

Proof Sketch. In the proof, we decompose $|\rho_t - \rho^*|$ into two terms, with the first term corresponding to the *empirical iterate error*, i.e., the difference between the absolute value of error at a given time and its mean, and the second term corresponding to the *expectation error*:

$$|\rho_t - \rho^*| = (|\rho_t - \rho^*| - \mathbb{E}[|\rho_t - \rho^*|]) + \mathbb{E}[|\rho_t - \rho^*|] \tag{3}$$

To derive the probabilistic tail bound for the $|\rho_t - \rho^*|$, we bound the empirical iterate error $|\rho_t - \rho^*| - \mathbb{E}[|\rho_t - \rho^*|]$ and the expectation error $\mathbb{E}[|\rho_t - \rho^*|]$ separately using martingale concentration bounds. □

Remark 4.4. *We would like to offer a few observations to help interpret the convergence bound*[1]*. In particular,*

- *when* $t \to \infty$, $\epsilon_t \to 0$,
- *when* $\bar{\lambda} \in (0, 1/2)$, $\sum_{i=1}^t L_i = \mathcal{O}(t^{-2\bar{\lambda}})$, *and*
- *when* $\bar{\lambda} \in [1/2, \infty)$, $\sum_{i=1}^t L_i = \mathcal{O}(1/t)$.

So we can characterize the bound in two regions based on whether $\bar{\lambda} \geq 1/2$. As a special case, when user feedback is unbiased, i.e., $F(\theta, \rho) = \theta$, we have $\bar{\lambda} = 1$, and the bound reduces to $\mathbb{P}(|\rho_t - \rho^*|) \geq \epsilon) \leq \mathcal{O}(e^{-\epsilon^2 t})$, the same as the standard Chernoff bound. Moreover, in our setting, since $F(\theta, \rho)$ is non-decreasing in ρ, i.e., $\nabla_\rho F \geq 0$. We have $\nabla_\rho^2 G = 1 - \nabla_\rho F \leq 1$. Therefore, while our bound holds for the region $\bar{\lambda} \in (0, \infty)$, in our setting, we focus on the region $\bar{\lambda} \in (0, 1]$.

Note that in this theorem, the convergence rate is a function of $\bar{\lambda}$, which is the property of the function G (hence the property of the feedback model F). As an intuitive interpretation, recall that $\nabla_\rho^2 G \geq \bar{\lambda}$ and $\nabla_\rho G = \rho - F(\theta, \rho)$. Therefore, small $\bar{\lambda}$ implies large $\partial F/\partial\rho$, which means user' feedback is influenced more by the other users' feedback and relatively less by the arm quality. When users' feedback depends less on the arm quality, it requires more feedback to infer the arm quality, and therefore the convergence is slower. This intuition aligns with the theorem, in which smaller $\bar{\lambda}$ leads to a slower convergence rate.

4.2 Designing bandit algorithms

Given the convergence bound in Theorem 4.3, we can design a UCB-like algorithm that achieves sublinear regret. We assume the learner has knowledge of the feedback model F. Note that since F models the behavior of feedback generation for the *user population* the platform is interacting

[1] The detailed derivations are included in the appendix of the full paper.

with, this assumption only requires the platform to have knowledge of the population instead of any particular users[2].

In each round of our algorithm, the learner maintains an estimator $\hat{\theta}_{k,t}$ of arm k's quality from the observation of average feedback $\rho_{k,t}$. From Lemma 4.1, an asymptotically unbiased and consistent estimator of arm's quality $\hat{\theta}_{k,t}$ can be obtained by solving the following equation.

$$\hat{\theta}_{k,t} = \max\{\min(\{\hat{\theta}_{k,t} : F(\hat{\theta}_{k,t}, \rho_{k,t}) = \rho_{k,t}\}, 1), 0\} \tag{4}$$

Intuitively, the solutions of the above equation represent the set of local optimal points of G. Moreover, we can show that the estimator $\hat{\theta}_{k,t}$ is unique for every $\rho_{k,t}$ if *A1* is satisfied.

Lemma 4.5. *Suppose A1 is satisfied, for any $\rho_{k,t}$, there exists a unique $\hat{\theta}_{k,t}$ that satisfies Equation (4).*

Given the convergence bounds and the estimator $\hat{\theta}_{k,t}$, we are ready to describe our proposed UCB-like algorithm Avg-UCB, as specified in Algorithm 1. The key differences to the standard UCB algorithms are that: First, we maintain a quality estimate $\hat{\theta}_{k,t}$ for each arm k at each time t by solving Equation (4) instead of using empirical average feedback. Second, the confidence interval in the UCB index is derived from the convergence rates as specified in Theorem 4.3. Our algorithm takes as input parameters β and $\bar{\lambda}$. β plays a similar role as the constant in UCB confidence radius to balance exploration and exploitation. $\bar{\lambda}$ is the parameter of the problem instance. Note that our algorithm only requires to find some $\bar{\lambda}$ such that $\nabla^2_\rho G \geq \bar{\lambda}$.

Algorithm 1 Avg-UCB for Avg-Herding Feedback Model

Input: β, $\bar{\lambda}$, K.
Initializations: first K rounds, play each arm once
for $t = K+1, ..., T$ **do**
  for each $k \in \{1, ..., K\}$ **do**
    Compute $\hat{\theta}_{k,t-1}$ from (4).
    $\mathtt{UCB}_{k,t} = \hat{\theta}_{k,t-1} + \sqrt{\frac{\beta \ln(t-1)}{n_{k,t-1}^{\min\{1, 2\bar{\lambda}\}}}}$.
  Choose arm $I_t \in \operatorname{argmax}_{k=1,...,K} \mathtt{UCB}_{k,t}$.
  (Ties are broken in some consistent way)
  Receive feedback X_t.
  $\rho_{I_t,t} \leftarrow (\rho_{I_t,t-1} \times n_{I_t,t-1} + X_t)/(n_{I_t,t-1} + 1)$
  $\rho_{k,t} = \rho_{k,t-1}$, $\forall k \neq I_t$.
  $n_{I_t,t} \leftarrow n_{I_t,t-1} + 1$
  $n_{k,t} \leftarrow n_{k,t-1}$, $\forall k \neq I_t$.

The following theorem gives the regret bound for the algorithm Avg-UCB.

[2]In practice, this assumption can be approximately satisfied through market research or behavioral experiments, which study the connection between users' real experience (i.e., Z_t) and reported feedback (i.e., X_t). Moreover, our results are robust to small estimation noises of F.

Theorem 4.6. *Suppose A1 and A2 are satisfied and $L_F^\rho < 1$. Let $\bar{\lambda}' = \max\{1, 1/(2\bar{\lambda})\}$, $\Delta_k = \theta^* - \theta_k$. With appropriately chosen β [3] the expected regret for Avg-UCB is bounded by:*

$$\mathbb{E}[R(T)] \leq \sum_{k \neq I^*} \Delta_k (4 \ln T / (C\Delta_k^2))^{\bar{\lambda}'} + K\pi^2/6,$$

where C is a constant that is dependent on the properties of feedback function F.

We introduce an additional notion $\bar{\lambda}' = \max\{1, 1/(2\bar{\lambda})\}$ to simplify the presentation due to the different convergence rates on whether $\bar{\lambda} < 1/2$ as discussed in Remark 4.4. Similar to the discussion on the convergence rate, the dependency of the above upper regret bound on $\bar{\lambda}'$ implies that it is harder to learn the quality of an arm if users are biased more by the historical information rather than the arm quality.

The above regret bound is a gap-dependent bound. In particular, let $\Delta_{\min} = \min_{k:k\neq I^*} \Delta_k$. The regret bound can be written as: $\mathbb{E}[R(T)] = \mathcal{O}\left(\frac{(\ln T)^{\bar{\lambda}'}}{\Delta_{\min}^{2\bar{\lambda}'-1}}\right)$. Observe that $\lim_{T\to\infty} \mathbb{E}[R(T)]/T \to 0$ for any $\bar{\lambda}' > 0$. Therefore, the algorithm achieves sublinear regret as long as G is strongly convex (i.e., $\bar{\lambda}' > 0$).

Moreover,we can derive gap-independent bounds from the above bound. When $\bar{\lambda} \geq 1/2$ (which includes the unbiased feedback setting with $\bar{\lambda} = 1$), we can show that $\mathbb{E}[R(T)] = \mathcal{O}(\sqrt{T \ln T})$, which matches the standard regret bound without biased feedback.

What if G is not convex. Our algorithm relies on the assumption that the latent function G is convex, i.e., $L_F^\rho < 1$. This assumption implies that users' feedback is not influenced too heavily by the change of feedback history. While this assumption seems mild, it is natural to wonder whether we can obtain similar results when G is not convex.

We would like to note that even in settings when G is non-convex, the statements of Lemma 4.1 and 4.5 still hold. This means the average user feedback for each arm still converges to some point, and we can infer the arm quality from the converged average feedback. The main obstacle to overcome is to derive the convergence rate as in Theorem 4.3. This problem is challenging as it is equivalent to deriving the convergence rate of optimization for non-convex functions. There have been recent works focusing on deriving the convergence rates in non-convex optimization in different settings [1, 13]. As long as one could characterize the convergence rate of ρ_t for non-convex function G, our bandit strategy can be adapted to generate a sublinear regret strategy (by changing the "confidence interval" in the UCB index based on the derived convergence rate).

5 Bandits with Beta-Herding Feedback Model

In the previous section, we explore avg-herding feedback model, in which user feedback is biased only by the average feedback of the selected arm. We show that, under some mild conditions, the average feedback for an arm almost surely converges to some value, and we can infer the arm quality from the average feedback, and therefore we can design a UCB-like algorithm for achieving sublinear regret.

[3] The choice of β depends on the parameters of $F(\theta, \rho)$. The detailed derivation is tied with the proof and is included in the appendix of the full paper.

However, in some scenarios, user feedback may be biased by not only the average feedback but also the number of feedback instances of the arm. In this section, we explore another natural feedback model, beta-herding feedback model, and prove impossibility results. In particular, we assume users give feedback in a Bayesian manner. They treat the feedback history as the prior, i.e., for an arm with history (n, ρ), there are $n\rho$ positive signals and $n(1-\rho)$ negative signals for the arm. After they experience the binary reward (drawn according to the arm's quality distribution), they update their posterior by treating their experience as m signals and then provide feedback according to the posterior. Therefore, in expectation, the probability for them to provide positive feedback for an arm with quality θ and history (n, ρ) is $\texttt{Feedback}(\theta, \rho, n) = (m\theta + n\rho)/(m+n)$.

5.1 Stochastic process of feedback generation

The first natural attempt is to replace $F(\theta, \rho_t)$ with $\texttt{Feedback}(\theta, \rho, n)$ in Equation (2) and apply similar analysis using stochastic approximation. However, when $\texttt{Feedback}(\theta, \rho, n)$ follows beta-herding feedback model, one can not directly apply this approach. Briefly speaking, the update rule in Equation (2) aims to find the equilibrium points of the feedback function. However, in beta-herding feedback model, the feedback function is changing over time, and it is not trivial whether the converged points satisfy the set of properties as derived with avg-herding feedback model.

