Protecting Global Properties of Datasets with Distribution Privacy Mechanisms

Michelle Chen
The University of Melbourne

Olga Ohrimenko
The University of Melbourne

ABSTRACT

Alongside the rapid development of data collection and analysis techniques in recent years, there is increasingly an emphasis on the need to address information leakage associated with such usage of data. To this end, much work in the privacy literature is devoted to the protection of individual users and contributors of data. However, many situations instead require a different notion of data confidentiality, involving global properties aggregated over the records of a dataset. Such notions of information protection are, for instance, particularly applicable in settings involving business and organization data, where global properties may reflect trade secrets and demographic data, which can be harmful if mishandled. Recent work on property inference attacks further illustrates the susceptibility of data analysis algorithms to leaking these global properties of data, highlighting the importance of developing mechanisms that can protect such information.

In this work, we demonstrate how a *distribution privacy* framework can be applied to formalize the problem of protecting global properties of datasets. Given this framework, we investigate several mechanisms and their tradeoffs for providing this notion of data confidentiality. We first demonstrate how the Wasserstein Mechanism from the established Pufferfish privacy framework can be adapted to provide distribution privacy. Then, we propose an Expected Value Mechanism to address some of the Wasserstein Mechanism's utility and efficiency limitations. We analyze the theoretical protection guarantees offered by these mechanisms under various data assumptions, then implement and empirically evaluate these mechanisms for several data analysis tasks. The results of our experiments show that our mechanisms can indeed reduce the effectiveness of practical property inference attacks while providing utility substantially greater than a crude group differential privacy baseline. Our work thus provides the groundwork for theoretically supported mechanisms for protecting global properties of datasets.

1 INTRODUCTION

Recent years have seen an influx of readily accessible data, spurring the development of a variety of sophisticated technologies in the field of data analysis. In turn, this has enabled advancements in an assortment of application areas ranging from agriculture [23] and healthcare [37] to cybersecurity [7] and natural language processing [33]. Alongside these advancements, however, there is increasingly an emphasis on the need to also address privacy issues associated with the use of data. Indeed, data is often sensitive in nature, with its misuse potentially having implications such as leaking personal information, revealing trade secrets, or causing security issues. This raises the question of how data analysis can be effectively performed without compromising sensitive properties of data.

Various notions of privacy have been proposed and studied, especially from the viewpoint of protecting individual users and contributors of data [13, 18, 26, 36]. Of these, differential privacy [18] currently reigns as one of the most commonly accepted and widely used, having been applied, for instance, to protect census data [2] and user data collected by technological companies [12]. The idea behind differential privacy is that the outcome of a data analysis algorithm cannot harm an individual if the algorithm's output is the same as if the individual's data were not used in the analysis. On this basis, a variety of differential privacy mechanisms have been developed using noise to obscure the effect of including or excluding individual records of data [16, 19].

There are, however, many situations where it is desirable to protect global properties of a dataset, rather than properties of individual records. This notion is particularly relevant in the context of business data, where properties aggregated over the records of a dataset may reflect trade secrets or other intellectual property [26]. For instance, consider a speech recognition software company seeking to release a machine learning model trained on voice data. Properties such as the overall proportion of particular accents or speech patterns in training data may have contributed to the effectiveness of the model and could hence be considered proprietary information [4]. It would thus be in the company's interests to ensure that their released machine learning model does not leak these properties so as to prevent this information from being exploited by their competitors.

The need to protect global properties of datasets can moreover extend beyond the setting of business data. Mishandling of demographic data, for instance, can contribute to prejudice or discrimination both directly and indirectly [30]. Thus, as an example, a hospital looking to share patient treatment data may desire to protect patient demographic details to avoid unfounded claims of correlations between diseases and certain demographics. Sharing of demographic data in some situations can furthermore exacerbate sensitive political issues, such as in the case of the 1932 Lebanon census, which revealed the religious makeup of the Lebanese population. Indeed, a national census has not been conducted in Lebanon since 1932 [41], due to the sensitivity of the matter of religious balance in the region [29].

Each of these described situations demonstrates a need to protect global properties aggregated over an entire dataset, a notion differing from the privacy of individuals typically considered in the literature. Moreover, the presence of possible correlations between sensitive information and other attributes of data implies that simply omitting sensitive attributes during analysis can be insufficient to protect this information [44]. In particular, sensitive properties may still be derivable by exploiting known correlations with the remaining, supposedly non-sensitive attributes.

Recently, several works have developed attacks demonstrating how data analysis algorithms can leak these global properties of datasets [4, 21, 28, 31, 35, 40, 45]. Attacks targeting such properties are known as property inference attacks, and concern a form of information leakage outside the privacy model considered by differential privacy [4, 21]. Ateniese et al. [4] were the first to consider this form of information leakage, observing that a machine learning model's learned parameters and behavior could unintentionally capture global properties of training data. They moreover established the insufficiency of differential privacy in protecting against this form of leakage.

Despite the development of these property inference attacks, there is very limited work on frameworks and mechanisms for defending against such attacks [40, 44]. In particular, the only defenses which have been evaluated are specific to particular property inference attacks [21, 31] and lack theoretical guarantees. On the other hand, theoretical frameworks aiming to capture protection of global properties of datasets, such as attribute privacy [44], require restrictive assumptions and have not yet been evaluated in practice. As such, the general applicability of these frameworks to protecting global properties of datasets is not well understood.

Our work aims to address this issue, establishing the groundwork for a theoretically supported approach for protecting against property inference attacks. Specifically, our contributions are:

- We demonstrate how the distribution privacy framework [24] can be applied to formalize the notion of protecting global properties of datasets.
- We demonstrate how the Wasserstein Mechanism for Pufferfish privacy [39] can be adapted to satisfy arbitrary instantiations of distribution privacy.
- We propose the Expected Value Mechanism, an alternative mechanism for attaining distribution privacy which addresses some of the utility and efficiency limitations of the Wasserstein Mechanism.
- We investigate the data assumptions underlying the Wasserstein Mechanism and Expected Value Mechanism, and analyze how these assumptions can be relaxed while maintaining each mechanism's theoretical privacy guarantees.
- We implement each of our theoretically analyzed mechanisms and empirically evaluate the privacy-utility tradeoff they offer. Our mechanisms attain substantially greater utility than a group differential privacy baseline.
- We evaluate the effectiveness of our mechanisms in protecting global properties against property inference attacks. Our results show that such attacks can easily infer sensitive global properties if not protected, while our mechanisms can significantly reduce attack accuracy.

2 RELATED WORK

Existing literature on data protection has largely focused on ensuring privacy for individual users and contributors of data [1, 10, 18, 32]. Much of this work is built on differential privacy [18], a framework based on the view that a data analysis algorithm cannot harm an individual if its output is the same whether or not the individual's data was used in the analysis. Many variants of differential privacy have been proposed to satisfy the privacy and utility requirements of different application contexts [1, 10, 22, 32, 43], however, these are still designed to provide individual-level privacy rather than protection for global properties of datasets.

On the other hand, recent work has developed various property inference attacks which demonstrate the vulnerability of data analysis algorithms to leaking these global properties of datasets [4, 21, 31, 35, 40, 45]. Ateniese et al. [4] was the first to formulate such an attack, demonstrating that machine learning models could unintentionally leak global properties of training data. In contrast to privacy attacks such as membership inference attacks [38] and model inversion attacks [20], their attack aimed to discover properties aggregated over all records in a dataset rather than properties of individual records. For instance, in the context of a speech recognition engine, Ateniese et al. [4] demonstrated that their attack could infer the proportion of training data generated by Indian speakers. Further property inference attacks have since been developed, demonstrating the threat of global property leakage for deep neural networks [21], large convolutional networks [40], in federated learning setting [31], in black-box settings [45] and in settings where an attacker can poison the data [28].

Currently, there is limited work on defense against property inference attacks [40, 44]. Ganju et al. [21] and Melis et al. [31] suggested intuitive defenses based on regularization or adding noise, but performed only preliminary empirical evaluation and provided no guarantees on the effectiveness of their suggested defenses. On the other hand, Suri and Evans [40] proposed a formal model of property inference attacks as a cryptographic game, but did not propose possible defenses against such attacks.

Similarly, while some privacy frameworks have considered the issue of protecting global properties of datasets, these still lack applicable rigorous defense mechanisms. Knowledge hiding [14, 42] takes a syntactic approach to hiding sensitive properties aggregated over a dataset, but does not provide rigorous privacy guarantees [26]. Some theoretical, rigorous privacy frameworks can capture protection of global properties of datasets, such as Pufferfish privacy [26], distribution privacy [24] and attribute privacy [44]. However, these have been studied in specific settings or under restrictive data assumptions, and are accompanied by little to no empirical evaluation in the context of protecting global properties of datasets.

From a technical point of view, our work also has similarities to various other privacy frameworks. Distributional differential privacy [5] treats datasets as random variables with some class of possible dataset distributions, and we consider a similar definition of privacy in which datasets are treated as random variables. However, while Bassily et al. [5] use these distributions to model adversarial uncertainty, our work uses them to model the generation of datasets with differing sensitive global properties. Noiseless privacy [6] also treats datasets as random variables, to analyze functions of datasets which may not require additional noise to achieve privacy. We emphasize that both distributional differential privacy and noiseless privacy only consider protecting individual records in a dataset, as opposed to the global properties of interest in our work.

3 PRIVACY FRAMEWORK

In this section, we formalize the problem of protecting global properties of datasets using a variation of differential privacy known as distribution privacy [24]. As motivation, let us consider the following example application.

Example: Census Data. A government body seeks to release census data for various local regions. However, they are concerned that particular financial statistics, such as the proportion of low income earners, may validate or create stigma towards certain regions if published. Since such stigma can harm business and discourage people from moving into the area, the government body would like to withhold these statistics. Their goal is thus to release a selective summary of census data which ensures that sensitive properties related to the population's financial status remain protected.

In this example, the proportion of low income earners is global, in the sense that it is an aggregate over all the records in a dataset. Thus, established privacy frameworks which focus on privacy for individual records of data are not directly applicable to this setting. We demonstrate how a variation of differential privacy known as distribution privacy can be used as a suitable alternative.

3.1 Notation and Terminology

We assume the existence of a *data curator* who owns a dataset $\mathfrak{Data}$. This data curator would like to release a *query*, or a function of their dataset, $F(\mathfrak{Data})$, while protecting particular global properties of $\mathfrak{Data}$. To achieve this goal, the data curator applies a *randomized mechanism* $\mathcal{M}$ which performs computations on $\mathfrak{Data}$ and produces some randomized output $\mathcal{M}(\mathfrak{Data})$.

To an *attacker* who does not have access to the dataset, $\mathfrak{Data}$ is a random variable. We refer to possible values of $\mathfrak{Data}$ as *dataset instances* and use D to denote a possible dataset instance. We use θ to denote a possible probability distribution of $\mathfrak{Data}$ and, for a given distribution θ and function F, we use f_θ to represent the probability distribution of $F(\mathfrak{Data})$ given $\mathfrak{Data} \sim \theta$.

3.2 Differential Privacy

Differential privacy [18] is a framework of privacy definitions and mechanisms based on the view that the outcome of a data analysis algorithm does not harm an individual if the algorithm's output is the same as if the individual's data were not used in the analysis.

Definition 3.1 (Differential Privacy). A mechanism $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy if for all datasets D and D' which differ on at most one record and all subsets $S \subseteq \text{Range}(\mathcal{M})$,

$$\Pr(\mathcal{M}(D) \in S) \leq \exp(\epsilon) \times \Pr(\mathcal{M}(D') \in S) + \delta.$$

Differential privacy does not apply directly to our problem, as it protects individual records rather than global properties aggregated over many records. Group differential privacy [19] somewhat addresses this issue, extending protection to groups of records of data. However group differential privacy mechanisms generally only provide meaningful utility when protecting small groups relative to the size of the whole dataset. Such mechanisms are thus unsuitable to protect sensitive global properties, which may involve all records of a dataset, as we also empirically demonstrate in Section 6.2.

To address this issue, we instead adopt the distribution privacy framework [24], a variation which applies the principles of differential privacy to protect the underlying distribution of a dataset.

3.3 Distribution Privacy

Distribution privacy [24] is similar to differential privacy in that it takes a probabilistic view on privacy defined in terms of a randomized mechanism's output distribution. The key distinction, however, is that distribution privacy is defined using the possible underlying distributions of a dataset, rather than the presence or absence of individual records. The problem of protecting global properties of datasets can then be captured by modeling possible dataset distributions according to the presence or values of particular global properties.

Formally, the distribution privacy framework is instantiated by specifying a set Θ of data distributions along with a subset $\Psi \subseteq \Theta \times \Theta$ of pairs of distributions. Each distribution $\theta \in \Theta$ may be viewed as a possible probability distribution of the dataset $\mathfrak{Data}$, intuitively representing an attacker's beliefs on how $\mathfrak{Data}$ may have been generated. The pairs of distributions Ψ should then reflect sensitive properties to be protected from the attacker.

Definition 3.2 (Distribution Privacy). A mechanism $\mathcal{M}$ satisfies (ϵ, δ)-distribution privacy with respect to a set of distribution pairs $\Psi \subseteq \Theta \times \Theta$ if for all pairs $(\theta_i, \theta_j) \in \Psi$ and all subsets $S \subseteq \text{Range}(\mathcal{M})$,

$$\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_i) \leq \exp(\epsilon) \times \Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_j) + \delta,$$

where the expression $\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta)$ denotes the probability that $\mathcal{M}(\mathfrak{Data}) \in S$ given $\mathfrak{Data} \sim \theta$.

We emphasize that $\mathfrak{Data}$ is a random variable rather than a fixed dataset instance. Thus, the probabilities in the above equation are w.r.t. randomness from both the mechanism $\mathcal{M}$ and the distribution of $\mathfrak{Data}$. This is in contrast to those in the definition of differential privacy (see Definition 3.1), which depend only on randomness in $\mathcal{M}$. In this manner, distribution privacy can be considered to share similarities with noiseless privacy [6], Pufferfish privacy [26] and other notions of privacy [5, 27] which are defined over possible distributions of a dataset rather than fixed dataset instances.

Using the distribution privacy framework, we can model confidentiality of global properties of a dataset $\mathfrak{Data}$ by identifying the possible underlying distributions of $\mathfrak{Data}$ given particular values for the sensitive properties of interest.

Example: Census Data. Consider how distribution privacy can be used for the scenario as described at the beginning of this section. The government body owns a dataset $\mathfrak{Data}$ of complete census data and would like to release a summary $F(\mathfrak{Data})$ while protecting sensitive properties related to the population's financial status. The government body could formalize their data confidentiality needs (i.e., hiding the proportion of low income earners) by first identifying data distributions θ_p modeling data generation scenarios given each possible proportion p of low income earners. They may then decide that an attacker should not be able to infer the proportion p to within an absolute error of k. Then distribution privacy is defined w.r.t. Ψ, where each pair of distributions in Ψ takes the form takes the form $(\theta_{r \mp d}, \theta_{r \pm d})$ with $|d| \leq k$.

3.4 Relation to Other Privacy Frameworks

While we adopt the distribution privacy definition of Kawamoto and Murakami [24], our work applies their definition in a different context. Kawamoto and Murakami [24] focused on protecting the distribution of individual records for location based services, corresponding to the case where $\mathfrak{Data}$ is a single record of data. In contrast, we consider $\mathfrak{Data}$ to be an entire dataset and investigate how global properties of such datasets can be protected by modeling such properties using the distribution privacy framework.

In this context, the distribution privacy framework may also be viewed as a generalization of the distributional attribute privacy framework proposed by Zhang et al. [44] for protecting sensitive attributes aggregated across a dataset. However, while distributional attribute privacy requires conditional marginal distributions to be known for all attributes of a dataset, the distribution privacy framework requires such distributions to be known only for the sensitive properties of interest. Distribution privacy thus has the capacity to capture a wider range of global properties while requiring fewer data assumptions.

From a technical perspective, distribution privacy can be considered as an instantiation of the Pufferfish framework [26]. The two privacy frameworks may in fact be viewed as expressively equivalent, in the following sense.

Theorem 3.1. *Every instantiation of distribution privacy can be equivalently expressed as an instantiation of Pufferfish privacy and every instantiation of Pufferfish privacy can be equivalently expressed as an instantiation of distribution privacy.*

An overview of Pufferfish privacy along with details of the proof of Theorem 3.1 are provided in Appendix A.1.

Theorem 3.1 suggests that properties and mechanisms designed for Pufferfish privacy can generally be adapted for distribution privacy, and vice versa. In our case, we use the distribution privacy framework as it allows more straightforward modeling of privacy needs involving global properties of datasets. Nonetheless, the distribution privacy results we prove throughout this work can be readily modified to apply to the Pufferfish framework as well.

Properties of Distribution Privacy. Finally, distribution privacy has various key properties which justify its status as a formal and rigorous privacy framework. In particular, it satisfies two properties regarded by some as fundamental privacy axioms, the post-processing property and convexity [24, 25]. We provide a summary of the key such properties of distribution privacy in Appendix A.2. Note again that Theorem 3.1 implies that such properties include those shown to be satisfied by Pufferfish privacy [26, 39], along with those satisfied by distribution privacy [24].

4 GENERAL MECHANISMS

In this section, we discuss a general mechanism for providing distribution privacy. Specifically, we demonstrate how the Wasserstein Mechanism for Pufferfish privacy [39] can be adapted to satisfy an arbitrary instantiation of distribution privacy. We then identify limitations of this mechanism and suggest a variation, the Approximate Wasserstein Mechanism, to address some of these limitations.

