Item-Oriented Personalized Local Differential Privacy for Discrete Distribution Estimation

Xin Li, Hong Zhu[0000−0001−9815−3934], Zhiqiang Zhang, and Meiyi Xie[0000−0002−1973−7470]

School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, 430074, China
{lising,zhuhong,kylinzhang,xiemeiyi}@hust.edu.cn

Abstract. Discrete distribution estimation is a fundamental statistical tool, which is widely used to perform data analysis tasks in various applications involving sensitive personal information. Due to privacy concerns, individuals may not always provide their raw information, which leads to unpredictable biases in the final results of estimated distribution. Local Differential Privacy (LDP) is an advanced technique for privacy protection of discrete distribution estimation. Currently, typical LDP mechanisms provide same protection for all items in the domain, which imposes unnecessary perturbation on less sensitive items and thus degrades the utility of final results. Although, several recent works try to alleviate this problem, the utility can be further improved. In this paper, we propose a novel notion called Item-Oriented Personalized LDP (IPLDP), which independently perturbs different items with different privacy budgets to achieve personalized privacy protection. Furthermore, to satisfy IPLDP, we propose the Item-Oriented Personalized Randomized Response (IPRR) based on the observation that the sensitivity of data shows an inverse relationship with the population size of respective individuals. Theoretical analysis and experimental results demonstrate that our method can provide fine-grained privacy protection and improve data utility simultaneously.

Keywords: Discrete distribution estimation · Local differential privacy · Item-oriented personalization · Randomized response

1 Introduction

Discrete distribution estimation is widely used as a fundamental statistics tool and has achieved significant performance in various data analysis tasks, including frequent pattern mining [16], histogram publication [39], and heavy hitter identification [33]. With the deepening and expansion of application scenarios, these data analysis tasks inevitably involve more and more sensitive personal data. Due to the privacy concerns, individuals may not always be willing to truthfully provide their personal information. When dealing with such data, however, discrete distribution estimation is difficult to play its due role. For instance, a health organization plans to make statistics about two epidemic diseases: HIV

and Hepatitis, so they issued a questionnaire survey containing three options: HIV, Hepatitis, and None, to inquire whether the citizens suffer from these two diseases. Undoubtedly, this question is highly sensitive, especially for people who actually have these two diseases. As a result, they have a high probability to give false information when filling the the questionnaire, and this will eventually lead to unpredictable biases in the estimation of the distribution of diseases. Therefore, under the requirement of privacy protection, how to conduct discrete distribution estimation is increasingly drawn the attention of researchers.

Differential Privacy (DP) [12, 13] is an advanced and promising technique for privacy protection. Benefiting from its rigorous mathematical definition and lightweight computation demand, DP has rapidly become one of the trend in the field of privacy protection. Generally, we can categorize DP into Centralized DP (CDP) [6, 10, 13, 21, 23] and Local DP (LDP) [8, 11, 32]. Compared with the former, the latter does not require a trusted server, and hence it is much more appropriate for privacy protection in the tasks of the discrete distribution estimation. Based on Randomized Response (RR) [37] mechanism, LDP provides different degrees of privacy protection through the assignment of different privacy budget. Currently, typical LDP mechanisms, such as K-ary RR (KRR) [19] and RAPPOR [14], perturb all items in the domain with the same privacy budget, thus providing uniform protection strength. However, in practical scenarios, each item's sensitivity is different rather than fixed, and the number of individuals involved is also inversely proportional to their sensitivity level. For example, in the questionnaire mentioned above, HIV undoubtedly has a much higher sensitivity than Hepatitis, but a relatively less population of individuals with it than that of the latter. Additionally, None is a non-sensitive option, which naturally accounts for the largest population. Therefore, if we provide privacy protection for all items at the same level without considering their distinct sensitivity, unnecessary perturbation will be imposed on those less sensitive (and even non-sensitive) items that account for a much more population, which severely degrades the data utility of the final result.

Recently, several works proposed to improve the utility by providing different levels of protection according to various sensitivities of items. Murakami et al. introduced the Utility-Optimized LDP (ULDP) [28], which partitions personal data into sensitive and non-sensitive data and ensures privacy protection for the sensitive data only. While ULDP have better utility than KRR and RAPPOR by distinguishing sensitive data from non-sensitive data, it still protects all the sensitive data at the same level without considering the different sensitivities among them. After that, Gu et al. proposed the Input-Discriminative LDP (ID-LDP) [15], which further improved utility by providing fine-grained privacy protection for items with different privacy budgets of inputs. However, under ID-LDP, the strength of perturbation is severely restricted by the minimum privacy budget. As the minimum privacy budget decreases, the corresponding perturbations imposed on different items will approach the maximum level, which greatly weakens the improvement of utility brought by differentiating handling for each item with a independent privacy budget, thereby limiting the applicability of this method.

Therefore, the current methods of discrete distribution estimation in local privacy setting leave much room to improve the utility. In this paper, we propose a novel notion of LDP named Item-Oriented Personalized LDP (IPLDP). Unlike previous works, IPLDP independently perturbs different items with different privacy budgets to achieve personalized privacy protection and utility improvement simultaneously. Through independent perturbation, the strength of perturbation imposed on those less sensitive items will never be influenced by the sensitivity of others. To satisfy IPLDP, we propose a new mechanism called Item-Oriented Personalized RR (IPRR), and it uses the direct encoding method as in KRR to guarantee the equivalent protection for inputs and outputs simultaneously.

Our main contributions are:

1. We propose a novel LDP named IPLDP, which independently perturbs different items with different privacy budgets to achieve personalized privacy protection and utility improvement simultaneously.
2. We propose IPRR mechanism to provide equivalent protection for inputs and outputs simultaneously using the direct encoding method.
3. By calculating the $l1$ and $l2$ losses through the unbiased estimator of the gound-truth distribution under IPRR, we theoretically prove that our method has tighter upper bounds than that of existing direct encoding mechanisms.
4. We evaluate our IPRR on a synthetic and a real-world dataset with the comparison with the existing methods. The results demonstrate that our method has better performance than existing methods in data utility.

The remainder of this paper is organized as follows. Section 2 lists the related works. Section 3 provides an overview of several preliminary concepts. Section 4 presents the definition of IPLDP. Section 5 discusses the design of our RR mechanism and its empirical estimator. Section 6 analyzes the utility of the proposed RR method. Section 7 shows the experimental results. Finally, in Section 8, we draw the conclusions.

2 Related Work

Since DP was firstly proposed by Dwork [12], it has attracted much attention from researchers, and numerous variants of DP have been studied including d-privacy [7], Pufferfish privacy [21], dependent DP [25], Bayesian DP [40], mutual information DP [10], R´enyi DP [27], Concentrated DP [6], and distribution privacy [20]. However, all of these methods require a trusted central server. To address this issue, Duchi et al. [11] proposed LDP, which quickly became popular in a variety of application scenarios, such as frequent pattern mining [9, 31, 32], histogram publication [5], heavy-hitter identification [4, 33], and graph applications [24, 30, 38]. Based on RR [37] mechanism, LDP provides different degrees of privacy protection through the assignment of privacy budget. Currently, typical RR mechanisms, such as KRR [19] and RAPPOR [14], perturb all items in the domain with the same privacy budget, thus providing uniform protection strength.

In recent years, several fine-grained privacy methods have been developed for both centralized and local settings. For example, in the centralized setting, Personalized DP [17, 29], Heterogeneous DP [3], and One-sided DP [22] have been studied. In the local setting, Murakami et al. proposed ULDP [28], which partitions the value domain into sensitive and non-sensitive sub-domains. While ULDP optimizes utility by reducing perturbation on non-sensitive values, it does not fully consider the distinct privacy requirements of sensitive values. Gu et al. introduced ID-LDP [15], which protects privacy according to the distinct privacy requirements of different inputs. However, the perturbation of each value is influenced by the minimum privacy budget. As the minimum privacy budget decreases, the perturbations of different items approach the maximum, which degrades the improvement of utility.

3 Preliminaries

In this section, we formally describe our problem. Then, we describe the definitions of LDP and ID-LDP. Finally, we introduce the distribution estimation and utility evaluation methods.

3.1 Problem Statement

A data collector or a server desires to estimate the distribution of several discrete items from n users. The set of all personal items held by these users and its distribution are denoted as $\mathcal{D}$ and $\boldsymbol{p} \in \mathbb{S}^{|\mathcal{D}|}$, respectively, where $\mathbb{S}$ stands for a probability simplex and $|\cdot|$ is the cardinality of a set. For each $x \in \mathcal{D}$, we use $\boldsymbol{p}_x$ to denote its respective probability. We also have a set of random variables $X^n = \{X_1, ..., X_n\} \in \mathcal{D}$ held by n users, which are drawn i.i.d. according to $\boldsymbol{p}$. Additionally, since the items may be sensitive or non-sensitive for users, we divide $\mathcal{D}$ into two disjoint partitions: $\mathcal{D}_S$, which contains sensitive items, and $\mathcal{D}_N$, which contains non-sensitive items.

Because of privacy issues, users perturb their items according to a privacy budget set $\mathcal{E} = \{\varepsilon_x\}_{x \in \mathcal{D}_S}$, where ε_x is the corresponding privacy budget of $x \in \mathcal{D}_S$. After perturbation, the data collector can only estimate $\boldsymbol{p}$ from users by observing $Y^n = \{Y_1, ..., Y_n\}$ which is the perturbed version of X^n through a mechanism $\boldsymbol{Q}$, and the mechanism $\boldsymbol{Q}$ maps an input item $x \in \mathcal{D}$ to an output $y \in \mathcal{D}$ with probability $\boldsymbol{Q}(y|x)$.

Our goals are: (1) to design $\boldsymbol{Q}$ that maps inputs $\forall x \in \mathcal{D}$ to outputs $\forall y \in \mathcal{D}$ according to the corresponding $\varepsilon_y \in \mathcal{E}$, and improves data utility as much as possible; (2) to estimate the distribution vector $\boldsymbol{p}$ from Y^n.

We assume that the data collector or the server is untrusted, and users never report their data directly but randomly choose an item from $\mathcal{D}$ to send, where $\mathcal{D}$ is shared by both the server and users. $\mathcal{E}$ should be also public with $\mathcal{D}$, so that users can calculate $\boldsymbol{Q}$ for perturbation, and the server can calibrate the result according to $\boldsymbol{Q}$.

3.2 Local Differential Privacy

In LDP [11], each user perturbs its data randomly and then send the perturbed data to the server. The server can only access these perturbed results, which

guarantees the privacy. In this section, we list two definitions of LDP notions, that is, the standard LDP and the ID-LDP [15].

Definition 1 (ε-LDP). *A randomized mechanism $\boldsymbol{Q}$ satisfies ϵ-LDP if, for any pair of inputs x, x', and any output y:*

$$e^{-\varepsilon} \leqslant \frac{\boldsymbol{Q}(y|x)}{\boldsymbol{Q}(y|x')} \leqslant e^{\varepsilon}, \tag{1}$$

where $\varepsilon \in \mathbb{R}^+$ is the privacy budget that controls the level of confidence an adversary can distinguish the output from any pair of inputs. Smaller ε means that an adversary feels less confidence for distinguishing y from x or x', which naturally provides a stronger privacy protection.

