Bandit Learning with Biased Human Feedback

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 principal 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 algorithms achieving a *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, we show that there exist no bandit algorithms that could achieve sublinear regrets. Our results demonstrate the importance of understanding human behavior when applying bandit approaches in systems with humans in the loop.

Introduction

In a multi-armed bandit problem, a principal sequentially selects from a set of arms. Each arm is associated with some unknown reward distribution. The principal can observe the realized reward for the selected arm. The goal of the principal 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, there have been a wide range of applications, such as medical trials, online auctions, or web advertisements.

In many application domains, however, the realized reward might not be observable, and the principal might only receive biased reports of the realized rewards. For example, consider platforms that rely on user ratings/feedback (e.g., Yelp, reddit, or Amazon) to make recommendation decisions. the platforms are essentially facing a bandit problem of balancing exploration (recommending items with fewer ratings to acquire more information) and exploitations (recommending items with the highest empirical ratings). However, as the empirical studies suggest (Muchnik, Aral, and Taylor 2013; Salganik, Dodds, and Watts 2006; Sipos, Ghosh, and Joachims 2014), user feedback are 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., herding behavior). These empirical evidences suggest a different stochastic model in that each observed feedback might be biased by the feedback history. Moreover, this biased user feedback introduces an additional challenge. 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 regrets from only observing biased feedback?

In this paper, we study a variant of the multi-armed bandit problem. In this problem, the principal 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 principal is to maximize the total realized rewards for the selected arms while only having access to human-generated feedback.

We explore two natural user feedback models. 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 (Robbins and Monro 1951). 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 connec-

tion, we show that under some mild conditions, the average feedback for each arm will converge to some value (which may not be the average rewards of the arm), and we quantify the rate of convergence. These convergence results further enable us to design a bandit algorithm based on UCB and achieve sublinear regrets.

While the results of 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, but also the number of feedbacks the arm has so far. This model captures a natural scenario that users might be biased more heavily if there exists more feedback 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 for the arm, the principal 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 exists no bandit algorithms that can achieve sublinear regrets when user feedback follows beta-herding feedback model.

Related Work

Our work is a variant of well-studied multi-armed bandit problem (Lai and Robbins 1985). Bandit problems traditionally assume the rewards generated by each arm at each round are directly observable, and the research focus has been divided into studying either rewards are independent and identically distributed (i.i.d.) (Auer, Cesa-Bianchi, and Fischer 2002) or adversarial (Auer et al. 1995; Audibert and Bubeck 2009). There exist other works that assume rewards are neither i.i.d. drawn nor adversarial. For example, bandits with Markovian rewards (Neu et al. 2010; Ortner et al. 2012) assume the state of each arm evolves according to a Markov process. Other non-stationary bandit problems (Besbes, Gur, and Zeevi 2014; Garivier and Moulines 2011) 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 (Ghosh and Hummel 2013; Liu and Ho 2018). Our work differs from the above work 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 have access to biased feedback while previous work assume the realized rewards are observable.

There have been recent works exploring bandit learning with humans in the loop (Kremer, Mansour, and Perry 2013; Frazier et al. 2014; Mansour, Slivkins, and Syrgkanis 2015; Papanastasiou, Bimpikis, and Savva 2017). In the setting of these works, the principal 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 suboptimal 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 (Frazier et al. 2014) or utilizing information asymmetry to lead agents to explore by designing what information to show to each agent (Kremer, Mansour, and Perry 2013; Mansour, Slivkins, and Syrgkanis 2015; Papanastasiou, Bimpikis, and Savva 2017). This idea of utilizing information asymmetry has been borrowed from Bayesian Persuasion (Kamenica and Gentzkow 2011) in economics. Our work has focused on a parallel aspect of human involvements in bandit learning, in which humans are involved in *feedback generation*.

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 (Salganik, Dodds, and Watts 2006; Muchnik, Aral, and Taylor 2013; Sipos, Ghosh, and Joachims 2014). For example, Muchnik, Aral, and Taylor (2013) 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 (Banerjee 1992; Bikhchandani, Hirshleifer, and Welch 1992; Smith and Sørensen 2000). They discuss the setting in which users' decisions might be influenced by other users' decisions. Therefore, under certain conditions, users sometimes might follow what other users do and collectively make the bad decision. 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 under two natural models and explore the impacts on the design of bandit algorithms.

