Responsible Bandit Learning via Privacy-Protected Mean-Volatility Utility

Shanshan Zhao[1], Wenhai Cui[1], Bei Jiang[2], Linglong Kong[2], Xiaodong Yan[1, 3*]

[1]Mathematics Discipline, Shandong University
[2] Mathematics Discipline, University of Alberta
[3] Shandong National Center for Applied Mathematics
{sszhao, cuiwenhai}@mail.sdu.edu.cn, {bei1, lkong}@ualberta.ca, yanxiaodong@sdu.edu.cn

Abstract

For ensuring the safety of users by protecting the privacy, the traditional privacy-preserving bandit algorithm aiming to maximize the mean reward has been widely studied in scenarios such as online ride-hailing, advertising recommendations, and personalized healthcare. However, classical bandit learning is irresponsible in such practical applications as they fail to account for risks in online decision-making and ignore external system information. This paper firstly proposes privacy protected mean-volatility utility as the objective of bandit learning and proves its responsibility, because it aims at achieving the maximum probability of utility by considering the risk. Theoretically, our proposed responsible bandit learning is expected to achieve the fastest convergence rate among current bandit algorithms and generates more statistical power than classical normality-based test. Finally, simulation studies provide supporting evidence for the theoretical results and demonstrate stronger performance when using stricter privacy budgets.

Introduction

As users become increasingly concerned with the privacy and security of their online information and activities, including medical records (Price and Cohen 2019), shopping records (Martin and Murphy 2017; Petrescu and Krishen 2018), browsing history (Talukder et al. 2010), and other internet usage, systems such as recommendation engines and advertising distributors require large amounts of user data to provide personalized recommendations and better services (Shin et al. 2018). This creates a dilemma as such data collection may compromise user privacy. To address this issue, it is important to develop algorithms that can balance the needs of the system with the protection of user privacy. One widely-used algorithm that provides high-level privacy protection is differential privacy (Dwork, Roth et al. 2014), which theoretically guarantees that any attacker cannot infer changes or additions/deletions of individual records based on the output. Currently, differential privacy has been widely applied to multi-armed bandit algorithms (Huang et al. 2011; Tossou and Dimitrakakis 2016; Wang et al. 2020). However, due to the fact that differential privacy cannot prevent attacks by the centroid curator, a local differential privacy algorithm for the multi-armed bandit field was proposed to provide comprehensive privacy protection for users (Ren et al. 2020; Han et al. 2021).

By incorporating differential privacy into the study of K-armed bandit algorithms, many trustworthy bandit algorithms have been developed (Basu, Dimitrakakis, and Tossou 2019). While these algorithms demonstrate fast convergence rates $O(\frac{K}{\epsilon\sqrt{n}})$, they also come with significant limitations: (i) Using the statistics or data obtained from traditional bandit processes for statistical inference is challenging. Taking the example of the adaptive testing of the two-armed bandit algorithm under the privacy mechanism, we assume that the objective is the average of cumulative rewards obtained through the strategy θ_ϵ during the n rounds of arm-pulling process, denoted as $Q_n^{\theta_\epsilon}$. The distribution of $Q_n^{\theta_\epsilon}$ at the n-th time step is determined by the strategy and the rewards obtained in the first $n-1$ rounds within the privacy mechanism. As a result, the accurate distribution of $Q_n^{\theta_\epsilon}$ is complex and difficult to explicitly express. Moreover, employing the standard test on data derived from the bandit algorithm would result in overlooking crucial information like strategy and sample characteristics. Particularly when dealing with limited sample sizes (n), utilizing normal tests on data safeguarded by privacy measures can yield unreliable conclusions and an increased likelihood of Type I errors (Williams et al. 2021). It is challenging to draw statistically significant conclusions from the results (Villar, Bowden, and Wason 2015; Yao et al. 2021). (ii) The accountability of these privacy-preserving bandit algorithms to agents (i.e., the reliability of decision-making at each moment) is unknown. Due to the traditional two-armed bandit being designed from the perspective of maximizing the amount of information within the system (mean), it does not take into account the uncertainty brought by external information. The agent cannot perceive the risks associated with fluctuations outside the system. Strategies formulated under this objective often lead the model into local optima due to neglecting uncertainty (volatility). For instance, when the average reward of the left arm is only slightly higher than that of the right arm, but the volatility of the left arm's rewards is significantly greater than that of the right arm. In this situation, the agent tends to choose the right arm to minimize extreme losses and ensure that the obtained rewards remain

*Xiaodong Yan is the corresponding author.

within the expected range. Traditional bandit algorithms, on the other hand, tend to choose the left arm to ensure the maximum long-term cumulative rewards. Because these algorithms prioritize achieving the maximum average reward and overlook the stability of arm rewards, in certain contexts, agents might consider them 'irresponsible'.

Therefore, this paper considers the following objective:

$$T_n^{\theta_\epsilon} = \frac{1}{n}\sum_{i=1}^{n} Z_i^{\theta_\epsilon} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{Z_i^{\theta_\epsilon} - \mu_i^{\theta_\epsilon}}{\sigma_\epsilon}.$$

The first part of the above equation represents average rewards, and the second part represents average volatility. By introducing the average volatility (external information) as a component of the objective utility, we construct the responsible bandit learning algorithm under privacy mechanisms. Figure 1 (left) illustrates the variation of average cumulative reward with the change in the number of arm pulls n, when applying the ϵ-greedy algorithm and the responsible bandit algorithm respectively. It is evident that the objective of the traditional bandit model is the cumulative reward and does not reflect the real-time risk of rewards. By setting the hyperparameter c in Theorem 1, our strategy can achieve the objectives of traditional bandit problems. Figure 1 (right) illustrates the variation of our objective utility with the change in the number of arm pulls n, when applying the ϵ-greedy algorithm and the responsible bandit algorithm respectively. When one arm yields an extreme reward, the responsible bandit algorithm will select another arm as a penalty, thereby maximizing the probability of the objective utility falling within the specified interval $[c-a, c+a]$ to minimize risk. By considering comprehensive information, we perceive the bandit algorithm to be responsible, with this reliability being reflected through probabilities. Meanwhile, our objective utility under the optimal strategy is statistically significant based on the Central Limit Theorem for Strategy (Chen and Epstein 2022).

