Parameswaran Kamalaruban*, Victor Perrier, Hassan Jameel Asghar, and Mohamed Ali Kaafar

Not All Attributes are Created Equal: $d_{\mathcal{X}}$-Private Mechanisms for Linear Queries

Abstract: Differential privacy provides strong privacy guarantees simultaneously enabling useful insights from sensitive datasets. However, it provides the same level of protection for all elements (individuals and attributes) in the data. There are practical scenarios where some data attributes need more/less protection than others. In this paper, we consider $d_{\mathcal{X}}$-privacy, an instantiation of the privacy notion introduced in [6], which allows this flexibility by specifying a separate privacy budget for each pair of elements in the data domain. We describe a systematic procedure to tailor any existing differentially private mechanism that assumes a query set and a sensitivity vector as input into its $d_{\mathcal{X}}$-private variant, specifically focusing on linear queries. Our proposed meta procedure has broad applications as linear queries form the basis of a range of data analysis and machine learning algorithms, and the ability to define a more flexible privacy budget across the data domain results in improved privacy/utility tradeoff in these applications. We propose several $d_{\mathcal{X}}$-private mechanisms, and provide theoretical guarantees on the trade-off between utility and privacy. We also experimentally demonstrate the effectiveness of our procedure, by evaluating our proposed $d_{\mathcal{X}}$-private Laplace mechanism on both synthetic and real datasets using a set of randomly generated linear queries.

Keywords: database privacy, linear queries, differential privacy

DOI 10.2478/popets-2020-0007
Received 2019-05-31; revised 2019-09-15; accepted 2019-09-16.

***Corresponding Author: Parameswaran Kamalaruban:** École Polytechnique Fédérale de Lausanne (work done while Kamalaruban was a Postgraduate Researcher at Data61, CSIRO), E-mail: kamalaruban.parameswaran@epfl.ch
Victor Perrier: ISAE-SUPAERO & Data61, CSIRO, E-mail: v.perrier0@gmail.com
Hassan Jameel Asghar: Macquarie University & Data61, CSIRO, E-mail: hassan.asghar@mq.edu.au
Mohamed Ali Kaafar: Macquarie University & Data61, CSIRO, E-mail: dali.kaafar@mq.edu.au

1 Introduction

Differential privacy [8] is a formal notion of privacy that allows a trustworthy data curator, in possession of sensitive data from several individuals, to approximately answer a set of queries submitted by an analyst while maintaining individual privacy. Intuitively, differential privacy guarantees that query answers with or without any individual's data are (almost) indistinguishable. One common mechanism for achieving differential privacy is to inject random noise to the query answers, carefully calibrated according to the sensitivity of the query and a global privacy budget ϵ. Sensitivity here is defined as the maximum amount of change in the query answer considering all neighboring datasets, *i.e.*, datasets differing in the data of one individual (one row), or equivalently having Hamming distance of one. One limitation of this definition is that it provides the same level of protection for all attributes of the dataset, i.e., all elements in the data universe $\mathcal{X}$.

In many scenarios, a more flexible notion of neighboring datasets may be more useful. For instance, in some domains it might be more natural to measure the distinguishability between two datasets by some generic metric $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ instead of just Hamming distance. A case in point is location-based systems where it might be acceptable to disclose coarse-grained information about an individual's location instead of his/her exact location. In this case, geographical distance would be an appropriate measure of distinguishability [2]. There are other scenarios where some attributes of the dataset may need more protection than others, and vice versa. As an example, consider a classification problem with instance space $\mathcal{X} \subset \mathbb{R}^d$ where specific features of $\mathcal{X}$ are more sensitive than others (maybe due to fairness requirements [7]). In this case, $d_{\mathcal{X}}(u, v) = \sum_{i=1}^{d} \epsilon_i [\![u_i \neq v_i]\!], \forall u, v \in \mathcal{X}$ might be a reasonable choice for the metric, where ϵ_i is the privacy budget for the ith feature.[1]

In the applications mentioned above, standard differential privacy (with a global privacy budget) is too

[1] Note that $[\![P]\!] = 1$ if predicate P is true.

strong privacy guarantee and as a result compromises much in utility. To address this limitation, several relaxations of differential privacy have been proposed recently [6, 16]. Despite the existence of these alternative proposals, they have gained little traction among practitioners. Part of the reason, we believe, is the dearth of standard procedures to develop algorithms satisfying these alternative definitions, as compared to differential privacy. In this work, we attempt to bridge this gap, by building on the privacy notion of *$d_{\mathcal{X}}$-privacy*, introduced in [6]. Intuitively, $d_{\mathcal{X}}$-privacy allows specifying a separate privacy budget for each pair of elements u, v in the data universe $\mathcal{X}$, given by the value $d_{\mathcal{X}}(u, v)$.

We propose a generic strategy to tailor any differentially private mechanism to satisfy $d_{\mathcal{X}}$-privacy for linear queries. Given *any* data universe $\mathcal{X}$ and *any* choice of the metric $d_{\mathcal{X}}$, the procedure shows how to convert a differentially private mechanism into its $d_{\mathcal{X}}$-private counterpart tailored to the given utility measure. The resulting mechanism then provides a better trade-off between utility and privacy. Our main contributions are summarized as follows:

- We describe a meta procedure (for any metric) to tailor any existing differentially private mechanism into a $d_{\mathcal{X}}$-private variant for the case of linear queries. The main challenge is that the privacy budget, *i.e.*, $d_{\mathcal{X}}$, is specified on the input universe $\mathcal{X}$, whereas the noise is added to the query response which belongs to the outcome space.
- The main component of our approach is a pre-processing optimization step, which depends on the utility measure of interest, to choose model parameters of the mechanism. We provide explicit formulation of this pre-processing optimization problem for some commonly used utility measures (under any $d_{\mathcal{X}}$-metric). In general, these problems are non-convex and computationally challenging. But we show that for certain loss functions the optimization problem can be approximately solved using heuristic approaches (cf. Algorithm 3 in Section 4.2).
- Based on the meta procedure, we describe $d_{\mathcal{X}}$-private variants of several well-known online and offline ϵ-differentially private algorithms. In particular, we illustrate $d_{\mathcal{X}}$-private variants of the Laplace [8] and Exponential [21] mechanisms, as well as the SmallDB [5] and MWEM [14] mechanisms, the latter two being mechanisms for releasing synthetic data. We remark that the choice of these algorithms is merely for demonstration. Our meta procedure can similarly be applied to other differentially private mechanisms.
- We demonstrate the effectiveness of $d_{\mathcal{X}}$-privacy in terms of utility, by evaluating the proposed $d_{\mathcal{X}}$-private Laplace mechanism on both synthetic and real datasets using a set of randomly generated linear queries. In both cases we define the $d_{\mathcal{X}}$ metric as the Euclidean distance between elements in the data universe. Our results show that the utility from the $d_{\mathcal{X}}$-private Laplace mechanism is higher than its *vanilla* counterpart, with some specific queries showing significant improvement.
- Finally, we demonstrate how $d_{\mathcal{X}}$-privacy generalizes and relates to other alternative privacy notions proposed in literature by extending our techniques to Blowfish [16] privacy (without constraints).
- Our work is the first to propose $d_{\mathcal{X}}$-private mechanisms for linear queries over histograms in the centralized model. This is in contrast to the most related work to ours, i.e., [2] where the authors focus on location-based systems in the local model, and [6] which only considers universally optimal mechanisms under some specific $d_{\mathcal{X}}$ metrics.

2 Background and $d_{\mathcal{X}}$-Privacy

This section gives the background on differential privacy and associated concepts of linear queries, sensitivity, and utility. We also introduce $d_{\mathcal{X}}$-privacy and its relation to other privacy notions.

2.1 Notation

Let $[n] := \{1, \ldots, n\}$ for $n \in \mathbb{N}$, and $\mathbb{R}_+ := [0, \infty)$. We write $[\![P]\!] = 1$ if P is true and $[\![P]\!] = 0$ otherwise. Let x_i denote the ith coordinate of the vector x, and $A_{i,:}$ denote the ith row of the matrix A. We denote the inner product of two vectors $x, y \in \mathbb{R}^n$ by $\langle x, y \rangle$. The k-element vector of all ones is denoted $\mathbf{1}_k := (1, \ldots, 1)^\top$. For two vectors $a, b \in \mathbb{R}^n$, the operation $a \odot b$ represents element-wise multiplication. For $a \in \mathbb{R}^n$ and $B \in \mathbb{R}^{n \times d}$, the operation $a \odot B$ represents row-wise scalar multiplication of B by the associated entry of a. For a vector $a \succeq 0$ represents that the vector is element-wise non-negative. Hamming distance is defined as $\|x - y\|_H := \sum_{i=1}^{n} [\![x_i \neq y_i]\!]$. The ℓ_p-norms are denoted by $\|\cdot\|_p$. For a matrix A, define $\|A\|_p := \left(\sum_i \|A_{i,:}\|_p^p\right)^{1/p}$.

2.2 Differential Privacy

Let $\mathcal{X}$ denote the data universe and $N := |\mathcal{X}|$ its size. A database D of n rows is modelled as a histogram $x \in \mathbb{N}^N$ (with $\|x\|_1 = n$), where x_i encodes the number of occurrences of the ith element of the universe $\mathcal{X}$.[2] Two neighboring databases D and D' (from $\mathcal{X}^n$) that differ in a single row ($\|D - D'\|_H = 1$) correspond to two histograms x and x' (from $\mathbb{N}^N$) satisfying $\|x - x'\|_1 = 2$. A mechanism $\mathcal{M} : \mathbb{N}^N \times \mathcal{Q} \rightsquigarrow \mathcal{Y}$ (where $\mathcal{Y}$ is the outcome space, and $\mathcal{Q}$ is the query class) is a randomized algorithm which takes a dataset $x \in \mathbb{N}^N$ and a query $q : \mathbb{N}^N \to \mathcal{Y}$, and answers with some $a \in \mathcal{Y}$.

Definition 1 (Differential Privacy, [8])**.** *A mechanism $\mathcal{M} : \mathbb{N}^N \times \mathcal{Q} \rightsquigarrow \mathcal{Y}$ is called ϵ-differentially private if for all $x, x' \in \mathbb{N}^N$ such that $\|x - x'\|_1 \le 2$, for every $q \in \mathcal{Q}$, and for every measurable $S \subseteq \mathcal{Y}$, we have*

$$\mathbb{P}\left[\mathcal{M}(x, q) \in S\right] \le \exp(\epsilon)\, \mathbb{P}\left[\mathcal{M}\left(x', q\right) \in S\right].$$

Here $\epsilon > 0$ is a parameter that measures the strength of the privacy guarantee (smaller ϵ being a stronger guarantee).

2.3 $d_{\mathcal{X}}$-Privacy

We consider a more flexible privacy notion, which is a particular case of the definition from [6], for statistical databases. Given a metric $d_{\mathcal{X}}$ on the data universe, a mechanism satisfies $d_{\mathcal{X}}$-privacy if the densities of the output distributions on input datasets $x, x' \in \mathbb{N}^N$ with $\|x - x'\|_1 \le 2$ and differing on i, j-th entries are pointwise within an $\exp\left(d_{\mathcal{X}}(i, j)\right)$ multiplicative factor of each other.

Definition 2 ($d_{\mathcal{X}}$-Privacy)**.** *Let $d_{\mathcal{X}} : [N] \times [N] \to \mathbb{R}_+$ be the privacy budget (such that $d_{\mathcal{X}}(i,j) \ge 0$, $d_{\mathcal{X}}(i,j) = d_{\mathcal{X}}(j,i)$, $d_{\mathcal{X}}(i,i) = 0$, and $d_{\mathcal{X}}(i,j) \le d_{\mathcal{X}}(i,k) + d_{\mathcal{X}}(k,j)$, $\forall i,j,k \in [N]$) of the data universe $\mathcal{X}$. A mechanism $\mathcal{M} : \mathbb{N}^N \times \mathcal{Q} \rightsquigarrow \mathcal{Y}$ is said to be $d_{\mathcal{X}}$-private iff $\forall x, x' \in \mathbb{N}^N$ s.t. $\|x - x'\|_1 \le 2$, $x_i \ne x_i'$, and $x_j \ne x_j'$ (for some $i, j \in [N]$), $\forall S \subseteq \mathcal{Y}$ and $\forall q \in \mathcal{Q}$ we have*

$$\frac{\mathbb{P}\left[\mathcal{M}(x, q) \in S\right]}{\mathbb{P}\left[\mathcal{M}(x', q) \in S\right]} \le \exp\left(d_{\mathcal{X}}(i, j)\right).$$

When $d_{\mathcal{X}}(i,j) = \epsilon, \forall i, j \in [N]$, we recover the standard ϵ-differential privacy.

Most of the desirable properties of differential privacy is carried over to $d_{\mathcal{X}}$-privacy as well, with suitable generalization [6].

Fact 1 (Properties of $d_{\mathcal{X}}$-Privacy)**.** *$d_{\mathcal{X}}$-privacy satisfies the following properties:*

1. *Resistant to post-processing: If $\mathcal{M} : \mathbb{N}^N \times \mathcal{Q} \rightsquigarrow \mathcal{Y}$ is $d_{\mathcal{X}}$-private, and $f : \mathcal{Y} \to \mathcal{Y}'$ is any arbitrary (randomized) function, then the composition $f \circ \mathcal{M} : \mathbb{N}^N \times \mathcal{Q} \rightsquigarrow \mathcal{Y}'$ is also $d_{\mathcal{X}}$-private.*
2. *Composability: Let $\mathcal{M}_i : \mathcal{X}^n \times \mathcal{Q}_i \rightsquigarrow \mathcal{Y}_i$ be a $d^i_{\mathcal{X}}$-private algorithm for $i \in [k]$. If $\mathcal{M}_{[k]} : \mathcal{X}^n \times \Pi_{i=1}^k \mathcal{Q}_i \rightsquigarrow \Pi_{i=1}^k \mathcal{Y}_i$ is defined to be:*

   $$\mathcal{M}_{[k]}\left(x, \Pi_{i=1}^k q_i\right) = \left(\mathcal{M}_1(x, q_1), \ldots, \mathcal{M}_k(x, q_k)\right),$$

   then $\mathcal{M}_{[k]}$ is $\sum_{i=1}^k d^i_{\mathcal{X}}$-private.
3. *Group privacy: If $\mathcal{M} : \mathbb{N}^N \times \mathcal{Q} \rightsquigarrow \mathcal{Y}$ is a $d_{\mathcal{X}}$-private mechanism and $x, x' \in \mathbb{N}^N$ satisfy $\|x - x'\|_1 \le k$ (with $k \ge 2$), then $\forall S \subseteq \mathcal{Y}$ and $\forall q \in \mathcal{Q}$ we have*

   $$\frac{\mathbb{P}\left[\mathcal{M}(x, q) \in S\right]}{\mathbb{P}\left[\mathcal{M}(x', q) \in S\right]} \le \exp\left(k \cdot \max_{i,j \in V} d_{\mathcal{X}}(i, j)\right),$$

   where V is the set of indices in which x and x' differ.

$d_{\mathcal{X}}$-privacy can naturally express indistinguishability requirements that cannot be represented by the standard notion of Hamming distance (between neighbouring datasets). But the metric $d_{\mathcal{X}}$ in the above definition must be appropriately defined to achieve meaningful privacy goals. We present some examples in Section 5. In this work, we mainly focus on how to convert an existing differentially private algorithm into $d_{\mathcal{X}}$-private equivalent, given an already appropriately defined $d_{\mathcal{X}}$-metric.

2.4 Linear Queries and Sensitivity

Our focus is on the trade-off between privacy and accuracy when answering a large number of *linear queries* over histograms. Linear queries include some natural classes of queries such as range queries [18, 19] and contingency tables [3, 10], and serve as the basis of a wide range of data analysis and learning algorithms (*e.g.*, Perceptron, K-means clustering, PCA [4]). Formally, given a query vector $q \in \mathbb{R}^N$, a linear query over the dataset $x \in \mathbb{N}^N$ is defined as $q(x) = \langle q, x \rangle$. A set of k linear queries can be represented by a *query matrix*

[2] Note that each element of the universe is composed of attribute values from all attributes (columns) in the tuple-wise representation of the database.

$Q \in \mathbb{R}^{k \times N}$ with the vector $Qx \in \mathbb{R}^k$ giving the correct answers to the queries.

For $d_{\mathcal{X}}$-privacy, we generalize the notion of *global sensitivity*, defined in [8], as follows:

Definition 3. *For $i, j \in [N]$ (with $i \neq j$), the generalized global sensitivity of a query $q \in \mathcal{Q}$ (w.r.t. $\|\cdot\|$) is defined as follows*

$$\Delta^q_{\|\cdot\|}(i,j) \ := \max_{\substack{x,x' \in \mathbb{N}^N : \|x-x'\|_1 \leq 2, \\ x_i \neq x'_i, x_j \neq x'_j \ for \ i,j \in [N]}} \left\| q(x) - q\left(x'\right) \right\|.$$

Also define $\Delta^q_{\|\cdot\|} := \max_{i,j \in [N]} \Delta^q_{\|\cdot\|}(i,j)$ (the usual global sensitivity). When $\|\cdot\| = \|\cdot\|_p$, we simply write Δ^q_p.

