Pufferfish Privacy: An Information-Theoretic Study

Theshani Nuradha, *Graduate Student Member, IEEE*, and Ziv Goldfeld, *Member, IEEE*

***Abstract*— Pufferfish privacy (PP) is a generalization of differential privacy (DP), that offers flexibility in specifying sensitive information and integrates domain knowledge into the privacy definition. Inspired by the illuminating formulation of DP in terms of mutual information due to Cuff and Yu, this work explores PP through the lens of information theory. We provide an information-theoretic formulation of PP, termed mutual information PP (MI PP), in terms of the conditional mutual information between the mechanism and the secret, given the public information. We show that MI PP is implied by the regular PP and characterize conditions under which the reverse implication is also true, recovering the relationship between DP and its information-theoretic variant as a special case. We establish convexity, composability, and post-processing properties for MI PP mechanisms and derive noise levels for the Gaussian and Laplace mechanisms. The obtained mechanisms are applicable under relaxed assumptions and provide improved noise levels in some regimes. Lastly, applications to auditing privacy frameworks, statistical inference tasks, and algorithm stability are explored.**

***Index Terms*— Auditing for privacy, differential privacy, information measures, privacy mechanisms, Pufferfish privacy.**

I. Introduction

With the exponential increase in personal data shared online and recent advancements in data mining techniques, privacy concerns have become more pressing than ever. Statistical privacy frameworks seek to address these threats in a principled manner subject to formal guarantees [2]. Differential privacy (DP) [3] is a popular framework, which preserves the privacy of individual records while enabling aggregate queries about a database. However, DP only deals with one type of private information (individual records modeled by rows of the database) and does not allow to incorporate domain knowledge into the framework. To address these limitations, a versatile generalization of DP called Pufferfish Privacy (PP) was proposed in [4]. PP enables customization of what constitutes private information and explicitly integrates distributional assumptions into its definition. Nevertheless, the flexibility of the PP definition comes at a cost as the general framework is hard to work with and derive mechanisms for. This work aims to circumvent this impasse by proposing a new structured PP framework along with a natural information-theoretic formulation thereof, which lends well for analysis and enables devising mechanisms and exploring various additional applications.

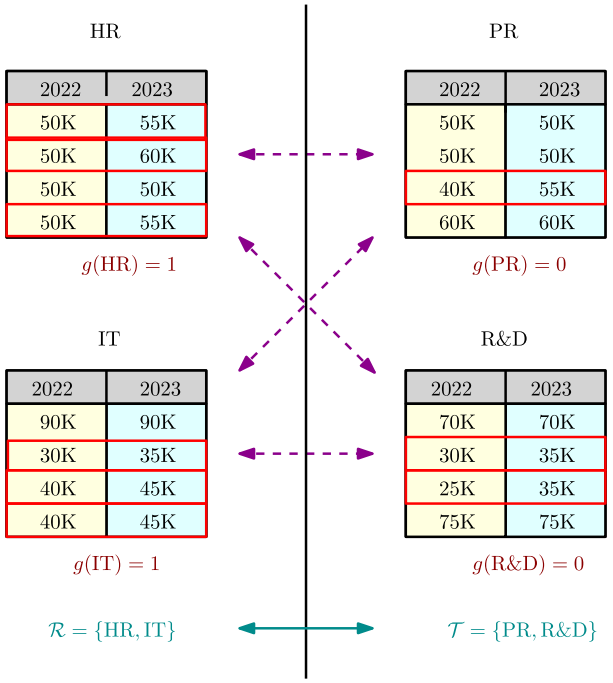


Fig. 1. 2022-2033 salary data in four departments: HR, IT, PR, R&D. The goal is to publish the average 2023 salary in each department (the average of the blue cells) while hiding whether the number raises (marked by red frames) is ≤ 2 corresponding to $g(\cdot) = 0$ or > 2 corresponding to $g(\cdot) = 1$. The average 2022 salaries (yellow cells) are public knowledge.

A. Pufferfish Privacy

Consider salary data from 2022-2023 at a company with four departments: HR, IT, PR, and R&D. The company wants to publish the average 2023 salary in each department while concealing whether more or less than m employees got a raise. The average salaries from 2022 are publicly available. More formally, the goal is to publish $f(x) = \frac{1}{n}\sum_{i=1}^{n} x(i, 2023)$ while privatizing whether $g(x) = \mathbb{1}_{\mathcal{A}_m}$, where $\mathcal{A}_m = \big\{|\{i : x(i, 2022) < x(i, 2023)\}| > m\big\}$, $x \in \mathcal{X} := \{\text{HR}, \text{IT}, \text{PR}, \text{R\&D}\}$, and $x(i, j)$ is the salary of the ith employee during year j in department x. The average salary from 2022, i.e., $w(x) = \frac{1}{n}\sum_{i=1}^{n} x(i, 2022))$, is public knowledge. See Fig. 1 for an instance of the described scenario ($n = 4$ and $m = 2$).

DP operates by making any pair of neighboring databases indistinguishable, with the definition of neighbors being

Manuscript received 22 October 2022; revised 3 May 2023; accepted 11 July 2023. Date of publication 17 July 2023; date of current version 20 October 2023. The work of Ziv Goldfeld was supported in part by NSF under Grant CCF-1947801, Grant CCF-2046018, and Grant DMS-2210368; and in part by the 2020 International Business Machines Corporation (IBM) Academic Award. An earlier version of this paper was presented in part at the 2022 IEEE International Symposium on Information Theory [DOI: 10.1109/ISIT50566.2022.9834520]. *(Corresponding author: Theshani Nuradha.)*

The authors are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14850 USA (e-mail: pt388@cornell.edu; goldfeld@cornell.edu).

Communicated by A. Sarwate, Associate Editor for Security and Privacy.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2023.3296288.

Digital Object Identifier 10.1109/TIT.2023.3296288

up to the privacy mechanism designer. In the scenario above, one may apply a DP-based approach by pairing as neighbors every two departments between which there is a difference in the function value g (whether number of employees getting a raise is more than m). For the example from Fig. 1 this amounts to the set of pairs $\{(\mathrm{HR},\mathrm{PR}),(\mathrm{HR},\mathrm{R\&D}),(\mathrm{IT},\mathrm{PR}),(\mathrm{IT},\mathrm{R\&D})\}$ (marked by the dashed purple arrows in the figure). However, by following this approach, we guarantee a stricter privacy requirement than necessary. Indeed, upon observing the privatized version of the published query f and assuming w is publicly known, we only need to make the sets $g^{-1}(0)=\{\mathrm{PR},\mathrm{R\&D}\}$ and $g^{-1}(1)=\{\mathrm{HR},\mathrm{IT}\}$ indistinguishable (marked by the solid dark cyan arrow in Fig. 1). The benefit of targeting this relaxed notion of privacy is that it enables to achieve improved accuracy and utility. The PP framework is designed to do just that, by enabling full customization of the events that are regarded as private. In addition, PP allows integrating into the framework domain knowledge on the distribution over databases; by considering the set of all possible distributions, this reduces back to the worst-case requirement of DP.

However, the generality of the PP also has drawbacks. For starters, PP does not satisfy general composability [4], [5], which is regarded as a privacy axiom—a property that any privacy mechanism should possess. Hence, the outputs of two PP mechanisms can not always be combined to satisfy PP together (as a multi-query output), which limits its usage in practice. In addition, there is a shortage of mechanisms that guarantee PP. The main attempt in that direction is the Wasserstein mechanism from [6], which is computationally burdensome as it requires computing the ∞-Wasserstein distance between all pairs of conditional distributions of the mechanism's output given any pair of secrets events. Lastly, we note that formal guarantees for PP mechanisms pertaining to privacy-utility tradeoffs, sample complexity bounds for private inference tasks, etc., are largely unavailable due to the hardship of analyzing this framework in full generality. Our goal is to address these shortcomings by introducing some structure into the PP framework to make it more tractable while preserving versatility, and then study the structured variant using tools from information theory.

B. Contributions

We first propose a novel structured PP framework, where the private and public information is modeled as pairs of functions of the database that are coupled via a bipartite graph. This framework captures various privacy notions, from DP [3] to attribute privacy (AP) [7],[1] as special cases, while lending well for analysis via tools from information theory. We provide an information-theoretic formulation of the structured PP framework in terms of the conditional mutual information between the mechanism and the secret function given the public one. Generally, the ϵ-mutual information PP (ϵ-MI PP) criteria is implied by ϵ-PP, but we further show that it is sandwiched between ϵ-PP and (ϵ,δ)-PP in terms of strength under appropriate distributional assumptions and parameter values. The proof relies on representing PP constraints as bounds on certain divergences,[2] and comparing those to the Kullback-Leibler (KL) divergence (and thus mutual information) via tools like Pinsker's inequality and the minimax redundancy capacity theorem.

The information-theoretic formulation of the structured PP framework enables a comprehensive analysis of properties, mechanisms, and applications. We begin by establishing properties of ϵ-MI PP mechanisms, encompassing convexity, post-processing, and composability. This shows that our MI PP definition satisfies all the axioms required from a privacy framework [5], [8]. In particular, while standard PP mechanisms are generally not composable [4], [5], our composability results for ϵ-MI PP offer greater flexibility especially in the non-adaptive query setting.

We next study ϵ-MI PP mechanisms, which is another aspect where the standard PP framework is lacking (the main available mechanisms for standard PP is the Wasserstein mechanism [6], which is computationally intractable). We derive sufficient conditions on the injected noise level for the Laplace and Gaussian mechanisms that guarantee ϵ-MI PP, and thus also ϵ-MI DP as a special case. The derivation of MI PP mechanisms relies on controlling mutual information via maximum entropy arguments and the entropy power inequality. The resulting noise parameter bounds depend on the conditional variance of the query, which differs from classical results that typically depend on the ℓ^1- or ℓ^2-sensitivity of the query; cf. e.g., [9], [10], [11], and [12]. Variance-based parameter bounds are particularly desirable under the PP framework as it encodes prior knowledge on the data distribution. Indeed, it may be the case that sensitivity explodes (e.g., for query functions with unbounded range) but variance is finite due to concentration properties of the distribution class. One drawback of the proposed mechanisms (as well as sensitivity-based ones) is that the injected noise level grows linearly with the dimension. To circumvent this effect, we also propose a Gaussian mechanism that first projects the high-dimensional data onto a low-dimensional space and then adds noise. We obtain parameter bounds for this projection mechanism in terms of the operator norm of conditional covariance matrices and the conditional mean vectors.

Several applications of ϵ-MI PP are explored, starting from auditing for DP [13], [14], [15]. Auditing black-box mechanisms to certify whether they satisfy a target DP guarantee is challenging, especially in high-dimensional settings. To address this problem, we observe that to audit for DP violations, it suffices to test whether a relaxed privacy notion violates the target privacy level [15]. We then propose a rigorous hypothesis testing framework for DP violations using the information-theoretic formulation of DP as our test statistic. Since estimating mutual information

[1] AP guarantees privacy of functions associated with possibly sensitive attributes in a database, e.g., race or gender. For instance, referring back to the example in Fig. 1, if the secret function was the maximum salary of the year 2023 and no public information of the average 2022 salary was available, then that setting would fall under the AP framework. We note that the current example, however, is not captured under AP since the private function is related to two attributes of the database rather than a single attribute as in AP.

[2] ∞-Rényi divergence for ϵ-PP and total variation distance for $(0,\delta)$-PP.

between high-dimensional variables is statistically burdensome, we introduce a further relaxation to privacy based on sliced mutual information (SMI) [16], [17], which enjoys parametric empirical convergence rates in arbitrary dimension. Our auditing approach naturally extends to the PP framework. Beyond privacy auditing, we also explore multivariate mean estimation under ϵ-MI PP and derive sample complexity bounds that adapt to the domain knowledge in the PP framework. Lastly, we study privacy-utility tradeoffs using ϵ-MI PP and explore its connections to algorithmic stability, which is a standard tool for establishing generalization bounds in statistical learning theory [18], [19].

C. Related Work

Connections between statistical privacy and information theory have gained increased attention [20], [21], [22], [23], [24], [25], [26] as they enable borrowing tools and ideas from one discipline to make progress in the study of the other. In particular, [24] established a two-sided connection between DP and the conditional mutual information between the mechanism and any individual record, given the rest of the database. This mutual information-based privacy notion (hereafter abbreviated as MI DP) lends well for an information-theoretic analysis and quantifies privacy via a common currency using which privacy-utility tradeoffs may be explored [27]. Privacy metrics based on mutual information have been leveraged to analyze and provide guarantees for various inference and learning tasks. MI DP has been used in [28] to study fundamental privacy-utility tradeoffs in linear regression problems. Variants of MI DP have also been used in the context of federated learning study the generalization error and privacy leakage [29], [30], as well as convergence of privacy-preserving training algorithms [31]. Mutual information-based privacy leakage metrics have also been used in other applications, such as optimal battery charging policies subject to privacy constraints [32], multiple hypothesis testing [33], and online location tracing [34].

Other widely used average-case privacy notion is Kullback-Leibler (KL) DP [5], [35] and Rényi DP [26], both of which serve as relaxations of the classical DP framework. While the main appeal of such average notions is their analytic tractability, they have also been utilized for various applications. KL DP has been applied in settings ranging from collaborative schemes [36], [37] and smart grids [38] to the industrial internet of things [39]. It has also been used in tandem with worst-case privacy notions such as DP [38], by employing them in different stages of the algorithm/scheme of interest. This highlights that even in applications where average-case privacy requirements are not sufficient by themselves, combining them in certain (less sensitive) stages of the system is beneficial, e.g., in terms of utility. KL DP is a special case of Rényi DP [26] with $\alpha = 1$. Rényi DP has been applied to keep track of the privacy budgets in applications including optimization [40], deep learning [41], and generative adversarial networks [42]. The utility and tractability of privacy notions like KL DP, Rényi DP, and MI DP serve as inspiration for the information-theoretic formulation of PP proposed herein.

D. Organization

The rest of the paper is organized as follows. In Section II, we introduce notation and preliminaries. The structured PP framework, its information-theoretic formulation, and the relation to ϵ- and (ϵ, δ)-PP are the focus of Section III. Properties of the ϵ-MI PP framework and mechanisms are treated in Sections IV and V, respectively. In Section VI we design a sample-efficient hypothesis test for privacy auditing based on ϵ-MI PP. Additional applications to private mean estimation, algorithmic stability, and privacy-utility tradeoffs are covered in Section VII. Proofs are given in Section VIII, while Section IX provides concluding remarks and future directions.

II. Background and Preliminaries

We set up the notation used throughout this paper, present the DP framework along with its information-theoretic formulation from [24], and introduce the PP paradigm.

A. Notation

Sets are denoted by calligraphic letters, e.g. $\mathcal{X}$. For $k, n \in \mathbb{N}$, we use $\mathcal{X}^{n\times k}$ for the database space of $n \times k$ matrices (columns correspond to different attributes while rows to different individuals). The (i, j)th entry of $x \in \mathcal{X}^{n\times k}$ is $x(i, j)$. The ith row and jth column of x are $x(i, \cdot)$ and $x(\cdot, j)$, respectively. The image of a function $g : \mathcal{X}^{n\times k} \to \mathbb{R}^d$ is denoted by $\mathrm{Im}(g)$. For $p \geq 1$, $\|\cdot\|_p$ designates the ℓ^p norm on $\mathbb{R}^d$; we omit the subscript when $p = 2$ (which is our typical use case). The operator norm for matrices is denoted by $\|\cdot\|_{\mathrm{op}}$. For two numbers a and b, we use the notation $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$

We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the underlying probability space on which all random variables (RVs) are defined, with $\mathbb{E}$ designating expectation. RVs are denoted by upper case letters, e.g., X, with P_X representing the corresponding probability law. For $X \sim P_X$, we interchangeably use $\mathrm{spt}(X)$ and $\mathrm{spt}(P_X)$ for the support. The joint law of (X, Y) is denoted by P_{XY}, while $P_{Y|X}$ designates the (regular) conditional probability of Y given X. Conventions for $n \times k$-dimensional random variables are the same as for deterministic elements. The space of all Borel probability measures on $\mathcal{S} \subseteq \mathbb{R}^d$ is denoted by $\mathcal{P}(\mathcal{S})$. We write $P \ll Q$ to denote that P is absolutely continuous with respect to (w.r.t.) Q. The n-fold product measure of $P \in \mathcal{P}(\mathcal{S})$ is $P^{\otimes n}$. Indicator function of a measurable event $\mathcal{A} \in \mathcal{F}$ is denoted by $\mathbb{1}_{\mathcal{A}}$.

For $(X, Y) \sim P_{XY}$, the mutual information between X and Y is denoted by $\mathsf{I}(X; Y)$. The differential entropy of X is $\mathsf{h}(X)$. Conditional versions of the above given a third (correlated) RV Z are denoted by Z by $\mathsf{I}(X; Y|Z)$ and $\mathsf{h}(X|Z)$, respectively. The Kullback-Leibler (KL) divergence between $P, Q \in \mathcal{P}(\mathcal{X})$ with $P \ll Q$ is

$$\mathsf{D}_{\mathsf{KL}}(P\|Q) := \mathbb{E}_P\left[\log\left(\frac{dP}{dQ}\right)\right],$$

where $\frac{dP}{dQ}$ is the Radon-Nikodym derivative of P w.r.t. Q. The total variation (TV) distance is defined as

$$\|P - Q\|_{\mathsf{TV}} := \sup_{\mathcal{A}} \big|P(\mathcal{A}) - Q(\mathcal{A})\big|,$$

where the supremum is over all measurable sets $\mathcal{A}$. Both the KL divergence and the TV distance are jointly convex in (P, Q), and are related to one another via Pinsker's inequality [43]: $\|P - Q\|_{\mathsf{TV}} \leq \sqrt{0.5 \mathsf{D}_{\mathsf{KL}}(P\|Q)}$. Also recall that $\mathsf{I}(X;Y)$ can be expressed in terms of KL divergence as $\mathsf{I}(X;Y) = \mathsf{D}_{\mathsf{KL}}(P_{XY}\|P_X \otimes P_Y)$, where P_X and P_Y are the respective marginals of X and Y. For $1 \leq p < \infty$, the p-Wasserstein distance between $P, Q \in \mathcal{P}(\mathcal{X})$ with finite pth absolute moments, i.e., $\mathbb{E}_P[\|X\|^p], \mathbb{E}_Q[\|Y\|^p] < \infty$, is $\mathsf{W}_p(P,Q) := \inf_{\pi \in \Pi(P,Q)} \left(\mathbb{E}_\pi\left[\|X - Y\|^p\right]\right)^{1/p}$, where $\Pi(P,Q)$ is the set of couplings of P and Q. The ∞-Wasserstein distance is given by $\mathsf{W}_\infty(P,Q) := \lim_{p\to\infty} \mathsf{W}_p(P,Q) = \inf_{\pi\in\Pi(P,Q)} \sup_{x,y\in \mathrm{spt}(\pi)} \|x - y\|$.

For multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0$, the partial derivative operator of order $\|\alpha\|_1$ is $D^\alpha = \frac{\partial^{\alpha_1}}{\partial^{\alpha_1} x_1} \cdots \frac{\partial^{\alpha_d}}{\partial^{\alpha_d} x_d}$. For an open set $\mathcal{U} \subseteq \mathbb{R}^d$ and $s \in \mathbb{N}_0$, let $\mathsf{C}^s(\mathcal{U})$ be the class of functions whose partial derivatives up to order s all exist and are continuous on $\mathcal{U}$. The Hölder function class of smoothness $s \in \mathbb{N}_0$ and radius $b \geq 0$ is then defined as $\mathsf{C}_b^s(\mathcal{U}) := \{f \in \mathsf{C}^s(\mathcal{U}) : \max_{\alpha:\|\alpha\|_1 \leq s} \|D^\alpha f\|_{\infty,\mathcal{U}} \leq b\}$. The restriction of $f : \mathbb{R}^d \to \mathbb{R}$ to $\mathcal{X} \subseteq \mathbb{R}^d$ is denoted by $f|_{\mathcal{X}}$. For compact $\mathcal{X}$, slightly abusing notation, we set $\|\mathcal{X}\| := \sup_{x\in\mathcal{X}} \|x\|$.