Problem Setting

We explore a multi-armed bandit problem with biased, history-dependent human feedback. In our setting, each arm is associated with an intrinsic reward distribution. At each time step, a random user arrives, and the principal needs to select an arm for the user. When an arm is selected, the user receives a reward randomly drawn from the arm's reward distribution. She then provides feedback, a biased report that depends on both the received reward and the feedback history of the arm. The principal can only observe the user feedback but not the realized reward. The goal of the principal is to maximize the rewards users receive over time while only having access to the biased feedback.

Formally, 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, the principal 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

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

The reward is not observable to the principal. However, after receiving the reward, each user provides a binary feed-

back $X_t \in \{0, 1\}$ about this arm. The goal of the principal is to maximize the total rewards users receive while observing only the (potentially biased) feedback.

User feedback models. Users' feedback depend on both the realized rewards and the feedback history of the arms. We assume users have access to the historical feedback information for the arms selected in their rounds. However, they have no access to the information of the other 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 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 easily extend 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. Note that if users provide unbiased feedback, we should have $X_t = Z_t$ for all t. However, in practice, user feedback might be biased by other users' feedback (i.e., the feedback history).

In this paper, we use a feedback function to model the probability of obtaining a positive feedback from a user in the population. Note that a feedback function can be interpreted as describing the characteristic of the *user population* the platform is interacting with, instead of describing specific users. In particular, we use $\texttt{Feedback}(\theta, \rho, n)$ to model the probability of obtaining a positive feedback from users 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 n.

Naturally, when $\texttt{Feedback}(\theta, \rho, n) = \theta$, 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

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

  In the discussion later, we study the stochastic process of user feedback specified by F and discuss the impacts 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. In particular, we assume users update their beliefs about the arm quality in a Bayesian manner. They treat the historical ratings as the prior signals 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. Therefore, 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 (user weights times arm quality) and the second term is the number of positive signals from other users. The total number of signals is $m + n$.

  Using standard Bayesian rule, the probability of obtaining positive feedback for arm k at round t can be written as

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

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

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

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

where the expectation is taken over the randomness of the 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.

Bandits with Avg-Herding Feedback Model

In this section, we explore the bandit learning problem when the 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.

Stochastic process of feedback generation

In the following discussion, we explore the feedback dynamics of a single arm. 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 repeated selected and therefore $n_t \equiv t$ when studying the stochastic process for a single arm.

Recall that in avg-herding feedback model, when the quality of the arm is θ and the average feedback for 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 reduced 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

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

Define the noise term $\xi_t = \mathbb{E}[X_t|\mathcal{F}_{t-1}] - X_t = F(\theta, X_t) - X_t$, where $\mathcal{F}_t = \sigma(\{X_t\}_{t\geq 1})$ is the filtration of the stochastic process. Note that 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.

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

Connection to stochastic approximation. Equation 2 is in the form of an update rule in stochastic approximation (Robbins and Monro 1951; Frikha, Menozzi, and others 2012). 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:

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

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.

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. All random variables involved in this model are defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

A1. $F(\theta, \rho)$ is twice differentiable, strictly increasing in θ, and strictly increasing in ρ;

A2. $F(\theta, x)$ is $L_F^\rho-$Lipschitz continuous with respect to ρ, where $L_F^\rho > 0$;

A1 implies that, condition on the same quality (average feedback), an arm with better average feedback (quality) receives more positive feedback in expectation, and *A2* assumes the improvement is smooth. With these assumptions, we can formally characterize the convergence of ρ_t. [1]

Lemma 1. *Suppose A1 is satisfied. 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 S_θ and characterizes the points in S_θ. Intuitively, S_θ is the set of the local optimal points for the latent function G. This intuition suggests that, when the latent function G is strictly 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. In particular, when $L_F^\rho < 1$, by definition, we have $\nabla_\rho^2 G(\theta, \rho) = \nabla_\rho(1 - F(\theta, \rho)) > 0$ for all $\rho \in [0, 1]$. Therefore, when $L_F^\rho < 1$, G is strictly convex. Below we formally characterize the convergence of ρ_t when G is strictly convex.