For an instantiation of distribution privacy specified by a set of pairs of distributions $\Psi \subseteq \Theta \times \Theta$, let $F(\mathfrak{Data})$ denote a query, or function of the data which the data curator would like to release. We will assume that F takes values in $\mathbb{R}^m$. Then, a mechanism satisfying distribution privacy should intuitively apply enough noise to obscure the differences in $F(\mathfrak{Data})$ caused by possible differences in the underlying distribution of $\mathfrak{Data}$. Letting f_θ denote the distribution of $F(\mathfrak{Data})$ given $\mathfrak{Data} \sim \theta$, we thus need to apply enough noise to prevent an attacker from determining whether $F(\mathfrak{Data})$ was drawn from f_{θ_i} or f_{θ_j}, for each pair of distributions $(\theta_i, \theta_j) \in \Psi$. For Pufferfish privacy, Song et al. [39] identifies ∞-Wasserstein distance as one suitable measure for determining the amount of noise required.

Definition 4.1 (∞-Wasserstein Distance Using the L_1 Norm). Let μ and ν be two distributions on $\mathbb{R}^m$, and let $\Gamma(\mu, \nu)$ be the set of all joint distributions with marginals μ and ν. The ∞-Wasserstein distance $W_\infty(\mu, \nu)$ between μ and ν is defined as

$$W_\infty(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \max_{(x,y) \in \mathrm{supp}(\gamma)} \|x - y\|_1.$$

Intuitively, each $\gamma \in \Gamma(\mu, \nu)$ may be interpreted as a way of transforming μ into ν by shifting probability mass between the two distributions. The expression $\max_{(x,y)\in\mathrm{supp}(\gamma)} \|x - y\|_1$ can then be interpreted as the cost of γ, representing the maximum L_1 distance traveled by a probability mass in the shifting described by γ. Thus, the ∞-Wasserstein distance between distributions μ and ν is intuitively the maximum distance traveled by a probability mass when transforming μ into ν in the most cost-efficient manner possible.

The Wasserstein Mechanism of Song et al. [39], adapted for the distribution privacy framework, scales Laplace noise to a notion of sensitivity $\Delta_W(\Psi, F)$ defined by the maximum ∞-Wasserstein distance between pairs of distributions of interest

$$\Delta_W(\Psi, F) = \sup_{(\theta_i,\theta_j)\in\Psi} W_\infty(f_{\theta_i}, f_{\theta_j}).$$

When this quantity is well-defined, the Wasserstein Mechanism achieves distribution privacy by adding Laplace noise with scale $\Delta_W(\Psi, F)/\epsilon$ independently to each component of $F(\mathfrak{Data})$.

Theorem 4.1. *The mechanism $\mathcal{M}$ which adds Laplace noise $Z_k \sim \mathrm{Lap}(\Delta_W(\Psi, F)/\epsilon)$ independently to each component of $F(\mathfrak{Data})$ satisfies $(\epsilon, 0)$-distribution privacy with respect to Ψ.*

Recalling the equivalence between distribution privacy and Pufferfish privacy established in Theorem 3.1, the proof of Theorem 4.1 is immediate from the Pufferfish privacy guarantees Song et al. [39] already prove for the Wasserstein Mechanism.

4.1 Approximate Wasserstein Mechanism

The Wasserstein Mechanism is very versatile, as it is applicable to general instantiations of distribution privacy with almost no restrictions on the distributions and functions involved. It does, however, implicitly assume that the ∞-Wasserstein distance between f_{θ_i} and f_{θ_j} is well-defined for each pair of distributions $(\theta_i, \theta_j) \in \Psi$. This can be an issue if the query function F is unbounded, as each distribution f_θ may then have unbounded support and the relevant ∞-Wasserstein distances then may not be well-defined.

In many practical settings it is reasonable to assume that F is bounded (e.g., average age). However, the relevant ∞-Wasserstein

distances can still be very large if F takes on a large range of values, even if most values only occur with very low probability (e.g., average salary). In such scenarios, the Wasserstein Mechanism may apply more noise than is necessary to provide privacy for the majority of cases. To address this limitation, we propose a variation of the Wasserstein Mechanism which requires less noise in such cases in exchange for a small probability of loss of privacy.

Our proposed variation of the Wasserstein Mechanism essentially allows a low probability set of possible values of F to be disregarded when computing ∞-Wasserstein distances. In doing so, it can achieve (ϵ, δ)-distribution privacy with a smaller amount of noise and thus potentially greater utility.

Formally, define a measure for scaling noise based on a generalization of ∞-Wasserstein distance as follows. Given two distributions μ and ν, we will say that μ and ν are (W, δ)-*close* if there exists a distribution $\gamma \in \Gamma(\mu, \nu)$ and a subset $R \subseteq \text{supp}(\gamma)$ such that

$$\|x - y\|_1 \leq W \qquad \forall\, (x, y) \in R$$

and

$$\int \int_{(x,y)\in R} \gamma(x, y)\, dx\, dy \geq 1 - \delta.$$

Recalling the view of each $\gamma \in \Gamma(\mu, \nu)$ being a transformation of probability distributions, we may interpret (W, δ)-closeness as the ability to transform one distribution into another by shifting each probability mass by a distance of at most W, with exceptions of mass at most δ. Observe that the case $\delta = 0$ corresponds exactly to ∞-Wasserstein distance, since the definitions imply that two distributions μ and ν are $(W, 0)$-close if and only if $W_\infty(\mu, \nu) = W$. Thus, (W, δ)-closeness may be viewed as a generalization of ∞-Wasserstein distance which allows for a small amount of probability mass to be disregarded when computing closeness.

The parameter δ allows for two distributions μ and ν to be considered (W, δ)-close for values of W that are potentially much smaller than $W_\infty(\mu, \nu)$. To illustrate this with an example, consider distributions μ and ν on the set $\{1, 2, \ldots, 100\}$ such that μ assigns probabilities 0.6, 0.2, 0, 0.2 and ν assigns probabilities 0.4, 0.3, 0.2, 0.1 to the elements 1, 2, 3, 100 respectively (and both assign probability 0 for elements 4, 5, …, 99). Then, as shown in Figure 1, the ∞-Wasserstein distance $W_\infty(\mu, \nu) = 97$ is large since probability mass must be shifted from $\mu(100)$ to $\nu(3)$. However, by disregarding the shift associated with this particular probability mass, we see that μ and ν are $(W, 0.1)$-close with $W = 1$.

This difference suggests that scaling noise according to (W, δ)-closeness can sometimes result in significantly less noise than scaling to ∞-Wasserstein distance. Moreover, the possibility for a small mass δ to be disregarded allows (W, δ)-closeness to be well-defined even in cases where ∞-Wasserstein distance may not be. Thus, using the notion of (W, δ)-closeness, we propose the Approximate Wasserstein Mechanism.

Theorem 4.2. *Suppose that, for all pairs $(\theta_i, \theta_j) \in \Psi$, the distributions f_{θ_i} and f_{θ_j} are (W, δ)-close. Then, the mechanism $\mathcal{M}$ which adds Laplace noise $Z_k \sim \text{Lap}(W/\epsilon)$ independently to each component of $F(\mathfrak{Data})$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ.*

Proof. Let $(\theta_i, \theta_j) \in \Psi$ be a pair of distributions. Since f_{θ_i} and f_{θ_j} are (W, δ)-close, there exists a distribution $\gamma \in \Gamma(f_{\theta_i}, f_{\theta_j})$ and a subset $R \subseteq \text{supp}(\gamma)$ such that

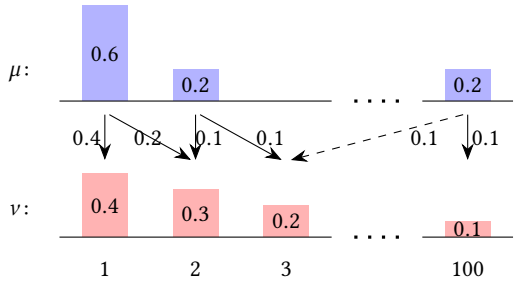


Figure 1: Example demonstrating the distinction between ∞-Wasserstein distance and (W, δ)-closeness. $W_\infty(\mu, \nu) = 97$ is large due to the distance of the shift represented by the dashed arrow, while disregarding this arrow shows that μ and ν are $(W, 0.1)$-close with $W = 1$.

$$\|t - s\|_1 \leq W \qquad \forall\, (t, s) \in R \tag{1}$$

and

$$\int \int_{(t,s)\in R} \gamma(t, s)\, dt\, ds \geq 1 - \delta. \tag{2}$$

Now, for all $S \subseteq \text{Range}(\mathcal{M})$,

$$\begin{aligned}
&\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_i) \\
&= \int_t \Pr(F(\mathfrak{Data}) = t \mid \theta_i) \Pr(Z + t \in S)\, dt \\
&= \int_t \int_s \gamma(t, s) \Pr(Z + t \in S)\, ds\, dt \\
&= \int \int_{(t,s)\in R} \gamma(t, s) \Pr(Z + t \in S)\, ds\, dt \\
&\quad + \int \int_{(t,s)\notin R} \gamma(t, s) \Pr(Z + t \in S)\, ds\, dt.
\end{aligned}$$

Since each component of noise $Z_k \sim \text{Lap}(W/\epsilon)$ is independently sampled, the Laplace Mechanism from differential privacy [18] implies that, for all $t, s \in \mathbb{R}^m$ such that $\|t - s\|_1 \leq W$,

$$\Pr(Z + t \in S) \leq \exp(\epsilon) \Pr(Z + s \in S).$$

Using (1), we then have

$$\begin{aligned}
&\int \int_{(t,s)\in R} \gamma(t, s) \Pr(Z + t \in S)\, ds\, dt \\
&\leq \exp(\epsilon) \int \int_{(t,s)\in R} \gamma(t, s) \Pr(Z + s \in S)\, ds\, dt.
\end{aligned} \tag{3}$$

Moreover, from (2), we have

$$\begin{aligned}
&\int \int_{(t,s)\notin R} \gamma(t, s) \Pr(Z + t \in S)\, ds\, dt \\
&\leq \int \int_{(t,s)\notin R} \gamma(t, s)\, ds\, dt \\
&= 1 - \int \int_{(t,s)\in R} \gamma(t, s)\, ds\, dt \\
&\leq 1 - (1 - \delta) = \delta.
\end{aligned} \tag{4}$$

Finally, putting (3) and (4) together gives

$$\begin{aligned}
&\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_i) \\
&\leq \exp(\epsilon) \int \int_{(t,s)\in R} \gamma(t,s) \Pr(Z + s \in S)\, ds\, dt + \delta \\
&\leq \exp(\epsilon) \int_s \Pr(F(\mathfrak{Data}) = s \mid \theta_j) \Pr(Z + s \in S)\, ds + \delta \\
&= \exp(\epsilon) \Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_j) + \delta.
\end{aligned}$$

It follows that $\mathcal{M}$ satisfies (ϵ, δ)-distribution privacy. □

We remark that (W, δ)-closeness can be also viewed as an approximation of the ∞-Wasserstein distance between two distributions using a subset of their supports. More precisely, given two distributions μ and ν, suppose that there exist subsets $R_1 \subseteq \text{supp}(\mu)$ and $R_2 \subseteq \text{supp}(\nu)$ such that $\Pr(\mu \in R_1) \geq 1 - \delta/2$ and $\Pr(\nu \in R_2) \geq 1 - \delta/2$. The union bound can then be used to show that $\Pr(\gamma \in R_1 \times R_2) \geq 1 - \delta$ for every coupling γ of μ and ν. Hence, μ and ν are (W, δ)-close, where

$$W = \inf_{\gamma \in \Gamma(\mu,\nu)} \max_{(x,y) \in \text{supp}(\gamma) \cap (R_1 \times R_2)} \|x - y\|_1.$$

Comparing with the definition of ∞-Wasserstein distance, the above expression can be interpreted as only considering the movement of probability masses between the subsets $R_1 \subseteq \text{supp}(\mu)$ and $R_2 \subseteq \text{supp}(\nu)$. Thus, (W, δ)-closeness formalizes the notion that $W_\infty(\mu, \nu)$ can be approximated using subsets of the supports of μ and ν.

This is significant since methods for computing ∞-Wasserstein distance typically proceed by first approximating continuous distributions with bounded or discrete distributions [34]. Thus, the Approximate Wasserstein Mechanism shows that such approximations of computing the distance can be applied in practice while maintaining formal privacy guarantees.

4.2 Functions Bounded with High Probability

As an example to demonstrate the applicability of the Approximate Wasserstein Mechanism, consider settings in which $F(\mathfrak{Data})$ takes some fixed range of values with high probability. That is, suppose that there exists some constant c such that

$$\|F(\mathfrak{Data}) - \mathbb{E}[F(\mathfrak{Data})]\|_1 \leq c$$

with probability at least $1 - \delta/2$ for each $\mathfrak{Data} \sim \theta \in \Theta$. Then, each distribution f_θ is bounded with high probability.

Now, consider a notion of sensitivity $\Delta_E(\Psi, F)$ defined by the worst case difference between the expected values of f_{θ_i} and f_{θ_j},

$$\Delta_E(\Psi, F) = \sup_{(\theta_i,\theta_j)\in\Psi} \|\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]\|_1.$$

Proposition 4.1. *Suppose that $\|F(\mathfrak{Data}) - \mathbb{E}[F(\mathfrak{Data})]\|_1 \leq c$ with probability at least $1 - \delta/2$ for each $\mathfrak{Data} \sim \theta \in \Theta$. Then, the distributions f_{θ_i} and f_{θ_j} are $(\Delta_E(\Psi, F) + 2c, \delta)$-close for all pairs $(\theta_i, \theta_j) \in \Psi$.*

The proof of Proposition 4.1 is provided in Appendix C.1. Given Proposition 4.1, Theorem 4.2 implies that applying Laplace noise with scale $(\Delta_E(\Psi, F) + 2c)/\epsilon$ achieves (ϵ, δ)-distribution privacy with respect to Ψ. As the quantity $\Delta_E(\Psi, F) + 2c$ requires only expected values and bounds for f_θ to be computed, the Approximate Wasserstein Mechanism thus circumvents the high computational cost associated with computing ∞-Wasserstein distances in the Wasserstein Mechanism [39] in this scenario.

5 EFFICIENT MECHANISM FOR SPECIALIZED SETTINGS

While the Wasserstein Mechanism is theoretically applicable to general instantiations of distribution privacy, it can be too computationally expensive to use in practice [39]. Indeed, given a pair of distributions μ and ν, computing or even approximating $W_\infty(\mu, \nu)$ generally requires solving a difficult optimal transport problem and can thus be highly computationally expensive [8, 11, 34]. The Approximate Wasserstein Mechanism we proposed can accommodate more efficient approximations in some situations, however, in general, (W, δ)-closeness may still be difficult to compute.

We now discuss an alternative, efficient mechanism which can achieve distribution privacy under more restricted settings. We first propose the Expected Value Mechanism and show that it can provide distribution privacy when the pairs of distributions to be protected are translations of each other. We then describe various possible modifications for improving the utility provided by the Expected Value Mechanism. In particular, we analyze variations of the mechanism using directional assumptions on how changes in a sensitive property affect a query, and assumptions on adversarial uncertainty based on possible distributions of the dataset. Finally, we discuss how these assumptions can be relaxed, showing that our mechanisms can be used even when the assumptions underlying our analysis do not exactly hold.

5.1 Expected Value Mechanism

Consider a distribution privacy setting in which the pairs of distributions to be protected are translations of each other. This occurs, for instance, when each distribution can be assumed to be Gaussian with the same variance but different means. We propose the Expected Value Mechanism, which can achieve distribution privacy in such settings by applying Laplace noise or Gaussian noise scaled using the expected values of the distributions involved.

Formally, let us assume that the possible distributions f_{θ_i} and f_{θ_j} of the query function F are translations of each other for each pair $(\theta_i, \theta_j) \in \Psi$. That is, for each pair $(\theta_i, \theta_j) \in \Psi$, there exists some $c \in \mathbb{R}^m$ such that $f_{\theta_i}(t) = f_{\theta_j}(t + c)$ for all $t \in \mathbb{R}^m$. The Laplace variant of the Expected Value Mechanism achieves distribution privacy by applying Laplace noise proportional to the worst case L_1 distance between expected values

$$\Delta_{E,1}(\Psi, F) = \sup_{(\theta_i,\theta_j)\in\Psi} \|\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]\|_1.$$

Theorem 5.1. *Suppose that f_{θ_i} is a translation of f_{θ_j} for every pair $(\theta_i, \theta_j) \in \Psi$. Then, the mechanism $\mathcal{M}$ which adds Laplace noise $Z_k \sim \text{Lap}(\Delta_{E,1}(\Psi, F)/\epsilon)$ independently to each component of $F(\mathfrak{Data})$ satisfies $(\epsilon, 0)$-distribution privacy with respect to Ψ.*

The Laplace variant of the Expected Value Mechanism may be viewed as a special case of the Wasserstein Mechanism described in Section 4. Indeed, if a pair of distributions are translations of each other, then their ∞-Wasserstein distance is given by the norm of the difference of their expected values. Theorem 5.1 thus follows

from the distribution privacy guarantees proved for the Wasserstein Mechanism in Theorem 4.1.