Definition 2 ($\mathcal{E}$-ID-LDP). *For a given privacy budget set $\mathcal{E} = \{\varepsilon_x\}_{x\in\mathcal{D}} \in \mathbb{R}_+^{|\mathcal{D}|}$, a randomized mechanism $\boldsymbol{Q}$ satisfies $\mathcal{E}$-ID-LDP if, for any pair of inputs $x, x' \in \mathcal{D}$, and any output $y \in Range(\boldsymbol{Q})$:*

$$e^{-r(\varepsilon_x,\varepsilon_{x'})} \leqslant \frac{\boldsymbol{Q}(y|x)}{\boldsymbol{Q}(y|x')} \leqslant e^{r(\varepsilon_x,\varepsilon_{x'})}, \tag{2}$$

where $r(\cdot,\cdot)$ is a system function of two privacy budgets.

Generally, we use $\mathcal{E}$-MinID-LDP in practical scenarios, where $r(\varepsilon_x, \varepsilon_{x'}) = \min(\varepsilon_x, \varepsilon_{x'})$.

3.3 Distribution Estimation Method

The empirical estimation [18] and the maximum likelihood estimation [18, 34] are two types of useful methods for estimating discrete distribution in local privacy setting. We use the former method in our theoretical analysis and use both in our experiments. Here, we explain the details of the empirical estimation.

Empirical estimation method The empirical estimation method calculates the emprical estimate $\hat{\boldsymbol{p}}$ of $\boldsymbol{p}$ using the empirical estimate $\hat{\boldsymbol{m}}$ of the distribution $\boldsymbol{m}$, where $\boldsymbol{m}$ is the distribution of the output of the mechanism $\boldsymbol{Q}$. Since both $\boldsymbol{p}$ and $\boldsymbol{m}$ are $|\mathcal{D}|$-dimensional vectors, $\boldsymbol{Q}$ can be viewed as a $|\mathcal{D}| \times |\mathcal{D}|$ conditional stochastic matrix. Then, the relationship between $\boldsymbol{p}$ and $\boldsymbol{m}$ can be given by $\boldsymbol{m} = \boldsymbol{pQ}$. Once the data collector obtains the observed estimation $\hat{\boldsymbol{m}}$ of $\boldsymbol{m}$ from Y^n, the estimation of $\boldsymbol{p}$ can be solved by $\hat{\boldsymbol{m}} = \hat{\boldsymbol{p}}\boldsymbol{Q}$. As n increases, $\hat{\boldsymbol{m}}$ remains unbiased for $\boldsymbol{m}$, and hence $\hat{\boldsymbol{p}}$ converges to $\boldsymbol{p}$ as well. However, when the sample count n is small, some elements in $\hat{\boldsymbol{p}}$ can be negative. To address this problem, several normalization methods [34] can be utilized to truncate and normalize the result.

3.4 Utility Evaluation Method

In this paper, the l_2 and l_1 losses is utilized for our theoretical analysis of utility. Mathematically, they are defined as $l_2(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \sum_{x\in\mathcal{D}}(\hat{\boldsymbol{p}}_x - \boldsymbol{p}_x)^2$, and $l_1(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \sum_{x\in\mathcal{D}} |\hat{\boldsymbol{p}}_x - \boldsymbol{p}_x|$. Both l_2 and l_1 losses evaluate the total distance between the estimate value and the ground-truth value. The shorter the distance, the better the data utility.

4 Item-Oriented Personalized LDP

In this section, we first introduce the definition of our proposed IPLDP. Then, we discuss the relationship between IPLDP and LDP. Finally, we compare IPLDP with MinID-LDP.

4.1 Privacy Definition

The standard LDP provides the same level of protection for all items using a uniform privacy budget, which can result in excessive perturbation for less sensitive items and lead to poor utility. To improve the utility, ID-LDP uses distinct privacy budgets for the perturbation of different inputs to provide fine-grained protection. Since all perturbations are influenced by the minimum privacy budget, the strength of all perturbations will be forced to approach the maximum value as the minimum privacy budget decreases. To avoid this problem, IPLDP uses different privacy budgets for outputs of the mechanism to provide independent protection for each item. However, using the output as the protection target may not provide equal protection for the input items. Therefore, in IPLDP, we force the input and output domains to be the same $\mathcal{D}$. Formally, IPLDP is defined as follows.

Definition 3 (**$(\mathcal{D}_S, \mathcal{E})$-IPLDP**). *For a privacy budget set $\mathcal{E} = \{\varepsilon_1, \cdots, \varepsilon_{|\mathcal{D}_S|}\} \in \mathbb{R}_+^{|\mathcal{D}_S|}$, a randomized mechanism $\boldsymbol{Q}$ satisfies $(\mathcal{D}_S, \mathcal{E})$-IPLDP if and only if it satisfies following conditions:*

1. *for any $x, x' \in \mathcal{D}$ and for any $x_i \in \mathcal{D}_S (i = 1, \cdots, |\mathcal{D}_S|)$,*

$$e^{-\varepsilon_i} \leqslant \frac{\boldsymbol{Q}(x_i|x)}{\boldsymbol{Q}(x_i|x')} \leqslant e^{\varepsilon_i}, \tag{3}$$

2. *for any $x \in \mathcal{D}_N$ and for any $x' \in \mathcal{D}$,*

$$\boldsymbol{Q}(x|x') > 0 \text{ and } \boldsymbol{Q}(x|x') = 0 \text{ for any } x \neq x' \tag{4}$$

Since non-sensitive items need no protection, the corresponding privacy budget can be viewed as an infinity value. However, we cannot set the privacy budget to infinity in practice. Hence, inspired by ULDP, IPLDP handles $\mathcal{D}_S$ and $\mathcal{D}_N$ separately.

According to the definition, IPLDP guarantees that the adversary's ability to distinguish any $y \in \mathcal{D}_S$ whether it is from any pair of inputs $x, x' \in \mathcal{D}$ would not exceed the range determined by the respected ε_y. That is to say, for $\forall x \in \mathcal{D}_S$, it should satisfy ε_x-LDP. For $\forall x \in \mathcal{D}_N$, It can only be perturbed to any $x \in \mathcal{D}_S$ or itself.

4.2 Relationship with LDP

We hereby assume $\mathcal{D} = \mathcal{D}_S$. Then, the obvious difference between LDP and IPLDP is the number of the privacy budgets. A special case is that, when all the privacy budgets are identical, i.e. $\varepsilon_x = \varepsilon$ for all $x \in \mathcal{D}$, then IPLDP becomes the general ε-LDP. Without loss of generality, on the one hand, if a mechanism that

satisfies ε-LDP, it also satisfies $(\mathcal{D}, \mathcal{E})$-IPLDP for all $\mathcal{E}$ with $\min\{\mathcal{E}\} = \varepsilon$. On the other hand, if a mechanism satisfies $(\mathcal{D}, \mathcal{E})$-IPLDP, it also satisfies $\max\{\mathcal{E}\}$-LDP. Therefore, IPLDP can be viewed as a relaxed version of LDP. Noticeably, the relaxation does not mean that IPLDP is weaker than LDP in terms of the privacy protection, but LDP is too strong for items with different privacy needs. IPLDP has the ability to guarantee the personalized privacy for each item.

4.3 Comparison with MinID-LDP

According to the definition of notion, the main difference between IPLDP and MinID-LDP lies in the corresponding target of the privacy budget. Our IPLDP controls the distinguishability according to the output, while MinID-LDP focuses on the any pair of inputs. Both notions can be considered as a relaxed version of LDP. However, from Lemma 1 in [15], $\mathcal{E}$-MinID-LDP relaxes LDP in $\varepsilon = 2\min\{\mathcal{E}\}$ at most, which means that the degree of relaxation is much lower than IPLDP with the same $\mathcal{E}$. Therefore, as the minimum privacy budget of $\mathcal{E}$ decreases, the utility improvement under MinID-LDP is limited, and we will further experimentally verify this in Section 7.

5 Item-Oriented Personalized Mechanisms and Distribution Estimation

In this section, to provide personalized protection, we first propose our IPRR mechanism for the sensitive domain $\mathcal{D}_S = \mathcal{D}$. We then extend the mechanism to be compatible with the non-sensitive domain $\mathcal{D}_N$. Finally, we present the unbiased estimator of IPRR using the empirical estimation method.

5.1 Item-Oriented Personalized Randomized Response

According to our definition of IPLDP, it focuses on the indistinguishability of the mechanism's output. Then, the input and output domains should keep the same to ensure the equivalent protection for both inputs and outputs. Therefore, the only way to design the mechanism $\boldsymbol{Q}$ is to use the same direct encoding method as in KRR. To use such method, we need to calculate $|\mathcal{D}|^2$ different probabilities for the $|\mathcal{D}| \times |\mathcal{D}|$ stochastic matrix of $\boldsymbol{Q}$. However, it is impossible to directly calculate these probabilities which make $\boldsymbol{Q}$ invertible and satisfy IPLDP constraints simultaneously. To calculate all the probabilities of $\boldsymbol{Q}$, a possible way is to find an optimal solution of minimizing the expectation of $l_2(\hat{\boldsymbol{p}}, \boldsymbol{p})$ subject to the constraints of IPLDP, i.e.

$$\min_{\boldsymbol{Q}} \mathop{\mathbb{E}}_{Y^n \sim \boldsymbol{m}(\boldsymbol{Q})} [l_2(\hat{\boldsymbol{p}}, \boldsymbol{p})] \quad \text{s.t.} \quad \ln |\boldsymbol{Q}(y|x)/\boldsymbol{Q}(y|x')| \leqslant \varepsilon_y, (\forall x, x', y \in \mathcal{D}). \tag{5}$$

Nevertheless, we still can not directly solve this optimization problem. Firstly, it is complicated to calculate a close-form of $\boldsymbol{Q}$, since the objective function is likely to be non-convex and all constraints are non-linear inequalities. Secondly, even if we solve this problem numerically, the complexity of each iteration will become very large as the cardinality of the items increases, since we have to calculate an inverse matrix of $\boldsymbol{Q}$ to calculate the objective function in (5).