Consider a multi-linear query $Q : \mathbb{N}^N \to \mathcal{Y} \subseteq \mathbb{R}^k$ defined as $Q(x) = Qx$, where $Q \in \mathbb{R}^{k \times N}$. Then the generalized global sensitivity of Q (for $i, j \in [N]$) is given by $\Delta^Q_{\|\cdot\|}(i,j) = \|Q_{:,i} - Q_{:,j}\|$. When $k = 1$, *i.e.*, for a single linear query $q(x) = \langle q, x \rangle$, we have $\Delta^q_{\|\cdot\|}(i,j) = |q_i - q_j|$. Thus, *the generalized notion is defined separately for each pair i, j of elements in $\mathcal{X}$.*

2.5 Laplace and Exponential Mechanism

Definition 4 (Laplace Mechanism, [8])**.** *For a query function $q : \mathbb{N}^N \to \mathbb{R}^k$ with ℓ_1-sensitivity Δ^q_1, Laplace mechanism will output*

$$\mathsf{Z} \ = \ \mathcal{M}_{\mathrm{Lap}, \frac{\Delta^q_1}{\epsilon} \cdot \mathbf{1}_k}(x, q) \ := \ q(x) + (\mathsf{Y}_1, \ldots, \mathsf{Y}_k), \quad (1)$$

where $\mathsf{Y}_i \overset{\text{iid}}{\sim} \mathrm{Lap}\left(\frac{\Delta^q_1}{\epsilon}\right)$, and $\mathrm{Lap}(\lambda)$ is a distribution with probability density function $f(x) = \frac{1}{2\lambda} e^{-\frac{|x|}{\lambda}}, \quad \forall x \in \mathbb{R}$.

The Laplace mechanism satisfies ϵ-differential privacy, but it satisfies $d_{\mathcal{X}}$-privacy only with $\epsilon \leq \min_{i,j \in [N]} d_{\mathcal{X}}(i,j)$. This would result in large noise addition, and eventually unnecessary compromise on overall utility.

Given some arbitrary range $\mathcal{R}$, the exponential mechanism is defined with respect to some utility function $u : \mathbb{N}^N \times \mathcal{R} \to \mathbb{R}$, which maps database/output pairs to utility scores. The sensitivity notion that we are interested here is given by:

Definition 5. *For $i, j \in [N]$ (with $i \neq j$) and $u : \mathbb{R}^N \times \mathcal{R} \to \mathbb{R}$, the generalized utility sensitivity is defined as follows*

$$\Delta u(i,j) := \max_{r \in \mathcal{R}} \max_{\substack{x,x' \in \mathbb{N}^N : \|x-x'\|_1 \leq 2, \\ x_i \neq x'_i, x_j \neq x'_j \ for \ i,j \in [N]}} \left| u(x,r) - u\left(x', r\right) \right|.$$

Also define $\Delta u := \max_{i,j \in [N]} \Delta u(i,j)$.

Formally, the exponential mechanism is:

Definition 6 (The Exponential Mechanism, [21])**.** *The exponential mechanism $\mathcal{M}_{\mathrm{Exp}, \frac{\Delta u}{\epsilon}}(x, u)$ selects and outputs an element $r \in \mathcal{R}$ with probability proportional to $\exp\left(\frac{\epsilon u(x,r)}{2\Delta u}\right)$.*

The exponential mechanism satisfies ϵ-differential privacy. The resulting mechanism also satisfies $d_{\mathcal{X}}$-privacy only if we set $\epsilon \leq \min_{i,j \in [N]} d_{\mathcal{X}}(i,j)$.

2.6 Utility

In the differential privacy literature, the performance of a mechanism is usually measured in terms of its worst-case total expected *error*, defined as follows:

Definition 7 (Error)**.** *Let $q : \mathbb{N}^N \to \mathcal{Y} \subseteq \mathbb{R}^k$ and $\ell : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_+$. We define the ℓ-error of a mechanism $\mathcal{M} : \mathbb{N}^N \times \mathcal{Q} \rightsquigarrow \mathcal{Y}$ as*

$$\mathrm{err}_\ell(\mathcal{M}, q) = \sup_{x \in \mathbb{N}^N} \mathop{\mathbb{E}}_{\mathsf{Z} \sim \mathcal{M}(x,q)} [\ell(\mathsf{Z}, q(x))]. \quad (2)$$

Here the expectation is taken over the internal coin tosses of the mechanism itself.

In this paper, we are mainly interested in the worst case expected ℓ_p-error defined by

$$\ell_p(y, \hat{y}) \ := \ \|y - \hat{y}\|_p \ = \ \left(\sum_{i=1}^k |y_i - \hat{y}_i|^p \right)^{\frac{1}{p}},$$

for $p \in \{1, 2, \infty\}$, and ℓ_2^2-error (given by $\ell_2^2(y, \hat{y}) := \|y - \hat{y}\|_2^2$). It is also common to analyze high probability bounds on the accuracy of the privacy mechanisms.

Definition 8 (Accuracy)**.** *Given a mechanism $\mathcal{M} : \mathbb{N}^N \times \mathcal{Q} \rightsquigarrow \mathcal{Y}$, query $q : \mathbb{N}^N \to \mathcal{Y} \subseteq \mathbb{R}^k$, sensitive dataset (histogram) $x \in \mathbb{N}^N$, and parameters $\alpha > 0$ and $\beta \in (0,1)$, the mechanism $\mathcal{M}$ is (α, β)-accurate for q on x under the $\|\cdot\|$-norm if $\mathbb{P}[\|\mathcal{M}(q,x) - q(x)\| \geq \alpha] \leq \beta$ where $\|\cdot\|$-norm can be any vector norm definition. In our analysis, we consider the $\|\cdot\|_1$-norm and the $\|\cdot\|_\infty$-norm.*

3 $d_{\mathcal{X}}$-Private Mechanisms for Linear Queries

In this section, we design $d_{\mathcal{X}}$-private mechanisms by extending some of the well known ϵ-differentially private (noise adding) mechanisms. Before delving into our approach, we exemplify potential technical issues when defining a variable privacy budget across attributes, and how $d_{\mathcal{X}}$-privacy provides a solution.

Example 1. Consider a simple domain $\mathcal{X}$ with three binary attributes described below.

Gender	Native	Age
Male (M)	Yes (Y)	Above 18 (A)
Female (F)	No (N)	Below 18 (B)

Figure 1 shows an example dataset x (as a histogram) from this domain, where we have used abbreviated attribute value names to describe each element x_i, $i \in [N] = [8]$. Assume that the attribute value "Native = Y" is considered sensitive and all other values non-sensitive. Using differential privacy, the data custodian might wish to use some privacy budget $\epsilon = \epsilon_0$ for the attribute value Y, and $\epsilon = \infty$ for all other attribute values. Any linear query $q \in \mathbb{R}^8$ with a non-zero value for coordinates 1, 2, 5 or 6 (containing attribute value Y), would be answered with the budget ϵ_0, and all remaining queries with budget ∞ (i.e., noiseless answers). While this may sound reasonable, notice that an analyst can obtain noiseless answer to the query (MNA, MNB, FNA, FNB) = (N),[3] and get the answer to (MYA, MYB, FYA, FYB) = (Y) without noise (since $\|x\|_1 = n$ is assumed to be publicly known). This simple example shows why attribute-wise privacy budget allocation should obey properties of a distance metric.

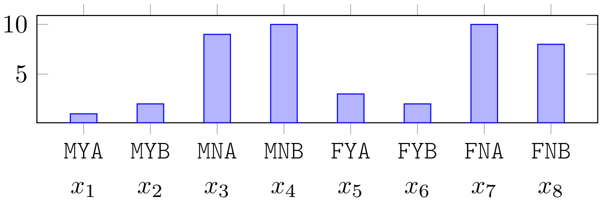


Fig. 1. Histogram representation of the dataset x used in Example 1; horizontal axis contains domain elements, while the vertical axis shows counts.

We can solve this using $d_{\mathcal{X}}$-privacy as follows. Denote by $\epsilon(X)$ the privacy budget allocated to attribute value X. Then, set $\epsilon(\texttt{Y}) = \epsilon_0$, and $\epsilon(X) = \epsilon_1 > \epsilon_0$, for all $X \neq \texttt{Y}$. Denote each of the x_i's as $x_i = X_i^{(1)} X_i^{(2)} X_i^{(3)}$, where $X_i^{(k)}$ denotes the kth attribute value of x_i. Finally, we define

$$d_{\mathcal{X}}(i,j) = \sum_{k=1}^{3} \min\{\epsilon(X_i^{(k)}), \epsilon(X_j^{(k)})\} [\![X_i^{(k)} \neq X_j^{(k)}]\!]. \tag{3}$$

For instance, $d_{\mathcal{X}}(1,1) = 0$, $d_{\mathcal{X}}(1,2) = \epsilon_1 \cdot 0 + \epsilon_0 \cdot 0 + \epsilon_1 \cdot 1 = \epsilon_1$, $d_{\mathcal{X}}(1,3) = \epsilon_0$, and $d_{\mathcal{X}}(1,8) = \epsilon_0 + 2\epsilon_1$. It is easy to verify that this is indeed a distance metric. As we will show in Section 4.1, given any linear query $q \in \mathbb{R}^8$, one way to answer through our framework is to let $c = \max_{i,j} \frac{|q_i - q_j|}{d_{\mathcal{X}}(i,j)}$, and then add Laplace noise of scale c to the answer. Thus, for instance, if $q = (\texttt{MNA}, \texttt{MNB}, \texttt{FNA}, \texttt{FNB}) = (\texttt{N})$, then the maximum is achieved at $i = 1, j = 3$ (not uniquely), which gives us $c = 1/\epsilon_0$. We have already seen that this is indeed a sensitive query, and is equivalent to answering the query (Y). Thus, the scale of noise is justified. Likewise, for the query $q = (1,1,1,1,0,0,0,0) = (\texttt{M})$, which is not considered sensitive, the maximum is achieved at $i = 1, j = 5$ (again, not uniquely), giving us $c = 1/\epsilon_1$. Thus, less noise is added to the answer to this query. Finally, notice that with the convention $a + \infty = \infty$ for all $a \in \mathbb{R}^+$, if we set $\epsilon_1 = \infty$, Eq. 3 still defines a distance metric. Thus we can even get noiseless answers to the "non-sensitive" queries. □

Consider a query $q : \mathbb{N}^N \to \mathbb{R}$, and the Laplace mechanism (Definition 4). To achieve ϵ-differential privacy, we compute the difference in the answers to the query over all neighbouring databases, given by the global sensitivity $\Delta^q_{\|\cdot\|}$, and then add noise of scale $\Delta^q_{\|\cdot\|}/\epsilon$. The corresponding quantity in $d_{\mathcal{X}}$-privacy is $\Delta^q_{\|\cdot\|}(i,j)/d_{\mathcal{X}}(i,j)$; but this is potentially different for all $i, j \in [N]$. Therefore, to obtain a $d_{\mathcal{X}}$-private counterpart, we need to obtain a single optimum noise scale $c \geq 0$ subject to the constraints $c \geq \Delta^q_{\|\cdot\|}(i,j)/d_{\mathcal{X}}(i,j)$, for all $i, j \in [N]$ (to guarantee privacy). Equivalently, if we introduce a parameter q' and set it approximately equal to q/c, i.e., $q' \approx q/c$, then we perturb the answer to the *scaled query* $cq' \approx q$ with Laplace noise of scale c subject to the con-

[3] We are omitting zeroed coordinates in this equivalent notation.

dition $\Delta^{q'}_{\|\cdot\|}(i,j) \leq d_{\mathcal{X}}(i,j)$, for all $i,j \in [N]$. The parameters c and q' can be optimized for a given measure of utility (within the privacy constraints imposed by the $d_{\mathcal{X}}$ metric). Note that the two parameters are not dependent on the input data, and can be optimized using pre-processing without compromising privacy. To summarize, given a query q, we have a two-step process: (a) obtain parameters c and q' through optimization (via pre-processing), (b) convert the ϵ-differentially private Laplace mechanism into its $d_{\mathcal{X}}$-private counterpart by replacing input $(q, \Delta^{q}_{\|\cdot\|}/\epsilon)$ by input (cq', c). The exact form of the optimization problem depends on the utility measure under consideration. We generalize this procedure in the following.

Given a dataset $x \in \mathbb{N}^N$, and a query $q : \mathbb{N}^N \to \mathcal{Y} \subseteq \mathbb{R}^k$, our approach (meta procedure) to design a $d_{\mathcal{X}}$-private (noise adding) mechanism is as follows:

1. Choose the (approximately optimal) model parameters $c \in \mathbb{R}^k$ and $q' : \mathbb{N}^N \to \mathbb{R}^k$ such that $\Delta^{q'}_{\|\cdot\|}(i,j) \leq d_{\mathcal{X}}(i,j)$, $\forall i,j \in [N]$, and $c \succeq 0$.
2. Then use an existing ϵ-differentially private mechanism with $(c \odot q', c)$ in place of $\left(q, \frac{\Delta^{q}_{\|\cdot\|}}{\epsilon}\mathbf{1}_k\right)$.

The model parameters q' and c are chosen by (approximately) solving the following pre-processing optimization problem (i.e. $(q', c) := F_{\text{pre-opt}}(q, n, d_{\mathcal{X}}(\cdot,\cdot), \ell)$):

$$\begin{aligned} \underset{q',c}{\text{minimize}} \quad & f_{\ell,\mathcal{M}}\left(q', c; q, n\right) \\ \text{subject to} \quad & \Delta^{q'}_{\|\cdot\|}(i,j) \leq d_{\mathcal{X}}(i,j), \quad \forall i,j \in [N] \\ & c \succeq 0, \end{aligned} \tag{4}$$

where $f_{\ell,\mathcal{M}}(q', c; q, n)$ is a surrogate function of the utility measure that we are interested in. Note that this pre-processing optimization depends only on the data universe $\mathcal{X}$ (or $[N]$), the query set $\mathcal{Q}$, and the database size n, but not on the dataset x. Thus we don't compromise any privacy during the optimization procedure. More over, we have to do the pre-processing optimization only once in an offline manner (for given $\mathcal{X}$, $\mathcal{Q}$, and n). The number of constraints in the optimization problem (4) can be exponentially large ($\approx 2^N$), but depending on the structure of the $d_{\mathcal{X}}$-metric the constraint count can be significantly reduced (cf. Appendix C).

Remark 1: Note that the pre-processing optimization problem (4) to choose the model parameters of our new strategy depends on the 1) utility measure 2) $d_{\mathcal{X}}$-metric 3) ϵ-differentially private mechanism that we want to transform. Every ϵ-differentially private mechanism requires a separate (utility) analysis to derive $f_{\ell,\mathcal{M}}$ of (4). In this work, we consider fundamental online and offline (synthetic data generation) mechanisms (Laplace, Exponential, SmallDB, and MWEM) under specific utility measures and any metric. We can similarly extend our analysis to advanced ϵ-differentially private mechanisms (for multi-linear queries) such as the Matrix [18], and K-norm [15] mechanisms. We leave it as future work.

Next we apply the above described abstract meta procedure in extending some ϵ-differential privacy mechanisms under different loss measures such as squared loss and absolute loss. We first show that the resulting mechanisms are in fact $d_{\mathcal{X}}$-private, and then we formulate the appropriate pre-processing optimization problems (4) for them.

3.1 $d_{\mathcal{X}}$-Private Laplace Mechanism

For a given query $q : \mathbb{R}^N \to \mathcal{Y} \subset \mathbb{R}^k$ over the histogram $x \in \mathbb{R}^N$, consider the following variant of Laplace mechanism (with the model parameters $q' : \mathbb{R}^N \to \mathbb{R}^k$, and $c \in \mathbb{R}^k$ which depend on the utility function of the task):

$$\mathsf{Z} = \mathcal{M}_{\text{Lap},c}\left(x, c \odot q'\right) := c \odot q'(x) + \left(\mathsf{Y}_1, \ldots, \mathsf{Y}_k\right), \tag{5}$$

where $\mathsf{Y}_i \overset{\perp}{\sim} \text{Lap}(c_i)$. When $q(x) = Qx$ (i.e., q is a multi-linear query), we choose $Q' \in \mathbb{R}^{k \times N}$ and $c \in \mathbb{R}^k$ as the model parameters *i.e.*, $q'(x) = Q'x$. Below we show that the above variant of Laplace mechanism satisfies $d_{\mathcal{X}}$-privacy under a sensitivity bound condition.