Instead, we make the observation that the stochastic process of beta-herding feedback model is similar to the urn process [15]. We utilize the property of *exchangeability* for the feedback history to give the characterization of ρ_t process. Below we formally characterize the stochastic process of ρ_t with beta-herding feedback model.

Lemma 5.1. *Consider the stochastic process in Equation (2) with the feedback model described in Equation (1), $\lim_{t\to\infty} \rho_t$ converges almost surely to a random variable specified by a beta distribution. In particular,*

$$\lim_{t\to\infty} \rho_t \sim \text{Beta}(m\theta, m(1-\theta)).$$

Proof Sketch. Let $S_t = \sum_{i=1}^{t} x_i$, where x_i is the realization of the feedback random variable X_i. It is easy to show that the sequence of random variables X_i satisfies the exchangeable property. Therefore, the probability that $S_t = l$, i.e., there are l positive feedback among t feedback, can be written as

$$\frac{\prod_{i=0}^{l-1}(m\theta + i) \cdot \prod_{j=0}^{t-1-l}(m(1-\theta) + j)}{\prod_{i=0}^{t-1}(m+i)}.$$

By Stirling's approximation, when $t \to \infty$, we have

$$\mathbb{P}(S_t = l) = \frac{l^{m\theta-1}}{B(m\theta, m(1-\theta))} \cdot (t-l)^{m(1-\theta)-1} \cdot t^{1-m},$$

where $B(\cdot)$ is the Beta function. Denote $l = \rho t$ for some $0 < \rho < 1$,

$$\mathbb{P}(\frac{S_t}{t} \le \rho) = \sum_{i=0}^{\lfloor t\rho \rfloor} \mathbb{P}(\frac{S_t}{t} = \frac{i}{t})$$

When $t \to \infty$, the summation can be written as an integral

$$\mathbb{P}(\frac{S_t}{t} \leq \rho) = t \int_0^{\rho} \mathbb{P}(\frac{S_t}{t} = u) du.$$

Plug in the above $\mathbb{P}(S_t = l)$ expression and replace l with ρt,

$$\mathbb{P}(\frac{S_t}{t} \leq \rho) = \frac{1}{B(m\theta, m(1-\theta))} \int_0^{\rho} u^{m\theta-1}(1-u)^{m(1-\theta)-1} du,$$

which is the CDF of the beta distribution. This completes the proof. □

Note that when the feedback is unbiased, i.e., when $m \to \infty$, the beta distribution will shrink to a Dirac delta function which has the point mass exactly in θ.

5.2 The impossibility result

In this section, we show that there exist no bandit algorithms that achieve sublinear regret if user feedback follows beta-herding feedback model.

Lemma 5.1 implies that, even if we obtain an infinite number of feedback instances for an arm, we cannot accurately infer the arm quality with high probability from the empirical average feedback ρ_∞. A natural next question to ask is, if we take into account all the feedback generated in the process, whether it is possible to infer the true arm quality. Below we use the notion of Fisher information to answer the question. In short, Fisher information provides a way to quantify the amount of information about the latent parameter θ we can obtain for observing each sample of a random variable X_i. Since Fisher information is additive, we can show that,

Lemma 5.2. *Consider the stochastic process in Equation (2) with the feedback model described in Equation (1). Let $\mathcal{I}_t(\theta)$ denote the Fisher information of θ for observing $t-$th sample. We have*

$$\lim_{t\to\infty} \sum_{i=1}^{t} \mathcal{I}_i(\theta) = \mathcal{O}(1).$$

Proof Sketch. Let $f(x|\theta)$ be the probability mass function of random variable X and x_t be the realization of X_t. Considering the stochastic process specified in Equation (1),

$$f(x_t|\theta) = (\frac{m\theta + S_{t-1}}{m+t-1})^{x_t} \cdot (1 - \frac{m\theta + S_{t-1}}{m+t-1})^{1-x_t},$$

where $x_t = 1$ or $x_t = 0$, $S_t = \sum_{i=1}^{t} X_i$. Define $l(x_t|\theta) = \log f(x_t|\theta)$. By definition of Fisher information for a single observation, and the chain rule for multiple observations, we have:

$$\begin{aligned}\sum_{i=1}^{t} \mathcal{I}_i(\theta) &= \sum_{i=1}^{t} -\mathbb{E}[l''(x_i|\theta)] \\ &= \sum_{i=1}^{t} \frac{m^2}{m+i-1} (\mathbb{E}[\frac{1}{m\theta+S_{i-1}}] + \mathbb{E}[\frac{1}{m(1-\theta)+i-1-S_{i-1}}])\end{aligned}$$

Since we know that $\{X_t\}_{t\geq 1}$ are exchangeable random variables,

$$\begin{aligned}\mathbb{E}[\frac{1}{m\theta+S_t}] &= \sum_{l=0}^{t}\binom{t}{l}\frac{\prod_{i=0}^{l-1}(m\theta+i)\cdot\prod_{j=0}^{t-l-1}(m(1-\theta)+j)}{\prod_{i=0}^{t-1}(m+i)}\cdot\frac{1}{m\theta+l}\\ &= \mathcal{O}(t^{-1})\end{aligned}$$

Similarly, we also have:

$$\mathbb{E}[\frac{1}{m(1-\theta)+t-1-S_{t-1}}] = \mathcal{O}(t^{-1})$$

Thus:

$$\begin{aligned}\lim_{t\to\infty}\sum_{i=1}^{t}\mathcal{I}_i(\theta) &= \lim_{t\to\infty}\sum_{i=1}^{t}\frac{m^2}{m+i-1}\mathcal{O}(i^{-1})\\ &= \mathcal{O}(1)\end{aligned}$$

where $\mathcal{O}(1)$ is a constant. This completes the proof. □

Using this fact, by the general Cramér-Rao bound, we know that, for any estimator $\hat{\theta}_t$, the variance of $\hat{\theta}_t$ must follow:

$$\mathrm{Var}(\hat{\theta}_t) \geq \Theta\left(\frac{1}{\sum_{i=1}^{t}\mathcal{I}_i(\theta)}\right)$$

Since $\lim_{t\to\infty}\sum_{i=1}^{t}\mathcal{I}_i(\theta)$ is bounded, the variance of any estimator will not shrink to zero even with infinitely many observations. Therefore, the learner cannot accurately infer the arm quality with high probability in the beta-herding feedback model and therefore cannot guarantee to identify the best arms even with infinitely many feedback instances. Since the learner only observes the feedback, we can conclude the following.

Theorem 5.3. *If users' feedback follows beta-herding feedback model, there exists no bandit algorithm that can achieve sublinear regrets in our setting.*

Proof Sketch. We prove this by contradiction. Consider the case with two arms. Without loss of generality, assume arm 1 is optimal and arm 2 is suboptimal, i.e., $\theta_1 > \theta_2$, and suppose there exists an algorithm $\mathcal{A}$ which can achieve sublinear regret, i.e., $\mathbb{E}(R_{\mathcal{A}}(T)) = o(T)$. Let k_t denote the arm chosen by algorithm $\mathcal{A}$ at time t. One must have $\lim_{t\to\infty}\mathbb{P}(k_t = 1) = 1$. Let $\hat{\theta}_1^t, \hat{\theta}_2^t$ be the algorithm's estimators on θ_1, θ_2 given the history information accumulated till time round t. The ability to almost surely choose arm 1 by algorithm $\mathcal{A}$ when $t \to \infty$ indicates that we are able to differentiate the two arms, i.e.,

$$\lim_{t\to\infty}\mathbb{P}(\hat{\theta}_1^t > \hat{\theta}_2^t) = 1$$

However, as shown in Lemma 5.2, since the fisher information on the estimator are always bounded even when given infinitely many observations. It implies the estimators are not consistent, and that $\lim_{t\to\infty}\mathbb{P}(\hat{\theta}_1^t < \hat{\theta}_2^t) > 0$. This leads to the contradiction and completes the proof. □

We note that the technique used in the proof can be extended to a more general feedback model for impossibility results. The intuition is to use Fisher information to quantify how informative a given data is with respect to a set of parameters and the influence of the data itself on the estimate. For different models, if the amount information for each feedback can be quantified, the same techniques can be applied.

5.3 An alternative approach: Designing information structures

Theorem 5.3 presents a strong impossibility result: if all feedback instances are generated according to beta-herding feedback model, we cannot design any bandit algorithms to achieve sublinear regret. A natural approach to get over this impossibility results is to break the assumption by taking interventions. Inspired by Bayesian persuasion [17], which designs the information structure to *persuade* agents to take certain actions, we explore whether we could design information structures to induce certain types of "feedback". For example, in the extreme case, if we do not show any historical information to users, and assume users provide unbiased feedback when no information is presented, then the problem reduces to standard bandit settings. However, in practice, we might not want to dramatically change the whole platform and might want to take as few interventions as possible. This leads to an interesting research question on whether we can minimally intervene the existing design of information structure, such that it is possible to design bandit algorithms with sublinear regrets.

In this section, we present a simple algorithm as a *toy example* to demonstrate the idea. A full study along this direction requires a careful and thorough modeling and is out of the scope of this paper. We consider the constrained setting in which the platform can only choose among two information design in each round, either showing all history information to users (and assuming users' feedback follow beta-herding feedback model) or showing no history information (and assuming users provide unbiased feedback). Our goal is to minimize the number of rounds that show no information to users while achieving sublinear regret. In particular, we propose a *two-stage policy*, as described in Algorithm 2, which shows no historical information for the first $\lfloor T^\alpha \rfloor$ rounds and resumes to standard design afterwards.

The regret bound of Algorithm 2 is given as follows.

Algorithm 2 *two-stage policy*

Input: learning rounds parameter $\alpha \in (0,1)$, exploration parameter $\beta > 0$, number of arms K.
Initializations: first K rounds, play each arm once
for $t = K+1, \ldots, \lfloor T^\alpha \rfloor$ **do**
    for each $k \in \{1, \ldots, K\}$ **do**
        $\mathsf{UCB}_{k,t} = \hat{\theta}'_{k,t-1} + \sqrt{\frac{\beta \ln(t-1)}{n_{k,t-1}}}$, where $\hat{\theta}'_{k,t-1} = \frac{\sum_{s=1}^{n_{k,t-1}} \mathbb{1}\{I_s = k\} X_{t-1}}{n_{k,t-1}}$.
    Choose arm $I_t \in \operatorname{argmax}_{k=1,\ldots,K} \mathsf{UCB}_{k,t}$.
    Present arm I_t without showing its history information to the user, and get feedback X_t.
    $\rho_{I_t,t} \leftarrow (\rho_{I_t,t-1} \times n_{I_t,t-1} + X_t)/(n_{I_t,t-1} + 1)$.
    $\rho_{k,t} = \rho_{k,t-1}$ for $k \neq I_t$.
    $n_{I_t,t} \leftarrow n_{I_t,t-1} + 1$.
    $n_{k,t} \leftarrow n_{k,t-1}$ for $k \neq I_t$.
Let $I_\tau \in \operatorname{argmax}_{k=1,\ldots,K} n_{k,\lfloor T^\alpha \rfloor}$.
Present arm I_τ with associated history information to the user in the remaining rounds.
(all ties broken in some consistent way)

Theorem 5.4. *Let $\Theta = \{\theta_1, \ldots, \theta_K\}$ be a bandit instance, and $\alpha \geq \ln(K(K+2))/\ln T$, then the expected*

regret of two-stage policy, where $\beta > 1$, is bounded from above by:

$$\mathbb{E}[R(T)] \leq \sum_{k \neq I^*} \left(\frac{4\alpha\beta \ln T}{\Delta_k} + 8\beta\Delta_k \right) + (T - T^\alpha) \left(\sqrt{\frac{4K\alpha\beta \ln T}{T^\alpha - K}} + \frac{K}{\beta - 1} \left(\frac{T^\alpha - K}{K} \right)^{2-2\beta} \right)$$

where the second term is in an order of $\mathcal{O}\left((T - T^\alpha) \sqrt{\frac{K\beta\alpha \ln T}{T^\alpha}} \right)$.