B. Differential Privacy

DP allows answering queries about aggregate quantities while protecting the individual entries in the database [3]. To that end, the output of differentially private mechanism should be indistinguishable for neighboring databases—those that differ only in a single record (row). Formally, we say that $x, x' \in \mathcal{X}^{n\times k}$ are neighbors, denoted $x \sim x'$, if $x(i,\cdot) \neq x'(i,\cdot)$ for some $i = 1, \ldots, n$, and agree on all other rows.

Definition 1 (Differential Privacy)**.** *Fix $\epsilon, \delta > 0$.* A randomized mechanism[3] $M : \mathcal{X}^{n\times k} \to \mathcal{Y}$ *is* (ϵ, δ)-differentially private if for all $x \sim x'$ with $x, x' \in \mathcal{X}^{n\times k}$, and $\mathcal{A} \subseteq \mathcal{Y}$ measurable, we have

$$\mathbb{P}\big(M(x) \in \mathcal{A}\big) \leq e^\epsilon \, \mathbb{P}\big(M(x') \in \mathcal{A}\big) + \delta. \tag{1}$$

The formulation of DP can be extended to arbitrary neighboring relations between databases, which is known as generic DP [5]. Namely, neighbors can be defined as pairs that agree on all entries except any prespecified portion of the database (as opposed to just the rows, as in standard DP). For instance, viewing databases that agree up to their columns as neighbors, gives rise to a variant of the AP framework [7].

An information-theoretic formulation of DP was proposed in [24] in terms of the conditional mutual information between each row of the database and the mechanism, given the rest of the rows. We next define ϵ-mutual information DP (ϵ-MI DP) and then recall the main equivalence result of [24].

[3]A randomized mechanism is described by a (regular) conditional probability distribution given the data, i.e., $P_{M|X}$.

Definition 2 (ϵ-MI DP)**.** *A Randomized mechanism $M : \mathcal{X}^{n\times k} \to \mathcal{Y}$ is ϵ-MI DP, if*

$$\sup_{\substack{P_X \in \mathcal{P}(\mathcal{X}^{n\times k}),\\ i=1,\ldots,n}} \mathsf{I}\big(X(i,\cdot); M(X) \big| \big(X(j,\cdot)\big)_{j\neq i}\big) \leq \epsilon. \tag{2}$$

Theorem 1 of [24] states that ϵ-DP (i.e., (ϵ, δ)-DP with $\delta = 0$) implies ϵ-MI DP, which further implies $(\epsilon', \sqrt{2\epsilon})$-DP, for any $\epsilon' \geq 0$. Thus, ϵ-MI DP is in fact sandwiched between ϵ-DP and (ϵ, δ)-DP in terms of its strength. It was also shown in [24] that ϵ-MI DP satisfies properties such as convexity, post-processing, and adaptive/non-adaptive composition.

C. Pufferfish Privacy

For a database space $\mathcal{X}^{n\times k}$, the PP framework [4] consists of three components: (i) a set of secrets $\mathcal{S}$, that contains measurable subsets of $\mathcal{X}^{n\times k}$; (ii) a set of secret pairs $\mathcal{Q} \subseteq \mathcal{S} \times \mathcal{S}$ that needs to be statistically indistinguishable in the (ϵ, δ) sense (see (3)); and (iii) a class of data distributions $\Theta \subseteq \mathcal{P}(\mathcal{X}^{n\times k})$, that captures prior beliefs or domain knowledge. As formulated next, the goal of PP is to make all secret pairs in $\mathcal{Q}$ indistinguishable w.r.t. those prior beliefs $P_X \in \Theta$.

Definition 3 (Pufferfish privacy)**.** *Fix $\epsilon, \delta > 0$.* A randomized mechanism $M : \mathcal{X}^{n\times k} \to \mathcal{Y}$ is (ϵ, δ)-private in the pufferfish framework $(\mathcal{S}, \mathcal{Q}, \Theta)$ if for all $P_X \in \Theta$, $(\mathcal{R}, \mathcal{T}) \in \mathcal{Q}$ with $P_X(\mathcal{R}), P_X(\mathcal{T}) > 0$, and $\mathcal{A} \subseteq \mathcal{Y}$ measurable, we have

$$\mathbb{P}\big(M(X) \in \mathcal{A} \big| \mathcal{R}\big) \leq e^\epsilon \, \mathbb{P}\big(M(X) \in \mathcal{A} \big| \mathcal{T}\big) + \delta. \tag{3}$$

DP from Definition 1 is a special case of PP when $\mathcal{S} = \mathcal{X}^{n\times k}$, $\mathcal{Q}$ contains all neighboring pairs of databases, and $\Theta = \mathcal{P}(\mathcal{X}^{n\times k})$ (i.e., no distributional assumptions are made, and privacy is guaranteed in the worst case). PP also subsumes any other famework under generic DP [5] (i.e., alternative neighboring relations) by choosing $\mathcal{Q}$ accordingly. Another special case of PP is AP [7], which privatizes global statistical properties of data attributes. In this case, $\mathcal{S}$ is the value of a function evaluated on the data, $\mathcal{Q}$ contains pairs of function values, and Θ captures assumptions on how the data was sampled and correlations across attributes thereof. These special cases are discussed in detail in Remark 2 ahead.

III. Pufferfish Privacy and Mutual Information

Towards an information-theoretic characterization of PP, it is convenient to focus on a slightly more structured formulation that explicitly decomposes pairs of secrets into private and public parts. The considered PP framework is presented next, followed by an information-theoretic characterization.

A. Structured Pufferfish Privacy Framework

We focus on a special case of the general framework, where pairs $(\mathcal{R}, \mathcal{T}) \in \mathcal{Q}$ are decomposed into a private part (on which they should be indistinguishable) and a common part (interpreted as public information). In Remark 2 we demonstrate how the considered formulation reduces to popular privacy notions like DP [3] and AP [7]. Our formulation is constructed as follows:

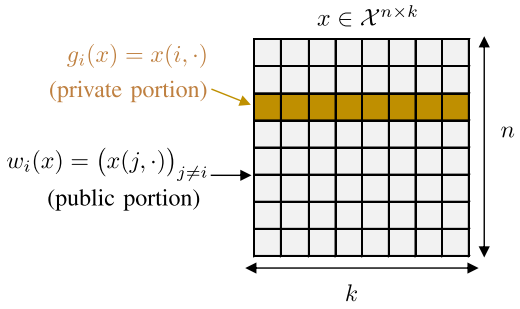


Fig. 2. Function pairs for DP: The ith row of $x \in \mathcal{X}^{n\times k}$ is the private portion, while the rest of the database is the corresponding public part.

1) Private/public functions: Let $\mathcal{G}$ and $\mathcal{W}$ be finite sets, containing functions on $\mathcal{X}^{n\times k}$. For $g \in \mathcal{G}$, we interpret $g(X)$ as a private feature of the database $X \sim P_X \in \Theta$, while $w(X)$, $w \in \mathcal{W}$, represents publicly available information.

2) Function pairs: To encode which private-public function pairs constitute a secret (i.e., an element of $\mathcal{S}$) we use a bipartite graph. Consider the graph $(\mathcal{G}, \mathcal{W}, \mathcal{E})$, where $\mathcal{E}$ is a given edge set between the two partitions $\mathcal{G}$ and $\mathcal{W}$. We write $g \sim w$ if $\{g, w\} \in \mathcal{E}$ for some $g \in \mathcal{G}$ and $w \in \mathcal{W}$. The operational meaning of an edge $g \sim w$ is that $g(X)$ must be concealed even if the adversary has access to $w(X)$. For example, for DP we take $g_i(x) = x(i,\cdot)$ as a specific row of the database and $w_i(x) = \big(x(j,\cdot)\big)_{j\neq i}$ as the rest of the database, where $i = 1, \ldots, n$; then set $\mathcal{G} = \{g_i\}_{i=1}^n$, $\mathcal{W} = \{w_i\}_{i=1}^n$, and $\mathcal{E} = \big\{\{g_i, w_i\}\big\}_{i=1}^n$, as depicted in Figure 2.)

3) Secret event: Each secret event (namely, an element of $\mathcal{S}$) corresponds to a specific value that a private-public function pair takes, i.e., for $\mathcal{G} \ni g \sim w \in \mathcal{W}$, $a \in \mathrm{Im}(g)$, and $c \in \mathrm{Im}(w)$, $\mathcal{S}$ comprises all events of the form $\mathcal{A}_{g,w}(a,c) := \big\{g(X) = a,\, w(X) = c\big\}$.

4) Secret event pairs: Elements of $\mathcal{Q} \subseteq \mathcal{S}\times\mathcal{S}$ are pairs that share the same public information (i.e., the value for $w(X)$) but differ in their private portions (the value of $g(X)$).

We are now ready to define the structured PP framework.

Definition 4 (Structured PP Framework)**.** *Fix $\epsilon, \delta > 0$ and consider a bipartite graph $(\mathcal{G}, \mathcal{W}, \mathcal{E})$ with sets of functions $\mathcal{G}$ and $\mathcal{W}$ as described above. A randomized mechanism $M : \mathcal{X}^{n\times k} \to \mathcal{Y}$ is (ϵ, δ)-private in the structured pufferfish framework $(\mathcal{G}, \mathcal{W}, \mathcal{E}, \Theta)$ if it satisfies Definition 3 with*

$$\mathcal{S} = \big\{\mathcal{A}_{g,w}(a,c) : \mathcal{G} \ni g \sim w \in \mathcal{W},\, a \in \mathrm{Im}(g),\, c \in \mathrm{Im}(w)\big\}$$
$$\mathcal{Q} = \big\{\{\mathcal{A}_{g,w}(a,c), \mathcal{A}_{g,w}(b,c)\} : \mathcal{G} \ni g \sim w \in \mathcal{W},\, c \in \mathrm{Im}(w),\, a, b \in \mathrm{Im}(g),\, a \neq b\big\}$$

and a set of data distributions $\Theta \subseteq \mathcal{P}(\mathcal{X}^{n\times k})$.

The structured PP framework captures various prominent privacy notions, such as DP [3] and AP [7].

Remark 1 (Semantics of the Structured PP Framework)**.** Structured PP framework provides the following privacy guarantee: for any database X generated from a distribution in the class Θ, an adversary that knows the function value $w(X)$, for $w \in \mathcal{W}$, and the output of the privatization mechanism $M(X)$ draws the same conclusions regardless of the value of the private function $g(X)$, $g \in \mathcal{G}$. This applies to many realistic scenarios, such as the one described in Section I-A (see also Fig. 1). Noting that this example indeed falls under the structured PP framework, we have that the value of g (whether more than half of the employees in each department received a raise) is protected even if the adversary has the access to w (average salaries of year 2022 across the department) and M (the privatized average 2023 salary).

Remark 2 (Special Cases)**.** The structured PP framework reduces to various important privacy notions. We provide two such examples pertaining to DP and AP.

1) DP corresponds to a structured PP framework with $\Theta = \mathcal{P}(\mathcal{X}^{n\times k})$, private functions $g_i(x) = x(i,\cdot)$, public functions $w_i(x) = \big(x(j,\cdot)\big)_{j\neq i}$, where $i = 1, \ldots, n$, and an edge set $\mathcal{E} = \big\{\{g_i, w_i\}\big\}_{i=1}^n$. This construction naturally extends to any privacy framework where secret events are singletons (databases). Then, each private function g acts on some prespecified portion of the database, and is connected by an edge to a public function w that acts on the remaining data entries.

2) The AP framework privatizes attributes of the database, which are captured by certain functions $\tilde{g}_j : \mathcal{X}^k \to \mathbb{R}$ of the columns $j = 1, \ldots, k$. Following the setup of [3], we take $g_j(x) = \tilde{g}_j\big(x(\cdot, j)\big)$, for $j = 1, \ldots, k$; as AP includes no public information we set $\mathcal{W} = \mathcal{E} = \emptyset$ and let Θ be the class of distributions of interest. Alternatively, one may consider a variant of AP with public information, which is the portion of the database except the considered column. In that case, $\mathcal{W}$ is a set of functions $w_j(x) = \big(x(\cdot, i)\big)_{i\neq j}$, where $j = 1, \ldots, k$, and $\mathcal{E} = \big\{\{g_j, w_j\}\big\}_{j=1}^k$.

Given the definition of the structured PP framework, it is natural to ask for mechanisms that attain it. The Wasserstein mechanism from [6] can be used to guarantee general PP. However, that approach is computationally intractable. A simpler mechanism can be devised by following the approach of [7, Theorem 1] for AP under appropriate Gaussianity assumptions. Specifically, fix a query $f : \mathcal{X}^{n\times k} \to \mathbb{R}$ and suppose that for any $X \sim P_X \in \Theta$, $g \in \mathcal{G}$, and $w \in \mathcal{W}$ with $g \sim w$, we have that the conditional distribution of $f(X)$ given $\big(g(X), w(X)\big)$ is Gaussian. Under this assumption, the conditional variance $\mathrm{Var}\big(f(X)\big|g(X) = a, w(X) = c\big)$ does not depend on the values (a, b) and the Gaussian noise-injection mechanism $M(X) = f(X) + Z$, with $Z \sim \mathcal{N}(0, \sigma^2)$ satisfies (ϵ, δ)-structured PP whenever

$$\sigma^2 \geq \sup_{\substack{P_X\in\Theta,\\ g\in\mathcal{G}, w\in\mathcal{W}:\\ g\sim w,\\ c\in\mathrm{Im}(w)}} 2\left(\epsilon^{-1}\Delta_{f,g,w}^{P_X}(c)\right)^2 \log(1.25/\delta) - \mathrm{Var}\big(f(X)\big|g(X) = a_0, w(X) = c_0\big),$$

where $(a_0, c_0) \in \mathrm{Im}(g) \times \mathrm{Im}(w)$ are arbitrary and

$$\Delta_{f,g,w}^{P_X}(c) := \sup_{a,b\in\mathrm{Im}(g)} \big|\mathbb{E}\big[f(X)\big|g(X) = a, w(X) = c\big] - \mathbb{E}\big[f(X)\big|g(X) = b, w(X) = c\big]\big|.$$

While the above Gaussianity requirement may hold by assuming, e.g., Gaussian data and linear functions, it is generally quite restrictive. To devise tractable mechanisms beyond this Gaussian setting, we next propose an information-theoretic reformulation of the structured PP framework. This reformulation lends well for analysis and enables deriving sufficient conditions on parameters of noise-injection mechanism that guarantee structured PP in general.

B. Information-Theoretic Formulation

To provide an information-theoretic formulation of the structured PP framework, we first define ϵ-MI PP.

Definition 5 (ϵ-MI PP)**.** *Let $(\mathcal{G},\mathcal{W},\mathcal{E})$ be a bipartite graph as in Definition 4 and $\Theta \subseteq \mathcal{P}(\mathcal{X}^{n\times k})$. A randomized mechanism $M:\mathcal{X}^{n\times k}\to\mathcal{Y}$ is ϵ-MI PP in the framework $(\mathcal{G},\mathcal{W},\mathcal{E},\Theta)$ if*

$$\sup_{\substack{P_X\in\Theta,\\ g\in\mathcal{G},w\in\mathcal{W}:\\ g\sim w}} \mathsf{I}\big(g(X);M(X)|w(X)\big)\le\epsilon.$$

Evidently, ϵ-MI PP as defined above recovers the notion of ϵ-MI DP from [24] (see Definition 2) by taking $(\mathcal{G},\mathcal{W},\mathcal{E},\Theta)$ as described in Part 1 of Remark 2.

Remark 3 (Revisiting Semantics of the Structured PP)**.** *The ϵ-MI PP formulation explicitly encodes the semantics of the structured PP framework, as explained in Remark 1. Namely, for any database distribution $P_X\in\Theta, \mathcal{G}\ni g\sim w\in\mathcal{W}$, the mechanism's output $M(X)$ should not convey more than ϵ information bits about any secret function $g(X)$, $g\in\mathcal{G}$, even when $w(X)$, $w\in\mathcal{W}$, is available as side information.*

The following theorem characterizes the relative strength of the structured PP framework from Definition 4 compared to ϵ-MI PP, showing that the latter lies between ϵ-PP (i.e., (ϵ,δ)-PP with $\delta=0$) and (ϵ,δ)-PP for appropriate parameter values.

Theorem 1 (Relative Strength)**.** *Consider the structured (ϵ,δ)-PP framework $(\mathcal{G},\mathcal{W},\mathcal{E},\Theta)$ from Definition 4. Let $\epsilon'>0$ be arbitrary and set $\epsilon''=\epsilon\wedge\frac{1}{2}\epsilon^2$. Then*

$$\epsilon\text{-PP}\implies\epsilon''\text{-MI PP}\ .$$

and if $\Theta=\mathcal{P}(\mathcal{X}^{n\times k})$, then we further have

$$\epsilon\text{-PP}\implies\epsilon''\text{-MI PP}\implies(\epsilon',\sqrt{2\epsilon''})\text{-PP}.$$

Moreover, the inverse implication

$$(\epsilon,\delta)\text{-PP}\implies\epsilon^\star\text{-MI PP}$$

holds under either of the following conditions:

1) $\big|\mathrm{spt}\big(M(X)\big)\big|<\infty$ *or* $\max_{g\in\mathcal{G}}|\mathrm{Im}(g)|<\infty$, *whence*

$$\epsilon^\star=2\mathsf{h}_b(\delta')+2\delta'\log\left(\big|\mathrm{spt}\big(M(X)\big)\big|\wedge\Big(\max_{g\in\mathcal{G}}|\mathrm{Im}(g)|+1\Big)\right)$$

where $\mathsf{h}_b(\alpha)=-\alpha\log(\alpha)-(1-\alpha)\log(1-\alpha)$, for $\alpha\in[0,1]$, is the binary entropy function in nats and $\delta'=1-2(1-\delta)/(e^\epsilon+1)\in[0,1]$.

2) *The joint density $f_{M(X),g(X),w(X)}$ and conditional density $f_{M(X)|g(X),w(X)}$ exists, whence $\epsilon^\star$ is given in (4), as shown at the bottom of the page, where*

$$\begin{aligned}\alpha_{a,b,c}^{-1}&=\sup_{x\in\mathrm{spt}(M(X))}\frac{f_{M(X)|g(X),w(X)}(x|a,c)}{f_{M(X)|g(X),w(X)}(x|b,c)}\\ \beta_{a,b,c}&=\inf_{x\in\mathrm{spt}(M(X))}\frac{f_{M(X)|g(X),w(X)}(x|a,c)}{f_{M(X)|g(X),w(X)}(x|b,c)}\\ u_{a,,c}&=\sup_{x\in\mathrm{spt}(M(X))}f_{M(X),g(X),w(X)}(x,a,c)\\ \ell_{a,c}&=\inf_{x\in\mathrm{spt}(M(X))}f_{M(X)|g(X),w(X)}(x,a,c).\end{aligned}$$

Theorem 1 is proven in Section VIII-A. The first implication follows by reformulating ϵ-PP in terms of the ∞-Rényi divergence, translating that to an ϵ bound on the corresponding KL divergence, and then use joint convexity to arrive at ϵ''-MI PP. When Θ is the set of all database distributions, the second implication is derived via the minimax redundancy capacity representation and Pinsker's inequality. The inverse implications first translates (ϵ,δ)-PP into a bound on the TV distance between corresponding conditional distributions and then employs either continuity of entropy w.r.t. the TV distance to control mutual information or the reverse Pinsker inequality.Note that the privacy guarantee provided by ϵ-PP is stronger than that of ϵ-MI PP, as evident from the implication ϵ-PP $\implies$ ϵ-MI PP.