Theorem 1. *If $\Delta^{q'}_1(i,j) \leq d_{\mathcal{X}}(i,j)$, $\forall i,j \in [N]$, then the mechanism $\mathcal{M}_{\text{Lap},c}(\cdot, c \odot q')$ given by (5) satisfies $d_{\mathcal{X}}$-privacy.*

The sensitivity bound condition of the above theorem for a multi-linear query $Q'(x) = Q'x$ can be written as: $\Delta^{Q'}_1(i,j) = \left\|Q'_{:,i} - Q'_{:,j}\right\|_1 \leq d_{\mathcal{X}}(i,j), \quad \forall i,j \in [N]$. The next theorem characterizes the performance of the $\mathcal{M}_{\text{Lap},c}(\cdot, c \odot q')$ mechanism under different choices of utility measures:

Theorem 2. *Let $Q : \mathbb{R}^N \to \mathbb{R}^k$ be a multi-linear query of the form $Q(x) = Qx$, and let $\mathsf{Z} = \mathcal{M}_{\text{Lap},c}(x, c \odot Q') = c \odot Q'x + \mathsf{Y}$ with $\mathsf{Y}_i \overset{\perp}{\sim} \text{Lap}(c_i)$.*

1. *When $\ell_2^2(y, y') = \|y - y'\|_2^2$, we have*

$$\begin{aligned} \text{err}_{\ell_2^2}\left(\mathcal{M}_{\text{Lap},c}\left(\cdot, c \odot Q'\right), Q\right) \leq\ & 2n^2 \left\|c \odot Q' - Q\right\|_2^2 \\ & + 4\left\|c\right\|_2^2, \end{aligned}$$

where $\text{err}_\ell(\mathcal{M}, Q)$ is defined in (2).

2. *When* $\ell_p(y, y') = \|y - y'\|_p$, *we have*

$$\mathrm{err}_{\ell_p}\left(\mathcal{M}_{\mathrm{Lap},c}\left(\cdot, c \odot Q'\right), Q\right) \leq n \left\|c \odot Q' - Q\right\|_p + \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\|\mathsf{Y}\|_p\right].$$

Note that $\mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\|\mathsf{Y}\|_1\right] = \|c\|_1$.

3. $\forall \delta \in (0, 1]$, *with probability at least* $1 - \delta$ *we have*

$$\|Qx - \mathsf{Z}\|_\infty \leq n \left\|c \odot Q' - Q\right\|_\infty + \ln\left(\frac{k}{\delta}\right) \cdot \|c\|_\infty.$$

Proofs of Theorem 1 and 2 are given in Appendix A.1. Based on the upper bounds that we obtained in the previous theorem, we can formulate the pre-processing optimization problem $F_{\text{pre-opt}}(q, n, d_{\mathcal{X}}(\cdot, \cdot), \ell)$ to select the model parameters c and Q' of the $\mathcal{M}_{\mathrm{Lap},c}(\cdot, c \odot Q')$ mechanism as follows:

$$\begin{aligned} \underset{Q', c}{\text{minimize}} \quad & f_{\ell, \mathcal{M}_{\mathrm{Lap},c}(\cdot, c \odot Q')}\left(Q', c; Q, n\right) \\ \text{subject to} \quad & \left\|Q'_{:,i} - Q'_{:,j}\right\|_1 \leq d_{\mathcal{X}}(i, j), \forall i, j \in [N] \\ & c \succeq 0. \end{aligned} \tag{6}$$

The objective function of the above optimization problem depends on the utility function that we are interested in. For example, when $\ell_2^2(y, y') = \|y - y'\|_2^2$, we can choose $f_{\ell_2^2, \mathcal{M}_{\mathrm{Lap},c}(\cdot, c \odot Q')}(Q', c; Q, n) = n^2 \|c \odot Q' - Q\|_2^2 + 2\|c\|_2^2$. In summary, the $d_{\mathcal{X}}$-private Laplace mechanism, under ℓ_2^2-error function, can be described as follows:

1. Choose the model parameters (Q', c) by approximately solving the pre-processing optimization problem $F_{\text{pre-opt}}$ $\left(Q, n, d_{\mathcal{X}}(\cdot, \cdot), \ell_2^2\right)$ given by (6).
2. Release the response $\mathsf{Z} = \mathcal{M}_{\mathrm{Lap},c}(x, c \odot Q')$ given by (5).

Observe that, when $d_{\mathcal{X}}(i, j) = \epsilon, \forall i, j \in [N]$, the choices $c_i = c = \frac{\Delta_1^Q}{\epsilon}, \forall i \in [k]$ and $Q' = \frac{1}{c} Q$ satisfy the constraints of the optimization problem (6) under squared loss. In fact these choices correspond to the standard Laplace mechanism $\mathcal{M}_{\mathrm{Lap}, \frac{\Delta_1^q}{\epsilon} \cdot \mathbf{1}_k}$, and thus our framework is able to recover standard ϵ-differential privacy mechanisms as well.

The optimization problem (6) is in general non-convex, which is indeed hard to optimize. However, certain instances of this problem (instantiated by the utility function) allow efficient solutions in light of recent results. We discuss this in Appendix B. Also note that even if globally optimal solutions are infeasible to obtain, an approximate solution might still yield good utility in practice. We show this in Section 4.

3.2 $d_{\mathcal{X}}$-Private Exponential Mechanism

For a given utility function $u : \mathbb{N}^N \times \mathcal{R} \to \mathbb{R}$ over the histogram $x \in \mathbb{R}^N$, consider the following variant of exponential mechanism (with the model parameters $u' : \mathbb{N}^N \times \mathcal{R} \to \mathbb{R}$, and $c \in \mathbb{R}$ which will be chosen later based on the utility function):

Definition 9. *The mechanism* $\mathcal{M}_{\mathrm{Exp},c}(x, u')$ *selects and outputs an element* $r \in \mathcal{R}$ *with probability proportional to* $\exp\left(\frac{u'(x,r)}{2c}\right)$.

Here we note that for ease of presentation, we do not consider using $c_r \in \mathbb{R}$ for each $r \in \mathcal{R}$. The following theorem provides a sufficient condition for the above mechanism to satisfy $d_{\mathcal{X}}$-privacy.

Theorem 3. *If* $\Delta u'(i, j) \leq c d_{\mathcal{X}}(i, j), \forall i, j \in [N]$, *then the mechanism* $\mathcal{M}_{\mathrm{Exp},c}(\cdot, u')$ *satisfies* $d_{\mathcal{X}}$*-privacy.*

For a given histogram x and a given utility measure $u : \mathbb{R}^N \times \mathcal{R} \to \mathbb{R}$, let $\star_u(x) = \max_{r \in \mathcal{R}} u(x, r)$ denote the maximum utility score of any element $r \in \mathcal{R}$ with respect to histogram x. Below we generalize the Theorem 3.11 from [9]:

Theorem 4. *Fixing a database* x, *let* $\mathcal{R}_{\star_{u'}} = \{r \in \mathcal{R} : u'(x, r) = \star_{u'}(x)\}$ *denote the set of elements in* $\mathcal{R}$ *which attain utility score* $\star_{u'}(x)$. *Also define* $\delta_{u,u'} := \max_{x,r} |u(x, r) - u'(x, r)|$. *Then for* $\mathsf{Z} = \mathcal{M}_{\mathrm{Exp},c}(x, u')$, *with probability at least* $1 - e^{-t}$, *we have*

$$u(x, \mathsf{Z}) > \delta_{u,u'} + \star_{u'}(x) - 2c\left\{\ln\left(\frac{|\mathcal{R}|}{\left|\mathcal{R}_{\star_{u'}}\right|}\right) + t\right\}.$$

Since we always have $\left|\mathcal{R}_{\star_{u'}}\right| \geq 1$, *we get*

$$\mathbb{P}\left[u(x, \mathsf{Z}) \leq \delta_{u,u'} + \star_{u'}(x) - 2c\{\ln(|\mathcal{R}|) + t\}\right] \leq e^{-t}.$$

The proofs of the above two theorems are given in Appendix A.2. The exponential mechanism is a natural building block for designing complex ϵ-differentially private mechanisms. Next we consider two data release mechanisms (i.e., offline synthetic data generation mechanisms) which use the Laplace and/or the expoential mechanism as building blocks. These are the $d_{\mathcal{X}}$-private variants of the small database mechanism [5], and multiplicative weights exponential mechanism [14].

3.3 $d_{\mathcal{X}}$-Private Small Database Mechanism

Here we consider the problem of answering a large number of real valued linear queries $q : \mathbb{N}^N \to \mathbb{R}$ of the form $q(x) = \langle q, x \rangle$ (where $q \in \mathbb{R}^N$, and $x \in \mathbb{N}^N$) from class $\mathcal{Q}$ via synthetic histogram/database release. For this problem [5] have proposed and studied a simple ϵ-differentially private small database mechanism, which is an instantiation of exponential mechanism. They have used a utility function $u : \mathbb{N}^N \times \mathcal{R} \to \mathbb{R}$ (with $\mathcal{R} = \left\{ y \in \mathbb{N}^N : \|y\|_1 = \frac{\log|\mathcal{Q}|}{\alpha^2} \right\}$) defined as $u(x, y) := -\max_{q \in \mathcal{Q}} |q(x) - q(y)|$.

Now we extend the mechanism developed in [5] to obtain a $d_{\mathcal{X}}$-private version of it using the model parameters $\mathcal{Q}'$ and $c \in \mathbb{R}$ (which are determined later). Algorithm 1 is a modified version of Algorithm 4 from [9], where the transformation from $\mathcal{Q}$ to $\mathcal{Q}'$ is one-to-one (thus we have $|\mathcal{Q}'| = |\mathcal{Q}|$). When answering a query $q \in \mathcal{Q}$ over x, we need to output $cq'(y)$ where $q' \in \mathcal{Q}'$ is the matching element of q and y is the output of the $d_{\mathcal{X}}$-private small database mechanism (Algorithm 1). The following theorem provides the $d_{\mathcal{X}}$-privacy characterization of the small database mechanism.

Algorithm 1 `Small Database Mechanism` [5]: SmallDB$(x, \mathcal{Q}', c, \alpha)$

```
Let $\mathcal{R} \leftarrow \left\{ y \in \mathbb{N}^N : \|y\|_1 = \frac{\log|\mathcal{Q}'|}{\alpha^2} \right\}$
Let $u' : \mathbb{N}^N \times \mathcal{R} \to \mathbb{R}$ be defined to be:
    $u'(x, y) := -c \max_{q' \in \mathcal{Q}'} |q'(x) - q'(y)|.$    (7)
Sample And Output $y \in \mathcal{R}$ with the mechanism $\mathcal{M}_{\mathrm{Exp},c}(x, u')$
```

Theorem 5. *If $\left|q'_i - q'_j\right| \le d_{\mathcal{X}}(i, j), \forall i, j \in [N]$ and $\forall q' \in \mathcal{Q}'$, then the small database mechanism is $d_{\mathcal{X}}$-private.*

The following proposition and theorem characterize the performance of the $d_{\mathcal{X}}$-private small database mechanism.

Proposition 1 (Proposition 4.4, [9])**.** *Let $\mathcal{Q}$ be any class of linear queries. Let y be the database output by* SmallDB$(x, \mathcal{Q}', c, \alpha)$. *Then with probability $1 - \beta$:*

$$\max_{q \in \mathcal{Q}} \left|q(x) - cq'(y)\right| \le n \max_{q \in \mathcal{Q}} \left\|q - cq'\right\|_\infty + \alpha n + 2c \left\{ \frac{\log N \log|\mathcal{Q}|}{\alpha^2} + \log\left(\frac{1}{\beta}\right) \right\}.$$

Theorem 6 (Theorem 4.5, [9])**.** *By the appropriate choice of α, letting y be the database output by* SmallDB$\left(x, \mathcal{Q}', c, \frac{\alpha}{2}\right)$, *we can ensure that with probability $1 - \beta$:*

$$\max_{q \in \mathcal{Q}} \left|q(x) - cq'(y)\right| \le n \max_{q \in \mathcal{Q}} \left\|q - cq'\right\|_\infty + \left(cn^2\gamma\right)^{1/3},$$

where $\gamma = 16 \log N \log|\mathcal{Q}| + 4 \log\left(\frac{1}{\beta}\right)$. Equivalently, for any c such that $c \le \frac{\alpha^3 n}{\gamma}$ with probability $1 - \beta$: $\max_{q \in \mathcal{Q}} |q(x) - cq'(y)| \le n \max_{q \in \mathcal{Q}} \|q - cq'\|_\infty + \alpha n$.

Proofs of these claims are given in Appendix A.3. From the upper bound of the above theorem, the model parameters $\mathcal{Q}'$ and c of the small database mechanism can be chosen through the following pre-processing optimization problem:

$$\begin{aligned} \underset{\mathcal{Q}', c}{\text{minimize}} \quad & f\left(\mathcal{Q}', c; \mathcal{Q}, n\right) \\ \text{subject to} \quad & \left|q'_i - q'_j\right| \le d_{\mathcal{X}}(i, j), \forall i, j \in [N], q' \in \mathcal{Q}' \\ & c \ge 0, \end{aligned} \tag{8}$$

where $f(\mathcal{Q}', c; \mathcal{Q}, n) = n \max_{q \in \mathcal{Q}} \|q - cq'\|_\infty + \left(cn^2\gamma\right)^{1/3}$. Once again the optimization problem (8) is non-convex. See Appendix B, for a brief discussion on the (non-convex) pre-processing optimization problems (6), and (8).

3.4 $d_{\mathcal{X}}$-Private Multiplicative Weights Exponential Mechanism

As in the case of small database mechanism, here also we consider the problem of answering a large number of real valued linear queries in $d_{\mathcal{X}}$-private manner via synthetic histogram/database release. Algorithm 2 is a simple modification of Algorithm 1 from [14]. The following theorem provides the $d_{\mathcal{X}}$-privacy characterization of the MWEM mechanism.

Theorem 7. *If $\left|q'_i - q'_j\right| \le d_{\mathcal{X}}(i, j), \forall i, j \in [N]$ and $\forall q' \in \mathcal{Q}'$, then the MWEM mechanism is $d_{\mathcal{X}}$-private.*

The following theorem characterizes the performance of the MWEM mechanism.

Algorithm 2 `Multiplicative Weights Exponential Mechanism` [14]: MWEM$(x, \mathcal{Q}', c, T)$

```
Input: histogram $x$ over a universe $[N]$, set $\mathcal{Q}'$ of linear queries, privacy parameter $c > 0$, and number of iterations $T \in \mathbb{N}$.
Let $n$ denote $\|x\|_1$, the number of records in $x$. Let $y^0$ denote $n$ times the uniform distribution over $[N]$.
for $t = 1, ...T$ do
    1. Exponential Mechanism: Sample a query $q'_t \in \mathcal{Q}'$ using the $\mathcal{M}_{\text{Exp},2cT}(x, u'_t)$ mechanism and the score function $u'_t : \mathbb{N}^N \times \mathcal{Q}' \to \mathbb{R}$ given by
           $u'_t(x, q') := c \left| q'\left(y^{t-1}\right) - q'(x) \right|.$
    2. Laplace Mechanism: Let measurement $m_t = cq'_t(x) + \mathsf{Y}$ with $\mathsf{Y} \sim \text{Lap}(2cT)$.
    3. Multiplicative Weights: Let $y^t$ be $n$ times the distribution whose entries satisfy $\forall i \in [N]$,
           $y^t_i \propto y^{t-1}_i \times \exp\left(\left(q'_t\right)_i \times \left(m_t - q'_t\left(y^{t-1}\right)\right) / 2n\right).$
end for
Output $y = \text{avg}_{t<T} y^t$
```

Theorem 8 (Theorem 2.2, [14])**.** *For any dataset x, set of linear queries $\mathcal{Q}$, $T \in \mathbb{N}$, and $c > 0$, with probability at least $1 - 2T/|\mathcal{Q}|$, MWEM produces y such that*

$$\max_{q \in \mathcal{Q}} \left| cq'(y) - q(x) \right| \leq 2n\sqrt{\frac{\log N}{T}} + 10Tc\log|\mathcal{Q}| + n \max_{q \in \mathcal{Q}} \left\| cq' - q \right\|_\infty .$$

By setting $2n\sqrt{\frac{\log N}{T}} = 10Tc\log|\mathcal{Q}|$, we get

$$\max_{q \in \mathcal{Q}} \left| cq'(y) - q(x) \right| \leq n \max_{q \in \mathcal{Q}} \left\| cq' - q \right\|_\infty + \frac{20}{5^{2/3}} \left(n^2 \log N \log|\mathcal{Q}|\right)^{1/3} c^{1/3}.$$

Proofs of both theorems are given in Appendix A.4. The model parameters $\mathcal{Q}'$ and c of the MWEM mechanism can be chosen through the optimization problem 8 with

$$f\left(\mathcal{Q}', c; \mathcal{Q}, n\right) = n \max_{q \in \mathcal{Q}} \left\| cq' - q \right\|_\infty + \frac{20}{5^{2/3}} \left(n^2 \log N \log|\mathcal{Q}|\right)^{1/3} c^{1/3}.$$

4 Experiments

In this section, we experimentally evaluate the effectiveness of our framework on both synthetic and real data. We will show that in many situations, we can drastically improve the accuracy of the noisy answers compared to the traditional differentially private mechanisms. The datasets considered in these experiments are geographic in nature. More specifically, for the ensuing experiments, the data universes considered consist of points in Euclidean space which allow an intuitive Euclidean distance-based $d_{\mathcal{X}}$-metric. Under this metric, fine-grained location information is protected while larger regions provide better utility. As stated in [6], when dealing with geographic locations, it might be acceptable to disclose the region of an individual. On the other hand, disclosing the precise location (town) of the individual is less desirable. Thus it is useful to have a distinguishability level that depends on the geographic distance.