Our contributions and the advantages of the privacy responsible bandit learning model are diverse, including:

- First-time study on the behavior and characteristics of the responsible bandit learning under the privacy mechanism. Under the Laplace, Bernoulli, and Gaussian mechanisms, the convergence rates for achieving maximum probability are $O(\frac{1}{\epsilon\sqrt{n}})$, $O(\frac{1}{e^\epsilon+1})$, and $O(\frac{\sqrt{\ln(1/\delta)}}{\epsilon\sqrt{n}})$.
- This approach allows for the existence of a limiting distribution for the objective utility, making the conclusions more statistically significant.
- First-time investigation of hypothesis testing for the bandit process under the LDP mechanism. Compared to normality test, this hypothesis testing method exhibits higher statistical power.

The following sections are organized as follows: In Section 1, we introduce the new objective utility and optimal strategy within the context of two-armed bandits. In Section 2, we provide the implementation algorithms for three noise mechanisms: Laplace, Bernoulli, and Gaussian, and present the theoretical results on asymptotic distribution and convergence rates. In Section 3, hypothesis testing is introduced as a motivated example. We obtain higher p-value through the asymptotic distribution of the objective utility under the noise mechanism. In Section 4, we conduct a simulation study to verify the results of Sections 2 and 3. Discussion is provided in Section 5, and the technical proofs are given in the Appendix.

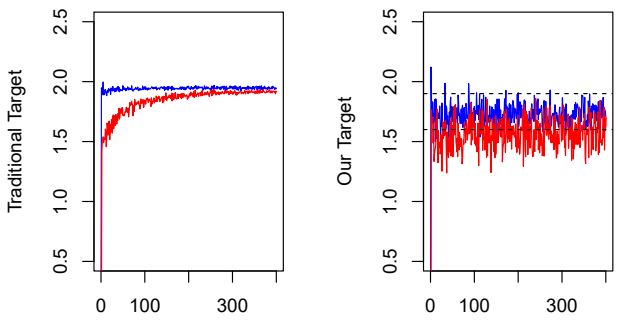

Figure 1: The variation of the objective over time n for the ϵ-greedy strategy (red line) and the responsible bandit learning strategy (blue line)

A New Framework for LDP TAB

Local Differential Privacy for Safe Bandit Model

In the two-armed bandit model, the agent has the option to choose either arm L or arm R at each stage. If they choose arm L, they will receive a reward of W^L, and if they choose arm R, they will receive a reward of W^R. To better understand the process of the agent making these choices, we have introduced a sequence strategy represented by the random variable sequence $\theta = \{\vartheta_1, \cdots, \vartheta_n\}$ for the first n stages. When $\vartheta_i = 1$, the agent has chosen arm L and received a reward of W_i^L, and when $\vartheta_i = 2$, the agent has chosen arm R and received a reward of W_i^R. The reward obtained by the agent at each moment is defined as

$$Z_i^\theta = \begin{cases} W_i^L, & \text{if } \vartheta_i = 1, \\ W_i^R, & \text{if } \vartheta_i = 2. \end{cases}$$

In this case, the choice of arm made by the agent at each stage, represented by ϑ_i, is determined based on the reward history. However, in systems like ride-hailing, advertising recommendations, and health data collection, users are likely to lose trust in agents. For instance, in an advertising recommendation system, the agent may attempt to suggest products based on the user's past purchase history in chronological order. Some users may be hesitant to share this information due to concerns about potential privacy leaks (e.g., inferring health conditions from someone purchasing a significant quantity of anti-hypertensive drugs). Differential privacy introduced by Dwork (2008) is often incorporated by centroid curator into the rewards obtained by agents.

As users lose trust in the centroid curator, having one's own curator becomes necessary. We introduce the concept of the Local Differential Privacy (LDP) bandit model in Definition 1, as suggested by Ren et al. (2020). Figure 2 shows

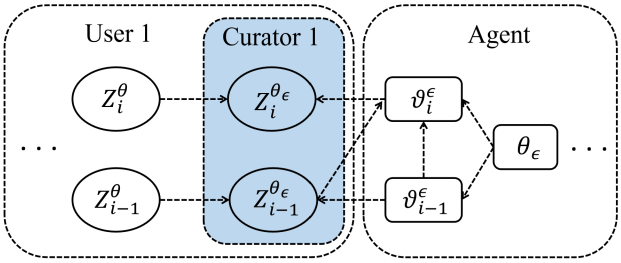

Figure 2: Local Differential Privacy

how the curators apply privacy mechanisms to the rewards obtained from each arm pull, without aggregating them, ensuring that the agent remains unaware of the actual rewards. In Figure 2, we assume that each user has their own curator, which is embedded in the user's device or terminal in the form of software or a plugin, thereby preventing non-private data from leaving the user's control. The agents can only observe the privacy-protected rewards $Z_i^{\theta_\epsilon}$ for each user at each time step. They base their strategies θ_ϵ and arm selections ϑ_i^ϵ on these rewards at each time step.

Definition 1. (ϵ-LDP Bandit Model). $\mathcal{D}$ is the finite reward domain. For $\epsilon > 0$, the bandit model is said to be ϵ-LDP if

i) a randomized mapping $M : \mathcal{D} \to \mathbb{R}$ on $\mathcal{D} \subset \mathbb{R}^k$ if for any neighboring x, x' in $\mathcal{D}$ and a measurable subset S of $\mathbb{R}$, we have

$$\mathbb{P}\{M(x) \in S\} \leq e^\epsilon \mathbb{P}\{M(x') \in S\},$$

ii) $\vartheta_{i+1}^\epsilon \in \sigma(\vartheta_j^\epsilon, M(Z_j^{\theta_\epsilon}) : 1 \leq j \leq i)$ for any time i.

If we swap x and x', the inequality stated above must still be true. This definition indicates that after an ϵ-LDP bandit mechanism is applied, the statistical characteristics of any two adjacent records will be similar. It becomes difficult for any party to determine the source of a particular output, providing protection against all attacks. In the rest of paper, we assume that the bound of reward difference is κ.