[1] Most of the proofs in this paper are included in the appendix due to space constraints.

Corollary 1. *Suppose A1 and A2 are satisfied. Given $L_F^\rho < 1$, i.e., G is strictly convex, there exists an 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 rates of ρ_t and focus on the case when G is strictly convex. In particular, we introduce $\bar{\lambda} > 0$, such that $\nabla_\rho^2 G \geq \bar{\lambda} > 0$.

Theorem 2. *Suppose A1 and A2 are satisfied. Assume $L_F^\rho < 1$, i.e., G is strictly convex. $\forall \delta > 0$, we have:*

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

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

Remark 1. *We would like to offer a few observations to help interpret the convergence bound. The detailed derivations are included in the appendix. In particular,*

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

So we can characterize the bound in two regions based on whether $\bar{\lambda} \geq 1/2$. Moreover, when user feedback is unbiased, i.e., $F(\theta, \rho) = \theta$, we have $\bar{\lambda} = 1$, and the bound reduces to $\mathbb{P}(|\rho_t - \rho^*|) \geq \delta) \leq \mathcal{O}(e^{-\delta^2 t})$, the same as the standard Chernoff bound.

Designing bandit algorithms

Given the convergence bound in Theorem 2, we can design a UCB-like algorithm that achieves sublinear regrets. We assume the principal has knowledge of the feedback model F. Note that since F models the behavior of providing feedback for the *user population* the platform is interacting with. This assumption only requires the platform to have knowledge of the population instead of any particular users.

In each round, the principal maintains an estimator $\hat{\theta}_{k,t}$ of arm k's quality from the observation of average feedback $\rho_{k,t}$. From Lemma 1, an asymptotically unbiased and consistent estimator of arm's quality $\hat{\theta}_{k,t}$ can be obtained by solving the equation:

$$F(\hat{\theta}_{k,t}, \rho_{k,t}) = \rho_{k,t} \qquad (3)$$

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 3. *Suppose A1 is satisfied. For any $\rho_{k,t}$, there exists a unique $\hat{\theta}_{k,t}$ that satisfies Equation 3.*

Given the convergence bounds and the estimator $\hat{\theta}_{k,t}$, we are ready to describe our proposed UCB-like algorithm Avg-UCB(β), as specified in Algorithm 1. The key differences to the standard UCB algorithms are that (1) we maintain a quality estimate $\hat{\theta}_{k,t}$ for each arm k at each time t, and (2)

the confidence interval in the UCB index is derived from the convergence rates as specified in Theorem 2. Our algorithm takes as input a parameter $\beta > 0$ to ensure the number of pulls of suboptimal arms with a logarithmic times.

Algorithm 1 Avg-UCB(β) 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 (3).
    $U_{k,t} = \hat{\theta}_{k,t-1} + \sqrt{\frac{\beta \ln(t-1)}{n_{k,t-1}^{2\bar{\lambda}}}}$.
  Choose arm $I_t \in \text{argmax}_{k=1,...,K} U_{k,t}$.
  (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(β).

Theorem 4. *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 [2], the expected regret for Avg-UCB(β) 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 the constant that is dependent on properties of function F.

The above regret bound is a gap-dependent bound. Let $\Delta_{\min} = \min_{k:k \neq I^*} \Delta_k$. The regret bound can be written as

$$\mathbb{E}[R(T)] = \mathcal{O}\left(\frac{(\ln T)^{\bar{\lambda}'}}{\Delta_{\min}^{2\bar{\lambda}'-1}}\right)$$

Given any constant $\bar{\lambda}' > 0$, $\lim_{T\to\infty} \mathbb{E}[R(T)]/T \to 0$. Therefore, the algorithm achieves sublinear regrets as long as G is strictly convex (i.e., $\bar{\lambda}' > 0$).