Remark 4 (ϵ-KL PP)**.** *ϵ-KL DP [5], [35] can be generalized to the setting of structured PP as follows. Let $(\mathcal{G},\mathcal{W},\mathcal{E})$ be a bipartite graph as in Definition 4 and $\Theta\subseteq\mathcal{P}(\mathcal{X}^{n\times k})$. A randomized mechanism $M:\mathcal{X}^{n\times k}\to\mathcal{Y}$ is ϵ-KL PP in the framework $(\mathcal{G},\mathcal{W},\mathcal{E},\Theta)$ if $\forall P_X\in\Theta$, $\mathcal{G}\ni g\sim w\in\mathcal{W}$, $a,b\in\mathrm{Im}(g)$, $c\in\mathrm{Im}(w)$*

$$\mathsf{D}_{\mathsf{KL}}\big(P_{M(X)|\mathcal{A}_{g,w}(a,c)}\big\|P_{M(X)|\mathcal{A}_{g,w}(b,c)}\big)\le\epsilon,$$

where $\mathcal{A}_{g,w}(a,c)=\big\{g(X)=a,\,w(X)=c\big\}$. ϵ-KL PP as defined above sits between the structured PP from Definition 4 and ϵ-MI PP from Definition 5 in terms of strength. This is evident from the proof of the first implication in Theorem 1, which effectively shows that ϵ-PP $\implies$ ϵ''-KL PP $\implies$ ϵ''-MI PP, for $\epsilon''=\epsilon\wedge\frac{1}{2}\epsilon^2$ (see (9) in Section VIII-A). The above

$$\epsilon^\star=\left(1-\frac{2(1-\delta)}{e^\epsilon+1}\right)\left\{\sup_{\substack{P_X\in\Theta,\\ (g,w)\in\mathcal{G}\times\mathcal{W}:\,g\sim w,\\ a,b\in\mathrm{Im}(g),\,c\in\mathrm{Im}(w)}}\frac{1}{2}\left(\frac{\log\left(\alpha_{a,b,c}^{-1}\right)}{1-\alpha_{a,b,c}}-\beta_{a,b,c}\right)\wedge\sup_{\substack{P_X\in\Theta,\\ (g,w)\in\mathcal{G}\times\mathcal{W}:\,g\sim w,\\ a\in\mathrm{Im}(g),\,c\in\mathrm{Im}(w)}}\log\left(\frac{u_{a,c}}{\ell_{a,c}}\right)\right\}\tag{4}$$

observation also applies for an extension of α-Rényi DP [26] to the structured PP setting.

IV. Properties of MI PP

We now explore the properties of ϵ-MI PP, encompassing convexity, post-processing, and composability (adaptive and non-adaptive). Modern guidelines for privacy frameworks [8] pose properties such as convexity and post-processing (also known as transformation invariance) as base requirements. Composability is another important property that implies that the joint distribution of the outputs of (possibly adaptively chosen) privacy mechanisms is in itself private. These properties are shown to hold for the general ϵ-PP framework in [4]. The next theorem shows ϵ-MI PP satisfies them as well.

Theorem 2 (Properties of ϵ-MI PP Mechanisms)**.** *The following properties hold*:

1) Convexity: Let $\epsilon > 0$, and $M_1, \ldots, M_k$ be ϵ-MI PP mechanisms. Take I as a k-ary categorical random variable with parameters $(p_1, \ldots, p_k)$. Then the mechanism $M := M_I$ (i.e., $M = M_i$ with probability p_i, for $i = 1, \ldots, k$) also satisfies ϵ-MI PP.

2) Post-processing: If mechanism $M : \mathcal{X}^{n\times k} \to \mathcal{Y}$ satisfies ϵ-MI PP, then for any randomized function $A : \mathcal{Y} \to \mathcal{Z}$, the processed mechanism $A \circ M$ also satisfies ϵ-MI PP.

3) Adaptive composability: Let $M_1, \ldots, M_k$ be sequentially and adaptively chosen $\epsilon_1 \ldots, \epsilon_k$-MI PP mechanisms, i.e., $\forall i = 1, \ldots, k$

$$\sup_{\substack{P_X\in\Theta,\\ g\in\mathcal{G},w\in\mathcal{W}:\\ g\sim w}} \mathsf{I}\big(g(X); M_i(X)\big|w(X), M_1(X), \ldots, M_{i-1}(X)\big) \le \epsilon_i.$$

Then the composition $M^k = (M_1, \ldots, M_k)$ satisfies $\big(\sum_{i=1}^k \epsilon_i\big)$-MI PP.

Theorem 2 is proven in Section VIII-B using basic properties of mutual information, such as the chain rule, the data processing inequality, and its nullification under independence. The simplicity of the argument highlights the virtue of the information-theoretic formulation of the PP framework.

We move to discuss non-adaptive composition. In this case, the mechanisms $M_1, \ldots, M_k$ from property (3) of Theorem 2 are chosen conditionally independent given the database. This instance is of practical importance since it includes noise injection mechanisms (e.g., Gaussian and Laplace), that are the focus on the next section. The following proposition is proven in Section VIII-C.

Proposition 1 (Non-Adaptive Composability)**.** *Let $M_1, \ldots, M_k$ be MI PP mechanisms with the parameters $\epsilon_1, \ldots, \epsilon_k$, respectively, which are chosen non-adaptively, i.e., $P_{M_1,\ldots,M_K|X} = \prod_{i=1}^k P_{M_i|X}$. Then the composition M^k is $\big(\sum_{i=1}^k \epsilon_i + \eta\big)$-MI PP, where*

$$\eta = \sup_{\substack{P_X\in\Theta,\\ g\in\mathcal{G},w\in\mathcal{W}:\\ g\sim w}} \sum_{i=2}^k \mathsf{I}\big(M_i(X); M^{i-1}(X)|w(X), g(X)\big),$$

and $M^{i-1}(X) = \big(M_1(X), \ldots, M_{i-1}(X)\big)$.

Proof of Proposition 1 (in Section VIII-C) follows from the repetitive application of chain rule for mutual information and by the fact, entropy being reduced by conditioning.

Remark 5 (Bounds on η)**.** *If $|\mathrm{spt}\big(M_i(X)\big)| < \infty$ for each $i = 2, \ldots, k$, then $\eta \le \sum_{i=2}^k \log\big|\mathrm{spt}\big(M_i(X)\big)\big|$. Thus, if the cardinality of the output of the mechanism is small, then so is η. Alternatively, if each mechanism conditioned on the database X is log-concave (this is satisfied by the Laplace and Gaussian noise injection mechanisms introduced in Section V) and its output is one-dimensional, then*

$$\eta \le \sup_{\substack{P_X\in\Theta,\\ g\in\mathcal{G},w\in\mathcal{W}:\\ g\sim w}} \frac{1}{2}\sum_{i=2}^k \mathbb{E}\left[\log\left(\frac{\pi e \mathrm{Var}\big(M_i(X)\big|g(X), w(X)\big)}{4\mathrm{Var}\big(M_i(X)\big|X\big)}\right)\right].$$

This follows from the Gaussian distribution achieving maximum entropy under a variance constraint, and the lower bound for the entropy of log-concave distributions [44].

Remark 6 (Composition When Secret Pairs are Databases)**.** *It was shown in [4] that standard ϵ-PP mechanisms compose in PP frameworks in which secret pairs $(\mathcal{R}, \mathcal{T}) \in \mathcal{Q}$ correspond to pairs of databases (i.e., when $\mathcal{S}$ contains only singletons; see Definition 3). This also holds for ϵ-MI PP mechanisms by observing that in this case we have $\eta = 0$ in Proposition 1. Indeed, as $\big(g(X), w(X)\big)$ specify a database, the conditional independence of the mechanism given X nullifies the mutual information.*

The general non-adaptive setting, without assuming that secrets specify databases, was studied in [4], where it was shown that composability does not hold in general. Reference [4] then identified a (rather restrictive) sufficient condition on the class of distributions Θ, termed *universally composable* (UC) distributions, under which non-adaptive composability holds for ϵ-PP. The class of UC distributions is defined next.

Definition 6 (UC Distributions)**.** *The class Θ_{UC} of UC distributions contains all $P_X \in \mathcal{P}(\mathcal{X}^{n\times k})$, such that for all $\mathcal{G} \ni g \sim w \in \mathcal{W}$ and $(a, c) \in \mathrm{Im}(g) \times \mathrm{Im}(w)$ with $P_X\big(A_{g,w}(a, c)\big) > 0$, we have $P_{X|A_{g,w}(a,c)} = \delta_x$, for some $x \in \mathcal{X}^{n\times k}$, where δ_x is the Dirac measure at x.*

In words, UC distributions are ones under which the database is specified by non-null secret events.

ϵ-MI PP also composes under the UC condition, but turns out to be more stable than the standard PP framework w.r.t. addition on non-UC distributions to Θ. The next corollary, which follows directly from Proposition 1, quantifies this fact.

Corollary 1 (Universal Composability)**.** *Let $M_1, \ldots, M_k$ be mutual information PP mechanisms with the parameters $\epsilon_1, \ldots, \epsilon_k$, respectively, which are chosen non-adaptively. Then the composition M^k is $\big(\sum_{i=1}^k \epsilon_i\big)$-MI PP, provided either of the following conditions holds*:

(i) $\Theta \subseteq \Theta_{\mathsf{UC}}$; or
(ii) $\Theta_{\mathsf{UC}} \subseteq \Theta$ and $M_1, \ldots, M_k$ satisfy standard PP with the same $\epsilon_1, \ldots, \epsilon_k$ parameters.

The proof of Proposition 1 is given in Proof VIII-D. Notably, Case (ii) above states that non-adaptive composability of ϵ-PP mechanisms holds in the sense of ϵ-MI PP whenever Θ contains all UC distributions. This means that ϵ-MI PP non-adaptive composability is stable to addition of non-UC distributions to Θ, so long that all UC distributions are there (e.g., when $\Theta = \mathcal{P}(\mathcal{X}^{n\times k})$). Standard ϵ-PP does not share this stability: even if Θ contains all UC distribution, adding even a single non-UC distribution to this set will compromise the composability of the classic PP framework.

V. Mechanisms

This section leverages the information-theoretic formulation to devise Laplace and Gaussian noise-injection ϵ-MI PP mechanisms whose noise level is specified in terms of elementary quantities. As a special case of MI PP, we obtain mechanisms for MI DP—a framework proposed and studied in [24], but mechanisms were not considered in that work. Mechanisms achieving MI DP tailored for specific applications, such as linear regression and coded federated learning, were developed in [28] and [29]. In contrast, this sequel provides mechanisms that are applicable in general, under minimal assumptions on the setting.

A. Laplace Mechanism

Given a query function $f : \mathcal{X}^{n\times k} \to \mathbb{R}^d$ and a database $X \sim P_X \in \Theta$, a noise-injection mechanism for privately publishing $f(X)$ outputs $M(X) = f(X) + Z$, where Z is a noise variable that follows a prescribe distribution with appropriately turned parameters. The following theorem characterizes parameter values for the Laplace mechanism (i.e., when Z follows the Laplace distribution) that guarantee ϵ-MI PP.

Theorem 3 (Laplace Mechanism)**.** Fix $\epsilon > 0$ and a structured PP framework $(\mathcal{G}, \mathcal{W}, \mathcal{E}, \Theta)$. Let $f : \mathcal{X}^{n\times k} \to \mathbb{R}^d$ be the query for privatization and consider the Laplace mechanism $M_{\mathsf{L}}(X) := f(X) + Z_{\mathsf{L}}$, where $Z_{\mathsf{L}} \sim \mathrm{Lap}(0, b)^{\otimes d}$ is a d-dimensional product Laplace distribution with the scale parameter $b > 0$. If

$$b \geq \sup_{P_X\in\Theta,\, w\in\mathcal{W}^\star} \frac{\sum_{j=1}^d \mathbb{E}\left[\sqrt{\mathrm{Var}\big(f_j(X)|w(X)\big)}\right]}{d(e^{\frac{\epsilon}{d}}-1)},$$

where $f_j(X)$ is the jth entry of $f(X) = \big(f_1(X), \ldots, f_d(X)\big)$ and $\mathcal{W}^\star = \{w \in \mathcal{W} : \exists\, g \in \mathcal{G},\, g \sim w\}$, then M_{L} is ϵ-MI PP.

The derivation of Theorem 3 is presented in Section VIII-E and relies on the fact that the Laplace distribution maximizes differential entropy subject to an expected absolute deviation constraint.

Remark 7 (Comparison With Wasserstein Mechanism)**.** Compared to computing ∞-Wasserstein distances for each secret pair for each distribution, variance is an elementary quantity that can be computed with relative ease. There are also scenarios where the noise level induced by Wasserstein mechanism is infinite and thus infeasible, while our Laplace mechanism derives a feasible, finite noise level. For instance, consider the setup of AP [7]: if Θ contains a distribution with respect to which $f(X)$ and $g(X)$ are jointly Gaussian for some $g \in \mathcal{G}$, variance is finite while ∞-Wasserstein distance may diverge.[4]

The next corollary specializes Theorem 3 to ϵ-MI DP (see Part 1 of Remark 2) by controlling the conditional variance in terms of ℓ^1-sensitivity. The proof is deferred to Section VIII-F.

Corollary 2 (Laplace Mechanism for ϵ-MI DP)**.** Under the setup of Theorem 3, Laplace mechanism with

$$b \geq \sup_{\substack{P_X\in\mathcal{P}(\mathcal{X}^{n\times k}),\\ i=1,\ldots,n}} \frac{\sum_{\ell=1}^d \mathbb{E}\left[\sqrt{\mathrm{Var}\big(f_\ell(X)|\big(X(j,\cdot)\big)_{j\neq i}\big)}\right]}{d(e^{\frac{\epsilon}{d}}-1)},$$

is ϵ-MI DP. Furthermore, the the statement remains true if the right-hand-side (RHS) above is replaced with $\frac{\Delta_1(f)}{\sqrt{2}d(e^{\frac{\epsilon}{d}}-1)}$, where $\Delta_1(f) := \max_{x\sim x'} \|f(x) - f(x')\|_1$.

Remark 8 (Classic Laplace Mechanisms for DP)**.** Classical Laplace mechanisms achieve ϵ-DP when $b \geq \Delta_1(f)/\epsilon$ [9], and (ϵ, δ)-DP when $b \geq \Delta_1(f)/\big(\epsilon - \log(1-\delta)\big)$ [10]. Evidently, for small ϵ (termed the 'high privacy regime'), both the classic and the ϵ-MI DP mechanism from Corollary 2 induce noise of order $O(1/\epsilon)$.

Remark 9 (Domain Knowledge for DP)**.** Compared to the classic sensitivity-based Laplace mechanisms for DP, the bound in Corollary 2 depends on the variance of f and allows to incorporate domain knowledge. Consider the product Gaussian family

$$\Theta_{\mathsf{G}}(m, s) = \Big\{ \prod_{i=1}^n \theta_i :\ \theta_i = \mathcal{N}(\mu_i, \sigma_i^2),\ |\mu_i| \leq m,\ \sigma_i^2 \leq s \Big\},$$

and let $f(X) = n^{-1}\sum_{i=1}^n X_i$ be the average of the database entries (the argument holds for any linear query). The noise derived from our mechanism is $\sqrt{s}/\big(n(e^\epsilon - 1)\big) < \infty$, while $\Delta_1(f) = \infty$ here since X has unbounded support. Thus, the sensitivity-based mechanisms are vacuous for this case, while our bound provides feasible noise levels. In these situations, the classic approach involves truncating the space [45] which is not necessary under our framework, whenever the variance is finite. In Section VII-A we also discusses the benefits of domain knowledge for private mean estimation tasks.

B. Gaussian Mechanism

We next characterize parameter values for the Gaussian ϵ-MI PP mechanism.

Theorem 4 (Gaussian Mechanism)**.** Fix $\epsilon > 0$ and a structured PP framework $(\mathcal{G}, \mathcal{W}, \mathcal{E}, \Theta)$. Let $f : \mathcal{X}^{n\times k} \to \mathbb{R}^d$ and consider the Gaussian mechanism $M_{\mathsf{G}}(X) := f(X) + Z_{\mathsf{G}}$, where $Z_{\mathsf{G}} \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ is a d-dimensional isotropic Gaussian of parameter

[4]Let the said joint distribution $P_{f(X),g(X)}$ be $\mathcal{N}(\mu, \Sigma)$ with $\mu = (\mu_f, \mu_g)^{\mathsf{T}}$ and $\Sigma = [(\sigma_f^2, \rho\sigma_f\sigma_g); (\rho\sigma_f\sigma_g, \sigma_g^2)]$, Then, $W_2^2\big(P_{f(X)|g(X)=a}, P_{f(X)|g(X)=b}\big) = |a-b|^2\sigma_f^2\rho^2/\sigma_g^2 - (1-\rho^2)\sigma_f^2$. When supremized over $(a, b) \in \mathrm{Im}(g)$, 2-Wasserstein distance diverges. Due to the monotonicity of Wasserstein distance, indeed ∞-Wasserstein distance explodes.

$\sigma > 0$. If

$$\sigma^2 \geq \sup_{P_X \in \Theta,\, w \in \mathcal{W}^\star} \frac{\sum_{j=1}^{d} \mathbb{E}\left[\mathrm{Var}\big(f_j(X)\big|w(X)\big)\right]}{d(e^{\frac{2\epsilon}{d}}-1)},$$

with $\mathcal{W}^\star = \{w \in \mathcal{W} : \exists\, g \in \mathcal{G},\, g \sim w\}$, then M_{G} is ϵ-MI PP.

The derivation of Theorem 4 (Section VIII-G) and uses the fact that the Gaussian distribution maximizes differential entropy subject to a second moment constraint.

Remark 10 (Comparison of Laplace and Gaussian Mechanisms for ϵ-MI PP)**.** In the high-privacy regime (i.e., when ϵ is small), we have $e^{\epsilon}-1 = \Theta(\epsilon)$. The expected noise variance of Laplace mechanism is $\mathbb{E}[\|Z_{\mathsf{L}}\|^2] = \Theta(d/\epsilon^2)$, while the Gaussian mechanism has noise variance $\mathbb{E}[\|Z_{\mathsf{G}}\|^2] = \Theta(d/\epsilon)$. The Gaussian ϵ-MI PP mechanism thus injects noise with a lower variance than the Laplace mechanism in this case. This is illustrated in Fig. 3a for the setting where f is the average of each column of a database in the space $\{0,1\}^{n\times d}$ with fixed $n = 100$ and varying d. Specifically, the figure shows Laplace and Gaussian noise variance needed to achieve ϵ-MI DP for $d = 1, 2, 5, 10, 30$.

Corollary 3 (Gaussian Mechanism for DP)**.** Under the setup of Theorem 4, Gaussian mechanism with

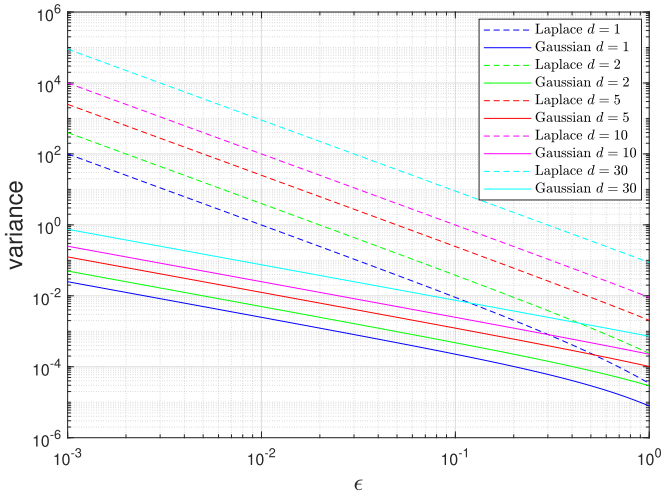
$$\sigma^2 \geq \sup_{\substack{P_X \in \mathcal{P}(\mathcal{X}^{n\times k}),\\ i=1,\ldots,n}} \frac{\sum_{\ell=1}^{d} \mathbb{E}\left[\mathrm{Var}\big(f_\ell(X)\big|\big(X(j,\cdot)\big)_{j\neq i}\big)\right]}{d(e^{\frac{2\epsilon}{d}}-1)},$$

is ϵ-MI DP. Furthermore, the the statement remains true if the RHS above is replaced with $\frac{\Delta_2^2(f)}{2d(e^{\frac{2\epsilon}{d}}-1)}$, where $\Delta_2(f) := \max_{x\sim x'} \|f(x)-f(x')\|$. Additionally, if $\mathcal{X}$ is compact and $f : \mathcal{X}^{n\times k} \to \mathbb{R}$ is continuous, then $\sigma^2 \geq \Delta_2^2(f)/\big(4(e^{2\epsilon}-1)\big)$ is sufficient to achieve ϵ-MI DP.