Responsible Two-Armed Bandit

In reality, the focus is often not only on maximizing cumulative rewards but also on the volatility of each reward (e.g., controlling a patient's blood pressure using medication or managing risk in investments). We introduce the average volatility to construct a new objective utility under the privacy mechanism as suggested by Cui et al. (2023), which is defined by Equation (1). The first half of Equation (1) represents the average of cumulative rewards, while the second half represents the accumulated normalized volatility. As a result, the objective utility can reflect the level of risk in online decision-making.

$$T_n^{\theta_\epsilon} = \frac{1}{n}\sum_{i=1}^{n} Z_i^{\theta_\epsilon} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{Z_i^{\theta_\epsilon} - \mu_i^{\theta_\epsilon}}{\sigma_\epsilon}, \tag{1}$$

where

$$\mu_i^{\theta_\epsilon} = \bar{\mu}_\epsilon I_{\{\vartheta_i^\epsilon = 1\}} + \underline{\mu}_\epsilon I_{\{\vartheta_i^\epsilon = 2\}}.$$

Let $\bar{\mu}_\epsilon$ represent the higher reward of the two arms, and $\underline{\mu}_\epsilon$ represent the lower reward. Without loss of generality, we assume that the arm L yields a higher reward, and the variances of both arms are equal, denoted by σ_ϵ^2. When the variances of the arm L and arm R are unequal in reality, it is natural to equalize them through standardization. It is clear that the reward $T_n^{\theta_\epsilon}$ achieved by using strategy θ_ϵ will fall within a range close to arm L if a larger proportion of the elements are equal to 1, and fall within a range close to arm R if a larger proportion of the elements are equal to 2.

Since traditional bandit strategies cannot achieve the maximization of the utility $T_n^{\theta_\epsilon}$ falling within the given interval, as shown in Figure 1 (right), it is therefore necessary to propose new strategies. To construct an optimal strategy driven by statistics, it is necessary to keep track of historical information and define

$$T_m^{\theta_\epsilon} = \frac{1}{n}\sum_{i=1}^{m} Z_i^{\theta_\epsilon} + \frac{1}{\sqrt{n}}\sum_{i=1}^{m} \frac{Z_i^{\theta_\epsilon} - \mu_i^{\theta_\epsilon}}{\sigma_\epsilon}. \tag{2}$$

The next lemma provides the optimal strategy $\theta_\epsilon^* = \{\vartheta_1^{\epsilon *}, \cdots, \vartheta_n^{\epsilon *}\}$ that maximizes the probability on a given interval under the LDP mechanism for utility $T_n^{\theta_\epsilon^*}$.

Lemma 1. (Optimal Strategy θ_ϵ^*) For any $c \in \mathbb{R}, n \geq 1$, we can construct strategy $\theta_\epsilon^* = (\vartheta_1^{\epsilon *}, \cdots, \vartheta_n^{\epsilon *})$ as follows,

$$\vartheta_m^{\epsilon *} = \begin{cases} 1, & T_{m-1}^{\theta_\epsilon^*} \leq c - (1 - \frac{m-1}{n})\frac{\bar{\mu}_\epsilon + \underline{\mu}_\epsilon}{2}, \\ 2, & T_{m-1}^{\theta_\epsilon^*} > c - (1 - \frac{m-1}{n})\frac{\bar{\mu}_\epsilon + \underline{\mu}_\epsilon}{2}, \end{cases} \quad \text{for } m \geq 1.$$

The optimal strategy is the recursive function of $T_{m-1}^{\theta_\epsilon^*}$. When the utility within a certain period is higher than the mean c of the given interval $[c-a, c+a]$, the strategy considers potential risks beyond that interval and implements corresponding strategies to reduce it. Conversely, when the utility is lower, the strategy perceives potential risks below the given interval and applies strategies to increase it, thus maximizing the probability of the objective utility falling within the specified interval $[c-a, c+a]$. The optimal strategy also endows desirable statistical properties to the utility $T_n^{\theta_\epsilon^*}$. For instance, if we take $c = (\bar{\mu}_\epsilon + \underline{\mu}_\epsilon)/2$, then

$$\begin{aligned} &\lim_{n\to\infty} P(c - a \leq T_n^{\theta_\epsilon^*} \leq c + a) \\ &= \lim_{n\to\infty} \sup_{\theta_\epsilon \in \Theta} P(c - a \leq T_n^{\theta_\epsilon} \leq c + a) \\ &= \Phi(d_1/2 + a) - e^{-ad_1}\Phi(d_1/2 - a) \\ &\geq 2\Phi(a) - 1 = P(-a \leq Z \leq a). \end{aligned} \tag{3}$$

The first equation always holds true, and the probability in the second equation is derived under the condition $\bar{\mu}_\epsilon - \underline{\mu}_\epsilon = d_1$. $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The above inequality indicates that when applying the optimal strategy, the probability of the initial statistical data falling into any given interval under the LDP mechanism can be explicitly expressed. Meanwhile, the Equality in (3) holds if and only if $d_1 = 0$, indicating that when there are differences in rewards between arm L and arm R, using the optimal strategy for hypothesis testing will outperform the normality test and specific analyses presented in Section 3.

Algorithm 1: Responsible Bandit Learning under Local Differential Privacy (RB-LDP)

1: Define: $\mu_i^{\theta_\epsilon^*} = \bar{\mu}_\epsilon I_{\{\vartheta_i^{\epsilon *}=1\}} + \underline{\mu}_\epsilon I_{\{\vartheta_i^{\epsilon *}=2\}}$;
2: Define: $T_{m-1}^{\theta_\epsilon^*} = \frac{1}{n}\sum_{i=1}^{m-1} Z_i^{\theta_\epsilon^*} + \frac{1}{\sqrt{n}}\sum_{i=1}^{m-1} \frac{Z_i^{\theta_\epsilon^*} - \mu_i^{\theta_\epsilon^*}}{\sigma_\epsilon}$;
3: Initialize $m = 1$ and utility $T_0^{\theta_\epsilon^*} = 0$;
4: **While** $m \leq n$ **do**
5: **if** $\vartheta_m^{\epsilon *} = 2 - I_{\left\{T_{m-1}^{\theta_\epsilon^*} \leq c - \left(1 - \frac{m-1}{n}\right)\frac{\bar{\mu}\epsilon + \underline{\mu}_\epsilon}{2}\right\}} = 1$ **then**
Pull arm L once and receive the private response from Curator;
6: **else** Pull arm R once and receive the private response from Curator;
7: **end if**
8: Update $\mu_m^{\theta_\epsilon^*}, T_m^{\theta_\epsilon^*}$; $m \leftarrow m + 1$;
9: **end while**

Algorithms and Theoretical Results

This section introduces the responsible bandit learning algorithm under privacy mechanisms, denoted as Algorithm 1. By introducing different privacy mechanisms into Algorithm 1, we theoretically investigate the asymptotic properties of $T_n^{\theta_\epsilon^*}$ and the convergence rate to achieve maximum reliability under the optimal strategy.