To interpret the bound, when $\alpha \geq 1/2$, the above regret bound is in the order of $\mathcal{O}(\sqrt{\alpha T^\alpha \ln T})$, while when $\alpha < 1/2$, the above regret bound is in the order of $\mathcal{O}(\sqrt{\alpha T^{1-\alpha} \ln T})$.

Algorithm 2 presents an example that we can achieve sublinear regrets by modifying the information structures presented to users. In particular, we only need to hide the historical information from T^α users, with $\alpha < 1$, out of T users to achieve subliner regrets. Note that we only consider a naive approach in a styled model, i.e., showing no information at all in some rounds, and assume simple user feedback models. We hope our results will encourage research that considers more fine-tuned information design and more thorough models of user feedback and platform utility.

6 Discussion on the Applications

In this section, we provide discussion on the applications of our setting. As the motivating example of this paper, we consider user-generated content platforms that need to learn content qualities through user feedback. Our analysis and results naturally extend to platforms that rely on user reviews to provide recommendations (such as Yelp or Amazon). However, to formulate the recommendation problem as a bandit learning problem, we need to make a simplifying assumption, as made in prior work [14, 20], that users are going to *follow* the recommendations. While this assumption seems strong, in practice, it approximates users' behavior to a certain degree. In particular, empirical studies demonstrate that the probability for a users to view an item drops significantly when the position of the item decreases [26, 9, 16]. These empirical observations suggest that a significant amount of users are indeed following recommendations (since recommended items are ranked higher). Moreover, there have been recent studies on incentivizing exploration using information asymmetry [18, 21, 25] which demonstrate it is possible to make recommendations that users will *choose* to follow. The techniques in this paper can be applied in that line of work to explore the dynamics of feedback generation.

In addition to the above example, our setting applies to scenarios when the platform cannot observe the true objective but can only use (potentially biased) estimates as the proxy for the objective. Consider the following illustrating scenario: the police station needs to decide which area to send police officers to patrol at each time step. Each area i has an intrinsic, unknown crime rate p_i. When sending police officers to an area i, the police station obtains an unobserved reward $u(p_i)$, representing the value of increased safety for the area. Assume $u(p_i)$ is increasing in p_i. After the patrol, police officers need to report the amount of criminal activities during their patrol. However, these reports might be biased by the history of *reported* crime rate of the area. For example, if there are more reports of illegal activities in the area in the history, they might stop more people for inspection. This creates biases in the reports. If the goal is to maximize the sum of $u(p_i)$, this problem can be formulated using our setting, since the objective is a function

of *true* crime rates, while the decision maker only has access to *reported* crime rates. Now assume the feedback model follows beta-herding feedback model. According to our results, without additional interventions, the police station might make *unfair* decisions in where to patrol using only the biased feedback, since it is impossible for them to infer the true crime rate from the reports. This example further emphasizes the importance of understanding human behavior in learning problems, especially when the corresponding actions have significant impacts on humans.

7 Conclusion and Future Work

We explore bandit problems with biased human feedback under two different feedback models. In avg-herding feedback model, where users' feedback is biased only by the average feedback of the arm, we show that the updates of average feedback over time is mathematically equivalent to users collectively performing stochastic gradient descent. With this connection, we design a UCB-like algorithm that achieves sublinear regret under some mild conditions. However, in beta-herding feedback model, where users' feedback is biased both by the average feedback and the number of feedback instances of the arm, using arguments from information theory, we show that there exist no bandit algorithms that can achieve sublinear regret.

We hope our work will open more discussion on better understanding human behavior when designing algorithms for systems with humans in the loop. Our results also point to potentially future research directions on designing interfaces (e.g., in terms of how information is exchanged) between humans and machine learning algorithms to leverage the power of both ends.

References

[1] Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning (ICML)*, pages 699–707, 2016. 10

[2] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *The 22nd Conference on Learning Theory (COLT)*, pages 217–226, 2009. 3

[3] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 322–331. IEEE, 1995. 3

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

[5] Abhijit V Banerjee. A simple model of herd behavior. *The quarterly journal of economics*, 107(3):797–817, 1992. 4

[6] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems (NIPS)*, pages 199–207, 2014. 3

[7] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5):992–1026, 1992. 4

[8] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011. 32

[9] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on Web Search and Data Mining (WSDM)*, pages 87–94. ACM, 2008. 15

[10] Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation (EC)*, 2014. 3, 4

[11] Noufel Frikha, Stéphane Menozzi, et al. Concentration bounds for stochastic approximations. *Electronic Communications in Probability*, 17, 2012. 6

[12] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, pages 174–188, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-24412-4. 3

[13] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsâĂŤonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory (COLT)*, pages 797–842, 2015. 10

[14] Arpita Ghosh and Patrick Hummel. Learning and incentives in user-generated content: Multi-armed bandits with endogenous arms. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science (ITCS)*, 2013. 2, 3, 15

[15] Bruce M Hill, David Lane, and William Sudderth. A strong law for some generalized urn processes. *The Annals of Probability*, pages 214–226, 1980. 11

[16] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, volume 51, pages 4–11. ACM, 2017. 15

[17] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6): 2590–2615, 2011. 4, 14

[18] Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce (EC)*, pages 605–606, 2013. 3, 4, 15

[19] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985. 3

[20] Yang Liu and Chien-Ju Ho. Incentivizing high quality user contributions: New arm generation in bandit learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2, 3, 15

[21] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation (EC)*, pages 565–582. ACM, 2015. 2, 3, 4, 15

[22] Lev Muchnik, Sinan Aral, and Sean J Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013. 2, 4

[23] Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1804–1812, 2010. 3

[24] Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless markov bandits. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 214–228. Springer, 2012. 3

[25] Yiangos Papanastasiou, Kostas Bimpikis, and Nicos Savva. Crowdsourcing exploration. *Management Science*, 2017. 2, 3, 4, 15

[26] Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web (WWW)*, pages 521–530. ACM, 2007. 15

[27] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 2, 6

[28] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006. 2, 4

[29] Ruben Sipos, Arpita Ghosh, and Thorsten Joachims. Was this review helpful to you?: It depends! context and voting patterns in online content. In *Proceedings of the 23rd International Conference on World Wide Web (WWW)*, pages 337–348, 2014. 2, 4

[30] Lones Smith and Peter SÃ ÿrensen. Pathological outcomes of observational learning. *Econometrica*, 68 (2):371–398, 2000. 4

A Useful Lemmas

In this section, we list some useful lemmas in our analysis.

Lemma A.1. ***(Stirling's approximation for gamma function quotient)*** *Given $\alpha > 0, \beta > 0$ and when $x \to \infty$, we have*

$$\frac{\Gamma(x+\beta)}{\Gamma(x+\alpha)} = x^{\beta-\alpha}.$$

Proof. From Stirling's Approximation, and let $\gamma = \beta - \alpha$

$$\begin{aligned}\frac{\Gamma(x+1+\beta)}{\Gamma(x+1+\alpha)} &= \frac{\sqrt{2\pi(x+\beta)}(\frac{x+\beta}{e})^{x+\beta}}{\sqrt{2\pi(x+\alpha)}(\frac{x+\alpha}{e})^{x+\alpha}} \\ &= (1+\frac{\gamma}{x+\alpha})^{x+\alpha+1/2}(1+\frac{\beta}{x})^{\gamma}(\frac{x}{e})^{\gamma}.\end{aligned}$$

Since $\lim_{x\to\infty}(1+y/x)^x = e^y$ and $\lim_{x\to\infty}(1+y/x) = 1$, we have

$$\lim_{x\to\infty}\frac{\Gamma(x+1+\beta)}{\Gamma(x+1+\alpha)} = x^{\beta-\alpha}.$$

□

Lemma A.2. ***(Concentration property for bounded random variable)*** *Given any bounded random variable Y, and a L–Lipschitz function $g(\cdot)$, then for $\forall\lambda \in \mathbb{R}$,*

$$\mathbb{E}[\exp(\lambda g(Y))] \le \exp(\lambda^2 L^2/2).$$

B Proofs and Simulations in Bandits with Avg-Herding Feedback Model

B.1 Proof of Lemma 4.1

Proof. Our goal is to prove $\mathbb{P}(\lim_{t\to\infty}\rho_t \in \mathcal{S}_\theta) = 1$ for $\mathcal{S}_\theta := \{\rho : \rho - F(\theta,\rho) = 0\}$. Since $F(\theta,\rho)$ is continuous, for all $\epsilon > 0$, we define the following two sets:

$$\begin{aligned}U_\epsilon &:= \{\rho : F(\theta,\rho) - \rho > \epsilon\}, \\ D_\epsilon &:= \{\rho : F(\theta,\rho) - \rho < -\epsilon\}.\end{aligned}$$