To address this problem, we reconsider the relationship between the privacy budget and the data utility. The privacy budget determines the indistinguishability of each item, i.e., $y \in \mathcal{D}$, by controlling the bound of $|\ln[\boldsymbol{Q}(y|x)/\boldsymbol{Q}(y|x')]|$ for any $x, x' \in \mathcal{D}$. Among all inputs, the contribution to the data utility comes from the honest answers (when $y = x$). Therefore, within the range controlled by the privacy budget, as long as the more honest answer can be distinguished from the dishonest ones, the more the utility can be improved. In other words, the ratio of $\boldsymbol{Q}(x|x)$ (denote as q_x) and $\boldsymbol{Q}(x|x')$ (denote as $\bar{q}_x$) should be as large as possible within the bound dominated by ε_x. Hence, we can reduce the computation complexity of probabilities from $|\mathcal{D}|^2$ to $2|\mathcal{D}|$ by making a tradeoff of forcing all $\bar{q}_x$ to be identical for all $x' \neq x$, and

$$q_x = e^{\varepsilon_x}\bar{q}_x, x \in \mathcal{D}. \tag{6}$$

Then, we can calculate each element $\boldsymbol{p}_x$ of $\boldsymbol{p}$ through each element $\boldsymbol{m}_x$ of $\boldsymbol{m}$ for all $x \in \mathcal{D}$ as follows:

$$\boldsymbol{m}_x = \boldsymbol{p}_x q_x + (1 - \boldsymbol{p}_x)\bar{q}_x = \boldsymbol{p}_x(e^{\varepsilon_x} - 1)\bar{q}_x + \bar{q}_x. \tag{7}$$

Next, we use the estimate $\hat{\boldsymbol{m}}$ and $\hat{\boldsymbol{p}}$ with (7) to calculate our objective function in (5). Since $n\hat{\boldsymbol{m}}_x$ follows the binomial distribution with parameters n and $\boldsymbol{m}_x$, its mean and variance are $\mathbb{E}(n\hat{\boldsymbol{m}}) = n\boldsymbol{m}_x$ and $\mathrm{Var}(n\hat{\boldsymbol{m}}) = n\boldsymbol{m}_x(1 - \boldsymbol{m}_x)$. We now can calculate the objective function in (5) according to (7):

$$\begin{aligned}\mathop{\mathbb{E}}_{Y^n \sim \boldsymbol{m}(\boldsymbol{Q})}[\ell_2(\hat{\boldsymbol{p}}, \boldsymbol{p})] &= \sum_{x\in\mathcal{D}} \mathbb{E}\left[(\hat{\boldsymbol{p}}_x - \boldsymbol{p}_x)^2\right] = \sum_{x\in\mathcal{D}} \frac{1}{(e^{\varepsilon_x} - 1)^2 \bar{q}_x^2} \cdot \frac{\boldsymbol{m}_x - \boldsymbol{m}_x^2}{n} \\ &= \sum_{x\in\mathcal{D}} \frac{\boldsymbol{p}_x (e^{\varepsilon_x} - 1) + 1}{n (e^{\varepsilon_x} - 1)^2 \bar{q}_x} - \sum_{x\in\mathcal{D}} \frac{[\boldsymbol{p}_x (e^{\varepsilon_x} - 1) + 1]^2}{n (e^{\varepsilon_x} - 1)^2}.\end{aligned} \tag{8}$$

The second term and n in (8) can be viewed as constants since they are irrelevant to $\bar{q}_x$. Therefore, by omitting these two constants, our final optimization problem can be given as

$$\min_{\{q_x, \bar{q}_x\}_{x\in\mathcal{D}}} \sum_{x\in\mathcal{D}} \frac{\boldsymbol{p}_x (e^{\varepsilon_x} - 1) + 1}{(e^{\varepsilon_x} - 1)^2 \bar{q}_x} \quad \text{s.t.} \quad q_x + \sum_{x'\in\mathcal{D}\setminus\{x\}} \bar{q}_{x'} = 1, \forall x \in \mathcal{D}. \tag{9}$$

Since the objective is a convex function of $\bar{q}_x$ for all $x \in \mathcal{D}$ and all the constraints are linear equations, we can efficiently calculate all the q_x and $\bar{q}_x$ for all $x \in \mathcal{D}$ via the Sherman-Morrison formula [2] at the intersection point of the hyper planes formed by the constraints in (9). After solving the linear equaltion groups, we can finally define our Item-Oriented Personalized RR (IPRR) mechanism as follows.

Definition 4 ($(\mathcal{D}, \mathcal{E})$-IPRR). *Let $\mathcal{D} = \{x_1, \cdots, x_{|\mathcal{D}|}\}, \mathcal{E} = \{\varepsilon_1, \cdots, \varepsilon_{|\mathcal{D}|}\} \in \mathbb{R}_+^{|\mathcal{D}|}$, Then $(\mathcal{D}, \mathcal{E})$-IPRR is a mechanism that maps $x' \in \mathcal{D}$ to $x \in \mathcal{D}$ with the probability $\boldsymbol{Q}_{\mathrm{IPRR}}(x|x')$ defined by*

$$\boldsymbol{Q}_{\mathrm{IPRR}}(x|x') = \begin{cases} q_x & \text{if } x = x', \\ \bar{q}_x & \text{otherwise}, \end{cases} \tag{10}$$

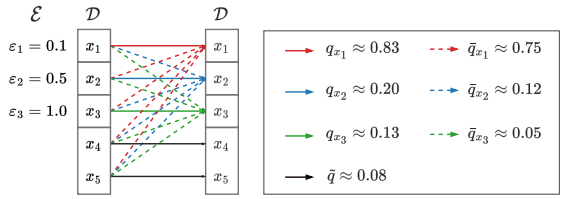

Fig. 1: Item-Oriented Personalized RR with $\mathcal{D}_S = \{x_1, x_2, x_3\}, \mathcal{D}_N = \{x_4, x_5\}$, and $\mathcal{E} = \{\varepsilon_1, \varepsilon_2, \varepsilon_3\} = \{0.1, 0.5, 1.0\}$. For instance, x_1 = HIV, x_2 = Cancer, x_3 = Hepatitis, x_4 = Flu, and x_5 = None.

where $\bar{q}_x = [(e^{\varepsilon_x} - 1)(1 + \sum_{x \in \mathcal{D}_S}(e^{\varepsilon_x} - 1)^{-1})]^{-1}$ *and* $q_x = e^{\varepsilon_x}\bar{q}_x$.

In addition, the special case is that, when all the elements in $\mathcal{E}$ are identical, IPRR becomes KRR.

Theorem 1. *$(\mathcal{D}, \mathcal{E})$-IPRR satisfies $(\mathcal{D}, \mathcal{E})$-IPLDP.*

5.2 IPRR with Non-sensitive Items

We hereby present a full version of IPRR that incorporates the non-sensitive domain $\mathcal{D}_N$. For all $x \in \mathcal{D}_N$, privacy protection is not needed, which is equivalent to $\varepsilon_x \to \infty$. Thus, to maximize the ratio of q_x and $\bar{q}_x$, we set $\bar{q}_x$ to zero. Then, inspired by URR, we define IPRR with the non-sensitive domain as follows.

Definition 5 ($(\mathcal{D}_S, \mathcal{D}_N, \mathcal{E})$-IPRR). *Let $\mathcal{D}_S = \{x_1, \cdots, x_{|\mathcal{D}_S|}\}, \mathcal{E} = \{\varepsilon_1, \cdots, \varepsilon_{|\mathcal{D}_S|}\} \in \mathbb{R}_+^{|\mathcal{D}_S|}$, then $(\mathcal{D}_S, \mathcal{D}_N, \mathcal{E})$-IPRR is a mechanism that maps $x \in \mathcal{D}$ to $x_i \in \mathcal{D}$ with the probability $\boldsymbol{Q}_{\mathrm{IPRR}}(x|y)$ defined by:*

$$\boldsymbol{Q}_{\mathrm{IPRR}}(x|x') = \begin{cases} q_x & \text{if } x \in \mathcal{D}_S \wedge x = x', \\ \bar{q}_x & \text{if } x \in \mathcal{D}_S \wedge x \neq x', \\ \tilde{q} & \text{if } x \in \mathcal{D}_N \wedge x = x', \\ 0 & \text{otherwise}, \end{cases} \tag{11}$$

where $\tilde{q} = 1 - \sum_{x \in \mathcal{D}_S} \bar{q}_x = (1 + \sum_{x \in \mathcal{D}_S} \frac{1}{e^{\varepsilon_x} - 1})^{-1}$.

In addition, the special case is that, when all the elements in $\mathcal{E}$ are identical (denoted as ε), it is equivalent to the $(\mathcal{D}_S, \varepsilon)$-URR.

Theorem 2. *$(\mathcal{D}_S, \mathcal{D}_N, \mathcal{E})$-IPRR satisfies $(\mathcal{D}_S, \mathcal{E})$-IPLDP.*

Fig. 1 depicts an example of the $(\mathcal{D}_S, \mathcal{D}_N, \mathcal{E})$-IPRR, which illustrates the perturbation of our IPRR, and also shows detailed values of all involved probabilities under $\mathcal{E} = \{0.1, 0.5, 1.0\}$. As shown in Fig. 1, as the privacy budget decreases, users with sensitive items are more honest, which is different from the perturbation style of mainstream RR mechanisms. In those mechanisms, the probability of an honest answer will decrease as the privacy budget decreases. However, data utility is hard to be improved if we follow the mainstream methods to achieve independent personalized protection. Inspired by Mangat's RR

[26], data utility can be further improved through a different style of RR while guaranteeing the privacy as long as we obey the definition of LDP. Mangat's RR requires users in the sensitive group always answer honestly and uses dishonest answers from other users to contribute to the perturbation. Then, the data collector still can not distinguish one response whether it is an honest answer or not. Furthermore, in practical scenerios, the sensitivity of data shows an inverse relationship with the population size of respective individuals. As a result, indistinguishability can be guaranteed by a large proportion of dishonest responses from less sensitive or non-sensitive groups, even if individuals in the sensitive group are honest. Therefore, in our privacy scheme, we can guarantees the privacy of the clients with improvement of utility as long as both server and clients reach an agreement on this protocol.

5.3 Empirical Estimation under IPRR

In this subsection, we show the details of the emprical estimate of $\boldsymbol{p}$ under our $(\mathcal{D}_S, \mathcal{D}_N, \mathcal{E})$-IPRR mechanism. To calculate the estimate, we define a vector $\boldsymbol{r}$ and a function $\mathcal{S}$ for convenience, which are given by

$$\boldsymbol{r}_x = \begin{cases} \frac{1}{e^{\varepsilon_x}-1} & \text{if } x \in \mathcal{D}_S \\ 0 & \text{if } x \in \mathcal{D}_N \end{cases}, \text{ and } \mathcal{S}_\cdot = \frac{1}{\sum_{x \in \cdot} \boldsymbol{r}_x + 1},$$

where $\boldsymbol{r}_x$ is the corresponding element of $\boldsymbol{r}$ to $x \in \mathcal{D}$, and "$\cdot$" can be any domains, i.e., $\mathcal{S}_{\mathcal{D}_S} = [\sum_{x \in \mathcal{D}_S} \boldsymbol{r}_x + 1]^{-1}$. Then, for all $x \in \mathcal{D}_S$, we have $\bar{q}_x = \frac{1}{(e^{\varepsilon_x}-1)} \cdot \frac{1}{\sum_{i=1}^{|\mathcal{E}|}(e^{\varepsilon_i}-1)^{-1}+1} = \boldsymbol{r}_x \mathcal{S}_{\mathcal{D}_S}$, and, based on (7), we can calculate $\boldsymbol{m}$ and the estimate $\hat{\boldsymbol{p}}$ with each element $\boldsymbol{m}_x$ and $\hat{\boldsymbol{p}}_x$ for all $x \in \mathcal{D}$ as

$$\boldsymbol{m}_x = \boldsymbol{p}_x \mathcal{S}_{\mathcal{D}_S} + \boldsymbol{r}_x \mathcal{S}_{\mathcal{D}_S} \Rightarrow \hat{\boldsymbol{p}}_x = \hat{\boldsymbol{m}}_x / \mathcal{S}_{\mathcal{D}_S} - \boldsymbol{r}_x. \tag{12}$$

As the sample count n increases, $\hat{\boldsymbol{m}}$ remains unbiased for $\boldsymbol{m}$, and hence $\hat{\boldsymbol{p}}$ converges to $\boldsymbol{p}$ as well.