Laplace Mechanism

To provide privacy protection in differential privacy, continuous random variables are often added to the noise of the Laplace mechanism (Dwork, Roth et al. 2014). For any $b > 0$, the probability density function of the Laplace(b) distribution is defined as

$$\text{Laplace}(b) : l(x \mid b) = (2b)^{-1} \exp(-|x|/b). \tag{4}$$

The curator described under the Laplace mechanism is outlined in Algorithm 2. When the true rewards for the left and right arms of the bandit model are $\bar{\mu}$ and $\underline{\mu}$, and the true variance is σ^2, under the Laplace privacy mechanism, the corresponding parameters in Equation (1) are as follows:

$$\bar{\mu}_\epsilon = \bar{\mu},\ \underline{\mu}_\epsilon = \underline{\mu},\ \sigma_\epsilon^2 = \sigma^2 + \sigma_l^2. \tag{5}$$

Wherein, $\sigma_l^2 = 2\kappa^2/\epsilon^2$. Adding Laplace privacy protection will increase the variance of arm rewards, with stronger privacy protection resulting in greater variance. The agent will have less information about the user preferences. The corresponding responsible bandit learning algorithm, denoted as Algorithm 1, receives rewards from CTL(ϵ). The strategy relies on information gathered through previous arm pulls, using a recursive function. The addition of privacy measures will alter the rewards received by the agent for each arm pull, potentially affecting the arm selection. Nonetheless, as the number of attempts n increases, the probability of $T_n^{\theta_\epsilon^*}$ falling within the given interval will approach the optimal result at the fastest rate.

We next provide the asymptotic distribution of $T_n^{\theta_\epsilon^*}$ under the privacy mechanism and its convergence rate based on the nonlinear central limit theorem (Chen, Feng, and Zhang 2022; Chen et al. 2023). The proof and the corresponding density function of Theorem 1 are given in the Appendix.

Algorithm 2: Convert-to-Laplace (ϵ)(CTL(ϵ))

On receiving a reward Z_i^θ from the user:
return $M_L(Z_i^\theta) = Z_i^\theta + L$, where $L \sim$ Laplace (κ/ϵ) distribution;

Theorem 1. Under the Laplace mechanism, if the hypothesis $(\mu_L, \mu_R) = (\bar{\mu}, \underline{\mu})$, we have

$$T_n^{\theta_\epsilon^*} \xrightarrow{\mathcal{L}} \eta_1,\ \eta_1 \sim \mathcal{B}\left(\frac{\underline{\mu} - \bar{\mu}}{2}, \frac{\bar{\mu} + \underline{\mu}}{2}, c\right). \tag{6}$$

The convergence rate of $P(|T_n^{\theta_\epsilon^*} - c| \leq a)$ to $P(|\eta_1 - c| \leq a)$ is $O(\frac{1}{\epsilon\sqrt{n}})$ (i.e., regret in this bandit model is $O\left(\epsilon^{-1}\sqrt{n}\right)$). If $(\mu_R, \mu_L) = (\bar{\mu}, \underline{\mu})$, we will get a similar conclusion.

Bernoulli Mechanism

In addition to the Laplace mechanism, the Bernoulli mechanism is also commonly used in the LDP bandit model. This technique converts bounded rewards, typically 0 or 1, into Bernoulli responses (Gajane, Urvoy, and Kaufmann 2018). In our study, we allow for rewards to take on any value within the range of [0,1].

Meanwhile, Algorithm 1 receives the privacy response from the curator, as governed by Algorithm 3 CTB(ϵ). Its corresponding parameters in Equation (1) are as follows:

$$\mu_\epsilon = \frac{1}{2} + (2\mu - 1)\frac{e^\epsilon - 1}{2(e^\epsilon + 1)}, \tag{7}$$

$$\sigma_\epsilon^2 = \frac{1}{4} - (2\mu - 1)^2 \frac{(e^\epsilon - 1)^2}{4(e^\epsilon + 1)^2}. \tag{8}$$

Through Equations (7) and (8), the mean μ_ϵ and variance σ_ϵ^2 of privacy response are affected by both the true mean μ and the privacy budget ϵ. As the privacy budget ϵ approaches 0, the mean and variance of data under the privacy protection mechanism tend to $\frac{1}{2}$ and $\frac{1}{4}$, respectively, making it difficult to distinguish the data even in a large sample. Since the variance at this point is a finite quantity between $(0, \frac{1}{4})$, the impact of variance fluctuations can be ignored. According to the density function of the limit distribution in the Appendix, as the difference between the arm L and arm R rewards d_1 becomes smaller, the probability of the objective utility falling into the same interval becomes smaller. The impact of the Bernoulli mechanism on power will be more worthy of exploration. Corollary 1 is derived based on Theorem 1 and Curator 3.

Corollary 1. Under the Bernoulli mechanism, if the hypothesis $(\mu_L, \mu_R) = (\bar{\mu}_\epsilon, \underline{\mu}_\epsilon)$, we can conclude that

$$T_n^{\theta_\epsilon^*} \xrightarrow{\mathcal{L}} \eta_2,\ \eta_2 \sim \mathcal{B}\left(\frac{\underline{\mu}_\epsilon - \bar{\mu}_\epsilon}{2}, \frac{\bar{\mu}_\epsilon + \underline{\mu}_\epsilon}{2}, c\right). \tag{9}$$

As ϵ decreases, the probability of $T_n^{\theta_\epsilon^*}$ falling into the same interval decreases at the rate of $O(\frac{1}{e^\epsilon + 1})$. If $(\mu_R, \mu_L) = (\bar{\mu}_\epsilon, \underline{\mu}_\epsilon)$, we will obtain a similar conclusion.