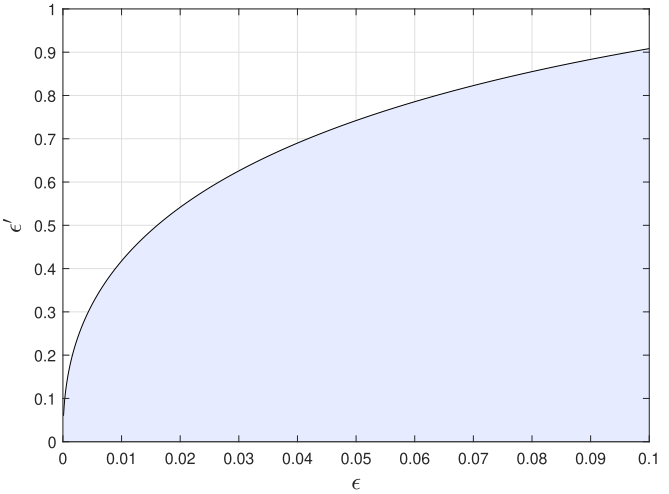
Remark 11 (Classic Gaussian Mechanisms for DP)**.** By Theorem 1 with the condition $\Theta = \mathcal{P}(\mathcal{X}^{n\times k})$, which holds in the DP setting, we have that ϵ-MI DP implies regular $(\epsilon', \sqrt{2\epsilon})$-DP, for any $\epsilon' \geq 0$. For comparison, the classical Gaussian mechanism achieves $(\epsilon', \sqrt{2\epsilon})$-DP with $\sigma^2 \geq 2\log(1.25/\sqrt{2\epsilon})\Delta_2^2(f)/{\epsilon'}^2$ for $\epsilon' \leq 1$ [9]. It can be shown that our mechanism requires a lower noise level than the classic one whenever $\epsilon' < 2\big(d(e^{2\epsilon/d}-1)\log(1.25/\sqrt{2\epsilon})\big)^{1/2}$. Fig. 3b shows the region of (ϵ', ϵ) values for which our mechanism injects noise of a lower variance for $d = 30$. We also note that this region is monotonically decreasing (in the sense of inclusion) with d.

The noise levels derived in Theorem 4 scales linearly with the increasing dimension of the query. This may result in large noise values which may compromise utility. Projecting the high-dimensional queries to a low-dimensional space may provide a better privacy-utility tradeoff if the dimension of the projection space is chosen appropriately.

Theorem 5 (Gaussian Mechanism With Projections)**.** Fix $\epsilon > 0$, a structured PP framework $(\mathcal{G}, \mathcal{W}, \mathcal{E}, \Theta)$. Let $f : \mathcal{X}^{n\times k} \to \mathbb{R}^d$ be the query function, $\mathrm{A} \in \mathbb{R}^{d\times \ell}$ be a projection matrix with $\ell \leq d$, and consider the Gaussian mechanism



(a)



(b)

Fig. 3. (a) Laplace and Gaussian noise variance injected to achieve ϵ-MI DP for the following setting where f is the average of each column of the database in the space $\{0,1\}^{n\times d}$ with fixed $n = 100$ and varying $d = 1, 2, 5, 10, 30$. (b) The region where the MI DP Gaussian noise injection mechanism adds noise with smaller variance compared to classical mechanism for achieving $(\epsilon', \sqrt{2\epsilon})$-DP with $d = 30$.

$M_{\mathsf{G}}^{\mathsf{proj}}(X) := \mathrm{A}^{\intercal} f(X) + Z_{\mathsf{G}}$, where $Z_{\mathsf{G}} \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_\ell)$. Then $M_{\mathsf{G}}^{\mathsf{proj}}$ is ϵ-MI PP if either of the following conditions hold:

1) $\mathrm{A} = [\phi_1, \ldots, \phi_\ell]$ is deterministic and

$$\sigma^2 \geq \sup_{P_X \in \Theta,\, w \in \mathcal{W}^\star} \frac{\mathbb{E}\left[\|\Sigma_{f|w}\|_{\mathrm{op}}\right] \max_{1\leq j\leq \ell} \|\phi_j\|^2}{(e^{\frac{2\epsilon}{\ell}}-1)},$$

where $\|\cdot\|_{\mathrm{op}}$ is the operator norm, $\Sigma_{f|w}$ is the conditional covariance matrix of $f(X)$ given $w(X)$, and $\mathcal{W}^\star$ is as in Theorem 3.

2) $\mathrm{A} = [\Phi_1, \ldots, \Phi_\ell]$ is random, chosen independently of the database X and the mechanism M, with $\mathbb{E}[\|\Phi_j\|^2] = 1$ and $\mathbb{E}[\Phi_j] = \mathbf{0}$ for all $j = 1, \ldots, \ell$, and

$$\sigma^2 \geq \sup_{P_X \in \Theta,\, w \in \mathcal{W}^\star} \frac{\mathbb{E}\left[\|\Sigma_{f|w}\|_{\mathrm{op}} + \|\mu_{f|w}\|_2^2\right]}{(e^{\frac{2\epsilon}{\ell}}-1)},$$

where $\mu_{f|w} := \mathbb{E}\big[f(X)\big|w(X)\big]$.

Remark 12 (Gaussian Projection Matrix)**.** The random matrix whose entries are sampled independently from the Gaussian distribution with 0 mean and $1/d$ variance satisfies the requirements of Theorem 5, Part (2). A similar approach was proposed in [46] for the task of estimating distances between users without being leaking private information. Their ϵ-DP mechanism first projected the query onto a random lower-dimensional space via Johnson-Lindenstrauss transform and then injected Gaussian noise. Theorem 5 thus enables using ϵ-MI PP in such settings.

Remark 13 (Projection Dimension)**.** For a d-dimensional query, the Gaussian mechanism without projections (Theorem 4) adds noise proportional to d, which may be prohibitive when $d \gg 1$. Theorem 5 shows that by incorporating a projection matrix, one may inject noise that is proportional to ℓ, where $\ell \ll d$. In practice, the projection dimension ℓ should be tuned to optimize the privacy-utility tradeoff, keeping in mind that larger ℓ would typically necessitate larger σ^2 values to guarantee privacy at a prescribed level [46].

C. Mechanisms With Explicit Dependence on Private Functions

The noise levels derived in Theorem 4 and 3 depend on the private function class $\mathcal{G}$ only through $\mathcal{W}^\star$. However, it may be desirable to capture the dependence on $\mathcal{G}$ more explicitly. This is particularly relevant when there is no public information (e.g., the AP framework from [7]; cf. Remark 2 Part 2) or if there is a single public function w corresponding to all private $g \in \mathcal{G}$. The following theorem provides noise levels with such explicit dependence.

Theorem 6 (Gaussian Mechanism With Dependence on $\mathcal{G}$)**.** Under the setup from Theorem 4 and assuming $\inf_{\substack{g\in\mathcal{G},w\in\mathcal{W}:\\ g\sim w}} \mathsf{h}\big(f(X)\big|g(X),w(X)\big) > -\infty$, the Gaussian mechanism M_{G} achieves ϵ-MI PP, if

$$\sigma^2 \geq \sup_{\substack{P_X\in\Theta,\\ g\in\mathcal{G},w\in\mathcal{W}:\\ g\sim w}} \frac{A - d\, e^{2\epsilon/d} B}{d(e^{2\epsilon/d}-1)} \vee 0\,,$$

with $A = \sum_{j=1}^{d} \mathbb{E}\big[\mathrm{Var}\big(f_j(X)\big|w(X)\big)\big]$ and $B = \frac{1}{2\pi}\exp\big(\frac{2}{d}\mathsf{h}\big(f(X)\big|g(X),w(X)\big) - 1\big)$.

In addition to maximum entropy arguments, the proof of Theorem 6 uses the entropy power inequality. We may replace the conditional entropy in B by any lower bound that may be easier to compute (cf. e.g., [44]), and ϵ-MI PP will still hold.

Remark 14 (Free ϵ-MI PP Regime)**.** The bound in Theorem 6 suggests that if $A \leq d\, e^{2\epsilon/d} B$ over the entire optimization domain, ϵ-MI PP holds without noise injection (i.e., $\sigma = 0$). It can be shown that $A \geq dB$ for any $P_X \in \mathcal{P}(\mathcal{X}^{n\times k})$ and functions f, g, and w. The free privacy regime therefore corresponds to cases where ϵ is large compared to $d/2$. Since large ϵ values are rarely of interest in practice, we conclude that a positive noise level is generally needed for ϵ-MI PP. For fixed ϵ and d, the above condition is related to how correlated the query and the private functions are, given the public information. For instance, if $d = 1$ and $f(X)$, $g(X)$, and $w(X)$ are jointly Gaussian, we have $A \leq d\, e^{2\epsilon/d} B$ whenever the conditional correlation coefficient between $f(X)$ and $g(X)$ given $\{w(X) = c\}$ satisfies $\rho\big(f(X), g(X)\big|w(X) = c\big) \leq \sqrt{(e^{2\epsilon}-1)e^{-2\epsilon}}$. Accordingly, weak correlation may lead to free privacy since the query leaks little information about the secret to begin with. Proofs related to the above arguments are presented in Section VIII-K

A Gaussian mechanism with explicit dependence on the secret functions was proposed for AP in [7], under a rather stringent setting. In their AP formulation there are not public functions (i.e., $\mathcal{W} = \emptyset$) and taking $d = 1$, they assume that $f(X)$ conditioned on $g(X)$ is Gaussian with a constant variance, i.e., $\mathrm{Var}\big(f(X)\big|g(X) = a\big) = \mathrm{Var}\big(f(X)\big|g(X) = b\big)$, for all $a, b \in \mathrm{Im}(g)$. Theorem 1 of [7] then shows that (ϵ, δ)-AP is achieved by the Gaussian mechanism with

$$\sigma^2 \geq \sup_{\substack{P_X\in\Theta,\\ g\in\mathcal{G}}} \left[\left(\frac{C\Delta_{\mathsf{AP}}(f)}{\epsilon}\right)^2 - \mathrm{Var}\big(f(X)\big|g(X)=a\big)\right] \vee 0,$$

where $C = \sqrt{2\log(1.25/\delta)}$ and

$$\Delta_{\mathsf{AP}}(f) = \max_{a,b\in\mathrm{Im}(g)} \big|\mathbb{E}\big[f(X)\big|g(X)=a\big] - \mathbb{E}\big[f(X)\big|g(X)=b\big]\big|.$$

By means of comparison, the following corollary specializes our Theorem 6 to the setting from [7].

Corollary 4 (Gaussian Mechanism for AP)**.** *Under the above setting, the Gaussian mechanism with the variance parameter*

$$\sigma^2 \geq \sup_{\substack{P_X\in\Theta,\\ g\in\mathcal{G}}} \left[\frac{\mathrm{Var}(f(X)) - \ e^{2\epsilon}\mathrm{Var}\big(f(X)\big|g(X)=a\big)}{e^{2\epsilon}-1}\right] \vee 0$$

satisfies ϵ-MI attribute privacy.

Note that both the mechanisms enter the free privacy regime when $f(X)$ is independent of $g(X)$ (Remark 14 above argues that this holds for our mechanism even when there is a weak dependence between $g(X)$ and $f(X)$). Under a multivariate extension of the product Gaussian family from Remark 9, i.e.,

$$\Theta^k_{\mathsf{G}}(m,s) = \Big\{\mathcal{N}(\mu,\Sigma)^{\otimes n}:\ \mu = (\mu_1 \dots \mu_k)^{\mathsf{T}}, |\mu_j| \leq m, \\ \Sigma(i,j) \leq s,\ \forall i = 1,\dots,n,\ j = 1,\dots,k\Big\}$$

and for f and g linear functions of the database columns (say, average of the column entries), $\Delta_{\mathrm{AP}}(f)$ is proportional to $\max_{a,b\in\mathrm{Im}(g)}|a-b|$ and thus diverges to infinity. The variance-based bound from Corollary 4, on the other hand, is finite and feasible.

VI. Auditing for Privacy

Privacy auditing aims to detect violations in privacy guarantees, reject incorrect algorithms, and provide counterexamples. This concept has been gaining recent attention for the special case of DP auditing. In [14], an heuristic approach based on poisoning attacks was proposed for auditing privacy violations of DP based stochastic gradient descent. The idea of formulating DP auditing as a hypothesis test was originally explored in [13] for univariate queries. An extension to the

multivariate query setting was proposed in [15] by relaxing DP to a privacy notion based on kernel Rényi divergence, and using an empirical version of the latter as a test statistic. However, all these approaches are tailored for classical DP and are not applicable beyond that setting, e.g., to PP auditing.

We propose a hypothesis testing pipeline to audit for ϵ-MI DP which readily extends to the PP setting (see Remark 16 ahead). From relative strength considerations, our approach can, in turn, audit for any stricter privacy notion, such as ϵ-KL DP, (α, ϵ)-Rényi DP, or ϵ-DP itself. Indeed, since ϵ-MI DP is a relaxation of ϵ-DP, any mechanism that violates the former must also violate the latter. Our audit tests between the null hypothesis $\mathcal{H}_0$, that ϵ-MI DP holds, and the alternative $\mathcal{H}_1$ using an estimate of the mutual information from (2) as the test statistic. If the estimate is larger than the threshold, we reject the null and declare the mechanism as violating ϵ-MI DP (and thus also ϵ-DP). When auditing for MI DP, the only source of error in the decision is the statistical estimation error. When auditing DP, on the other hand, an extra slackness may arise from the gap between the DP constraint (say, in terms of ∞-Renyi divergence) and the relaxed mutual information-based one.

The main challenge of the proposed approach lies in the inherent hardness of estimating mutual information. For continuous, high-dimensional random variables (which is often the regime of interest in modern privacy applications), the sample complexity of estimating mutual information grows exponentially with dimension [47], [48], making tests based on ϵ-MI DP infeasible. Fortunately, for auditing purposes we may further relax the ϵ-MI DP privacy notion in order to gain sample efficiency of the test. To that end, we propose to employ sliced mutual information (SMI). Note that this relaxation also comes at a cost in terms of test power due to the gap between mutual information and SMI.

A. Sliced Mutual Information

SMI was introduced in [16] as an information measure that preserves many properties of classic (Shannon) mutual information, while being amenable to scalable estimation from high-dimensional samples. SMI is defined as an average of mutual information terms between one-dimensional projections of the considered random variables, namely, for $(X, Y) \sim P_{XY} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ it is given by

$$\mathsf{SI}(X;Y) := \int_{\mathbb{S}^{d_x-1}} \int_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\intercal X; \phi^\intercal Y) d\sigma_{d_x}(\theta) d\sigma_{d_y}(\phi), \quad (5)$$

where $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\| = 1\}$ is the unit sphere in $\mathbb{R}^d$ and σ_d is the uniform distribution on it (cf. [17] for an extension to k-dimensional projections). Proposition 1 of [16] shows that SMI satisfies many properties akin to classic mutual information, such as identification of independence, (sliced) entropy-based decompositions, tensorization, variational forms, and more. By the data processing inequality, SMI is always upper bounded by classic mutual information, i.e., $\mathsf{SI}(X;Y) \leq \mathsf{I}(X;Y)$, which enables using it to define a relaxed privacy notion.

We also recall the sliced entropy and conditional SMI. For $(X, Y, Z) \sim P_{XYZ} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \mathbb{R}^{d_z})$ and $(\Theta, \Phi, \Psi) \sim \sigma_{d_x} \otimes \sigma_{d_y} \otimes \sigma_{d_z}$, the sliced entropy of X, its conditional version given Y, and the condition SMI between X and Y given Z are defined, respectively, by

$$\begin{aligned} \mathsf{sh}(X) &:= \mathsf{h}(\Theta^\intercal X|\Theta) \\ \mathsf{sh}(X|Y) &:= \mathsf{h}(\Theta^\intercal X|\Theta, \Phi, \Phi^\intercal Y) \\ \mathsf{SI}(X;Y|Z) &:= \mathsf{I}(\Theta^\intercal X; \Phi^\intercal Y|\Theta, \Phi, \Psi, \Psi^\intercal Z). \end{aligned}$$

With these definitions, we have $\mathsf{SI}(X;Y) = \mathsf{sh}(X) - \mathsf{sh}(X|Y)$ and $\mathsf{SI}(X;Y,Z) = \mathsf{SI}(X;Y) + \mathsf{SI}(X;Z|Y)$, among others.

B. Sliced Mutual Information Differential Privacy

A randomized mechanism M is said to satisfy ϵ-SMI DP (w.r.t. the distribution class $\Theta \subseteq \mathcal{P}(\mathcal{X}^{n\times k})$ if

$$\sup_{\substack{P_X \in \Theta, \\ i=1,\ldots,n}} \mathsf{SI}\big(X(i,\cdot); M(X) \big| \big(X(j,\cdot)\big)_{j\neq i}\big) \leq \epsilon; \quad (6)$$

here we unfold $\big(X(j,\cdot)\big)_{j\neq i}$ into a vector of size $k(n-1)$ and then project it. Evidently, this is similar to the definition of ϵ-MI DI from [24] (Definition 2) but with SMI replacing MI and while allowing for distributional assumptions on the database (as in the structured PP framework). We henceforth make the simplifying assumption that the database is independent across records. Namely, denoting the marginal distribution of the ith row $X(i,\cdot)$ by P_i, we assume that $X \sim P_X = \prod_{i=1}^n P_i$. While the subsequent results are derived under this restriction, we expect the ideas to naturally extend to general data distributions and PP frameworks. The following proposition states the ϵ-SMI DP is a relaxation of ϵ-MI DP.

Proposition 2 (ϵ-MI DP Relaxation)**.** *If a randomized mechanism $M : \mathcal{X}^{n\times k} \to \mathcal{Y}$ is ϵ-MI DP then it is also ϵ-SMI DP.*

The proof is immediate since under the independence assumption (and hence it is omitted), we have

$$\begin{aligned} &\mathsf{SI}\big(X(i,\cdot); M(X) \big| \big(X(j,\cdot)\big)_{j\neq i}\big) \\ &= \mathsf{SI}\big(X(i,\cdot); M(X), \big(X(j,\cdot)\big)_{j\neq i}\big) \\ &\leq \mathsf{I}\big(X(i,\cdot); M(X), \big(X(j,\cdot)\big)_{j\neq i}\big) \\ &= \mathsf{I}\big(X(i,\cdot); M(X) \big| \big(X(j,\cdot)\big)_{j\neq i}\big). \end{aligned}$$

Thus, we deduce that if M violates ϵ-SMI DP then it cannot be ϵ-MI DP nor ϵ-DP. In fact, it suffices to find a single distribution $P_X \in \Theta$ for which (6) does not hold to reject the null hypothesis and declare violation of ϵ-DP. These are the main observations for devising the SMI-based audit for DP (Section VI-D). Another key component of the test is scalability with which SMI can be estimated, which is formulated next.

C. Sliced Mutual Information Estimation

We consider estimating of SMI statistic from the left-hand-side (LHS) of 6, but for a fixed distribution $P_X \in \Theta$ (a further relaxation). Defining the shorthands $X_i = X(i,\cdot)$, $Y = M(X)$, and $Z_i = \big(X(j,\cdot)\big)_{j\neq i}$, our objective is thus $\max_{i=1,\ldots,n} \mathsf{SI}_i$, where $\mathsf{SI}_i := \mathsf{SI}(X_i; Y|Z_i)$. We first describe a Monte Carlo based estimation procedure for SMI, employing a generic mutual information estimator between scalar

variables. Afterwards, we instantiate the generic estimator via the neural estimation framework of [49] and provide formal convergence guarantees.

