Stochastic Multi-Player Bandit Learning from Player-Dependent Feedback

Zhi Wang[†] ZHIWANG@ENG.UCSD.EDU
Manish Kumar Singh[†] MKSINGH@ENG.UCSD.EDU
Chicheng Zhang[§] CHICHENGZ@CS.ARIZONA.EDU
Laurel D. Riek[†] LRIEK@ENG.UCSD.EDU
Kamalika Chaudhuri[†] KAMALIKA@CS.UCSD.EDU
[†]*Department of Computer Science and Engineering, University of California, San Diego*
[§]*Department of Computer Science, University of Arizona*

Abstract

We investigate robust data aggregation in a multi-agent online learning setting. In reality, multiple online learning agents are often deployed to perform similar tasks and receive similar feedback. We study how agents can improve their collective performance by sharing information among each other. In this paper, we formulate the *ϵ-multi-player multi-armed bandit* problem, in which a set of M players that have similar reward distributions for each arm play concurrently. We develop an upper confidence bound-based algorithm that adaptively aggregates rewards collected by different players. To our best knowledge, we are the first to develop such a scheme in a multi-player bandit learning setting. We show that under the assumption that pairwise distances between the means of the player-dependent distributions for each arm are small, we improve the (collective) regret bound by nearly a factor of M, in comparison with a baseline algorithm in which the players learn individually using the UCB-1 algorithm (Auer et al., 2002). Our algorithm also exhibits a fallback guarantee, namely, if our task similarity assumption fails to hold, our algorithm still has a performance guarantee that cannot be worse than the baseline by a constant factor. Empirically, we validate our algorithm on synthetic data.

Keywords: Stochastic multi-player bandit learning, heterogeneous data aggregation

1. Introduction

Online learning has many important real-world applications (see Villar et al., 2015; Shen et al., 2015; Li et al., 2010, for a few examples). In practice, a group of online learning agents are often deployed for similar tasks, and they learn to perform these tasks in similar yet nonidentical environments. One natural question arises: Can the agents collaborate to achieve a better collective reward? In this paper, we study robust aggregation of data collected by multiple online learning agents that perform similar tasks.

Consider the following application scenario: a group of assistive robots are deployed to provide personalized cognitive training to people with dementia (PwD), e.g., by teaching metacognitive skills, supporting healthy lifestyle choices, and playing stimulating games (Kubota et al., 2020). In order to ensure the health intervention is useful, adopted, and adhered to, it is critical that it is *tailored* to the individual, and able to adapt to the person

as they change over time (a key characteristic of PwD) (Riek, 2017; Woodworth et al., 2018). The goal of each robot is to learn the preferences of its paired individual in an *online* setting—each robot seeks to discover and recommend activities that a PwD favors and enjoys based on how the PwD reacts to and is engaged with the task (as captured by sensors on the robots) (Kubota et al., 2020). As PwD may have similar preferences and may therefore exhibit similar reactions, the robots as a multi-agent system can potentially learn to perform their respective tasks faster by sharing information with each other.

In this paper, we generalize the the multi-armed bandit problem (Auer et al., 2002) and formulate the ϵ-multi-player multi-armed bandit (ϵ-MPMAB) problem, which models *heterogeneous multi-task learning* in a multi-agent bandit learning setting. In an ϵ-MPMAB problem instance, a set of M players are deployed to perform similar tasks—simultaneously they interact with a set of actions/arms, and they receive feedback from different reward distributions for taking the same action/pulling the same arm. In the above assistive robotics example, each player corresponds to a robot; each arm corresponds to one of the cognitive activities to choose from; for each player and each arm, there is a separate reward distribution which reflects a PwD's personal preferences. $\epsilon \geq 0$ is a discrepancy parameter that upper bounds the pairwise distances between different reward distributions for different players on the same arm. The players can communicate and share information among each other, with a goal of minimizing their collective regret.

While multi-player bandit learning has been studied extensively in the literature (e.g., (Landgren et al., 2016; Cesa-Bianchi et al., 2013; Gentile et al., 2014)), and warm-starting bandit learning using a different feedback source has also been investigated recently (Zhang et al., 2019), to our best knowledge, there is no prior work that studies bandit learning for similar tasks in a multi-player setting with a focus on *robust* data aggregation based upon the (dis)similarities between the sources of data. We believe this problem is an important addition to the literature on collaborative online learning and multi-task bandit learning.

It is worth noting that naively utilizing the data collected by other players may substantially hurt a player's regret (Zhang et al., 2019), if there are large disparities between the sources of feedback. This is also known as *negative transfer* in the transfer learning literature (Rosenstein et al., 2005; Brunskill and Li, 2013).

In this paper, we propose an upper confidence bound (UCB)-based algorithm that *adaptively* aggregates rewards collected by different players and is *robust* against negative transfer. To our best knowledge, this is the first such algorithm for multi-player bandit learning. We provide performance guarantees for our algorithm. We show that when the discrepancy parameter ϵ is small, we improve the collective regret bound by nearly a factor of M, in comparison with a baseline algorithm in which the players learn individually using the UCB-1 algorithm (Auer et al., 2002). Our algorithm also exhibits robustness—we show a fallback guarantee: when ϵ is large and it is unsafe for the players to aggregate data aggressively, our algorithm still has a performance guarantee no worse than that of the baseline algorithm by a constant factor. We validate our algorithm empirically on synthetic data.

2. Problem Specification

We formulate the ϵ-MPMAB problem, building on the standard model of stochastic multi-armed bandits (Lai and Robbins, 1985; Auer et al., 2002).

We consider an ϵ-MPMAB problem instance with a set of M players, labeled as $1, 2, \ldots, M$. *Concurrently*, the players interact with a set of K arms, labeled as $1, 2, \ldots, K$. For each player $p \in [M]$, each arm $i \in [K]$ is associated with an unknown reward distribution $\mathcal{D}_i^p$ with support $[0, 1]$ and mean μ_i^p. The reward distributions of the same arm are not necessarily identical for different players, but we assume them to be similar.

Assumption 1 *For every pair of players $p, q \in [M]$, $\max_{i\in[K]} |\mu_i^p - \mu_i^q| \leq \epsilon$, where $\epsilon \in [0, 1]$.*

In each round $t = 1, 2, \ldots, T$, each player $p \in [M]$ pulls an arm i_t^p, and observes an independent and identically distributed reward $r_{i_t,t}^p \sim \mathcal{D}_{i_t}^p$.[1] Once all the players finish pulling an arm in round t, each decision, i_t^p, together with the corresponding reward received, $r_{i_t,t}^p$, is instantaneously shared with every other player in $[M]$.

Let $\mu_*^p = \max_{i\in[K]} \mu_i^p$ for every player $p \in [M]$. Denote $n_i^p(t)$ as the number of pulls of arm i by player p after t rounds, and $\Delta_i^p = \mu_*^p - \mu_i^p \geq 0$ as the gap between the means of the reward distributions associated with the optimal arm i_*^p and arm i for player p. For simplicity, we assume that for each player, there exists one unique optimal arm. Then, the expected regret of player p can be stated as $\mathbb{E}[\mathcal{R}^p(T)] = \sum_{i=1}^K \Delta_i^p \cdot \mathbb{E}[n_i^p(T)]$. In an ϵ-MPMAB problem, the goal is to to minimize the players' expected collective regret, defined as $\mathbb{E}[\mathcal{R}(T)] = \sum_{p=1}^M \mathbb{E}[\mathcal{R}^p(T)]$.

3. Related Work and Comparisons

In this section, we compare existing multi-agent bandit learning problems with the ϵ-MPMAB problem. We provide a more detailed review of the literature in Appendix A.

A large portion of prior studies (Kar et al., 2011; Szörényi et al., 2013; Landgren et al., 2016; Kolla et al., 2018; Sankararaman et al., 2019; Wang et al., 2019) focuses on the setting where a network of players collaboratively work on one bandit learning problem instance, i.e., the reward distributions of an arm are identical across all players. In contrast, we study multi-agent bandit learning where the reward distributions across players can be different.

Multi-agent bandit learning with player-dependent rewards has also been covered by previous studies. In (Shahrampour et al., 2017), a group of players seek to find the arm with the largest average reward over all the players; however, in each round, the players have to reach a consensus and choose the same arm. (Cesa-Bianchi et al., 2013) studies a network of linear contextual bandit players with player-dependent rewards—the players propagate information based on their affinity which is specified by a graph. In (Gentile et al., 2014), players are clustered on the fly and share feedback information with other players in the same cluster. However, neither of these papers focuses on robust aggregation of data shared by other players. In this paper, we study how data can be safely and adaptively aggregated based on a pre-defined discrepancy parameter.