If we can show that $\mathbb{P}(\lim_{t\to\infty}\rho_t \notin U_\epsilon) = 1$ and $\mathbb{P}(\lim_{t\to\infty}\rho_t \notin D_\epsilon) = 1$ for any arbitrarily small ϵ, the proof is completed. We first establish the following fact: If there exists some $t_0 \ge 0$ such that $\rho_{t_0} \in U_\epsilon$, we must have

$$\mathbb{P}(\lim_{t\to\infty}\rho_t \notin U_\epsilon) = 1. \tag{5}$$

To prove (5), consider the first exit time τ of $\{\rho_t\}$ from U_ϵ, namely, τ is the smallest time round such that $\rho_t \notin U_\epsilon$ (or ∞ if $\{\rho_t\}$ never leaves U_ϵ). Let $\tau_t = \min\{\tau, t\}$ denote the minimum of τ and $t \geq t_0$. It is easy to see that $\forall k \geq t_0 + 1$, the event $\{\tau_t \geq k\}$ is $\mathcal{F}_{k-1}-$ measurable, thus

$$\begin{aligned}
1 \geq \mathbb{E}[\rho_{\tau_t}] \geq \mathbb{E}[\rho_{\tau_t} - \rho_{t_0}] &= \mathbb{E}[\rho_{t_0+1} - \rho_{t_0} + \rho_{t_0+2} - \rho_{t_0+1} + \ldots + \rho_{\tau_t} - \rho_{\tau_t - 1}] \\
&= \mathbb{E}\Bigg[\sum_{k=t_0+1}^{t} (\rho_k - \rho_{k-1}) \mathbb{1}\{\tau_t \geq k\}\Bigg] \\
&\geq \mathbb{E}\Bigg[\sum_{k=t_0+1}^{t} \mathbb{E}\Big[\rho_k - \rho_{k-1} | \mathcal{F}_{k-1}\Big] \mathbb{1}\{\tau = \infty\}\Bigg],
\end{aligned}$$

where $\mathbb{1}\{\mathcal{E}\}$ is the indicator function of event $\mathcal{E}$. We aslo have

$$\begin{aligned}
\mathbb{E}[\rho_k - \rho_{k-1} | \mathcal{F}_{k-1}] &= \mathbb{E}[\rho_k - \rho_{k-1} | \rho_{k-1}] \\
&= \mathbb{E}\Big[\eta_k \Big(F(\theta, \rho_{k-1}) - \rho_{k-1}\Big) | \rho_{k-1}\Big] \\
&\qquad \text{By the update rule defined in (2) and } \mathbb{E}[\xi_k | \mathcal{F}_{k-1}] = 0 \\
&\geq \epsilon / k. \qquad \text{By the fact that } \rho_{k-1} \in U_\epsilon
\end{aligned}$$

Then for $\forall t \geq t_0$, $\forall t_0 \geq 0$,

$$\epsilon \sum_{k=t_0+1}^{t} \frac{\mathbb{P}(\tau = \infty)}{k} \leq 1.$$

Since $\sum_k 1/k$ is divergent, then the probability that $\tau = \infty$ must be zero, i.e., $\mathbb{P}(\tau = \infty) = 0$. This completes the proof of (5).

Similarly, we can also prove $\mathbb{P}(\lim_{t\to\infty} \rho_t \notin D_\epsilon) = 1$. Since ϵ is arbitrarily selected, we have $\mathbb{P}(\lim_{t\to\infty} |\rho_t - F(\theta, \rho_t)| \leq \epsilon) = 1$, which completes the proof. □

B.2 Proof of Corollary 4.2

Proof. Since G is strongly convex with respect to ρ, we have $\nabla^2_\rho G > 0$, i.e., $\nabla_\rho F \leq 1$. By Banach fixed-point theorem, we know that $F(\theta, \rho)$ has a unique fixed point ρ^*_θ in $(0, 1)$.

Define $h(\theta, \rho) = \rho - F(\theta, \rho)$. We now proceed to prove that this fixed point ρ^*_θ is globally asymptotically stable for $h(\theta, \rho)$. Consider a Lyapunov function $V(\theta, \rho) : \rho \to \frac{1}{2}(\rho - \rho^*_\theta)^2$ for the ODE defined in (2). We have $V(\theta, \rho) \geq 0$ for $\rho \in (0, 1)$. And $V(\theta, \rho) = 0$ if and only if $\rho = \rho^*_\theta$. Furthermore, we have

$$\frac{d}{d_t} V(\theta, \rho) = (\rho - \rho^*_\theta) \frac{d}{d_t} \rho = (\rho^*_\theta - \rho) h(\theta, \rho).$$

By assumption, $F(\theta, \rho)$ is a contraction mapping function, it is easy to see that $h(\theta, \rho)$ is strictly increasing in ρ, i.e., $\partial h(\theta, \rho)/\partial \rho > 0$. So we have $h(\theta, \rho) \geq (\leq) 0$ for $\rho \geq (\leq) \rho^*_\theta$, which means $dV(\theta, \rho)/d_t \leq 0$ for all $\rho \in (0, 1)$ and $dV(\theta, \rho)/d_t < 0$ for all $\rho \in (0, 1) \setminus \rho^*_\theta$. This proves that ρ^*_θ is the asymptotically stable point of $h(\theta, \rho)$. □

B.3 Proof of Theorem 4.3

We can decompose $z_t := |\rho_t - \rho^*|$ for each $t \geq 0$ into two parts. the *empirical iterate error* $|\rho_t - \rho^*| - \mathbb{E}[|\rho_t - \rho^*|$ and the *expectation error* $\mathbb{E}[|\rho_t - \rho^*|]$:

$$z_t := |\rho_t - \rho^*| = (|\rho_t - \rho^*| - \mathbb{E}[|\rho_t - \rho^*|]) + \mathbb{E}[|\rho_t - \rho^*|]. \tag{6}$$

To derive the probabilistic tail bound for $|\rho_t - \rho^*|$, we bound the empirical iterate error and expectation error separately. Below, we first give the high probability bound for the empirical iterate error:

Lemma B.1. *Given the average feedback dynamics $\{\rho_t\}_{t\geq 0}$ (ρ_0 is the prior information) defined in (2), and under the assumptions of (A1 - A2), and G is strongly convex, then for any $\delta > 0$, we have:*

$$\mathbb{P}(|\rho_t - \rho^*| - \mathbb{E}[|\rho_t - \rho^*|] \geq \delta) \leq \exp\left(\frac{-\delta^2}{2\sum_{i=1}^{t}\eta_i^2\left(\prod_{j=i}^{t-1}(1 - 2\bar{\lambda}\eta_{j+1} + \eta_{j+1}^2(L_h^\rho)^2)\right)}\right),$$

where $L_h^\rho = 1 - L_F^\rho$.

Proof. Notice that by introducing a telescoping sum of martingale differences, we could rewrite $|\rho_t - \rho^*| - \mathbb{E}[|\rho_t - \rho^*|]$ as follows

$$\begin{aligned}
|\rho_t - \rho^*| - \mathbb{E}[|\rho_t - \rho^*|] &= \sum_{i=1}^{t} \mathbb{E}[|z_t||\mathcal{F}_i] - \mathbb{E}[|z_t||\mathcal{F}_{i-1}] \\
&= \sum_{i=1}^{t} g_i(\rho_{i-1}, \xi_i) - \mathbb{E}[g_i(\rho_{i-1}, \xi_i)|\mathcal{F}_{i-1}],
\end{aligned}$$

where $g_i(\rho_i, \xi) = \mathbb{E}[|\rho_t - \rho^*||\rho_i, \mathcal{F}_{i-1}]$. Let $L_h^\rho = 1 - L_F^\rho$ and $\mathcal{G}_i = g_i - \mathbb{E}[g_i|\mathcal{F}_{i-1}]$. Recall that $\mathcal{F}_i := \sigma(\xi_j, j \leq i)$ and $\{\mathcal{F}_i\}_{i\in\mathbb{N}}$ denotes the natural filtration.

Let $H(\theta, \rho, \xi) = \rho - F(\theta, \rho) + \xi$. We then reduce the proof of empirical iterate error by establishing a Lipschitz continuous property of $\mathcal{G}_i$ on martingale difference ξ_i. We also write out the superscript of ρ_j as $\rho_j^{\rho,i}$ to explicitly express the dependency of $\{\rho_t\}$ on a given initial starting time i such that $\rho_i = \rho$. Recall that $h(\theta, \rho) = \mathbb{E}[H(\theta, \rho, \xi)]$. By introducing a martingale difference $\Delta_{j+1}^{\rho,i} = H(\theta, \rho_j^{\rho,i}, \xi_{j+1}) - h(\theta, \rho_j^{\rho,i})$, we then have

$$\begin{aligned}
|\rho_{j+1}^{\rho,i} - \rho_{j+1}^{\rho',i}|^2 &= |\rho_j^{\rho,i} - \rho_j^{\rho',i} - \eta_{j+1}(H(\theta, \rho_j^{\rho,i}, \xi_{j+1}) - H(\theta, \rho_j^{\rho',i}, \xi_{j+1}))|^2 \\
&= (\rho_j^{\rho,i} - \rho_j^{\rho',i})^2 - 2\eta_{j+1}(\rho_j^{\rho,i} - \rho_j^{\rho',i})\Big(H(\theta, \rho_j^{\rho,i}, \xi_{j+1}) - H(\theta, \rho_j^{\rho',i}, \xi_{j+1})\Big) \\
&\quad + \eta_{i+1}^2\Big(H(\theta, \rho_j^{\rho,i}, \xi_{j+1}) - H(\theta, \rho_j^{\rho',i}, \xi_{j+1})\Big)^2 \\
&= (\rho_j^{\rho,i} - \rho_j^{\rho',i})^2 - 2\eta_{j+1}(\rho_j^{\rho,i} - \rho_j^{\rho',i})(h(\theta, \rho_j^{\rho,i}) - h(\theta, \rho_j^{\rho',i})) \\
&\quad - 2\eta_{j+1}(\rho_j^{\rho,i} - \rho_j^{\rho',i})(\Delta_{j+1}^{\rho,i} - \Delta_{j+1}^{\rho',i}) + \eta_{j+1}^2\Big(H(\theta, \rho_j^{\rho,i}, \xi_{j+1}) - H(\theta, \rho_j^{\rho',i}, \xi_{j+1})\Big)^2.
\end{aligned}$$

Applying the Lipschitz continuity of $H(\cdot)$ w.r.t. ρ, by the strongly convex property of $G(\cdot)$, the above equation gives us following

$$\begin{aligned}|\rho_{j+1}^{\rho,i} - \rho_{j+1}^{\rho',i}|^2 &\le (\rho_j^{\rho,i} - \rho_j^{\rho',i})^2 - 2\bar{\lambda}\eta_{j+1}(\rho_j^{\rho,i} - \rho_j^{\rho',i})^2 \\ &\quad - 2\eta_{j+1}(\rho_j^{\rho,i} - \rho_j^{\rho',i})(\Delta_{j+1}^{\rho,i} - \Delta_{j+1}^{\rho',i}) + \eta_{i+1}^2(L_h^\rho)^2(\rho_j^{\rho,i} - \rho_j^{\rho',i})^2 \\ &= (\rho_j^{\rho,i} - \rho_j^{\rho',i})^2(1 - 2\bar{\lambda}\eta_{j+1} + \eta_{j+1}^2(L_h^\rho)^2) - 2\eta_{j+1}(\rho_j^{\rho,i} - \rho_j^{\rho',i})(\Delta_{j+1}^{\rho,i} - \Delta_{j+1}^{\rho',i}).\end{aligned}$$

Then taking induction on j from i to t, we have

$$\begin{aligned}|\rho_t^{\rho,i} - \rho_t^{\rho',i}|^2 &\le (\rho - \rho')^2 \prod_{j=i}^{t-1}(\eta_{j+1}^2(L_h^\rho)^2 - 2\bar{\lambda}\eta_{j+1} + 1) \\ &\quad - 2\prod_{j=i}^{t-1}(1 - 2\bar{\lambda}\eta_{j+1} + \eta_{j+1}^2(L_h^\rho)^2)\sum_{j=1}^{t-1}\frac{\eta_{j+1}}{\prod_{l=i}^{j}(1 - 2\bar{\lambda}\eta_{l+1} + \eta_{l+1}^2(L_h^\rho)^2)}(\rho_j^{\rho,i} - \rho_j^{\rho',i})(\Delta_{j+1}^{\rho,i} - \Delta_{j+1}^{\rho',i}).\end{aligned}$$