Algorithm 3: Convert-to-Bernoulli $(\epsilon)(\mathrm{CTB}(\epsilon))$

On receiving a reward Z_i^θ from the user:

return $M_B(Z_i^\theta)$ = an independent sample of Bernoulli $(\frac{Z_i^\theta e^\epsilon + 1 - Z_i^\theta}{1+e^\epsilon})$;

Remark 1.1. The Bernoulli mechanism does not affect the convergence rate of the asymptotically optimal distribution but directly alters the asymptotically optimal distribution. However, our model still conforms to the maximum likelihood framework under the Bernoulli mechanism, its superiority remains unchanged.

Gaussian Mechanism

When measuring the privacy sensitivity of data queries using L_2-sensitivity, the Gaussian privacy mechanism is often employed. Prior to introducing the Gaussian mechanism, we present the relaxed Local Differential Privacy (LDP) bandit model (Dwork, Roth et al. 2014).

Definition 2. ((ϵ, δ)-LDP Bandit Model). $\mathcal{D}$ is the finite reward domain. For $\epsilon > 0$, the bandit model is said to be (ϵ, δ)-LDP if

i) a randomized mapping $M : \mathcal{D} \to \mathbb{R}$ on $\mathcal{D} \subset \mathbb{R}^k$ if for any neighboring x, x' in $\mathcal{D}$ and a measurable subset S of $\mathbb{R}$, we have

$$\mathbb{P}\{M(x) \in S\} \leq e^\epsilon \mathbb{P}\{M(x') \in S\} + \delta.$$

ii) $\vartheta_{i+1}^\epsilon \in \sigma(\vartheta_j^\epsilon, M(Z_j^{\theta_\epsilon}) : 1 \leq j \leq i)$ for any time i.

In the relaxed LDP bandit model, the new privacy parameter δ represents the probability that privacy protection will fail. When the probability is 1-δ, we have the same level of protection as ϵ-LDP, but when the probability is δ, there is no guarantee of protection. As a result, it is common to require δ to be very small, often less than $\frac{1}{n^2}$, where n is the size of the data set. The Gaussian mechanism is a form of relaxed LDP bandit model.

The curator mechanism with Gaussian privacy is described in Algorithm 4. Meanwhile, Algorithm 1 receives rewards from $\mathrm{CTG}(\epsilon, \delta)$. Its corresponding parameters in Equation (1) are as follows:

$$\bar{\mu}_\epsilon = \bar{\mu}, \underline{\mu}_\epsilon = \underline{\mu}, \sigma_\epsilon^2 = \sigma^2 + \sigma_g^2. \tag{10}$$

Herein, $\sigma_g^2 = \frac{2\kappa^2 \ln(1.25/\delta)}{\epsilon^2}$. Obviously, our strategy $\{\theta_\epsilon^*\}$ is impacted by both the privacy budget ϵ and the failure probability δ. Leveraging Theorem 1 and Algorithm 4, we establish Corollary 2.

Corollary 2. Under the Gaussian mechanism, if the hypothesis $(\mu_L, \mu_R) = (\bar{\mu}, \underline{\mu})$, we can conclude that

$$T_n^{\theta_\epsilon^*} \xrightarrow{\mathcal{L}} \eta_1,\ \eta_1 \sim \mathcal{B}\left(\frac{\underline{\mu} - \bar{\mu}}{2}, \frac{\bar{\mu} + \underline{\mu}}{2}, c\right). \tag{11}$$

The convergence rate of $P(|T_n^{\theta_\epsilon^*} - c| \leq a)$ to $P(|\eta_1 - c| \leq a)$ is $O(\frac{\sqrt{\ln(1/\delta)}}{\epsilon\sqrt{n}})$. If $(\mu_R, \mu_L) = (\bar{\mu}, \underline{\mu})$, we will obtain a similar conclusion.

Algorithm 4: Convert-to-Gaussian$(\epsilon, \delta)(\mathrm{CTG}(\epsilon, \delta))$

On receiving a reward Z_i^θ from the user:

return $M_G(Z_i^\theta) = Z_i^\theta + G$, $G \sim N(0, \frac{2\kappa^2 \ln(1.25/\delta)}{\epsilon^2})$ distribution;

Hypothesis Test

A motivated example of sequential hypothesis testing as the responsible bandit model is presented in this section. Agents consider not only the best arm based on user preference, but also the gap $d_1 = \bar{\mu}_\epsilon - \underline{\mu}_\epsilon$ between the best and suboptimal arms in order to develop strategies and manage risks. In other words we would like to conduct the hypothesis test:

$$\mathbf{H}_0 : \bar{\mu}_\epsilon - \underline{\mu}_\epsilon \geq d_0; \quad \mathbf{H}_1 : \bar{\mu}_\epsilon - \underline{\mu}_\epsilon < d_0. \tag{12}$$

Without loss of generality, we assume that the sum of the average rewards for arm L and arm R, $\bar{\mu}_\epsilon + \underline{\mu}_\epsilon = d > 0$, is a constant. The constant d signifies our focus on understanding how the distance between the two arms affects the distribution (i.e., the concentration of information). In other words, as the total rewards of the arms reach a certain level, the agent becomes more concerned about the differences in arm rewards to manage risk. Each corresponds to a left or right margin:

$$\mathbf{H}_{L0} : \bar{\mu}_\epsilon \geq (d + d_0)/2; \quad \mathbf{H}_{L1} : \bar{\mu}_\epsilon < (d + d_0)/2;$$
$$\mathbf{H}_{R0} : \underline{\mu}_\epsilon \leq (d - d_0)/2; \quad \mathbf{H}_{R1} : \underline{\mu}_\epsilon > (d - d_0)/2.$$

The agent can naturally utilize traditional statistical tests based on the normal distribution (Fisher 1992) :

$$\frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \frac{W_{i,\epsilon}^L - \bar{\mu}_\epsilon}{\sigma_\epsilon} \quad \text{and} \quad \frac{1}{\sqrt{n_2}} \sum_{i=1}^{n_2} \frac{W_{i,\epsilon}^R - \underline{\mu}_\epsilon}{\sigma_\epsilon}. \tag{13}$$

In this hypothesis testing, we typically select arm L for the first n_1 trials to obtain rewards $\{W_{i,\epsilon}^L, i = 1, 2, \cdots, n_1\}$ and arm R for the remaining $n_2 = n - n_1$ trials obtain rewards $\{W_{i,\epsilon}^R, i = 1, 2, \cdots, n_2\}$. However, the test statistics do not take the strategy or sample performance into account and do not use prior information (Chen, Yan, and Zhang 2023). Under the privacy mechanism, the naive approach exhibits lower statistical power and requires more samples.