Similarities in reward distributions are explored in (Zhang et al., 2019), which studies a warm-start scenario, in which data are provided as history (Shivaswamy and Joachims, 2012) for an learning agent to explore faster. In this paper, however, we study the multi-player setting, where all players learn continually and concurrently.

In the interest of space, we defer the rest of our comparisons to Appendix A.

1. We will use $\mathcal{D}_{i_t}^p$, $\mu_{i_t}^p$, and $r_{i_t,t}^p$ in place of $\mathcal{D}_{i_t^p}^p$, $\mu_{i_t^p}^p$, and $r_{i_t^p,t}^p$, respectively.

Algorithm 1: Robust Learning in ϵ-MPMAB

Input: Distribution discrepancy parameter $\epsilon \in [0, 1]$;

Initialization: Set $n_i^p = 0$ for all $p \in [M]$ and all $i \in [K]$.

for $t = 1, 2 \ldots, T$ **do**

  for $p \in [M]$ **do**

    for $i \in [K]$ **do**

      Let $m_i^p = \sum_{q \in [M]: q \neq p} n_i^q$;

      Let $\overline{n_i^p} = \max(1, n_i^p)$ and $\overline{m_i^p} = \max(1, m_i^p)$;

      Let $F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon) = 8\sqrt{6 \ln T [\frac{\lambda^2}{\overline{n_i^p}} + \frac{(1-\lambda)^2}{\overline{m_i^p}}]} + (1 - \lambda)\epsilon$;

      Compute $\lambda^* = \text{argmin}_{\lambda \in [0,1]} F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon)$;

      Let

$$\zeta_i^p(t) = \frac{1}{\overline{n_i^p}} \sum_{\substack{s<t \\ i_s^p = i}} r_{i_s,s}^p,\ \eta_i^p(t) = \frac{1}{\overline{m_i^p}} \sum_{\substack{q \in [M] \\ q \neq p}} \sum_{\substack{s<t \\ i_s^q = i}} r_{i_s,s}^q, \text{and } \kappa_i^p(t, \lambda) = \lambda \zeta_i^p(t) + (1-\lambda)\eta_i^p(t);$$

      Compute the upper confidence bound of the reward of arm i for player p:

$$\text{UCB}_i^p(t) = \kappa_i^p(t, \lambda^*) + F(\overline{n_i^p}, \overline{m_i^p}, \lambda^*, \epsilon).$$

    Let $i_t^p = \text{argmax}_{i \in [K]} \text{UCB}_i^p(t)$;

    Player p pulls arm i_t^p and observes reward $r_{i_t,t}^p$;

  for $p \in [M]$ **do**

    Let $i = i_t^p$ and set $n_i^p = n_i^p + 1$.

4. Algorithm

In this section, we provide an algorithm, namely Algorithm 1, that robustly aggregates samples collected by different players in the ϵ-MPMAB problem. We first provide some intuition of the algorithm. In any round, a player may decide to take advantage of data from other players, depending on the sample size of their own collected rewards, the sample sizes of other players' rewards, as well as the discrepancy between the players' reward distributions. These factors are extensively discussed in (Ben-David et al., 2010). Our algorithm is built upon this insight of trading off (i) a smaller deviation of the empirical mean from the true mean due to an increased number of samples against (ii) the inaccuracy of our estimate due to the discrepancy in the distributions.

We consider an adaptive aggregation and weighting of samples collected by the players, which can lead to tighter confidence bounds on mean rewards. As is shown in (Auer, 2002), confidence bounds is instrumental for designing stochastic bandit algorithms. More specifically, our algorithm is based on the UCB-1 algorithm (Auer et al., 2002) and maintains a confidence interval for each mean μ_i^p such that "with high probability," the empirical estimate of the mean (a weighted combination) always lies in the confidence interval. We use the number of samples collected by each player, the discrepancy parameter ϵ, and a

weighting factor λ to minimize the width of confidence intervals (similar to (Ben-David et al., 2010)).[2] This data aggregation procedure can potentially allow us to have tighter confidence intervals for the players' reward estimates, compared to confidence intervals constructed solely on the player's own data, and therefore achieve better regret guarantees.

This idea of assigning weights to samples have also been studied in (Zhang et al., 2019) for warm starting contextual bandits from misaligned distributions, and in (Russac et al., 2019) for online learning in non-stationary environments. To our best knowledge, we are the first to investigate *adaptive* aggregation of data in a multiplayer bandit learning setting.

5. Performance Guarantees

In this section, we provide regret analyses for Algorithm 1. We first provide an upper bound on the collective regret of Algorithm 1 under a further assumption. Due to space constraints, the proofs are deferred to Appendices B, C and D.

Assumption 2 *For any player $p \in [M]$ and any nonoptimal arm $i \neq i_*^p \in [K]$, $\Delta_i^p = \mu_*^p - \mu_i^p > 2\epsilon$.*

Fact 1 *Under Assumption 2, the optimal arms for each player $p \in [M]$ have the same index, henceforth denoted as $i_* \in [K]$.*

Theorem 1 *Let Algorithm 1 run on an ϵ-MPMAB problem instance. Define $\Delta_i^{\min} = \min_p \Delta_i^p > 2\epsilon$, and similarly, $\Delta_i^{\max} = \max_p \Delta_i^p$. Then, under Assumption 2, the expected collective regret in a horizon of T rounds satisfies*

$$\mathbb{E}[\mathcal{R}(T)] \leq O\left(\sum_{\substack{i\in[K]\\ i\neq i_*}} \frac{\ln T}{\left(\Delta_i^{\min} - 2\epsilon\right)^2} \cdot \Delta_i^{\max} + KM\right).$$

Consider an algorithm that runs the UCB-1 algorithm (Auer et al., 2002) individually for each player—hereinafter, we refer to this algorithm as the *baseline* algorithm.

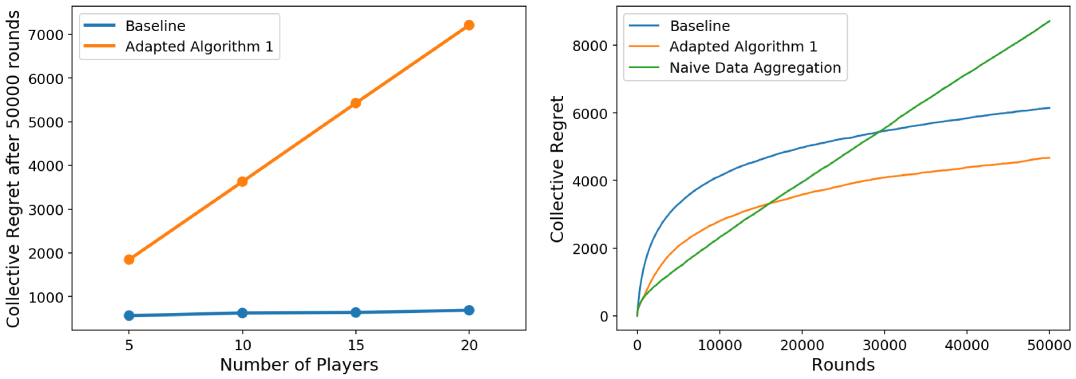
Remark 2 *By (Auer et al., 2002, Theorem 1 thereof), the expected collective regret of the baseline algorithm satisfies $\mathbb{E}[\mathcal{R}(T)] \leq O\left(\sum_{p\in[M]} \sum_{\substack{i\in[K]\\ \mu_i^p<\mu_*^p}} \frac{\ln T}{\Delta_i^p} + KM\right)$.*

It is easy to observe that, compared to the baseline algorithm, we improve the regret bound by nearly a factor of M—if we set aside the $O(KM)$ term, then the expected collective regret in Theorem 1 does not have a dependence on M.

The following lemma shows that the regret guarantee of Algorithm 1 is never worse than that of the baseline algorithm by a constant factor, even when Assumption 2 does not hold.