Taking the expectation on both sides, applying the tower property of expectation and by the fact that $\mathbb{E}[\Delta_j^{\rho,i}] = 0$, we have

$$\mathbb{E}[|\rho_t^{\rho,i} - \rho_t^{\rho',i}|^2] \le (\rho - \rho')^2 \prod_{j=i}^{t-1}(\eta_{j+1}^2(L_h^\rho)^2 - 2\bar{\lambda}\eta_{j+1} + 1).$$

Back to our error decomposition, we have the following Lipschitz bound for the function $g_i(\cdot)$,

$$\begin{aligned}|g_i(\rho,\xi) - g_i(\rho,\xi')| &= \left|\mathbb{E}\left[|\rho_t^{\rho+\eta_i H(\theta,\rho,\xi),i} - \rho^*|\right] - \mathbb{E}\left[|\rho_t^{\rho+\eta_i H(\theta,\rho,\xi'),i} - \rho^*|\right]\right| \\ &\le \mathbb{E}\left[|\rho_t^{\rho+\eta_i H(\theta,\rho,\xi),i} - \rho_t^{\rho+\eta_i H(\theta,\rho,\xi'),i}|\right] \\ &\le \eta_i|\xi - \xi'|\left(\prod_{j=i}^{t-1}(\eta_{j+1}^2(L_h^\rho)^2 - 2\bar{\lambda}\eta_{j+1} + 1\right)^{1/2}.\end{aligned}$$

The above inequality shows $g_i(\cdot)$ is a Lipschitz continuous function defined on random variable ξ given $\mathcal{F}_{i-1}$ with the Lipschitz constant equaling to $L_{g_i} = \eta_i\left(\prod_{j=i}^{t-1}(\eta_{j+1}^2(L_h^\rho)^2 - 2\bar{\lambda}\eta_{j+1} + 1\right)^{1/2}$. Thus,

$$\begin{aligned}\mathbb{P}(|\rho_t - \rho^*| - \mathbb{E}[|\rho_t - \rho^*|] \ge \delta) &= \mathbb{P}(\sum_{i=1}^{t}\mathcal{G}_i \ge \delta) \\ &\le \exp(-\gamma\delta)\mathbb{E}[\exp(\gamma\sum_{i=1}^{t}\mathcal{G}_i)] \\ &= \exp(-\gamma\delta)\mathbb{E}[\exp(\gamma\sum_{i=1}^{t-1}\mathcal{G}_i)]\mathbb{E}[\exp(\gamma\mathcal{G}_t)|\mathcal{F}_{t-1}].\end{aligned}$$

Now it shows that $\mathcal{G}_t$ is a L_{g_t}-Lipschitz function conditional on $\mathcal{F}_{t-1}$. By invoking a martingale concentration bound in Lemma A.2, we have:

$$\mathbb{E}[\exp(\gamma\mathcal{G}_t)|\mathcal{F}_{t-1}] \le \exp\left(\frac{\gamma^2 L_{g_t}^2}{2}\right).$$

By induction on i, we have

$$\mathbb{P}(|\rho_t - \rho^*| - \mathbb{E}[|\rho_t - \rho^*|] \geq \delta) \leq \exp(-\gamma\delta)\exp\left(\frac{\gamma^2 \sum_{i=1}^t L_{g_i}^2}{2}\right).$$

We can then finish the proof by optimizing with respect to γ.

□

We now proceed to bound the expectation error $\mathbb{E}[|\rho_t - \rho^*|]$:

Lemma B.2. *Given the ratio dynamics $\{\rho_t\}_{t\geq 0}$ (ρ_0 is the prior information) defined in (2), and under the assumptions of (A1 -A2), and G is strongly convex, then we have*

$$\mathbb{E}[|\rho_t - \rho^*|] \leq \exp(-\bar{\lambda} S_t)|\rho_0 - \rho^*| + \sqrt{\sum_{i=0}^{t-1} \eta_{i+1}^2 \exp(-2\bar{\lambda}(S_t - S_{i+1}))},$$

where $S_t = \sum_{i=1}^t \eta_i$.

Proof. We define the following

$$\begin{aligned} z_{t+1} := \rho_{t+1} - \rho^* &= \rho_t - \rho^* - \eta_{t+1} H(\theta, \rho_t, \xi_{t+1}) \\ &= \rho_t - \rho^* - \eta_{t+1}(h(\theta, \rho_t) - \Delta Y_{t+1}), \end{aligned}$$

where $\Delta Y_{t+1} = h(\theta, \rho_t) - H(\theta, \rho_t, \xi_{t+1}) = \mathbb{E}[H(\theta, \rho_t, \xi_{t+1})|\mathcal{F}_t] - H(\theta, \rho_t, \xi_{t+1})$. Rewriting above equation as follows

$$z_{t+1} = \rho_t - \rho^* - \eta_{t+1}(\rho_t - \rho^*) \int_0^1 (\partial h(\theta, \rho^* + \alpha(\rho_t - \rho^*))/\partial\rho) d\alpha + \eta_{t+1}\Delta Y_{t+1}.$$

Let $\mathcal{J}_t = \int_0^1 (\partial h(\theta, \rho^* + \alpha(\rho_t - \rho^*))/\partial\rho) d\alpha$, then we have

$$\begin{aligned} z_{t+1} &= \rho_t - \rho^* - \eta_{t+1}(\rho_t - \rho^*)\mathcal{J}_t + \eta_{t+1}\Delta Y_{t+1} \\ &= z_t(1 - \eta_{t+1}\mathcal{J}_t) + \eta_{t+1}\Delta Y_{t+1}. \end{aligned}$$

Taking a square operator on both sides, expanding and then taking expectation, we can get

$$\begin{aligned} \mathbb{E}[|z_{t+1}|^2] &= \mathbb{E}[|z_t(1 - \eta_{t+1}\mathcal{J}_t) + \eta_{t+1}\Delta Y_{t+1}|^2] \\ &= \mathbb{E}[|z_t(1 - \eta_{t+1}\mathcal{J}_t)|^2] + 2\mathbb{E}[z_t(1 - \eta_{t+1}\mathcal{J}_t)\eta_{t+1}\Delta Y_{t+1}] + \mathbb{E}[|\eta_{t+1}\Delta Y_{t+1}|^2] \\ &= (1 - \eta_{t+1}\mathcal{J}_t)^2 \mathbb{E}[|z_t|^2] + \eta_{t+1}^2 \mathbb{E}[|\Delta Y_{t+1}|^2], \end{aligned}$$

where the last equality is due to the martingale difference property of ΔY_{t+1}. Notice that by the property of strongly convex G, namely, there exists a global stable equilibrium point of h, we have $|1 - \eta_{t+1}\mathcal{J}_t| \leq \exp(-\bar{\lambda}\eta_{t+1})$. Thus,

$$\mathbb{E}[|z_{t+1}|^2] \leq \exp(-2\bar{\lambda}\eta_{t+1})\mathbb{E}[|z_t|^2] + \eta_{t+1}^2.$$

Finally, taking the induction from $t = 1$ will give us the following

$$\mathbb{E}[|z_t|^2] \leq z_0^2 \exp(-2\bar{\lambda} S_t) + \sum_{i=0}^{t-1} \eta_{i+1}^2 \exp(-2\bar{\lambda}(S_t - S_{i+1})),$$

where $S_t = \sum_{i=1}^t \eta_i$. □

Combining the above empirical iterate error and expectation error completes the proof of Theorem 4.3.

Convergence Analysis Let $M_t = \prod_{i=1}^t (1-2\bar{\lambda}/i + (L_h^\rho)^2/i^2)$, then $\sum_{i=1}^t L_i = M_t \sum_{i=1}^t \eta_i^2/M_i$. By $1+x < e^x$, it is immediate to see that $M_t \le \prod_{i=1}^t \exp\Big(-2\bar{\lambda}/i + (L_h^\rho)^2/i^2\Big) = \exp\Big(\sum_{i=1}^t (-2\bar{\lambda}/i + (L_h^\rho)^2/i^2)\Big) = \exp\Big(-2\bar{\lambda}\ln t + (L_h^\rho)^2\pi^2/6\Big) = \exp\Big((L_h^\rho)^2\pi^2/6\Big)t^{-2\bar{\lambda}}$. Thus,

- when $\bar{\lambda} \in (0, 1/2]$, we have $i^2\prod_{j=1}^i (1-2\bar{\lambda}/j + (L_h^\rho)^2/j^2) > i^2(1-2\bar{\lambda})\prod_{j=2}^i (1-1/j) = i(1-2\bar{\lambda})$. Thus, $\sum_{i=1}^t \eta_i^2/M_i$ is summable, and $\sum_{i=1}^t \eta_i^2/M_i \le \sum_{i=1}^t \frac{1}{i^2\prod_{j=1}^i (1-2\bar{\lambda}/j)} \le C_0$, where $C_0 = \lim_{t\to\infty}\sum_{i=1}^t \frac{1}{i^2\prod_{j=1}^i (1-2\bar{\lambda}/j)}$. For simplicity, let $C_1 = C_0\exp((L_h^\rho)^2\pi^2/6)$, then we have $\sum_{i=1}^t L_i \le C_1 t^{-2\bar{\lambda}} = \mathcal{O}(t^{-2\bar{\lambda}})$;
- when $\bar{\lambda} \in (1/2, \infty)$, it can be proved by comparisons with integrals that $\sum_{i=1}^t \eta_i^2/M_i \le Ct^{(2\bar{\lambda}-1)}$, where C is a constant which is only dependent on $\bar{\lambda}$. Thus, let $C_2 = \frac{(2\bar{\lambda}+1)\exp((L_h^\rho)^2\pi^2/6)}{4(2^{2\bar{\lambda}-1}-1)}$, we'll have $\sum_{i=1}^t L_i \le C_2 t^{-1} = \mathcal{O}(t^{-1})$.

For the expectation error δ_t, we know that $S_t = \Theta(\ln t)$. Thus, we have $\sum_{i=0}^{t-1}\eta_{i+1}^2\exp(-2\bar{\lambda}(S_t - S_{i+1})) = \sum_{k=1}^t (1/k^2)\exp(-2\bar{\lambda}\sum_{i=k}^t 1/i) \le \sum_{k=1}^t (1/k^2)\exp(-2\bar{\lambda}\ln t/k) \le t^{-2\bar{\lambda}}\sum_{k=1}^t 1/k^{2-2\bar{\lambda}}$. By comparing the sums with integrals,

- when $\bar{\lambda} \in (0, 1/2)$, we have $\lim_{t\to\infty} t^{-2\bar{\lambda}}\sum_{k=1}^t 1/k^{2-2\bar{\lambda}} = \mathcal{O}(t^{-\bar{\lambda}})$;
- when $\bar{\lambda} = 1/2$, we have $\lim_{t\to\infty} t^{-2\bar{\lambda}}\sum_{k=1}^t 1/k^{2-2\bar{\lambda}} = \Theta(t^{-1}\ln t)$;
- when $\bar{\lambda} \in (1/2, \infty)$, we have $\lim_{t\to\infty} t^{-2\bar{\lambda}}\sum_{k=1}^t 1/k^{2-2\bar{\lambda}} = \mathcal{O}(1/\sqrt{t})$.