1) Monte Carlo Estimate of SMI: Fix $i = 1, \dots, n$, let $\{(X_i^j, Y^j, Z_i^j)\}_{j=1}^m$ be m i.i.d. samples of (X_i, Y, Z_i), and proceed as follows:

1) Draw $\{(\Theta_j, \Phi_j, \Psi_j)\}_{j=1}^p$ i.i.d. projection triples from $\sigma_k \otimes \sigma_d \otimes \sigma_{(n-1)k}$.
2) For each $j = 1, \dots, m$ and $\ell = 1, \dots, p$, compute $(\Theta_\ell^\intercal X_i^j, \Phi_\ell^\intercal Y^j, \Psi_\ell^\intercal Z_i^j)$.
3) For each $\ell = 1, \dots, p$, we estimate the mutual information $\mathsf{I}\big(\Theta_\ell^\intercal X_i; \Phi_\ell^\intercal Y, \Psi_\ell^\intercal Z_i\big)$ using the estimate $\hat{\mathsf{I}}\big((\Theta_\ell^\intercal X_i)^m, (\Phi_\ell^\intercal Y)^m, (\Psi_\ell^\intercal Z_i)^m\big)$ (to be described shortly), where
$$(\Theta_\ell^\intercal X_i)^m := (\Theta_\ell^\intercal X_i^1, \dots, \Theta_\ell^\intercal X_i^m)$$
and $(\Phi_\ell^\intercal Y)^m$, $(\Psi_\ell^\intercal Z_i)^m$ are defined similarly.
4) Take a Monte-Carlo average of the above estimates, resulting in the SMI estimator:
$$\widehat{\mathsf{SI}}_i^{m,p} := \frac{1}{p} \sum_{\ell=1}^{p} \hat{\mathsf{I}}\big((\Theta_\ell^\intercal X_i)^m, (\Phi_\ell^\intercal Y)^m, (\Psi_\ell^\intercal Z_i)^m\big). \tag{7}$$
5) Set the SMI DP statistic estimator as
$$\widehat{\mathsf{SI}}^{m,p} := \max_{i=1,\dots,n} \widehat{\mathsf{SI}}_i^{m,p}. \tag{8}$$

2) Neural Estimation of Mutual Information: We now instantiate the generic mutual information estimator $\hat{\mathsf{I}}(\cdot,\cdot,\cdot)$ in Step 3 via the neural estimation framework of [49], and obtain explicit estimation error bounds in terms of m, p, and the size of the neural network. The appeal of this approach is twofold: (i) the SMI neural estimator [17] lower bounds the population objective in the large sample limit, thus serving as a further relaxation which is in line with the auditing pipeline; and (ii) neural estimators are efficiently computable via standard gradient-based optimizers and have low memory footprint, even for high-dimensional data and massive sample sets.

Neural estimation of mutual information relies on the Donsker-Varadhan (DV) variational form, whereby
$$\mathsf{I}(U;V) = \sup_{f:\mathbb{R}^{d_u}\times\mathbb{R}^{d_v}\to\mathbb{R}} \mathbb{E}[f(U,V)] - \log\left(\mathbb{E}\big[e^{f(\tilde{U},\tilde{V})}\big]\right),$$
where $(U,V) \sim P_{UV} \in \mathcal{P}(\mathbb{R}^{d_u} \times \mathbb{R}^{d_v})$, $(\tilde{U},\tilde{V}) \sim P_U \otimes P_V$, and f is a measurable function for which the expectations above are finite. Given i.i.d. data $(U_1,V_1), \dots, (U_m,V_m)$ from P_{UV}, the neural estimator parameterizes the DV potential f by an ℓ-neuron shallow network and approximates expectations by sample means,[5] resulting in the estimate
$$\hat{\mathsf{I}}_\ell(U^m, V^m) := \sup_{g\in\mathcal{G}_\ell} \frac{1}{m}\sum_{i=1}^m g(U_i,V_i) - \log\left(\frac{1}{n}\sum_{i=1}^m e^{g(U_i,V_{\sigma(i)})}\right),$$
where the neural network function class is defined as
$$\mathcal{G}_\ell(a) := \left\{ \begin{array}{c} g:\mathbb{R}^3\to\mathbb{R}: \\ g(z) = \sum_{i=1}^\ell \beta_i \phi\left(\langle w_i, z\rangle + b_i\right) + \langle w_0, z\rangle + b_0, \\ \max_{1\le i\le \ell} \|w_i\|_1 \vee |b_i| \le 1, \\ \max_{1\le i\le \ell} |\beta_i| \le \frac{a}{2\ell}, \\ |b_0|, \|w_0\|_1 \le a \end{array} \right\}$$
with $\phi(z) = z \vee 0$ as the ReLU activation, and we set the shorthand $\mathcal{G}_\ell = \mathcal{G}_\ell(\log\log\ell \vee 1)$. Inserting $\hat{\mathsf{I}}_\ell$ with $U^m = (\Theta_\ell^\intercal X_i)^m$ and $V^m = \big((\Phi_\ell^\intercal Y)^m, (\Psi_\ell^\intercal Z_i)^m\big)$ as the generic mutual information estimator in (7), the neural estimator of SI_i is
$$\widehat{\mathsf{SI}}_{i,\mathsf{NE}}^{\ell,m,p} := \frac{1}{p}\sum_{j=1}^p \hat{\mathsf{I}}_\ell\big((\Theta_j^\intercal X_i)^m, (\Phi_j^\intercal Y)^m, (\Psi_j^\intercal Z_i)^m\big),$$
and the corresponding SMI DP statistic estimator is $\widehat{\mathsf{SI}}_{\mathsf{NE}}^{\ell,m,p} := \max_{i=1,\dots,n} \widehat{\mathsf{SI}}_{i,\mathsf{NE}}^{\ell,m,p}$. Note the $\widehat{\mathsf{SI}}_{i,\mathsf{NE}}^{\ell,m,p}$ is readily implemented by parallelizing m ℓ-neuron ReLU nets with inputs in $\mathbb{R}^3$ and scalar outputs.

3) Formal Guarantees: Drawing upon the results of [49] for neural estimation of f-divergences, we provide non-asymptotic error bounds for $\widehat{\mathsf{SI}}_{\mathsf{NE}}^{\ell,m,p}$, subject to certain regularity assumptions on $P_{X,M(X)}$. Suppose that $\mathcal{X}^{n\times k} \subset \mathbb{R}^{n\times k}$ and $\mathcal{Y} \subset \mathbb{R}^d$ are compact sets and that P_{XY} (recall that $Y = M(X)$) has a Lebesgue density f_{XY} supported on $\mathcal{X}^{n\times k} \times \mathcal{Y}$. For $b, \eta \ge 0$, let $\mathcal{F}(b,\eta) \subseteq \mathcal{P}(\mathcal{X}^{n\times k} \times \mathcal{Y})$ be the distribution class that contains all P_{XY} as above that also satisfy the following property: $\exists\, r \in \mathsf{C}_b^4(\mathcal{U})$ for some open set $\mathcal{U} \supset \mathcal{X}^{n\times k} \times \mathcal{Y}$, such that $\log f_{XY} = r|_{\mathcal{X}^{n\times k}\times\mathcal{Y}}$, and $\max_{i=1,\dots n} \mathsf{I}(X_i; M(X), Z_i) \le \eta$ (namely, the log-density has a smooth extension to an open set $\mathcal{U}$ containing the support). In particular, this class contains distributions whose densities are smooth and bounded from above and below and admit the aforementioned smooth extension condition. This includes uniform distributions, truncated Gaussians, truncated Cauchy distributions, etc.

We next provide convergence rates for the SMI DP statistic $\widehat{\mathsf{SI}}_{\mathsf{NE}}^{\ell,m,p}$, uniformly over the class $\mathcal{F}(b,\eta)$, characterizing the dependence of the error on s, m, and ℓ. To simplify the bound we assume $\|\mathcal{X}\| = \|\mathcal{Y}\| = 1$, i.e., that the feature and the mechanism output spaces are normalized. The results readily extend to arbitrary compact domains.

Proposition 3 (Neural Estimation Error)**.** *For any $\eta, b \ge 0$, we have*
$$\sup_{P_{XY}\in\mathcal{F}(b,\eta)} \mathbb{E}\left[\left|\max_{i=1,\dots,n} \mathsf{SI}_i - \widehat{\mathsf{SI}}_{\mathsf{NE}}^{\ell,m,p}\right|\right] \le C\, n^3 k^2 \big(\ell^{-\frac{1}{2}} + m^{-\frac{1}{2}} + p^{-\frac{1}{2}}\big),$$
where C is a constant that depend only on η and b.

Proposition 3 follows by bounding
$$\mathbb{E}\left[\left|\max_{i=1,\dots,n} \mathsf{SI}_i - \widehat{\mathsf{SI}}_{\mathsf{NE}}^{\ell,m,p}\right|\right] \le \sum_{i=1}^n \mathbb{E}\left[\left|\mathsf{SI}_i - \widehat{\mathsf{SI}}_{i,\mathsf{NE}}^{\ell,m,p}\right|\right],$$
and then applying the SMI neural estimation bound from [17] to the RHS above (cf. [49, Proposition 2]). Further details are omitted for brevity.

[5]Negative samples, i.e., from $P_U \otimes P_V$, can be obtained from the positive ones via $(U_1, V_{\sigma(1)}), \dots, (U_m, V_{\sigma(m)})$, where $\sigma \in S_m$ is a permutation such that $\sigma(i) \neq i$, for all $i = 1, \dots, m$.

D. Auditing Differential Privacy via SMI DP

We now present the hypothesis testing pipeline for auditing ϵ-SMI DP. Notably, violations on the latter also implies a violation of ϵ-DP. We consider a composite hypothesis test between the null $H_0: \max_{i=1,\ldots,n} \mathsf{SI}_i \leq \epsilon$ (i.e., ϵ-SMI DP holds), and the alternative H_1. Given samples $\{(X_i^j, Y^j, Z_i^j)\}_{(i,j)=(1,1)}^{(n,m)}$ from the database and the mechanism, we use the maximized SMI estimator $\widehat{\mathsf{SI}}_{\mathsf{NE}}^{\ell,m,p}$ as our test statistic. An immediate consequence of Proposition 3 and Markov's inequality is the following Type 1 error bound.

Proposition 4 (Type-I Error)**.** *Fix arbitrary $\epsilon, r > 0$ and consider the above setup. We have*

$$\mathbb{P}\left(\widehat{\mathsf{SI}}_{\mathsf{NE}}^{\ell,m,p} > \epsilon + r \,\middle|\, H_0\right) \leq C\frac{n^3k^2}{r}\left(\ell^{-\frac{1}{2}} + m^{-\frac{1}{2}} + p^{-\frac{1}{2}}\right),$$

where C is the constant from Proposition 3.

Choosing r and the estimator parameters such that the error probability is not significant, for instance, $r \asymp C\frac{n^3k^2}{\alpha}\left(\ell^{-\frac{1}{2}} + m^{-\frac{1}{2}} + p^{-\frac{1}{2}}\right)$ with $\alpha \in (0,1)$, and $\ell = m = p \asymp n^6k^4$, we obtain an hypothesis test with level α significance. Indeed, under the null, the rejection probability of this test is below α. A power analysis of the proposed test (namely, the Type II error) is also of significant interest since it provides guarantees for identifying privacy violating mechanisms. This, however, requires more advanced machinery such as a limit distribution theory for the test statistic using which a power against local alternatives can be quantified. Since a refined statistical analysis of SMI estimators is beyond the scope of this work, we leave the power analysis of the above test for future work.

Remark 15 (Auditing Variants of DP Frameworks)**.** The above hypothesis test can be used to audit for various privacy framework, including ϵ-DP, (α,ϵ)-Rényi DP [26] with $\alpha \geq 1$, ϵ-MI DP, etc. This is since all these framework are stronger than (and hence imply) ϵ-SMI DP. Furthermore, when the database distribution has finite support and a bounded density that satisfies the conditions from Theorem 1, the implication $(\epsilon',\delta') - \text{DP} \implies \epsilon^* - \text{MI DP}$ holds and the hypothesis test can also audit for (ϵ,δ)-DP algorithms.

Remark 16 (Auditing PP Frameworks)**.** These ideas readily extended to auditing of PP frameworks. In particular, when $\mathcal{W} = \mathcal{E} = \emptyset$ in the structured PP framework from Definition 4, the above procedure can be adapted even without requiring that the database rows are i.i.d. (as needed in the case of DP).

VII. Additional Applications

A. Private Mean Estimation

Differentially-private mean estimation is a basic private statistical inference tasks, which was widely studied under the classic DP paradigm [45], [50], [51]. We now revisit this problem and quantify the sample complexity of ϵ-MI DP multivariate mean estimation. Potential gains of incorporating domain knowledge into the framework are also discussed. Let $X \sim P_X \in \mathcal{P}(\mathbb{R}^d)$ be d-dimensional random variable with mean $\mathbb{E}[X] = \mu$. Given n i.i.d. samples of X, the goal is obtain an accurate estimate $\hat{\mu}_n$ of its mean that also satisfies ϵ-MI DP. We propose Algorithm 1 as the procedure for doing so.

Algorithm 1 ϵ-MI DP Algorithm for Mean Estimation

Input: n i.i.d. data samples $X_1, \ldots, X_n$; ϵ; β
$m \leftarrow 200\log(1/\beta)$
$k \leftarrow \lfloor n/m \rfloor$
$\sigma^2 \leftarrow \frac{dm^2}{2n^2\epsilon}$
for $p = 1 : m$ **do**
  $\tilde{\mu}_p \leftarrow \frac{1}{k}\left(X_{(p-1)k+1} + \ldots + X_{pk}\right) + Z_p,\ Z_p \sim \mathcal{N}(0, \sigma^2\mathrm{I}_d)$
end for
$\hat{\mu}_n \leftarrow \operatorname{argmin}_{y\in\mathbb{R}^d} \sum_{p=1}^{m} \|y - \tilde{\mu}_p\|$
Output: $\hat{\mu}_n$

Proposition 5 (Mean Estimation Under ϵ-MI DP)**.** Fix $\alpha, \beta, \epsilon > 0$. Let $X \sim P_X \in \mathcal{P}(\mathbb{R}^d)$ have mean $\mathbb{E}[X] = \mu$ and a bounded absolute second moment $\mathbb{E}\left[\|X - \mu\|^2\right] < \infty$. Then Algorithm 1 fed with $n \geq n_0$ data samples, where

$$n_0 = O\left(\log(1/\beta)\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\sqrt{\epsilon}}\right)\right),$$

outputs an estimated $\hat{\mu}_n$ that satisfies ϵ-MI DP and achieves $\mathbb{P}\left(\|\hat{\mu}_n - \mu\| \leq \alpha\right) \geq 1 - \beta$.

Proposition 5 is proven in Section VIII-L. The argument uses Theorem 4 to derive noise levels under which Algorithm 1 attains ϵ-MI DP, and then applies median of means techniques (specifically geometric median) to improve the accuracy of the estimate.

Remark 17 (Sensitivity-Based Mechanisms)**.** Private mean estimation under other variants of DP, such as ϵ-DP, (ϵ,δ)-DP, and ρ-concentrated DP [52], typically employs standard, sensitivity-based noise injection mechanisms. However, these require knowledge of an upper bound on the mean, i.e., R such that $\|\mu\| \leq R$. For example, [45] propose a mechanism that is initiated using a rough proxy of R which is iteratively refined using the data samples. The ϵ-MI DP mechanism proposed herein, on the other hand, does not require boundedness so long that the distribution of interest has finite variance. This is due to the fact that ϵ-MI DP incorporates domain knowledge regarding the class of distributions, as opposed to the aforementioned DP variants that guarantee privacy in the worst-case (and are hence stricter).

Remark 18. [Computational Complexity of Algorithm 1] For $d = 1$, Algorithm 1 boils down to computing the median of the means obtained from m rounds. In the multivariate case, our algorithm requires evaluating the geometric median, which is a computationally hard problem. Nevertheless, there are near-linear time algorithms for approximate geometric median computation, i.e., with complexity $O\left(dm(\log(1/\alpha))^3\right)$ where $\alpha > 0$ is the approximation parameter [53].[6] To reduce the computational burden, one may consider coordinate-wise median, whose complexity is $O\left(dm\log(m)\right)$; the linear dependence on d can further be alleviated by parallelization.

[6]Another approach for computing a multivariate median is the smallest-ball median method, but it is computationally intractable [54].

However, the sample complexity needed to to achieve the same level of accuracy α using the coordinate-wise median worsens to $n_0 = O\big(\log(d/\beta)\big(d\alpha^{-2} + d\alpha^{-1}\epsilon^{-1/2}\big)\big)$.

Remark 19 (Sample Complexity of ϵ-DP Algorithms)**.** The sample complexity of ϵ-differentially private mean estimation was evaluated in [45], under the assumption that $\|\mu\| \leq R$, for some known R, and that the $k \geq 2$ moment of $X \sim P_X$ is bounded. Their result reads as

$$n_0 = O\left(\log(d/\beta)\left(\frac{d}{\alpha^2} + \frac{d}{\epsilon\alpha^{k/k-1}} + \frac{d\log(R)}{\epsilon}\right)\right).$$

The achieving algorithm first truncates the data into a bounded region so as to limit the amount of injected noise needed for privacy, and then performs mean estimation in a differentially private manner. To find the truncation range they use an iterative procedure which depends on the known R. This approach, however, becomes computationally infeasible when dimension is large. While the ϵ-DP sample complexity bound above grows like $1/\epsilon$, the one corresponding to ϵ-MI DP is on the order of $1/\sqrt{\epsilon}$ (see Proposition 5). Nevertheless, it is important to note that the privacy guarantees provided by these approaches are different, with ϵ-DP being strictly stronger.

B. Algorithmic Stability

Conditional mutual information was used in [19] to study algorithmic stability and, in turn, generalization of machine learning models. We next recall the setup from [19], outline their main stability results, and demonstrate that an algorithm satisfying ϵ-MI DP is stable in that sense. Consider a (possibly randomized) learning algorithm $A : \mathcal{X}^{n\times k} \to \mathcal{Y}$ that takes as input n samples, each with k features, and outputs an hypothesis from the class $\mathcal{Y}$ (the precise task or loss function are inconsequential here). Let $Q \in \mathcal{P}(\mathcal{X}^k)$ be the data distribution and draw a $2n$-sized dataset $X \sim Q^{\otimes 2n}$. Let $B = (B_1, \ldots, B_n) \sim \mathrm{Ber}(0.5)^{\otimes n}$ be a n-lengthed string of i.i.d. random bits independent of everything else. Using B, we define an n-sized database X_B, which will be fed into the learning algorithm as follows. Set $X_B(i,\cdot) = X\big(iB_i + (n+i)(1-B_i)\big)$, for $i = 1, \ldots, n$, i.e., the ith entry of X_B is $X(i,\cdot)$ of $B_i = 1$ and $X(n+i,\cdot)$ otherwise. Note that by symmetry, the distribution of this database is $X_B \sim Q^{\otimes n}$.

The following conditional mutual information measure, abbreviated CMI, quantifies how informative the output of an algorithm is about the selected samples (which are determined by the string B), given the entire $2n$-sized original database.

Definition 7 (CMI [19])**.** *Under the setup above, the CMI of the algorithm A w.r.t. the data distribution Q is given by*

$$\mathsf{CMI}_Q(A) := \mathsf{I}\big(A(X_B); B|X\big),$$

while its distribution-free CMI is

$$\mathsf{CMI}(A) := \sup_{x\in\mathcal{X}^{2n\times k}} \mathsf{I}\big(A(x_B); B\big).$$

Several generalization bounds based on the above definition were proven in [19], which roughly behave as $\big(\mathsf{CMI}_Q(A)/n\big)^{1/2}$ (or the distribution-free analogue); cf., e.g., Theorems 1.2 and 1.3 therein. In addition, [19] showed that DP algorithms have bounded distribution-free CMI. Specifically, an algorithm A that satisfies $\sqrt{2\epsilon}$-DP has $\mathsf{CMI}(A) \leq \epsilon n$. The following result shows that ϵ-MI DP also entails a similar conclusion.

Proposition 6 (CMI Bound via ϵ-MI DP)**.** *If the algorithm $A : \mathcal{X}^{n\times k} \to \mathcal{Y}$ satisfies ϵ-MI DP, then we have $\mathsf{CMI}(A) \leq \epsilon n$.*

This result follows because ϵ-MI DP algorithms also satisfy ϵ-mutual information stability[7] [18], which by [19, Theorem 4.7] further implies the CMI bound above. Consequently, learning algorithms that satisfy ϵ-MI DP follow the generalization bounds from [19].