Moreover, as shown in the proof, we can derive gap-independent bounds from the above bound. For example, 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 if G is non-convex, the statements of Lemma 1 and 3 still hold, which means the stochastic process will still converge to some point, and we can infer the true arm quality from the converged average feedback. The main obstacle for us to overcome is to derive the convergence rate as in Theorem 2. This problem is challenging as it is equivalent to deriving the convergence rate for optimization for non-convex functions. Recently, there have been some recent work focusing on the derivation of the convergence rates for non-convex optimization in different settings (Allen-Zhu and Hazan 2016; Ge et al. 2015). As long as one could characterize the convergence rate of ρ_t even non-convex G, our bandit strategy can be adapted to generate a sublinear regret strategy (by changing the "confidence interval" in the UCB index according to the convergence rate).

[2]The choice of β depends on the parameters of $F(\theta, \rho)$. The detailed derivation is include in the appendix. β plays a similar role as the constant in UCB confidence radius to balance exploration and exploitation.

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 regrets.

However, in some scenarios, user feedback may be biased by not only the average feedback but also the number of feedback of the arm. It is natural to ask whether we can derive similar results by substituting $F(\theta, \rho)$ with more general $\texttt{Feedback}(\theta, \rho, n)$ in the update rule as specified in Equation 2. However, it turns out the stochastic process becomes a lot more complicated with beta-herding feedback model.

In this section, we explore another natural feedback model within the class of beta-herding feedback model and prove some 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 $\texttt{Feedback}(\theta, \rho, n) = (m\theta + n\rho)/(m+n)$.

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, when feedback

functions following 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 (Hill, Lane, and Sudderth 1980). We utilize the property of *exchangeability* for the feedback history to give the characterization of ρ_t process. The detailed discussion is included in appendix. Below we formally characterize the stochastic process of ρ_t given the above feedback model.

Lemma 5. *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,*

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

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.

The impossibility result

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

Lemma 5 implies that, even if we obtain an infinite number of feedback for an arm, we cannot accurately infer the arm quality with high probability from the empirical average feedback ρ_∞. The natural question to ask is, whether it is possible to infer the arm quality by taking into account all the feedback generated in the process? 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. Let $\mathcal{I}_t(\theta)$ denotes the Fisher information of θ for observing $t-$th sample. Since Fisher information is additive, we can show that

Lemma 6. $\lim_{t\to\infty} \sum_{i=1}^{t} \mathcal{I}_i(\theta) = \mathcal{O}(1)$

Lemma 6 shows that Fisher information about θ is bounded by a constant even if we observe infinitely many feedback. From the general Cramér–Rao bound, we know that, for any estimator $\hat{\theta}_t$, the variance of $\hat{\theta}_t$ must follow:

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

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 principal 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. Since the principal only observes the feedback, we can conclude the following.

Theorem 7. *If users' feedback follows beta-herding feedback model, there exists no bandit algorithm that can achieve sublinear regrets in our setting.*

Proof. 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 differentiable 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 6, 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. □

An alternative approach: Designing information structures

Theorem 7 presents a strong impossibility result: if all feedback are generated according to beta-herding feedback model, we cannot design any bandit algorithms to achieve sublinear regrets. A natural approach to get over this impossibility results is to break the assumption by taking interventions. Inspired by Bayesian persuasion (Kamenica and Gentzkow 2011), which designs the information structure to *persuade* agents to take certain actions, it is interesting to explore whether we could design information structure to induce certain types of "feedback". For example, in the extreme case, if we do not show any historical information to users, and we assume users provide unbiased feedback when no information is presented, then the problem reduces to standard bandit problem with existing solutions. However, in practice, we might not want to dramatically change the whole system and might want to take as few interventions as possible. This leads to an interesting research direction 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 a binary choice in information design, either showing all history information to users (and users' feedback follow beta-herding feedback model) or showing no history information (and users provide unbiased feedback). Our goal is to minimize the number of rounds showing no information to users while still being able to achieve sublinear regrets. In particular, we propose UCB(β)-MPA, 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:

Theorem 8. *Let* $\Theta = \{\theta_1, ..., \theta_K\}$ *be a bandit instance, and* $\alpha \geq \ln(K(K+2))/\ln T$*, then the expected regret of*