Hence, we have $\delta_t \to 0$ when $t \to \infty$.

B.4 Proof of Theorem 4.6

We first prove that a small deviation of $\rho_{k,t}$ leads to a small deviation of the quality estimator $\hat{\theta}_{k,t}$, as summarized in the following lemma:

Lemma B.3. *Assume there exist $D_\theta > 0$ and $D_\rho \in (0,1)$, such that for any $0 < \theta_2 < \theta_1 < 1$, $D_\theta(\theta_1 - \theta_2) \le F(\theta_1, \rho) - F(\theta_2, \rho)$; and for any $0 < \rho_2 < \rho_1 < 1$, $D_\rho(\rho_1 - \rho_2) \le F(\theta, \rho_1) - F(\theta, \rho_2)$. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be the quality estimates for ρ_1 and ρ_2, as specified in Equation (4). Then the following holds*

$$|\hat{\theta}_1 - \hat{\theta}_2| \le \frac{1 - D_\rho}{D_\theta}|\rho_1 - \rho_2|.$$

Proof. Since the quality estimate $\hat{\theta}$ is chosen such that $F(\hat{\theta}, \rho) = \rho$, we have

$$F(\hat{\theta}_1, \rho_1) - F(\hat{\theta}_2, \rho_2) = \rho_1 - \rho_2$$

From Corollary 4.2, we know that when $\rho_1 = \rho_2$, $\hat{\theta}_1 = \hat{\theta}_2$. Therefore the lemma statement is true. Below we discuss the case $\rho_1 > \rho_2$. We first argue that when $\rho_1 > \rho_2$, $\hat{\theta}_1 > \hat{\theta}_2$. Assume by contradiction that $\hat{\theta}_1 < \hat{\theta}_2$. We have

$$\begin{aligned} F(\hat{\theta}_1, \rho_1) - F(\hat{\theta}_2, \rho_2) &= F(\hat{\theta}_1, \rho_1) - F(\hat{\theta}_2, \rho_1) + F(\hat{\theta}_2, \rho_1) - F(\hat{\theta}_2, \rho_2) \\ &\le -D_\theta(\hat{\theta}_2 - \hat{\theta}_1) + L_F^\rho(\rho_1 - \rho_2), \end{aligned}$$

where L_F^ρ is the Lipschitz constant of F. Since $F(\hat{\theta}_1, \rho_1) - F(\hat{\theta}_2, \rho_2) = \rho_1 - \rho_2$,

$$\hat{\theta}_2 - \hat{\theta}_1 \leq \frac{L_F^\rho - 1}{D_\theta}(\rho_1 - \rho_2). \tag{7}$$

Since $F(\theta, \rho)$ is contraction mapping w.r.t. ρ, i.e., $L_F^\rho < 1$. The above inequality leads to contradiction.

Now we focus on the case when $\hat{\theta}_1 > \hat{\theta}_2$, we can get

$$\begin{aligned} F(\hat{\theta}_1, \rho_1) - F(\hat{\theta}_2, \rho_2) &= F(\hat{\theta}_1, \rho_1) - F(\hat{\theta}_2, \rho_1) + F(\hat{\theta}_2, \rho_1) - F(\hat{\theta}_2, \rho_2) \\ &\geq D_\theta(\hat{\theta}_1 - \hat{\theta}_2) + D_\rho(\rho_1 - \rho_2). \end{aligned}$$

Again, we get

$$\hat{\theta}_1 - \hat{\theta}_2 \leq \frac{1 - D_\rho}{D_\theta}(\rho_1 - \rho_2).$$

Following the same argument, when $\rho_1 < \rho_2$, we must have $\hat{\theta}_2 > \hat{\theta}_1$, and moreover:

$$\hat{\theta}_2 - \hat{\theta}_1 \leq \frac{1 - D_\rho}{D_\theta}(\rho_2 - \rho_1).$$

Combining above two cases will complete the proof. □

Armed with the above small deviation result, we now proceed to prove the regret bound in Theorem 4.6.

Proof. We will restrict to prove the regret bound for the case $\bar{\lambda} \in (0, 1/2)$, and the proof also holds when $\bar{\lambda} \in [1/2, \infty)$. By the small deviation connection between $\rho_{k,t}$ and $\hat{\theta}_{k,t}$, the following holds

$$P(|\rho_{k,t} - \rho_k^*| \geq \delta) \geq \mathbb{P}\left(|\hat{\theta}_{k,t} - \theta_k| \geq \frac{1 - D_\rho}{D_\theta}\delta\right).$$

By the convergence analysis of $\rho_{k,t}$, we have the following concentration inequality for the estimator $\hat{\theta}_{k,t}$

$$\mathbb{P}(|\hat{\theta}_{k,t} - \theta_k| \geq \delta) \leq \exp\left(-\frac{D_\theta^2}{2C_1(1 - D_\rho)^2}\delta^2 n_{k,t}^{2\bar{\lambda}}\right),$$

where C_1 is a constant dependent on $\bar{\lambda}$ (defined in the above convergence analysis) and $n_{k,t}$ is the number of pulls of arm k till up to round t. Therefore, for each arm k at time t, we have the following

$$|\hat{\theta}_{k,t} - \theta_k| \leq \sqrt{\frac{\beta \ln t}{n_{k,t}^{2\bar{\lambda}}}},$$

with probability at least $1 - t^{-\beta C'}$, where $C' = \frac{D_\theta^2}{2C_1(1-D_\rho)^2}$. From this, it is immediate to get the following two useful bounds: With probability at least $1 - t^{-\beta C'}$, we have

$$\mathrm{UCB}_{k,t} > \theta_k. \tag{8}$$

Furthermore, given $n_{k,t} \geq (4\beta \ln t/\Delta_k^2)^{1/(2\bar{\lambda})}$, where $\Delta_k = \theta^* - \theta_k$, we have

$$\hat{\theta}_{k,t} < \theta_k + \Delta_k/2. \tag{9}$$

The above two bounds implies the optimistic property of our constructed UCB algorithm. Particularly, (8) implies UCB value should be probably as large as the true arm quality. And (9) implies that given enough samples (at least $(4\beta \ln t/\Delta_k^2)^{1/(2\bar{\lambda})}$), then the quality estimator $\hat{\theta}_{k,t}$ would not exceed the true arm quality by more than $\Delta_k/2$. In words, above two bounds can be used to get following guarantee on finding out a suboptimal arm

$$\mathbb{P}(I_t = k | n_{k,t} \geq (4\beta \ln t/\Delta_k^2)^{1/(2\bar{\lambda})}) \leq t^{-2\beta C'}. \tag{10}$$

Above property is due to:

$$\begin{aligned}
\mathrm{UCB}_{k,t} = \hat{\theta}_{k,t} + \sqrt{\beta \ln t / n_{k,t}^{2\bar{\lambda}}} &\leq \hat{\theta}_{k,t} + \Delta_k/2 \\
&< \theta_k + \Delta_k/2 + \Delta_k/2 \\
&= \theta^* < \hat{\theta}_{I^*,t} + \sqrt{\beta \ln t / n_{I^*,t}^{2\bar{\lambda}}} \\
&= \mathrm{UCB}_{I^*,t}.
\end{aligned}$$

The first inequality is coming from $n_{k,t} \geq (4\beta \ln t/\Delta_k^2)^{1/(2\bar{\lambda})}$ and second inequality coming from (9), the third equality coming from $\Delta_k = \theta^* - \theta_k$ and the forth inequality is due to (8).
Till now, we can bound the number of pulls for suboptimal arm k

$$\begin{aligned}
\mathbb{E}[n_{k,T}] &= 1 + \mathbb{E}[\sum_{t=K}^{T} \mathbb{1}(I_{t+1} = k)] \\
&= 1 + \mathbb{E}[\sum_{t=K}^{T} \mathbb{1}(I_{t+1} = k, n_{k,t} < (4\beta \ln t/\Delta_k^2)^{1/(2\bar{\lambda})})] + \mathbb{E}[\sum_{t=K}^{T} \mathbb{1}(I_{t+1} = k, n_{k,t} \geq (4\beta \ln t/\Delta_k^2)^{1/(2\bar{\lambda})}] \\
&\leq (4\beta \ln T/\Delta_k^2)^{1/(2\bar{\lambda})} + \mathbb{E}[\sum_{t=K}^{T} \mathbb{1}(I_{t+1} = k, n_{k,t} \geq (4\beta \ln t/\Delta_k^2)^{1/(2\bar{\lambda})})] \\
&= (4\beta \ln T/\Delta_k^2)^{1/(2\bar{\lambda})} + \sum_{t=K}^{T} \mathbb{P}(I_{t+1} = k, n_{k,t} \geq (4\beta \ln t/\Delta_k^2)^{1/(2\bar{\lambda})}) \\
&= (4\beta \ln T/\Delta_k^2)^{1/(2\bar{\lambda})} + \sum_{t=K}^{T} \mathbb{P}(I_{t+1} = k | n_{k,t} \geq (4\beta \ln t/\Delta_k^2)^{1/(2\bar{\lambda})}) \mathbb{P}(n_{k,t} \geq (4\beta \ln t/\Delta_k^2)^{1/(2\bar{\lambda})}) \\
&\leq (4\beta \ln T/\Delta_k^2)^{1/(2\bar{\lambda})} + \sum_{t=K}^{T} t^{-2\beta C'} \\
&\leq (\frac{4 \ln T}{C' \Delta_k^2})^{\frac{1}{2\bar{\lambda}}} + \pi^2/6.
\end{aligned}$$

where the first equality is for adding 1 initial pull for every arm. For the first inequality, suppose the indicator $\mathbb{1}(I_{t+1} = k, n_{k,t} < N)$ takes value 1 at more than $N-1$ time rounds, where $N = (4 \ln t/\Delta_k^2)^{1/(2\bar{\lambda})}$. And let t' be the time step where $\mathbb{1}(I_{t+1} = k, n_{k,t} < N) = 1$ for $(N-1)^{\text{th}}$ round.

Thus, including the initial pull, arm k has been pulled at least N rounds until time t'. Then for any $t > t'$, $n_{k,t} > N$ which implies $n_{k,t} > (4\ln t/\Delta_k^2)^{1/(2\bar{\lambda})}$. Thus, the indication cannot be 1 for any $t > t'$, contradicting the assumption that the indicator takes values 1 for more than $N-1$ rounds. The second inequality is coming from the tail bound for number of pulls for suboptimal to bound the first conditional term. The second probability is bounded by 1. The last inequality is by choosing $\beta = 1/C' = \frac{2C_1(1-D_\rho)^2}{D_\theta^2}$, where C_1 is defined as above in convergence analysis.