6 Utility Analysis

In this section, we first evaluate the data utility of IPRR based on the l_2 and l_1 losses of the empirical estimate $\hat{\boldsymbol{p}}$. Then, for each loss, we calculate its tight upper bound independent to the unknown distribution $\boldsymbol{p}$. Finally, we discuss the upper bound in both high and low privacy regimes.

First, we evaluate the expectation of l_2 and l_1 losses under our IPRR mechanism.

Theorem 3. *(l_2 and l_1 losses of $(\mathcal{D}_S, \mathcal{D}_N, \mathcal{E})$-IPRR) According to the Definition 5 and the empirical estimator given in* (12)*, for all $\mathcal{E}$, the expected l_2 and l_1 losses of the $(\mathcal{D}_S, \mathcal{D}_N, \mathcal{E})$-IPRR are given by*

$$\mathbb{E}[l_2(\boldsymbol{p}, \hat{\boldsymbol{p}})] = \mathbb{E}\left[\sum_{x \in \mathcal{D}} (\hat{\boldsymbol{p}}_x - \boldsymbol{p}_x)^2\right] = \frac{1}{n} \sum_{x \in \mathcal{D}} \left[(\boldsymbol{p}_x + \boldsymbol{r}_x)\left(\mathcal{S}_{\mathcal{D}_S}^{-1} - (\boldsymbol{p}_x + \boldsymbol{r}_x)\right)\right], \tag{13}$$

and for large n,

$$\mathbb{E}[l_1(\boldsymbol{p}, \hat{\boldsymbol{p}})] = \mathbb{E}\left[\sum_{x\in\mathcal{D}} |\hat{\boldsymbol{p}}_x - \boldsymbol{p}_x|\right] \approx \sqrt{\frac{2}{n\pi}} \sum_{x\in\mathcal{D}} \sqrt{(\boldsymbol{p}_x + \boldsymbol{r}_x)(\mathcal{S}_{\mathcal{D}_S}^{-1} - (\boldsymbol{p}_x + \boldsymbol{r}_x))}, \tag{14}$$

where $a_n \approx b_n$ represents $\lim_{n\to\infty} a_n/b_n = 1$.

According to (13) and (14), we can see that the two losses share the similar structure. Hence, to conveniently discuss the property of the losses, we define a general loss L as follows.

Definition 6 (general loss of $(\mathcal{D}_S, \mathcal{D}_N, \mathcal{E})$-IPRR). *The general loss L of $(\mathcal{D}_S, \mathcal{D}_N, \mathcal{E})$-IPRR is can be defined as*

$$L(\mathcal{E}; \boldsymbol{p}, \hat{\boldsymbol{p}}, \mathcal{D}) = C \sum_{x\in\mathcal{D}} g \circ f_x, \tag{15}$$

where $f_x = (\boldsymbol{p}_x + \boldsymbol{r}_x)\left(\mathcal{S}_{\mathcal{D}_S}^{-1} - (\boldsymbol{p}_x + \boldsymbol{r}_x)\right)$, g is any monotonically increasing concave function with $g(0) = 0$, and C is a non-negative constant.

With this definition, we show that, for any distribution $\boldsymbol{p}$, privacy budget set $\mathcal{E}$, $\mathcal{D}_S$, and $\mathcal{D}_N$, both $l2$ and $l1$ losses of $(\mathcal{D}_S, \mathcal{D}_N, \mathcal{E})$-IPRR are lower than that of $(\mathcal{D}_S, \min\{\mathcal{E}\})$-URR.

Through the assignment of fine-grained privacy budgets to each items in IPRR, the emprical estimator in Section 5.3 is a general version for mechanisms which use the direct encoding method. Due to the generality of our estimator, we can use it to calculate the empirical estimate of URR or KRR as long as all the privacy budgets are identical in $\mathcal{E}$. Therefore, based on the general empirical estimator, (13) and (14) are also applicable to these two mechanisms, even other mechanisms that use direct encoding method. To show that the losses of IPRR are lower than that of URR, we first give a lemma below.

Lemma 1 *Let $\tilde{\cdot}$ be a sorted version of any given set $\cdot$. For any two privacy budget sets $\mathcal{E}_1$ and $\mathcal{E}_2$ with same dimension k, $L(\mathcal{E}_1; \boldsymbol{p}, \hat{\boldsymbol{p}}, \mathcal{D}) \leqslant L(\mathcal{E}_2; \boldsymbol{p}, \hat{\boldsymbol{p}}, \mathcal{D})$, if $\mathcal{E}_1 \leqslant \mathcal{E}_2$, where $A \leqslant B$ stands for that, for all $a_i \in \tilde{A}$ and $b_i \in \tilde{B} (i = 1, \cdots k)$, we have $a_i \leqslant b_i$.*

Based on Lemma 1, for any distribution $\boldsymbol{p}$, privacy budget set $\mathcal{E}$, $\mathcal{D}_S$, and $\mathcal{D}_N$, since $\mathcal{E} \geqslant \{\min\{\mathcal{E}\}\} \in \mathbb{R}_+^{|\mathcal{E}|}$, the general loss L of $(\mathcal{D}_S, \mathcal{D}_N, \mathcal{E})$-IPRR are lower than that of $(\mathcal{D}_S, \min\{\mathcal{E}\})$-URR. Since l_2 and l_1 losses are specific versions of L when $g(x) = x$ and $g(x) = \sqrt{x}, respectively$, both two losses of IPRR also lower than that of URR in the same setting.

Next, we evaluate the worst case of the loss L. Observe that L is closely related to the original distribution $\boldsymbol{p}$. However, since $\boldsymbol{p}$ is unknown for theoretical analysis, we need to calculate a tight upper bound of the loss that does not depend on the unknown $\boldsymbol{p}$. Then, to obtain the tight upper bound, we need

to find an optimal $\boldsymbol{p}$ that maximizes L. To address this issue, we convert this problem to an optimization problem subject to $\boldsymbol{p}$ being a probability simplex as

$$\max_{\boldsymbol{p}} \sum_{x\in\mathcal{D}} g\circ f_x \qquad \text{s.t. } \boldsymbol{p}\in\mathbb{S}^{|\mathcal{D}|}. \tag{16}$$

Then, the optimal solution can be given as the following lemma.

Lemma 2 *Let $\mathcal{D}^*$ be a subset of $\mathcal{D}$. For all $x\in\mathcal{D}$, if $\boldsymbol{r}_x$ satisfies*

$$\begin{cases} (|\mathcal{D}^*|\mathcal{S}_{\mathcal{D}^*})^{-1}-1<\boldsymbol{r}_x<(|\mathcal{D}^*|\mathcal{S}_{\mathcal{D}^*})^{-1} & \text{if } x\in\mathcal{D}^*,\\ \boldsymbol{r}_x\geqslant(|\mathcal{D}^*|\mathcal{S}_{\mathcal{D}^*})^{-1} & \text{otherwise,}\end{cases} \tag{17}$$

$\boldsymbol{p}^*$ is the optimal solution that maximizes the objective function in (16)*, which is given by*

$$\boldsymbol{p}_x^*=\begin{cases}(|\mathcal{D}^*|\mathcal{S}_{\mathcal{D}^*})^{-1}-\boldsymbol{r}_x & \text{if } x\in\mathcal{D}^*,\\ 0 & \text{otherwise.}\end{cases} \tag{18}$$

According to Lemma 2, we can obtain the following general upper bounds of (13) and (14) as

Theorem 4 (General upper bound of l_2 and l_1 losses of IPRR). *Eq.* (13) *and* (14) *can be maximized by $\boldsymbol{p}^*$:*

$$\mathbb{E}[l_2(\boldsymbol{p},\hat{\boldsymbol{p}})]\leqslant\mathbb{E}[l_2(\boldsymbol{p}^*,\hat{\boldsymbol{p}})]=\frac{1}{n}\left(\frac{1}{\mathcal{S}_{\mathcal{D}}^2}-\frac{1}{|\mathcal{D}^*|\mathcal{S}_{\mathcal{D}^*}^2}-\sum_{x\in\mathcal{D}\setminus\mathcal{D}^*}\boldsymbol{r}_x^2\right); \tag{19}$$

$$\begin{aligned}\mathbb{E}[l_1(\boldsymbol{p},\hat{\boldsymbol{p}})]&\lesssim\mathbb{E}[l_1(\boldsymbol{p}^*,\hat{\boldsymbol{p}})]\\ &=\sqrt{\frac{2}{n\pi}}\left[\sum_{x\in\mathcal{D}\setminus\mathcal{D}^*}\sqrt{\boldsymbol{r}_x(\mathcal{S}_{\mathcal{D}}^{-1}-\boldsymbol{r}_x)}+\sqrt{\mathcal{S}_{\mathcal{D}^*}^{-1}(|\mathcal{D}^*|\mathcal{S}_{\mathcal{D}}^{-1}-\mathcal{S}_{\mathcal{D}^*}^{-1})}\right],\end{aligned} \tag{20}$$

where $a_n\lesssim b_n$ represents $\lim_{n\to\infty}a_n/b_n\leqslant 1$.

Finally, we discuss the losses in the high and low privacy regimes based on the general upper bounds. Let $\varepsilon_{\min}=\min\{\mathcal{E}\}$ and $\varepsilon_{\max}=\max\{\mathcal{E}\}$.

Theorem 5 (l_2 and l_1 losses in high privacy regime). *When $\varepsilon_{\max}$ is close to 0, for all $x\in\mathcal{D}$, we have $e^{\varepsilon_x}-1\approx\varepsilon_x$. Then, the worst case of l_2 and l_1 losses are:*

$$\mathbb{E}[l_2(\boldsymbol{p},\hat{\boldsymbol{p}})]\leqslant\mathbb{E}[l_2(\boldsymbol{p}^*,\hat{\boldsymbol{p}})]\approx\frac{1}{n}\sum_{x\in\mathcal{D}_S}\sum_{x'\in\mathcal{D}_S\setminus\{x\}}\frac{1}{\varepsilon_x\varepsilon_{x'}}; \tag{21}$$

$$\mathbb{E}[l_1(\boldsymbol{p},\hat{\boldsymbol{p}})]\lesssim\mathbb{E}[l_1(\boldsymbol{p}^*,\hat{\boldsymbol{p}})]\approx\sqrt{\frac{2|\mathcal{D}_S|}{n\pi}\sum_{x\in\mathcal{D}_S}\sum_{x'\in\mathcal{D}_S\setminus\{x\}}\frac{1}{\varepsilon_x\varepsilon_{x'}}}. \tag{22}$$

Table 1: Synthetic and Real-world Datasets

Datasets	# Users	# Items
Zipf	100000	20
Kosarak [1]	646510	100

Table 2: Parameter Settings

#	Mechanisms to Compare	$\varepsilon_{\min}$	$\varepsilon_{\max}$	LC	NR	SR
1	KRR,URR	0.1	1	4	0.5	0.2-1.0
2	KRR,URR	1	10	4	0.5	0.2-1.0
3	KRR,URR	0.1	10	4	0.5	0.2-1.0
4	URR	0.1	10	4	0.0-0.8 * 0.1-0.9 **	1.0
5	IDUE	0.1	10	4	0	0.1-1.0

* is used on Zipf dataset and ** is used on Kosarak dataset

According to [28], in high privacy regime, the expectation of l_2 and l_1 losses of $(\mathcal{D}_S, \varepsilon_{\min})$-URR are $\frac{|\mathcal{D}_S|(|\mathcal{D}_S|-1)}{n\varepsilon_{\min}^2}$ and $\sqrt{\frac{2}{n\pi}} \cdot \frac{|\mathcal{D}_S|\sqrt{|\mathcal{D}_S|-1}}{\varepsilon_{\min}}$, accordingly. Thus, the losses of our method is much smaller than that of URR in current setting.