Lemma 3 (Fallback Gaurantee) *The expected collective regret of Algorithm 1 satisfies $\mathbb{E}[\mathcal{R}(T)] \leq O\left(\sum_{p\in[M]} \sum_{\substack{i\in[K]\\ \mu_i^p<\mu_*^p}} \frac{\ln T}{\Delta_i^p}\right)$.*

2. See Appendix F for an analytical solution to the optimal weighting factor λ^*.



(a) Experiment-1 (b) Experiment-2

Figure 1: (a) shows the collective regrets of adapted Algorithm 1 and the baseline algorithm after $T = 50000$ rounds, with $M \in \{5, 10, 15, 20\}$; (b) shows the collective regrets of adapted Algorithm 1, adapted Algorithm 1 assuming $\epsilon = 0$ (a naive data aggregation algorithm), and the baseline algorithm, over a horizon of $T = 50000$ rounds.

6. Empirical Validation

Since standard generalization bounds are loose, we performed experiments on a more practical and aggressive algorithm adapted from Algorithm 1.[3] A pseudocode for this slightly different algorithm (hereinafter called "adapted Algorithm 1") can be found in Appendix E.

We provide an empirical validation of the performance of adapted Algorithm 1 in terms of its collective regret and robustness. We performed two experiments using synthetic data with $K = 10$ arms. In the interest of space, we defer a more detailed description of the experimental setup to Appendix E.

Experiment 1. We study the dependence of collective regret on the number of players. For $M = 5, 10, 15$, and 20 players, we each generated $C = 30$ ϵ-MPMAB instances under Assumptions 1 and 2 with $\epsilon = 0.1$. On each problem instance, we ran adapted Algorithm 1 and the baseline algorithm. Figure 1a compares the averaged collective regrets of the algorithms over C instances after $T = 50000$ rounds for different choices of M. The results show that the collective regret of adapted Algorithm 1 is insensitive to the number of players.

Experiment 2. We investigate the robustness of Algorithm 1. With $M = 10$ players, we generated $C = 20$ ϵ-MPMAB problem instances under Assumption 1 only (with $\epsilon = 0.2$). We ran (1) adapted Algorithm 1 given $\epsilon = 0.2$; (2) adapted Algorithm 1 given $\epsilon = 0$ (naive data aggregation); and (3) the baseline algorithm. Figure 1b compares the collective regrets of the three algorithms averaged over C instances. The result indicates the importance of robust data aggregation as well as the robustness of our approach.

3. We added an initialization phase to be consistent with the baseline algorithm which includes one (this also ensures that $n_i^p, m_i^p > 0$ for all p) and used a more aggressive upper confidence bound with length $\min_{\lambda \in [0,1]} \sqrt{2 \ln T [\frac{\lambda^2}{n_i^p} + \frac{(1-\lambda)^2}{m_i^p}]} + (1-\lambda)\epsilon$.

Acknowledgments

We thank Geelon So and Gaurav Mahajan for insightful discussions. We also thank the National Science Foundation under 1915734 for research support.

References

Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

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

Baruch Awerbuch and Robert D Kleinberg. Competitive collaborative learning. In *International Conference on Computational Learning Theory*, pages 233–248. Springer, 2005.

Yogev Bar-On and Yishay Mansour. Individual regret in cooperative nonstochastic multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3110–3120, 2019.

Peter L Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. 2008.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Emma Brunskill and Lihong Li. Sample complexity of multi-task reinforcement learning. *UAI*, 2013.

Sébastien Bubeck and Thomas Budzinski. Coordination without communication: optimal regret in two players multi-armed bandits, 2020.

Swapna Buccapatnam, Atilla Eryilmaz, and Ness B Shroff. Stochastic bandits with side observations on networks. In *The 2014 ACM international conference on Measurement and modeling of computer systems*, pages 289–300, 2014.

Stéphane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. Leveraging side observations in stochastic bandits. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, page 142–151, Arlington, Virginia, USA, 2012. AUAI Press. ISBN 9780974903989.

Nicolo Cesa-Bianchi, Claudio Gentile, and Giovanni Zappella. A gang of bandits. In *Advances in Neural Information Processing Systems*, pages 737–745, 2013.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Delay and cooperation in nonstochastic bandits. *The Journal of Machine Learning Research*, 20(1):613–650, 2019.

Konstantina Christakopoulou and Arindam Banerjee. Learning to interact with users: A collaborative-bandit approach. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 612–620. SIAM, 2018.

Aniket Anand Deshmukh, Urun Dogan, and Clay Scott. Multi-task learning for contextual bandits. In *Advances in neural information processing systems*, pages 4848–4856, 2017.

Raphael Feraud, Reda Alami, and Romain Laroche. Decentralized exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1901–1909, 2019.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *International Conference on Machine Learning*, pages 757–765, 2014.

Soummya Kar, H Vincent Poor, and Shuguang Cui. Bandit problems in networks: Asymptotically efficient distributed allocation rules. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 1771–1778. IEEE, 2011.

Ravi Kumar Kolla, Krishna Jagannathan, and Aditya Gopalan. Collaborative learning of stochastic bandits over a social network. *IEEE/ACM Transactions on Networking*, 26 (4):1782–1795, 2018.

Alyssa Kubota, Emma IC Peterson, Vaishali Rajendren, Hadas Kress-Gazit, and Laurel D Riek. Jessie: Synthesizing social robot behaviors for personalized neurorehabilitation and beyond. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 121–130, 2020.

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

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 167–172. IEEE, 2016.

Peter Chal Landgren. *Distributed Multi-Agent Multi-Armed Bandits*. PhD thesis, Princeton University, 2019.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 661–670, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772758. URL `https://doi.org/10.1145/1772690.1772758`.

Shuai Li, Wei Chen, Shuai Li, and Kwong-Sak Leung. Improved algorithm on online clustering of bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2923–2929. AAAI Press, 2019.

Keqin Liu and Qing Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, 58(11):5667–5681, 2010.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pages 684–692, 2011.

Laurel D Riek. Healthcare robotics. *Communications of the ACM*, 60(11):68–78, 2017.

Michael T. Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G. Dietterich. To transfer or not to transfer. In *NIPS 2005 workshop on transfer learning*, 2005.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, pages 12017–12026, 2019.

Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–35, 2019.

Shahin Shahrampour, Alexander Rakhlin, and Ali Jadbabaie. Multi-armed bandits in multi-agent networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2786–2790. IEEE, 2017.

Nihal Sharma, Soumya Basu, Karthikeyan Shanmugam, and Sanjay Shakkottai. Warm starting bandits with side information from confounded data. *arXiv preprint arXiv:2002.08405*, 2020.

Weiwei Shen, Jun Wang, Yu-Gang Jiang, and Hongyuan Zha. Portfolio choices with orthogonal bandit learning. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 974–980. AAAI Press, 2015. ISBN 9781577357384.

Pannagadatta Shivaswamy and Thorsten Joachims. Multi-armed bandit problems with history. In *Artificial Intelligence and Statistics*, pages 1046–1054, 2012.

Aleksandrs Slivkins. Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1):2533–2568, 2014.

Linqi Song, Cem Tekin, and Mihaela Van Der Schaar. Online learning in large-scale contextual recommender systems. *IEEE Transactions on Services Computing*, 9(3):433–445, 2014.

Balázs Szörényi, Róbert Busa-Fekete, István Hegedűs, Róbert Ormándi, Márk Jelasity, and Balázs Kégl. Gossip-based distributed stochastic bandit algorithms. In *Journal of Machine Learning Research Workshop and Conference Proceedings*, volume 2, pages 1056–1064. International Machine Learning Society, 2013.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Huazheng Wang, Qingyun Wu, and Hongning Wang. Factorization bandits for interactive recommendation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017a.

Qing Wang, Chunqiu Zeng, Wubai Zhou, Tao Li, S Sitharama Iyengar, Larisa Shwartz, and Genady Ya Grabarnik. Online interactive collaborative filtering using multi-armed bandit with dependent arms. *IEEE Transactions on Knowledge and Data Engineering*, 31(8):1569–1580, 2018.

Xin Wang, Steven CH Hoi, Chenghao Liu, and Martin Ester. Interactive social recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 357–366, 2017b.

Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: How much communication is needed to achieve (near) optimal regret. *arXiv preprint arXiv:1904.06309*, 2019.

Bryce Woodworth, Francesco Ferrari, Teofilo E Zosa, and Laurel D Riek. Preference learning in assistive robotics: Observational repeated inverse reinforcement learning. In *Machine Learning for Healthcare Conference*, pages 420–439, 2018.

Qingyun Wu, Huazheng Wang, Quanquan Gu, and Hongning Wang. Contextual bandits in a collaborative environment. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 529–538, 2016.

Yifan Wu, András György, and Csaba Szepesvári. Online learning with gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*, pages 1360–1368, 2015.