Remark 20 (Domain Knowledge)**.** Algorithmic stability on restricted classes of distributions may be of interest when information about the data generating process is available. The above ideas and results readily extended to this case by replacing the set of all possible distributions by a subset $\Theta \subset \mathcal{P}(\mathcal{X}^{n\times k})$.

C. Utility of PP and MI PP Mechanisms

Generally, it may be hard to assess the utility of a given PP mechanism. However, the MI PP framework can be used as a proxy to understand the privacy-utility tradeoff in the mechanism design phase. We showcase another scenario here where we can borrow the results connected to information theory in making progress in privacy research via MI PP. In this subsection, we focus on the setup where $\Theta = \{P_X\}, \mathcal{G} = \{g\}$ and $\mathcal{W} = \emptyset$, namely, when the database is drawn from a given distribution and we want to hide one specific function thereof. Let $f(X)$ be the query and $M(X)$ the output of the private mechanism. We assume that the ranges of f and g are finite and cast $\mathsf{I}(f(X); M(X))$ as the utility metric, i.e., how much information the mechanism's output has of the query.

Consider the maximal utility of an ϵ-MI PP mechanism $M(X)$ in two situations: when $g(X)$ is or is not available at the mechanism design step. Namely, we define

$$\mathcal{U}_1^\epsilon(X) := \sup_{\substack{P_{M(X)|f(X),g(X)},\\ \mathsf{I}(M(X);g(X))\leq\epsilon}} \mathsf{I}\big(f(X); M(X)\big)$$

$$\mathcal{U}_2^\epsilon(X) := \sup_{\substack{P_{M(X)|f(X)},\\ g(X)\to f(X)\to M(X),\\ \mathsf{I}(M(X);g(X))\leq\epsilon}} \mathsf{I}\big(f(X); M(X)\big)$$

The following proposition, which relies on [55, Lemma 8], gives an upper bound on the maximal utility.

Proposition 7 (Lemma 8 of [55])**.** *For any $0 \leq \epsilon < \mathsf{I}(g(X); f(X))$, we have*

$$\mathcal{U}_2^\epsilon(X) \leq \mathcal{U}_1^\epsilon(X) \leq \mathsf{H}\big(f(X)|g(X)\big) + \epsilon.$$

From Proposition 7, we observe that the maximum utility which could be expected from any ϵ-MI PP mechanism (regardless of the accessibility to $g(X)$) is at most $\mathsf{H}\big(f(X)|g(X)\big) + \epsilon$. Since ϵ-PP implies ϵ-MI PP, all PP

[7]Given a data set $X = (X_1, \ldots, X_n)$ that comprises n i.i.d. samples from $Q \in \mathcal{P}(\mathcal{X}^k)$, an algorithm A is said to be ϵ-mutual information stable if $\sup_{Q\in\mathcal{P}(\mathcal{X}^k)} \frac{1}{n}\sum_{i=1}^n \mathsf{I}\big(A(X); X(i,\cdot)\big|\big(X(j,\cdot)\big)_{j\neq i}\big) \leq \epsilon$..

mechanisms are included in the optimization and we obtain a maximal utility bound for them as well. In particular, the bound shows that the increase in utility is at most linear in ϵ.

Through the relation to MI PP, we can also show the existence of MI PP mechanisms that achieve a certain utility lower bound.

Proposition 8 (Theorem 2 of [55]). *For any* $0 \leq \epsilon < \mathsf{I}(g(X); f(X))$, *we have*

$$\mathcal{U}_1^\epsilon(X) \geq L_1^\epsilon \vee L_2^\epsilon,$$

where $L_1^\epsilon = \mathsf{H}\big(f(X)|g(X)\big) - \mathsf{H}\big(g(X)|f(X)\big) + \epsilon$ *and* $L_2^\epsilon = \mathsf{H}\big(f(X)|g(X)\big) - \alpha\mathsf{H}\big(g(X)|f(X)\big) + \epsilon - (1-\alpha)\big(\log(\mathsf{I}(g(X); f(X)) + 1) + 4\big)$ *with* $\alpha = \epsilon/\mathsf{H}(g(X))$.

Proposition 8 implies that there exists an ϵ-MI PP mechanism with utility of at least $L_1^\epsilon \vee L_2^\epsilon$. Notice that when $\mathsf{H}\big(g(X)|f(X)\big) = 0$ (i.e., $g(X)$ is deterministic given $f(X)$), this lower bound is tight. Under this setting, combining Propositions 7 and 8, one may deduce that there exists an ϵ-MI PP mechanism with utility $\mathcal{U}_1^\epsilon(X) = \mathsf{H}\big(f(X)|g(X)\big) + \epsilon$. We stress, however, that this is merely an existence claim that does not reveal how to design a mechanism that achieves this maximum utility.

VIII. Proofs

A. Proof of Theorem 1

For the first implication, note that ϵ-PP implies that

$$\sup_{\mathcal{A}} \log\left(\frac{\mathbb{P}\big(M(X) \in \mathcal{A}|\mathcal{R}\big)}{\mathbb{P}\big(M(X) \in \mathcal{A}|\mathcal{T}\big)}\right) \leq \epsilon, \quad \forall\ (\mathcal{R}, \mathcal{T}) \in \mathcal{Q}.$$

The left-hand side above is the infinite order Rényi divergence. By monotonicity of Rényi divergences w.r.t. their order [56], we have $\mathsf{D}_{\mathsf{KL}}\left(P_{M(X)|\mathcal{R}} \| P_{M(X)|\mathcal{T}}\right) \leq \epsilon$. Then,

$$\begin{aligned}
&\mathsf{I}\big(g(X); M(X)|w(X)\big) \\
&= \mathbb{E}\left[\mathsf{D}_{\mathsf{KL}}\left(P_{M(X)|g(X),w(X)} \middle\| P_{M(X)|w(X)}\right)\right] \\
&\leq \mathbb{E}\left[\mathsf{D}_{\mathsf{KL}}\left(P_{M(X)|g(X),w(X)} \middle\| P_{M(X)|g(X)',w(X)}\right)\right] \qquad (9)
\end{aligned}$$

where the inequality uses convexity of KL divergence, with $g(X)'$ as an i.i.d. copy of $g(X)$. Recalling that under the structured PP framework secret pairs are $\big(A_{g,w}(a,c), A_{g,w}(b,c)\big)$, with $A_{g,w}(a,c) = \big\{g(X) = a,\ w(X) = c\big\}$, ϵ-MI PP follows by the KL divergence bound.

Assuming $\Theta = \mathcal{P}(\mathcal{X}^{n\times k})$, the second implication follows by the minimax redundancy capacity theorem [57], which gives

$$\begin{aligned}
&\sup_{\Theta} \mathsf{I}\big(g(X); M(X)|w(X) = c\big) \leq \epsilon \implies \\
&\inf_{Q} \max_{a} \mathsf{D}_{\mathsf{KL}}\left(P_{M(X)|g(X)=a,w(X)=c} \middle\| Q\right) \leq \epsilon. \qquad (10)
\end{aligned}$$

Let $Q^\star$ achieve the infimum on the RHS. Since M is ϵ-MI PP by assumption, we have $\mathsf{D}_{\mathsf{KL}}\left(P_{M(X)|g(X)=a,w(X)=c} \middle\| Q^\star\right) \leq \epsilon$, for all $a \in \mathrm{Im}(g)$. Applying Pinsker's inequality together with the triangle inequality, we obtain

$$\left\|P_{M(X)|g(X)=a,w(X)=c} - P_{M(X)|g(X)=b,w(X)=c}\right\|_{\mathsf{TV}} \leq \sqrt{2\epsilon},$$

which implies that M is $(0, \sqrt{2\epsilon})$-PP and hence $(\epsilon', \sqrt{2\epsilon})$-PP (recall Definition 3 for $\epsilon' > 0$.

For the third implication, in both part we rely on an adaptation of Property 3 from [24] from DP (as considered therein) to PP. Specifically, that (ϵ, δ)-PP implies $(0, \delta')$-PP, with $\delta' = 1 - 2(1-\delta)/(e^\epsilon + 1)$. Having that, for Part (1), we follow the argument from the proof of [24, Lemma 3] to show that if $\big|\mathrm{spt}\big(M(X)\big)\big| < \infty$ or $\max_{g\in\mathcal{G}} |\mathrm{Im}(g)| < \infty$, then $(0,\delta)$-PP implies $\epsilon^\star$-MI PP with $\epsilon^\star$ as stated in Theorem 1.

For Part (2), we recall that $(0,\delta')$-PP is equivalent to the TV bound $\forall (a,b) \in \mathrm{Im}(g),\ c \in \mathrm{Im}(w)$

$$\left\|P_{M(X)|g(X)=a,w(X)=c} - P_{M(X)|g(X)=b,w(X)=c}\right\|_{\mathsf{TV}} \leq \delta',$$

and then control the mutual information term of interest by the above TV norm. To obtain the first component of the $\epsilon^\star$ expression, we use the reverse Pinsker inequality from [58, Theorem 1] to translate the above TV bound into the following bound the KL divergence:

$$\begin{aligned}
&\mathsf{D}_{\mathsf{KL}}\left(P_{M(X)|g(X)=a,w(X)=c} \middle\| P_{M(X)|g(X)'=b,w(X)=c}\right) \\
&\leq \frac{\delta'}{2} \sup_{\substack{P_X\in\Theta,\\ (g,w)\in\mathcal{G}\times\mathcal{W}:\, g\sim w,\\ a,b\in\mathrm{Im}(g),\, c\in\mathrm{Im}(w)}} \left(\frac{\log\left(\alpha_{a,b,c}^{-1}\right)}{1-\alpha_{a,b,c}} - \beta_{a,b,c}\right) =: \epsilon_1^\star.
\end{aligned}$$

Having that, we bound the mutual information as in (9) to obtain $\epsilon_1^\star$-MI PP. For the second component of the $\epsilon^\star$ expression, we use the following Lemma which follows directly from Corollary 12 of [59].

Lemma 1. *If joint density* $f_{M(X),g(X),w(X)}$ *exists, then we have*

$$\begin{aligned}
&\mathsf{I}\big(g(X); M(X)|w(X) = c\big) \\
&\leq \gamma(c\,; g, w, P_X) \\
&\quad \times \mathbb{E}\left[\|P_{M(X)|w(X)=c} - P_{M(X)|w(X)=c,g(X)}\|_{\mathsf{TV}}\right], \qquad (11)
\end{aligned}$$

where

$$\begin{aligned}
&\gamma(c\,; g, w, P_X) \\
&:= \sup_{a\in\mathrm{Im}(g)} \log\left(\frac{\sup_{z\in\mathrm{spt}(M(X))} f_{M(X),g(X),w(X)}(z,a,c)}{\inf_{z'\in\mathrm{spt}(M(X))} f_{M(X),g(X),w(X)}(z',a,c)}\right).
\end{aligned}$$

Lemma 1 together with the fact that (ϵ,δ)-PP $\implies$ $(0,\delta')$-PP implies MI PP with parameter

$$\epsilon_2^\star := \delta' \sup_{\substack{P_X\in\Theta,\\ (g,w)\in\mathcal{G}\times\mathcal{W}:\, g\sim w,\\ c\in\mathrm{Im}(w)}} \gamma(c\,; g, w, P_X).$$

Taking $\epsilon^\star = \epsilon_1^\star \wedge \epsilon_2^\star$ yields the result.

B. Proof of Theorem 2

For (1), let I be a k-ary categorical random variable with the probabilities $p_1, \ldots, p_k$. Then

$$\begin{aligned}
\mathsf{I}\big(g(X); M(X)\big|w(X)\big) &\leq \mathsf{I}\big(g(X); M(X), I\big|w(X)\big) \\
&\overset{(a)}{=} \mathsf{I}\big(g(X); M(X)\big|w(X), I\big) \\
&\overset{(b)}{=} \sum_{i=1}^{k} p_i\ \mathsf{I}\big(g(X); M_i(X)\big|w(X)\big)
\end{aligned}$$

where (a) and (b) follow from the independence of I and $\big(X, M_1(X), \ldots, M_k(X)\big)$. The claim now follows since $M_1, \ldots, M_k$ satisfy ϵ-MI PP.

Part (2) directly follows from the chain rule of mutual information

$$\mathsf{I}\big(g(X); M^k \big| w(X)\big) = \sum_{i=1}^{k} \mathsf{I}\big(g(X); M_i(X) \big| w(X), M_1, .., M_{i-1}\big),$$

while Part (3) is a consequence of the data processing inequality

$$\mathsf{I}\big(g(X); A(M(X)) \big| w(X)\big) \leq \mathsf{I}\big(g(X); M(X) \big| w(X)\big).$$

C. Proof of Proposition 1

First use induction to prove that

$$\mathsf{I}\big(g(X); M^k(X) | w(X)\big) \leq \sum_{i=1}^{m} \epsilon_i + \eta', \tag{12}$$

where

$$\eta' = \sum_{i=2}^{k} \mathsf{I}\big(M_i(X); M_1(X), \ldots, M_{i-1}(X) | w(X), g(X)\big).$$

Given this inequality, supremizing over $P_X \in \Theta$ and $\mathcal{G} \ni g \sim w \in \mathcal{W}$ yields the result of Theorem 1.

For $k = 2$, consider

$$\begin{aligned} &\mathsf{I}\big(M_1(X), M_2(X); g(X) | w(X)\big) \\ &= \mathsf{I}\big(M_1(X); g(X) | w(X)\big) && (13) \\ &\quad + \mathsf{I}\big(M_2(X); g(X) | w(X), M_1(X)\big) && (14) \\ &\leq \epsilon_1 + \mathsf{I}\big(M_2(X); g(X) | w(X), M_1(X)\big), && (15) \end{aligned}$$

where the last step uses the fact that M_1 is ϵ_1-MI PP. For the second above, we have

$$\begin{aligned} &\mathsf{I}\big(M_2(X); g(X) \big| w(X), M_1(X)\big) \\ &= \mathsf{h}\big(M_2(X) | w(X), M_1(X)\big) \\ &\quad - \mathsf{h}\big(M_2(X) | w(X), M_1(X), g(X)\big) \\ &\leq \mathsf{h}\big(M_2(X) \big| w(X)\big) - \mathsf{h}\big(M_2(X) \big| w(X), M_1(X), g(X)\big) \\ &\quad \pm \mathsf{h}\big(M_2(X) \big| w(X), g(X)\big) \\ &= \mathsf{I}\big(M_2(X); g(X) \big| w(X)\big) \\ &\quad + \mathsf{I}\big(M_1(X); M_2(X) \big| w(X), g(X)\big) \\ &\leq \epsilon_2 + \mathsf{I}\big(M_1(X); M_2(X) \big| w(X), g(X)\big), \end{aligned}$$

where the last step is since M_2 is ϵ_2-MI PP. Collecting the terms, proves the claim for $k = 2$.

Assume that the result holds for $k = m$, i.e.,

$$\mathsf{I}\big(g(X); M^m(X) | w(X)\big) \leq \sum_{i=1}^{m} \epsilon_i + \eta'' \tag{16}$$

with

$$\eta'' = \sum_{i=2}^{m} \mathsf{I}\big(M_i(X); M_1(X), \ldots, M_{i-1}(X) | w(X), g(X)\big),$$

and consider the following for $k = m + 1$:

$$\begin{aligned} &\mathsf{I}\big(g(X); M^{m+1}(X) | w(X)\big) \\ &= \mathsf{I}\big(g(X); M^m(X) | w(X)\big) \\ &\quad + \mathsf{I}\big(M_{m+1}(X); g(X) | M^m(X), w(X)\big) \\ &\overset{(a)}{\leq} \sum_{i=1}^{m} \epsilon_i + \eta'' + \mathsf{I}\big(M_{m+1}(X); g(X) | M^m(X), w(X)\big) \\ &\overset{(b)}{\leq} \sum_{i=1}^{m} \epsilon_i + \eta'' + \epsilon_{m+1} \\ &\quad + \mathsf{I}\big(M_{m+1}(X); g(X) | M^m(X), w(X)\big), \end{aligned}$$

where (a) uses the induction assumption, while (b) follows since M_{m+1} is ϵ_{m+1}-MI PP. This establishes (12) and the proof is concluded by supermizing the RHS over $P_X \in \Theta$ and $\mathcal{G} \ni g \sim w \in \mathcal{W}$.

D. Proof of Proposition 1

Part (i): We prove this part by induction. Fix $P_X \in \Theta$ and $\mathcal{G} \ni g \sim w \in \mathcal{W}$. For $k = 2$, we use the chain rule together with the fact that M_1 is ϵ_1-MI PP to first obtain

$$\begin{aligned} &\mathsf{I}\big(g(X); M_1(X), M_2(X) \big| w(X)\big) \\ &\qquad \leq \epsilon_1 + \mathsf{I}\big(g(X); M_2(X) \big| w(X), M_1(X)\big). \end{aligned} \tag{17}$$

Then, notice that under the assumption that $\Theta \subseteq \Theta_{\mathsf{UC}}$, for any $P_X \in \Theta$, we have

$$\begin{aligned} &\mathsf{I}\big(g(X); M_2(X) \big| w(X), M_1(X)\big) \\ &\leq \mathsf{h}\big(M_2(X) \big| w(X)\big) - \mathsf{h}\big(M_2(X) \big| w(X), g(X), M_1(X)\big) \\ &= \mathsf{I}\big(g(X); M_2(X) \big| w(X)\big) \\ &\leq \epsilon_2, \end{aligned}$$

where the penultimate equality follows since $M_2(X) \leftrightarrow \big(g(X), w(X)\big) \leftrightarrow M_1(X)$ forms a Markov chain whenever P_X is a UC distribution, while the last step is due to M_2 satisfying ϵ_2-MI PP. Combining the above, the statement for $k = 2$ follows.

Assume that for $k = m$, we have

$$\mathsf{I}\big(g(X); M^m(X) \big| w(X)\big) \leq \sum_{i=1}^{m} \epsilon_i.$$

For $k = m + 1$, consider

$$\begin{aligned} &\mathsf{I}\big(g(X); M^{m+1}(X) \big| w(X)\big) \\ &= \mathsf{I}\big(g(X); M^m(X), M_{m+1} \big| w(X)\big) \\ &\leq \sum_{i=1}^{m} \epsilon_i + \epsilon_{m+1}, \end{aligned}$$

where the inequality follows from the $k = 2$ case applied to the mechanisms (M^m, M_{m+1}). By induction, we deduce that

$$\mathsf{I}\big(g(X); M^k(X) | w(X)\big) \leq \sum_{i=1}^{k} \epsilon_i,$$

and the claim follows by supremizing the LHS over $P_X \in \Theta$ and $\mathcal{G} \ni g \sim w \in \mathcal{W}$.