Algorithm 2 UCB(β)-MPA for Beta-Herding Feedback Model

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 $t = K+1, ..., \lfloor T^\alpha \rfloor$ **do**
  for each $k \in \{1, ..., K\}$ **do**
    $U_{k,t} = \hat{\theta}'_{k,t-1} + \sqrt{\frac{\beta \ln(t-1)}{n_{k,t-1}}}$, 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 $I_t \in \operatorname{argmax}_{k=1,...,K} U_{k,t}$.
  Present arm I_t without showing its history information to the agent, 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 $I_\tau \in \operatorname{argmax}_{k=1,...,K} n_{k,\lfloor T^\alpha \rfloor}$.
Present arm I_τ with associated history information to the agent in the remaining rounds.
(all ties broken in some consistent way)

UCB(β)-MPA, 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 in the RHS has an order of $\mathcal{O}\left((T - T^\alpha) \sqrt{\frac{K\beta\alpha \ln T}{T^\alpha}} \right)$.

When $\alpha \geq 1/2$, above regret bound has an order of $\mathcal{O}(\sqrt{\alpha T^\alpha \ln T})$, while when $\alpha < 1/2$, above regret about has 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. While we only consider the most naive approach, i.e., showing no information at all in some rounds, more fine-tuned ways of information design (e.g., only showing the average feedback) are also worth exploring.

Discussion

In this section, we discuss a few applications of our setting. First, as the motivating example of this paper, we consider platforms (such as Reddit, Yelp, or Amazon) that rely on user reviews to provide recommendations. To formulate the recommendation problem as a bandit learning problem, we have made a simplifying assumption, as made in prior work (Ghosh and Hummel 2013; Liu and Ho 2018), that users are going to follow the recommendations. In practice, this assumption approximates users' behavior reasonably well. In particular, empirical studies demonstrate that the probability for a users to check an item drops significantly when the position of the item decreases (Richardson, Dominowska, and Ragno 2007; Craswell et al. 2008; Joachims et al. 2017). As a concrete example, most users do not go beyond the first page when presented with multiple pages of items. 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 along the line of incentivizing exploration using information asymmetry (Kremer, Mansour, and Perry 2013; Mansour, Slivkins, and Syrgkanis 2015; Papanastasiou, Bimpikis, and Savva 2017) which demonstrate it is possible to make recommendations that users will *choose* to follow. The techniques studied in this paper can directly applied in that line of work to explore the dynamics of feedback generation.

In addition to the above example, our setting could be applied to other bandit learning problems when feedback is generated by humans and could be biased. Below we mention one application we find particular interesting. Consider the following illustrating scenario: the police station needs to decide which area to send polices to patrol at each time step. Each area i has an intrinsic, unknown crime rate p_i. When sending polices 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 monotone increasing with p_i. When polices are sent to an area, they have access to the history of *reported* crime rate of the area, and this information could bias their behavior and their reports. For example, they might stop more people for inspection if there are more reports of illegal activities in the area in the history. This creates biases in polices' reports. If the goal is to maximize the sum of $u(p_i)$, this problem can be formulated using our setting. One particular interesting point of this example is: assume the feedback model follows beta-herding feedback model or other time varying models. 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.

Conclusion and Future Work

We explore bandit problems under two different feedback models. We show that, in one feedback model, the updates of average feedback is mathematically equivalent to users collectively performing stochastic gradient descent. With this connection, we are able to design a UCB-like algorithm that achieves sublinear regrets under certain conditions. However, in another natural model, we show that there exists no bandit algorithms that can achieve sublinear regrets.

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

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

[Audibert and Bubeck 2009] Audibert, J.-Y., and Bubeck, S. 2009. Minimax policies for adversarial and stochastic bandits. In *The 22nd Conference on Learning Theory (COLT)*, 217–226.

[Auer et al. 1995] Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 1995. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, 322–331. IEEE.

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

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

[Besbes, Gur, and Zeevi 2014] Besbes, O.; Gur, Y.; and Zeevi, A. 2014. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems (NIPS)*, 199–207.

