

# (Private) Kernelized Bandits with Distributed Biased Feedback

Fengjiao Li  
Virginia Tech  
Blacksburg, VA, USA  
fengjiaoli@vt.edu

Xingyu Zhou  
Wayne State University  
Detroit, MI, USA  
xingyu.zhou@wayne.org

Bo Ji  
Virginia Tech  
Blacksburg, VA, USA  
boji@vt.edu

## ABSTRACT

We study kernelized bandits with distributed biased feedback. This problem is motivated by several real-world applications (such as dynamic pricing, cellular network configuration, and policy making), where users from a large population contribute to the reward of the action chosen by a central entity, but it is difficult to collect feedback from all users. Instead, only biased feedback (due to user heterogeneity) from a subset of users may be available. In addition to such biased feedback, we are also faced with two practical challenges due to communication cost and computation complexity. To tackle these challenges, we carefully design a new *distributed phase-then-batch-based elimination* (DPBE) algorithm, which samples users in phases for collecting feedback to reduce the bias and employs *maximum variance reduction* to select actions in batches within each phase. By properly choosing the phase length, the batch size, and the confidence width used for eliminating suboptimal actions, we show that DPBE achieves a sublinear regret of  $\tilde{O}(T^{1-\alpha/2} + \sqrt{YT})$ , where  $\alpha \in (0, 1)$  is the user-sampling parameter one can tune. Moreover, DPBE can significantly reduce both communication cost and computation complexity in distributed kernelized bandits, compared to some variants of the state-of-the-art algorithms (originally developed for standard kernelized bandits). Furthermore, by incorporating various *differential privacy* models, we generalize DPBE to provide privacy guarantees for users participating in the distributed learning process. The algorithm design, analyses, and numerical experiments are provided in the full version of this paper [4].

## KEYWORDS

Kernelized bandits, Distributed feedback, Bias, Regret, Communication cost, Computation complexity, Privacy

### ACM Reference Format:

Fengjiao Li, Xingyu Zhou, and Bo Ji. 2023. (Private) Kernelized Bandits with Distributed Biased Feedback. In *Abstract Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '23 Abstracts)*, June 19–23, 2023, Orlando, FL, USA. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3578338.3593565>

## 1 INTRODUCTION

Bandit optimization is a popular online learning paradigm for sequential decision-making. It has been widely used in a wide variety of real-world applications, including hyperparameter tuning [6],

recommendation systems [5], and dynamic pricing [7]. In such problems, if chosen, each decision point (called an arm or action) yields an unknown reward. The agent's goal is to maximize the cumulative reward by making proper decisions sequentially. An important way to capture general (e.g., *non-linear* and even *non-convex*) unknown objective functions is to consider a smoothness condition specified by a small norm of a Reproducing Kernel Hilbert Space (RKHS) associated with a kernel function. This setup is often referred to as *kernelized bandits*. Thanks to the strong link between RKHS functions and Gaussian processes (GP) [1], an extensive line of work has exploited GP models to estimate an unknown function  $f$  given a set of (noisy) evaluations of its values  $f(\mathbf{x})$  at chosen actions  $\mathbf{x}$ .

**Motivation.** We are motivated to consider a different but commonly seen scenario where the value  $f(\mathbf{x})$  represents an overall effect of action  $\mathbf{x}$  on a large population of users where it is difficult for the learning agent to make direct observations; yet, the agent could collect some biased feedback from the distributed users in the population due to user heterogeneity (e.g., different preferences). Here, the observed feedback at each user  $u$  in the population is associated with a local function value  $f_u(\mathbf{x})$ . Consider the dynamic pricing problem [7]. When a company sets a pricing mechanism  $\mathbf{x}$ , this decision influences all the customers, and every customer, based on her individual demand and preference, makes a choice (purchase or not), which contributes to the total profits  $f(\mathbf{x})$ . Without knowing products' demand curves in advance, the company makes a sequence of pricing decisions with the goal of *maximizing profits while learning*. That is, the company aims to infer the expected demand and thus the expected profits  $f$  by collecting feedback from customers in each decision epoch. Note that it might be difficult for the company to collect feedback from *all* the customers - since purchases may take place at many local stores at different locations. For example, it is impractical for McDonald's headquarters to collect sales information from all of the nationwide customers within each decision epoch. Instead, the headquarter might be able to get feedback (i.e., sales information) from a subset of the customers. However, each customer's choice depends not only on her own preference towards the products and their prices but also on several other factors (location, competitors, promotion events, etc.), which is often *biased* feedback for the overall profits.

