Protecting Global Properties of Datasets with Distribution Privacy Mechanisms

Michelle Chen
The University of Melbourne

Olga Ohrimenko
The University of Melbourne

Abstract

We consider the problem of ensuring confidentiality of dataset properties aggregated over many records of a dataset. Such properties can encode sensitive information, such as trade secrets or demographic data, while involving a notion of data protection different to the privacy of individual records typically discussed in the literature. In this work, we demonstrate how a *distribution privacy* framework can be applied to formalize such data confidentiality. We extend the Wasserstein Mechanism from Pufferfish privacy and the Gaussian Mechanism from attribute privacy to this framework, then analyze their underlying data assumptions and how they can be relaxed. We then empirically evaluate the privacy-utility tradeoffs of these mechanisms and apply them against a practical property inference attack which targets global properties of datasets. The results show that our mechanisms can indeed reduce the effectiveness of the attack while providing utility substantially greater than a crude group differential privacy baseline. Our work thus provides groundwork for theoretical mechanisms for protecting global properties of datasets along with their evaluation in practice.

1 INTRODUCTION

While many notions of privacy have been proposed for protecting individual contributors of data (Samarati and Sweeney, 1998; Dinur and Nissim, 2003; Dwork et al., 2006; Kifer and Machanavajjhala, 2014), there are situations where it is instead desirable to protect global properties of a dataset. A hospital looking to share patient treatment data may, for instance, want to protect overall patient demographics to avoid unfounded claims of corre-

lations between diseases and certain demographics. Similarly, sharing of such data can exacerbate sensitive political issues, as occurred when the 1932 Lebanon census revealed the population's religious makeup. Indeed, a national census has not been conducted in Lebanon since 1932 (U.S. Department of State, 2019) due to the sensitivity of the matter of religious balance in the region (Maktabi, 1999).

Recent works have developed various *property inference attacks* which demonstrate how data analysis algorithms can leak these global properties of datasets. In contrast to privacy attacks such as membership inference attacks (Shokri et al., 2017) and model inversion attacks (Fredrikson et al., 2015), these attacks aim to discover properties aggregated over all records in a dataset rather than properties of individual records. Ateniese et al. (2015) was the first to formulate such an attack, and further property inference attacks have since been developed against deep neural networks (Ganju et al., 2018), large convolutional networks (Suri and Evans, 2021), in federated learning settings (Melis et al., 2019), in black-box settings (Zhang et al., 2021) and in settings where an attacker can poison the data (Mahloujifar et al., 2022).

Currently, there is a lack of rigorous mechanisms for defending against property inference attacks (Suri and Evans, 2021; Zhang et al., 2022). Intuitive defenses based on regularization or adding noise (Ganju et al., 2018; Melis et al., 2019) lack guarantees of protection. On the other hand, Suri and Evans (2021) proposed a formal model of property inference attacks as a cryptographic game, but did not propose possible defense mechanisms.

Similarly, while some privacy frameworks have considered notions of global properties of datasets, they still lack rigorous mechanisms for protecting such properties. Knowledge hiding (Verykios and Gkoulalas-Divanis, 2008; Domadiya and Rao, 2013) uses a syntactic approach to hide sensitive properties aggregated over a dataset, but does not provide privacy guarantees (Kifer and Machanavajjhala, 2014). On the other hand, theoretical frameworks such as Pufferfish privacy (Kifer and Machanavajjhala, 2014) and distribution privacy (Kawamoto and Murakami, 2019) can model protection of these properties, but have few general mechanisms for this purpose (Song et al., 2017; Zhang et al., 2022). To address this, Zhang et al. (2022) propose at-

tribute privacy for protecting global properties of datasets. However, their definitions and mechanisms involve restrictive data assumptions, consider queries with a single output and have not been evaluated on real data for their impact on utility or how well they protect against attacks.

We aim to contribute groundwork for a theoretically supported approach for protecting against property inference attacks while overcoming some of these shortcomings of previous works. Our main contributions are as follows:

- We demonstrate how distribution privacy (Kawamoto and Murakami, 2019) can be applied to formalize protection of global properties of datasets. This provides an arguably simpler and more general alternative to attribute privacy (Zhang et al., 2022) as it can capture a wide range of global properties using fewer data assumptions.