4.1 Single Linear Queries over Synthetic Data

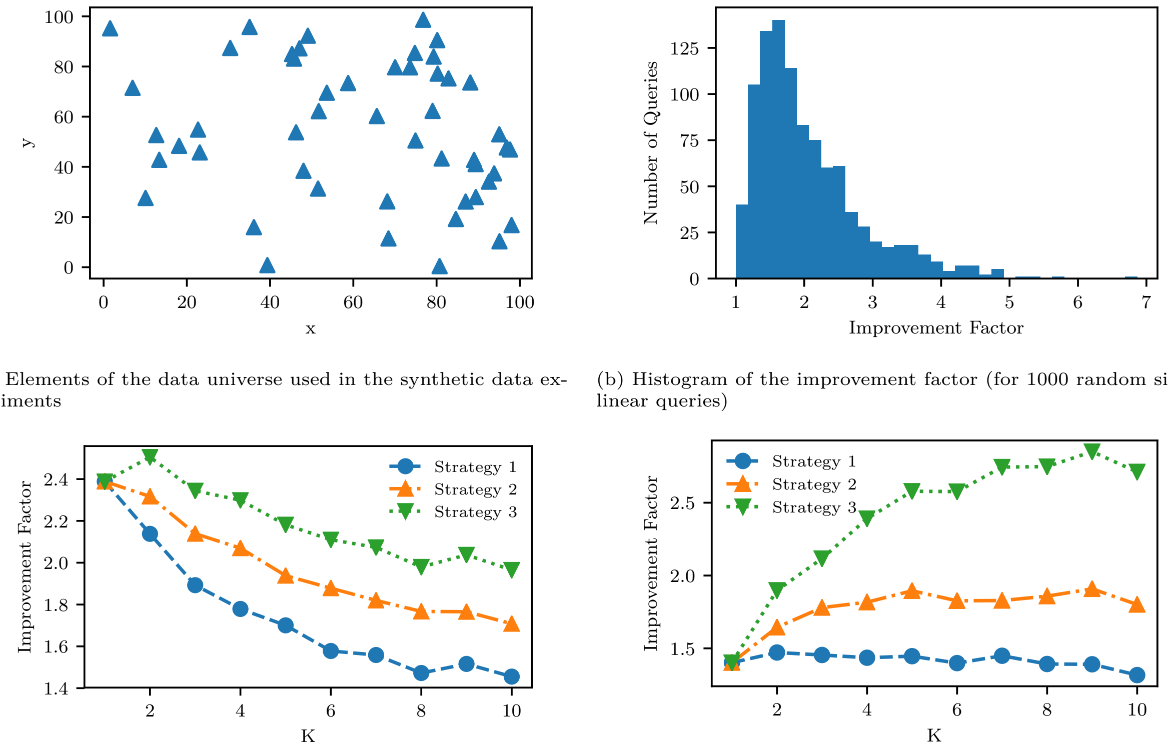
We first consider randomly generated single linear queries ($q : \mathbb{N}^N \to \mathbb{R}$), and compare the following two mechanisms: (a) the ϵ-differentially private Laplace mechanism (with $\epsilon = \min_{i,j} d_{\mathcal{X}}(i,j)$): $\mathcal{M}_{\text{Lap},\frac{\Delta_1^q}{\epsilon}}(x, q) = q(x) + \mathsf{Y}$, where $\mathsf{Y} \sim \text{Lap}\left(\frac{\Delta_1^q}{\epsilon}\right)$, and (b) the $d_{\mathcal{X}}$-private Laplace mechanism (with the model parameters $c \in \mathbb{R}$ and $q' \in \mathbb{R}^N$):

$$\mathcal{M}_{\text{Lap},c}\left(x, cq'\right) = cq'(x) + \mathsf{Y},$$

where $\mathsf{Y} \sim \text{Lap}(c)$, under the experimental setup given below.

Data and Privacy Metric: We generate a random dataset (histogram) with $n = 10,000$ records from a data universe of size $N = 50$. We then randomly sample N distinct two-dimensional points $\{(u_i, v_i)\}_{i=1}^N$ from the set $S = [0, 100] \times [0, 100] \subseteq \mathbb{R}^2$, and associate each point (u_i, v_i) with an element ($i \in [N]$) of the data universe. Note that this simulates geographic locations over a region, e.g., user locations in a city.[4] The sampled data universe elements are shown in Figure 2a. We

[4] Note that the data universe is fixed at $N = 50$ locations, each location exhibiting zero or more of the $n = 10,000$ records. Since this is a synthetic dataset, we choose a random data universe as well, by randomly sampling N locations. In practice, these N locations could be N hotspots in a city. Privacy is provided for the $n = 10,000$ subjects who can be in any of the N locations in the data universe, with higher privacy for nearby locations.



(a) Elements of the data universe used in the synthetic data experiments

(b) Histogram of the improvement factor (for 1000 random single linear queries)

(c) Improvement factor of multi-linear queries (random coefficients from real interval $[0, 1]$)

(d) Improvement factor of multi-linear queries (random binary coefficients from $\{0, 1\}$)

Fig. 2. Synthetic experiment with $N = 50$, and $d_{\mathcal{X}}$ metric defined based on Euclidean distance.

define the privacy metric $d_{\mathcal{X}} : [N] \times [N] \rightarrow \mathbb{R}$ based on the Euclidean distance (metric) on 2-dimensional space. Specifically, for any $i, j \in [N]$, define $d_{\mathcal{X}}(i, j) := \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}$.

Random Queries: We evaluate the two mechanisms over 1000 random single linear queries, where the query coefficients are randomly drawn from a uniform distribution over the real interval $[0, 1]$.

Performance Measure: We measure the individual performance of the mechanisms by the *root mean squared error* (RMSE; between the private response and the actual output) on the above generated data, *i.e.*, we consider the squared loss function $\ell(y, y') = \|y - y'\|_2^2$. Then the model parameters c and q' of the $d_{\mathcal{X}}$-private Laplace mechanism can be obtained by solving the following pre-processing optimization problem (for each query q):

$$\begin{aligned} \underset{c, q'}{\text{minimize}} \quad & f\left(c, q'\right) := n^2 \left\|cq' - q\right\|_2^2 + 2c^2 \\ \text{subject to} \quad & \left|q'_i - q'_j\right| \leq d_{\mathcal{X}}(i, j), \quad \forall i, j \in [N] \\ & c > 0. \end{aligned}$$

Since n is very large, by fixing $cq' = q$ in the above problem, we obtain an approximately optimal (closed form) solution given by $c = \max_{i,j} \frac{|q_i - q_j|}{d_{\mathcal{X}}(i,j)}$, and $q' = \frac{1}{c} q$.

Improvement Factor: We define another measure for cross-comparison of the two mechanisms. For a given single linear query q, the *improvement factor* of the $d_{\mathcal{X}}$-private Laplace mechanism compared to the baseline (ϵ-differentially private Laplace) mechanism is defined as $\mathrm{IF}(q) := \frac{\Delta_1^q / \epsilon}{c}$. This factor is simply the ratio between the scales (λ) of the noise $(\mathrm{Lap}(\lambda))$ added by these two mechanisms. Then for each random query (1000 in total), we compute the improvement factor. The resulting values are presented in a histogram form in Figure 2b, where the $d_{\mathcal{X}}$-private mechanism exhibits significant improvement in utility compared to the baseline mechanism. Notice that IF does not depend on ϵ, since ϵ is set to $\min_{i,j} d_{\mathcal{X}}(i, j)$, and c is set to be inversely related to $\min_{i,j} d_{\mathcal{X}}(i, j)$. Thus, ϵ effecitvely "cancels out" in the definition of IF. As a result, the IF results in Figure 2b hold for any $\epsilon > 0$. We checked this for multiple values of ϵ and obtained similar plots.

Algorithm 3 `Parameter Selection Algorithm:` PSA$(d_{\mathcal{X}}, Q)$

```
Input: privacy metric $d_{\mathcal{X}}$, and query matrix $Q \in \mathbb{R}^{K \times N}$.
Let $R = \mathbf{0}_K$, $T = \mathbf{1}_K$.
while $T \neq \mathbf{0}_K$ do
1.    $c'_k = \max_{i,j} \frac{|Q_{k,i} - Q_{k,j}|}{d_{\mathcal{X}}(i,j)}$, $\forall k \in [K]$. {near optimal scale of the noise for query $Q_{k,:}$, if the whole privacy budget is consumed by it.}
2.    $d^k_{\mathcal{X}}(i,j) = d_{\mathcal{X}}(i,j) \cdot \frac{\frac{1}{c'_k}|Q_{k,i} - Q_{k,j}|}{\sum_{l=1}^{K} \frac{1}{c'_l}|Q_{l,i} - Q_{l,j}|}$, $\forall k \in [K]$, $\forall i, j \in [N]$.{distribute the privacy budget between each query, based on $c'_k$'s.}
3.    $c_k = \max_{i,j} \frac{|Q_{k,i} - Q_{k,j}|}{d^k_{\mathcal{X}}(i,j)}$, $\forall k \in [K]$. {calculate the scale of the noise for each single linear query by considering the privacy budget allocated to them.}
4.    $d_{\mathcal{X}}(i,j) = d_{\mathcal{X}}(i,j) - \sum_{k=1}^{K} \frac{1}{c_k} \left|Q_{k,i} - Q_{k,j}\right|$, $\forall i, j \in [N]$. {calculate the remaining (total) privacy budget.}
5.    $T = \left[\frac{1}{c_1}, \dots, \frac{1}{c_K}\right]$, and $R = R + T$. {accumulate the share gained at this step.}
end while
Output: $c = \left[\frac{1}{R_1}, \dots, \frac{1}{R_K}\right]$
```

We note that IF is not a reasonable performance measure when the spread of the elements of the data universe is *profoundly* non-uniform (*e.g.*, two points are infinitesimally close to each other), in which case the traditional Laplace mechanism may get heavily penalized. But in both our real and synthetic data, the elements are (roughly) uniformly spread.

4.2 Multi-Linear Queries over Synthetic Data

Next we consider random multi-linear queries given by $Q \in \mathbb{R}^{K \times N}$, where we vary K from 1 to 10. We consider the same data, privacy metric, and performance measure (squared loss) used in Section 4.1. We consider two types of query matrices: the first type consists of matrices whose entries are drawn from a uniform distribution over the real interval $[0, 1]$, and the second type has matrices whose entries are random binary numbers, i.e., elements of the set $\{0, 1\}$.

Again, we compare the $d_{\mathcal{X}}$-private Laplace mechanism (5) with the ϵ-differentially private Laplace mechanism (1), with $\epsilon = \min_{i,j} d_{\mathcal{X}}(i,j)$. The model parameters $Q' \in \mathbb{R}^{K \times N}$ and $c \in \mathbb{R}^K$ of the $d_{\mathcal{X}}$-private Laplace mechanism (5) can be obtained from the optimization problem (6) with loss function $\ell(y, y') = \|y - y'\|_2^2$. Since n is considerably large, by imposing the constraint $c \odot Q' = Q$, the resulting optimization problem can be written as follows (for each query Q):

$$
\begin{aligned}
\underset{c}{\text{minimize}} \quad & \|c\|_2^2 = \sum_{k=1}^{K} c_k^2 \\
\text{subject to} \quad & \sum_{k=1}^{K} \frac{1}{c_k} \left|Q_{k,i} - Q_{k,j}\right| \leq d_{\mathcal{X}}(i,j), \forall i, j \in [N] \\
& c_k \geq 0, \forall k \in [K].
\end{aligned}
$$

In particular, we consider the following three different strategies to choose $c \in \mathbb{R}^K$ (with $Q'_{k,:} = \frac{1}{c_k} Q_{k,:}, \forall k \in [K]$), which satisfy the constraints of the optimization problem above:

1. *Strategy 1:* $c_k = \max_{i,j} \frac{|Q_{k,i} - Q_{k,j}|}{d_{\mathcal{X}}(i,j)/K}$, $\forall k \in [K]$, *i.e.*, we share the privacy budget equally ($\frac{d_{\mathcal{X}}(i,j)}{K}$) between the queries.
2. *Strategy 2:* $c_k = \max_{i,j} \frac{\|Q_{:,i} - Q_{:,j}\|_1}{d_{\mathcal{X}}(i,j)}$, $\forall k \in [K]$, *i.e.*, we add same scale noise to all the query response components.
3. *Strategy 3:* We obtain c via Algorithm 3, which distributes the budget between queries proportional to their privacy budget requirements.

For a given multi-linear query $Q \in \mathbb{R}^{K \times N}$, the *improvement factor* of the $d_{\mathcal{X}}$-private Laplace mechanism (5) compared to the baseline (ϵ-differentially private Laplace, (1)) mechanism is defined as IF$(Q) := \left\{\frac{\Delta_1^Q/\epsilon}{c_1} \cdot \frac{\Delta_1^Q/\epsilon}{c_2} \cdots \frac{\Delta_1^Q/\epsilon}{c_K}\right\}^{1/K}$, i.e., as a geometric mean of the individual improvement factors. For each $K \in [10]$, we randomly draw 100 query matrices $Q \in \mathbb{R}^{K \times N}$, and compute the (averaged) improvement factor IF(Q) for the above three different choices of c. The results are shown in Figure 2c and 2d. Some interesting insights are in order.

- Strategy 3 outperforms other strategies for both types of query matrices. This is understandable, as this strategy uses a smarter way of allocating budget between queries. More significantly, the strategy performs much better for the query matrix with binary coefficients (cf. Figure 2d). This is true since there is a high likelihood that two query coefficients are the same (i.e., $q_i = q_j$), resulting in no depletion of the privacy budget $d_{\mathcal{X}}(i,j)$.
- Strategy 1 has only marginal gain (IF$(Q) \leq 1.5$) for binary coefficient query matrices. This is because $|Q_{k,i} - Q_{k,j}| \leq 1$ for such matrices and therefore the

noise scale is essentially $c_k = K/\min_{i,j}(d_{\mathcal{X}}(i,j))$, when the query coefficients do not cancel each other out, i.e., $|Q_{k,i} - Q_{k,j}| \neq 0$. This is the same scale as the vanilla Laplace mechanism. The slight improvement is due to the cases where $|Q_{k,i} - Q_{k,j}| = 0$, which does not result in budget depletion in the case of $d_{\mathcal{X}}$-privacy.

4.3 Single Linear Queries over Real Data

Next we empirically evaluate $d_{\mathcal{X}}$-private Laplace mechanism for random single linear queries on a real-world geolocation dataset with longitude, latitude, and elevation attributes. The dataset is based on the United States Cities Database [23] which, among other attributes, contains the location (latitude and longitude) and population count of the cities in the United States (US). From this dataset, we extract the location and population count of cities with more than 50k inhabitants, resulting in a total of 741 cities. We further augment this dataset with elevation information by querying the Google Maps Elevation API [20] with the corresponding latitude and longitude values. We translate this dataset into a histogram over the cities (with $N = 741$). The 2D-locations (longitude and latitude wise) of the towns are presented in Figure 3a. We define the privacy budget $d_{\mathcal{X}}$ based on the Euclidean distance on this 2D-representation.

We generate and evaluate 1000 random linear queries over this dataset. The improvement factors of these queries are presented in Figure 3b. The average of the IF values lies between 2 to 3, with some queries showing an improvement factor of more than 7.5. We also wanted to test the improvement factor over queries with an obvious real-world interpretation. One such query is the "average elevation of a US resident's house" (the coefficients are simply the elevation of each city). In this case our algorithm performed particularly well, with an improvement factor of 202. This scale of improvement is due to the fact that there is a strong correlation between the query and the distance map: two nearby cities (*i.e.,* having strong privacy requirement) also have similar elevation. Note that even if the solutions shown above are sub-optimal, we still perform better than the baseline for synthetic data, and outperform it depending on the query structure and real data.