Gaussian noise can similarly be applied to achieve distribution privacy by considering the worst case L_2 distance

$$\Delta_{E,2}(\Psi, F) = \sup_{(\theta_i,\theta_j)\in\Psi} \|\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]\|_2.$$

Theorem 5.2. *Suppose that f_{θ_i} is a translation of f_{θ_j} for every pair $(\theta_i, \theta_j) \in \Psi$. Let $\sigma \geq c\Delta_{E,2}(\Psi, F)/\epsilon$, where $\epsilon \in (0, 1)$ and $c = \sqrt{2\ln(1.25/\delta)}$ for some $\delta > 0$. Then, the mechanism $\mathcal{M}$ which adds Gaussian noise $Z_k \sim \mathcal{N}(0, \sigma^2)$ independently to each component of $F(\mathfrak{Data})$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ.*

The proof of Theorem 5.2 follows from an application of the Gaussian Mechanism from differential privacy [19], using the condition that each pair of distributions $f_{\theta_i}, f_{\theta_j}$ are translations of each other. Details of the proof are provided in Appendix C.2.

5.2 Variant Using Directional Assumptions

We now describe our first variant of the Expected Value Mechanism, which uses assumptions on the direction of change of a query as sensitive properties vary. The main insight is that such assumptions can lead to distribution privacy being achievable while applying noise only in particular directions or to particular components of a query. For instance, consider a company seeking to release a set of customer statistics while protecting their proportion of female customers. Suppose that particular statistics, such as customer income, insurance claims, and investments, are known to increase by certain amounts as the proportion p of female customers varies. If the company models these statistics as changing in an approximately constant direction v as the proportion p changes, it may suffice to add noise only in the direction v to ensure privacy of this particular global property.

Expressed using the distribution privacy framework, we assume as before that the distributions f_{θ_i} and f_{θ_j} are translations of each other but now add noise only in the possible directions of $\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]$. We formalize this with the Directional Expected Value Mechanism, described in Algorithm 1. For simplicity, the mechanism is presented for the case where the differences of mean vectors $\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]$ are all parallel to some unit vector v. Slight modifications to the mechanism can relax this assumption, as we outline later in this section. Note also that Algorithm 1 can readily be modified to use Gaussian noise instead of Laplace noise.

Algorithm 1: Directional Expected Value Mechanism (dataset $\mathfrak{Data}$, query F, distribution pairs $\Psi \subseteq \Theta \times \Theta$, privacy parameters ϵ, δ, unit vector v)

1 **for** *each* $(\theta_i, \theta_j) \in \Psi$ **do**
2 Set $f_{\theta_i} = \Pr(F(\mathfrak{Data}) = \cdot \mid \theta_i)$, $f_{\theta_j} = \Pr(F(\mathfrak{Data}) = \cdot \mid \theta_j)$.
3 Calculate $\|\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]\|_2$.
4 **end**
5 Set $\Delta_{E,2}(\Psi, F) = \sup_{(\theta_i,\theta_j)\in\Psi} \|\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]\|_2$.
6 **return** $F(\mathfrak{Data}) + Yv$, where $Y \sim \mathrm{Lap}(\Delta_{E,2}(\Psi, F)/\epsilon)$.

Theorem 5.3. *Suppose that f_{θ_i} is a translation of f_{θ_j} and furthermore that the vector $\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]$ is parallel to the unit vector v for each $(\theta_i, \theta_j) \in \Psi$. Then, the mechanism $\mathcal{M}$ described in Algorithm 1 satisfies $(\epsilon, 0)$-distribution privacy with respect to Ψ.*

The proof of Theorem 5.3 makes use of the Laplace Mechanism from differential privacy [19] in a manner similar to the Expected Value Mechanism from Section 5.1. The details are provided in Appendix C.3.

In the case that the differences of mean vectors $\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]$ are not all parallel to some vector v, privacy can still be ensured by applying a small amount of additional noise in other directions as necessary. One approach is to identify a set of orthogonal vectors $v_1, v_2, \ldots, v_d$ which span the possible directions of $\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]$ as the pairs $(\theta_i, \theta_j) \in \Psi$ vary. Distribution privacy can then be achieved by applying noise proportional to $\Delta_{E,2}(\Psi, F)/\epsilon$ in the directions of $v_1, v_2, \ldots, v_d$. If the number of directions d is small in comparison to the total dimension m of the query, this can still improve the utility provided by the Expected Value Mechanism. Alternatively, approximations $\tilde{f_\theta}$ can be chosen to satisfy the directional assumptions and used in place of the exact distributions f_θ. In this case, analysis can proceed as for our general results on relaxing mechanism assumptions, discussed later in Section 5.4.

5.3 Variant Using Adversarial Uncertainty

We now exploit adversarial uncertainty in the data distributions, modifying the Expected Value Mechanism to add less noise where possible. As an example, consider a scenario in which a company wishes to release the average income of their customers, but considers their proportion of female customers to be a sensitive property. Though there may be a correlation between income and gender, the variance associated with average income may be enough to prevent accurate estimation of the proportion of female customers using the customers' average income. Thus, the company may be able to safely release this statistic with little to no applied noise.

For this modification, we assume the distribution f_θ of the query function to be multivariate Gaussian with mean μ_θ and covariance matrix Σ_θ for each $\theta \in \Theta$. We also assume that f_{θ_i} and f_{θ_j} have the same covariance matrix $\Sigma_{\theta_i} = \Sigma_{\theta_j}$ for each pair $(\theta_i, \theta_j) \in \Psi$. Thus, the distributions in concern are now not only translations of each other but also assumed to be normally distributed. Note that we consider possible relaxations of these conditions later in Section 5.4, where we discuss use of these mechanisms with approximations $\tilde{f_\theta}$ in place of the true distributions f_θ.

We first identify a set of conditions under which a query $F(\mathfrak{Data})$ may be safely released without additional noise.

Theorem 5.4. *Suppose that $f_\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$ for each $\theta \in \Theta$ and that $\Sigma_{\theta_i} = \Sigma_{\theta_j}$ for each $(\theta_i, \theta_j) \in \Psi$. Let $c = \sqrt{2\ln(1.25/\delta)}$. Then, the mechanism $\mathcal{M}$ which simply outputs $F(\mathfrak{Data})$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ as long as*

$$(\mu_{\theta_i} - \mu_{\theta_j})^{\mathrm{T}} \Sigma_{\theta_i}^{-1} (\mu_{\theta_i} - \mu_{\theta_j}) \leq (\epsilon/c)^2$$

for all $(\theta_i, \theta_j) \in \Psi$.

The key observation underlying Theorem 5.4 is that every multivariate Gaussian variable can be expressed as a transformation $X = AZ + \mu$ of a standard normal vector Z and, moreover, $\Pr(X =$

Algorithm 2: Eigenvector Gaussian Mechanism (dataset $\mathfrak{Data}$, query F, distribution pairs $\Psi \subseteq \Theta \times \Theta$, privacy parameters ϵ, δ, normalized eigenvectors $v_1, v_2, \ldots, v_m$)

Set $c = \sqrt{2\ln(1.25/\delta)}$.
for *each* $\theta \in \Theta$ **do**
  Set $\mu_\theta = \mathbb{E}[F(\mathfrak{Data}) \mid \mathfrak{Data} \sim \theta]$,
  $\Sigma_\theta = \mathrm{Cov}(F(\mathfrak{Data}) \mid \mathfrak{Data} \sim \theta)$.
end
for *each* $(\theta_i, \theta_j) \in \Psi$ **do**
  Calculate $\|\mu_{\theta_i} - \mu_{\theta_j}\|_2$.
end
Set $\Delta_{E,2}(\Psi, F) = \sup_{(\theta_i,\theta_j)\in\Psi} \|\mu_{\theta_i} - \mu_{\theta_j}\|_2$.
for *each eigenvector* v_k **do**
  for *each* $\theta \in \Theta$ **do**
    Set $\lambda^2_{\theta,k} = v_k^{\mathrm{T}} \Sigma_\theta v_k$.
    Set $\sigma^2_{\theta,k} = \max\left(0, (c\Delta_{E,2}(\Psi, F)/\epsilon)^2 - \lambda^2_{\theta,k}\right)$.
  end
  Set $\sigma_k^2 = \max_{\theta\in\Theta} \sigma^2_{\theta,k}$.
end
Set $\Sigma = \sum_k \sigma_k^2 v_k v_k^{\mathrm{T}}$.
return $F(\mathfrak{Data}) + Z$, where $Z \sim \mathcal{N}(0, \Sigma)$.

$x) = \frac{1}{|A|}\Pr(Z = A^{-1}x)$ for all $x \in \mathbb{R}^m$. The proof of the theorem then amounts to transforming the distributions into standard normal distributions, then proceeding as for the analysis of the Gaussian Mechanism for differential privacy [19]. For completeness, the details are provided in Appendix B.

Intuitively, Theorem 5.4 shows that the query $F(\mathfrak{Data})$ can be safely released long as all the eigenvalues of each covariance matrix Σ_θ are large enough. Thus, privacy can be ensured by adding noise only in the directions of eigenvectors corresponding to small eigenvalues. We formalize this with the Eigenvector Gaussian Mechanism, described in Algorithm 2. The mechanism is presented for the case where the eigenvectors of Σ_θ are the same for each $\theta \in \Theta$, but may be modified to account for cases where they differ slightly, for instance by using approximations of the query function.

Theorem 5.5. *Suppose that $f_\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$ for each $\theta \in \Theta$ and that $\Sigma_{\theta_i} = \Sigma_{\theta_j}$ for each $(\theta_i, \theta_j) \in \Psi$. Suppose furthermore that the normalized eigenvectors $v_1, v_2, \ldots, v_m$ of Σ_θ are the same for each $\theta \in \Theta$. Then, the mechanism $\mathcal{M}$ described in Algorithm 2 satisfies (ϵ, δ)-distribution privacy with respect to Ψ.*

The proof of Theorem 5.5 is in Appendix B.1.

Example: Consider an instantiation of distribution privacy given by $\Theta = \{\theta_1, \theta_2\}$ and $\Psi = \{(\theta_1, \theta_2), (\theta_2, \theta_1)\}$, and suppose that f_{θ_1} and f_{θ_2} are normally distributed with

$$\mu_{\theta_1} = \begin{bmatrix}100\\101\end{bmatrix}, \quad \mu_{\theta_2} = \begin{bmatrix}99\\102\end{bmatrix}, \quad \Sigma_{\theta_1} = \Sigma_{\theta_2} = \begin{bmatrix}22 & -6\\-6 & 13\end{bmatrix}.$$

Suppose that we require $\epsilon = 1$ and $\delta = 0.001$. Then, to ensure (ϵ, δ)-distribution privacy, the Gaussian variant of the Expected Value Mechanism would simply add Gaussian noise with variance $(c\sqrt{2}/\epsilon)^2 \approx 28.52$, where $c = \sqrt{2\ln(1.25/\delta)}$, to each component of $F(\mathfrak{Data})$.

On the other hand, for the Eigenvector Gaussian Mechanism, observe that the eigenvectors of $\Sigma_{\theta_1} = \Sigma_{\theta_2}$ are $v_1 = (1/\sqrt{5}, 2/\sqrt{5})$ and $v_2 = (2/\sqrt{5}, -1/\sqrt{5})$ with the corresponding eigenvalues being 10 and 25. Since $\Delta_{E,2}(\Psi, F) = \sqrt{2}$, the Eigenvector Gaussian Mechanism only needs to increase these eigenvalues to $(c\sqrt{2}/\epsilon)^2$. When $\epsilon = 1$ and $\delta = 0.001$, we have that $(c\sqrt{2}/\epsilon)^2 \approx 28.52$. Hence, the Eigenvector Gaussian Mechanism is equivalent to adding Gaussian noise with variance $\sigma_1^2 \approx 18.52$ in the direction of v_1 and variance $\sigma_2^2 \approx 3.52$ in the direction of v_2, instead of 28.52 for each.

Note that Zhang et al. [44] propose a similar Gaussian Mechanism for attribute privacy which exploits adversarial uncertainty. However they only consider the one-dimensional case, thus, Algorithm 2 may be viewed as a generalization of their mechanism to higher dimensions.

Finally, we remark that it is also possible to exploit inherent randomness while also using the directional assumptions discussed in Section 5.2. Details of the resulting variant of the Expected Value Mechanism are described in Appendix B.2.

5.4 Privacy Guarantees Under Relaxed Assumptions

Our analysis so far establishes a set of privacy guarantees for the Expected Value Mechanism which hold when particular assumptions are imposed on the distributions involved. In particular, we adopted various translation assumptions, directional assumptions, and inherent randomness assumptions. We now examine how these assumptions may be relaxed and analyze the corresponding effects on privacy guarantees.

Consider the scenario in which approximations $\tilde{f}_\theta$ are used in place of the exact distributions f_θ to determine the noise $\widetilde{Z}$ to be added. We are then interested in the privacy guarantees of the approximate mechanism $\widetilde{\mathcal{M}}(\mathfrak{Data}) = F(\mathfrak{Data}) + \widetilde{Z}$, where the query $F(\mathfrak{Data})$ follows a true distribution f_θ for some $\theta \in \Theta$, but $\widetilde{Z}$ is computed based on the approximations $\tilde{f}_\theta$. To this end, we can use max-divergence to quantify the deviation between the approximations $\tilde{f}_\theta$ and the true distributions f_θ.

Definition 5.1 (δ-Approximate Max-Divergence). Let μ and ν be two distributions. The δ-approximate max-divergence $D^\delta_\infty(\mu \parallel \nu)$ between μ and ν is defined as

$$D^\delta_\infty(\mu \parallel \nu) = \sup_{S\subseteq \mathrm{supp}(\mu): \Pr(\mu\in S)\geq\delta} \ln \frac{\Pr(\mu \in S) - \delta}{\Pr(\nu \in S)}.$$

We can then quantify the privacy loss when approximations are used in place of the true query distribution. The following result applies generally for mechanisms $\mathcal{M}$ which apply noise based on the possible distributions f_θ of the query $F(\mathfrak{Data})$, as is the case for all of the mechanisms we have discussed so far.

Theorem 5.6. *Suppose that $\mathcal{M}$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ and that $\mathcal{M}(\mathfrak{Data}) \mid F(\mathfrak{Data})$ is independent of $\mathfrak{Data}$. For each $\theta \in \Theta$, let $\tilde{f}_\theta$ be an approximation of f_θ and suppose that*

$$\max\left(D^\eta_\infty(\tilde{f}_\theta \parallel f_\theta), D^\eta_\infty(f_\theta \parallel \tilde{f}_\theta)\right) \leq \lambda$$

for all $\theta \in \Theta$. Then, the mechanism $\widetilde{\mathcal{M}}$ which uses the approximations $\tilde{f}_\theta$ in place of the true distributions f_θ satisfies (ϵ', δ')-distribution privacy with respect to Ψ, where $\epsilon' = \epsilon + 2\lambda$ and $\delta' = (1 + \exp(\epsilon + \lambda))\eta + \exp(\lambda)\delta$.

The proof of Theorem 5.6 appears in Appendix C.4. Theorem 5.6 is similar to the guarantees against close attackers Song et al. [39] prove for Pufferfish privacy, but with privacy loss quantified in terms of the query distributions f_θ rather than the data distributions θ. That is, the theorem guarantees that $\widetilde{\mathcal{M}}$ achieves reasonable privacy guarantees as long as the approximations $\tilde{f}_\theta$ are close to the true distributions f_θ, as measured by approximate max-divergence. Hence, the Expected Value Mechanism can provide formal guarantees when used in situations where the distributions involved may not be exact translations of each other, but can be approximated by distributions with this property. Similarly, Theorem 5.6 may be applied to the results in Section 5.3 to to use the mechanism for distributions that are only approximately Gaussian.

We can also use ∞-Wasserstein distance to analyze the privacy guarantees offered when approximations are used to apply our mechanisms. In particular, privacy can be ensured by applying a small amount of additional noise to account for possible deviations between the assumed approximations and the true distributions. Let us consider mechanisms $\mathcal{M}$ of the form $\mathcal{M}(\mathfrak{Data}) = F(\mathfrak{Data}) + Z$, where Z is independent of $\mathfrak{Data}$.

Theorem 5.7. *Suppose that mechanism $\mathcal{M}(\mathfrak{Data}) = F(\mathfrak{Data}) + Z$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ. For each $\theta \in \Theta$, let $\tilde{f}_\theta$ be an approximation of f_θ and let*

$$W = \sup_{\theta \in \Theta} W_\infty(f_\theta, \tilde{f}_\theta).$$

Let $Z' = (Z'_1, Z'_2, \ldots, Z'_m)$, where $Z'_k \sim \mathrm{Lap}(W/\lambda)$ for each k, and let $\widetilde{\mathcal{M}}$ be the mechanism which applies $\mathcal{M}$ using the approximations $\tilde{f}_\theta$ in place of the true distributions f_θ. Then the mechanism which outputs $\widetilde{\mathcal{M}}(\mathfrak{Data}) + Z'$ satisfies (ϵ', δ')-distribution privacy with respect to Ψ, where $\epsilon' = \epsilon + 2\lambda$ and $\delta' = \exp(\lambda)\delta$.

The proof of Theorem 5.7 is in Appendix C.4. Intuitively, if the approximations $\tilde{f}_\theta$ are similar to the true distributions f_θ, as measured by ∞-Wasserstein distance, then W will be small. Thus, Theorem 5.7 suggests that the amount of additional Laplace noise Z' required to ensure privacy will be small as long as the approximations are close to the true distributions. Note also that by considering ∞-Wasserstein distance based on the L_2 norm rather than the L_1 norm, we can produce a similar result for mechanisms which use Gaussian noise.