Theorem 6 **(l_2 and l_1 losses in low privacy regime).** *When $\varepsilon_{\min} > \ln(|\mathcal{D}_N| + 1)$, for all $x \in \mathcal{D}$, the worst case of l_2 and l_1 losses are:*

$$\mathbb{E}[l_2(\boldsymbol{p}, \hat{\boldsymbol{p}})] \leqslant \mathbb{E}[l_2(\boldsymbol{p}^*, \hat{\boldsymbol{p}})] = \left[\sum_{x \in \mathcal{D}_S} \frac{|\mathcal{D}_S| + e^{\varepsilon_x} - 1}{|\mathcal{D}_S|(e^{\varepsilon_x} - 1)}\right]^2 \left(1 - \frac{1}{|\mathcal{D}|}\right); \tag{23}$$

$$\mathbb{E}[l_1(\boldsymbol{p}, \hat{\boldsymbol{p}})] \lesssim \mathbb{E}[l_1(\boldsymbol{p}^*, \hat{\boldsymbol{p}})] = \sqrt{\frac{2(|\mathcal{D}| - 1)}{n\pi}} \cdot \sum_{x \in \mathcal{D}_S} \frac{|\mathcal{D}_S| + e^{\varepsilon_x} - 1}{|\mathcal{D}_S|(e^{\varepsilon_x} - 1)}. \tag{24}$$

According to [28], in low privacy regime, the expectation of l_2 and l_1 losses of $(\mathcal{D}_S, \varepsilon_{\min})$-URR are $\frac{(|\mathcal{D}_S| + e^{\varepsilon_{\min}} - 1)^2}{n(e^{\varepsilon_{\min}} - 1)^2}\left(1 - \frac{1}{|\mathcal{D}|}\right)$ and $\sqrt{\frac{2(|\mathcal{D}|-1)}{n\pi}} \cdot \frac{|\mathcal{D}_S| + e^{\varepsilon_{\min}} - 1}{e^{\varepsilon_{\min}} - 1}$, accordingly. Thus, the losses of our method is much smaller than that of URR in current setting.

7 Evaluation

In this section, we evaluate the performance of our IPRR based on the emprical estimation method with the Norm-sub (NS) truncation method and maximum likelihood estimation (MLE) method, and compare it with the the KRR, URR, and Input-Discriminative Unary Encoding (IDUE) [15] satisfying ID-LDP.

7.1 Experimental Setup

Datasets We conducted experiments over two datasets, and show their details in Table 1. The first dataset, Zipf, was generated by sampling from a Zipf distribution with an exponential parameter $\alpha = 2$, followed by filtering the results using a specific threshold to control the size of the item domain and the number of users. The second dataset, Kosarak, is one of the largest real-world datasets, which contains millions of records related to the click-stream of news portals from users (e.g. see [35, 9, 36]). For Kosarak dataset, we randomly selected an item for every user to serve as the item they hold, and then applied the same filtering process used for the Zipf dataset.

Metrics We use Mean Squar Error (MSE) and Relative Error (RE) as the metrics of the performance, which are defined as

$$\text{MSE} = \sum_{x \in \mathcal{D}} \frac{(f_x - \hat{f}_x)^2}{n}, \quad \text{RE} = \sum_{x \in \mathcal{D}} \frac{|f_x - \hat{f}_x|}{f_x}, \tag{25}$$

where f_x (resp. $\hat{f}_x$) is the true (resp. estimated) frequency count of x. We take the sample mean of one hundred repeated experiments for analysis.

Settings We conduct five experiments for both Zipf and Kosarak datasets, where the experiments #1∼#3 compare the utility of IPRR with that of KRR and URR under various privacy level groups with different sample ratios, the experiment #4 compares IPRR with URR under different $|\mathcal{D}_N|$, and the experiment #5 compares IPRR with IDUE under different sample ratios. We use SR (sample ratio) to calculate $SR \cdot |\mathcal{D}|$ as the sample count n. In each experiment, we evaluate the utility under various privacy levels. Since the sensitivity of data shows an inverse relationship with the population size of respective individuals, we first sort the dataset based on the count of each item, and items with smaller counts are assigned to higher privacy levels. Then, we choose items with larger size as $\mathcal{D}_N$ (others as $\mathcal{D}_S$) by using NR (non-sensitive ratio), which controls the ratio of $|\mathcal{D}_N|$ over $|\mathcal{D}|$. For $\mathcal{D}_S$, we use $\varepsilon_{\min}$, $\varepsilon_{\max}$, LC (level count) to divide $\mathcal{D}_S$ into different privacy levels, where LC divides $\mathcal{D}_S$ and a range $[\varepsilon_{\min}, \varepsilon_{\max}]$ evenly to assign the privacy level, accordingly. For example, assume we have $\mathcal{D} = \{A, B, C, D, E, F\}$ with sizes 6, 5, 4, 3, 2, and 1 for each item. Under $\varepsilon_{\min} = 0.1$, $\varepsilon_{\max} = 0.3$, $LC = 3$, and $NR = 0.5$, we can obtain $\mathcal{D}_N = \{A, B, C\}$, with privacy budgets of 0.3, 0.2, and 0.1 assigned to items D, E, and F, respectively. In practical scenarios, it is unnecessary to assign a unique privacy level to every item. Thus, we set $LC = 4$ for all experiments in our settings. Additionally, in all experiments, KRR and URR satisfy $\varepsilon_{\min}$-LDP and $(\mathcal{D}_S, \mathcal{D}_S, \varepsilon_{\min})$-ULDP, respectively. Table 2 shows the details of the parameter settings for all experiments.

7.2 Experimental Results

Utility under various privacy level groups In Fig. 2 we illustrate the results of the experiments #1∼#3. We conducted these experiments to compare the utility of our IPRR with other types of direct encoding mechanisms under various combinations of privacy budgets with fixed NR. Fristly, Fig. 2(a) and (d) show the comparison of the utility in a high privacy regime among IPRR, URR and KRR on both datasets. As we can see in the high privacy regime, our method outperforms the others by approximate one order of magnitude. As the sample count n increases, the loss decreases as well. Noticeably, in the figure, the results of KRR under both the NS and MLE methods are almost indistinguishable, while our results of the latter method are improved significantly compared with that of the former method. The reason is that the former method may truncate the empirical estimate for lacking samples. Furthermore, the high privacy regime will also escalate the degree of truncation for the emprical estimate, which naturally

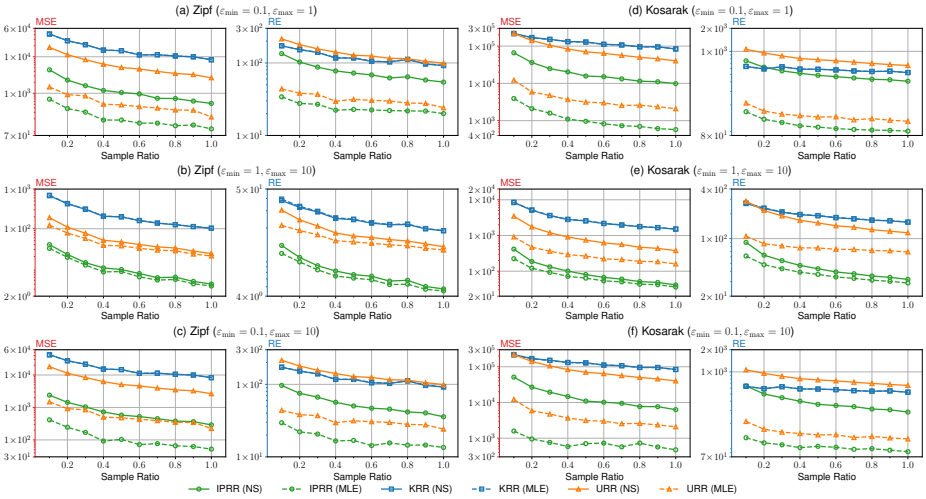

Fig. 2: Utility under Various Privacy Levels.

introduces more errors than the latter method. Secondly, the results of Fig. 2(b) and (e) present the comparison of the utility in a low privacy regime. It is clear that our method also has better performance than the others. Notably, in the current setting, the improvement of the MLE method over the NS method is less significant compared with that in the high privacy regime. We argue that a large privacy budget does not result in much truncation for the emprical estimate, so the results are close to each other. Finally, Fig. 2(c) and (f) give the results of the hybrid high and low privacy regimes. The results are close to Fig. 2(a) and (d), accordingly. Although the improvement is limited in this setting, it does not mean that all perturbations in our method are greatly influenced by the minimum privacy budget similar to IDUE.

Fig.3 shows the results of the experiment #4. In this experiment, we compare the utility of our IPRR with URR in the same privacy budget set to check the influence of different $|\mathcal{D}_N|$. We only compare with URR because only these two mechanisms support non-sensitive items. We restricted the maximum value of NR to 80% for the Zipf dataset since $|\mathcal{D}_S|$ will be less than $LC = 4$ if NR exceeds 0.8. As one can see, as $|\mathcal{D}_N|$ increases, our method outperforms URR, and all metrics decrease.

After all, our IPRR shows better performance than URR and KRR on the two corresponding metrics in the two datasets, which verifies our theoretical analysis. Compared with KRR, URR reduces the perturbation for non-sensitive items by coarsely dividing the domain into sensitive and non-sensitive subsets, resulting in lower variance than that of KRR under the identical sample ratio. Thus, URR has better overall performance than KRR. Our method further divides the sensitive domain into finer-grained subsets with personalized perturbation for each item, which reduces the perturbation for less sensitive items. Therefore,

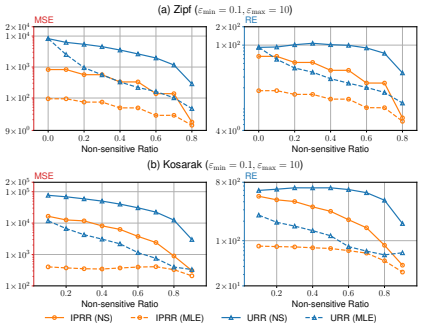

Fig. 3: Unility under Different NR.