4.4 Experiments with Blowfish Privacy

In this section, we demonstrate that the $d_{\mathcal{X}}$-privacy notion can generalize some of the other alternative privacy notions as well, and hence our techniques can be applied to these other notions. This is true since our general pre-processing strategy applies to *any metric.* Thus, for instance, our techniques can be extended to the Blowfish [16] privacy (without constraints) notion as well. We can carefully define a $d_{\mathcal{X}}$ metric for any privacy policy considered in the Blowfish framework. First, we define $d_{\mathcal{X}}$ such that $d_{\mathcal{X}}(i,j) = \infty, \forall i \neq j$, and $d_{\mathcal{X}}(k,k) = 0, \forall k$. Then for each pair of neighbors (i,j), we check if there is a secret to be protected with the Blowfish policy. If so we just set $d_{\mathcal{X}}(i,j) = \epsilon$ (the privacy budget). Finally, we need to make sure (possibly by some transformations) that the resulting $d_{\mathcal{X}}$ satisfies the triangular inequality (a necessary condition for a distance metric).

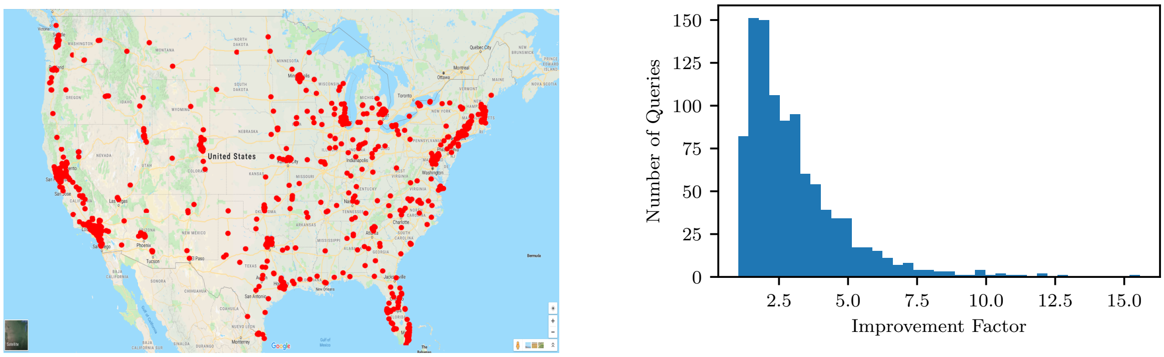
We consider the same data, single linear queries, mechanism ($d_{\mathcal{X}}$-private Laplace) and performance measure (squared loss) used in Section 4.1. But here we work with two different privacy metrics. Given a threshold T, and a privacy parameter ϵ, define:

1. $d_{\mathcal{X}}^{\mathrm{Blow}}$ s.t. $d_{\mathcal{X}}^{\mathrm{Blow}}(i,j) = \epsilon$ if $d_{\mathcal{X}}^{\mathrm{Euc}}(i,j) \leq T,$ and $d_{\mathcal{X}}^{\mathrm{Blow}}(i,j) = \infty$ otherwise
2. $d_{\mathcal{X}}^{\mathrm{Smooth}}$ s.t. $d_{\mathcal{X}}^{\mathrm{Smooth}}(i,j) = \epsilon$ if $d_{\mathcal{X}}^{\mathrm{Euc}}(i,j) \leq T,$ and $d_{\mathcal{X}}^{\mathrm{Smooth}}(i,j) = \frac{\epsilon d_{\mathcal{X}}^{\mathrm{Euc}}(i,j)}{T}$ otherwise,

where $d_{\mathcal{X}}^{\mathrm{Euc}}(i,j) := \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}$.

The first metric assigns privacy budget ϵ for any pair of points within distance T, and ∞ otherwise. The second metric "smoothly" increases the privacy budget proportional to the distance between the pair of points. Our base method for comparison is the ϵ-differentially private Laplace mechanism. First, we compute the average RMSE over 1000 random single linear queries under both privacy metrics defined above (for different values of ϵ and T). The results are shown in Figure 4. We can see that the results under both metrics are roughly the same. The higher the threshold T (*i.e.*, more neighbors are protected), the higher is the average error. The $d_{\mathcal{X}}^{\mathrm{Smooth}}$ metric behaves like the $d_{\mathcal{X}}^{\mathrm{Euc}}$ metric after the threshold value. Thus it induces tighter (and smoother) privacy than $d_{\mathcal{X}}^{\mathrm{Blow}}$, and results in a higher average error for the same threshold.

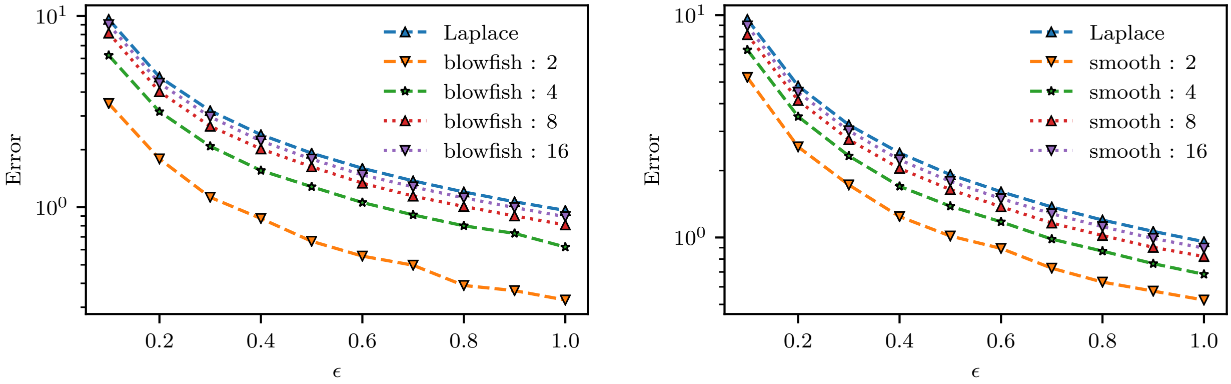
Then we fix $\epsilon = 1$, and for each random query (1000 in total), we compute the improvement factor. The resulted values are presented in a distribution form in Figure 5. Observe that for higher threshold values the distributions under both metrics are roughly similar, but



(a) US cities with more than 50k inhabitants

(b) Histogram of the improvement factor (for 1000 random single linear queries)

Fig. 3. Real-data (US cities [23]) experiment with $N = 741$, and $d_{\mathcal{X}}$ metric defined based on Euclidean distance.



(a) Error under $d_{\mathcal{X}}^{\text{Blow}}$-privacy

(b) Error under $d_{\mathcal{X}}^{\text{Smooth}}$-privacy

Fig. 4. Average RMSE (over 1000 random single linear queries) under different privacy metrics (with $N = 50$, and $T = 2, 4, 8, 16$).

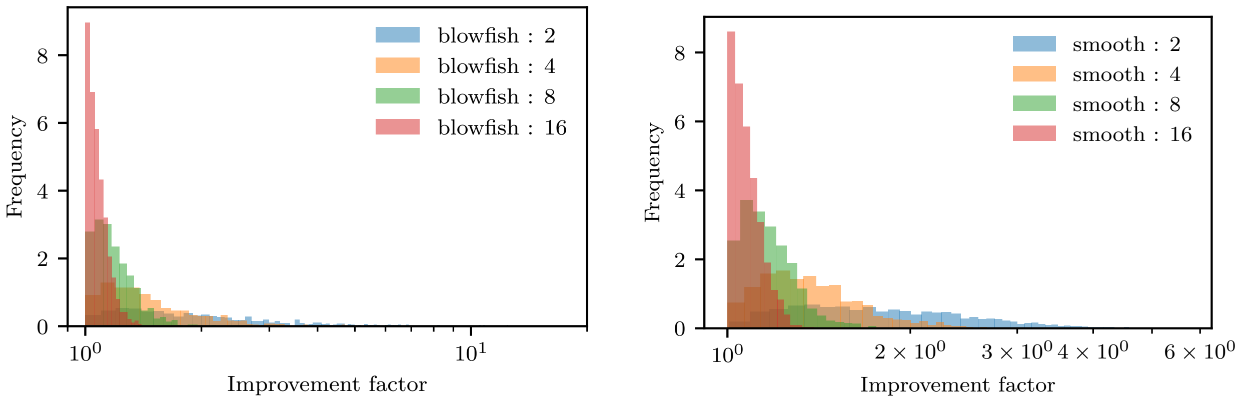
for lower threshold values (*e.g.* $T = 2$), the improvement factor under $d_{\mathcal{X}}^{\text{Blow}}$ is better than under $d_{\mathcal{X}}^{\text{Smooth}}$.

5 Discussion

Example $d_{\mathcal{X}}$-metric Instantiations: The main contribution of this paper is a meta procedure that converts an existing differentially private mechanism to its $d_{\mathcal{X}}$-private counterpart, given any metric $d_{\mathcal{X}}$. The interpretation of the privacy guarantees of the resulting mechanism is tied to how well the metric translates a given set of privacy requirements. Here we show some examples of appropriate $d_{\mathcal{X}}$-metrics for different privacy requirements.

Location Privacy: We have already presented some location privacy specific $d_{\mathcal{X}}$-metrics, i.e., the Euclidean distance based metric in Section 4.1 where nearby points are required to be more indistinguishable than distant points, and the distance threshold metrics in Section 4.4 (based on an example of sensitive information specification for the Blowfish framework [16]), which provides higher indistinguishability for points that are within a given distance threshold.

Heterogeneous Privacy for Tabular Data: Notably, location privacy is not the only application for $d_{\mathcal{X}}$-privacy. We have shown one such instance in Example 1 where the metric defines some attribute values as more sensitive than others. First, for binary datasets (each attribute having a cardinality of two), the metric in Example 1 can be generalized for any number of attributes. This does not generalize to attributes with more than



(a) Distribution of the Improvement Factor under $d_{\mathcal{X}}^{\text{Blow}}$-privacy

(b) Distribution of the Improvement Factor under $d_{\mathcal{X}}^{\text{Smooth}}$-privacy

Fig. 5. Distribution of the improvement factors (for 1000 random single linear queries) under different privacy metrics (with $N = 50$, $\epsilon = 1$, and $T = 2, 4, 8, 16$).

2 values, as the min function used in the metric does not satisfy the triangle inequality in such a case. An example metric, for the same privacy requirement (i.e., providing higher privacy to individuals having selected attribute values), can be defined as:

$$d_{\mathcal{X}}(i,j) = \sum_{k=1}^{d} \left(\epsilon(X_i^{(k)}) + \epsilon(X_j^{(k)}) \right) [\![X_i^{(k)} \neq X_j^{(k)}]\!].$$

where d is the total number of attributes in the dataset and $\epsilon(X_i^{(k)})$ defines the privacy budget allocated to the kth attribute value. One can then set $\epsilon(X_i^{(k)}) = \epsilon_k$ for all $i \in [N]$, where ϵ_k can be set to be lower for more sensitive attribute values. This closely resembles the metric $d_{\mathcal{X}}(i,j) = \sum_{k=1}^{d} \epsilon_k [\![X_i^{(k)} \neq X_j^{(k)}]\!]$ (discussed in the introduction) which allows to set sensitivity of the entire attribute via assigning the same ϵ_k for all values of the attribute. Note that this privacy metric relates to the notion of heterogeneous differential privacy [1] in which a user (owner of d items) chooses a separate privacy budget ϵ_k for its kth item. Similarly, if we require more privacy for some individuals in the dataset (modelled as elements in the histogram representation of the dataset), we can use the metric $d_{\mathcal{X}}(i,j) = (\epsilon_i + \epsilon_j)[\![i \neq j]\!]$, and assign lower privacy budgets for more sensitive elements.

Data Generalization: Several other examples of $d_{\mathcal{X}}$-metrics are given in [6]. One example is when the exact date of a particular event is considered sensitive, but releasing a slightly generalized date, say within a T-day period, might be appropriate. In this case the scaled metric $d_{\mathcal{X}}(i,j) = \epsilon \frac{|u_i - u_j|}{T}$ can be used, where u_i is the exact date (say, number of days since January 1, 2000) associated with the ith element in the data universe [6].

Privacy of Time-Series Data: Another example is protecting time-series data (e.g., smart energy data) where the privacy requirement is to only prevent fine-grained inference of the time-series. Here, an l_∞ norm based $d_{\mathcal{X}}$-metric is appropriate which is the maximum of the distances between each component of the time-series [6].

Privacy in Social Networks: Another natural metric based on a minimum spanning tree is given in [16]: vertices represent elements of the data universe, with edges between them having equal weights. Here the adversary may better distinguish points farther apart in the tree, than those that are closer. If some elements of the data universe are highly sensitive than others, non-uniform edge weights can capture the requirement. This metric is suitable for privacy in social networks.

A comprehensive treatment of privacy requirements and a suitable choice of $d_{\mathcal{X}}$-metric for each of them is beyond the scope of this work. The above examples show that $d_{\mathcal{X}}$-privacy can be used in many different applications. We stress however that the metric $d_{\mathcal{X}}$ must be appropriately defined to achieve meaningful privacy goals. A wrong choice of $d_{\mathcal{X}}$-metric may adversely impact privacy. For instance, if we replace the min function with the max function in Eq. 3 of Example 1, then even though the resulting function is still a metric, it does not satisfy the privacy requirement of providing more protection to more sensitive attributes. In particular, the query $q = (\texttt{MNA}, \texttt{MNB}, \texttt{FNA}, \texttt{FNB}) = (\texttt{N})$ will now be answered with noise of scale $c = 1/\epsilon_1$.

Correlated Data: If the database contains correlated data, it may be possible to infer about sensitive attributes even if the mechanism is $d_{\mathcal{X}}$-private. For instance, in Example 1, it may be known that males above 18 years of age are 90 percent more likely to be native. Then the answer to the query $(\mathtt{MA}) = (\mathtt{MYA}, \mathtt{MNA})$ will have less noise added to it as the query is non-sensitive (noise scale will be $c = 1/\epsilon_1$ in the example). Multiplying the answer by 0.9 gives us a much more accurate approximate number of native men above the age of 18, then what would have been possible through the query $(\mathtt{MYA})$ (noise scale $c = 1/\epsilon_0$). Protecting the answers from such correlations requires broadening the scope of $d_{\mathcal{X}}$-privacy to take such information as input, possibly in the form of constraints, as is done in the Blowfish privacy framework [16]. We note that susceptibility of $d_{\mathcal{X}}$-privacy under these *column-wise* correlations is similar to the case of differential privacy with correlated rows. Just like how differential privacy provides privacy for atypical rows (uncorrelated rows), $d_{\mathcal{X}}$-privacy guarantees privacy for atypical attribute values.

Unbounded Differential Privacy: In many instantiations of our meta procedure (4), we have used the bounded differential privacy model (in which the number of elements, i.e., n, in the dataset is public information). Our procedure can also be applied to unbounded differential privacy by spending some privacy budget to query database size n, similar to the conversion between the two flavours of differential privacy [22, p. 358]. For the Laplace mechanism (6), by setting $cq' = q$, we could get rid of the dependence on n, which makes it applicable to unbounded differential privacy as well. For SmallDB/MWEM, it's better to exploit the knowledge n, as we need to do the pre-processing only once for a given query set Q and dataset x.

6 Related Work

In [2] the notion of geo-indistinguishability is proposed which protects a user's exact location while allowing approximate information for location-based services. Some mechanisms to achieve privacy under this notion are also proposed which are variations of the Laplace mechanism for differential privacy. Geo-indistinguishability can be considered as an example of $d_{\mathcal{X}}$-privacy where the Euclidean metric within the discrete Cartesian plane is used as the data universe. Compared to [2], where only a few variations of the Laplace mechanism are given, we have proposed a general procedure to convert any differential privacy mechanism to its $d_{\mathcal{X}}$-privacy equivalent for linear queries. Furthermore, the focus of [2] is on location based services in the local model, whereas our work targets $d_{\mathcal{X}}$-private mechanisms for linear queries over histograms in the centralized model.

As mentioned earlier, the definition of $d_{\mathcal{X}}$-privacy is an instance of the notion of *generalized privacy* with a metric $d_{\mathcal{X}}$ which was proposed in [6] for the case of statistical databases (where each user's data is one row of the database). In addition to proposing the definition, the authors in [6] have only constructed *universally optimal mechanisms* [12][5] under some specific $d_{\mathcal{X}}$ metrics (such as Manhattan metric) for some particular class of queries such as count, sum, average, and percentage queries. In comparison, we propose a generic strategy to tailor any differentially private mechanism to satisfy $d_{\mathcal{X}}$-privacy for linear queries (which encompass a broad range of queries including the aforementioned).

Blowfish privacy [16] is a class of definitions that aims to strengthen differential privacy by the use of privacy policies that include a set of secrets (i.e., information deemed sensitive in the dataset, akin to what is modelled by the $d_{\mathcal{X}}$-metric) and a set of constraints that model an adversary's background knowledge or public knowledge about the dataset. There are some recent results on generalizing differentially private mechanisms to the Blowfish privacy equivalent under a given privacy policy [13]. In contrast to [13], we (a) consider any instance of the $d_{\mathcal{X}}$-metric (which covers Blowfish [16] privacy notion without constraints), and (b) preprocess the query alone (and not the input database) – thus we only need to do pre-processing once for a given data domain, i.e., not having to redo pre-processing for database changes. Currently, our proposed procedure applies to only a special case of the Blowfish that does not introduce deterministic constraints (modelling public knowledge), and extending our results to general Blowfish which deals with correlations is an interesting future direction.