[Bikhchandani, Hirshleifer, and Welch 1992] Bikhchandani, S.; Hirshleifer, D.; and Welch, I. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy* 100(5):992–1026.

[Bubeck, Munos, and Stoltz 2011] Bubeck, S.; Munos, R.; and Stoltz, G. 2011. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science* 412(19):1832–1852.

[Craswell et al. 2008] Craswell, N.; Zoeter, O.; Taylor, M.; and Ramsey, B. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on Web Search and Data Mining (WSDM)*, 87–94. ACM.

[Frazier et al. 2014] Frazier, P.; Kempe, D.; Kleinberg, J.; and Kleinberg, R. 2014. Incentivizing exploration. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation (EC)*.

[Frikha, Menozzi, and others 2012] Frikha, N.; Menozzi, S.; et al. 2012. Concentration bounds for stochastic approximations. *Electronic Communications in Probability* 17.

[Garivier and Moulines 2011] Garivier, A., and Moulines, E. 2011. On upper-confidence bound policies for switching bandit problems. In Kivinen, J.; Szepesvári, C.; Ukkonen, E.; and Zeugmann, T., eds., *Algorithmic Learning Theory*, 174–188. Berlin, Heidelberg: Springer Berlin Heidelberg.

[Ge et al. 2015] Ge, R.; Huang, F.; Jin, C.; and Yuan, Y. 2015. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory (COLT)*, 797–842.

[Ghosh and Hummel 2013] Ghosh, A., and Hummel, P. 2013. 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)*.

[Hill, Lane, and Sudderth 1980] Hill, B. M.; Lane, D.; and Sudderth, W. 1980. A strong law for some generalized urn processes. *The Annals of Probability* 214–226.

[Joachims et al. 2017] Joachims, T.; Granka, L.; Pan, B.; Hembrooke, H.; and Gay, G. 2017. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, volume 51, 4–11. ACM.

[Kamenica and Gentzkow 2011] Kamenica, E., and Gentzkow, M. 2011. Bayesian persuasion. *American Economic Review* 101(6):2590–2615.

[Kremer, Mansour, and Perry 2013] Kremer, I.; Mansour, Y.; and Perry, M. 2013. Implementing the "wisdom of the crowd". In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce (EC)*, 605–606.

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

[Liu and Ho 2018] Liu, Y., and Ho, C.-J. 2018. Incentivizing high quality user contributions: New arm generation in bandit learning. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[Mansour, Slivkins, and Syrgkanis 2015] Mansour, Y.; Slivkins, A.; and Syrgkanis, V. 2015. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation (EC)*, 565–582. ACM.

[Muchnik, Aral, and Taylor 2013] Muchnik, L.; Aral, S.; and Taylor, S. J. 2013. Social influence bias: A randomized experiment. *Science* 341(6146):647–651.

[Neu et al. 2010] Neu, G.; Antos, A.; György, A.; and Szepesvári, C. 2010. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems (NIPS)*, 1804–1812.

[Ortner et al. 2012] Ortner, R.; Ryabko, D.; Auer, P.; and Munos, R. 2012. Regret bounds for restless markov bandits. In *International Conference on Algorithmic Learning Theory (ALT)*, 214–228. Springer.

[Papanastasiou, Bimpikis, and Savva 2017] Papanastasiou, Y.; Bimpikis, K.; and Savva, N. 2017. Crowdsourcing exploration. *Management Science*.

[Pemantle 1991] Pemantle, R. 1991. When are touchpoints limits for generalized pólya urns? *Proceedings of the American Mathematical Society* 113(1):235–243.

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

[Robbins and Monro 1951] Robbins, H., and Monro, S. 1951. A stochastic approximation method. *The annals of mathematical statistics* 400–407.

[Salganik, Dodds, and Watts 2006] Salganik, M. J.; Dodds, P. S.; and Watts, D. J. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762):854–856.

[Sipos, Ghosh, and Joachims 2014] Sipos, R.; Ghosh, A.; and Joachims, T. 2014. 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)*, 337–348.

[Smith and Sørensen 2000] Smith, L., and Sørensen, P. 2000. Pathological outcomes of observational learning. *Econometrica* 68(2):371–398.