Due to the challenges in using traditional normal tests for statistical inference under the LDP mechanism, Equation (1) is introduced as a test statistic within the bandit framework. In reality, since agents are unaware of the specific privacy-preserving methods applied to the data, we refrain from making explicit assumptions about the privacy mechanisms. This sequential test statistic enables a more comprehensive exploration of its asymptotic distribution in a broader probability space, exhibiting increased concentration under the null hypothesis and reduced concentration under the alternative hypothesis, as demonstrated in Corollary 3.

Corollary 3. Under the LDP mechanism, if $(\mu_L, \mu_R) = (\bar{\mu}_\epsilon, \underline{\mu}_\epsilon)$, then we have

$$T_n^{\theta_\epsilon^*} \xrightarrow{\mathcal{L}} \sigma_d \eta_n,\ \eta_n \sim \mathcal{B}(\alpha_n, \beta_n, \gamma_n), \tag{14}$$

where $\beta_n = d/2\sigma_d$, $\gamma_n = c/\sigma_d$, $\alpha_n = -d_1/2 - \sqrt{n}(d_1 - d_0)/2\sigma_\epsilon$, $\sigma_d = \sqrt{1 + (d_1 - d_0)^2/4\sigma_\epsilon^2}$. In particular, for any $a \in \mathbb{R}$, let $c = d/2$, we have

$$\lim_{n\to\infty} P\left(\left|T_n^{\theta_\epsilon^*} - \frac{d}{2}\right| \le a\right) = \Phi\left(-\alpha_n + \frac{a}{\sigma_d}\right) - e^{\frac{2a\alpha_n}{\sigma_d}}\Phi\left(-\alpha_n - \frac{a}{\sigma_d}\right). \quad (15)$$

When assuming $\mathbf{H}_0$ has a true value of $\bar{\mu}_\epsilon - \underline{\mu}_\epsilon = d_0$, the parameter $\sigma_d = 1$ and $T_n^{\theta_\epsilon^*}$ follows the spike distribution $\mathcal{B}(-\frac{d_0}{2}, \frac{d}{2}, \frac{d}{2})$. Hence, we can reject the null hypothesis by the occurrence of event

$$\left\{\left|T_n^{\theta_\epsilon^*} - \frac{d}{2}\right| > z_{\frac{\alpha}{2}}\right\},$$

where $z_{\frac{\alpha}{2}}$ is the upper αth of the distribution $\mathcal{B}(-\frac{d_0}{2}, 0, 0)$. The related statistical efficiency can be calculated by

$$1 - \alpha = \Phi\left(\frac{d_0}{2} + z_{\frac{\alpha}{2}}\right) - e^{-d_0 z_{\frac{\alpha}{2}}}\Phi\left(\frac{d_0}{2} - z_{\frac{\alpha}{2}}\right). \quad (16)$$

When the distance between the two arms, $\bar{\mu}_\epsilon - \underline{\mu}_\epsilon = d_1 > d_0$, the first parameter

$$\alpha_n = -\frac{d_1}{2} - \frac{\sqrt{n}(d_1 - d_0)}{2\sigma_\epsilon} < -\frac{d_1}{2}. \quad (17)$$

This is because, even though α_n is currently dependent on σ_ϵ and μ_ϵ, the rate of change of α_n under the Gaussian mechanism and Laplace mechanism converges with the same convergence rates as the maximization probability ($O(\frac{\ln(1/\delta)}{\epsilon\sqrt{n}})$ and $O(\frac{1}{\epsilon\sqrt{n}})$, respectively). Meanwhile, the Bernoulli mechanism does not alter the relative magnitude of rewards. The above inequality holds true under LDP mechanism. The Bandit distribution will be steeper and the information will be more concentrated as α_n increases that is shown in Appendix, leading to a greater statistical efficiency.

When $\mathbf{H}_1$ is true with a value of $\bar{\mu}_\epsilon - \underline{\mu}_\epsilon = d_1 < d_0$, the associated statistical power is calculated as follows:

$$1 - \beta_1 = \lim_{n\to\infty} P\left(|T_n^{\theta_\epsilon^*} - \frac{d}{2}| > z_{\frac{\alpha}{2}}|\mathbf{H_1}\right) = 1 - \Phi\left(-\alpha_n + \frac{z_{\alpha/2}}{\sigma_d}\right) + e^{\frac{2\alpha_n z_{\alpha/2}}{\sigma_d}}\Phi\left(-\alpha_n - \frac{z_{\alpha/2}}{\sigma_d}\right). \quad (18)$$

Similar to Inequality (17), we can deduce that the parameter $\alpha_n > 0$. When α_n is positive, the statistic follows a bimodal distribution, as shown in the Appendix, significantly increasing the tail probability. This substantial reduction in noise interference makes it a much better alternative to assumption testing using a normal distribution.

Within this responsible bandit learning framework, hypothesis testing is conducted in a way that harnesses both prior information and the learning nature of the bandit process to enhance the test's power and convergence rate. Moreover, the statistical attributes of parameter testing are leveraged to offer clear insights into the statistical inference outcomes across various hypotheses and sample sizes. This becomes particularly crucial within the realm of privacy frameworks.