6 EXPERIMENTS

We now empirically evaluate our mechanisms for protecting global properties of datasets, measuring their privacy-utility tradeoffs and how well they defend against property-inference attacks. Specifically, our goal in this section is to address the following questions:

(1) What are the privacy-utility tradeoffs offered by our mechanisms, and how do these compare to baselines from the privacy literature?

Table 1: Query components for Adult dataset.

Attribute	Statistic of Interest
Age	Average
Years of education	Average
Marital status	Num. of never married individuals
Sex	Num. of female individuals
Hours worked per week	Average

(2) Under which scenarios do our noise optimizations for the Expected Value Mechanism produce noticeable improvements to utility in practice?
(3) How effective are our mechanisms in reducing the accuracy of practical property inference attacks?

6.1 Methodology

Dataset. We use the Adult dataset from the UCI Machine Learning repository [15], a well-established dataset which has been used both in early work in privacy [3, 9] as well as more recently to demonstrate property inference attacks [21, 45]. The dataset is a subset of the 1994 US Census database and contains 48 842 records with 14 attributes. For our purposes, we remove records with missing attribute values, leaving a total of 45 222 records. Of these, we randomly allocate 10 000 records as auxiliary data to implement property inference attacks, 10 000 records as testing data to evaluate the accuracy of the attacks, and the remainder for modeling data distributions to implement our mechanisms.

Task & Sensitive properties. We consider a scenario where a data curator wishes to release statistics on a dataset of 100 records (i.e., a subset of records from Adult dataset) on the commonly reported attributes age, years of education, marital status, sex, and hours worked per week. The specific statistics of interest for each attribute are shown in Table 1. The goal is to compute and release approximations of these statistics for this subset of 100 records while protecting sensitive properties of the subset.

We consider two cases for sensitive properties — the proportion p_I of individuals in each subset whose income exceeds \$50K, and the proportion p_W who are private sector workers. Note that these sensitive properties are both global, in the sense that they are aggregate quantities over all records in the given subset rather than properties of individual records.

Modeling and Assumptions. We let each $\theta \in \Theta$ model a possible distribution over subsets of data $\mathfrak{Data}$ given particular values of the sensitive properties p_I and p_W. Then, the set of pairs of distributions $\Psi \subseteq \Theta \times \Theta$ represent pairs of data generation scenarios where the sensitive property value p differs by at most some value Δp.

We assume that each distribution f_θ is approximately multivariate Gaussian with a shared covariance matrix $\Sigma_{\theta_i} = \Sigma_{\theta_j}$ for all pairs $(\theta_i, \theta_j) \in \Psi$. To calculate each Gaussian approximation, we randomly sample subsets of 100 records from the Adult dataset with the global sensitive property values as specified by θ, for each distribution f_θ. We compute the query F over 1000 such subsets for each f_θ and use these to determine a mean vector μ_θ and covariance matrix Σ_θ for the approximation $\mathcal{N}(\mu_\theta, \Sigma_\theta)$. We note that the

Table 2: Mechanisms implemented in the evaluation.

Variant of Expected Value Mechanism	Shorthand
Laplace (Section 5.1)	ExpM (L)
Laplace, Directional (Section 5.2)	DirM (L)
Gaussian (Section 5.1)	ExpM (G)
Gaussian, Eigenvector (Section 5.3)	EigM (G)
Gaussian, Directional with Adversarial Uncertainty (Section 5.3)	DauM (G)

Baseline GroupDP Mechanism	Shorthand
Laplace Mechanism (GroupDP)	GDP (L)
Gaussian Mechanism (GroupDP)	GDP (G)

Table 3: L_2 error incurred by variants of the Expected Value Mechanism and group differential privacy baselines for various values of ϵ. Error values are reported for protecting the income property p_I to an additive factor of $\Delta p_I = 0.1$.

	L_2 Error		
Mechanism	$\epsilon = 0.2$	$\epsilon = 1$	$\epsilon = 5$
ExpM (L)	99.83	21.58	4.13
DirM (L)	35.22	7.24	1.38
ExpM (G)	177.28	34.98	7.11
EigM (G)	175.65	34.87	4.89
DauM (G)	69.85	13.40	1.24
GDP (L)	5528.52	1063.39	213.61
GDP (G)	7394.67	1539.93	293.17

variances of each component are not all exactly the same, contrary to the conditions required for our mechanisms. However, the variances are close enough to apply our mechanisms with the caveat of slightly weaker privacy guarantees as discussed in Section 5.4.

Mechanisms and Baselines. We implement and evaluate five of our proposed variants of the Expected Value Mechanism, summarized in Table 2. Note that the two variants which use Laplace noise are able to guarantee $(\epsilon, 0)$-distribution privacy, while the three variants which use Gaussian noise are only able to guarantee (ϵ, δ)-distribution privacy for $\delta > 0$.

As a baseline for privacy and utility, we compare our mechanisms to the Laplace and Gaussian Mechanisms for group differential privacy [19]. Note that in order for this baseline to ensure distribution privacy for global properties, we need to choose the group size k to be $k = n = 100$, the number of records in the subset we query.

While we do not explicitly evaluate the Wasserstein Mechanism, we remark that under our experimental settings, it reduces to the Expected Value Mechanism. In particular, recall that we model the possible distributions f_θ of the query $F(\mathfrak{Data})$ as being exact translations of each other for every $\mathfrak{Data} \sim \theta \in \Theta$. Thus, our experimental results for the Expected Value Mechanism may be viewed as also indicative of the privacy and utility able to be attained by the Wasserstein Mechanism under our assumed setting.

For each mechanism, we vary parameter ϵ from 0.1 to 10, representing a high and low degrees of privacy, respectively. For the mechanisms which use Gaussian noise, we vary the parameter δ from 0.0001 to 0.01. Where not specified, we take $\delta = 0.001$.

6.2 Privacy-Utility Tradeoffs

We evaluate the privacy-utility tradeoff offered by each mechanism, based on incurred L_2 error. In what follows, we report L_2 error values averaged over 50 repetitions of sampling subsets of the Adult dataset, computing our statistics of interest, then applying each mechanism on these statistics.

Figure 2 shows the privacy-utility tradeoffs of our mechanisms as they are applied to protect the income and work class properties for $\Delta p_I = 0.1$ and $\Delta p_W = 0.04$, respectively. The group differential privacy baselines are not shown in this figure since they incur significantly higher error values than the Expected Value Mechanism. Indeed, as shown in Table 3, the error values incurred by the group differential privacy mechanisms are at least an order of magnitude larger than for the variants of the Expected Value Mechanism.

Since the group differential privacy baselines produce more noise than is reasonable for most practical applications, we also use the Laplace and Gaussian Mechanisms for differential privacy as points of comparison. Note that these mechanisms do not provide theoretical distribution privacy guarantees and apply the same amount of noise regardless of the sensitive global property of interest. Nonetheless, Figure 3 shows the error incurred by Gaussian variants of the Expected Value Mechanisms as Δp varies, with the error incurred by differential privacy mechanisms shown for comparison. We omit Laplace variants as they display similar behavior.

Figure 3a show that the DauM variant of the Expected Value Mechanism incurs less error than differential privacy if the income property only needs to be protected to an additive factor of $\Delta p_I \leq 0.12$. Thus, for instance, the Directional Gaussian Mechanism can prevent an attacker from distinguishing between data with proportion $p_I = 0.44$ of individuals with income > \$50K and data with proportion $p_I = 0.56$ of such individuals, while incurring less noise than the Gaussian Mechanism for differential privacy. Figure 3b show that the work class property can be protected to even larger additive factors Δp_W while incurring less error than the differential privacy mechanisms. We remark that these figures also demonstrate an advantage of our mechanisms over differential privacy mechanisms, in being able to scale noise differently according to the specific sensitive property of interest and degree to which it needs to be protected.

6.3 Property Inference Attack

We now evaluate how well our mechanisms can protect against a property inference attack. We consider the case where an attacker is aiming to determine whether the sensitive property has value $p_1 = 0.5 - \Delta p/2$ or $p_2 = 0.5 + \Delta p/2$, for $\Delta p \in [0.02, 0.4]$. For example, when Δp is 0.2, the attacker is trying to determine if the proportion of people with income more than \$50K is 0.4 or 0.6. Our attack is based on the standard meta-classifier technique [4, 21, 45], using logistic regression as the meta-classifier.

6.3.1 Attack setup. To train the meta-classifier for the attack, we use the auxiliary data to sample 200 shadow datasets, each of size $n = 100$ records with half of the sampled datasets satisfying $p = p_1$

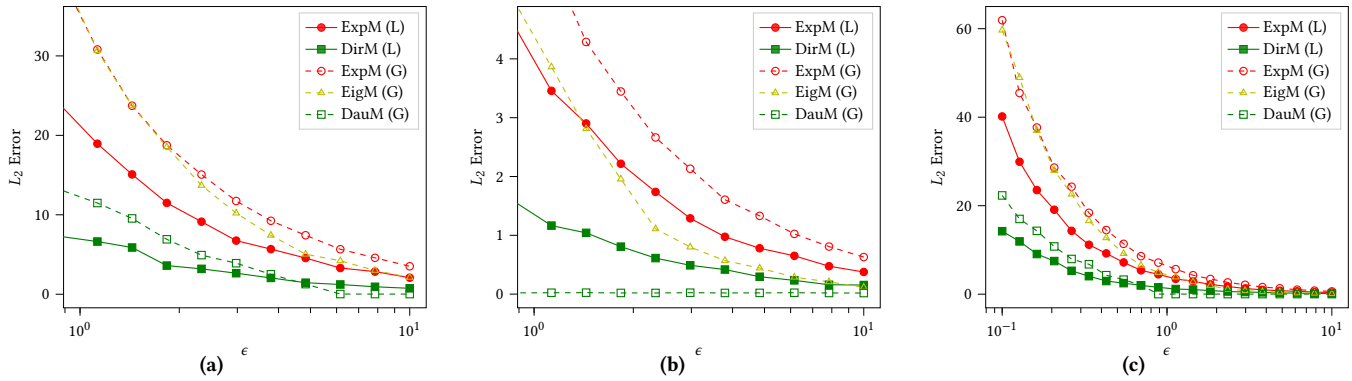
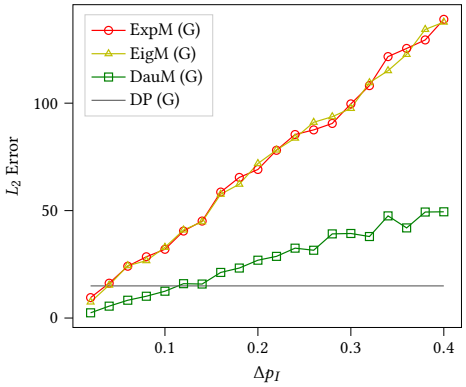
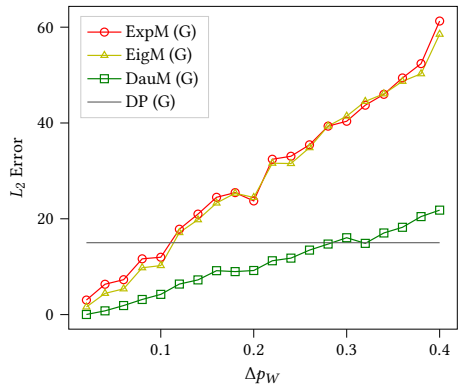


Figure 2: Privacy-utility tradeoffs of variants of our mechanisms, as measured by L_2 error, when (a) protecting income, $\Delta p_I = 0.1$; and protecting work class, $\Delta p_W = 0.04$, with (b) a low privacy ϵ range and (c) a greater range of ϵ (higher privacy level).



(a) Gaussian variants, protecting income.



(b) Gaussian variants, protecting work class.

Figure 3: Error incurred by variants of the Expected Value Mechanism as Δp varies, with $\epsilon = 1$, and $\delta = 0.001$. The error incurred by differential privacy (DP) mechanisms are shown for comparison, however, note that these do not provide distribution privacy guarantees.

and half satisfying $p = p_2$. We then compute the query F to use as training features for each shadow dataset and use the value of the global sensitive property p as labels.

Table 4: Accuracy of the attack on income when Gaussian variants of the Expected Value Mechanism are applied with varying values of ϵ and δ. Here, $\Delta p_I = 0.1$. That is, the attack aims to distinguish between subsets of data satisfying $p_I = 0.45$ and subsets satisfying $p_I = 0.55$.

Mechanism		Attack Accuracy		
		$\delta = 0.0001$	$\delta = 0.001$	$\delta = 0.01$
ExpM (G)	$\epsilon = 0.2$	0.492	0.500	0.502
	$\epsilon = 1$	0.506	0.511	0.520
	$\epsilon = 5$	0.530	0.539	0.549
EigM (G)	$\epsilon = 0.2$	0.507	0.501	0.512
	$\epsilon = 1$	0.510	0.512	0.503
	$\epsilon = 5$	0.537	0.550	0.545
DauM (G)	$\epsilon = 0.2$	0.517	0.508	0.511
	$\epsilon = 1$	0.548	0.545	0.562
	$\epsilon = 5$	0.714	0.739	0.744

To evaluate the accuracy of the trained meta-classifier, we use 200 similarly sampled subsets of size $n = 100$ from the testing data, half satisfying $p = p_1$ and half satisfying $p = p_2$. We repeat the entire process 50 times, with a different set of auxiliary data and testing data each time. The accuracy of the attack is then taken to be the average meta-classifier accuracy over all 50 repetitions.

6.3.2 Attack accuracy. Figure 4a shows the resultant attack accuracy for each sensitive property when no defense mechanisms are applied. The accuracy of the attack is close to 100% when $\Delta p_I = 0.4$ for the attack on income, and close to 90% when $\Delta p_W = 0.4$ for the attack on work class. These results show that global properties can be easily inferred by an attacker if not protected.

On the other hand, Figures 4b and 4c show the accuracy of the attack when Gaussian variants of the Expected Value Mechanism are applied with $\epsilon \in [0.1, 10]$. All of our mechanisms provide strong defense when applied with $\epsilon = 0.1$; every mechanism reduces attack accuracy to near 50% in this case. Table 4 furthermore shows how the accuracy of the attack changes with varying values of ϵ and δ. The Laplace variants demonstrate similar trends.

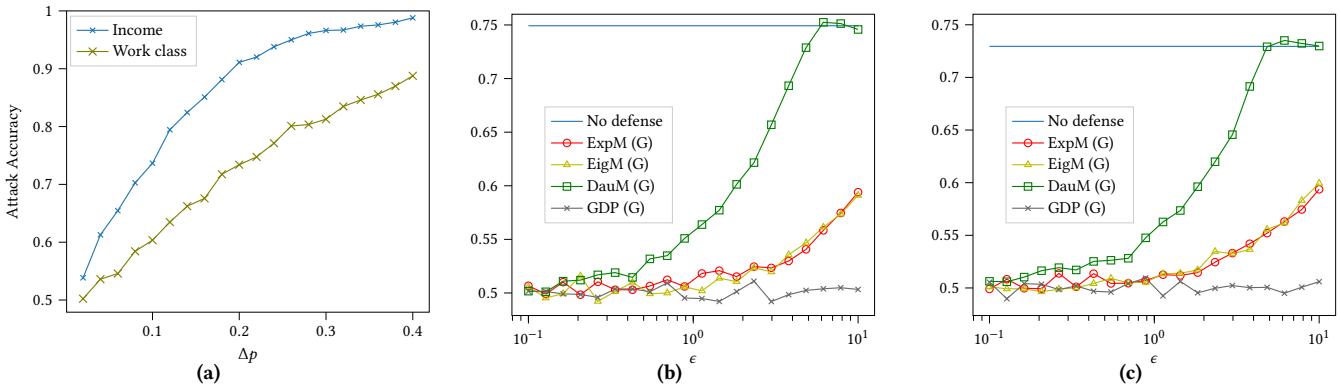


Figure 4: (a) Accuracy of a property inference attack on inferring the proportion p of records with particular sensitive values for income and work class attributes when no defense mechanism is used. The value shown for each Δp is the accuracy of the attack in distinguishing between $p = 0.5 - \Delta p/2$ and $p = 0.5 + \Delta p/2$. Accuracy of the attack when Gaussian variants of the Expected Value Mechanism are applied as defense (b) for income when $\Delta p_I = 0.1$ and (c) for work class when $\Delta p_W = 0.2$. $\delta = 0.001$.

We note that the group differential privacy mechanisms appear to provide the strongest defense out of all of the implemented mechanisms. However, Figure 4 shows that group differential privacy applies enough noise to reduce the attack accuracy to near 50% even with a low degree of privacy with $\epsilon = 10$. Since the group differential privacy mechanisms and the Expected Value Mechanism only differ in how they scale Laplace or Gaussian noise, this suggests that the former are not appropriately calibrated for (ϵ, δ)-distribution privacy, adding more noise than is necessary. This is in contrast to the variants of the Expected Value Mechanism, which demonstrate a trend of high to low defense as ϵ increases.

6.4 Observations

We now address the questions posed at the beginning of Section 6. The privacy-utility tradeoffs presented in Section 6.2 show that our mechanisms provide significantly greater utility than crude group differential privacy baselines. Our experiments suggest that the privacy-utility tradeoffs incurred by our mechanisms are sometimes even comparable to those of standard differential privacy mechanisms. We do note, however, that the precise amount of noise applied by our mechanisms depends greatly on the specific sensitive properties of interest and the degree to which they need to be protected. On the one hand, this shows that our mechanisms have the capacity to be adapted to protect different sensitive properties as necessary. On the other hand, our mechanisms may then require considerable amounts of noise to protect particular sensitive properties, as illustrated for the larger values of Δp in Figure 3.