The concept of *heterogeneous differential privacy* is proposed in [1] in the the user profile setting where a database itself is attributed to a single user. They consider the case where a user does not have homogeneous privacy requirements for all his/her items. We note that our meta procedure idea can be extended to the user profile setting as well. This extension would require slight modification in privacy and sensitivity def-

[5] Roughly, a mechanism is universally optimal if it provides the same utility to all users, regardless of their background information and (legal) loss function (modeling utility loss), as would a mechanism that is specifically tailored to each user.

initions and utility analysis. In particular, the metric $d_{\mathcal{X}}(u,v) = \sum_{i=1}^{d} \epsilon_i [\![u_i \neq v_i]\!]$ (that we discussed in the introduction and at the end of Appendix C) is closely related to the privacy definition in [1]. For linear queries, the stretching mechanism in [1] also transforms the original query vector q into q' (similar to our meta procedure), but their noise term is fixed and depends on the *global* sensitivity (in that sense our meta procedure is more general than theirs with noise parameter c). Moreover, the transformation $q \mapsto q'$ in [1] is not utility dependent. In the context of user profiles, detailed investigation of the connection between our (extended) meta procedure and the stretching mechanism is indeed an interesting future work. Similarly, [17] considers *personalized differential privacy* (PDP) where different users have different privacy expectations in the context of statistical databases. However, unlike $d_{\mathcal{X}}$-privacy, a user can only set the same privacy budget for all items (column-wise privacy). Moreover, a general *sampling* mechanism to convert any differential privacy mechanism to its PDP counterpart is also proposed in [17], which samples rows from the original dataset based on the privacy requirement of each user. This introduces an additional error term (due to sampling) [17]. In both these prior works, the utility measure of interest is not taken into consideration while distributing the privacy budget, whereas our meta procedure explicitly focuses on the utility measure.

7 Conclusion

In this paper, we developed new $d_{\mathcal{X}}$-private mechanisms for linear queries by extending the standard ϵ-differentially private mechanisms. These new mechanisms fully utilize the privacy budgets of different elements and maximize the utility of the private response. We have empirically shown that carefully selecting the model parameters of the $d_{\mathcal{X}}$-private mechanisms (depending on the utility function and $d_{\mathcal{X}}$-metric) can result in substantial improvement over the baseline mechanisms in terms of utility. Note that our analysis can be extended to advanced ϵ-differentially private mechanisms such as the Matrix [18], and K-norm [15] mechanisms. We leave it as future work. Finally, we would like to remark that for statistical queries (a special case of linear queries), which are (loosely) defined as the sum of predicates over the rows of the input dataset, we can design $d_{\mathcal{X}}$-private mechanisms more efficiently by exploiting the sum-structure. We refer the reader to Appendix C for more details.

8 Acknowledgment

We would like to thank the anonymous reviewers for their helpful comments and feedback. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. When the work was done, both Parameswaran Kamalaruban, and Victor Perrier were working at Data61, CSIRO.

References

[1] M. Alaggan, S. Gambs, and A. M. Kermarrec. Heterogeneous differential privacy. *arXiv preprint arXiv:1504.06998*, 2015.

[2] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. *Proceedings of the ACM SIGSAC conference on Computer & communications security*, pages 901–914, 2013.

[3] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. *Proceedings of the ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 273–282, 2007.

[4] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. *Proceedings of the ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 128–138, 2005.

[5] A. Blum, K. Ligett, and A. Roth. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):12, 2013.

[6] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi. Broadening the scope of differential privacy using metrics. *International Symposium on Privacy Enhancing Technologies Symposium*, pages 82–102, 2013.

[7] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. *Proceedings of the Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating Noise to Sensitivity in Private Data Analysis. *Proceedings of the Conference on Theory of Cryptography*, pages 265–284, 2006.

[9] C. Dwork, and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, pages 211–407, 2014.

[10] S. E. Fienberg, A. Rinaldo, and X. Yang. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *International Conference on Privacy in Statistical Databases*, pages 187–199, 2010.

[11] A. Ghosh, and A. Roth. Selling privacy at auction. *Games and Economic Behavior*, pages 334–346, 2015.

[12] A. Ghosh, T. Roughgarden, and M. Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, pages 1673–1693, 2012.

[13] S. Haney, A. Machanavajjhala, and B. Ding. Design of policy-aware differentially private algorithms. *Proceedings of the VLDB Endowment*, pages 264–275, 2015.

[14] M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. *Advances in Neural Information Processing Systems*, pages 2339–2347, 2012.

[15] M. Hardt, and K. Talwar. On the geometry of differential privacy. *Proceedings of the ACM symposium on Theory of computing*, pages 705–714, 2010.

[16] X. He, A. Machanavajjhala, and B. Ding. Blowfish privacy: Tuning privacy-utility trade-offs using policies. *Proceedings of the ACM SIGMOD international conference on Management of data*, pages 1447–1458, 2014.

[17] Z. Jorgensen, T. Yu, and G. Cormode. Conservative or liberal? Personalized differential privacy. *International Conference on Data Engineering*, pages 1023–1034, 2015.

[18] C. Li, and G. Miklau. Efficient batch query answering under differential privacy. *arXiv preprint arXiv:1103.1367*, 2011.

[19] C. Li, and G. Miklau. An adaptive mechanism for accurate query answering under differential privacy. *Proceedings of the VLDB Endowment*, pages 514–525, 2012.

[20] Google Maps. Google Elevation API. https://developers.google.com/maps/documentation/elevation/intro, 2018.

[21] F. McSherry, and K. Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science*, pages 94–103, 2007.

[22] V. Vadhan. The complexity of Differential Privacy. *Tutorials on the Foundations of Cryptography*, pages 347–450, Springer, 2017.

[23] SimpleMaps. United States Cities Database. https://simplemaps.com/data/us-cities, 2018.

[24] S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, pages 3–34, 2015.

[25] Y. Xu, and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, pages 1758–1789, 2013.

A Proofs

A.1 Laplace Mechanism

Theorem 1. If $\Delta_1^{q'}(i,j) \leq d_{\mathcal{X}}(i,j)$, $\forall i,j \in [N]$, then the mechanism $\mathcal{M}_{\mathrm{Lap},c}(\cdot, c \odot q')$ given by (5) satisfies $d_{\mathcal{X}}$-privacy.

Proof. Let $x, x' \in \mathbb{R}^N$ s.t. $\|x - x'\|_1 \leq 2$, $x_i \neq x'_i$, and $x_j \neq x'_j$, and let $q \in \mathcal{Q}$. Let p_x and $p_{x'}$ denote the probability density functions of $\mathcal{M}_{\mathrm{Lap},c}(x,q)$ and $\mathcal{M}_{\mathrm{Lap},c}(x',q)$ respectively. Then for any $z \in \mathcal{Y}$ we have

$$\begin{aligned}
\frac{p_x(z)}{p_{x'}(z)} &= \Pi_{i=1}^k \left(\frac{\exp\left(-\frac{|c_i q'(x)_i - z_i|}{c_i}\right)}{\exp\left(-\frac{|c_i q'(x')_i - z_i|}{c_i}\right)} \right) \\
&= \Pi_{i=1}^k \exp\left(\frac{|c_i q'(x')_i - z_i| - |c_i q'(x)_i - z_i|}{c_i} \right) \\
&\overset{(i)}{\leq} \Pi_{i=1}^k \exp\left(\frac{|c_i q'(x')_i - c_i q'(x)_i|}{c_i} \right) \\
&= \exp\left(\left\| q'(x) - q'(x') \right\|_1 \right) \\
&\overset{(ii)}{\leq} \exp\left(\Delta_1^{q'}(i,j) \right) \\
&\leq \exp\left(d_{\mathcal{X}}(i,j) \right),
\end{aligned}$$

where (i) follows from the triangle inequality and (ii) follows from the definition of generalized global sensitivity and due to the choice of x and x'. That $\frac{p_x(z)}{p_{x'}(z)} \geq \exp(-d_{\mathcal{X}}(i,j))$, follows by symmetry. □

Theorem 2. Let $Q : \mathbb{R}^N \to \mathbb{R}^k$ be a multi-linear query of the form $Q(x) = Qx$, and let $\mathsf{Z} = \mathcal{M}_{\mathrm{Lap},c}(x, c \odot Q') = c \odot Q'x + \mathsf{Y}$ with $\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)$.

1. When $\ell_2^2(y,y') = \|y - y'\|_2^2$, we have

$$\mathrm{err}_{\ell_2^2}\left(\mathcal{M}_{\mathrm{Lap},c}\left(\cdot, c \odot Q'\right), Q\right) \leq 2n^2 \left\| c \odot Q' - Q \right\|_2^2 + 4\|c\|_2^2,$$

where $\mathrm{err}_\ell(\mathcal{M}, Q)$ is defined in (2).

2. When $\ell_p(y,y') = \|y - y'\|_p$, we have

$$\mathrm{err}_{\ell_p}\left(\mathcal{M}_{\mathrm{Lap},c}\left(\cdot, c \odot Q'\right), Q\right) \leq n \left\| c \odot Q' - Q \right\|_p + \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\|\mathsf{Y}\|_p \right].$$

Note that $\mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\|\mathsf{Y}\|_1 \right] = \|c\|_1$.

3. $\forall \delta \in (0,1]$, with probability at least $1 - \delta$ we have

$$\|Qx - \mathsf{Z}\|_\infty \leq n \left\| c \odot Q' - Q \right\|_\infty + \ln\left(\frac{k}{\delta}\right) \cdot \|c\|_\infty .$$

Proof. **Part 1.** Consider

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathsf{Z}} \left[\ell\left(\mathsf{Z}, Q(x)\right) \right] \\
&= \mathop{\mathbb{E}}_{\mathsf{Z}} \left[\|\mathsf{Z} - Qx\|_2^2 \right] \\
&= \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\left\| c \odot Q'x + \mathsf{Y} - Qx \right\|_2^2 \right] \\
&\overset{(i)}{\leq} \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\left(\left\| c \odot Q'x - Qx \right\|_2 + \|\mathsf{Y}\|_2 \right)^2 \right] \\
&\overset{(ii)}{\leq} 2 \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\left\| c \odot Q'x - Qx \right\|_2^2 + \|\mathsf{Y}\|_2^2 \right] \\
&= 2 \left\{ \left\| c \odot Q'x - Qx \right\|_2^2 + \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\|\mathsf{Y}\|_2^2 \right] \right\} \\
&= 2 \left\{ \sum_{i=1}^k \left| \left\langle c_i Q'_{i,:} - Q_{i,:}, x \right\rangle \right|^2 + \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\|\mathsf{Y}\|_2^2 \right] \right\} \\
&\overset{(iii)}{\leq} 2 \left\{ \sum_{i=1}^k \left\| c_i Q'_{i,:} - Q_{i,:} \right\|_2^2 \|x\|_2^2 + \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\|\mathsf{Y}\|_2^2 \right] \right\} \\
&= 2 \left\{ \|x\|_2^2 \sum_{i=1}^k \left\| c_i Q'_{i,:} - Q_{i,:} \right\|_2^2 + \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\|\mathsf{Y}\|_2^2 \right] \right\} \\
&\overset{(iv)}{\leq} 2 \left\{ n^2 \sum_{i=1}^k \left\| c_i Q'_{i,:} - Q_{i,:} \right\|_2^2 + \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\|\mathsf{Y}\|_2^2 \right] \right\} \\
&= 2 \left\{ n^2 \left\| c \odot Q' - Q \right\|_2^2 + \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\|\mathsf{Y}\|_2^2 \right] \right\} \\
&\overset{(v)}{=} 2 \left\{ n^2 \left\| c \odot Q' - Q \right\|_2^2 + 2\|c\|_2^2 \right\}
\end{aligned}$$

where (i) is by triangle inequality, (ii) is due to the fact that $(a+b)^2 \leq 2a^2 + 2b^2$, (iii) is by Hölder's Inequality, (iv) is due to the fact that $\|x\|_2 \leq \|x\|_1 = n$, and (v) is due to the fact that $\mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\|\mathsf{Y}\|_2^2 \right] = 2\|c\|_2^2$ (since $\mathop{\mathbb{E}}_{\mathsf{X} \sim \mathrm{Lap}(\lambda)} \left[\mathsf{X}^2 \right] = 2\lambda^2$ for $\mathsf{X} \in \mathbb{R}$). This completes the proof of first part.

Part 2. Consider (by the similar reasoning as of Part 1)

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathsf{Z}} \left[\ell\left(\mathsf{Z}, Q(x)\right) \right] \\
&= \mathop{\mathbb{E}}_{\mathsf{Z}} \left[\|\mathsf{Z} - Qx\|_p \right] \\
&= \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\left\| c \odot Q'x + \mathsf{Y} - Qx \right\|_p \right] \\
&\leq \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\left\| c \odot Q'x - Qx \right\|_p + \|\mathsf{Y}\|_p \right] \\
&= \left\| c \odot Q'x - Qx \right\|_p + \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\|\mathsf{Y}\|_p \right] \\
&= \left\| c \odot Q'x - Qx \right\|_p + \mathop{\mathbb{E}}_{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)} \left[\|\mathsf{Y}\|_p \right]
\end{aligned}$$

$$
\begin{aligned}
&= \left(\sum_{i=1}^{k} \left|\left\langle c_i Q'_{i,:} - Q_{i,:}, x\right\rangle\right|^p\right)^{1/p} + \underset{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)}{\mathbb{E}} \left[\|\mathsf{Y}\|_p\right] \\
&\leq \left(\sum_{i=1}^{k} \left\|c_i Q'_{i,:} - Q_{i,:}\right\|_p^p \|x\|_q^p\right)^{1/p} + \underset{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)}{\mathbb{E}} \left[\|\mathsf{Y}\|_p\right] \\
&= \|x\|_q \left(\sum_{i=1}^{k} \left\|c_i Q'_{i,:} - Q_{i,:}\right\|_p^p\right)^{1/p} + \underset{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)}{\mathbb{E}} \left[\|\mathsf{Y}\|_p\right] \\
&\leq n \left(\sum_{i=1}^{k} \left\|c_i Q'_{i,:} - Q_{i,:}\right\|_p^p\right)^{1/p} + \underset{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)}{\mathbb{E}} \left[\|\mathsf{Y}\|_p\right] \\
&= n \left\|c \odot Q' - Q\right\|_p + \underset{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)}{\mathbb{E}} \left[\|\mathsf{Y}\|_p\right].
\end{aligned}
$$

Note that $\underset{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)}{\mathbb{E}} \left[\|\mathsf{Y}\|_1\right] = \|c\|_1$ (since $\underset{\mathsf{X} \sim \mathrm{Lap}(\lambda)}{\mathbb{E}} \left[|\mathsf{X}|\right] = \lambda$ for $\mathsf{X} \in \mathbb{R}$).

Part 3. We will use the fact that if $\mathsf{Y} \sim \mathrm{Lap}(b)$, then $\mathbb{P}\left[|\mathsf{Y}| \geq t \cdot b\right] = \exp(-t)$. We have:

$$
\begin{aligned}
&\mathbb{P}\left[\left\|c \odot Q'x - \mathsf{Z}\right\|_\infty \geq \ln\left(\frac{k}{\delta}\right) \cdot \|c\|_\infty\right] \\
&= \mathbb{P}\left[\max_{i \in [k]} |\mathsf{Y}_i| \geq \ln\left(\frac{k}{\delta}\right) \cdot \|c\|_\infty\right] \\
&\leq k \cdot \mathbb{P}\left[|\mathsf{Y}_i| \geq \ln\left(\frac{k}{\delta}\right) \cdot \|c\|_\infty\right] \\
&\leq k \cdot \mathbb{P}\left[|\mathsf{Y}_i| \geq \ln\left(\frac{k}{\delta}\right) \cdot c_i\right] \\
&= k \cdot \left(\frac{\delta}{k}\right) \\
&= \delta
\end{aligned}
$$

where the first inequality is due to union bound, and the second to last equality follows from the fact that each $\mathsf{Y}_i \sim \mathrm{Lap}(c_i)$. That is with probability at least $1 - \delta$ we have

$$
\begin{aligned}
&\|Qx - \mathsf{Z}\|_\infty \\
&\leq \left\|c \odot Q'x - Qx\right\|_\infty + \ln\left(\frac{k}{\delta}\right) \cdot \|c\|_\infty \\
&= \max_{i \in [k]} \left|\left\langle c_i Q'_{i,:} - Q_{i,:}, x\right\rangle\right| + \ln\left(\frac{k}{\delta}\right) \cdot \|c\|_\infty \\
&\leq \max_{i \in [k]} \left\|c_i Q'_{i,:} - Q_{i,:}\right\|_\infty \|x\|_1 + \ln\left(\frac{k}{\delta}\right) \cdot \|c\|_\infty \\
&= n \max_{i \in [k]} \left\|c_i Q'_{i,:} - Q_{i,:}\right\|_\infty + \ln\left(\frac{k}{\delta}\right) \cdot \|c\|_\infty \\
&= n \left\|c \odot Q' - Q\right\|_\infty + \ln\left(\frac{k}{\delta}\right) \cdot \|c\|_\infty.
\end{aligned}
$$

□

A.2 Exponential Mechanism

Theorem 3. If $\Delta u'(i,j) \leq c d_{\mathcal{X}}(i,j), \forall i,j \in [N]$, then the mechanism $\mathcal{M}_{\mathrm{Exp},c}(\cdot, u')$ satisfies the $d_{\mathcal{X}}$-privacy.