X. Xu, S. Vakili, Q. Zhao, and A. Swami. Online learning with side information. In *MILCOM 2017 - 2017 IEEE Military Communications Conference (MILCOM)*, pages 303–308, 2017.

Chicheng Zhang, Alekh Agarwal, Hal Daumé Iii, John Langford, and Sahand Negahban. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. In *International Conference on Machine Learning*, pages 7335–7344, 2019.

Zhenyu Zhu, Liusheng Huang, and Hongli Xu. Collaborative thompson sampling. *Mobile Networks and Applications*, 2020. doi: 10.1007/s11036-019-01453-x.

Appendix A. Related Work and Comparisons

In this section, we review the literature on online learning problems that involve multiple interacting players (see also (Landgren, 2019) Section 1.3.2 thereof for a survey). We then comment on how existing problem formulations compare with the problems studied in this paper.

Identical Reward Distributions. A large portion of prior studies focuses on the setting where a network of players collaboratively work on one bandit learning problem instance, i.e., the arms and their corresponding reward distributions are identical for all the players. For example, (Landgren et al., 2016) applies running consensus algorithms to study distributed cooperative MABs in which agents communicate with their neighbors in a predefined graph. (Kolla et al., 2018) studies collaborative stochastic bandits over different structures of social networks, through which players are connected. Peer-to-peer networks are explored in (Szörényi et al., 2013), in which limited communications are allowed with a few random other players. (Kar et al., 2011) studies a networked bandit problem, in which only one major agent observes rewards, whereas the other agents only have access to the sampling pattern of the major agent. Multi-agent bandit learning with limited communication is investigated in (Sankararaman et al., 2019). (Wang et al., 2019) studies the communication complexity in multi-agent multi-armed bandits.

Player-Dependent Reward Distributions. Previous studies have also covered cases where players have different reward distributions. In (Shahrampour et al., 2017), a group of players seek to find the arm with the largest average reward over all the players. This setting differs from our problem formulation in two ways. On one hand, in each round, the players have to reach a consensus and choose the same arm; on the other hand, the goal is to identify the best arm averaging out all the players instead of finding the optimal arm for each player. (Cesa-Bianchi et al., 2013) studies a network of linear contextual bandit players with player-dependent rewards—the players propagate information based on their affinity which is specified by a graph. In (Gentile et al., 2014), players are clustered on the fly and share feedback information with other players in the same cluster. Unfortunately, neither of these papers focuses on the robust aggregation of data shared by other players. In this paper, we study how data can be safely and adaptively aggregated based on a pre-defined discrepancy parameter. Similarities in reward distributions are explored in (Zhang et al., 2019), which studies a warm-start scenario, in which data are provided as history (Shivaswamy and Joachims, 2012) for an learning agent to explore faster. In this paper, however, we study the multi-player setting, where all players learn continually and concurrently.

Social Influence in Reward Generation. Social influence has been incorporated in reward generation for bandit learning problems. In (Wu et al., 2016), a player's reward is a compound of one's own preferences as well as the preferences of other players; the affinity relationship is modeled using a graph network. Such a mixture of rewards from neighboring players is also explored in (Wang et al., 2017a). (Wang et al., 2017b) adaptively learns the weights of players' degree of trust for each other, which determines the each player's reward. In this paper, our problem formulation differs from those in this line of research,

as we focus on robust aggregation of data from similar reward functions rather than how rewards are generated based on a social network.

Clustering of Players or Arms. Several studies addresses multi-player collaboration by clustering players and/or arms such that players and/or arms in one cluster share similar features. For instance, (Li et al., 2019) introduces schemes that employ adaptive clustering of players, and each player in a cluster shares the same parameters for rewards. Similarly, (Zhu et al., 2020) dynamically clusters players to group and then utilizes a Thompson Sampling-based approach for dynamic environments. Neither of these papers explicitly discuss the closeness in reward distributions for different players within a cluster. In this paper, however, we study how to aggregate data given the closeness between similar yet nonidentical distributions. In (Song et al., 2014), arm-cluster trees are constructed, and online decisions are made in a hierarchical fashion such that a cluster is chosen first and then an arm is chosen within the cluster. (Wang et al., 2018) also models dependencies of different arms in a bandit learning problem as clusters.

Side Information. Models in which learning agents observe side information have also been studied in prior works—one can consider data collected by other players in multiplayer bandit learning as side observations (Landgren, 2019). (a) In some models, a player observes side information for a subset of arms that are not chosen in the current round. These models differ from ours, as the additional data are for other arms in the same bandit learning problem instance, whereas we focus on learning by aggregating biased data from others players. Stochastic bandits with side information are studied in (Caron et al., 2012; Buccapatnam et al., 2014; Wu et al., 2015; Deshmukh et al., 2017), and in (Mannor and Shamir, 2011; Alon et al., 2017) for adversarial bandits. (b) In (Xu et al., 2017), closeness in reward distributions are presented as side information; however, similar to the above models, such closeness is between different arms in one bandit learning problem rather than between different players solving different problems. (c) Upper and lower bounds on the means of reward distributions are used as side information in (Sharma et al., 2020). Further, side information can also refer to "context" in contextual bandits (Slivkins, 2014); however, herein we focus on data aggregation in a multi-armed setting.

Collisions in Multi-Player Bandit Learning. Another set of models (see (Bubeck and Budzinski, 2020; Liu and Zhao, 2010) for two examples) popularly studied in the community focuses on collision avoidance. In such models, if two players choose the same arm in one round, i.e., they collide, they receive no rewards. Such models have a wide range of applications in, for example, multi-channel radio networks. However, in our paper, we focus primarily on minimizing rewards through information sharing among each player.

Other Multi-Player Online Learning Topics. Many other topics that are related to multi-player bandit learning have also been explored. (Christakopoulou and Banerjee, 2018) learns latent features collaboratively across players and arms to address top-K recommendations; Nonstochastic bandit learning of communicating agents is studied in (Bar-On and Mansour, 2019; Cesa-Bianchi et al., 2019); Privacy protection in decentralized exploration is investigated in (Feraud et al., 2019); (Awerbuch and Kleinberg, 2005) studies competitive collaborative learning, in which a set of players are uncandid. The goals of these papers do not align closely with ours in this paper.

Appendix B. Proof of Fact 1

Proof Suppose there exist two players $p, q \in [M]$ such that their optimal arms do not have the same index. Let $i \in [K]$ be the index of the optimal arm for p, and $j \in [K]$ for q. It follows that $\mu_i^p \geq \mu_j^p$. Since $\mu_i^q \geq \mu_i^p - \epsilon$ by Assumption 1, we have $\mu_j^q > \mu_i^q + 2\epsilon \geq \mu_i^p + \epsilon$, where the first inequality follows from Assumption 2. Then, by Assumption 1, we have $\mu_j^p \geq \mu_j^q - \epsilon$. It follows that $\mu_j^p \geq \mu_j^q - \epsilon > \mu_i^p + \epsilon - \epsilon = \mu_i^p$, which leads to a contradiction. ■

Appendix C. Proof of Theorem 1

C.1 Proof Overview.

In Appendix C.2 and Appendix C.3, we focus on showing that in a "clean" event $\mathcal{E}$ (defined in C.3), the upper confidence bound $\text{UCB}_i^p(t) = \kappa_i^p(t, \lambda) + F(\overline{n_i^p}, \overline{m_i^p}, \lambda, \epsilon)$ (line 11 of Algorithm 1)[4] holds for every $t \in [T], i \in [K], p \in [M]$ and $\lambda \in [0, 1]$; and the "clean" event $\mathcal{E}$ occurs with $1 - 2/T$ probability.

Then, in Appendix C.4, we provide a proof for Theorem 1.

C.2 Event $\mathcal{Q}_i(t)$

Let $\overline{z} = \max\{z, 1\}$. Recall that $n_i^p(t-1)$ is the number of pulls of arm i by player p after the first $(t-1)$ rounds. Let $m_i^p(t-1) = \sum_{q \in [M]: q \neq p} n_i^q(t-1)$.

We now define the following event.