**Problem Statement.** We introduce a new kernelized bandit problem where the unknown function represents the overall reward over a large population containing an infinite number of users. The unknown reward function  $f : \mathcal{D} \rightarrow \mathbb{R}$  is assumed to be in an RKHS associated with kernel  $k$ , denoted by  $\mathcal{H}_k$ . At round  $t$ , the agent chooses an action  $\mathbf{x}_t \in \mathcal{D}$ , leading to a reward with mean  $f(\mathbf{x}_t)$ . This reward is unknown to the agent but captures the overall effectiveness of action  $\mathbf{x}_t$  over the entire population  $\mathcal{U}$ , thus called *global reward*. Meanwhile, each user  $u$  in the population observes a (noisy)

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGMETRICS '23 Abstracts, June 19–23, 2023, Orlando, FL, USA

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0074-3/23/06.

<https://doi.org/10.1145/3578338.3593565>

*local reward*:  $y_{u,t} = f_u(\mathbf{x}_t) + \eta_{u,t}$  with mean  $f_u(\mathbf{x}_t)$ , where  $\eta_{u,t}$  is the noise, and  $f_u : \mathcal{D} \rightarrow \mathbb{R}$  is the local reward function, assumed to be an (unknown) realization of a random function with mean  $f$ . Assume  $f_u \sim \mathcal{GP}(f(\cdot), k(\cdot, \cdot))$ . In this setting, the exact global reward corresponding to the entire population cannot be observed; only biased local reward feedback is available to the agent. We refer to this setting as *kernelized bandits with distributed biased feedback*.

The goal of the agent is to maximize the cumulative global reward, or equivalently, to minimize the regret defined as follows:

$$R(T) \triangleq \sum_{t=1}^T \left( \max_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x}) - f(\mathbf{x}_t) \right). \quad (1)$$

As stated before, the agent has to learn the unknown function  $f$  by communicating (biased) feedback from some distributed users (i.e., *learning with communication*). We let the feedback communicated in phases and let  $L$  be the total number of phases in  $T$  rounds. Communication cost, as a critical factor in general distributed systems, is measured by the total quantity of communicated numbers (between the agent and all users). Let  $U_l$  be the set of users that report their feedback in the  $l$ -th phase and  $N_{u,l}$  be the number of scalars in user  $u$ 's feedback for  $u \in U_l$ . Then, the total communication cost is  $C(T) \triangleq \sum_{l=1}^L \sum_{u \in U_l} N_{u,l}$ .

**Challenges.** The above “communicating feedback in phases” naturally leads us to consider a phased elimination algorithm that gradually eliminates suboptimal actions by periodically aggregating and analyzing the local feedback from the participants. However, several new challenges arise in our setting compared to the standard phase elimination algorithm in linear bandits [2, 3]. First, *how to select actions for each phase?* Due to the possible infinite feature dimension of RKHS functions, it is nontrivial to adopt the so-called near-optimal experimental design (used in standard phased elimination algorithms and related to the (finite) dimension of actions) to kernelized bandit setting. We wonder if there is a simple and efficient method of selecting actions in each phase for our kernelized bandits setting. Second, *how to use biased feedback?* In order to reduce the impact of bias, an efficient user-sampling scheme is needed. However, how to incorporate this idea into the phase elimination algorithm is unclear. Last but not least, *how to deal with scalability?* In our setting, scalability refers to both computation complexity and communication cost. On the one hand, it is well-known that standard kernelized/GP bandits have a poor computation complexity (e.g.,  $O(T^3)$ ) due to the matrix inverse at each step for GP posterior update. On the other hand, due to the communication between the agent and the users, it is imperative to ensure a low communication cost.