We note that the above analysis also holds true when $\bar{\lambda} \in [1/2, \infty)$, and the key difference is that concentration inequality for estimator $\hat{\theta}_{k,t}$ will become following

$$\mathbb{P}(|\hat{\theta}_{k,t} - \theta_k| \geq \delta) \leq \exp\left(-\frac{D_\theta^2}{2C_2(1-D_\rho)^2}\delta^2 n_{k,t}\right),$$

and $\beta = \frac{2C_2(1-D_\rho)^2}{D_\theta^2}$ to ensure the number pulls for suboptimal arms with a logarithmic times.

Then the expected regret is obtained by summing for all suboptimal arms: $\mathbb{E}[R(T)] \leq \sum_{k \neq I^*} \mathbb{E}[n_{k,t}]\Delta_k$.

When $\bar{\lambda} \geq 1/2$ (which includes the unbiased feedback setting with $\bar{\lambda} = 1$) dividing the arms into two groups, group 1 contains "almost optimal" arms with $\Delta_k < \sqrt{\ln T/T}$, while group 2 contains "bad" arms with $\Delta_k \geq \sqrt{\ln T/T}$. Then the regret $\mathbb{E}[R(T)] \leq \sum_{k \in \text{Group } 1} \mathbb{E}[n_{k,t}]\Delta_k + \sum_{k \in \text{Group } 2} \mathbb{E}[n_{k,t}]\Delta_k \leq \sqrt{\ln T/T} \sum_{k \in \text{Group } 1} \mathbb{E}[n_{k,t}] + \sum_{k \in \text{Group } 2} (\frac{4\ln T}{C\Delta_k} + \pi^2/6)\Delta_k \leq T\sqrt{\ln T/T} + 4\sqrt{T \ln T}$, which shows a regret of $\mathcal{O}(\sqrt{T \ln T})$ over T rounds.

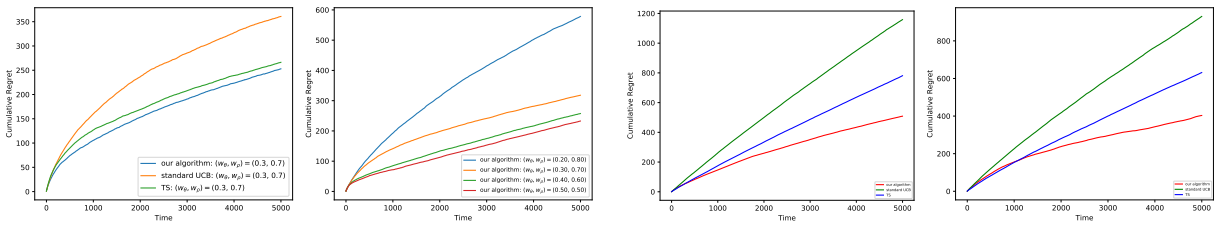
When $\bar{\lambda} \to 0$, i.e., $\partial F(\theta, \rho)/\partial \rho \to 1$, which means the information gain on updating estimator $\hat{\theta}_{k,t}$ will become negligible , thus becomes hard to differentiate the arms, which reflects suffering regret in above result. □

B.5 Experiments.

We conduct a simple simulation to evaluate our Algorithm Avg-UCB. For each experiment, we perform 50 independent trials up to time $T = 5000$ and report the average cumulative regret. For each independent trial, there are $K = 5$ arms with quality drawn uniformly at random from the unit range $(0, 1)$. We use the classic UCB and TS (Thompson Sampling), the two most popular and robust bandit algorithms, as the comparison baselines. In these baseline algorithms, the learner treats the biased feedback as the unbiased estimates of the true rewards. For the UCB algorithm, we set the exploration paramter $\beta = 2$ as the default setting..

Evaluate the performance. We start with evaluating the performance of our algorithm compared with the UCB and TS. We use the feedback function as provided in Example 1, i.e., $F(\theta, \rho) = w_\theta\theta + w_\rho\rho$, for any $w_\theta, w_\rho \geq 0$ and $w_\theta + w_\rho = 1$. In this function, it is easy to see that $\bar{\lambda} = w_\theta$. Note that when $w_\theta \in [1/2, 1]$, our algorithm will recover the standard UCB algorithm. Thus, we set $w_\theta = 0.3$ and compute $\beta = 1.2$ for Algorithm 1. Figure 1a, which shows the regret of three algorithms across time, demonstrates that our algorithm does achieve better performance than baseline algorithms that are oblivious of biased feedback.

Evaluate the performance with different $\bar{\lambda}$. In this experiment, we again use the Example 1 as the user's feedback function. As showed in our regret bound, $\bar{\lambda}$ reflects the learnability of the



(a) (b) (c) (d)

Figure 1: (a) & (b): Performance of Algorithm 1 on feedback function defined in (11). (a): Performance compared with UCB and TS; (b): Performance on different w_θ. (c) & (d): Performance of Algorithm 1 on feedback function defined in (11). (c): $k = 0.7, b = 0.4, \bar{\lambda} = 0.4535$; (d): $k = 0.8, b = 0.4, \bar{\lambda} = 0.5250$.

hidden parameter θ from the noise feedback, i.e., when $\bar{\lambda}$ (recall that in this particular feedback function, $\bar{\lambda} = w_\theta$) is larger, the learner can be more aggressive to learn θ. Thus, in Figure 1b, we compare the performance of our algorithm on different w_θ, i.e., we set $w_\theta = [0.20, 0.30, 0.40, 0.50]$, where we set β all equal to 1.2. The result shows that the regret get increased as the w_θ increases, this confirms our derived regret bound in Theorem 4.6 of the previous section, where larger $\bar{\lambda}$ (smaller $\bar{\lambda}'$) leads better regrets in the non-asymptotic regime.

Non-convex G**.** We further investigate how to adapt our algorithm when G is non-convex with respect to ρ, i.e., there exists some $\rho \in (0,1)$ such that $\nabla^2_\rho G = \nabla_\rho\big(\rho - F(\theta,\rho)\big) < 0$. We construct $F(\theta,\rho)$ from a generalized logistic function. Given the arm's history information ρ_t:

- if the user's private experience is positive, the probability for him to provide positive feedback is $f(\rho) = \frac{1}{1+\exp(-k(\rho-b))}$;
- if the user's private experience is negative, the probability for him to provide positive feedback is $f(1-\rho) = \frac{1}{1+\exp(-k((1-\rho)-b))}$;

Thus, the probability for a user to provide positive feedback is characterized by the following

$$F(\theta,\rho) = \frac{\theta}{1+\exp(-k(\rho-b))} + \frac{1-\theta}{1+\exp(-k((1-\rho)-b))}, \tag{11}$$

where k and b are the parameters which control the steepness of the curve and the midpoint of f, they're chosen to ensure the output of $f(\cdot)$ falls in $(0,1)$ for all $\theta \in (0,1)$. Recall that due to the non-convexity of G, we cannot directly apply our algorithm, since $\bar{\lambda}$ is smaller than 0. To adapt our algorithm in non-convex setting, we use the local convexity of equilibrium points ρ^* of $\rho - F(\theta,\rho)$ and we compute

$$\bar{\lambda} = 1 - \sup_{\forall\theta\in(0,1)} \max_{\rho^*\in\mathcal{S}_\theta} \nabla_{\rho^*} F(\theta,\rho).$$

In Figure 1c, we set $k = 0.7, b = 0.4$, and compute $\bar{\lambda} = 0.4535$, while for Figure 1d, we set $k = 0.8, b = 0.4$, and compute $\bar{\lambda} = 0.5250$ in order for matching the derived two regions of our regret bound. By exploiting the local convexity of equilibrium point of function G, the result demonstrates our algorithm is still robust on finding out optimal arm.

C Proofs in Bandits with Beta-Herding Feedback Model

C.1 Proof of Lemma 5.1

Proof. Let $S_t = \sum_{i=1}^{t} x_i$, where x_i is the realization of the feedback random variable X_i. For simplicity, let $\{n_0\rho_0, n_0(1-\rho_0)\} = \{a, b\}$ (in our current setting, $n_0 = \rho_0 = 0$, which implies $a = b = 0$, but our results hold even if they are nonzero.).

Before we characterize the asymptotic behavior of $\{X_t\}_{t\geq 0}$, we observe that $\{X_t\}_{t\geq 0}$ satisfies a so-called *exchangeable* property. This is summarized in following definition.

Definition C.1. *A sequence $\{X_t\}$ of random variables is exchangeable if for all $t \geq 2$*

$$X_1, ..., X_t \overset{\Delta}{=} X_{\pi(1)}, ..., X_{\pi(t)}, \forall \pi \in S(t).$$

where $S(t)$ is the symmetric group, the group of permutations.

We have the following

Lemma C.2. *Given the above defined learning process, the stochastic random process $\{X_t\}_{t\geq 0}$ is exchangeable.*

Proof. Suppose the principal has received a total of t feedback from the agent, then the probability that there are l positive feedback is given by

$$\frac{\prod_{i=0}^{l-1}(m\theta + a + i)\prod_{j=0}^{t-l-1}(m(1-\theta)b + j)}{\prod_{i=0}^{t-1}(m + a + b + i)}.$$

This shows the exchangeability of $\{X_t\}_{t\geq 0}$. □

Based on the exchangeable property of $\{X_t\}_{t\geq 0}$, we can establish that the asymptotic positive feedback ratio ρ_∞ converges almost surely to a random variable. Suppose $S_t = l$, by the above exchangeability property, we have

$$\mathbb{P}(X_1 = x_1, ..., X_t = x_t) = \frac{\prod_{i=0}^{l-1}(m\theta + a + i)\cdot\prod_{j=0}^{t-1-l}(m(1-\theta) + b + j)}{\prod_{i=0}^{t-1}(m + a + b + i)}.$$

Moreover,

$$\begin{aligned}
\mathbb{P}(S_t = l) &= \binom{t}{l}\frac{\prod_{i=0}^{l-1}(m\theta + a + i)\cdot\prod_{j=0}^{t-1-l}(m(1-\theta) + b + j)}{\prod_{i=0}^{t-1}(m + a + b + i)} \\
&= \binom{t}{l}\frac{\frac{\Gamma(m\theta+a+l)}{\Gamma(m\theta+a)}\cdot\frac{\Gamma(m(1-\theta)+b+t-l)}{\Gamma(m(1-\theta)+b)}}{\frac{\Gamma(m+a+b+t)}{\Gamma(m+a+b)}} \\
&= \frac{\Gamma(m+a+b)}{\Gamma(m\theta+a)\cdot\Gamma(m(1-\theta)+b)}\frac{\Gamma(l+a+m\theta)}{\Gamma(l+1)}\frac{\Gamma(t-l+b+m(1-\theta))}{\Gamma(t-l+1)}\frac{\Gamma(t+1)}{\Gamma(t+a+b+m)} \\
&= \frac{1}{B(m\theta+a, m(1-\theta)+b)}\frac{\Gamma(l+a+m\theta)}{\Gamma(l+1)}\frac{\Gamma(t-l+b+m(1-\theta))}{\Gamma(t-l+1)}\frac{\Gamma(t+1)}{\Gamma(t+a+b+m)}.
\end{aligned}$$

where $\Gamma(\cdot)$ and $B(\cdot)$ are Gamma function and Beta function, respectively.