Part (ii): Define the shorthand $f_{a,c} := P_{M_1(X),M_2(X)|g(X)=a,w(X)=c}$ and $f_{\tilde{a},c} := P_{M_1(X),M_2(X)|g(X)=\tilde{a},w(X)=c}$. We again use induction. Fix $P_X \in \Theta$ and $\mathcal{G} \ni g \sim w \in \mathcal{W}$, and for $k=2$ first note that

$$\begin{aligned}&\mathsf{I}\big(g(X); M_1(X), M_2(X)\big|w(X)=c\big)\\ &\le \int \mathsf{D}_{\mathsf{KL}}\big(f_{a,c}\big\|f_{\tilde{a},c}\big)\,\mathsf{d}P_{g(X)|w(X)}(a|c)\,\mathsf{d}P_{g(X)|w(X)}(\tilde{a}|c)\end{aligned}$$

which follows by convexity of the KL divergence. Observe that

$$\begin{aligned}&P_{M_1(X),M_2(X)|g(X),w(X)}(\cdot|a,c)\\ &\qquad= \int_{\mathcal{X}} P_{M_1(X),M_2(X)|X}(\cdot|x)\,\mathsf{d}P_{X|g(X),w(X)}(x|a,c)\end{aligned}$$

(which uses the conditional independence of $\big(M_1(X), M_2(X)\big)$ from $\big(g(X), w(X)\big)$ given X itself) and leverage convexity once more to bound

$$\begin{aligned}&\mathsf{D}_{\mathsf{KL}}\big(f_{a,c}\big\|f_{\tilde{a},c}\big)\\ &\le \int \mathsf{D}_{\mathsf{KL}}\big(P_{M_1(X),M_2(X)|X=x}\big\|P_{M_1(X),M_2(X)|X=\tilde{x}}\big)\\ &\qquad\qquad \mathsf{d}P_{X|g(X),w(X)}(x|a,c)\,\mathsf{d}P_{X|g(X),w(X)}(\tilde{x}|\tilde{a},c)\\ &\le \int \Big[\mathsf{D}_{\mathsf{KL}}\big(P_{M_1(x)}\big\|P_{M_1(\tilde{x})}\big)+\mathsf{D}_{\mathsf{KL}}\big(P_{M_2(x)}\big\|P_{M_2(\tilde{x})}\big)\Big]\\ &\qquad\qquad \mathsf{d}P_{X|g(X),w(X)}(x|a,c)\,\mathsf{d}P_{X|g(X),w(X)}(\tilde{x}|\tilde{a},c)\end{aligned}$$

where the last step is due to the conditional independence of $M_1(X)$ and $M_2(X)$ given X and tensorization of KL divergence. Recall that by definition of UC distributions, for any $(x,\tilde{x}) \in \mathrm{spt}(P_{X|g(X)=a,w(X)=c}) \times \mathrm{spt}(P_{X|g(X)=\tilde{a},w(X)=c})$ we have $\lambda\delta_x + (1-\lambda)\delta_{\tilde{x}} \in \Theta_{\mathsf{UC}}$ for $\lambda \in (0,1)$. Since Part (ii) assumes $\Theta_{\mathsf{UC}} \subseteq \Theta$, the fact that M_i, for $i=1,2$, is ϵ_i-PP in the framework with distribution class Θ, we obtain $\mathsf{D}_{\mathsf{KL}}\big(P_{M_i(x)}\big\|P_{M_i(\tilde{x})}\big) \le \epsilon_i$. Inserting this into the bounds above yields

$$\mathsf{I}\big(g(X); M_1(X), M_2(X)\big|w(X)\big) \le \epsilon_1 + \epsilon_2,$$

which is the desired claim for $k=2$.

The induction assumption for $k=m$ reads as

$$\mathsf{I}\big(g(X); M^n(X)\big|w(X)\big) \le \sum_{i=1}^{n} \epsilon_i,$$

and for $k=m+1$, we have

$$\begin{aligned}&\mathsf{I}\big(g(X); M^{m+1}(X)\big|w(X)\big)\\ &= \mathsf{I}\big(g(X); M^m(X), M_{m+1}(X)\big|w(X)\big)\\ &\le \sum_{i=1}^{m+1} \epsilon_i,\end{aligned}$$

where the inequality follows by the $k=2$ case as before. We conclude again by induction and taking the appropriate supremum.

E. Proof of Theorem 3

Let $Z_{\mathsf{L}} = (Z_1, \ldots, Z_d)$, with $Z_j \sim \mathrm{Lap}(0,b)$ i.i.d. We have

$$\begin{aligned}&\mathsf{h}\big(f(X)+Z_{\mathsf{L}}\big|w(X)\big)\\ &\overset{(a)}{\le} \sum_{j=1}^{d} \int \mathsf{h}\big(f_j(X)+Z_j-m_j(\cdot)\big|w(X)=\cdot\big)dP_{w(X)}\\ &\overset{(b)}{\le} \sum_{j=1}^{d} \int \log\Big(2e\mathbb{E}\big[|f_j(X)-m_j(\cdot)+Z_j|\big|w(X)=\cdot\big]\Big)dP_{w(X)}\\ &\overset{(c)}{\le} d\log\left(2e\left(\sum_{j=1}^{d}\frac{1}{d}\,\mathbb{E}\left[\sqrt{\mathrm{Var}\big(f_j(X)\big|w(X)\big)}\right]+b\right)\right),\end{aligned} \tag{18}$$

where (a) uses the chain rule, the fact that conditioning cannot increase differential entropy, and its translation invariance with $m_j(c) := \mathbb{E}[f_j(X)|w(X)=c]$; (b) is because the Laplace distribution maximizes differential entropy subject to an expected absolute deviation constraint; while (c) follows from Jensen's inequality, along with $\mathbb{E}[|Z_j|]=b$ and $\mathbb{E}[|X|]^2 \le \mathbb{E}[X^2]$.

Combining (18) with $\mathsf{h}\big(f(X)+Z_{\mathsf{L}}\big|g(X),w(X)\big) \ge \mathsf{h}(Z_{\mathsf{L}}) = d\log(2be)$ yields:

$$\begin{aligned}&\mathsf{I}\big(g(X); M_{\mathsf{L}}(X)\big|w(X)\big)\\ &= \mathsf{h}\big(f(X)+Z_{\mathsf{L}}\big|w(X)\big) - \mathsf{h}\big(f(X)+Z_{\mathsf{L}}\big|g(X),w(X)\big)\\ &\le d\log\left(2e\left(\sum_{j=1}^{d}\frac{1}{d}\,\mathbb{E}\left[\sqrt{\mathrm{Var}\big(f_j(X)\big|w(X)\big)}\right]+b\right)\right)\\ &\qquad - d\log(2be).\end{aligned} \tag{19}$$

To conclude, further upper bound (19) by ϵ and solve for b.

F. Proof of Corollary 2

Using the fact that $\mathrm{Var}(Z) = \frac{1}{2}\mathbb{E}\big[(Z-Z')^2\big]$ for Z' an i.i.d. copy of Z, we have

$$\begin{aligned}&\sum_{j=1}^{d} \mathrm{Var}\Big(f_j(X)\Big|\big(X(k,\cdot)\big)_{k\neq i}=\big(x(k,\cdot)\big)_{k\neq i}\Big)\\ &= \frac{1}{2}\sum_{k=1}^{d}\mathbb{E}\Big[\big(f_j(X)-\tilde{f}_j(X)\big)^2\Big|\big(X(k,\cdot)\big)_{k\neq i}=\big(x(k,\cdot)\big)_{k\neq i}\Big]\\ &= \frac{1}{2}\,\mathbb{E}\left[\sum_{j=1}^{d}\big(f_j(X)-\tilde{f}_j(X)\big)^2\Big|\big(X(k,\cdot)\big)_{k\neq i}=\big(x(k,\cdot)\big)_{k\neq i}\right]\\ &\le \frac{\Delta_2^2(f)}{2},\end{aligned} \tag{20}$$

where the last step comes from the definition of ℓ^2- sensitivity. Insert (20) into the entropy bound from (18) and use the fact that $\Delta_2(f) \le \Delta_1(f)$ to obtain

$$\begin{aligned}&\mathsf{I}\big(g(X); M_{\mathsf{L}}(X)\big|w(X)\big)\\ &\qquad\le d\log\left(2e\left(\frac{\Delta_1(f)}{\sqrt{2}}+b\right)\right) - d\log(2be).\end{aligned}$$

To conclude, proceed as in the proof of Theorem 3 to find the parameter b.

G. Proof of Theorem 4

We follow a similar argument to that in the proof of Theorem 3. Denote the conditional covariance matrix of $f(X)$ given $\{w(X)=c\}$ by $\Sigma_{f|w=c}$ and consider

$$\begin{aligned}
&\mathsf{h}\big(f(X)+Z_{\mathsf{G}}\big|w(X)\big)\\
&\overset{(a)}{\leq}\int\frac{1}{2}\log\left((2\pi e)^d\det(\Sigma_{f|w=\cdot}+\sigma^2\mathrm{I}_d)\right)dP_{w(X)}\\
&\overset{(b)}{\leq}\int\frac{1}{2}\log\left((2\pi e)^d\prod_{j=1}^{d}\left(a_j(\cdot)+\sigma^2\right)\right)dP_{w(X)}\\
&\overset{(c)}{\leq}\frac{d}{2}\int\log\left(2\pi e\left(\frac{1}{d}\sum_{j=1}^{d}\mathrm{Var}\big(f_j(X)\big|w(X)=\cdot\big)+\sigma^2\right)\right)dP_{w(X)}\\
&\overset{(d)}{\leq}\frac{d}{2}\log\left(2\pi e\left(\frac{1}{d}\sum_{j=1}^{d}\mathbb{E}\Big[\mathrm{Var}\big(f_j(X)\big|w(X)\big)\Big]+\sigma^2\right)\right),
\end{aligned}\tag{21}$$

where (a) follows from the Gaussian distribution maximizing differential entropy subject to a variance constant, with $|K|$ denoting the determinant of K; (b) denotes $a_j(c):=\mathrm{Var}\big(f_j(X)|w(X)=c\big)$ and uses $|K|\leq\prod_{j=1}^{d}K(j,j)$, which applies to any positive semi-definite matrix; and (c)-(d) from concavity of $x\mapsto\log x$ and Jensen's inequality.

Combining (21) with $\mathsf{h}\big(f(X)+Z_{\mathsf{G}}\big|g(X),w(X)\big)\geq\mathsf{h}(Z_{\mathsf{G}})=\frac{d}{2}\log(2\pi e\sigma^2)$ yields

$$\begin{aligned}
&\mathsf{I}\big(g(X);M_{\mathsf{G}}(X)\big|w(X)\big)\\
&\leq\frac{d}{2}\log\left(2\pi e\left(\frac{1}{d}\sum_{j=1}^{d}\mathbb{E}\Big[\mathrm{Var}\big(f_j(X)\big|w(X)\big)\Big]+\sigma^2\right)\right)\\
&\quad-\frac{d}{2}\log(2\pi e\sigma^2).
\end{aligned}$$

Upper bounding the RHS above by ϵ and solving for σ^2 concludes the proof.

H. Proof of Corollary 3

We first insert the upper bound from (20) into (21) to obtain an ℓ^2-sensitivity bound on $\mathsf{h}\big(f(X)+Z_{\mathsf{G}}\big|w(X)\big)$. Combining this with $\mathsf{h}\big(f(X)+Z_{\mathsf{G}}\big|X\big)\geq\mathsf{h}(Z_{\mathsf{G}})=\frac{d}{2}\log(2\pi e\sigma^2)$ yields

$$\begin{aligned}
&\mathsf{I}\big(X(i,\cdot);M_{\mathsf{G}}(X)\big|\big(X(k,\cdot)\big)_{k\neq i}\big)\\
&\qquad\leq\frac{d}{2}\log\left(2\pi e\left(\frac{\Delta_2^2(f)}{2}+\sigma^2\right)\right)-\frac{d}{2}\log(2\pi e\sigma^2).
\end{aligned}$$

Upper bounding the RHS above by ϵ and solving for σ^2 produces the result.

For the case when $\mathcal{X}$ is compact and $f:\mathcal{X}^{n\times k}\to\mathbb{R}$ is of continuous, we apply the Popoviciu inequality for variance to obtain

$$\begin{aligned}
&\mathrm{Var}\left(f_j(X)\Big|\big(X(k,\cdot)\big)_{k\neq i}=\big(x(k,\cdot)\big)_{k\neq i}\right)\\
&\leq\frac{1}{4}\left(\sup_{x(i,\cdot)}f_j\Big(x(i,\cdot),\big(x(k,\cdot)\big)_{k\neq i}\Big)-\inf_{x(i,\cdot)}f_j\Big(x(i,\cdot),\big(x(k,\cdot)\big)_{k\neq i}\Big)\right)^2\\
&\leq\frac{\Delta_2^2(f)}{4},
\end{aligned}$$

Applying this relation and proceeding with the same argument as above leads to the desired result.

I. Proof of Theorem 5

Invoking Theorem 4 for the query function $g(X)=\mathrm{A}^{\intercal}f(X)$, where $\mathrm{A}=[\phi_1,\ldots,\phi_\ell]$, we obtain

$$\sigma^2\geq\sup_{P_X\in\Theta,\,w\in\mathcal{W}^\star}\frac{\sum_{j=1}^{\ell}\mathbb{E}\left[\mathrm{Var}\big(\phi_j^{\intercal}f(X)\big|w(X)\big)\right]}{\ell(e^{\frac{2\epsilon}{\ell}}-1)}.\tag{22}$$

Recall that $\Sigma_{f|w}$ denotes the conditional covariance matrix of $f(X)$ given $w(X)$ by, and bound $\forall j=1\ldots,\ell$

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Var}\big(\phi_j^{\intercal}f(X)\big|w(X)\big)\right]\\
&=\mathbb{E}\big[\phi_j^{\intercal}\Sigma_{f|w}\phi_j\big]\\
&\leq\mathbb{E}\big[\|\Sigma_{f|w}\|_{\mathrm{op}}\|\phi_j\|^2\big].
\end{aligned}$$

Combining the above with (22) shows the sufficiency of the variance parameter in Part (1) of the theorem.

For a random projection matrix $\mathrm{A}=[\Phi_1,\ldots,\Phi_\ell]$, since Φ_j is centered and independent of X, we have $\mathbb{E}\big[\phi_j^{\intercal}f(X)\big|w(X)\big]=0$. Recalling the notation $\mu_{f|w}:=\mathbb{E}\big[f(X)|w(X)\big]$, we consequently have

$$\begin{aligned}
&\mathbb{E}\big[\mathrm{Var}\big(\Phi_j^{\intercal}f(X)\big|w(X)\big)\big]\\
&=\mathbb{E}\Big[\Phi_j^{\intercal}\mathbb{E}\big[f(X)f(X)^{\intercal}\big|w(X)\big]\Phi_j\Big]\\
&=\mathbb{E}\Big[\Phi_j^{\intercal}\big(\Sigma_{f|w}+\mu_{f|w}\mu_{f|w}^{\intercal}\big)\Phi_j\Big]\\
&\leq\mathbb{E}\Big[\|\Sigma_{f|w}\|_{\mathsf{op}}+\|\mu_{f|w}\|^2\Big]
\end{aligned}$$

where the last step uses $\mathbb{E}\big[\|\Phi_j\|^2\big]=1$ and the fact that $\|aa^{\intercal}\|_{\mathsf{op}}\leq\|a\|^2$, for any vector $a\in\mathbb{R}^d$. Given the variance bound, we proceed as in the proof of the deterministic projection case to obtain the result.

J. Proof of Theorem 6

First, rewrite (21) in terms of A given in Theorem 6, to obtain

$$\mathsf{h}\big(f(X)+Z_{\mathsf{G}}\big|w(X)\big)\leq\frac{d}{2}\log\left(2\pi e\left(\frac{A}{d}+\sigma^2\right)\right).$$

From entropy power inequality, we have

$$e^{\frac{2}{d}\mathsf{h}(f(X)+Z_{\mathsf{G}}|g(X),w(X))}\geq e^{\frac{2}{d}\mathsf{h}(f(X)|g(X),w(X))}+e^{\frac{2}{d}\mathsf{h}(Z_{\mathsf{G}})}.$$

Noting that $\mathsf{h}(Z_{\mathsf{G}})=0.5d\log(2\pi e\sigma^2)$ and by the choice of B in the statement of the theorem, we arrive at

$$\mathsf{h}\big(f(X)+Z_{\mathsf{G}}\big|g(X),w(X)\big)\geq0.5\;d\log\left(2\pi e(B+\sigma^2)\right),$$

which combined with the above yields

$$\begin{aligned}
&\mathsf{I}\big(g(X);f(X)+Z_{\mathsf{G}}|w(X)\big)\\
&\qquad\leq\frac{d}{2}\log\left(2\pi e\left(\frac{A}{d}+\sigma^2\right)\right)-\frac{d}{2}\log\left(2\pi e(B+\sigma^2)\right).
\end{aligned}$$

Upper bounding the RHS by ϵ and solving for σ^2 completes the proof.

K. Proofs Related to Remark 14

We first show that $A \geq dB$ holds for any $P_X \in \mathcal{P}(\mathcal{X}^{n\times k})$ and functions f, $g \in \mathcal{G}$, and $w \in \mathcal{W}$ with $g \sim w$. Reusing the notation $\Sigma_{f|w}$ for the conditional covariance matrix of $f(X)$ given $w(X)$, we have

$$\begin{aligned}
&\mathsf{h}\big(f(X)|w(X),g(X)\big)\\
&\leq \mathsf{h}\big(f(X)|w(X)\big)\\
&\overset{(a)}{\leq} \frac{1}{2}\,\mathbb{E}\Big[\log\big((2\pi e)^d \det(\Sigma_{f|w})\big)\Big]\\
&\overset{(b)}{\leq} \frac{1}{2}\,\mathbb{E}\left[\log\left((2\pi e)^d \prod_{j=1}^{d} \mathrm{Var}\big(f_j(X)\big|w(X)\big)\right)\right]\\
&\leq \frac{d}{2}\,\mathbb{E}\left[\log\left(\frac{2\pi e}{d}\sum_{j=1}^{d} \mathrm{Var}\big(f_j(X)|w(X)\big)\right)\right]\\
&\leq \frac{d}{2}\,\log\left(\frac{2\pi e}{d}\sum_{j=1}^{d} \mathbb{E}\left[\mathrm{Var}\big(f_j(X)|w(X)\big)\right]\right)
\end{aligned}$$

where (a) is since the Gaussian distribution maximizing entropy under finite variance constraint; (b) uses $|K| \leq \prod_{j=1}^{d} K(j,j)$, which applies to any positive semidefinite matrix; and the last two steps follow from Jensen's inequality. Substituting the above bound in place of B in Theorem 6 yields the result.

Next, we show that when $d = 1$ and $\big(f(X), g(X), w(X)\big)$ are jointly Gaussian, we indeed have $A \leq d\,e^{2\epsilon/d}B$ under the said correlation coefficient bound. Under this Gaussian setting, we have $\forall c \in \mathrm{Im}(w)$

$$A = \mathbb{E}\big[\mathrm{Var}\big(f(X)|w(X)\big)\big] = \mathrm{Var}\big(f(X)|w(X)=c\big).$$

Similarly, joint Gaussianity of the involved variables implies $\forall (a,c) \in \mathrm{Im}(g)\times\mathrm{Im}(w)$

$$B = \mathrm{Var}\big(f(X)|w(X)=c, g(X)=a\big).$$

Inserting this into the inequality $A \leq d\,e^{2\epsilon/d}B$ with $d = 1$, while observing that

$$\begin{aligned}
&\mathrm{Var}\big(f(X)\big|w(X)=c, g(X)=a\big)\\
&=\Big(1-\big(\rho\,\big(f(X),g(X)\big|w(X)=c\big)\big)^2\Big)\,\mathrm{Var}\big(f(X)|w(X)=c\big)
\end{aligned}$$

completes the proof.

L. Proof of Proposition 5

We show that the estimate produced by Algorithm 1 satisfies ϵ-MI DP, and establish that given $n \geq n_0$ samples, the estimator also achieves accuracy α with high probability. Denote $c := \mathbb{E}\big[\|X-\mu\|^2\big] < \infty$ and notice that the noisy mean estimates $(\tilde{\mu}_1,\ldots,\tilde{\mu}_m)$ produced during the m rounds of Algorithm 1 are i.i.d. Further assume that $n = mk$ (otherwise, simple modifications of the subsequent argument using ceiling/floor operation may be needed). Furthermore denote $A_p := \{(p-1)k+1,\ldots,pk\}$ for $p = 1,\ldots,m$, and note that $\bigcup_{p=1}^{m} A_p = \{1,\ldots,n\}$. We first establish privacy of the mechanism and then analyze its accuracy.