Definition 4 *Let*

$$\mathcal{Q}_i(t) = \left\{ \forall p, \left|\zeta_i^p(t) - \mu_i^p\right| \leq 6\sqrt{\frac{5\ln T}{\overline{n_i^p}(t-1)}}, \quad \left|\eta_i^p(t) - \sum_{q \neq p} \frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)} \mu_i^q\right| \leq 8\sqrt{\frac{3\ln T}{\overline{m_i^p}(t-1)}} \right\},$$

where

$$\zeta_i^p(t) = \frac{\sum_{s=1}^{t-1} \mathbb{1}(i_t^p = i) r_{i,t}^p}{\overline{n_i^p}(t-1)},$$

and

$$\eta_i^p(t) = \frac{\sum_{s=1}^{t-1} \sum_{q=1}^{M} \mathbb{1}(q \neq p,\ i_t^q = i) r_{i,t}^q}{\overline{m_i^p}(t-1)}.$$

Lemma 5

$$\Pr(\mathcal{Q}_i(t)) \geq 1 - 2T^{-3}.$$

Proof For any fixed player p, we discuss the two inequalities separately. Lemma 5 then follows by a union bound over the two inequalities and over all $p \in [M]$.

4. Recall that $\overline{z} = \max\{z, 1\}$.

We first discuss the concentration of $\zeta_i^p(t)$. We define a filtration $\{\mathcal{B}_t\}_{t=1}^T$, where

$$\mathcal{B}_t = \sigma(\Big\{i_s^{p'}, r_{i_s,s}^{p'} : s \in [t], p' \in [M], i \in [K]\Big\} \cup \Big\{i_{t+1}^{p'} : p' \in [M]\Big\})$$

is the σ-algebra, generated by the historical observations up to time step t and the arm selection of all players at time step $t+1$.

Let random variable $X_t = \mathbb{1}(i_t^p = i)\left(r_{i_t,t}^p - \mu_i^p\right)$. We have $\mathbb{E}\left[X_t \mid \mathcal{B}_{t-1}\right] = 0$; in addition, $\mathbb{E}\left[X_t^2 \mid \mathcal{B}_{t-1}\right] \leq \mathbb{1}(i_t^p = i)$ and $|X_t| \leq 1$.

Applying Freedman's inequality (Bartlett et al., 2008, Lemma 2), we have that with probability at least $1 - T^{-4}$,

$$\left|\sum_{s=1}^{t-1} X_s\right| \leq 4\sqrt{\sum_{s=1}^{t-1} \mathbb{1}(i_t^p = i) \cdot \ln(T^4 \log_2 T) + 2\ln(T^4 \log_2 T)}. \tag{1}$$

We consider two cases:

1. If $n_i^p(t-1) = \sum_{s=1}^{t-1} \mathbb{1}(i_t^p = i) = 0$, we have $\overline{n_i^p}(t-1) = 1$ and $\zeta_i^p(t) = 0$. In this case, we trivially have
$$\left|\zeta_i^p(t) - \mu_i^p\right| \leq 1 \leq 6\sqrt{\frac{5\ln T}{\overline{n_i^p}(t-1)}}.$$

2. Otherwise, $n_i^p(t-1) \geq 1$. In this case, we have $\overline{n_i^p}(t-1) = n_i^p(t-1)$. Divide both sides of Equation (1) by $n_i^p(t-1)$, and use the fact that $\log T \leq T$, we have
$$\left|\frac{\sum_{s=1}^{t-1} \mathbb{1}(i_t^p = i) r_{i,t}^p}{n_i^p(t-1)} - \mu_i^p\right| \leq 4\sqrt{\frac{5\ln T}{n_i^p(t-1)}} + \frac{10\ln T}{n_i^p(t-1)}.$$
If $\frac{10\ln T}{n_i^p(t-1)} \geq 1$, $\left|\frac{\sum_{s=1}^{t-1} \mathbb{1}(i_t^p=i) r_{i,t}^p}{n_i^p(t-1)} - \mu_i^p\right| \leq 6\sqrt{\frac{5\ln T}{n_i^p(t-1)}}$ is trivially true. Otherwise, $\frac{10\ln T}{n_i^p(t-1)} \leq \sqrt{\frac{10\ln T}{n_i^p(t-1)}}$, which implies that $\left|\frac{\sum_{s=1}^{t-1} \mathbb{1}(i_t^p=i) r_{i,t}^p}{n_i^p(t-1)} - \mu_i^p\right| \leq (4\sqrt{5}+\sqrt{10})\sqrt{\frac{\ln T}{n_i^p(t-1)}} \leq 6\sqrt{\frac{5\ln T}{n_i^p(t-1)}}$.

In summary, in both cases, with probability at least $1 - T^{-4}$, we have

$$\left|\zeta_i^p(t-1) - \mu_i^p\right| \leq 6\sqrt{\frac{5\ln T}{\overline{n_i^p}(t-1)}}.$$

A similar application of Freedman's inequality also shows the concentration of $\eta_i^p(t)$. Similarly, we define a filtration $\{\mathcal{G}_{t,q}\}_{t\in[T],q\in[M]}$, where

$$\mathcal{G}_{t,q} = \sigma(\Big\{i_s^{p'}, r_{i_s,s}^{p'} : s \in [t], p' \in [M], i \in [K]\Big\} \cup \Big\{i_{t+1}^{p'} : p' \in [M], p' \leq q\Big\})$$

is the σ-algebra, generated by the historical observations up to time step t and the arm selection of players $1, 2, \ldots, q$ at time step $t+1$. We have

$$\mathcal{G}_{1,1} \subset \mathcal{G}_{1,2} \subset \ldots \subset \mathcal{G}_{1,M} \subset \mathcal{G}_{2,1} \subset \ldots \subset \mathcal{G}_{2,M} \subset \ldots \subset \mathcal{G}_{T,M}.$$

Let random variable $Y_{t,q} = \mathbb{1}(q \neq p,\ i_t^q = i)\left(r^q_{i_t,t} - \mu_i^q\right)$. We have $\mathbb{E}\left[Y_{t,q} \mid \mathcal{G}_{t-1,q}\right] = 0$; in addition, $\mathbb{E}\left[Y_{t,q}^2 \mid \mathcal{G}_{t-1,q}\right] \leq \mathbb{1}(q \neq p,\ i_t^q = i)$ and $\left|Y_{t,q}\right| \leq 1$.

Similarly, applying Freedman's inequality (Bartlett et al., 2008, Lemma 2), we have that with probability at least $1 - T^{-4}$,

$$\left|\sum_{s=1}^{t-1}\sum_{q=1}^{M} Y_{s,q}\right| \leq 4\sqrt{\sum_{s=1}^{t-1}\sum_{q=1}^{M} \mathbb{1}(q \neq p,\ i_t^q = i) \cdot \ln(T^4 \log_2(TM)) + 2\ln(T^4 \log_2(TM))}. \quad (2)$$

Again, we consider two cases. If $m_i^p(t-1) = 0$, then we have $\eta_i^p(t-1) = 0$ and

$$\left|\eta_i^p(t-1) - \frac{\sum_{q\neq p} n_i^q(t-1)}{m_i^p(t-1)}\mu_i^q\right| \leq 1 \leq 8\sqrt{\frac{3\ln T}{\overline{m_i^p}(t-1)}}$$

.

Otherwise, we have $\overline{m_i^p}(t-1) = m_i^p(t-1)$. Divide both sides of Equation (2) by $m_i^p(t-1)$, and use the fact that $\log_2(TM) \leq T^2$, we have

$$\left|\frac{\sum_{s=1}^{t-1}\sum_{q=1}^{M}\mathbb{1}(q \neq p,\ i_t^q = i) r_{i,t}^q}{m_i^p(t-1)} - \frac{\sum_{q\neq p} n_i^q(t-1)}{m_i^p(t-1)}\mu_i^q\right| \leq 4\sqrt{\frac{6\ln T}{m_i^p(t-1)}} + \frac{12\ln T}{m_i^p(t-1)}.$$

If $\frac{12\ln T}{m_i^p(t-1)} \geq 1$, $\left|\frac{\sum_{s=1}^{t-1}\mathbb{1}(i_t^p = i) r_{i,t}^p}{m_i^p(t-1)} - \mu_i^p\right| \leq 6\sqrt{\frac{5\ln T}{m_i^p(t-1)}}$ is trivially true. Otherwise, $\frac{12\ln T}{m_i^p(t-1)} \leq 2\sqrt{\frac{3\ln T}{m_i^p(t-1)}}$, which implies that

$$\begin{aligned}\left|\frac{\sum_{s=1}^{t-1}\sum_{q=1}^{M}\mathbb{1}(q \neq p,\ i_t^q = i) r_{i,t}^q}{m_i^p(t-1)} - \frac{\sum_{q\neq p} n_i^q(t-1)}{m_i^p(t-1)}\mu_i^q\right| &\leq (4\sqrt{6} + 2\sqrt{3})\sqrt{\frac{\ln T}{m_i^p(t-1)}} \\ &\leq 8\sqrt{\frac{3\ln T}{m_i^p(t-1)}}.\end{aligned}$$

In summary, in both cases, with probability at least $1 - T^{-4}$, we have

$$\left|\eta_i^p(t-1) - \frac{\sum_{q\neq p} n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right| \leq 8\sqrt{\frac{3\ln T}{\overline{m_i^p}(t-1)}}.$$

The lemma follows by taking a union bound over these two inequalities for each fixed p, and over all $p \in [M]$, given $M \leq T$. ■

C.3 Event $\mathcal{E}$

Let $\mathcal{E} = \cap_{t=1}^{T} \cap_{i=1}^{K} \mathcal{Q}_i(t)$. We present the following corollary and lemma regarding event $\mathcal{E}$.