By Stirling's approximation for gamma function quotient in Lemma A.1, we have

$$\mathbb{P}(S_t = l) = \frac{1}{B(m\theta + a, m(1-\theta) + b)} \cdot l^{a+m\theta-1} \cdot (t-l)^{b+m(1-\theta)-1} \cdot t^{1-a-b-m}. \tag{12}$$

Denote $l = \rho t$ for some $0 < \rho < 1$. Then we have

$$\mathbb{P}(\frac{S_t}{t} \leq \rho) = \mathbb{P}(\frac{S_t}{t} = 0) + \mathbb{P}(\frac{S_t}{t} = \frac{1}{t}) + ... + \mathbb{P}(\frac{S_t}{t} = \frac{\lfloor t\rho \rfloor}{t}).$$

Therefore,

$$\int_0^\rho \mathbb{P}(\frac{S_t}{t} = u)du = \lim_{t\to\infty} \frac{1}{t}[\mathbb{P}(\frac{S_t}{t} = 0) + \mathbb{P}(\frac{S_t}{t} = \frac{1}{t}) + ... + \mathbb{P}(\frac{S_t}{t} = \frac{\lfloor t\rho \rfloor}{t})]$$
$$\mathbb{P}(\frac{S_t}{t} \leq \rho) = t\int_0^\rho \mathbb{P}(\frac{S_t}{t} = u)du.$$

Replacing l with ρt in Equation (12), we have

$$\mathbb{P}(\frac{S_t}{t} \leq \rho) = \frac{1}{B(m\theta + a, m(1-\theta) + b)} \int_0^\rho u^{a+m\theta-1}(1-u)^{b+m(1-\theta)-1}du,$$

which completes the proof. □

C.2 Proof of Lemma 5.2

Proof. Let $f(x|\theta)$ be the probability mass function of random variable X and x_t be the realization of X_t. Consider the stochastic process specified in Equation (1), the probability mass function can be computed as $f(x_t|\theta) = (\frac{m\theta+S_{t-1}+a}{m+a+b+t-1})^{x_t} \cdot (1 - \frac{m\theta+S_{t-1}+a}{m+a+b+t-1})^{1-x_t}$, where $x_t = 1$ or $x_t = 0$, $S_t = \sum_{i=1}^t X_i$. Let $l(x_t|\theta)$ be the log-likelihood of $f(x_t|\theta)$, namely,

$$l(x_t|\theta) = x_t \log\left(\frac{m\theta + S_{t-1} + a}{m + a + b + t - 1}\right) + (1 - x_t)\log\left(1 - \frac{m\theta + S_{t-1} + a}{m + a + b + t - 1}\right).$$

According to the definition of Fisher information for a single observation, and by the chain rule for multiple observations, we have

$$\sum_{i=1}^t \mathcal{I}_i(\theta) = \sum_{i=1}^t -\mathbb{E}[l''(x_i|\theta)]$$
$$= \sum_{i=1}^t \frac{m^2}{m + a + b + i - 1}(\mathbb{E}[\frac{1}{m\theta + a + S_{i-1}}] + \mathbb{E}[\frac{1}{m(1-\theta) + b + i - 1 - S_{i-1}}]).$$

Since $\{X_t\}_{t\geq 1}$ are exchangeable random variables, then

$$\begin{aligned}
\mathbb{E}[\frac{1}{m\theta+a+S_t}] &= \sum_{l=0}^{t}\binom{t}{l}\frac{\prod_{i=0}^{l-1}(m\theta+a+i)\cdot\prod_{j=0}^{t-l-1}(m(1-\theta)+b+j)}{\prod_{i=0}^{t-1}(m+a+b+i)}\cdot\frac{1}{m\theta+a+l} \\
&= \sum_{l=0}^{t}\frac{1}{B(m\theta+a,m(1-\theta)+b)}l^{a+m\theta-1}(t-l)^{b+m(1-\theta)-1}t^{1-a-b-m}\cdot\frac{1}{m\theta+a+l} \\
&= \sum_{l=0}^{t}\frac{1}{B(m\theta+a,m(1-\theta)+b)}(l/t)^{a+m\theta-1}(1-l/t)^{b+m(1-\theta)-1}t^{-1}\cdot\frac{1}{m\theta+a+l} \\
&\leq \frac{1}{B(m\theta+a,m(1-\theta)+b)}t^{-1}\sum_{l=0}^{t}(l/t)^{a+m\theta-1}(1-l/t)^{b+m(1-\theta)-1} \\
&= \mathcal{O}(t^{-1}).
\end{aligned}$$

Similarly, we also have

$$\mathbb{E}[\frac{1}{m(1-\theta)+b+t-1-S_{t-1}}] = \mathcal{O}(t^{-1}).$$

Thus,

$$\begin{aligned}
\lim_{t\to\infty}\sum_{i=1}^{t}\mathcal{I}_i(\theta) &= \lim_{t\to\infty}\sum_{i=1}^{t}\frac{m^2}{m+a+b+i-1}\mathcal{O}(i^{-1}) \\
&\leq m^2\lim_{t\to\infty}\sum_{i=1}^{t}\frac{1}{i}\mathcal{O}(i^{-1}) \\
&= \mathcal{O}(1),
\end{aligned}$$

where $\mathcal{O}(1)$ is a constant. This completes the proof. □

C.3 Proof of Theorem 5.3

Proof. We prove this by contradiction. Consider a bandit model which has two arms. Without loss of generality, assume arm 1 is optimal and arm 2 is suboptimal, i.e., $\theta_1 > \theta_2$, and suppose there exists an algorithm $\mathcal{A}$ which can achieve sublinear regret, i.e., $\mathbb{E}(R_{\mathcal{A}}(T)) = o(T)$. Let k_t denote the arm chosen by algorithm $\mathcal{A}$ at time t. One must have $\lim_{t\to\infty}\mathbb{P}(k_t = 1) = 1$. Let $\hat{\theta}_1^t, \hat{\theta}_2^t$ be the algorithm's estimators on θ_1, θ_2 given the history information accumulated till time round t. The ability to almost surely choose arm 1 by algorithm $\mathcal{A}$ when $t \to \infty$ indicates that we are able to differentiate the two arms, i.e.,

$$\lim_{t\to\infty}\mathbb{P}(\hat{\theta}_1^t > \hat{\theta}_2^t) = 1$$

However, as shown in Lemma 5.2, since the fisher information on the estimator are always bounded even when given infinitely many observations. It implies the estimators are not consistent, and that $\lim_{t\to\infty}\mathbb{P}(\hat{\theta}_1^t < \hat{\theta}_2^t) > 0$. This leads to the contradiction and completes the proof. □

C.4 Proof of Theorem 5.4

Proof. We prove the upper regret bound by separately bounding the regret of two phases. The proof in the first learning phase is similar to the proof of upper regret bound in avg-herding feedback model. We include the proof here for completeness. By Chernoff Inequality, we have

$$|\hat{\theta}'_{k,t-1} - \theta'_k| < \sqrt{\frac{\beta \ln t}{n_{k,t}}},$$

with probability at least $1 - 2\beta/t^2$. This means, with probability at most $2\beta/t^2$,

$$U_{k,t} < \theta'_k. \tag{13}$$

Given that $n_{i,t} \geq 4\beta \ln t/\Delta_k^2$, with probability at least $1 - 2\beta/t^2$,

$$\hat{\theta}'_{k,t} < \theta'_k + \Delta_k/2. \tag{14}$$

The above inequality means, the quality estimator would not exceed the true quality by more than $\Delta_k/2$ with high probability. Thus, given a suboptimal arm k which has been pulled for $n_{k,t} > 4\beta \ln t/\Delta_k^2$ times, with probability at most $4\beta/t^2$, we have $U_{I^*,t} < U_{k,t}$, i.e.,

$$\mathbb{P}(I_{t+1} = k | n_{k,t} \geq 4\beta/\Delta_k^2) \leq 4\beta/t^2.$$

The above high probability bound is coming from $U_{k,t} = \hat{\theta}'_{k,t-1} + \sqrt{\beta \ln t/n_{k,t}} \leq \hat{\theta}'_{k,t-1} + \Delta_k/2 < \theta'_k + \Delta_k = \theta^{*\prime} \leq \hat{\theta}^{*\prime} + \sqrt{\beta \ln t/n_{k^*,t}} = U_{k^*,t}$ given both (13) and (14) hold true. Then, one can bound the expected number of pulls of arm k up to round T^α:

$$\begin{aligned}
\mathbb{E}[n_{k,T^\alpha}] &= 1 + \mathbb{E}[\sum_{t=K}^{T^\alpha} \mathbb{1}(I_{t+1} = k)] \\
&= 1 + \mathbb{E}[\sum_{t=K}^{T^\alpha} \mathbb{1}(I_{t+1} = k, n_{k,t} < 4\beta \ln t/\Delta_k^2)] + \mathbb{E}[\sum_{t=K}^{T^\alpha} \mathbb{1}(I_{t+1} = k, n_{k,t} \geq 4\beta \ln t/\Delta_k^2)] \\
&\leq 4\beta \ln T^\alpha/\Delta_k^2 + \sum_{t=K}^{T^\alpha} \mathbb{P}(I_{t+1} = k, n_{k,t} \geq 4\beta \ln t/\Delta_k^2) \\
&= 4\alpha\beta \ln T/\Delta_k^2 + \sum_{t=K}^{T^\alpha} \mathbb{P}(I_{t+1} = k | n_{k,t} \geq 4\beta \ln t/\Delta_k^2) \mathbb{P}(n_{k,t} \geq 4\beta \ln t/\Delta_k^2) \\
&\leq 4\alpha\beta \ln T/\Delta_k^2 + 8\beta.
\end{aligned}$$

Thus, the regret in the first learning phase could be bounded as follows

$$\mathbb{E}[R(T^\alpha)] = \sum_{k \neq I^*} \mathbb{E}[n_{k,T^\alpha}]\Delta_k \leq \sum_{k \neq I^*} \frac{4\alpha\beta \ln T}{\Delta_k} + 8\beta\Delta_k. \tag{15}$$

In the second phase, the algorithm recommends the arm I_τ for the reminder of the rounds. Denote the regret accumulated in the second phase be recommendation regret, i.e., $\mathbb{E}[r_\tau]$. To bound the recommendation regret, we note that the algorithms is essentially running UCB(β) in the first phase and then select the most played arm (MPA) in the second phase. The regret caused in the second phase has been derived by [8] and we rephrase it as follows.

Lemma C.3. *If we select most played arm (MPA) in the second phase after adopting UCB(β) in the first phase, for $\beta > 1$, and $\tau \geq K(K+2)$, then*

$$\mathbb{E}[r_\tau] \leq \sqrt{4K\beta \ln \tau/(\tau - K)} + \frac{K}{\beta - 1}(\tau/K - 1)^{2-2\beta}. \tag{16}$$

Combining the above two upper regret bounds and summing for all suboptimal arms and all rounds will give us the final regret bound on two-level policy. □