Privacy analysis: For privacy, we first show that each $\tilde{\mu}_p$, $p = 1,\ldots,m$, is ϵ-MI DP. Having that, we argue that $M^m(X) = (\tilde{\mu}_1,\ldots\tilde{\mu}_m)$ is private in the same sense via composition, and finally deduce the privacy of the geometric median $\hat{\mu}_n$ via post-processing (Property (2) of Theorem 2). Consider $\tilde{\mu}_1$ and notice that it satisfies ϵ-MI DP by Corollary 3. Indeed, for each $i = 1,\ldots,k$ and $\ell = 1,\ldots,d$, we have

$$\mathrm{Var}\left(\tilde{\mu}_1(\ell)\Big|\big(X_j\big)_{j\in A_1\setminus\{i\}}\right) = \frac{1}{k^2}\mathrm{Var}\big(X_i(\ell)\big) \leq \frac{m^2 c}{n^2},$$

where the first equality is since $X_1,\ldots,X_k$ are i.i.d. and the second uses the 2nd moment bound. Therefore,

$$\frac{\sum_{\ell=1}^{d}\mathrm{Var}\left(\tilde{\mu}_1(\ell)\Big|\big(X_j\big)_{j\in A_1\setminus\{i\}}\right)}{d(e^{\frac{2\epsilon}{d}}-1)} \leq \frac{dm^2c}{2n^2\epsilon},$$

where we have used $e^x \geq 1+x$. By Corollary 3 we now see that the noise level stated in Algorithm 1 suffices to guarantee ϵ-MI DP of $\tilde{\mu}_1$. By symmetry, the same hold for all $\tilde{\mu}_p$, $p = 1,\ldots,m$.

Next, we show that $(\tilde{\mu}_1,\ldots\tilde{\mu}_m)$ is ϵ-MI DP w.r.t. the entire database $(X_1,\ldots,X_n)$. Fix $i = 1,\ldots,n$ and let $p(i) \in \{1,\ldots,m\}$ be such that $i \in A_{p(i)}$. Then consider

$$\begin{aligned}
&\mathsf{I}\big(X_i;\tilde{\mu}_1,\ldots,\tilde{\mu}_m\big|(X_j)_{j\neq i}\big)\\
&= \mathsf{I}\big(X_i;\tilde{\mu}_{p(i)}\big|\big(X_j\big)_{j\neq i}\big) + \mathsf{I}\big(X_i;(\tilde{\mu}_q)_{q\neq p(i)}\big|(X_j)_{j\neq i},\tilde{\mu}_{p(i)}\big)\\
&\overset{(a)}{=} \mathsf{I}\big(X_i;\tilde{\mu}_{p(i)}\big|(X_j)_{j\in A_{p(i)}\setminus\{i\}}\big)\\
&\overset{(b)}{\leq} \epsilon
\end{aligned}$$

where (a) follows since $(X_i,\tilde{\mu}_{p(i)}) \leftrightarrow (X_j)_{j\in A_{p(i)}\setminus\{i\}} \leftrightarrow (X_j)_{j\notin A_{p(i)}}$ and $X_i \leftrightarrow \big((X_j)_{j\neq i},\tilde{\mu}_{p(i)}\big) \leftrightarrow (\tilde{\mu}_q)_{q\neq p(i)}$ form Markov chains, while (b) is since $\tilde{\mu}_{p(i)}$ satisfies ϵ-MI DP. Maximizing the LHS above over $i = 1,\ldots,n$ yields

$$\sup_{i=1,\ldots,n}\ \mathsf{I}\big(X_i;\tilde{\mu}_1,\ldots,\tilde{\mu}_m\big|(X_j)_{j\neq i}\big) \leq \epsilon,$$

which shows that $(\tilde{\mu}_1,\ldots,\tilde{\mu}_m)$ is ϵ-MI DP. Recalling that post-processing preserves ϵ-MI DP (cf. Property (3) of Theorem 2), we conclude that the geometric median of $(\tilde{\mu}_1,\ldots,\tilde{\mu}_m)$ also satisfies ϵ-MI DP.

Accuracy analysis: We first derive a lower bound on k (the number of samples used to evaluate $\tilde{\mu}_p$ for each $p = 1,\ldots,m$) such that the mean estimate $\tilde{\mu}_p$ is closer to true mean μ with high probability. Then, we invoke [60, Theorem 3.1] on confidence boosting via geometric medians to argue that the geometric median of these m estimates satisfies the accuracy level of accuracy stated in the theorem.

The following lemma states the lower bound on k for a single iteration of Algorithm 1, i.e., for each $1,\ldots,m$.

Lemma 2. [Accuracy of a single iteration of Algorithm 1] Fix $p = 1,\ldots,m$ and suppose that

$$k \geq O\left(\frac{d}{\alpha'^2} + \frac{d}{\alpha'\sqrt{\epsilon}}\right).$$

Then the mean estimate $\tilde{\mu}_p$ produced by the pth iteration of Algorithm 1 satisfies $\mathbb{P}\big(\|\tilde{\mu}_p - \mu\| \leq \alpha'\big) \geq 0.8$.

Proof: We first bound the empirical mean estimation error. By Chebyshev's inequality and since the second moment is bounded by c, we have

$$\begin{aligned}&\mathbb{P}\left(\left\|\mu - \frac{1}{k}\sum\nolimits_{i\in A_p} X_i\right\| > \frac{\alpha'}{2}\right)\\ &\leq \frac{4\mathbb{E}\left[\left\|\mu - \frac{1}{k}\sum_{i\in A_p} X_i\right\|^2\right]}{\alpha'^2}\\ &\leq \frac{4dc}{k\alpha'^2},\end{aligned}$$

Setting $k \geq 4dc/(0.1\alpha'^2)$ guarantees that $\left\|\mu - \frac{1}{k}\sum_{i\in A_p} X_i\right\| \leq \alpha'/2$ at least with probability 0.9.

Next, consider the error due to noise injection for privacy with parameter $\sigma^2 = dc/(2k^2\epsilon)$. Using the tail bounds for d-dimensional Gaussian vector we have

$$\mathbb{P}\left(\|Z_p\| > \frac{\alpha'}{2}\right) \leq 2\exp\left(-\frac{\alpha'^2}{8\sigma^2 d}\right).$$

Choosing $k \geq \frac{2dc}{\alpha\sqrt{\epsilon}}\sqrt{\log(20)}$ guarantees that $\|Z_p\| \leq \frac{\alpha'}{2}$ at least with probability 0.9.

The choice of k as stated in the Lemma is sufficient to satisfy both bounds. Now consider

$$\begin{aligned}&\mathbb{P}\big(\|\tilde{\mu}_p - \mu\| \leq \alpha'\big)\\ &\geq \mathbb{P}\left(\left\|\mu - \frac{1}{k}\sum\nolimits_{i\in A_p} X_i\right\| \leq \frac{\alpha'}{2}, \|Z_p\| \leq \frac{\alpha'}{2}\right)\\ &\geq 1 - \mathbb{P}\left(\left\|\mu - \frac{1}{k}\sum\nolimits_{i\in A_p} X_i\right\| > \frac{\alpha'}{2}\right) - \mathbb{P}\left(\|Z_p\| > \frac{\alpha'}{2}\right)\end{aligned}$$

The result follows by recalling that each term on the RHS is less than 0.1. □

From Lemma 2, we have $\mathbb{P}\big(\|\tilde{\mu}_p - \mu\| \geq \alpha'\big) \leq 0.2$ for all $p = 1, \ldots, m$, so long that k satisfies the prescribed lower bound. Applying Theorem 3.1 of [60] on boosting the confidence via geometric medians, with our choice of $m = 200\log(1/\beta)$ yields[8]

$$\mathbb{P}\big(\|\hat{\mu}_n - \mu\| \geq 1.04\alpha'\big) \leq \beta.$$

Setting $\alpha = 1.04\alpha'$ and noticing that $n = km$ completes the proof of the accuracy guarantee.

IX. Concluding Remarks and Future Directions

This work established an information-theoretic characterization, termed ϵ-MI PP, of the structured PP framework, where private information is modeled by functions of the database. The characterization was leveraged to derive properties of ϵ-MI PP and obtain sufficient conditions for noise-injection (Laplace and Gaussian) mechanisms. Our results highlight the virtues of an information-theoretic perspective on PP, enabling more flexible composability theorems and variance-dependent noise parameter bounds that exploit the distributional assumptions in the PP framework. As applications of ϵ-MI PP we explored auditing privacy frameworks, statistical inference tasks, and privacy-utility tradeoffs.

Future research directions are abundant. First, we aim to better and strengthen the reverse implication in Theorem 1, i.e., derive relaxed and general conditions under which ϵ-MI PP implies PP in the classic sense. This is a non-trivial endeavour as mutual information is an average quantity, while classic privacy notions are typically worst-case. We also target a power (namely, Type II error) analysis of the auditing hypothesis test in Section VI. A possible direction from which to tackle this question is to derive a limit distribution theory under the alternative for the test statistic and use that to analyze the power of the test under local alternatives via LeCam's Third Lemma. Lastly, we are interested in further exploring to privacy-utility tradeoffs [55], [61], [62] via ϵ-MI PP and connect those to tradeoffs for standard PP mechanisms (preliminary results in this direction are found in Section VII-C). In particular, we aim to characterize the achievable privacy-utility region for PP mechanisms and design optimal mechanisms for different points in that region.

[8]Specifically, adapting to their notation, we invoke [60, Theorem 3.1] with $p = 0.2$, $\epsilon = \alpha'$, and $\alpha = 0.21$.

Acknowledgment

The authors would like to thank Rachel Cummings for her feedback and helpful suggestions regarding an earlier version of the ϵ-MI PP idea.

References

[1] T. Nuradha and Z. Goldfeld, "An information-theoretic characterization of pufferfish privacy," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2022, pp. 2005–2010.

[2] S. Haney, A. Machanavajjhala, J. M. Abowd, M. Graham, M. Kutzbach, and L. Vilhuber, "Utility cost of formal privacy for releasing national employer-employee statistics," in *Proc. ACM Int. Conf. Manage. Data*, May 2017, pp. 1339–1354.

[3] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Conf. Theory Cryptogr. (TCC)*, Mar. 2006, pp. 265–284.

[4] D. Kifer and A. Machanavajjhala, "Pufferfish: A framework for mathematical privacy definitions," *ACM Trans. Database Syst.*, vol. 39, no. 1, pp. 1–36, Jan. 2014.

[5] B. Pejó and D. Desfontaines, *Guide to Differential Privacy Modifications: A Taxonomy of Variants and Extensions*. Cham, Switzerland: Springer, 2022.

[6] S. Song, Y. Wang, and K. Chaudhuri, "Pufferfish privacy mechanisms for correlated data," in *Proc. ACM Int. Conf. Manage. Data*, May 2017, pp. 1291–1306.

[7] W. Zhang, O. Ohrimenko, and R. Cummings, "Attribute privacy: Framework and mechanisms," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Jun. 2022, pp. 757–766.

[8] D. Kifer and B.-R. Lin, "An axiomatic view of statistical privacy and utility," *J. Privacy Confidentiality*, vol. 4, no. 1, pp. 1–36, Jul. 2012.

[9] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, Aug. 2014.

[10] N. Holohan, D. Leith, and O. Mason, "Differential privacy in metric spaces: Numerical, categorical and functional data under the one roof," *Inf. Sci.*, vol. 305, pp. 256–268, Jun. 2015.

[11] B. Balle and Y.-X. Wang, "Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2018, pp. 394–403.

[12] J. Zhao et al., "Reviewing and improving the Gaussian mechanism for differential privacy," 2019, *arXiv:1911.12060*.

[13] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer, "Detecting violations of differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2018, pp. 475–489.

[14] M. Jagielski, J. Ullman, and A. Oprea, "Auditing differentially private machine learning: How private is private SGD?" in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 22205–22216.

[15] C. Domingo-Enrich and Y. Mroueh, "Auditing differential privacy in high dimensions with the kernel quantum Rényi divergence," 2022, *arXiv:2205.13941*.

[16] Z. Goldfeld and K. Greenewald, "Sliced mutual information: A scalable measure of statistical dependence," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, Sep. 2021, pp. 17567–17578.

[17] Z. Goldfeld, K. Greenewald, T. Nuradha, and G. Reeves, "k-sliced mutual information: A quantitative study of scalability with dimension," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 15982–15995.

[18] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu, "Information-theoretic analysis of stability and bias of learning algorithms," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Sep. 2016, pp. 26–30.

[19] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information," in *Proc. 33rd Int. Conf. Learn. Theory (PMLR)*, 2020, pp. 3437–3452.

[20] G. Barthe and B. Kopf, "Information-theoretic bounds for differentially private mechanisms," in *Proc. IEEE 24th Comput. Secur. Found. Symp.*, Jun. 2011, pp. 191–204.

[21] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proc. IEEE 54th Annu. Symp. Found. Comput. Sci.*, Oct. 2013, pp. 429–438.

[22] S. Asoodeh, F. Alajaji, and T. Linder, "On maximal correlation, mutual information and data privacy," in *Proc. IEEE 14th Can. Workshop Inf. Theory (CWIT)*, Jul. 2015, pp. 27–31.

[23] W. Wang, L. Ying, and J. Zhang, "On the relation between identifiability, differential privacy, and mutual-information privacy," *IEEE Trans. Inf. Theory*, vol. 62, no. 9, pp. 5018–5029, Sep. 2016.

[24] P. Cuff and L. Yu, "Differential privacy as a mutual information constraint," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 43–54.

[25] I. Issa and A. B. Wagner, "Operational definitions for some common information leakage metrics," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 769–773.

[26] I. Mironov, "Rényi differential privacy," in *Proc. IEEE 30th Comput. Secur. Found. Symp. (CSF)*, Aug. 2017, pp. 263–275.

[27] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2014, pp. 501–505.

[28] M. Showkatbakhsh, C. Karakus, and S. Diggavi, "Privacy-utility trade-off of linear regression under random projections and additive noise," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 186–190.

[29] S. Yagli, A. Dytso, and H. V. Poor, "Information-theoretic bounds on the generalization error and privacy leakage in federated learning," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, May 2020, pp. 1–5.

[30] L. P. Barnes, A. Dytso, and H. V. Poor, "Improved information theoretic generalization bounds for distributed and federated learning," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2022, pp. 1465–1470.

[31] Y. Sun, J. Shao, S. Li, Y. Mao, and J. Zhang, "Stochastic coded federated learning with convergence and privacy guarantees," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2022, pp. 2028–2033.

[32] S. Li, A. Khisti, and A. Mahajan, "Information-theoretic privacy for smart metering systems with a rechargeable battery," *IEEE Trans. Inf. Theory*, vol. 64, no. 5, pp. 3679–3695, May 2018.

[33] C. Guo, A. Sablayrolles, and M. Sanjabi, "Analyzing privacy leakage in machine learning via multiple hypothesis testing: A lesson from Fano," 2022, *arXiv:2210.13662*.

[34] W. Zhang, M. Li, R. Tandon, and H. Li, "Online location trace privacy: An information theoretic approach," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 1, pp. 235–250, Jan. 2019.

[35] Y.-X. Wang, J. Lei, and S. E. Fienberg, "On-average KL-privacy and its equivalence to generalization for max-entropy mechanisms," in *Proc. Int. Conf. Privacy Stat. Databases UNESCO Chair Data Privacy*. Dubrovnik, Croatia: Springer, 2016, pp. 121–134.

[36] X. Wang, H. Ishii, J. He, and P. Cheng, "Dynamic privacy-aware collaborative schemes for average computation: A multi-time reporting case," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 3843–3858, 2021.

[37] X. Wang, J. He, P. Cheng, and J. Chen, "Privacy preserving collaborative computing: Heterogeneous privacy guarantee and efficient incentive mechanism," *IEEE Trans. Signal Process.*, vol. 67, no. 1, pp. 221–233, Jan. 2019.

[38] A. Wang, W. Liu, T. Dong, X. Liao, and T. Huang, "DisEHPPC: Enabling heterogeneous privacy-preserving consensus-based scheme for economic dispatch in smart grids," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 5124–5135, Jun. 2022.

[39] X. Wang, S. Garg, H. Lin, G. Kaddoum, J. Hu, and M. S. Hossain, "PPCS: An intelligent privacy-preserving mobile-edge crowdsensing strategy for industrial IoT," *IEEE Internet Things J.*, vol. 8, no. 13, pp. 10288–10298, Jul. 2021.

[40] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta, "Privacy amplification by iteration," in *Proc. IEEE 59th Annu. Symp. Found. Comput. Sci. (FOCS)*, Oct. 2018, pp. 521–532.

[41] M. Abadi et al., "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 308–318.

[42] R. Torkzadehmahani, P. Kairouz, and B. Paten, "DP-CGAN: Differentially private synthetic data and label generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 98–104.

[43] M. D. Reid and R. C. Williamson, "Generalised Pinsker inequalities," 2009, *arXiv:0906.1244*.

[44] A. Marsiglietti and V. Kostina, "A lower bound on the differential entropy of log-concave random vectors with applications," *Entropy*, vol. 20, no. 3, p. 185, Mar. 2018.

[45] G. Kamath, V. Singhal, and J. Ullman, "Private mean estimation of heavy-tailed distributions," in *Proc. 33rd Int. Conf. Learn. Theory (PMLR)*, 2020, pp. 2204–2235.

[46] K. Kenthapadi, A. Korolova, I. Mironov, and N. Mishra, "Privacy via the Johnson–Lindenstrauss transform," 2012, *arXiv:1204.2606*.

[47] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, no. 6, pp. 1191–1253, Jun. 2003.

[48] D. McAllester and K. Stratos, "Formal limitations on the measurement of mutual information," in *Proc. Int. Conf. Artif. Intell. Statist. (PMLR)*, 2020, pp. 875–884.

[49] S. Sreekumar and Z. Goldfeld, "Neural estimation of statistical divergences," *J. Mach. Learn. Res.*, vol. 23, no. 1, pp. 5460–5534, 2022.

[50] G. Kamath, J. Li, V. Singhal, and J. Ullman, "Privately learning high-dimensional distributions," in *Proc. Int. Conf. Learn. Theory (PMLR)*, 2019, pp. 1853–1902.

[51] V. Karwa and S. Vadhan, "Finite sample differentially private confidence intervals," 2017, *arXiv:1711.03908*.

[52] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Proc. Theory Cryptogr. Conf.* Beijing, China: Springer, 2016, pp. 635–658.

[53] M. B. Cohen, Y. T. Lee, G. Miller, J. Pachocki, and A. Sidford, "Geometric median in nearly linear time," in *Proc. 48th Annu. ACM Symp. Theory Comput.*, Jun. 2016, pp. 9–21.

[54] D. Hsu and S. Sabato, "Loss minimization and parameter estimation with heavy tails," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 543–582, 2016.

[55] A. Zamani, T. J. Oechtering, and M. Skoglund, "Bounds for privacy-utility trade-off with non-zero leakage," 2022, *arXiv:2201.08738*.

[56] T. van Erven and P. Harremos, "Rényi divergence and Kullback–Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.

[57] T. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[58] I. Sason, "On reverse Pinsker inequalities," 2015, *arXiv:1503.07118*.

[59] A. Xu, "Continuity of generalized entropy and statistical learning," 2021, *arXiv:2012.15829*.

[60] S. Minsker, "Geometric median and robust estimation in Banach spaces," *Bernoulli*, vol. 21, no. 4, pp. 2308–2335, Nov. 2015.

[61] B. Rassouli and D. Gündüz, "Optimal utility-privacy trade-off with total variation distance as a privacy measure," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 594–603, 2020.

[62] E. Erdemir, P. L. Dragotti, and D. Gündüz, "Active privacy-utility trade-off against a hypothesis testing adversary," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2660–2664.

Theshani Nuradha (Graduate Student Member, IEEE) received the B.Sc. degree (Hons.) from the Department of Electronic and Telecommunication Engineering (ENTC), University of Moratuwa, Sri Lanka, December, in 2018. She is currently pursuing the Ph.D. degree with the School of Electrical and Computer Engineering, Cornell University, USA. During her bachelor's studies, she was a full-time Research Intern with the Singapore University of Technology and Design, Singapore, in 2017. Prior to joining Cornell, she was a Lecturer on contract with ENTC from 2019 to 2020.

Ziv Goldfeld (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in Electrical and Computer Engineering from Ben-Gurion University, Israel, in 2012, 2014, and 2017, respectively. From 2017 to 2019, he was a Post-Doctoral Fellow with the Laboratory for Information and Decision Systems (LIDS), MIT. He is currently an Assistant Professor in Electrical and Computer Engineering with Cornell University. He was a recipient of several awards, such as the NSF CAREER Award, the International Business Machines Corporation (IBM) Academic Award, and the Rothschild Fellowship.