Corollary 6 *It follows from Lemma 5 that* $\Pr[\mathcal{E}] \geq 1 - 2/T$.

Lemma 7 *If $\mathcal{E}$ occurs, we have that for every $t \in [T]$, $i \in [K]$, $p \in [M]$, for all $\lambda \in [0,1]$,*

$$\left|\kappa_i^p(t,\lambda) - \mu_i^p\right| \leq 8\sqrt{6\ln T\left(\frac{\lambda^2}{\overline{n_i^p}(t-1)} + \frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}\right)} + (1-\lambda)\epsilon,$$

where $\kappa_i^p(t,\lambda) = \lambda\zeta_i^p(t) + (1-\lambda)\eta_i^p(t)$.

Proof If $\mathcal{E}$ occurs, for every $t \in [T]$ and $i \in [K]$, by the definition of event $\mathcal{Q}_i(t)$, we have

$$\left|\zeta_i^p(t) - \mu_i^p\right| < 6\sqrt{\frac{5\ln T}{\overline{n_i^p}(t-1)}}, \text{and } \left|\eta_i^p(t) - \sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right| \leq 8\sqrt{\frac{3\ln T}{\overline{m_i^p}(t-1)}}.$$

As $\kappa_i^p(t,\lambda) = \lambda\zeta_i^p(t) + (1-\lambda)\eta_i^p(t)$, we have:

$$\begin{aligned}\left|\kappa_i^p(t,\lambda) - \left[\lambda\mu_i^p + (1-\lambda)\sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q\right]\right| \leq& 6\lambda\sqrt{\frac{5\ln T}{\overline{n_i^p}(t-1)}} + 8(1-\lambda)\sqrt{\frac{3\ln T}{\overline{m_i^p}(t-1)}} \\ \leq& 8\sqrt{6\ln T\left(\frac{\lambda^2}{\overline{n_i^p}(t-1)} + \frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}\right)}, \quad (3)\end{aligned}$$

where the second inequality uses the elementary facts that $\sqrt{A} + \sqrt{B} \leq \sqrt{2(A+B)}$ and $6\sqrt{5} < 8\sqrt{3}$.

Furthermore, from Assumption 1, we have

$$\left|\sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q - \mu_i^p\right| \leq \sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\left|\mu_i^q - \mu_i^p\right| \leq \epsilon.$$

This shows that

$$\left|\mu_i^p - (\lambda\mu_i^p + (1-\lambda)\sum_{q\neq p}\frac{n_i^q(t-1)}{\overline{m_i^p}(t-1)}\mu_i^q)\right| \leq (1-\lambda)\epsilon.$$

Combining the above inequality with Equation (3), we get

$$\left|\kappa_i^p(t,\lambda) - \mu_i^p\right| \leq 8\sqrt{6\ln T\left(\frac{\lambda^2}{\overline{n_i^p}(t-1)} + \frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}\right)} + (1-\lambda)\epsilon.$$

This completes the proof. ■

C.4 Proof of Theorem 1

We now provide a proof for Theorem 1. We have

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)] &\leq \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] + \mathbb{E}[\mathcal{R}(T)|\overline{\mathcal{E}}]\Pr[\overline{\mathcal{E}}] \\ &\leq \mathbb{E}[\mathcal{R}(T)|\mathcal{E}] + 2M, \end{aligned} \tag{4}$$

where the second inequality uses the fact that $\mathbb{E}[\mathcal{R}(T)|\overline{\mathcal{E}}] \leq TM$, as the instantaneous regret for each player in each round is bounded by 1.

We now focus on the expected collective regret when the event $\mathcal{E}$ occurs.

Denote by $n_i(t) = \sum_{p=1}^{M} n_i^p(t)$ for every t and i. Following the strategy in (Auer et al., 2002) and (Landgren et al., 2016), we seek to bound the expected total number of pulls of each nonoptimal arm i by all the players in T rounds, which we denote as $n_i(T)$, conditional on the event $\mathcal{E}$. Note that it follows from Fact 1 that if an arm i is nonoptimal for a player p, then it is also nonoptimal for any other player q.

We have

$$\begin{aligned} n_i(T) &= \sum_{t=1}^{T}\sum_{p=1}^{M} \mathbb{1}\{i_t^p = i\} \\ &\leq M + \tau + \sum_{t=1}^{T}\sum_{p=1}^{M} \mathbb{1}\{i_t^p = i, n_i(t-1) > \tau\}. \end{aligned} \tag{5}$$

Here, $\tau \geq 1$ is an arbitrary integer. The term M is due to communication delay in the ϵ-MPMAB problem: Let s be the first round such that after round s, the total number of pulls $n_i(s) > \tau$. This implies that $n_i(s-1) \leq \tau$. Then in round s, there can be up to M pulls of arm i by all the players, which means that in round $(s+1)$ when the third term in Eq. 5 can first start counting, there could have been up to $\tau + M$ pulls of the arm i.

It then follows that

$$n_i(T) \leq M + \tau + \sum_{t=1}^{T}\sum_{p=1}^{M} \mathbb{1}\{\mathrm{UCB}_{i_*}^p(t) \leq \mathrm{UCB}_i^p(t),\ n_i(t-1) > \tau\}. \tag{6}$$

Let $\Delta_i^{\min} = \min_p \Delta_i^p$. With foresight, we choose $\tau = \lceil \frac{1536 \ln T}{(\Delta_i^{\min} - 2\epsilon)^2} \rceil$. Conditional on $\mathcal{E}$, we show that the event $\{\mathrm{UCB}_{i_*}^p(t) \leq \mathrm{UCB}_i^p(t), n_i(t-1) > \tau\}$ never happens. It suffices to show that if $n_i(t-1) > \tau$,

$$\mathrm{UCB}_{i_*}^p(t) \geq \mu_*^p, \tag{7}$$

and

$$\mathrm{UCB}_i^p(t) < \mu_*^p \tag{8}$$

happen simultaneously.

Equation (7) follows straightforwardly from the definition of $\mathcal{E}$ along with Lemma 7. For Equation (8), we have the following upper bound on $\mathrm{UCB}_i^p(t)$:

$$\begin{aligned}
\mathrm{UCB}_i^p(t) &= \kappa_i^p(t,\lambda^*) + F(\overline{n_i^p}, \overline{m_i^p}, \lambda^*, \epsilon) \\
&\le \mu_i^p + 2F(\overline{n_i^p}, \overline{m_i^p}, \lambda^*, \epsilon) \\
&= \mu_i^p + 2\Big[\min_{\lambda\in[0,1]} 8\sqrt{6\ln T[\frac{\lambda^2}{\overline{n_i^p}(t-1)} + \frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}]} + (1-\lambda)\epsilon\Big] \\
&\le \mu_i^p + 2\Big[8\sqrt{\frac{6\ln T}{\overline{n_i^p}(t-1) + \overline{m_i^p}(t-1)}} + \epsilon\Big] \\
&\le \mu_i^p + 2\Big[8\sqrt{\frac{6\ln T}{n_i(t-1)}} + \epsilon\Big] \\
&< \mu_i^p + 2\Big[8\sqrt{\frac{6\ln T(\Delta_i^p - 2\epsilon)^2}{1536\ln T}} + \epsilon\Big] = \mu_i^p + \Delta_i^p = \mu_*^p,
\end{aligned}$$

where the first inequality is from the definition of $\mathcal{E}$ and Lemma 7; the second inequality is from choosing $\lambda = \frac{\overline{n_i^p}(t-1)}{\overline{n_i^p}(t-1)+\overline{m_i^p}(t-1)}$; the third inequality is from the simple facts that $n_i^p(t-1) \le \overline{n_i^p}(t-1)$, $m_i^p(t-1) \le \overline{m_i^p}(t-1)$, and $n_i(t-1) = n_i^p(t-1) + m_i^p(t-1)$; the last inequality is from the premise that $n_i^p(t-1) > \tau \ge \frac{1536\ln T}{(\Delta_i^{\min}-2\epsilon)^2} \ge \frac{1536\ln T}{(\Delta_i^p-2\epsilon)^2}$.