**Contributions.** In this work, we address the above challenges and make the following contributions.

1) To the best of our knowledge, this is the first work that studies a new kernelized bandit setting with distributed biased feedback, where three key challenges (user heterogeneity, communication efficiency, and computation complexity) inherently arise in the design of sample-efficient, scalable learning algorithms. To solve this bandits problem, we propose the *learning with communication* framework where the biased feedback is communicated in phases and design a new phased elimination algorithm that aggregates the

distributed biased feedback and eliminates suboptimal actions in a computation-efficient manner.

2) Specifically, we design a new *distributed phase-then-batch-based elimination* algorithm which is carefully crafted to address all the aforementioned challenges. In particular, DPBE adds a *user-sampling* process to reduce the impact of bias from each individual user and selects actions according to *maximum variance reduction* within each phase. Moreover, a *batching* strategy is employed to improve both communication efficiency and computation complexity. That is, instead of selecting a new action at each round, DPBE plays the same action for a batch of rounds before switching to the next one. Not only does it save the number of times one needs to compute the next action via GP update, but it also reduces the dimensions of the vectors and matrices involved in both communication and computation via a reformulation.

3) We show that DPBE achieves a sublinear regret of  $\tilde{O}(T^{1-\alpha/2} + \sqrt{\gamma_T T})$  while incurring a communication cost of  $O(\gamma_T T^\alpha)$  and a computation complexity of  $O((|\mathcal{D}|\gamma_T^3 + \gamma_T^4) \log T + \gamma_T T^\alpha)$ , where  $\gamma_T$  is the *maximum information gain* associated with the kernel of the unknown function  $f$ ,  $\mathcal{D}$  is the decision set, and  $\alpha > 0$  is a user-sampling parameter that one can tune. It is worth noting that DPBE with  $\alpha \in (0, 1)$  has a better computation complexity than some variants of the state-of-the-art algorithms (originally developed for standard kernelized bandits without biased feedback). Specifically, DPBE achieves three significant improvements compared to the state-of-the-art algorithms: (i) user-sampling efficiency ( $O(T^\alpha)$  vs.  $T$ ), (ii) communication cost ( $O(\gamma_T T^\alpha)$  vs.  $T$ ), and (iii) computation complexity ( $O(\gamma_T T^\alpha)$  vs.  $O(T^3)$ ). We also conduct extensive simulations to validate our theoretical results and evaluate the empirical performance of regret, communication cost, and running time.

4) We generalize our phase-then-batch framework to incorporate various *differential privacy* (DP) models (including the central, local, and shuffle models) into DPBE, which ensures privacy guarantees for users participating in the distributed learning process.

## ACKNOWLEDGMENTS

We thank our shepherd, Giulia Fanti, and the anonymous paper reviewers for their insightful feedback. We also thank Duo Cheng for fruitful discussions. This work is supported in part by the NSF grants under CNS-2112694 and CNS-2153220.

## REFERENCES

- [1] Sayak Ray Chowdhury and Aditya Gopalan. 2017. On kernelized multi-armed bandits. In *International Conference on Machine Learning*. PMLR, 844–853.
- [2] Tor Lattimore and Csaba Szepesvári. 2020. *Bandit algorithms*. Cambridge University Press.
- [3] Fengjiao Li, Xingyu Zhou, and Bo Ji. 2022. Differentially Private Linear Bandits with Partial Distributed Feedback. *arXiv preprint arXiv:2207.05827* (2022).
- [4] Fengjiao Li, Xingyu Zhou, and Bo Ji. 2023. (Private) Kernelized Bandits with Distributed Biased Feedback. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 7, 1, Article 5 (mar 2023), 47 pages. <https://doi.org/10.1145/3579318>
- [5] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*. 661–670.
- [6] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. 2017. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research* 18, 1 (2017), 6765–6816.
- [7] Kanishk Misra, Eric M Schwartz, and Jacob Abernethy. 2019. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science* 38, 2 (2019), 226–252.