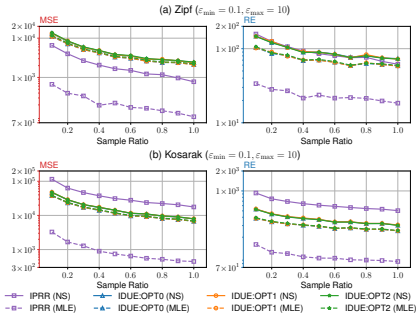

Fig. 4: Comparison between the IPRR and the IDUE (`opt0`, `opt1`, `opt2`).

with the same sample ratio, our method reduces much more total variance than URR, and thus our method performs better than other mechanisms.

Comparison with IDUE In Fig. 4, we show the results of the experiment #5. Since IDUE does not support non-sensitive items, we conducted this separate experiment to compare the utility of our method with IDUE in the same privacy setting. It is clear that IPRR owns better performance than IDUE over Zipf with both NS and MLE methods, while IDUE outperforms IPRR over Kosarak with the NS method. The reason is that the unary encoding method used by IDUE has more advantages when processing an item domain with a larger size, and this may reduce the truncation for the empirical estimate. However, under the MLE method without the influence of truncation, IPRR outperforms IDUE even over Kosarak with the larger $|\mathcal{D}|$. We think that our IPRR effectively reduces unnecessary perturbation for less sensitive items than IDUE since the strength of perturbation is highly affected by the minimum privacy budget. In the current setting of our experiment, the perturbation for items with the maximum privacy budget only needs to satisfy 10-LDP in our method, while IDUE can only relax their perturbation at most 0.2-LDP according to the Lemma 1 in [15].

8 Conclusion

In this paper, we first proposed a novel notion of LDP called IPLDP for discrete distribution estimation in local privacy setting. To improve utility, IPLDP perturbs items independently for personalized protection according to the outputs with different privacy budgets. Then, to satisfy IPLDP, we proposed a new mechanism called IPRR based on a common phenomenon that the sensitivity of data shows an inverse relationship with the population size of respective individuals. We prove that IPRR has tighter upper bound than that of existing direct encoding methods under both l_2 and l_1 losses of emprical estimate. Finally, we conducted related experiments on a synthetic and a real-world datasets. Both theoretical analysis and experimental results demonstrate that our scheme owns better performance than existing methods.

A Item-Oriented Personalized LDP

A.1 Proof of Theorem 1

Proof. Since $\mathcal{D}_N = \emptyset$, we only need to consider (3). Then, since $q_x/\bar{q}_x = e^{\varepsilon_x}$, the inequality (4) holds. □

A.2 Proof of Theorem 2

Proof. For all $x \in \mathcal{D}_S$, since $q_x/\bar{q}_x = e^{\varepsilon_x}$, the inequality (4) holds. Then, for all $x \in \mathcal{D}_N$, it follows from (3) that (4) also holds. □

B Utility Analysis

B.1 Proof of Theorem 3

Proof. 1. The l_2 loss of the estimate.

Since $n\hat{\boldsymbol{m}}_x$ follows the binomial distribution with parameters n and $\boldsymbol{m}_x$, its mean and variance are $\mathbb{E}(n\hat{\boldsymbol{m}}) = n\boldsymbol{m}_x$ and $\mathrm{Var}(n\hat{\boldsymbol{m}}) = n\boldsymbol{m}_x(1-\boldsymbol{m}_x)$. Then,

$$\begin{aligned}
\mathbb{E}_{Y^n\sim\boldsymbol{m}(\boldsymbol{Q})}[l_2(\hat{\boldsymbol{p}},\boldsymbol{p})] &= \mathbb{E}\left[\sum_{x\in\mathcal{D}}(\hat{\boldsymbol{p}}_x-\boldsymbol{p}_x)^2\right] \\
&= \sum_{x\in\mathcal{D}}\mathbb{E}\left[(\hat{\boldsymbol{p}}_x-\boldsymbol{p}_x)^2\right] \\
&= \sum_{x\in\mathcal{D}}\mathbb{E}\left[(\hat{\boldsymbol{m}}_x/\mathcal{S}_{\mathcal{D}_S}-\boldsymbol{m}_x/\mathcal{S}_{\mathcal{D}_S})^2\right] \\
&= \frac{1}{\mathcal{S}_{\mathcal{D}_S}^2}\sum_{x\in\mathcal{D}}[\mathbb{E}(\hat{\boldsymbol{m}}^2)-\boldsymbol{m}^2] \\
&= \frac{1}{\mathcal{S}_{\mathcal{D}_S}^2}\sum_{x\in\mathcal{D}}\frac{\boldsymbol{m}_x-\boldsymbol{m}_x^2}{n} \\
&= \frac{1}{n\mathcal{S}_{\mathcal{D}_S}^2}\sum_{x\in\mathcal{D}}\left[\boldsymbol{p}_x\mathcal{S}_{\mathcal{D}_S}+\boldsymbol{r}_x\mathcal{S}_{\mathcal{D}_S}-(\boldsymbol{p}_x\mathcal{S}_{\mathcal{D}_S}+\boldsymbol{r}_x\mathcal{S}_{\mathcal{D}_S})^2\right] \\
&= \frac{1}{n}\sum_{x\in\mathcal{D}}\left[(\boldsymbol{p}_x+\boldsymbol{r}_x)\left(\mathcal{S}_{\mathcal{D}_S}^{-1}-(\boldsymbol{p}_x+\boldsymbol{r}_x)\right)\right]
\end{aligned}$$

2. The l_1 loss of the estimate.

$$\begin{aligned}
\mathbb{E}_{Y^n\sim\boldsymbol{m}(\boldsymbol{Q})}[l_1(\hat{\boldsymbol{p}},\boldsymbol{p})] &= \mathbb{E}\left(\sum_{x\in\mathcal{D}}|\hat{\boldsymbol{p}}_x-\boldsymbol{p}_x|\right) \\
&= \frac{1}{\mathcal{S}_{\mathcal{D}_S}}\sum_{x\in\mathcal{D}}\mathbb{E}[|\hat{\boldsymbol{m}}-\boldsymbol{m}_x|] \\
&= \frac{1}{\mathcal{S}_{\mathcal{D}_S}}\sum_{x\in\mathcal{D}}\frac{\sqrt{\mathrm{Var}(n\hat{\boldsymbol{m}}_x)}}{n}\,\mathbb{E}\left[\left|\frac{n\hat{\boldsymbol{m}}_x-\mathbb{E}(n\hat{\boldsymbol{m}}_x)}{\sqrt{\mathrm{Var}(n\hat{\boldsymbol{m}}_x)}}\right|\right].
\end{aligned}$$

It follows from the central limit theorem that $\frac{n\hat{\boldsymbol{m}}_x - \mathbb{E}(n\hat{\boldsymbol{m}}_x)}{\sqrt{\mathrm{Var}(n\hat{\boldsymbol{m}}_x)}}$ converges to the normal distribution $\mathcal{N}(0,1)$. Hence,

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{Y^n\sim\boldsymbol{m}(\boldsymbol{Q})} \left[\left| \frac{n\hat{\boldsymbol{m}}_x - \mathbb{E}(n\hat{\boldsymbol{m}}_x)}{\sqrt{\mathrm{Var}(n\hat{\boldsymbol{m}}_x)}} \right|\right] = \sqrt{\frac{2}{n\pi}}$$

Therefore,

$$\begin{aligned}\mathbb{E}_{Y^n\sim\boldsymbol{m}(\boldsymbol{Q})}[l_1(\hat{\boldsymbol{p}},\boldsymbol{p})] &\approx \frac{1}{\mathcal{S}_{\mathcal{D}_S}}\sqrt{\frac{2}{n\pi}}\sum_{x\in\mathcal{D}}\sqrt{\boldsymbol{m}_x - \boldsymbol{m}_x^2} \\ &= \sqrt{\frac{2}{n\pi}}\sum_{x\in\mathcal{D}}\sqrt{(\boldsymbol{p}_x+\boldsymbol{r}_x)\left(\mathcal{S}_{\mathcal{D}_S}^{-1} - (\boldsymbol{p}_x+\boldsymbol{r}_x)\right)}\end{aligned}$$

□

B.2 Proof of Lemma 1

Proof.

$$f_x = (\boldsymbol{p}_x+\boldsymbol{r}_x)\left(\mathcal{S}_{\mathcal{D}_S}^{-1} - (\boldsymbol{p}_x+\boldsymbol{r}_x)\right) = (\boldsymbol{p}_x+\boldsymbol{r}_x)\left(\sum_{x\in\mathcal{D}\setminus\{x\}}\boldsymbol{r}_x + 1 - \boldsymbol{p}_x\right).$$

Then,

$$\begin{cases}\frac{\partial f_x}{\partial \boldsymbol{r}_x} = \sum_{x\in\mathcal{D}\setminus\{x\}}\boldsymbol{r}_x + 1 - \boldsymbol{p}_x > 0, \\ \frac{\partial f_x}{\partial \boldsymbol{r}_{x'}} = \boldsymbol{p}_x + \boldsymbol{r}_x > 0 \qquad\qquad \text{if } x'\in D\setminus\{x\}.\end{cases}$$

Apparently, f_x is a monotonically increasing function of $\boldsymbol{r}_x$ for all $x\in\mathcal{D}$. Then,

$$\frac{\partial L}{\partial \boldsymbol{r}_x} = Cg'\circ f_x\cdot\frac{\partial f_x}{\partial \boldsymbol{r}_x} + C\sum_{x'\in\mathcal{D}\setminus\{x\}} g'\circ f_{x'}\cdot\frac{\partial f_{x'}}{\partial \boldsymbol{r}_x} > 0$$

Therefore, L is a monotonically increasing function of $\boldsymbol{r}_x$ for all $x\in\mathcal{D}$. Moreover, due to the loss L is non-negative and $\boldsymbol{r}_x$ is a monotonically decreasing function of ε_x for $x\in\mathcal{D}$, the proposition holds. □

B.3 Proof of Lemma 2

Proof. To prove this lemma, we consider a more general optimization problem as

$$F(w) = \max_{\boldsymbol{\theta}}\sum_{i=1}^{K} g[(\theta_i+c_i)(C-(\theta_i+c_i))] \quad \text{s.t. } \boldsymbol{\theta}^\top\mathbf{1} = w,\ \mathbf{0}\leqslant\boldsymbol{\theta}\leqslant\mathbf{1}w,$$

where $\boldsymbol{\theta}$ and $\boldsymbol{c}$ are vectors with k-dimension, $\boldsymbol{c}$ is a constant vector, $w\in(0,1)$, and C is a large enough positive constant (e.g. $C\gg\boldsymbol{c}^\top\mathbf{1}+1$).