Continuing Equation (6), it then follows that

$$\mathbb{E}[n_i(T)|\mathcal{E}] \le \lceil \frac{1536\ln T}{(\Delta_i^{\min}-2\epsilon)^2} \rceil + M \le \frac{1536\ln T}{(\Delta_i^{\min}-2\epsilon)^2} + (M+1).$$

Since the instantaneous regret for arm i and any player p is upper bounded by $\Delta_i^{\max} = \max_p \Delta_i^p \le 1$, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(T)|\mathcal{E}] &\le \sum_{\substack{i\in[K]\\ i\neq i_*}} \mathbb{E}[n_i(T)|\mathcal{E}] \cdot \Delta_i^{\max} \\
&\le \sum_{\substack{i\in[K]\\ i\neq i_*}} \frac{1536\ln T}{(\Delta_i^{\min}-2\epsilon)^2} \cdot \Delta_i^{\max} + K(M+1).
\end{aligned}$$

It then follows from Eq. 4 that

$$\mathbb{E}[\mathcal{R}(T)] \le O\left(\sum_{\substack{i\in[K]\\ i\neq i_*}} \frac{\ln T}{(\Delta_i^{\min}-2\epsilon)^2} \cdot \Delta_i^{\max} + KM\right).$$

This completes the proof of Theorem 1.

Appendix D. Proof of Lemma 3

Proof We use arguments similar to the ones presented in Appendix C for Theorem 1. Again, we consider the expected collective regret conditional on event $\mathcal{E}$.

For any player $p \in [M]$, we seek to bound the number of pulls of any nonoptimal arm i by p in T rounds, where $\mu_i^p < \mu_*^p$. Since the optimal arm may be different for different players, we treat each player separately.

Recall that $n_i^p(t-1)$ is the number of pulls of arm i by player p after $(t-1)$ rounds. We have

$$\begin{aligned} n_i^p(T) &= \sum_{t=1}^{T} \mathbb{1}\{i_t^p = i\} \\ &\leq \tau + \sum_{t=\tau+1}^{T} \mathbb{1}\{i_t^p = i, n_i^p(t-1) > \tau\}, \end{aligned} \tag{9}$$

where $\tau \geq 1$ is an arbitrary integer. It then follows that

$$n_i^p(T) \leq \tau + \sum_{t=\tau+1}^{T} \mathbb{1}\{\mathrm{UCB}_{i_*^p}^p(t) \leq \mathrm{UCB}_i^p(t), n_i^p(t-1) > \tau\}.$$

With foresight, let $\tau = \lceil \frac{1536 \ln T}{(\Delta_i^p)^2} \rceil$. Conditional on $\mathcal{E}$, we show that the event $\{\mathrm{UCB}_{i_*^p}^p(t) \leq \mathrm{UCB}_i^p(t), n_i^p(t-1) > \tau\}$ never happens. It suffices to show that if $n_i^p(t-1) > \tau$,

$$\mathrm{UCB}_{i_*^p}^p(t) \geq \mu_*^p, \tag{10}$$

and

$$\mathrm{UCB}_i^p(t) < \mu_*^p \tag{11}$$

happen simultaneously.

Equation (10) follows straightforwardly from the definition of $\mathcal{E}$ along with Lemma 7. For Equation (11), we have the following upper bound on $\mathrm{UCB}_i^p(t)$:

$$\begin{aligned} \mathrm{UCB}_i^p(t) &= \kappa_i^p(t, \lambda^*) + F(\overline{n_i^p}, \overline{m_i^p}, \lambda^*, \epsilon) \\ &\leq \mu_i^p + 2F(\overline{n_i^p}, \overline{m_i^p}, \lambda^*, \epsilon) \\ &= \mu_i^p + 2\Big[\min_{\lambda \in [0,1]} 8\sqrt{6 \ln T [\frac{\lambda^2}{\overline{n_i^p}(t-1)} + \frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}]} + (1-\lambda)\epsilon\Big] \\ &\leq \mu_i^p + 2\Big[8\sqrt{\frac{6 \ln T}{\overline{n_i^p}(t-1)}}\Big] \\ &\leq \mu_i^p + 2\Big[8\sqrt{\frac{6 \ln T}{n_i^p(t-1)}}\Big] \\ &< \mu_i^p + 2\Big[8\sqrt{\frac{6 \ln T (\Delta_i^p)^2}{1536 \ln T}}\Big] = \mu_i^p + \Delta_i^p = \mu_*^p, \end{aligned}$$

where the first inequality is from the definition of event $\mathcal{E}$ and Lemma 7; the second inequality is from choosing $\lambda = 1$; the third inequality uses the basic fact that $n_i^p(t-1) \leq \overline{n_i^p}(t-1)$; the fourth inequality is by our premise that $n_i^p(t-1) > \tau \geq \frac{1536 \ln T}{(\Delta_i^p)^2}$.

It follows that conditional on $\mathcal{E}$, the second term in Eq. 9 is always zero, i.e., player p would not pull arm i again. Therefore, we have

$$\mathbb{E}[n_i^p(T)|\mathcal{E}] \leq \lceil \frac{1536 \ln T}{(\Delta_i^p)^2} \rceil \leq \frac{1536 \ln T}{(\Delta_i^p)^2} + 1.$$

It then follows that

$$\begin{aligned}\mathbb{E}[\mathcal{R}^p(T)|\ \mathcal{E}] &\leq \sum_{\substack{i \in [K] \\ i \neq i_*^p}} \mathbb{E}[n_i(T)|\mathcal{E}] \cdot \Delta_i^p \\ &\leq \sum_{\substack{i \in [K] \\ i \neq i_*^p}} \frac{1536 \ln T}{\Delta_i^p} + 1.\end{aligned}$$

Summing over all the players, we have

$$\mathbb{E}[\mathcal{R}(T)|\ \mathcal{E}] \leq \sum_{p \in [M]} \sum_{\substack{i \in [K] \\ i \neq i_*^p}} \frac{1536 \ln T}{\Delta_i^p} + 1.$$

It then follows from Eq. 4 that

$$\begin{aligned}\mathbb{E}[\mathcal{R}(T)] &\leq \sum_{p \in [M]} \sum_{\substack{i \in [K] \\ i \neq i_*^p}} \frac{1536 \ln T}{\Delta_i^p} + 3M. \\ &= O\left(\sum_{p \in [M]} \sum_{\substack{i \in [K] \\ i \neq i_*^p}} \frac{\ln T}{\Delta_i^p} \right)\end{aligned}$$

This completes the proof of Lemma 3. ■

Appendix E. Experimental Details

In Appendix E.1, we present a psuedocode for the algorithm used in the experiments. Then, in Appendix E.2, we describe the experimental setup in detail.

E.1 Adapted version of Algorithm 1

Algorithm 2: An adapted version of Algorithm 1

Input: A parameter $\epsilon \in [0, 1]$;

Initialization: Set $n_i^p = 0$ for all $p \in [M]$ and all $i \in [K]$.

for $t = 1, 2, \dots K$ **do**

  for $p \in [M]$ **do**

    Player p pulls arm $i_t^p = t$ and observes reward $r_{i_t}^p$;

    Set $n_i^p = n_i^p + 1$.

for $t = K + 1, K + 2 \dots, T$ **do**

  for $p \in [M]$ **do**

    for $i \in [K]$ **do**

      Let $m_i^p = \sum_{q \in [M]: q \neq p} n_i^q$;

      Let $F(n_i^p, m_i^p, \lambda, \epsilon) = \sqrt{2 \ln T [\frac{\lambda^2}{n_i^p} + \frac{(1-\lambda)^2}{m_i^p}]} + (1 - \lambda)\epsilon$;

      Compute $\lambda^* = \text{argmin}_{\lambda \in [0,1]} F(n_i^p, m_i^p, \lambda, \epsilon)$;

      Let

$$\zeta_i^p(t) = \frac{1}{n_i^p} \sum_{\substack{s<t \\ i_s^p = i}} r_{i_s,s}^p,\ \eta_i^p(t) = \frac{1}{m_i^p} \sum_{\substack{q \in [M] \\ q \neq p}} \sum_{\substack{s<t \\ i_s^q = i}} r_{i_s,s}^q, \text{and } \kappa_i^p(t, \lambda) = \lambda \zeta_i^p(t) + (1-\lambda)\eta_i^p(t);$$

      Compute the upper confidence bound of the reward of arm i for player p:

$$\text{UCB}_i^p(t) = \kappa_i^p(t, \lambda^*) + F(n_i^p, m_i^p, \lambda^*, \epsilon).$$

    Let $i_t^p = \text{argmax}_{i \in [K]} \text{UCB}_i^p(t)$;

    Player p pulls arm i_t^p and observes reward $r_{i_t,t}^p$;

  for $p \in [M]$ **do**

    Let $i = i_t^p$ and set $n_i^p = n_i^p + 1$.