- We adapt the Wasserstein Mechanism from Pufferfish privacy (Song et al., 2017) to distribution privacy. Since this mechanism requires computation of ∞-Wasserstein distances which may be large, not well defined or hard to compute in practice, we extend it to allow use of an approximation of ∞-Wasserstein distance, with corresponding privacy guarantees. This result is of independent interest and can be applied in other settings where the Wasserstein Mechanism is used (e.g., Song et al. (2017); Zhang et al. (2022)).

- We propose the Expected Value Mechanism for distribution privacy, which generalizes the Gaussian Mechanism of Zhang et al. (2022). In particular, we extend the mechanism to queries with multiple components (e.g., multiple statistics of interest), then analyze directional and adversarial uncertainty assumptions for reducing the noise required by the mechanism.

- We evaluate our mechanisms on privacy-utility trade-offs and effectiveness against a property inference attack. The results show that they can reduce attack accuracy while providing substantially greater utility than a group differential privacy baseline. To our knowledge, this is the first evaluation of utility and defense for mechanisms with theoretical guarantees on protecting global properties of real datasets.

2 PRIVACY FRAMEWORK

In this section, we formalize the protection of global properties of datasets using the distribution privacy framework (Kawamoto and Murakami, 2019).

2.1 Differential Privacy

We first outline differential privacy (Dwork et al., 2006), a widely used privacy framework 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 2.1 (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 $S \subseteq \mathrm{Range}(\mathcal{M})$,*

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

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 (Dwork and Roth, 2014) 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, making them unsuitable for protecting global properties of datasets.

2.2 Distribution Privacy

Distribution privacy (Kawamoto and Murakami, 2019) is similar to differential privacy, but is defined using the possible underlying distributions of a dataset rather than the presence of individual records. 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 a dataset $\mathfrak{Data}$, intuitively representing an attacker's beliefs on how $\mathfrak{Data}$ may have been generated. The pairs of distributions Ψ then reflect sensitive properties to be protected from the attacker.

Definition 2.2 (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 $S \subseteq \mathrm{Range}(\mathcal{M})$,*

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

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 $\mathfrak{Data}$. In this manner, distribution privacy can be considered to share similarities with noiseless privacy (Bhaskar et al., 2011), Pufferfish privacy (Kifer and Machanavajjhala, 2014) and other notions of privacy (Bassily et al., 2013; Machanavajjhala et al., 2009) 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 possible underlying distributions of $\mathfrak{Data}$ given particular values for the sensitive properties of interest.

Example 2.1 (Census Data). *Suppose a government body owns a dataset $\mathfrak{Data}$ of complete census data and seeks 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 (e.g., 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 $(\theta_{r \mp d}, \theta_{r \pm d})$ with $|d| \leq k$.*

The generality of the distribution privacy framework also allows it to simultaneously capture the protection of multiple properties (global or otherwise) by specifying pairs of data distributions in Ψ appropriately for each sensitive property.

Identifying appropriate data distributions Θ and pairs Ψ is an important data modeling step when using the distribution privacy framework. For our experiments in Section 5, we assume that aggregated Census data can be approximated by some multivariate Gaussian distribution and use data sampling to estimate possible parameters. In general, we take an approach similar to Pufferfish privacy (Kifer and Machanavajjhala, 2014) and assume the existence of a domain expert who can instantiate the framework according to the application requirements.

2.3   Relation to Other Privacy Frameworks

While we adopt the distribution privacy definition of Kawamoto and Murakami (2019), our work applies their definition in a different context. Kawamoto and Murakami (2019) 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, in which case the Tupling Mechanism of Kawamoto and Murakami (2019) and other distribution privacy mechanisms are not applicable.

In this context, the distribution privacy framework can be viewed as a generalization of the distributional attribute privacy framework proposed by Zhang et al. (2022) 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, our 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.

3   WASSERSTEIN MECHANISM

In this section, we adapt the Wasserstein Mechanism from Pufferfish privacy (Song et al., 2017) to distribution privacy. Since this mechanism requires computation of ∞-Wasserstein distances which may be large or not well defined in practice, we also propose a variation, the Approximate Wasserstein Mechanism, which can be applied in more general settings. We note that our analysis is of independent interest as it can also be used by other mechanisms that require ∞-Wasserstein distances.