Simulation Studies

The section will examine the performance of the proposed test statistic $T_n^{\theta_\epsilon^*}$ under the LDP mechanism. To test the asymptotic distribution and corresponding critical value at a significance level of $T_n^{\theta_\epsilon^*}$, the following treatment is carried out. For the Laplace and Gaussian mechanisms, it is assumed that W_i^L is drawn from random numbers generated within the interval $[2\bar{\mu} - 1, 1]$, following the distribution $N(\bar{\mu}, 1)$, and W_i^R is drawn from random numbers generated within the interval $[0, 2\underline{\mu}]$, following the distribution $N(\underline{\mu}, 1)$. For the Bernoulli mechanism, it is assumed that $W^L \sim U(2\bar{\mu} - 1, 1)$ and $W^R \sim U(0, 2\underline{\mu})$ to ensure rewards are within the interval $[0, 1]$. Here, $\kappa = 1$, $\bar{\mu} = 0.9, 0.8, 0.7$, and $\underline{\mu} = 0.1, 0.2, 0.3$. The privacy budget and failure probability are set as $\epsilon = 0.5, 1$, and $\delta = 0.02$, assuming that σ_ϵ is known. Referring to the maximum probability achieved by $T_n^{\theta_\epsilon^*}$ under the Laplace mechanism when $n > 2000$, as shown in Figure 3, we set $n = 5000$ to discuss the asymptotic distribution of $T_n^{\theta_\epsilon^*}$ under the LDP mechanism. With the true distance of $\Delta_\mu = \bar{\mu} - \underline{\mu} = 0.6$ and $d = 1$, the critical values and results of the asymptotic distribution at a significance level of $\alpha = 0.05$ are presented in Table 1 and Figure 4. It can be observed that under the responsible strategy of Lemma 1, $T_n^{\theta_\epsilon^*}$ can rapidly achieve maximum probability coverage over the given interval centered at c.

	d_0	$z^1_{\alpha/2}$	significant level		
			Laplace	Bernoulli	Gaussian
$\epsilon = 1$	0.8	1.64(1.81)	0.050	0.048	0.056
	0.6	1.72(1.85)	0.054	0.046	0.056
	0.4	1.80(1.88)	0.050	0.050	0.056
$\epsilon = 0.5$	0.8	1.64(1.88)	0.052	0.048	0.062
	0.6	1.72(1.90)	0.050	0.042	0.060
	0.4	1.80(1.92)	0.060	0.046	0.052

[1] The critical value under the corresponding Bernoulli mechanism is in parentheses.

Table 1: The estimated proportions $P(|T_n^{\theta_\epsilon^*} - \frac{d}{2}| > z_{\alpha/2})$ with $\bar{\mu} - \underline{\mu} = d_0$ in privacy mechanisms after 500 replicates.

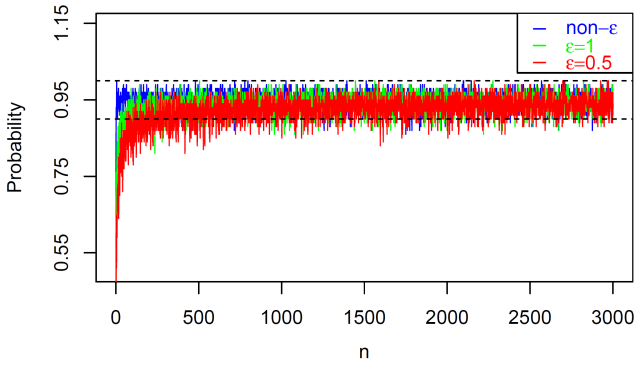

Figure 3: The rate at which the $T_n^{\theta_\epsilon^*}$ achieves maximum probability under the Laplace mechanism.

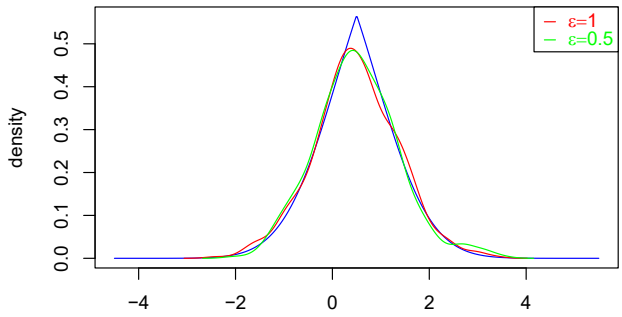

Figure 4: The true density plot (blue line) and the estimated density plot (red and green lines) of $T_n^{\theta^*_\epsilon}$ under the Laplace mechanism.

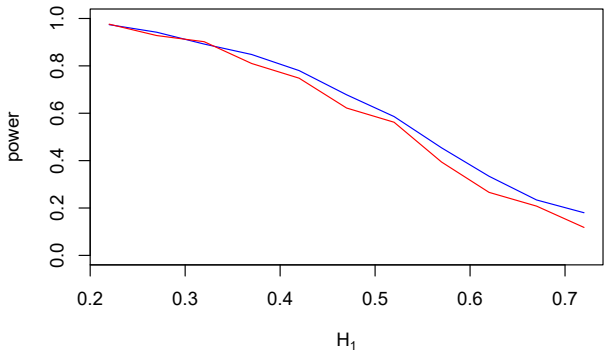

Figure 5: The power curve plots of the proposed test (blue line) and the normal test (red line) at different upper expectations $d_1 < d_0$ under the alternative hypothesis and $\epsilon = 1, \Delta_\mu = 0.8$.

The calculation of the normal test statistic is obtained by selecting arm L for the first $n/2$ trials and arm R for the remaining $n/2$ trials, as illustrated in Equation (13). When considering the alternative hypothesis, distinct values of d_1, ϵ, and d are chosen for comparing the power of the two tests. Leveraging the convergence rate of the empirical distribution of $T_n^{\theta^*_\epsilon}$ under the Bernoulli, Laplace, and Gaussian mechanisms, we repeat the procedure 500 times with $n = 100$, 300, and 900, while setting the distance $\Delta_\mu = 0.8$. In comparison with the power of the normality test, the power of the proposed test under the Laplace mechanism is depicted in Figures 5-7. Results for the Bernoulli and Gaussian mechanisms are presented in the Appendix. It is evident from Figures 5-7 that the superiority of the proposed test statistic persists as the distance and total sum between the two rewards vary. Furthermore, as more noise is introduced, the advantage of our statistic becomes more pronounced. And the superiority of the proposed test statistic remains equally significant under the Bernoulli and Gaussian mechanisms.