We also observe that the directional variants of the Expected Value Mechanism consistently incur the least error of our proposed mechanisms. Table 3 in particular demonstrates how these variants can apply comparatively less noise to certain statistics, as opposed to applying an equal amount of noise independently to every statistic. This illustrates the potential for these directional variants to significantly reduce overall error by essentially only applying noise to statistics with high correlation with the sensitive property being protected. Our optimizations based on using inherent noise were generally less noticeable. Figure 2 for instance shows that the Eigenvector Gaussian Mechanism produces noticeable improvements over the Gaussian Variant of the Expected Value Mechanism only for relatively large ϵ. Nonetheless, this optimization may still be useful in situations where the total amount of noise needed to provide privacy is relatively small, allowing for statistics to be released with low or no error if inherent adversarial uncertainty is sufficient to provide privacy.

Finally, the high accuracy of our implemented property inference attacks against unprotected queries reinforces the notion that statistical data analysis tasks can be susceptible to leaking sensitive global properties of datasets. On the other hand, Figure 4 shows that our Expected Value Mechanism variants are indeed able to substantially reduce the attack accuracy, demonstrating the practical implications of these mechanisms' theoretical (ϵ, δ)-distribution privacy guarantees.

7 CONCLUDING REMARKS

We presented an approach to protecting global properties of datasets using the distribution privacy framework and examined its theoretical properties in relation to other established privacy frameworks. We investigated two mechanisms for achieving distribution privacy — the Wasserstein Mechanism, adapted from Pufferfish privacy, and the Expected Value Mechanism, which addresses some of the utility and efficiency limitations of the Wasserstein Mechanism. Our results demonstrated that these mechanisms can substantially reduce the accuracy of a property inference attack while providing significantly better utility than a group differential privacy baseline.

REFERENCES

[1] Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security*. 308–318. https://doi.org/10.1145/2976749.2978318

[2] John M. Abowd. 2018. The U.S. Census Bureau Adopts Differential Privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2867. https://doi.org/10.1145/3219819.3226070

[3] Rakesh Agrawal, Ramakrishnan Srikant, and Dilys Thomas. 2005. Privacy Preserving OLAP. In *Proceedings of the 31st ACM SIGMOD International Conference on Management of Data*. 251–262.

[4] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. 2015. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks* 10, 3 (2015), 137–150. https://doi.org/10.1504/IJSN.2015.071829

[5] Raef Bassily, Adam Groce, Jonathan Katz, and Adam Smith. 2013. Coupled-Worlds Privacy: Exploiting Adversarial Uncertainty in Statistical Data Privacy. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. 439–448. https://doi.org/10.1109/FOCS.2013.54

[6] Raghav Bhaskar, Abhishek Bhowmick, Vipul Goyal, Srivatsan Laxman, and Abhradeep Thakurta. 2011. Noiseless Database Privacy. In *Proceedings of the 17th International Conference on the Theory and Application of Cryptology and Information Security*, Vol. 7073. 215–232.

[7] Anna L. Buczak and Erhan Guven. 2016. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys and Tutorials* 18, 2 (2016), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

[8] Thierry Champion, Luigi De Pascale, and Petri Juutinen. 2008. The ∞-Wasserstein Distance: Local Solutions and Existence of Optimal Transport Maps. *SIAM Journal on Mathematical Analysis* 40, 1 (2008), 1–20.

[9] Graham Cormode. 2011. Personal privacy vs population privacy: Learning to attack anonymization. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1253–1261.

[10] Graham Cormode, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang. 2018. Privacy at Scale: Local Differential Privacy in Practice. In *Proceedings of the 44th ACM SIGMOD International Conference on Management of Data*. 1655–1658. https://doi.org/10.1145/3183713.3197390

[11] Luigi De Pascale and Jean Louet. 2019. A study of the dual problem of the one-dimensional L^∞-optimal optimal transport problem with applications. *Journal of Functional Analysis* 276, 11 (2019), 3304–3324.

[12] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. 2017. Collecting Telemetry Data Privately. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 3571–3580.

[13] Irit Dinur and Kobbi Nissim. 2003. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. 202–210.

[14] Nikunj H. Domadiya and Udai Pratap Rao. 2013. Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database. In *Proceedings of the 3rd IEEE International Advanced Computing Conference*. 1306–1310. https://doi.org/10.1109/IAdCC.2013.6514417

[15] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[16] Cynthia Dwork. 2008. Differential Privacy: A Survey of Results. In *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation*. 1–19. https://doi.org/10.1007/978-3-540-79228-4_1

[17] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In *Proceedings of the 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, Vol. 4004. 486–503.

[18] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography* (New York, NY). 265–284. https://doi.org/10.1007/11681878_14

[19] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3-4 (2014), 211–407. https://doi.org/10.1561/0400000042

[20] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 1322–1333. https://doi.org/10.1145/2810103.2813677

[21] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. 2018. Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations. In *Proceedings of the 25th ACM SIGSAC Conference on Computer and Communications Security*. 619–633. https://doi.org/10.1145/3243734.3243834

[22] Bargav Jayaraman and David Evans. 2019. Evaluating Differentially Private Machine Learning in Practice. In *Proceedings of the 28th USENIX Security Symposium*. 1895–1912.

[23] Andreas Kamilaris, Andreas Kartakoullis, and Francesc X. Prenafeta-Boldu. 2017. A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture* 143 (2017), 23–37.

[24] Yusuke Kawamoto and Takao Murakami. 2019. Local Obfuscation Mechanisms for Hiding Probability Distributions. In *Proceedings of the 24th European Symposium on Research in Computer Security*, Vol. 11735. 128–148.

[25] Daniel Kifer and Bing-Rong Lin. 2012. An Axiomatic View of Statistical Privacy and Utility. *Journal of Privacy and Confidentiality* 4, 1 (2012).

[26] Daniel Kifer and Ashwin Machanavajjhala. 2014. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems* 39, 1 (2014), 3:1–3:36. https://doi.org/10.1145/2514689

[27] Ashwin Machanavajjhala, Johannes Gehrke, and Michaela Götz. 2009. Data Publishing against Realistic Adversaries. *Proceedings of the VLDB Endowment* 2, 1 (2009), 790–801.

[28] S. Mahloujifar, E. Ghosh, and M. Chase. 2022. Property Inference from Poisoning. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 1569–1569. https://doi.org/10.1109/SP46214.2022.00140

[29] Rania Maktabi. 1999. The Lebanese Census of 1932 Revisited. Who Are the Lebanese? *British Journal of Middle Eastern Studies* 26, 2 (1999), 219–241. http://www.jstor.org/stable/195924

[30] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (2021), 115:1–115:35.

[31] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting Unintended Feature Leakage in Collaborative Learning. In *Proceedings of the 40th IEEE Symposium on Security and Privacy*. 691–706. https://doi.org/10.1109/SP.2019.00029

[32] I. Mironov. 2017. Rényi Differential Privacy. In *Proceedings of the 30th IEEE Computer Security Foundations Symposium* (2017-08). 263–275. https://doi.org/10.1109/CSF.2017.11

[33] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2021. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems* 32, 2 (2021), 604–624. https://doi.org/10.1109/TNNLS.2020.2979670

[34] Victor M. Panaretos and Yoav Zemel. 2019. Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application* 6, 1 (2019), 405–431.

[35] Dario Pasquini, Giuseppe Ateniese, and Massimo Bernaschi. 2021. Unleashing the Tiger: Inference Attacks on Split Learning. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2113–2129.

[36] Pierangela Samarati and Latanya Sweeney. 1998. Generalizing Data to Provide Anonymity when Disclosing Information. In *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. 188.

[37] Sagar Sharma, Keke Chen, and Amit P. Sheth. 2018. Toward Practical Privacy-Preserving Analytics for IoT and Cloud-Based Healthcare Systems. *IEEE Internet Computing* 22, 2 (2018), 42–51.

[38] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *Proceedings of the 38th IEEE Symposium on Security and Privacy*. 3–18. https://doi.org/10.1109/SP.2017.41

[39] Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. 2017. Pufferfish Privacy Mechanisms for Correlated Data. In *Proceedings of the 43rd ACM SIGMOD International Conference on Management of Data*. 1291–1306. https://doi.org/10.1145/3035918.3064025

[40] Anshuman Suri and David Evans. 2021. Formalizing and Estimating Distribution Inference Risks. *CoRR* (2021). arXiv:2109.06024 https://arxiv.org/abs/2109.06024

[41] U.S. Department of State. 2019. International Religious Freedom Report: Lebanon. *2019 Report on International Religious Freedom* (2019). "https://www.state.gov/reports/2019-report-on-international-religious-freedom/lebanon/"

[42] Vassilios S. Verykios and Aris Gkoulalas-Divanis. 2008. A Survey of Association Rule Hiding Methods for Privacy. In *Privacy-Preserving Data Mining - Models and Algorithms*. Vol. 34. 267–289.

[43] Lei Yu, Ling Liu, Calton Pu, Mehmet Emre Gursoy, and Stacey Truex. 2019. Differentially Private Model Publishing for Deep Learning. In *Proceedings of the 40th IEEE Symposium on Security and Privacy*. 332–349. https://doi.org/10.1109/SP.2019.00019

[44] Wanrong Zhang, Olga Ohrimenko, and Rachel Cummings. 2022. Attribute Privacy: Framework and Mechanisms. *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)* (2022).

[45] Wanrong Zhang, Shruti Tople, and Olga Ohrimenko. 2021. Leakage of Dataset Properties in Multi-Party Machine Learning. In *Proceedings of the 30th USENIX Security Symposium*. 2687–2704.

A RELATED PRIVACY FRAMEWORKS

A.1 Pufferfish Privacy

The Pufferfish framework [26] is a generalization of differential privacy involving three components — a set of secrets $\mathbb{S}$, a set of secret pairs $\mathbb{S}_{pairs} \subseteq \mathbb{S} \times \mathbb{S}$, and a set of data distributions $\mathbb{D}$. Intuitively, it is a flexible privacy framework which can handle various privacy requirements through appropriate choice of each of these components. Formally, Pufferfish privacy is defined as follows.

Definition A.1 (Pufferfish Privacy). A mechanism $\mathcal{M}$ satisfies (ϵ, δ)-Pufferfish privacy with respect to $(\mathbb{S}, \mathbb{S}_{pairs}, \mathbb{D})$ if for all secret pairs $(s_i, s_j) \in \mathbb{S}_{pairs}$ and all subsets $S \subseteq \text{Range}(\mathcal{M})$,

$$\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid s_i, \theta) \leq \exp(\epsilon) \times \Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid s_j, \theta) + \delta$$

for all distributions $\theta \in \mathbb{D}$ for which $\Pr(s_i \mid \theta) \neq 0$ and $\Pr(s_j \mid \theta) \neq 0$.

Note that the original definition of Pufferfish privacy proposed by Kifer and Machanavajjhala [26] covered only $(\epsilon, 0)$-Pufferfish privacy. However, their definition naturally extends to (ϵ, δ)-Pufferfish privacy as we have described in Definition A.1.

From the definitions of distribution privacy and Pufferfish privacy, we observe that an equivalence between the two can naturally be drawn by mapping each combination of a secret $s \in \mathbb{S}$ and distribution $\theta \in \mathbb{D}$ under the Pufferfish privacy framework to a single distribution $\theta_s \in \Theta$ under the distribution privacy framework.

Theorem 3.1. *Every instantiation of distribution privacy can be equivalently expressed as an instantiation of Pufferfish privacy and every instantiation of Pufferfish privacy can be equivalently expressed as an instantiation of distribution privacy.*

Proof. We first show that every instantiation of distribution privacy can be expressed as an instantiation of Pufferfish privacy. Consider an arbitrary instantiation of distribution privacy, specified by a set of data distributions Θ and a set of pairs of distributions $\Psi \subseteq \Theta \times \Theta$. For each distribution $\theta \in \Theta$, let s_θ be the statement that the true distribution of $\mathfrak{Data}$ is θ, that is, that $\mathfrak{Data} \sim \theta$. For each pair of distributions $(\theta_i, \theta_j) \in \Psi$, let ϕ_{θ_i,θ_j} be the mixture distribution corresponding to the data generation scenario where $\mathfrak{Data}$ is generated by first choosing one of θ_i and θ_j uniformly at random, then sampling from the chosen distribution. Consider the instantiation of Pufferfish privacy specified by secrets $\mathbb{S}$, secret pairs $\mathbb{S}_{pairs}$, and data distributions $\mathbb{D}$ as follows:

$$\begin{aligned}\mathbb{S} &= \{s_\theta : \theta \in \Theta\},\\ \mathbb{S}_{pairs} &= \{(s_{\theta_i}, s_{\theta_j}) : (\theta_i, \theta_j) \in \Psi\},\\ \mathbb{D} &= \{\phi_{\theta_i,\theta_j} : (\theta_i, \theta_j) \in \Psi\}.\end{aligned}$$

For arbitrary mechanisms $\mathcal{M}$, subsets $S \subseteq \text{Range}(\mathcal{M})$, and pairs of distributions $(\theta_i, \theta_j) \in \Psi$,

$$\begin{aligned}&\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_i)\\ &\quad\leq \exp(\epsilon) \times \Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_j) + \delta\\ &\quad\Longleftrightarrow\\ &\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid s_{\theta_i}, \phi_{\theta_i,\theta_j})\\ &\quad\leq \exp(\epsilon) \times \Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid s_{\theta_j}, \phi_{\theta_i,\theta_j}) + \delta.\end{aligned}$$

Notice moreover that $\Pr(s_\theta \mid \phi_{\theta_i,\theta_j}) = 0$ unless $\theta = \theta_i$ or $\theta = \theta_j$. It follows that the described instantiations of distribution privacy and Pufferfish privacy are equivalent.

We now show that every instantiation of Pufferfish privacy can be expressed as an instantiation of distribution privacy. Consider an arbitrary instantiation of Pufferfish privacy, specified by a set of secrets $\mathbb{S}$, a set of secret pairs $\mathbb{S}_{pairs}$, and a set of data distributions $\mathbb{D}$. For each secret $s \in \mathbb{S}$ and distribution $\theta \in \mathbb{D}$ such that $\Pr(s \mid \theta) \neq 0$, let θ_s be the distribution θ conditioned on s. That is, $\Pr(\mathfrak{Data} = D \mid \theta_s) = \Pr(\mathfrak{Data} = D \mid s, \theta)$ for all dataset instances D. Consider the instantiation of distribution privacy specified by data distributions Θ and pairs of data distributions $\Psi \subseteq \Theta \times \Theta$ as follows:

$$\begin{aligned}\Theta &= \{\theta_s : s \in \mathbb{S},\ \theta \in \mathbb{D},\ \Pr(s \mid \theta) \neq 0\},\\ \Psi &= \{(\theta_{s_i}, \theta_{s_j}) : (s_i, s_j) \in \mathbb{S}_{pairs},\ \theta \in \mathbb{D},\ \Pr(s_i \mid \theta)\Pr(s_j \mid \theta) \neq 0\}.\end{aligned}$$

For arbitrary mechanisms $\mathcal{M}$, subsets $S \subseteq \text{Range}(\mathcal{M})$, secret pairs $(s_i, s_j) \in \mathbb{S}_{pairs}$, and distributions $\theta \in \mathbb{D}$ such that $\Pr(s_i \mid \theta) \neq 0$ and $\Pr(s_j \mid \theta) \neq 0$,

$$\begin{aligned}&\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid s_i, \theta) \leq \exp(\epsilon) \times \Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid s_j, \theta) + \delta\\ &\qquad\Longleftrightarrow\\ &\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_{s_i}) \leq \exp(\epsilon) \times \Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_{s_j}) + \delta.\end{aligned}$$

It follows that the described instantiations of Pufferfish privacy and distribution privacy are equivalent. □

A.2 Properties of Distribution Privacy

A.2.1 Privacy Axioms. We first examine two properties regarded by some as fundamental axioms which should be satisfied by all rigorous privacy definitions [25]. These axioms were proposed following initial work on differential privacy and subsequent discussion on properties that should be preserved when relaxing the definition of differential privacy.

The first of these two privacy axioms is known as *transformation invariance* or the *post-processing* property. Intuitively, this property ensures that computations performed on the output of a private mechanism cannot compromise more privacy than the original output. Thus, if the output of a mechanism is safe to release under distribution privacy, results of further analysis on this output are also safe to release.

Proposition A.2 (Post-Processing). *Let $\mathcal{M}_1$ be a mechanism which satisfies (ϵ, δ)-distribution privacy with respect to Ψ. Let $\mathcal{M}_2$ be a randomised mechanism which runs on the output of $\mathcal{M}_1$. Then, the composite mapping $\mathcal{M}_2 \circ \mathcal{M}_1$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ.*

The second privacy axiom is known as *convexity* and ensures that privacy can be ensured if a data curator randomly chooses between several private mechanisms.

Proposition A.3 (Convexity). *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be mechanisms which satisfy (ϵ, δ)-distribution privacy with respect to Ψ. Then, the mechanism $\mathcal{M}^p$ which runs $\mathcal{M}_1$ with probability p and $\mathcal{M}_2$ with probability $1 - p$ also satisfies (ϵ, δ)-distribution privacy with respect to Ψ.*

Both the post-processing and convexity properties were established for $(\epsilon, 0)$-Pufferfish privacy by Kifer and Machanavajjhala [26], and their proofs naturally extend to the case of (ϵ, δ)-Pufferfish privacy. Thus, from the equivalence we established between Pufferfish privacy and distribution privacy, it follows that (ϵ, δ)-distribution privacy satisfies both properties as well.