Algorithm 2 provides a pseudocode for the more practical algorithm used in the experiments. This algorithm is adapted from Algorithm 1 with few modifications. We added an initialization phase, and used a more aggressive upper confidence bound.

E.2 Experimental Setup

We now describe our experimental setup. For both experiments, we set the number of arms $K = 10$ and the time horizon $T = 50000$ rounds.

E.2.1 Experiment 1

Data Generation. We generated synthetic data that satisfy both Assumption 1 and Assumption 2 using the following procedure:

Let $\delta = 0.3$ and $\epsilon = 0.1$. We first sampled the means of the reward distributions for player 1. Without loss of generality, let arm 1 be the optimal arm for player 1. For every nonoptimal arm $i \in [2, K]$, we sampled $\mu_i^1 \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0, 1 - \delta]$, where $\mathcal{U}[a, b]$ is the uniform distribution with support $[a, b]$. Let $j = \text{argmax}_{2 \le i \le K} \mu_j^1$. We set $\mu_1^1 = \mu_j^1 + \delta$. It follows that $\forall i \neq 1, \Delta_i^1 \ge \delta$.

We then sampled the means of the reward distributions for every other player $p \neq 1$. For every $i \in [K]$ and every $p \in [2, M]$, we sampled $\mu_i^p \overset{\text{i.i.d.}}{\sim} \mathcal{U}\big(\max(0, \mu_i^1 - \frac{\epsilon}{2}), \min(\mu_i^1 + \frac{\epsilon}{2}, 1)\big]$.

It can be easily shown from the above procedure, Assumption 1 is satisfied. We now show that Assumption 2 is also satisfied. It follows that $\forall p \in [2, M], \forall i \in [2, K]$,

$$\mu_i^p \le \mu_j^1 + \frac{\epsilon}{2},$$

and

$$\mu_1^p > \mu_1^1 - \frac{\epsilon}{2}.$$

Since $\mu_1^1 \ge \mu_j^1 + \delta$, we have $\forall p \in [2, M], \forall i \in [2, K]$,

$$\mu_1^p > \mu_1^1 - \frac{\epsilon}{2} \ge \mu_j^1 + \delta - \frac{\epsilon}{2} \ge \mu_i^p + \delta - \epsilon = \mu_i^p + 0.2 = \mu_i^p + 2\epsilon.$$

It then follows that, for every player p and for every nonoptimal arm $i \neq i_*^p = 1$, $\Delta_i^p > 2\epsilon$.

Setup and Result We study the dependence of collective regret on the number of players. For $M = 5, 10, 15$, and 20 players, we each generated $C = 30$ ϵ-MPMAB instances. We then ran adapted Algorithm 1 and the baseline algorithm on each of the problem instances. Figure 1a shows the averaged collective regret after $T = 50000$ rounds for each choice of M, where the average is taken over $C = 30$ instances. The results show that the collective regret of adapted Algorithm 1 is insensitive to the number of players M, whereas the collective regret of the baseline algorithm grows "linearly" as M increases.

E.2.2 Experiment 2

Data Generation. We generated synthetic data that satisfy Assumption 1 using the following procedure. We note that Assumption 2 may or may not be satisfied.

Let $\epsilon = 0.2$. We first sampled the means of the reward distributions for player 1. For every arm $i \in [K]$, we sampled $\mu_i^1 \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0, 1]$. We then sampled the means of the reward distributions for every other player. For every $i \in [K]$ and every $p \in [2, M]$, we sampled $\mu_i^p \overset{\text{i.i.d.}}{\sim} \mathcal{U}\big[\max(0, \mu_i^1 - \frac{\epsilon}{2}), \min(\mu_i^1 + \frac{\epsilon}{2}, 1)\big]$.

Setup and Result We study the robustness of Algorithm 1. With $M = 10$ players, we generated $C = 20$ ϵ-MPMAB problem instances. On each problem instance, we ran

- adapted Algorithm 1 given $\epsilon = 0.2$;
- adapted Algorithm 1 given $\epsilon = 0$ (naive data aggregation); and
- the baseline algorithm.

We note that the naive data aggregation algorithm simply assumes that the data shared by other players are from the same distributions. Figure 1b shows the the collective regret of each algorithm over a time horizon of $T = 50000$ rounds, where the average is taken over $C = 20$ instances. From this figure, we can see the importance of robustness in data aggregation. This figure also demonstrates that adapted Algorithm 1 can still perform better than the baseline algorithm even when Assumption 2 may not hold.

Appendix F. Analytical Solution to λ^*

We present an analytical solution to λ^* (line 8 of Algorithm 1). A similar analysis is presented in Section 6 of (Ben-David et al., 2010) for classification using data from different domains.

We minimize

$$f(\lambda) = 8\sqrt{6(\ln T)[\frac{\lambda^2}{\overline{n_i^p}(t-1)} + \frac{(1-\lambda)^2}{\overline{m_i^p}(t-1)}]} + (1-\lambda)\epsilon \tag{12}$$

$$= 8\sqrt{6(\ln T)(\frac{1}{\overline{n_i^p}(t-1)} + \frac{1}{\overline{m_i^p}(t-1)})\lambda^2 - \frac{12\ln t}{\overline{m_i^p}(t-1)}\lambda + \frac{6\ln t}{\overline{m_i^p}(t-1)}} - \epsilon\lambda + \epsilon \tag{13}$$

Let $A = 384(\ln T)(\frac{1}{\overline{n_i^p}(t-1)} + \frac{1}{\overline{m_i^p}(t-1)}) > 0$, $B = -\frac{768\ln T}{\overline{m_i^p}(t-1)}$, $C = \frac{384\ln T}{\overline{m_i^p}(t-1)}$, $D = -\epsilon \leq 0$, and $E = \epsilon$.

Then, we have

$$f(\lambda) = [A\lambda^2 + B\lambda + C]^{\frac{1}{2}} + D\lambda + E.$$

By substituting λ with $\xi = \sqrt{A}\cdot\lambda + \frac{B}{2\sqrt{A}}$, we can write $f(\lambda)$ as

$$f(\xi) = [\xi^2 + H]^{\frac{1}{2}} + J\xi + L,$$

where $H = C - \frac{B^2}{4A}$, $J = \frac{D}{\sqrt{A}} \leq 0$, and $L = E - \frac{BD}{2A}$. It follows that

$$f'(\xi) = \frac{\xi}{\sqrt{\xi^2 + H}} + J.$$

We solve for $f'(\xi) = 0$, which implies that $\frac{\xi}{\sqrt{\xi^2+H}} = -J$. Since $\sqrt{\xi^2 + H} \geq 0$, ξ and J have different signs. It then follows that

$$\xi = \sqrt{\frac{J^2H}{1-J^2}}.$$

Substituting back $\lambda = \frac{\xi}{\sqrt{A}} - \frac{B}{2A}$, we obtain

$$\lambda = \sqrt{\frac{4ACD^2 - B^2D^2}{4A^3 - 4A^2D^2}} - \frac{B}{2A}. \tag{14}$$

Then, we have

$$\lambda^* = \min\left(\text{clip}\left\{\sqrt{\frac{4ACD^2 - B^2D^2}{4A^3 - 4A^2D^2}} - \frac{B}{2A}, 0, 1\right\}, 0, 1\right), \tag{15}$$

where $\text{clip}(x, \min, \max) = \begin{cases} x, & \text{if } x \in [\min, \max] \\ \min, & \text{if } x < \min \\ \max & \text{if } x > \max \end{cases}$.