Proof. For clarity, we assume $\mathcal{R}$ to be finite. Let $x, x' \in \mathbb{R}^N$ s.t. $\|x - x'\|_1 \leq 2$, $x_i \neq x'_i$ and $x_j \neq x'_j$. Then for any $r \in \mathcal{R}$ we have

$$
\begin{aligned}
&\frac{\mathbb{P}\left[\mathcal{M}_{\mathrm{Exp},c}(x,u') = r\right]}{\mathbb{P}\left[\mathcal{M}_{\mathrm{Exp},c}(x',u') = r\right]} \\
&= \frac{\left(\frac{\exp\left(\frac{u'(x,r)}{2c}\right)}{\sum_{r' \in \mathcal{R}} \exp\left(\frac{u'(x,r')}{2c}\right)}\right)}{\left(\frac{\exp\left(\frac{u'(x',r)}{2c}\right)}{\sum_{r' \in \mathcal{R}} \exp\left(\frac{u'(x',r')}{2c}\right)}\right)} \\
&= \frac{\exp\left(\frac{u'(x,r)}{2c}\right)}{\exp\left(\frac{u'(x',r)}{2c}\right)} \cdot \frac{\sum_{r' \in \mathcal{R}} \exp\left(\frac{u'(x',r')}{2c}\right)}{\sum_{r' \in \mathcal{R}} \exp\left(\frac{u'(x,r')}{2c}\right)} \\
&= \exp\left(\frac{u'(x,r) - u'(x',r)}{2c}\right) \cdot \frac{\sum_{r' \in \mathcal{R}} \exp\left(\frac{u'(x',r')}{2c}\right)}{\sum_{r' \in \mathcal{R}} \exp\left(\frac{u'(x,r')}{2c}\right)} \\
&\leq \exp\left(\frac{\Delta u'(i,j)}{2c}\right) \cdot \frac{\sum_{r' \in \mathcal{R}} \exp\left(\frac{u'(x,r') + \Delta u'(i,j)}{2c}\right)}{\sum_{r' \in \mathcal{R}} \exp\left(\frac{u'(x,r')}{2c}\right)} \\
&= \exp\left(\frac{\Delta u'(i,j)}{2c}\right) \cdot \exp\left(\frac{\Delta u'(i,j)}{2c}\right) \cdot (1) \\
&= \exp\left(d_{\mathcal{X}}(i,j)\right).
\end{aligned}
$$

Similarly, $\frac{\mathbb{P}\left[\mathcal{M}_{\mathrm{Exp},c}(x,u') = r\right]}{\mathbb{P}\left[\mathcal{M}_{\mathrm{Exp},c}(x',u') = r\right]} \geq \exp\left(-d_{\mathcal{X}}(i,j)\right)$ by symmetry. □

Theorem 4. Fixing a database x, let

$$
\mathcal{R}_{\star_{u'}} = \left\{r \in \mathcal{R} : u'(x,r) = \star_{u'}(x)\right\}
$$

denote the set of elements in $\mathcal{R}$ which attain utility score $\star_{u'}(x)$. Also define $\delta_{u,u'} := \max_{x,r} |u(x,r) - u'(x,r)|$. Then for $\mathsf{Z} = \mathcal{M}_{\mathrm{Exp},c}(x,u')$, we have

$$
\mathbb{P}\left[u(x,\mathsf{Z}) \leq \delta_{u,u'} + \star_{u'}(x) - 2c\left\{\ln\left(\frac{|\mathcal{R}|}{\left|\mathcal{R}_{\star_{u'}}\right|}\right) + t\right\}\right] \leq e^{-t}.
$$

Since we always have $\left|\mathcal{R}_{\star_{u'}}\right| \geq 1$, we get

$$
\mathbb{P}\left[u(x,\mathsf{Z}) \leq \delta_{u,u'} + \star_{u'}(x) - 2c\left\{\ln(|\mathcal{R}|) + t\right\}\right] \leq e^{-t}.
$$

Proof.

$$
\mathbb{P}\left[u'(x,\mathsf{Z}) \leq \alpha\right] \leq \frac{|\mathcal{R}| \exp(\alpha/2c)}{\left|\mathcal{R}_{\star_{u'}}\right| \exp\left(\star_{u'}(x)/2c\right)}
$$

$$= \frac{|\mathcal{R}|}{\left|\mathcal{R}_{\star_{u'}}\right|} \exp\left(\frac{\alpha - \star_{u'}(x)}{2c}\right).$$

The inequality follows from the observation that each $r \in \mathcal{R}$ with $u'(x,r) \leq \alpha$ has un-normalized probability mass at most $\exp(\alpha/2c)$, and hence the entire set of such "bad" elements r has total un-normalized probability mass at most $|\mathcal{R}|\exp(\alpha/2c)$. In contrast, we know that there exist at least $\left|\mathcal{R}_{\star_{u'}}\right| \geq 1$ elements with $u'(x,r) = \star_{u'}(x)$, and hence un-normalized probability mass $\left|\mathcal{R}_{\star_{u'}}\right| \exp(\star_{u'}(x)/2c)$, and so this is a lower bound on the normalization term. The proof is completed by plugging in the appropriate value for α, and by noting that

$$\begin{aligned} u(x,r) &\leq u'(x,r) + \left|u(x,r) - u'(x,r)\right| \\ &\leq u'(x,r) + \max_{x,r}\left|u(x,r) - u'(x,r)\right|. \end{aligned}$$

□

A.3 Small Database Mechanism

Theorem 5. If $\left|q'_i - q'_j\right| \leq d_{\mathcal{X}}(i,j), \forall i,j \in [N]$ and $\forall q' \in \mathcal{Q}'$, then the small database mechanism is $d_{\mathcal{X}}$-private.

Proof. First we will find the condition for $\Delta u'(i,j) \leq c d_{\mathcal{X}}(i,j), \forall i,j \in [N]$ and $\forall q' \in \mathcal{Q}'$:

$$\Delta u'(i,j) = \max_{y \in \mathcal{R}} \max_{\substack{x,x' \in \mathbb{R}^N : \left\|x - x'\right\|_1 \leq 2, \\ x_i \neq x'_i, x_j \neq x'_j \text{ for } i,j \in [N]}} \left|u'(x,y) - u'\left(x',y\right)\right|.$$

For some $x, x' \in \mathbb{N}^N$ such that $\|x - x'\|_1 \leq 2, x_i \neq x'_i, x_j \neq x'_j$ for some $i,j \in [N]$, we have:

$$\begin{aligned} &\left|u'(x,y) - u'\left(x',y\right)\right| \\ &= \left|c \max_{q' \in \mathcal{Q}'}\left|q'\left(x'\right) - q'(y)\right| - c\max_{q' \in \mathcal{Q}'}\left|q'(x) - q'(y)\right|\right| \\ &\overset{(i)}{\leq} c \max_{q' \in \mathcal{Q}'}\left|\left\{\left|q'\left(x'\right) - q'(y)\right| - \left|q'(x) - q'(y)\right|\right\}\right| \\ &\overset{(ii)}{\leq} c \max_{q' \in \mathcal{Q}'}\left|q'(x) - q'\left(x'\right)\right| \\ &= c\max_{q' \in \mathcal{Q}'}\left|\left\langle q', x - x'\right\rangle\right| \\ &\overset{(iii)}{\leq} c \max_{q' \in \mathcal{Q}'}\left|q'_i - q'_j\right|, \end{aligned}$$

where (i) due to the fact that $|\max_x |a(x)| - \max_x |b(x)|| \leq \max_x |\{|a(x)| - |b(x)|\}|$, (ii) is by triangle inequality, and (iii) is due to the choice of x and x'. Thus we require

$$\Delta u'(i,j) \leq c \max_{q' \in \mathcal{Q}'}\left|q'_i - q'_j\right| \leq c d_{\mathcal{X}}(i,j).$$

The Small Database mechanism is simply an instantiation of the $\mathcal{M}_{\mathrm{Exp},c}(\cdot, u')$ mechanism. Therefore, privacy follows from Theorem 3. □

We use the following theorem from [9] directly.

Theorem 9 (Theorem 4.2, [9])**.** *For any finite class of linear queries $\mathcal{Q}'$, if $\mathcal{R} = \left\{y \in \mathbb{N}^N : \|y\|_1 = \frac{\log|\mathcal{Q}'|}{\alpha^2}\right\}$ then for all $x \in \mathbb{N}^N$, there exists a $y \in \mathcal{R}$ such that:*

$$\max_{q' \in \mathcal{Q}'}\left|cq'(x) - cq'(y)\right| \leq \alpha n.$$

Proposition 1. Let $\mathcal{Q}$ be any class of linear queries. Let y be the database output by SmallDB $(x, \mathcal{Q}', c, \alpha)$. Then with probability $1 - \beta$:

$$\begin{aligned} \max_{q \in \mathcal{Q}}\left|q(x) - cq'(y)\right| \leq &n \max_{q \in \mathcal{Q}}\left\|q - cq'\right\|_\infty + \alpha n \\ &+ 2c\left\{\frac{\log N \log|\mathcal{Q}|}{\alpha^2} + \log\left(\frac{1}{\beta}\right)\right\}. \end{aligned}$$

Proof. Applying the utility bounds for the $\mathcal{M}_{\mathrm{Exp},c}(\cdot, u')$ mechanism (Theorem 4) with $-\star_{u'}(x) \geq \alpha n$ (which follows from Theorem 9), we find:

$$\mathbb{P}\left[\max_{q' \in \mathcal{Q}'}\left|cq'(x) - cq'(y)\right| \geq \alpha n + 2c\left\{\ln(|\mathcal{R}|) + t\right\}\right] \leq e^{-t}.$$

By noting that $\mathcal{R}$, which is the set of all databases of size at most $\log|\mathcal{Q}|/\alpha^2$ (since $|\mathcal{Q}'| = |\mathcal{Q}|$), satisfies $|\mathcal{R}| \leq |\mathcal{X}|^{\log|\mathcal{Q}|/\alpha^2}$ and by setting $t = \log\left(\frac{1}{\beta}\right)$, we get with probability $1 - \beta$:

$$\begin{aligned} &\max_{q' \in \mathcal{Q}'}\left|cq'(x) - cq'(y)\right| \\ &\leq \alpha n + 2c\left\{\frac{\log N \log|\mathcal{Q}|}{\alpha^2} + \log\left(\frac{1}{\beta}\right)\right\}. \end{aligned}$$

Thus with probability $1 - \beta$ we have ($q' \in \mathcal{Q}'$ is the one-to-one mapping of $q \in \mathcal{Q}$):

$$\begin{aligned} &\max_{q \in \mathcal{Q}}\left|q(x) - cq'(y)\right| \\ &\overset{(i)}{\leq} \max_{q \in \mathcal{Q}}\left\{\left|q(x) - cq'(x)\right| + \left|cq'(x) - cq'(y)\right|\right\} \\ &\overset{(ii)}{\leq} \max_{q \in \mathcal{Q}}\left|q(x) - cq'(x)\right| + \max_{q \in \mathcal{Q}}\left|cq'(x) - cq'(y)\right| \\ &= \max_{q \in \mathcal{Q}}\left|\left\langle q - cq', x\right\rangle\right| + \max_{q \in \mathcal{Q}}\left|cq'(x) - cq'(y)\right| \\ &\overset{(iii)}{\leq} \|x\|_1 \max_{q \in \mathcal{Q}}\left\|q - cq'\right\|_\infty + \max_{q \in \mathcal{Q}}\left|cq'(x) - cq'(y)\right| \\ &\overset{(iv)}{=} n\max_{q \in \mathcal{Q}}\left\|q - cq'\right\|_\infty + \max_{q' \in \mathcal{Q}'}\left|cq'(x) - cq'(y)\right| \\ &\leq n\max_{q \in \mathcal{Q}}\left\|q - cq'\right\|_\infty + \alpha n \end{aligned}$$

$$+ 2c \left\{ \frac{\log N \log |\mathcal{Q}|}{\alpha^2} + \log\left(\frac{1}{\beta}\right) \right\},$$

where (i) is by triangle inequality, (ii) is by the fact that $\max_x \{a(x) + b(x)\} \leq \max_x a(x) + \max_x b(x)$, (iii) is by the Hölder's Inequality, and (iv) is by the fact that $\|x\|_1 = n$. □

Theorem 6. By the appropriate choice of α, letting y be the database output by SmallDB $\left(x, \mathcal{Q}', c, \frac{\alpha}{2}\right)$, we can ensure that with probability $1 - \beta$:

$$\begin{aligned} &\max_{q \in \mathcal{Q}} \left| q\left(x\right) - cq'\left(y\right) \right| \\ &\leq n \max_{q \in \mathcal{Q}} \left\| q - cq' \right\|_\infty + \left(cn^2 \gamma\right)^{1/3}, \end{aligned} \tag{9}$$

where $\gamma = 16 \log N \log |\mathcal{Q}| + 4 \log\left(\frac{1}{\beta}\right)$. Equivalently, for any c such that

$$c \ \leq \ \frac{\alpha^3 n}{\gamma} \tag{10}$$

with probability $1 - \beta$: $\max_{q \in \mathcal{Q}} |q(x) - cq'(y)| \leq n \max_{q \in \mathcal{Q}} \|q - cq'\|_\infty + \alpha n$.

Proof. By Proposition 1, we get:

$$\begin{aligned} &\max_{q \in \mathcal{Q}} \left| q\left(x\right) - cq'\left(y\right) \right| \\ &\leq n \max_{q \in \mathcal{Q}} \left\| q - cq' \right\|_\infty + \frac{\alpha}{2} n \\ &\quad + 2c \left\{ \frac{4 \log N \log |\mathcal{Q}|}{\alpha^2} + \log\left(\frac{1}{\beta}\right) \right\}. \end{aligned}$$

Setting this quantity to be at most $n \max_{q \in \mathcal{Q}} \|q - cq'\|_p + \alpha n$ and solving for c yields (10). Solving for α yields (9). □

A.4 Multiplicative Weights Exponential Mechanism

Theorem 7. If $\left|q_i' - q_j'\right| \leq d_{\mathcal{X}}\left(i, j\right), \forall i, j \in [N]$ and $\forall q' \in \mathcal{Q}'$, then the MWEM mechanism is $d_{\mathcal{X}}$-private.

Proof. **Exponential Mechanism:** Consider the utility function $u' : \mathbb{N}^N \times \mathcal{Q}' \to \mathbb{R}$ given by

$$u'\left(x, q'\right) \ := \ c \left| q'\left(y\right) - q'\left(x\right) \right|, \text{ for some } y \in \mathbb{R}^N.$$

First we find a condition for $\Delta u'\left(i, j\right) \leq c d_{\mathcal{X}}\left(i, j\right), \forall i, j \in [N]$ and $\forall q' \in \mathcal{Q}'$:

$$\begin{aligned} &\Delta u'\left(i, j\right) \\ &= \max_{q' \in \mathcal{Q}'} \max_{\substack{x, x' \in \mathbb{N}^N : \left\|x - x'\right\|_1 \leq 2, \\ x_i \neq x_i', x_j \neq x_j' \text{ for } i, j \in [N]}} \left| u'\left(x, q'\right) - u'\left(x', q'\right) \right|. \end{aligned}$$

For some $x, x' \in \mathbb{N}^N$ such that $\|x - x'\|_1 \leq 2, x_i \neq x_i', x_j \neq x_j'$ for some $i, j \in [N]$, we have:

$$\begin{aligned} &\left| u'\left(x, q'\right) - u'\left(x', q'\right) \right| \\ &= \left| c \left| q'\left(y\right) - q'\left(x\right) \right| - c \left| q'\left(y\right) - q'\left(x'\right) \right| \right| \\ &\overset{(i)}{\leq} c \left| q'\left(x'\right) - q'\left(x\right) \right| \\ &= c \left| \left\langle q', x - x' \right\rangle \right| \\ &\overset{(ii)}{\leq} c \left| q_i' - q_j' \right|, \end{aligned}$$

where (i) is by triangle inequality, and (ii) is due to the choice of x and x'. That is we require

$$\Delta u'\left(i, j\right) \ \leq \ c \left| q_i' - q_j' \right| \ \leq \ c d_{\mathcal{X}}\left(i, j\right).$$

Thus with the above transformed class $\mathcal{Q}'$, if we use the $\mathcal{M}_{\text{Exp}, 2cT}\left(x, u'\right)$ mechanism, we get $\frac{d_{\mathcal{X}}(i,j)}{2T}$-privacy.