A.2.2 Composition. Along with the post-processing and convexity properties discussed in the previous section, it is also desirable for privacy definitions to satisfy certain *composition* properties. Informally, these concern how privacy guarantees may degrade when the outputs of multiple private mechanisms are combined together. Differential privacy, for instance, composes well, in that

combining the output of a (ϵ_1, δ_1)-differentially private mechanism with that of a (ϵ_2, δ_2)-differentially private mechanism produces a result which can effectively be regarded as the output of a $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-differentially private mechanism [17]. This allows cumulative privacy loss over separate data releases to be controlled and furthermore enables complex mechanisms to be designed and analyzed in a modular manner.

Unfortunately, Pufferfish privacy and distribution privacy do not always compose well [26]. Nonetheless, there are various conditions under which Pufferfish privacy and distribution privacy have been shown to have useful composition properties. Kifer and Machanavajjhala [26] identify one set of sufficient conditions for $(\epsilon, 0)$-Pufferfish privacy, intuitively corresponding to the scenario of an extremely knowledgeable attacker who believes that, for each sensitive property of concern, there is only one possible dataset for which the sensitive property holds. Interpreted in terms of the distribution privacy framework, these conditions correspond to the scenario where each distribution $\theta \in \Theta$ represents a single dataset instance. This essentially reduces the definition of distribution privacy to that of differential privacy, as the probabilities in Definition 3.2 no longer consider randomness in the true dataset $\mathfrak{Data}$. Composition of (ϵ, δ)-distribution privacy in this case therefore follows from the corresponding properties of differential privacy.

Proposition A.4. *Suppose Θ is a set of distributions such that for all $\theta \in \Theta$, there exists a dataset instance D for which $\Pr(\mathfrak{Data} = D \mid \theta) = 1$. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be mechanisms which satisfy (ϵ_1, δ_1)-distribution privacy and (ϵ_2, δ_2)-distribution privacy, respectively, with respect to $\Psi \subseteq \Theta \times \Theta$. Then, the mechanism $\mathcal{M}_{1,2}$ defined by $\mathcal{M}_{1,2}(\mathfrak{Data}) = (\mathcal{M}_1(\mathfrak{Data}), \mathcal{M}_2(\mathfrak{Data}))$ satisfies $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-distribution privacy with respect to Ψ.*

A related notion of composition concerns the combination of outputs from mechanisms computed on different datasets, drawn independently from possibly different distributions. This can be viewed as a generalisation of the previous case, where each distribution corresponded to a single dataset instance and datasets drawn from the same distribution would thus take on the same value. Kawamoto and Murakami [24] prove that distribution privacy composes well under this notion of independent sampling. More precisely, for distributions θ and θ', let $\theta \times \theta'$ denote the distribution of the pair $(\mathfrak{Data}, \mathfrak{Data}')$, where $\mathfrak{Data} \sim \theta$ and $\mathfrak{Data}' \sim \theta'$ are independently sampled. Furthermore, for a set of pairs of distributions Ψ, define $\Psi \diamond \Psi$ by

$$\Psi \diamond \Psi = \{(\theta_i \times \theta'_i, \theta_j \times \theta'_j) : (\theta_i, \theta_j), (\theta'_i, \theta'_j) \in \Psi\}.$$

Then, Kawamoto and Murakami [24] prove the following.

Proposition A.5. *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be mechanisms which satisfy (ϵ_1, δ_1)-distribution privacy and (ϵ_2, δ_2)-distribution privacy, respectively, with respect to Ψ. Let $\mathfrak{Data}$ and $\mathfrak{Data}'$ be datasets that are sampled independently. Then, the mechanism $\mathcal{M}_{1,2}$ defined by $\mathcal{M}_{1,2}(\mathfrak{Data}, \mathfrak{Data}') = (\mathcal{M}_1(\mathfrak{Data}), \mathcal{M}_2(\mathfrak{Data}'))$ satisfies $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-distribution privacy with respect to $\Psi \diamond \Psi$.*

Various other composition properties are known to hold for specific Pufferfish privacy and distribution privacy mechanisms [39] as well as for instantiations of distribution privacy where more is known about the pairs of distributions in Ψ [24]. As such, while distribution privacy may not compose well in general, there may nonetheless be composition properties that apply when distribution privacy is used in practice.

A.2.3 Guarantees Against Close Attackers. To conclude our discussion on properties of distribution privacy, we briefly examine the privacy guarantees that hold when an attacker's beliefs are not reflected exactly by the set of distributions Θ. This applies, for instance, when the true distribution of $\mathfrak{Data}$ is difficult to model exactly, or when an attacker is able to estimate the true distribution of $\mathfrak{Data}$ more precisely than anticipated. Concretely, consider that attacker may wish to determine whether $\mathfrak{Data}$ was sampled from $\tilde{\theta}_i$ or $\tilde{\theta}_j$, where $\tilde{\theta}_i \notin \Theta$ and $\tilde{\theta}_j \notin \Theta$. Then, distribution privacy can still be upheld as long as there is a pair $(\theta_i, \theta_j) \in \Psi$ such that θ_i is close to $\tilde{\theta}_i$ and θ_j is close to $\tilde{\theta}_j$, as quantified by a measure called max-divergence.

Definition A.6 (Max-Divergence). Let μ and ν be two distributions with the same support. The max-divergence $D_\infty(\mu \parallel \nu)$ between μ and ν is defined as

$$D_\infty(\mu \parallel \nu) = \sup_{S \subseteq \mathrm{supp}(\mu)} \ln \frac{\Pr(\mu \in S)}{\Pr(\nu \in S)}.$$

Proposition A.7. *Let $\mathcal{M}$ be a mechanism which satisfies (ϵ, δ)-distribution privacy with respect to $\Psi \subseteq \Theta \times \Theta$ and suppose that Θ is a closed set. Then, for all distributions $\tilde{\theta}_i$ and $\tilde{\theta}_j$ and all subsets $S \subseteq \mathrm{Range}(\mathcal{M})$*

$$\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \tilde{\theta}_i) \leq \exp(\epsilon + 2c) \times \Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \tilde{\theta}_j) + \delta$$

where

$$c = \inf_{(\theta_i, \theta_j) \in \Psi} \max \left(D_\infty(\tilde{\theta}_i \parallel \theta_i), D_\infty(\theta_j \parallel \tilde{\theta}_j) \right).$$

The proof of Proposition A.7 can be adapted from Song et al. [39], who prove an analogous guarantee against close attackers for Pufferfish privacy.

Intuitively, this result for close attackers ensures that distribution privacy can be maintained with only a small loss of privacy when data assumptions associated with Θ and Ψ do not hold exactly. For instance, in practical applications of distribution privacy, approximations may be used in place of the true distributions if they are unknown or inconvenient to work with. Proposition A.7 then ensures that privacy will largely be maintained as long as the approximations are close to the true distributions, as measured by max-divergence.

B ADVERSARIAL UNCERTAINTY RESULTS

Theorem 5.4. *Suppose that $f_\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$ for each $\theta \in \Theta$ and that $\Sigma_{\theta_i} = \Sigma_{\theta_j}$ for each $(\theta_i, \theta_j) \in \Psi$. Let $c = \sqrt{2 \ln(1.25/\delta)}$. Then, the mechanism $\mathcal{M}$ which simply outputs $F(\mathfrak{Data})$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ as long as*

$$(\mu_{\theta_i} - \mu_{\theta_j})^\mathrm{T} \Sigma_{\theta_i}^{-1} (\mu_{\theta_i} - \mu_{\theta_j}) \leq (\epsilon/c)^2$$

for all $(\theta_i, \theta_j) \in \Psi$.

Proof. Since $f_\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$ for each $\theta \in \Theta$, we can write each possible distribution of the query function in the form $F(\mathfrak{Data}) \mid \theta = A_\theta Z + \mu_\theta$, where Z is a standard normal vector and A_θ is an

arbitrary matrix satisfying $A_\theta A_\theta^{\mathrm{T}} = \Sigma_\theta$. Then, for each $\theta \in \Theta$ and $x \in \mathbb{R}^m$,

$$\begin{aligned}\Pr(F(\mathfrak{Data}) = x \mid \theta) &= \Pr(A_\theta Z + \mu_\theta = x) \\ &= \frac{1}{|A_\theta|} \Pr(Z = z_\theta),\end{aligned}$$

where $z_\theta = A_\theta^{-1}(x - \mu_\theta)$.

Let $(\theta_i, \theta_j) \in \Psi$ be a pair of distributions. By assumption, we have $\Sigma_{\theta_i} = \Sigma_{\theta_j}$, thus, we may choose $A_{\theta_i} = A_{\theta_j}$ to be the same matrix. Then,

$$\begin{aligned}\frac{\Pr(F(\mathfrak{Data}) = x \mid \theta_i)}{\Pr(F(\mathfrak{Data}) = x \mid \theta_j)} &= \frac{\frac{1}{|A_{\theta_i}|} \Pr(Z = z_{\theta_i})}{\frac{1}{|A_{\theta_j}|} \Pr(Z = z_{\theta_j})} \\ &= \frac{\Pr(Z = z_{\theta_i})}{\Pr(Z = z_{\theta_i} + (z_{\theta_j} - z_{\theta_i}))}.\end{aligned}$$

Now, a standard normal distribution is spherically symmetric, so the distribution of Z is independent of the orthonormal basis from which its components are drawn. Let us fix an orthonormal basis $b_1, b_2, \ldots, b_m$ such that b_1 is parallel to $z_{\theta_j} - z_{\theta_i}$. We may then express Z in the form

$$Z = \sum_{k=1}^{m} \lambda_k b_k,$$

where $\lambda_k \sim \mathcal{N}(0, 1)$ for each k are independently drawn. Writing $z_{\theta_i} = \sum_{k=1}^m \alpha_k b_k$, we then have

$$\begin{aligned}&\frac{\Pr\left(Z = z_{\theta_i}\right)}{\Pr\left(Z = z_{\theta_i} + (z_{\theta_j} - z_{\theta_i})\right)} \\ &= \frac{\Pr\left(\lambda_1 = \alpha_1\right)}{\Pr\left(\lambda_1 = \alpha_1 + \|z_{\theta_j} - z_{\theta_i}\|_2\right)} \prod_{k=2}^{m} \frac{\Pr\left(\lambda_k = \alpha_k\right)}{\Pr\left(\lambda_k = \alpha_k\right)} \\ &= \exp\left(-\frac{1}{2}\left(\alpha_1^2 - \left(\alpha_1 + \|z_{\theta_j} - z_{\theta_i}\|_2\right)^2\right)\right) \\ &= \exp\left(\alpha_1 \|z_{\theta_j} - z_{\theta_i}\|_2 + \frac{1}{2}\|z_{\theta_j} - z_{\theta_i}\|_2^2\right).\end{aligned}$$

The conditions of the theorem moreover give

$$\begin{aligned}&(\mu_{\theta_i} - \mu_{\theta_j})^{\mathrm{T}} \Sigma_{\theta_i}^{-1} (\mu_{\theta_i} - \mu_{\theta_j}) \le (\epsilon/c)^2 \\ \implies &(\mu_{\theta_i} - \mu_{\theta_j})^{\mathrm{T}} (A_{\theta_i} A_{\theta_i}^{\mathrm{T}})^{-1} (\mu_{\theta_i} - \mu_{\theta_j}) \le (\epsilon/c)^2 \\ \implies &\|A_{\theta_i}^{-1}(\mu_{\theta_i} - \mu_{\theta_j})\|_2 \le \epsilon/c,\end{aligned}$$

thus, $\|z_{\theta_j} - z_{\theta_i}\|_2 = \|A_{\theta_i}^{-1}(\mu_{\theta_j} - \mu_{\theta_i})\|_2 \le \epsilon/c$. It follows that

$$\frac{\Pr\left(Z = z_{\theta_i}\right)}{\Pr\left(Z = z_{\theta_i} + (z_{\theta_j} - z_{\theta_i})\right)} \le \exp\left(\frac{|\alpha_1|\epsilon}{c} + \frac{\epsilon^2}{2c^2}\right).$$

The above quantity is bounded by $\exp(\epsilon)$ whenever $|\alpha_1| \le c - \epsilon/2c$. We now show that this occurs with probability at least $1 - \delta$.

Since $\alpha_1 \sim \mathcal{N}(0, 1)$, the standard Gaussian tail bound gives

$$\Pr\left(|\alpha_1| > t\right) < \frac{\sqrt{2}\exp(-t^2/2)}{t\sqrt{\pi}}$$

for $t > 0$. Letting $t = c - \epsilon/2c$, we would like to ensure that $\Pr(|\alpha_1| > t) < \delta$. It suffices to have

$$\begin{aligned}&\frac{\sqrt{2}\exp(-t^2/2)}{t\sqrt{\pi}} < \delta \\ \iff &t\exp\left(\frac{t^2}{2}\right) > \frac{\sqrt{2}}{\sqrt{\pi}\delta} \\ \iff &\ln(t) + \frac{t^2}{2} > \ln\left(\frac{\sqrt{2}}{\sqrt{\pi}\delta}\right).\end{aligned}$$

We may assume that $\epsilon \le 1$ and $c = \sqrt{2\ln(1.25/\delta)} \ge 3/2$, so that the first term can be bounded by $\ln(t) = \ln(c - \epsilon/2c) \ge \ln(7/6)$. For the second term, $\epsilon \le 1$ implies that

$$\frac{t^2}{2} = \frac{1}{2}\left(c - \frac{\epsilon}{2c}\right)^2 = \frac{1}{2}\left(c^2 - \epsilon + \frac{\epsilon^2}{4c^2}\right) \ge \frac{c^2 - 1}{2}.$$

Thus, it suffices to have

$$\begin{aligned}&\ln\left(\frac{7}{6}\right) + \frac{c^2 - 1}{2} \ge \ln\left(\frac{\sqrt{2}}{\sqrt{\pi}\delta}\right) \\ \iff &c^2 \ge 2\ln\left(\frac{6\sqrt{2}\exp(1/2)}{7\sqrt{\pi}\delta}\right)\end{aligned}$$

which holds for $c = \sqrt{2\ln(1.25/\delta)}$ since $6\sqrt{2}\exp(1/2)/7\sqrt{\pi} < 1.25$. Thus, $\Pr(|\alpha_1| > c - \epsilon/2c) < \delta$.

To conclude the proof, write each $x \in \mathbb{R}^m$ in the form $x = A_{\theta_i}\sum_{k=1}^m \alpha_k b_k + \mu_{\theta_i}$ and partition $\mathbb{R}^m$ as $\mathbb{R}^m = R_1 \cup R_2$, where

$$\begin{aligned}R_1 &= \{x : |\alpha_1| \le c - \epsilon/2c\} \\ R_2 &= \{x : |\alpha_1| > c - \epsilon/2c\}.\end{aligned}$$

Then, $\Pr(F(\mathfrak{Data}) = x \mid \theta_i) \le \exp(\epsilon)\Pr(F(\mathfrak{Data}) = x \mid \theta_j)$ for all $x \in R_1$, and $\Pr(F(\mathfrak{Data}) \in R_2 \mid \theta_i) < \delta$. Fix an arbitrary subset $S \subseteq \mathbb{R}^m$, and define $S_1 = S \cap R_1$ and $S_2 = S \cap R_2$. We have

$$\begin{aligned}&\Pr(F(\mathfrak{Data}) \in S \mid \theta_i) \\ &= \Pr(F(\mathfrak{Data}) \in S_1 \mid \theta_i) + \Pr(F(\mathfrak{Data}) \in S_2 \mid \theta_i) \\ &\le \exp(\epsilon)\Pr(F(\mathfrak{Data}) \in S_1 \mid \theta_j) + \delta \\ &\le \exp(\epsilon)\Pr(F(\mathfrak{Data}) \in S \mid \theta_j) + \delta.\end{aligned}$$

If follows that $\mathcal{M}$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ. □

Theorem 5.4 gives a means of estimating the amount of privacy that can be guaranteed only using randomness inherent in the distributions f_θ for each $\theta \in \Theta$. However, additional noise may be needed to ensure a desirable level of privacy, especially in cases where there may not be much randomness inherent in the query. In such cases, we can add Gaussian noise to manipulate the covariance matrices Σ_θ to satisfy the conditions required by the theorem.

Corollary B.1. *Suppose that $f_\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$ for each $\theta \in \Theta$ and that $\Sigma_{\theta_i} = \Sigma_{\theta_j}$ for each $(\theta_i, \theta_j) \in \Psi$. Let $c = \sqrt{2\ln(1.25/\delta)}$. Then, the mechanism $\mathcal{M}$ which outputs $F(\mathfrak{Data}) + Z$, where $Z \sim \mathcal{N}(0, \Sigma)$, satisfies (ϵ, δ)-distribution privacy with respect to Ψ as long as*

$$(\mu_{\theta_i} - \mu_{\theta_j})^{\mathrm{T}} (\Sigma_{\theta_i} + \Sigma)^{-1} (\mu_{\theta_i} - \mu_{\theta_j}) \le (\epsilon/c)^2$$

for all $(\theta_i, \theta_j) \in \Psi$.