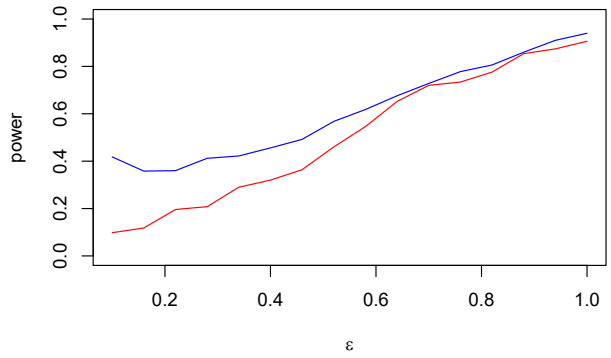

Figure 6: The power curve plots of the proposed test (blue line) and the normal test (red line) at different privacy budgets ϵ under the $d_1 = 0.3$ and $\Delta_\mu = 0.8$.

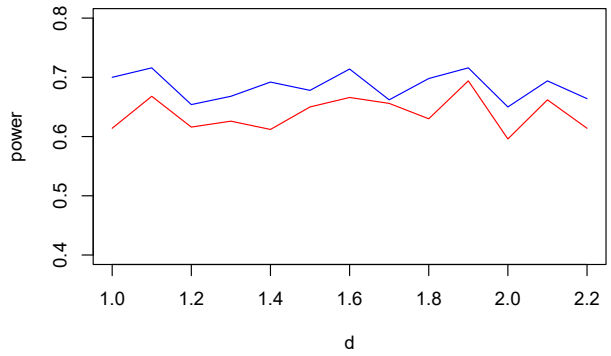

Figure 7: The power curve plots of the proposed test (blue line) and the normal test (red line) at different values of d under the $d_1 = 0.4$, $\epsilon = 1$, and $\Delta_\mu = 0.8$.

Discussion

To address the issue that existing privacy-preserving bandit algorithms cannot be responsible for real-time risk in decision-making, we introduce the mean-volatility as a new objective utility. We maximize the probability of this utility falling within a given interval to take responsibility for risk. By conducting optimal learning, the probability of $T_n^{\theta^*_\epsilon}$ falling within a given interval has been maximized at the fastest rate of $O(\frac{1}{\epsilon\sqrt{n}})$. Utilizing the learned statistical quantity for sequential testing also demonstrates higher power compared to normal testing. In the future, researchers can leverage these findings for practical applications, such as developing optimal privacy-responsible bandit learning algorithms for recommendation systems and clinical trial designs.

Acknowledgements

We are especially grateful to the anonymous reviewers for great feedback on the paper. In addition, Dr. Yan was supported by National Key R&D Program of China (No.

2023YFA1008701), the National Natural Science Foundation of China (No.12371292), the National Statistical Science Research Project (No. 2022LY080) and Jinan Science and Technology Bureau (No. 2021GXRC056).

References

Basu, D.; Dimitrakakis, C.; and Tossou, A. 2019. Privacy in multi-armed bandits: Fundamental definitions and lower bounds. *arXiv preprint arXiv:1905.12298*.

Chen, Z.; and Epstein, L. G. 2022. A central limit theorem for sets of probability measures. *Stochastic Processes and their Applications*, 152: 424–451.

Chen, Z.; Feng, S.; and Zhang, G. 2022. Strategy-Driven Limit Theorems Associated Bandit Problems. *arXiv preprint arXiv:2204.04442*.

Chen, Z.; Feng, X.; Liu, S.; and Yan, X. 2023. Optimal distributions of rewards for a two-armed slot machine. *Neurocomputing*, 518: 401–407.

Chen, Z.; Yan, X.; and Zhang, G. 2023. Strategic two-sample test via the two-armed bandit process. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkad061.

Cui, W.; Ji, X.; Kong, L.; and Yan, X. 2023. Opposite online learning via sequentially integrated stochastic gradient descent estimators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7270–7278.

Dwork, C. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, 1–19. Springer.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407.

Fisher, R. A. 1992. Statistical methods for research workers. In *Breakthroughs in statistics*, 66–70. Springer.

Gajane, P.; Urvoy, T.; and Kaufmann, E. 2018. Corrupt bandits for preserving local privacy. In *Algorithmic Learning Theory*, 387–412. PMLR.

Han, Y.; Liang, Z.; Wang, Y.; and Zhang, J. 2021. Generalized linear bandits with local differential privacy. *Advances in Neural Information Processing Systems*, 34: 26511–26522.

Huang, L.; Joseph, A. D.; Nelson, B.; Rubinstein, B. I.; and Tygar, J. D. 2011. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, 43–58.

Martin, K. D.; and Murphy, P. E. 2017. The role of data privacy in marketing. *Journal of the Academy of Marketing Science*, 45(2): 135–155.

Petrescu, M.; and Krishen, A. S. 2018. Analyzing the analytics: data privacy concerns.

Price, W. N.; and Cohen, I. G. 2019. Privacy in the age of medical big data. *Nature medicine*, 25(1): 37–43.

Ren, W.; Zhou, X.; Liu, J.; and Shroff, N. B. 2020. Multi-armed bandits with local differential privacy. *arXiv preprint arXiv:2007.03121*.

Shin, H.; Kim, S.; Shin, J.; and Xiao, X. 2018. Privacy enhanced matrix factorization for recommendation with local differential privacy. *IEEE Transactions on Knowledge and Data Engineering*, 30(9): 1770–1782.

Talukder, N.; Ouzzani, M.; Elmagarmid, A. K.; Elmeleegy, H.; and Yakout, M. 2010. Privometer: Privacy protection in social networks. In *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, 266–269. IEEE.

Tossou, A. C.; and Dimitrakakis, C. 2016. Algorithms for differentially private multi-armed bandits. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Villar, S. S.; Bowden, J.; and Wason, J. 2015. 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.

Wang, H.; Zhao, Q.; Wu, Q.; Chopra, S.; Khaitan, A.; and Wang, H. 2020. Global and local differential privacy for collaborative bandits. In *Proceedings of the 14th ACM Conference on Recommender Systems*, 150–159.

Williams, J. J.; Nogas, J.; Deliu, N.; Shaikh, H.; Villar, S. S.; Durand, A.; and Rafferty, A. 2021. Challenges in statistical analysis of data collected by a bandit algorithm: An empirical exploration in applications to adaptively randomized experiments. *arXiv preprint arXiv:2103.12198*.

Yao, J.; Brunskill, E.; Pan, W.; Murphy, S.; and Doshi-Velez, F. 2021. Power constrained bandits. In *Machine Learning for Healthcare Conference*, 209–259. PMLR.