Laplace Mechanism: If $\left|q_i' - q_j'\right| \leq d_{\mathcal{X}}\left(i, j\right), \forall i, j \in [N]$ and $\forall q' \in \mathcal{Q}'$, then the Laplace mechanism given by $m = cq'\left(x\right) + \text{Lap}\left(2cT\right)$ satisfies $\frac{d_{\mathcal{X}}(i,j)}{2T}$-privacy.

The composition rules for $d_{\mathcal{X}}$-privacy state that c values accumulate appropriately. We make T calls to the Exponential Mechanism with parameter $2cT$ and T calls to the Laplace Mechanism with parameter $2cT$, resulting in $d_{\mathcal{X}}$-privacy. □

Theorem 8. For any dataset x, set of linear queries $\mathcal{Q}$, $T \in \mathbb{N}$, and $c > 0$, with probability at least $1 - 2T / |\mathcal{Q}|$, MWEM produces y such that

$$\begin{aligned} &\max_{q \in \mathcal{Q}} \left| cq'\left(y\right) - q\left(x\right) \right| \\ &\leq 2n \sqrt{\frac{\log N}{T}} + 10Tc \log |\mathcal{Q}| + n \max_{q \in \mathcal{Q}} \left\| cq' - q \right\|_\infty. \end{aligned}$$

By setting $2n\sqrt{\frac{\log N}{T}} = 10Tc \log |\mathcal{Q}|$, we get

$$\begin{aligned} &\max_{q \in \mathcal{Q}} \left| cq'\left(y\right) - q\left(x\right) \right| \\ &\leq n \max_{q \in \mathcal{Q}} \left\| cq' - q \right\|_\infty + \frac{20}{5^{2/3}} \left(n^2 \log N \log |\mathcal{Q}|\right)^{1/3} c^{1/3}. \end{aligned}$$

Proof. The following inequality follows directly by replacing the ϵ by $\frac{1}{c}$ along the proof given in [14]:

$$\max_{q' \in \mathcal{Q}'} \left| cq'\left(y\right) - cq'\left(x\right) \right| \leq 2n \sqrt{\frac{\log N}{T}} + 10Tc \log |\mathcal{Q}|.$$

Then with probability at least $1 - 2T / |\mathcal{Q}|$, we have

$$\begin{aligned} &\max_{q \in \mathcal{Q}} \left| cq'\left(y\right) - q\left(x\right) \right| \\ &\leq \max_{q \in \mathcal{Q}} \left\{ \left| cq'\left(y\right) - cq'\left(x\right) \right| + \left| cq'\left(x\right) - q\left(x\right) \right| \right\} \end{aligned}$$

$$\begin{aligned}
&\leq \max_{q' \in \mathcal{Q}'} \left| cq'(y) - cq'(x) \right| + \max_{q \in \mathcal{Q}} \left| cq'(x) - q(x) \right| \\
&= \max_{q' \in \mathcal{Q}'} \left| cq'(y) - cq'(x) \right| + \max_{q \in \mathcal{Q}} \left| \left\langle cq' - q, x \right\rangle \right| \\
&\leq \max_{q' \in \mathcal{Q}'} \left| cq'(y) - cq'(x) \right| + \max_{q \in \mathcal{Q}} \left\| cq' - q \right\|_\infty \|x\|_1 \\
&\leq 2n\sqrt{\frac{\log N}{T}} + 10Tc\log|\mathcal{Q}| + n\max_{q\in\mathcal{Q}} \left\| cq' - q \right\|_\infty .
\end{aligned}$$

□

B Pre-processing Optimization

In Section 3, we have shown that by (approximately) solving certain pre-processing optimization problems (*e.g.* (6),(8)), we can obtain the model parameters of the $d_{\mathcal{X}}$-private mechanisms with enhanced utility. One can easily verify that these problems are non-convex optimization problems. Recently, in the optimization and machine learning community, there is a huge interest in developing efficient algorithms for non-convex optimization problems with provable guarantees. One can also observe that these pre-processing optimization problems exhibit coordinate friendly structures, and thus the coordinate descent family of algorithms [24] is a natural choice to solve them.

Consider the optimization problem (6) under squared loss. One can easily verify that $f(c, Q') = n^2 \left\| c \odot Q' - Q \right\|_2^2 + 2\|c\|_2^2$ is a (smooth) multi-convex function *i.e.* $f(\cdot, Q')$ is convex in c for any fixed Q', and $f(c, \cdot)$ is convex in Q' for any fixed c, but f is not jointly convex in (c, Q'). Recently [25] have shown that under certain conditions, the multi-convex optimization problem can be efficiently solved via a variant of cyclic block coordinate descent algorithm. Now consider the optimization problem (8) with the objective function $f(c, \mathcal{Q}') = n\max_{q\in\mathcal{Q}} \left\| cq' - q \right\|_\infty + \frac{20}{5^{2/3}} \left(n^2 \log N \log|\mathcal{Q}|\right)^{1/3} c^{1/3}$. In this case, the objective function is both non-smooth and non-convex, thus the resulting problem is very hard to optimize. However, in practice approximate solutions would still yield good utility.

C Statistical Queries

A statistical query on a data universe $\mathcal{X} \subset \mathbb{R}^d$ is defined by a mapping $q : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}^k$. Abusing notation, we define the evaluation of a statistical query q on the database $x \in \mathcal{X}^n$ to be the average of the predicate over the rows

$$q(x) = \frac{1}{n}\sum_{i=1}^{n} q(x_i). \tag{11}$$

When $q(u) = u$, $\forall u \in \mathcal{X}$, we call it d-way marginal query. We can actually treat the statistical query as a linear query over histogram ($y \in \mathbb{N}^N$) with query matrix $Q \in \mathbb{R}^{k\times N}$. But we can exploit the sum-structure ((11)) of it to design efficient algorithms.

Definition 10. *For $u, v \in \mathcal{X}$ (with $u \neq v$), define the generalized global sensitivity of a query $q \in \mathcal{Q}$ (w.r.t. $\|\cdot\|$) as*

$$\Delta^q_{\|\cdot\|}(u, v) \;:=\; \max_{\substack{x, x' \in \mathcal{X}^n : \left\|x - x'\right\|_H \leq 1, \\ x_i = u, x'_i = v \text{ for } i\in[n]}} \left\| q(x) - q(x') \right\|.$$

Also define $\Delta^q_{\|\cdot\|} := \max_{u,v\in\mathcal{X}} \Delta^q_{\|\cdot\|}(u, v)$ (the usual global sensitivity). When $\|\cdot\| = \|\cdot\|_p$, we simply write Δ^q_p.

The generalized global sensitivity (for $u, v \in \mathcal{X}$) of the statistical query q is given by

$$\begin{aligned}
&\max_{\substack{x, x' \in \mathcal{X}^n : \left\|x - x'\right\|_H \leq 1, \\ x_i = u, x'_i = v \text{ for } i\in[n]}} \left\| \frac{1}{n}\sum_{i=1}^{n} q(x_i) - \frac{1}{n}\sum_{i=1}^{n} q(x'_i) \right\| \\
&= \frac{\|q(u) - q(v)\|}{n}.
\end{aligned}$$

For the d-way marginal query q, we have $\Delta^q_{\|\cdot\|}(u, v) = \frac{\|u - v\|}{n}$.

Definition 11. *Let $\mathcal{X}$ (with $\phi \in \mathcal{X}$) be the data universe, $d_{\mathcal{X}} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be the privacy budget, and $q : \mathcal{X}^n \to \mathcal{Y}$ be the query. A mechanism $\mathcal{M} : \mathcal{X}^n \times \mathcal{Q} \rightsquigarrow \mathcal{Y}$ is said to be $d_{\mathcal{X}}$-private iff $\forall x, x' \in \mathcal{X}^n$ s.t. $\|x - x'\|_H \leq 1$, and $x_i \neq x'_i$ (for some $i \in [n]$), $\forall S \subseteq \mathcal{Y}$ and $\forall q \in Q$ we have*

$$\frac{\mathbb{P}\left[\mathcal{M}(x, q) \in S\right]}{\mathbb{P}\left[\mathcal{M}(x', q) \in S\right]} \;\leq\; \exp\left(d_{\mathcal{X}}\left(x_i, x'_i\right)\right).$$

When $d_{\mathcal{X}}(u, v) = \epsilon, \forall u, v \in \mathcal{X}$, we recover the standard ϵ-differential privacy, and when $d_{\mathcal{X}}(u, v) = \epsilon_u \wedge \epsilon_v$ for $u, v \in \mathcal{X}$, we recover the instance specific differential privacy notion introduced in [11].

For a given query $q : \mathcal{X}^n \to \mathcal{Y} \subset \mathbb{R}^k$ over the database $x \in \mathcal{X}^n$, consider the following variant of Laplace mechanism (with the mapping $\mathcal{X} \mapsto \mathcal{X}'$, and $c \in \mathbb{R}^k$):

$$\mathsf{Z} \;=\; \mathcal{M}_{\mathrm{Lap},c}\left(x', q\right) \;:=\; c \odot q\left(x'\right) + \left(\mathsf{Y}_1, \ldots, \mathsf{Y}_k\right), \tag{12}$$

where $\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)$. Below we show that the above variant of Laplace mechanism satisfies the $d_{\mathcal{X}}$-privacy under a sensitivity bound condition.

Theorem 10. *Let $q'(u) := q(u'), \forall u \in \mathcal{X}$ under the mapping $\mathcal{X} \mapsto \mathcal{X}'$. If $\Delta_1^{q'}(u,v) \leq d_{\mathcal{X}}(u,v)$, $\forall u,v \in \mathcal{X}$, then the mechanism $\mathcal{M}_{\mathrm{Lap},c}(x',q)$ given by (12) satisfies the $d_{\mathcal{X}}$-privacy.*

Proof. Proof is similar to that of Theorem 1. □

The sensitivity bound condition of the above theorem for a statistical query q can be written as follows:

$$\Delta_1^{q'}(u,v) = \frac{\|q(u') - q(v')\|_1}{n} \leq d_{\mathcal{X}}(u,v), \quad \forall u,v \in \mathcal{X}.$$

For the d-way marginal query, the above condition reduces to $\|u' - v'\|_1 \leq n d_{\mathcal{X}}(u,v), \forall u,v \in \mathcal{X}$. The next theorem characterizes the performance of the $\mathcal{M}_{\mathrm{Lap},c}(x',q)$ mechanism under different choices of utility measures:

Theorem 11. *Let $q : \mathcal{X}^n \to \mathbb{R}^k$ be a statistical query of the form $q(x) = \frac{1}{n}\sum_{i=1}^n q(x_i)$, and let $\mathsf{Z} = \mathcal{M}_{\mathrm{Lap},c}(x',q) = c \odot q(x') + \mathsf{Y}$ with $\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)$.*

1. *When $\ell_2^2(y,y') = \|y - y'\|_2^2$, we have*

$$\begin{aligned}&\mathrm{err}_{\ell_2^2}\left(\mathcal{M}_{\mathrm{Lap},c}, q\right)\\ &\leq 2\left\{\max_{u\in\mathcal{X}} \left\|c \odot q\left(u'\right) - q(u)\right\|_2^2 + 2\|c\|_2^2\right\}.\end{aligned}$$

2. *When $\ell_p(y,y') = \|y - y'\|_p$, we have*

$$\begin{aligned}&\mathrm{err}_{\ell_p}\left(\mathcal{M}_{\mathrm{Lap},c}, q\right)\\ &\leq \max_{u\in\mathcal{X}} \left\|c \odot q\left(u'\right) - q(u)\right\|_p + \underset{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)}{\mathbb{E}}\left[\|\mathsf{Y}\|_p\right].\end{aligned}$$

3. *$\forall \delta \in (0,1]$, with probability at least $1-\delta$ we have*

$$\begin{aligned}&\|\mathsf{Z} - q(x)\|_\infty\\ &\leq \max_{u\in\mathcal{X}} \left\|c \odot q\left(u'\right) - q(u)\right\|_\infty + \ln\left(\frac{k}{\delta}\right)\cdot\|c\|_\infty.\end{aligned}$$

Proof. Proof is similar to that of Theorem 2, but with the following change in the appropriate places (with $q'(u) := q(u'), \forall u \in \mathcal{X}$):

$$\begin{aligned}&\max_{x\in\mathcal{X}^n} \left\|c \odot q'(x) - q(x)\right\|_p\\ &= \max_{x\in\mathcal{X}^n} \frac{1}{n}\left\|c \odot \sum_{i=1}^n q\left(x_i'\right) - \sum_{i=1}^n q(x_i)\right\|_p\\ &\leq \max_{x\in\mathcal{X}^n} \frac{1}{n}\sum_{i=1}^n \left\|c \odot q\left(x_i'\right) - q(x_i)\right\|_p\\ &= \frac{1}{n}\sum_{i=1}^n \max_{x_i\in\mathcal{X}} \left\|c \odot q\left(x_i'\right) - q(x_i)\right\|_p\\ &= \max_{u\in\mathcal{X}} \left\|c \odot q\left(u'\right) - q(u)\right\|_p.\end{aligned}$$

□

Now we model the following optimization problem to select the model parameters c and $\mathcal{X} \mapsto \mathcal{X}'$ of the $\mathcal{M}_{\mathrm{Lap},c}(x',q)$ mechanism:

$$\begin{aligned}\underset{c,\mathcal{X}'}{\text{minimize}} \quad & f_{\ell,\mathcal{M}}\left(c,\mathcal{X}';q,n\right)\\ \text{subject to} \quad & \left\|q\left(u'\right) - q\left(v'\right)\right\|_1 \leq n d_{\mathcal{X}}(u,v), \forall u,v \in \mathcal{X}\\ & c \succeq 0.\end{aligned} \tag{13}$$

The objective function $f_{\ell,\mathcal{M}}(c,\mathcal{X}';q,n)$ depends on the utility function that we are interested in. For example, when $\ell_2^2(y,y') = \|y - y'\|_2^2$, we can choose $f_{\ell_2^2,\mathcal{M}}(c,\mathcal{X}';q,n) = \max_{u\in\mathcal{X}} \|c \odot q(u') - q(u)\|_2^2 + 2\|c\|_2^2$. In fact there are two ways to design $d_{\mathcal{X}}$-private mechanisms from existing ϵ differentially private mechanisms: either transform the query vector or the data universe. The approach we used above is $\mathcal{X} \mapsto \mathcal{X}'$ (that is $q(u) \to q'(u) = q(u')$). Thus we can reduce the number of variables in the pre-processing optimization by a factor of k.

Consider a privacy budget (metric) of the form $d_{\mathcal{X}}(u,v) = \sum_{i=1}^d d_i(u_i,v_i)$, where for example $d_i(u_i,v_i) = \epsilon_i[\![u_i \neq v_i]\!]$. In this case, if $|u_i' - v_i'| \leq n d_i(u_i,v_i), \forall u_i,v_i \in \mathcal{X}_i$ (for example, when $\mathcal{X} = \{-1,+1\}^d$, we have $\mathcal{X}_i = \{-1,+1\}$), then the d-way marginal query is $d_{\mathcal{X}}$-private. Moreover, when $\ell_1(y,y') = \|y - y'\|_1$, we have

$$f_{\ell_1,\mathcal{M}}\left(c,\mathcal{X}';q,n\right) = \sum_{i=1}^k f_i(c_i,\mathcal{X}_i)$$

with $f_i(c_i,\mathcal{X}_i) = \max_{u_i\in\mathcal{X}_i} |c_i u_i' - u_i| + c_i$ for d-way marginal queries (since $\max_{u\in\mathcal{X}} \|c \odot q(u') - q(u)\|_1 + \underset{\mathsf{Y}_i \overset{\perp}{\sim} \mathrm{Lap}(c_i)}{\mathbb{E}}[\|\mathsf{Y}\|_1] = \max_{u\in\mathcal{X}} \sum_{i=1}^k |c_i u_i' - u_i| + \|c\|_1 = \sum_{i=1}^k \max_{u_i\in\mathcal{X}_i} |c_i u_i' - u_i| + \sum_{i=1}^k c_i$). Thus in this setting, we can instantiate and *relax* the above optimization problem (13) into k independent optimization problems as follows:

$$\begin{aligned}\underset{c_i,\mathcal{X}_i'}{\text{minimize}} \quad & \max_{u_i\in\mathcal{X}_i} \left|c_i u_i' - u_i\right| + c_i\\ \text{subject to} \quad & \left|u_i' - v_i'\right| \leq n d_i(u_i,v_i), \quad \forall u_i,v_i \in \mathcal{X}_i\\ & c_i \geq 0.\end{aligned} \tag{14}$$