Proof. Since $f_\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$ and $Z \sim \mathcal{N}(0, \Sigma)$ are independent, we have that $F(\mathfrak{Data}) + Z \sim \mathcal{N}(\mu_\theta, \Sigma_\theta + \Sigma)$ for each $\mathfrak{Data} \sim \theta \in \Theta$. The modified query $F(\mathfrak{Data}) + Z$ then satisfies the conditions of Theorem 5.4 and the result follows. □

Alternatively, we can use the following set of sufficient conditions expressed in terms of the worst case L_2 distance between expected values, $\Delta_{E,2}(\Psi, F)$, defined in Section 5.1.

Corollary B.2. *Suppose that $f_\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$ for each $\theta \in \Theta$ and that $\Sigma_{\theta_i} = \Sigma_{\theta_j}$ for each $(\theta_i, \theta_j) \in \Psi$. Let $c = \sqrt{2\ln(1.25/\delta)}$. Then, the mechanism $\mathcal{M}$ which outputs $F(\mathfrak{Data}) + Z$, where $Z \sim \mathcal{N}(0, \Sigma)$, satisfies (ϵ, δ)-distribution privacy with respect to Ψ as long as*

$$\Sigma_\theta + \Sigma \geq (c\Delta_{E,2}(\Psi, F)/\epsilon)^2 I$$

for all $\theta \in \Theta$. Equivalently, $\mathcal{M}$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ as long as the minimum eigenvalue of $\Sigma_\theta + \Sigma$ is at least $(c\Delta_{E,2}(\Psi, F)/\epsilon)^2$ for all $\theta \in \Theta$.

Proof. We may assume that $\Delta_{E,2}(\Psi, F) > 0$, so that we have $(c\Delta_{E,2}(\Psi, F)/\epsilon)^2 I > 0$. Then, using properties of positive definite matrices, we have

$$\begin{aligned} &\Sigma_\theta + \Sigma \geq (c\Delta_{E,2}(\Psi, F)/\epsilon)^2 I \\ \implies &(\Sigma_\theta + \Sigma)^{-1} \leq (\epsilon/c\Delta_{E,2}(\Psi, F))^2 I. \end{aligned}$$

Let $(\theta_i, \theta_j) \in \Psi$. By definition of $\Delta_{E,2}(\Psi, F)$, we have $\|\mu_{\theta_i} - \mu_{\theta_j}\|_2 \leq \Delta_{E,2}(\Psi, F)$. Thus,

$$\begin{aligned} &(\mu_{\theta_i} - \mu_{\theta_j})^{\mathrm{T}} (\Sigma_{\theta_i} + \Sigma)^{-1} (\mu_{\theta_i} - \mu_{\theta_j}) \\ &\leq (\mu_{\theta_i} - \mu_{\theta_j})^{\mathrm{T}} ((\epsilon/c\Delta_{E,2}(\Psi, F))^2 I)(\mu_{\theta_i} - \mu_{\theta_j}) \\ &= (\epsilon/c\Delta_{E,2}(\Psi, F))^2 \|\mu_{\theta_i} - \mu_{\theta_j}\|_2^2 \\ &\leq (\epsilon/c)^2. \end{aligned}$$

The conditions of Corollary B.1 are thus satisfied, and it follows that $\mathcal{M}$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ. □

B.1 Eigenvector Gaussian Mechanism

Theorem 5.5. *Suppose that $f_\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$ for each $\theta \in \Theta$ and that $\Sigma_{\theta_i} = \Sigma_{\theta_j}$ for each $(\theta_i, \theta_j) \in \Psi$. Suppose furthermore that the normalized eigenvectors $v_1, v_2, \ldots, v_m$ of Σ_θ are the same for each $\theta \in \Theta$. Then, the mechanism $\mathcal{M}$ described in Algorithm 2 satisfies (ϵ, δ)-distribution privacy with respect to Ψ.*

Proof. Let $\theta \in \Theta$. By assumption, $v_1, v_2, \ldots, v_m$ are the normalized eigenvectors of Σ_θ, and since Σ_θ is symmetric, these eigenvectors are orthogonal. We may thus write $\Sigma_\theta = \sum_k \lambda_{\theta,k}^2 v_k v_k^{\mathrm{T}}$, where $\lambda_{\theta,k}^2 = v_k^{\mathrm{T}} \Sigma_\theta v_k$ is the eigenvalue of Σ_θ corresponding to the eigenvector v_k. Then, we have

$$\begin{aligned} \Sigma_\theta + \Sigma &= \sum_k (\lambda_{\theta,k}^2 + \sigma_k^2) v_k v_k^{\mathrm{T}} \\ \implies (\Sigma_\theta + \Sigma) v_r &= \sum_k (\lambda_{\theta,k}^2 + \sigma_k^2) v_k v_k^{\mathrm{T}} v_r \\ &= (\lambda_{\theta,r}^2 + \sigma_r^2) v_r, \end{aligned}$$

for each $1 \leq r \leq m$. The eigenvalues of $\Sigma_\theta + \Sigma$ are thus exactly the values $\lambda_{\theta,k}^2 + \sigma_k^2$. Now, from our choice of σ_k,

$$\lambda_{\theta,k}^2 + \sigma_k^2 \geq \lambda_{\theta,k}^2 + \sigma_{\theta,k}^2 \geq (c\Delta_{E,2}(\Psi, F)/\epsilon)^2,$$

so that the minimum eigenvalue of $\Sigma_\theta + \Sigma$ is at least $(c\Delta_{E,2}(\Psi, F)/\epsilon)^2$. The conditions of Corollary B.2 are thus satisfied, and it follows that $\mathcal{M}$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ. □

B.2 Adversarial Uncertainty with Directional Assumptions

We briefly discuss how directional assumptions (Section 5.2) can be used in conjunction with inherent randomness (Section 5.3). Essentially, we apply noise in the direction of $\mu_{\theta_i} - \mu_{\theta_j}$ while also taking into account the variance of each distribution f_θ. In doing so, we produce a variant of the Expected Value Mechanism which uses directional and adversarial uncertainty assumptions (DauM), described in Algorithm 3. We present the mechanism for the case where the differences of mean vectors $\mu_{\theta_i} - \mu_{\theta_j}$ are all parallel to some vector v. As for the Directional Expected Value Mechanism (Section 5.2) and Eigenvector Gaussian Mechanism (Section 5.3), this mechanism can also be modified to account for cases where this assumption may not exactly hold.

Algorithm 3: DauM (dataset $\mathfrak{Data}$, query F, distribution pairs $\Psi \subseteq \Theta \times \Theta$, privacy parameters ϵ, δ, unit vector v)

```
Set c = sqrt(2 ln(1.25/δ)).
for each θ ∈ Θ do
    Set μ_θ = E[F(Data) | Data ~ θ],
      Σ_θ = Cov(F(Data) | Data ~ θ).
end
for each (θ_i, θ_j) ∈ Ψ do
    Set α_{i,j} = (μ_{θ_i} − μ_{θ_j})^T v.
    Find σ²_{i,j} > 0 such that Σ_{θ_i} + (σ²_{i,j} − (α_{i,j} c/ε)²) v v^T > 0.
end
Set σ² = max_{(θ_i,θ_j)∈Ψ} σ²_{i,j}.
return F(Data) + Z, where Z ~ N(0, σ² v v^T).
```

To prove that Algorithm 3 satisfies distribution privacy, we will use the following lemma.

Lemma B.3. *Let v be a vector and let $\Sigma > 0$. Suppose that $\Sigma v = \lambda v$. Then, $v^{\mathrm{T}} \Sigma^{-1} v = \lambda^{-1} \|v\|_2^2$.*

Proof. We have

$$\Sigma v = \lambda v \implies \Sigma^{-1} v = \lambda^{-1} v \implies v^T \Sigma^{-1} v = \lambda^{-1} v^{\mathrm{T}} v = \lambda^{-1} \|v\|_2^2.$$

□

Theorem B.4. *Suppose that $f_\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$ for each $\theta \in \Theta$ and that $\Sigma_{\theta_i} = \Sigma_{\theta_j}$ for each $(\theta_i, \theta_j) \in \Psi$. Suppose furthermore that, for each $(\theta_i, \theta_j) \in \Psi$, the vector $\mu_{\theta_i} - \mu_{\theta_j}$ is parallel to the unit vector v. Then, the mechanism $\mathcal{M}$ described in Algorithm 3 satisfies (ϵ, δ)-distribution privacy with respect to Ψ.*

Proof. Let $(\theta_i, \theta_j) \in \Psi$. By Corollary B.1, it suffices to show that

$$(\mu_{\theta_i} - \mu_{\theta_j})^{\mathrm{T}}(\Sigma_{\theta_i} + \sigma^2 vv^{\mathrm{T}})^{-1}(\mu_{\theta_i} - \mu_{\theta_j}) \leq (\epsilon/c)^2. \tag{5}$$

By assumption, $\mu_{\theta_i} - \mu_{\theta_j}$ is parallel to v, hence $\mu_{\theta_i} - \mu_{\theta_j} = \alpha_{i,j} v$, where $\alpha_{i,j} = (\mu_{\theta_i} - \mu_{\theta_j})^{\mathrm{T}} v$. Then,

$$\begin{aligned}(\mu_{\theta_i} - \mu_{\theta_j})^{\mathrm{T}}(\Sigma_{\theta_i} + \sigma^2 vv^{\mathrm{T}})^{-1}(\mu_{\theta_i} - \mu_{\theta_j}) \\ = \alpha_{i,j}^2 v^{\mathrm{T}}(\Sigma_{\theta_i} + \sigma^2 vv^{\mathrm{T}})^{-1} v.\end{aligned}$$

We would thus like to show that $v^{\mathrm{T}}(\Sigma_{\theta_i} + \sigma^2 vv^{\mathrm{T}})^{-1} v \leq (\epsilon/c\alpha_{i,j})^2$. Now, from our choice of σ^2 and using properties of positive semi-definite matrices, we have

$$\Sigma_{\theta_i} + \sigma^2 vv^{\mathrm{T}} \geq \Sigma_{\theta_i} + \sigma_{i,j}^2 vv^{\mathrm{T}} > (\alpha_{i,j} c/\epsilon)^2 vv^{\mathrm{T}}.$$

Thus, there exists some $\eta > 0$ such that $\Sigma_{\theta_i} + \sigma^2 vv^{\mathrm{T}} - (\alpha_{i,j} c/\epsilon)^2 vv^{\mathrm{T}} > \eta I$. We then have

$$\begin{aligned}&\Sigma_{\theta_i} + \sigma^2 vv^{\mathrm{T}} > (\alpha_{i,j} c/\epsilon)^2 vv^{\mathrm{T}} + \eta I \\ \implies &(\Sigma_{\theta_i} + \sigma^2 vv^{\mathrm{T}})^{-1} < ((\alpha_{i,j} c/\epsilon)^2 vv^{\mathrm{T}} + \eta I)^{-1}\end{aligned}$$

Now, since $((\alpha_{i,j} c/\epsilon)^2 vv^{\mathrm{T}} + \eta I)v = ((\alpha_{i,j} c/\epsilon)^2 + \eta)v$, Lemma B.3 implies

$$v^{\mathrm{T}}((\alpha_{i,j} c/\epsilon)^2 vv^{\mathrm{T}} + \eta I)^{-1} v = ((\alpha_{i,j} c/\epsilon)^2 + \eta)^{-1} < (\epsilon/c\alpha_{i,j})^2.$$

Thus, $v^{\mathrm{T}}(\Sigma_{\theta_i} + \sigma^2 vv^{\mathrm{T}})^{-1} v < v^{\mathrm{T}}((\alpha_{i,j} c/\epsilon)^2 vv^{\mathrm{T}} + \eta I)^{-1} v < (\epsilon/c\alpha_{i,j})^2$, and (5) holds. The conditions of Corollary B.1 are thus satisfied, and it follows that $\mathcal{M}$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ. □

Note that the condition on line 7 of Algorithm 3 can always be satisfied by setting $\sigma_{i,j}^2 = (\alpha_{i,j} c/\epsilon)^2$. However, doing so essentially disregards Σ_{θ_i} and reduces Algorithm 3 to the Gaussian variant of the Directional Expected Value Mechanism described in Section 5.2. More generally a suitable and possibly smaller value of $\sigma_{i,j}^2$ can be found by solving $\det\left(\Sigma_{\theta_i} + (\sigma_{i,j}^2 - (\alpha_{i,j} c/\epsilon)^2) vv^{\mathrm{T}}\right) = 0$ then increasing $\sigma_{i,j}^2$ slightly to ensure that $\Sigma_{\theta_i} + (\sigma_{i,j}^2 - (\alpha_{i,j} c/\epsilon)^2) vv^{\mathrm{T}}$ is strictly positive definite. The noise reduction provided by Algorithm 3 in comparison to the Directional Expected Value Mechanism is then proportional to $(\alpha_{i,j} c/\epsilon)^2 - \sigma_{i,j}^2$.

C ADDITIONAL PROOFS

C.1 Approximate Wasserstein Mechanism

Proposition 4.1. *Suppose that $\|F(\mathfrak{Data}) - \mathbb{E}[F(\mathfrak{Data})]\|_1 \leq c$ with probability at least $1 - \delta/2$ for each $\mathfrak{Data} \sim \theta \in \Theta$. Then, the distributions f_{θ_i} and f_{θ_j} are $(\Delta_E(\Psi, F) + 2c, \delta)$-close for all pairs $(\theta_i, \theta_j) \in \Psi$.*

Proof. For each $\theta \in \Theta$, let $R_\theta = \{t : \|t - \mathbb{E}[f_\theta]\|_1 \leq c\}$. Then, $\Pr(F(\mathfrak{Data}) \in R_\theta \mid \theta) \geq 1 - \delta/2$ for each $\theta \in \Theta$.

Fix a pair $(\theta_i, \theta_j) \in \Psi$, and consider the distributions f_{θ_i} and f_{θ_j}. Let $\gamma \in \Gamma(f_{\theta_i}, f_{\theta_j})$ be an arbitrary joint distribution with marginals f_{θ_i} and f_{θ_j}, and define $R = R_{\theta_i} \times R_{\theta_j}$. Then, for all $(t, s) \in R$, the triangle inequality gives

$$\begin{aligned}\|t - s\|_1 &\leq \|\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]\|_1 + \|t - \mathbb{E}[f_{\theta_i}]\|_1 + \|s - \mathbb{E}[f_{\theta_j}]\|_1 \\ &\leq \Delta_E(\Psi, F) + 2c.\end{aligned}$$

By the union bound, we also have

$$\begin{aligned}\int\int_{(t,s)\in R} \gamma(t,s)\,dt\,ds &= 1 - \int\int_{(t,s)\notin R_{\theta_i}\times R_{\theta_j}} \gamma(t,s)\,dt\,ds \\ &\geq 1 - \int_{t\notin R_{\theta_i}} f_{\theta_i}(t)\,dt - \int_{s\notin R_{\theta_j}} f_{\theta_j}(s)\,ds \\ &\geq 1 - \delta/2 - \delta/2 \\ &= 1 - \delta.\end{aligned}$$

It follows that f_{θ_i} and f_{θ_j} are $(\Delta_E(\Psi, F) + 2c, \delta)$-close. □

C.2 Expected Value Mechanism

Theorem 5.2. *Suppose that f_{θ_i} is a translation of f_{θ_j} for every pair $(\theta_i, \theta_j) \in \Psi$. Let $\sigma \geq c\Delta_{E,2}(\Psi, F)/\epsilon$, where $\epsilon \in (0, 1)$ and $c = \sqrt{2\ln(1.25/\delta)}$ for some $\delta > 0$. Then, the mechanism $\mathcal{M}$ which adds Gaussian noise $Z_k \sim \mathcal{N}(0, \sigma^2)$ independently to each component of $F(\mathfrak{Data})$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ.*

Proof. Let $(\theta_i, \theta_j) \in \Psi$ be a pair of distributions and let $Z = (Z_1, Z_2, \ldots, Z_m)$. Since f_{θ_i} and f_{θ_j} are translations of each other, there exists some $b \in \mathbb{R}^m$ such that $f_{\theta_i}(t) = f_{\theta_j}(t + b)$ for all $t \in \mathbb{R}^m$. Then, for all $S \subseteq \mathrm{Range}(\mathcal{M})$,

$$\begin{aligned}\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_i) &= \int_t f_{\theta_i}(t) \Pr(Z + t \in S)\,dt \\ &= \int_s f_{\theta_j}(s) \Pr(Z + s - b \in S)\,ds.\end{aligned}$$

Now, note that $\|b\|_2 = \|\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]\|_2 \leq \Delta_{E,2}(\Psi, F)$, and that $\sigma \geq c\Delta_{E,2}(\Psi, F)/\epsilon$. Then, since each component of noise $Z_k \sim \mathcal{N}(0, \sigma^2)$ is independently sampled, the Gaussian Mechanism from differential privacy [19] implies that

$$\Pr(Z + s - b \in S) \leq \exp(\epsilon) \Pr(Z + s \in S) + \delta$$

for all $s \in \mathbb{R}^m$. Thus,

$$\begin{aligned}\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_i) &\leq \int_s f_{\theta_j}(s) \left(\exp(\epsilon) \Pr(Z + s \in S) + \delta\right) ds \\ &= \exp(\epsilon) \int_s f_{\theta_j}(s) \Pr(Z + s \in S)\,ds + \delta \\ &= \exp(\epsilon) \Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_j) + \delta.\end{aligned}$$

It follows that $\mathcal{M}$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ. □

C.3 Directional Expected Value Mechanism

Theorem 5.3. *Suppose that f_{θ_i} is a translation of f_{θ_j} and furthermore that the vector $\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]$ is parallel to the unit vector v for each $(\theta_i, \theta_j) \in \Psi$. Then, the mechanism $\mathcal{M}$ described in Algorithm 1 satisfies $(\epsilon, 0)$-distribution privacy with respect to Ψ.*

Proof. Let $(\theta_i, \theta_j) \in \Psi$ be a pair of distributions. Since f_{θ_i} and f_{θ_j} are translations of each other, we have $f_{\theta_i}(t) = f_{\theta_j}(t + b)$, where $b = \mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]$, for all $t \in \mathbb{R}^m$. Then, for all $S \subseteq \mathrm{Range}(\mathcal{M})$,

$$\begin{aligned}\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_i) &= \int_t f_{\theta_i}(t) \Pr(Yv + t \in S)\,dt \\ &= \int_s f_{\theta_j}(s) \Pr(Yv + s - b \in S)\,ds.\end{aligned} \tag{6}$$

By assumption, $b = \mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]$ is parallel to the unit vector v. Thus, letting $S' = S \cap (s + \text{span}(v))$, we have

$$\begin{aligned}\Pr(Yv + s - b \in S) &= \Pr(Yv + s - b \in S') \\ &= \Pr(Yv - b \in S' - s) \\ &= \Pr(Y - \|b\|_2 \in v^{\mathrm{T}}(S' - s)),\end{aligned}$$

where $v^{\mathrm{T}}(S' - s) = \{v^{\mathrm{T}}(x - s) : x \in S'\}$.