First, we find a proper constant vectors $\boldsymbol{c}$ to obtain the optimal $\boldsymbol{\theta}$ without zero elements. Since g is any monotonically increasing concave function with $g(0) = 0$, according to Jensen inequality, we have

$$\sum_{i=1}^{K} g[(\theta_i + c_i)(C - (\theta_i + c_i))] \leqslant g\left[K \sum_{i=1}^{K} [(\theta_i + c_i)(C - (\theta_i + c_i))]\right],$$

where the equality holds iff $\theta_1 + c_1 = \cdots = \theta_K + c_K$. Hence, to satisfy the equality condition, we have

$$\sum_{i=1}^{K} (\theta_i + c_i) = w + \sum_{i=1}^{K} c_i \Rightarrow \theta_i = \frac{w + \sum_{j=1}^{K} c_j}{|k|} - c_i.$$

Then, since $0 < \theta_i < w$,

$$0 < \frac{w + \sum_{j=1}^{K} c_j}{K} - c_i < w.$$

Therefore, if all elements of $\boldsymbol{c}$ satisfy

$$\frac{w + \sum_{j=1}^{K} c_j}{K} - w < c_i < \frac{w + \sum_{j=1}^{K} c_j}{K},$$

we can ensure that the optimal $\boldsymbol{\theta}$ has no zero elements, and the maximum value is

$$\begin{aligned} F(w) &= g\left[K \sum_{i=1}^{K} \left[\frac{w + \sum_{j=1}^{K} c_j}{K} \left(C - \frac{w + \sum_{j=1}^{K} c_j}{K}\right)\right]\right] \\ &= g\left[K^2 \cdot \frac{w + \sum_{i=1}^{K} c_i}{K} \left(C - \frac{w + \sum_{i=1}^{K} c_i}{K}\right)\right] \\ &= g\left[\left(w + \sum_{i=1}^{K} c_i\right)\left(KC - \left(w + \sum_{i=1}^{K} c_i\right)\right)\right] \end{aligned}$$

Next, we consider the general case, where the optimal $\boldsymbol{\theta}$ contains zero elements. Let

$$\tilde{F}(w) = \sum_{t=1}^{T} \tilde{F}_t(w) = \sum_{t=1}^{T} g[((1-w)\phi_t + c_t)(C - ((1-w)\phi_t + c_t))],$$

where c_t is a constant which satisfies $c_t \geqslant \frac{w + \sum_{i=1}^{K} c_i}{K}$, and ϕ_t is a constant which satisfies $\sum_{t=1}^{T} \phi_t = 1$ and $0 < \phi_t < 1$.

Let $H(w) = F(w) + \tilde{F}(w)$. Then,

$$\begin{aligned}
H' &= F' \cdot \left[KC - \left(w + \sum_{i=1}^{K} c_i \right) - \left(w + \sum_{i=1}^{K} c_i \right) \right] \\
&\quad + \sum_{t=1}^{T} \phi_t \tilde{F}_t' \cdot [((1-w)\phi_t + c_t) - (C - ((1-w)\phi_t + c_t))] \\
&= \frac{KC - 2\left(w + \sum_{i=1}^{K} c_i \right)}{1/F'} + \sum_{t=1}^{T} \frac{2[(1-w)\phi_t + c_t] - C}{1/\left(\phi_t \tilde{F}_t' \right)} \\
&= \sum_{t=1}^{T} \left[\frac{C - 2\frac{w + \sum_{i=1}^{K} c_i}{K}}{T/(KF')} + \frac{2[(1-w)\phi_t + c_t] - C}{1/\left(\phi_t \tilde{F}_t' \right)} \right] \\
&\geqslant \sum_{t=1}^{T} \frac{C - 2\frac{w + \sum_{i=1}^{K} c_i}{K} + 2[(1-w)\phi_t + c_t] - C}{\max\left[T/(KF'), 1/\left(\phi_t \tilde{F}_t' \right) \right]} \\
&= 2\sum_{t=1}^{T} \frac{(1-w)\phi_t + c_t - \frac{w + \sum_{i=1}^{K} c_i}{K}}{\max\left[T/(KF'), 1/\left(\phi_t \tilde{F}_t' \right) \right]} \\
&\geqslant 0,
\end{aligned}$$

where $F' = g'\left[\left(w + \sum_{i=1}^{K} c_i \right) \left(KC - \left(w + \sum_{i=1}^{K} c_i \right) \right) \right]$, and $\tilde{F}_t' = g'[((1-w)\phi_t + c_t)(C - ((1-w)\phi_t + c_t))]$ Since $H' \geqslant 0$, $H(w)$ reaches its maximum when $w = 1$.

Finally, because any general case of the optimization problem in this lemma can be convert to the function H, the lemma holds. □

B.4 Proof of Theorem 4

Proof. For save the pages, we won't give the details of the proof, since we just get the result by substituting $\boldsymbol{p}^*$ of Lemma 2 into (13) and (14). □

B.5 Proof of Theorem 5

Proof. As $\varepsilon_{\max}$ is close to 0, $\mathcal{D}^*$ will become $\mathcal{D}_N$. Hence, according to Lemma 2, when $\mathcal{D}^* = \mathcal{D}_N$, $\min\{\boldsymbol{r}\} = \frac{1}{e^{\varepsilon_{\max}} - 1} > \frac{1}{|\mathcal{D}_N|} \Rightarrow 0 < \varepsilon_{\max} < \ln(|\mathcal{D}_N| + 1)$. In this case, $\boldsymbol{p}^* = \boldsymbol{p}^{\mathrm{u}}$, where

$$\boldsymbol{p}_x^{\mathrm{u}} = \begin{cases} \frac{1}{|\mathcal{D}_N|} & \text{if } x \in \mathcal{D}_N, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{Y^n\sim\boldsymbol{m}(\boldsymbol{Q})}[l_2(\hat{\boldsymbol{p}},\boldsymbol{p})] &\leqslant \mathbb{E}_{Y^n\sim\boldsymbol{m}(\boldsymbol{Q})}[l_2(\hat{\boldsymbol{p}},\boldsymbol{p}^{\mathrm{u}})] \\
&= \frac{1}{n}\left[\sum_{x\in\mathcal{D}_S}\left[\boldsymbol{r}_x\left(\mathcal{S}_{\mathcal{D}_S}^{-1}-\boldsymbol{r}_x\right)\right]+\sum_{x\in\mathcal{D}_N}\left[\frac{1}{|\mathcal{D}_N|}\left(\mathcal{S}_{\mathcal{D}_S}^{-1}-\frac{1}{|\mathcal{D}_N|}\right)\right]\right] \\
&= \frac{1}{n}\left[\sum_{x\in\mathcal{D}_S}\left[\boldsymbol{r}_x\left(\sum_{x'\in\mathcal{D}}\boldsymbol{r}_{x'}+1-\boldsymbol{r}_x\right)\right]+\sum_{x\in\mathcal{D}}\boldsymbol{r}_x+1-\frac{1}{|\mathcal{D}_N|}\right] \\
&\approx \frac{1}{n}\left[\sum_{x\in\mathcal{D}_S}\left[\frac{1}{\varepsilon_x}\left(\sum_{x'\in\mathcal{D}}\frac{1}{\varepsilon_{x'}}+1-\frac{1}{\varepsilon_x}\right)\right]+\sum_{x\in\mathcal{D}}\frac{1}{\varepsilon_x}+1-\frac{1}{|\mathcal{D}_N|}\right] \\
&= \frac{1}{n}\left[\left(\sum_{x\in\mathcal{D}_S}\frac{1}{\varepsilon_x}\right)^2-\sum_{x\in\mathcal{D}_S}\frac{1}{\varepsilon_x^2}+\sum_{x\in\mathcal{D}}\frac{2}{\varepsilon_x}+1-\frac{1}{|\mathcal{D}_N|}\right] \\
&\approx \frac{1}{n}\left[\left(\sum_{x\in\mathcal{D}_S}\frac{1}{\varepsilon_x}\right)^2-\sum_{x\in\mathcal{D}_S}\frac{1}{\varepsilon_x^2}\right] \\
&= \frac{1}{n}\sum_{x\in\mathcal{D}_S}\sum_{x'\in\mathcal{D}_S\setminus\{x\}}\frac{1}{\varepsilon_x\varepsilon_{x'}},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{Y^n\sim\boldsymbol{Q}}[l_1(\hat{\boldsymbol{p}},\boldsymbol{p})] &\leqslant \mathbb{E}_{Y^n\sim\boldsymbol{Q}}\left[l_1\left(\hat{\boldsymbol{p}},\boldsymbol{p}^{\mathrm{u}}\right)\right] \\
&= \sqrt{\frac{2}{n\pi}}\left[\sum_{x\in\mathcal{D}_S}\sqrt{\boldsymbol{r}_x\left(\mathcal{S}_{\mathcal{D}_S}^{-1}-\boldsymbol{r}_x\right)}+\sqrt{|\mathcal{D}_N|\mathcal{S}_{\mathcal{D}_S}^{-1}-1}\right] \\
&\leqslant \sqrt{\frac{2}{n\pi}}\left[\sqrt{|\mathcal{D}_S|\sum_{x\in\mathcal{D}_S}\boldsymbol{r}_x\left(\mathcal{S}_{\mathcal{D}_S}^{-1}-\boldsymbol{r}_x\right)}+\sqrt{|\mathcal{D}_N|\mathcal{S}_{\mathcal{D}_S}^{-1}-1}\right] \\
&\approx \sqrt{\frac{2}{n\pi}}\left[\sqrt{|\mathcal{D}_S|\sum_{x\in\mathcal{D}_S}\left[\frac{1}{\varepsilon_x}\left(\sum_{x'\in\mathcal{D}}\frac{1}{\varepsilon_{x'}}+1-\frac{1}{\varepsilon_x}\right)\right]}\right. \\
&\qquad \left.+\sqrt{|\mathcal{D}_N|\sum_{x\in\mathcal{D}}\frac{1}{\varepsilon_x}+|\mathcal{D}_N|-1}\right] \\
&\approx \sqrt{\frac{2}{n\pi}}\sqrt{|\mathcal{D}_S|\sum_{x\in\mathcal{D}_S}\left[\frac{1}{\varepsilon_x}\left(\sum_{x'\in\mathcal{D}}\frac{1}{\varepsilon_{x'}}+1-\frac{1}{\varepsilon_x}\right)\right]} \\
&= \sqrt{\frac{2}{n\pi}}\sqrt{|\mathcal{D}_S|\left[\left(\sum_{x\in\mathcal{D}_S}\frac{1}{\varepsilon_x}\right)^2+\sum_{x\in\mathcal{D}_S}\frac{1}{\varepsilon_x}-\sum_{x\in\mathcal{D}_S}\frac{1}{\varepsilon_x^2}\right]}
\end{aligned}$$

$$\approx \sqrt{\frac{2}{n\pi}} \sqrt{|\mathcal{D}_S| \left[\left(\sum_{x \in \mathcal{D}_S} \frac{1}{\varepsilon_x} \right)^2 - \sum_{x \in \mathcal{D}_S} \frac{1}{\varepsilon_x^2} \right]}$$
$$= \sqrt{\frac{2}{n\pi}} \sqrt{|\mathcal{D}_S| \sum_{x \in \mathcal{D}_S} \sum_{x' \in \mathcal{D}_S \setminus \{x\}} \frac{1}{\varepsilon_x \varepsilon_{x'}}}$$