Now, observe that $\|\|b\|_2\|_1 = \|b\|_2 = \|\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]\|_2 \leq \Delta_{E,2}(\Psi, F)$. Since $Y \sim \text{Lap}(\Delta_{E,2}(\Psi, F)/\epsilon)$, the Laplace Mechanism from differential privacy [18] then gives that

$$\begin{aligned}\Pr(Y - \|b\|_2 \in v^{\mathrm{T}}(S' - s)) &\leq \exp(\epsilon) \Pr(Y \in v^{\mathrm{T}}(S' - s)) \\ &= \exp(\epsilon) \Pr(Yv + s \in S).\end{aligned}$$

Thus, $\Pr(Yv + s - b \in S) \leq \exp(\epsilon) \Pr(Yv + s \in S)$ for all $s \in \mathbb{R}^m$. Together with (6), this implies

$$\begin{aligned}\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_i) &\leq \exp(\epsilon) \int_s f_{\theta_j}(s) \Pr(Yv + s \in S)\, ds \\ &= \exp(\epsilon) \Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_j).\end{aligned}$$

It follows that $\mathcal{M}$ satisfies $(\epsilon, 0)$-distribution privacy with respect to Ψ. □

C.4 Approximations with Max-Divergence

Theorem 5.6. *Suppose that $\mathcal{M}$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ and that $\mathcal{M}(\mathfrak{Data}) \mid F(\mathfrak{Data})$ is independent of $\mathfrak{Data}$. For each $\theta \in \Theta$, let $\tilde{f}_\theta$ be an approximation of f_θ and suppose that*

$$\max\left(D_\infty^\eta(\tilde{f}_\theta \parallel f_\theta), D_\infty^\eta(f_\theta \parallel \tilde{f}_\theta)\right) \leq \lambda$$

for all $\theta \in \Theta$. Then, the mechanism $\widetilde{\mathcal{M}}$ which uses the approximations $\tilde{f}_\theta$ in place of the true distributions f_θ satisfies (ϵ', δ')-distribution privacy with respect to Ψ, where $\epsilon' = \epsilon + 2\lambda$ and $\delta' = (1 + \exp(\epsilon + \lambda))\eta + \exp(\lambda)\delta$.

Proof. For convenience of presentation, we write $F = F(\mathfrak{Data})$ and refer interchangeably between a random variable and its underlying distribution. By assumption, $\mathcal{M}(\mathfrak{Data}) \mid F$ is independent of $\mathfrak{Data}$, thus, $\widetilde{\mathcal{M}}(\mathfrak{Data}) \mid F$ is also independent of $\mathfrak{Data}$. Let us write $\widetilde{\mathcal{M}}_f$ to denote the distribution of $\widetilde{\mathcal{M}}(\mathfrak{Data}) \mid F \sim f$.

Since $\mathcal{M}$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ, it follows that $\widetilde{\mathcal{M}}$ will as well when the approximations $\tilde{f}_\theta$ are the true distributions. Thus, for all pairs $(\theta_i, \theta_j) \in \Psi$ and subsets $S \subseteq \text{Range}(\widetilde{\mathcal{M}})$,

$$\Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}} \in S) \leq \exp(\epsilon) \Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_j}} \in S) + \delta. \tag{7}$$

Since $\max\left(D_\infty^\eta(\tilde{f}_\theta \parallel f_\theta), D_\infty^\eta(f_\theta \parallel \tilde{f}_\theta)\right) \leq \lambda$ for all $\theta \in \Theta$, we also have

$$\begin{aligned}\Pr(\widetilde{\mathcal{M}}_{f_{\theta_i}} \in S) &= \int_t f_{\theta_i}(t) \Pr(\widetilde{\mathcal{M}}(\mathfrak{Data}) \in S \mid F = t)\, dt \\ &\leq \int_t (\exp(\lambda)\tilde{f}_{\theta_i}(t) + \eta) \Pr(\widetilde{\mathcal{M}}(\mathfrak{Data}) \in S \mid F = t)\, dt \\ &\leq \exp(\lambda) \int_t \tilde{f}_{\theta_i}(t) \Pr(\widetilde{\mathcal{M}}(\mathfrak{Data}) \in S \mid F = t)\, dt + \eta \\ &= \exp(\lambda) \Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}} \in S) + \eta.\end{aligned}$$

Similarly, $\Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_j}} \in S) \leq \exp(\lambda) \Pr(\widetilde{\mathcal{M}}_{f_{\theta_j}} \in S) + \eta$. Together with (7), this gives

$$\begin{aligned}&\Pr(\widetilde{\mathcal{M}}(\mathfrak{Data}) \in S \mid \theta_i) \\ &= \Pr(\widetilde{\mathcal{M}}_{f_{\theta_i}} \in S) \\ &\leq \exp(\lambda) \Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}} \in S) + \eta \\ &\leq \exp(\lambda)(\exp(\epsilon) \Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_j}} \in S) + \delta) + \eta \\ &\leq \exp(\lambda)(\exp(\epsilon)(\exp(\lambda) \Pr(\widetilde{\mathcal{M}}_{f_{\theta_j}} \in S) + \eta) + \delta) + \eta \\ &= \exp(\epsilon') \Pr(\widetilde{\mathcal{M}}_{f_{\theta_j}} \in S) + \delta' \\ &= \exp(\epsilon') \Pr(\widetilde{\mathcal{M}}(\mathfrak{Data}) \in S \mid \theta_j) + \delta'.\end{aligned}$$

It follows that $\widetilde{\mathcal{M}}$ satisfies (ϵ', δ')-distribution privacy with respect to Ψ. □

C.5 Approximations with Wasserstein Distance

Observe that the Wasserstein Mechanism of Section 4 can be viewed as a result on the effect of Laplace noise on the max-divergence between distributions. In particular, Theorem 4.1 may be reinterpreted as follows.

Lemma C.1. *Let F and $\widetilde{F}$ be two distributions on $\mathbb{R}^m$. Let $Z = (Z_1, Z_2, \ldots, Z_m)$, where $Z_k \sim \text{Lap}(W_\infty(F, \widetilde{F})/\epsilon)$ for each k. Then, $D_\infty(F + Z \parallel \widetilde{F} + Z) \leq \epsilon$.*

Using Lemma C.1, we can analyze the privacy guarantees offered by our mechanisms when their underlying assumptions may not exactly hold.

Theorem 5.7. *Suppose that mechanism $\mathcal{M}(\mathfrak{Data}) = F(\mathfrak{Data}) + Z$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ. For each $\theta \in \Theta$, let $\tilde{f}_\theta$ be an approximation of f_θ and let*

$$W = \sup_{\theta \in \Theta} W_\infty(f_\theta, \tilde{f}_\theta).$$

Let $Z' = (Z'_1, Z'_2, \ldots, Z'_m)$, where $Z'_k \sim \text{Lap}(W/\lambda)$ for each k, and let $\widetilde{\mathcal{M}}$ be the mechanism which applies $\mathcal{M}$ using the approximations $\tilde{f}_\theta$ in place of the true distributions f_θ. Then the mechanism which outputs $\widetilde{\mathcal{M}}(\mathfrak{Data}) + Z'$ satisfies (ϵ', δ')-distribution privacy with respect to Ψ, where $\epsilon' = \epsilon + 2\lambda$ and $\delta' = \exp(\lambda)\delta$.

Proof. For convenience of presentation, we write $F = F(\mathfrak{Data})$ and refer interchangeably between a random variable and its underlying distribution. Let us write $\widetilde{\mathcal{M}}_f$ to denote the distribution of $\widetilde{\mathcal{M}}(\mathfrak{Data}) \mid F \sim f$.

Since $\mathcal{M}$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ, it follows that $\widetilde{\mathcal{M}}$ will as well when the approximations $\tilde{f}_\theta$ are the true distributions. Thus, for all pairs $(\theta_i, \theta_j) \in \Psi$ and subsets $S \subseteq \text{Range}(\widetilde{\mathcal{M}})$,

$$\Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}} \in S) \leq \exp(\epsilon) \Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_j}} \in S) + \delta.$$

In turn, this implies

$$\begin{aligned}
&\Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}} + Z' \in S) \\
&= \int_z \Pr(Z' = z) \Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}} \in S - z)\, dz \\
&\leq \int_z \Pr(Z' = z)(\exp(\epsilon) \Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}} \in S - z) + \delta)\, dz \\
&\leq \exp(\epsilon) \Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_j}} + Z' \in S) + \delta. \qquad (8)
\end{aligned}$$

Now, using properties of ∞-Wasserstein distance defined using the L_1 norm, we have

$$\begin{aligned}
W_\infty(\widetilde{\mathcal{M}}_{f_{\theta_i}}, \widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}}) &= W_\infty(F + Z \mid F \sim f_{\theta_i}, F + Z \mid F \sim \tilde{f}_{\theta_i}) \\
&= W_\infty(F \mid F \sim f_{\theta_i}, F \mid F \sim \tilde{f}_{\theta_i}) \\
&= W_\infty(f_{\theta_i}, \tilde{f}_{\theta_i}) \\
&\leq W.
\end{aligned}$$

Hence, by Lemma C.1, we have $D_\infty(\widetilde{\mathcal{M}}_{f_{\theta_i}} + Z' \parallel \widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}} + Z') \leq \lambda$. Similarly, $D_\infty(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_j}} + Z' \parallel \widetilde{\mathcal{M}}_{f_{\theta_i}} + Z') \leq \lambda$. Together with (8), this gives

$$\begin{aligned}
&\Pr(\widetilde{\mathcal{M}}(\mathfrak{Data}) + Z' \in S \mid \theta_i) \\
&= \Pr(\widetilde{\mathcal{M}}_{f_{\theta_i}} + Z' \in S) \\
&\leq \exp(\lambda) \Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}} + Z' \in S) \\
&\leq \exp(\lambda)(\exp(\epsilon) \Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_j}} + Z' \in S) + \delta) \\
&\leq \exp(\lambda)(\exp(\epsilon)(\exp(\lambda) \Pr(\widetilde{\mathcal{M}}_{f_{\theta_j}} + Z' \in S)) + \delta) \\
&= \exp(\epsilon') \Pr(\widetilde{\mathcal{M}}_{f_{\theta_j}} + Z' \in S) + \delta' \\
&= \exp(\epsilon') \Pr(\widetilde{\mathcal{M}}(\mathfrak{Data}) + Z' \in S \mid \theta_j) + \delta'.
\end{aligned}$$

It follows that $\widetilde{\mathcal{M}} + Z'$ satisfies (ϵ', δ')-distribution privacy with respect to Ψ. □

We remark that the privacy guarantees in Theorem 5.7 may sometimes be attainable even without applying the additional noise Z'. Observe that the proof of Theorem 5.7 only requires Z' to satisfy $D_\infty^\eta(\mathcal{M}_{f_\theta} + Z' \parallel \mathcal{M}_{\tilde{f}_\theta} + Z') \leq \lambda$ and $D_\infty^\eta(\mathcal{M}_{\tilde{f}_\theta} + Z' \parallel \mathcal{M}_{f_\theta} + Z') \leq \lambda$. Thus, if these inequalities are satisfied for $Z' = 0$, then the Expected Value Mechanism may be applied safely using the approximations $\tilde{f}_\theta$ with no additional noise required.

Furthermore, in situations where the ∞-Wasserstein distances between f_θ and $\tilde{f}_\theta$ are not well defined, our analysis may be readily re-expressed in terms of the notion of (W, δ)-closeness discussed in Section 4.1. Thus, for instance, Theorem 5.7 and its Gaussian variant may be applied even in cases where unbounded functions are used to approximate bounded functions, as will often be the case when approximating real data with Gaussian distributions.

D ADDITIONAL EXPERIMENTAL RESULTS

D.1 Gaussian Modeling

Figure 5 provides a visualization of our Gaussian approximations for selected query components and underlying sensitive property values. We remark that the variance of each component does vary slightly as the underlying sensitive property value varies, suggesting that our assumption on the covariance matrices Σ_{θ_i} and Σ_{θ_j} being equal for all pairs $(\theta_i, \theta_j) \in \Psi$ cannot hold exactly. For instance, the variance of the number of females in each subset of data is ≈ 18.31 when $p_I = 0.45$, and ≈ 18.03 when $p_I = 0.55$. However, the variance and covariance values are similar enough that we may still apply our mechanisms, albeit with the caveat of slightly weaker privacy guarantees as discussed in Section 5.4.

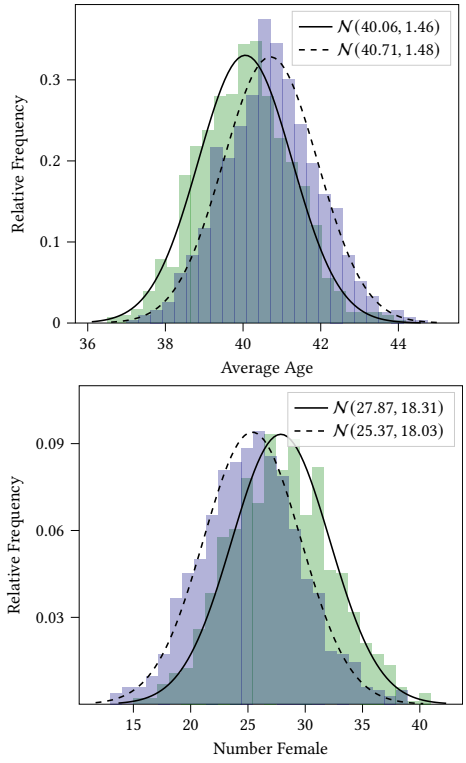
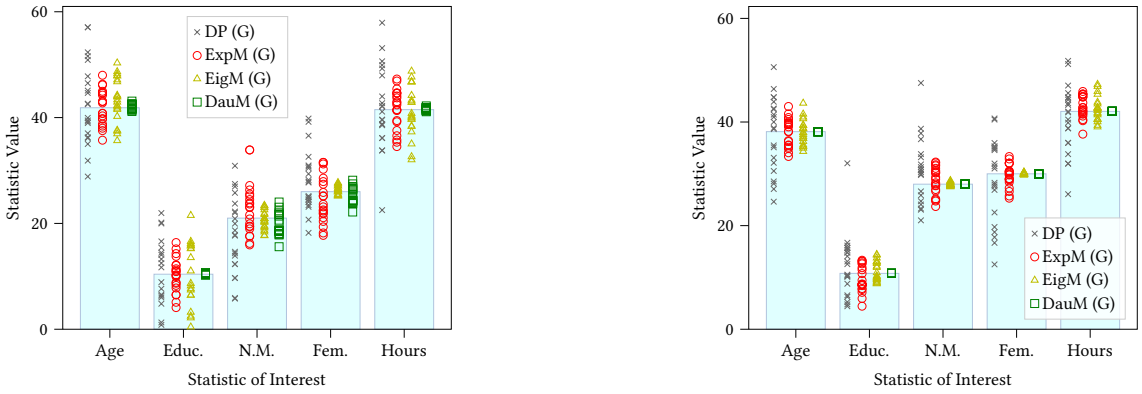


Figure 5: Gaussian approximations for selected statistics of interest over randomly sampled subsets of the Adult dataset. Here, the subsets are sampled to have $n = 100$ records with fixed proportions p_I of individuals having income $>$ \$50K, where green and purple histograms are used for $p_I = 0.45$ and $p_I = 0.55$, respectively.

D.2 Utility Visualizations

Figure 6 provides a visualization of the noise applied by each of our mechanisms in comparison to differential privacy mechanisms. Note again that the latter mechanisms do not provide theoretical distribution privacy guarantees. Moreover, recall that our previously discussed group differential privacy baselines require the amount of noise used by differential privacy mechanisms to be multiplied by $n = 100$ to guarantee distribution privacy.



(a) Gaussian variants, protecting income, $\Delta p_I = 0.02$.

(b) Gaussian variants, protecting work class, $\Delta p_W = 0.04$.

Figure 6: Visualization of variants of the Expected Value Mechanism. The mechanisms are each applied 20 times on a randomly sampled subset of 100 records, with $\epsilon = 1$, and $\delta = 0.001$ for the Gaussian variants. The error incurred by differential privacy (DP) mechanisms are shown for comparison, however, note that these do not provide distribution privacy guarantees.