□

B.6 Proof of Theorem 6

Proof. According to Lemma 2, when $\varepsilon_{\min} > \ln(|\mathcal{D}_N| + 1)$, $\boldsymbol{m}$ is a uniform distribution. In this case, $\mathcal{D}^* = \mathcal{D}$. Therefore,

$$\begin{aligned}
\mathbb{E}_{Y^n \sim \boldsymbol{m}(\boldsymbol{Q})}[l_2(\hat{\boldsymbol{p}}, \boldsymbol{p})] &\leqslant \mathbb{E}_{Y^n \sim \boldsymbol{m}(\boldsymbol{Q})}[l_2(\hat{\boldsymbol{p}}, \boldsymbol{p}^*)] \\
&= \frac{1}{n} \sum_{x \in \mathcal{D}} \left[(\mathcal{S}_{\mathcal{D}} |\mathcal{D}|)^{-1} \left(\mathcal{S}_{\mathcal{D}_S}^{-1} - (\mathcal{S}_{\mathcal{D}} |\mathcal{D}|)^{-1} \right) \right] \\
&= \frac{1}{n} |\mathcal{D}| \left[(\mathcal{S}_{\mathcal{D}} |\mathcal{D}|)^{-1} \left(\mathcal{S}_{\mathcal{D}_S}^{-1} - (\mathcal{S}_{\mathcal{D}} |\mathcal{D}|)^{-1} \right) \right] \\
&= \frac{1}{n} \left[\mathcal{S}_{\mathcal{D}}^{-1} \left(\mathcal{S}_{\mathcal{D}_S}^{-1} - (\mathcal{S}_{\mathcal{D}} |\mathcal{D}|)^{-1} \right) \right] \\
&= \frac{1}{n \mathcal{S}_{\mathcal{D}}} \left(\mathcal{S}_{\mathcal{D}_S}^{-1} - (\mathcal{S}_{\mathcal{D}} |\mathcal{D}|)^{-1} \right) \\
&= \frac{1}{n \mathcal{S}_{\mathcal{D}_S}^2} \left(1 - \frac{1}{|\mathcal{D}|} \right),
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{Y^n \sim \boldsymbol{Q}}[l_1(\hat{\boldsymbol{p}}, \boldsymbol{p})] &\leqslant \mathbb{E}_{Y^n \sim \boldsymbol{Q}} \left[l_1 \left(\hat{\boldsymbol{p}}, \boldsymbol{p}^* \right) \right] \\
&= \sqrt{\frac{2}{n\pi}} \sum_{x \in \mathcal{D}} \sqrt{(\mathcal{S}_{\mathcal{D}} |\mathcal{D}|)^{-1} \left[\mathcal{S}_{\mathcal{D}_S}^{-1} - (\mathcal{S}_{\mathcal{D}} |\mathcal{D}|)^{-1} \right]} \\
&= \sqrt{\frac{2}{n\pi}} |\mathcal{D}| \sqrt{(\mathcal{S}_{\mathcal{D}} |\mathcal{D}|)^{-1} \left[\mathcal{S}_{\mathcal{D}_S}^{-1} - (\mathcal{S}_{\mathcal{D}} |\mathcal{D}|)^{-1} \right]} \\
&= \sqrt{\frac{2}{n\pi}} \sqrt{\mathcal{S}_{\mathcal{D}}^{-1} \left(|\mathcal{D}| \mathcal{S}_{\mathcal{D}_S}^{-1} - \mathcal{S}_{\mathcal{D}}^{-1} \right)} \\
&= \sqrt{\frac{2}{n\pi}} \sqrt{\mathcal{S}_{\mathcal{D}}^{-2} (|\mathcal{D}| - 1)} \\
&= \sqrt{\frac{2(|\mathcal{D}| - 1)}{n\pi}} \mathcal{S}_{\mathcal{D}}^{-1}.
\end{aligned}$$

□

References

1. "kosarak dataset", http://fimi.uantwerpen.be/data/

2. Abstracts of Papers. The Annals of Mathematical Statistics **20**(4), 620 – 624 (1949). https://doi.org/10.1214/aoms/1177729959, https://doi.org/10.1214/aoms/1177729959
3. Alaggan, M., Gambs, S., Kermarrec, A.: Heterogeneous differential privacy. J. Priv. Confidentiality **7**(2) (2016)
4. Bassily, R., Nissim, K., Stemmer, U., Thakurta, A.G.: Practical locally private heavy hitters. In: NIPS. pp. 2288–2296 (2017)
5. Bassily, R., Smith, A.D.: Local, private, efficient protocols for succinct histograms. In: STOC. pp. 127–135. ACM (2015)
6. Bun, M., Steinke, T.: Concentrated differential privacy: Simplifications, extensions, and lower bounds. In: TCC (B1). Lecture Notes in Computer Science, vol. 9985, pp. 635–658 (2016)
7. Chatzikokolakis, K., Andrés, M.E., Bordenabe, N.E., Palamidessi, C.: Broadening the scope of differential privacy using metrics. In: Privacy Enhancing Technologies. Lecture Notes in Computer Science, vol. 7981, pp. 82–102. Springer (2013)
8. Chen, R., Li, H., Qin, A.K., Kasiviswanathan, S.P., Jin, H.: Private spatial data aggregation in the local setting. In: ICDE. pp. 289–300. IEEE Computer Society (2016)
9. Chen, Z., Wang, J.: Ldp-fpminer: Fp-tree based frequent itemset mining with local differential privacy. CoRR **abs/2209.01333** (2022)
10. Cuff, P., Yu, L.: Differential privacy as a mutual information constraint. In: CCS. pp. 43–54. ACM (2016)
11. Duchi, J.C., Jordan, M.I., Wainwright, M.J.: Local privacy and statistical minimax rates. In: FOCS. pp. 429–438. IEEE Computer Society (2013)
12. Dwork, C.: Differential privacy. In: ICALP (2). Lecture Notes in Computer Science, vol. 4052, pp. 1–12. Springer (2006)
13. Dwork, C., McSherry, F., Nissim, K., Smith, A.D.: Calibrating noise to sensitivity in private data analysis. In: TCC. Lecture Notes in Computer Science, vol. 3876, pp. 265–284. Springer (2006)
14. Erlingsson, Ú., Pihur, V., Korolova, A.: RAPPOR: randomized aggregatable privacy-preserving ordinal response. In: CCS. pp. 1054–1067. ACM (2014)
15. Gu, X., Li, M., Xiong, L., Cao, Y.: Providing input-discriminative protection for local differential privacy. In: ICDE. pp. 505–516. IEEE (2020)
16. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: SIGMOD Conference. pp. 1–12. ACM (2000)
17. Jorgensen, Z., Yu, T., Cormode, G.: Conservative or liberal? personalized differential privacy. In: ICDE. pp. 1023–1034. IEEE Computer Society (2015)
18. Kairouz, P., Bonawitz, K.A., Ramage, D.: Discrete distribution estimation under local privacy. In: ICML. JMLR Workshop and Conference Proceedings, vol. 48, pp. 2436–2444. JMLR.org (2016)
19. Kairouz, P., Oh, S., Viswanath, P.: Extremal mechanisms for local differential privacy. In: NIPS. pp. 2879–2887 (2014)
20. Kawamoto, Y., Murakami, T.: Differentially private obfuscation mechanisms for hiding probability distributions. CoRR **abs/1812.00939** (2018)
21. Kifer, D., Machanavajjhala, A.: Pufferfish: A framework for mathematical privacy definitions. ACM Trans. Database Syst. **39**(1), 3:1–3:36 (2014)
22. Kotsogiannis, I., Doudalis, S., Haney, S., Machanavajjhala, A., Mehrotra, S.: One-sided differential privacy. In: ICDE. pp. 493–504. IEEE (2020)
23. Lin, B., Kifer, D.: Information preservation in statistical privacy and bayesian estimation of unattributed histograms. In: SIGMOD Conference. pp. 677–688. ACM (2013)

24. Lin, W., Li, B., Wang, C.: Towards private learning on decentralized graphs with local differential privacy. IEEE Trans. Inf. Forensics Secur. **17**, 2936–2946 (2022)
25. Liu, C., Chakraborty, S., Mittal, P.: Dependence makes you vulnberable: Differential privacy under dependent tuples. In: NDSS. The Internet Society (2016)
26. Mangat, N.S.: An improved randomized response strategy. Journal of the Royal Statistical Society. Series B (Methodological) **56**(1), 93–95 (1994), http://www.jstor.org/stable/2346030
27. Mironov, I.: Rényi differential privacy. In: CSF. pp. 263–275. IEEE Computer Society (2017)
28. Murakami, T., Kawamoto, Y.: Utility-optimized local differential privacy mechanisms for distribution estimation. In: USENIX Security Symposium. pp. 1877–1894. USENIX Association (2019)
29. Nie, Y., Yang, W., Huang, L., Xie, X., Zhao, Z., Wang, S.: A utility-optimized framework for personalized private histogram estimation. IEEE Trans. Knowl. Data Eng. **31**(4), 655–669 (2019)
30. Qin, Z., Yu, T., Yang, Y., Khalil, I., Xiao, X., Ren, K.: Generating synthetic decentralized social graphs with local differential privacy. In: CCS. pp. 425–438. ACM (2017)
31. Wang, N., Xiao, X., Yang, Y., Hoang, T.D., Shin, H., Shin, J., Yu, G.: Privtrie: Effective frequent term discovery under local differential privacy. In: ICDE. pp. 821–832. IEEE Computer Society (2018)
32. Wang, T., Li, N., Jha, S.: Locally differentially private frequent itemset mining. In: IEEE Symposium on Security and Privacy. pp. 127–143. IEEE Computer Society (2018)
33. Wang, T., Li, N., Jha, S.: Locally differentially private heavy hitter identification. IEEE Trans. Dependable Secur. Comput. **18**(2), 982–993 (2021)
34. Wang, T., Lopuhaä-Zwakenberg, M., Li, Z., Skoric, B., Li, N.: Locally differentially private frequency estimation with consistency. In: NDSS. The Internet Society (2020)
35. Wang, T., Xu, M., Ding, B., Zhou, J., Hong, C., Huang, Z., Li, N., Jha, S.: Improving utility and security of the shuffler-based differential privacy. Proc. VLDB Endow. **13**(13), 3545–3558 (2020)
36. Wang, Z., Zhu, Y., Wang, D., Han, Z.: Fedfpm: A unified federated analytics framework for collaborative frequent pattern mining. In: INFOCOM. pp. 61–70. IEEE (2022)
37. Warner, S.L.: Randomized response: a survey technique for eliminating evasive answer bias. Publications of the American Statistical Association **60** (1965)
38. Wei, C., Ji, S., Liu, C., Chen, W., Wang, T.: Asgldp: Collecting and generating decentralized attributed graphs with local differential privacy. IEEE Trans. Inf. Forensics Secur. **15**, 3239–3254 (2020)
39. Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G.: Differentially private histogram publication. In: ICDE. pp. 32–43. IEEE Computer Society (2012)
40. Yang, B., Sato, I., Nakagawa, H.: Bayesian differential privacy on correlated data. In: SIGMOD Conference. pp. 747–762. ACM (2015)