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 assume that F takes values in $\mathbb{R}^m$. Letting f_θ denote the distribution of $F(\mathfrak{Data})$ given $\mathfrak{Data} \sim \theta$, we intuitively 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. (2017) identify ∞-Wasserstein distance as one suitable measure for determining the amount of noise required.

Definition 3.1 (∞-Wasserstein Distance). *Let μ and ν be two distributions on $\mathbb{R}^m$. 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*

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

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.

We adapt the Wasserstein Mechanism of Song et al. (2017) by scaling Laplace noise to the maximum ∞-Wasserstein distance between pairs of distributions of interest

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

Theorem 3.1. *The mechanism $\mathcal{M}(\mathfrak{Data}) = F(\mathfrak{Data}) + Z$, where $Z \sim \mathrm{Lap}(0, (\Delta_W(\Psi, F)/\epsilon)I)$, satisfies $(\epsilon, 0)$-distribution privacy with respect to Ψ.*

Now, this mechanism assumes that $W_\infty(f_{\theta_i}, f_{\theta_j})$ is well-defined for each pair $(\theta_i, \theta_j) \in \Psi$. This can be an issue

if, for instance, the query function F is unbounded. Similarly, if F takes on a large range of values with low probability (e.g., average salary), these ∞-Wasserstein distances can be large, resulting in more noise than is necessary for the majority of cases. Thus, we prove an variation of the Wasserstein Mechanism which can require less noise in exchange for a small probability of loss of privacy.

3.1 Approximate Wasserstein Mechanism

Our variation essentially allows a low probability set of possible values of F to be disregarded while still providing (ϵ, δ)-distribution privacy. Formally, we first generalize ∞-Wasserstein distance as follows.

Definition 3.2 $((W, \delta)$-closeness). *Let μ and ν be two distributions on $\mathbb{R}^m$. We say that μ and ν are (W, δ)-close if there exists a distribution $\gamma \in \Gamma(\mu, \nu)$ and subset $R \subseteq \mathrm{supp}(\gamma)$ 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 $\gamma \in \Gamma(\mu, \nu)$ as 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.

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, consider distributions μ and ν 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 elements 1, 2, 3, 100 respectively. Then, as shown in Figure 1, $W_\infty(\mu, \nu) = 97$ 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 shows how scaling noise according to (W, δ)-closeness can result in significantly less noise than scaling to ∞-Wasserstein distance. The ability to disregard a small mass δ also allows (W, δ)-closeness to be well-defined even when ∞-Wasserstein distance may not be.

Theorem 3.2. *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 $\mathcal{M}(\mathfrak{Data}) = F(\mathfrak{Data}) + Z$, where $Z \sim \mathrm{Lap}(0, (W/\epsilon)I)$, 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 $\gamma \in$
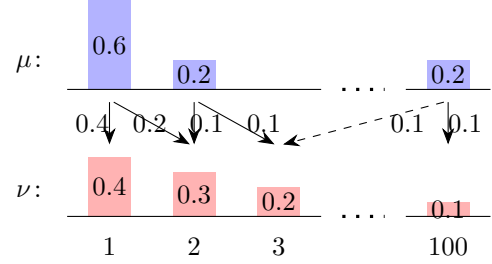

Figure 1: Distinction between ∞-Wasserstein distance and (W, δ)-closeness. $W_\infty(\mu, \nu) = 97$ due to the shift represented by the dashed arrow, while disregarding this shift shows that μ and ν are $(W, 0.1)$-close with $W = 1$.

$\Gamma(f_{\theta_i}, f_{\theta_j})$ and a subset $R \subseteq \mathrm{supp}(\gamma)$ as in the conditions of Definition 3.2. Now, for all $S \subseteq \mathrm{Range}(\mathcal{M})$,

$$\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_i)$$
$$= \int_t \Pr(F(\mathfrak{Data}) = t \mid \theta_i) \Pr(Z + t \in S)\, dt$$
$$= \int_t \int_s \gamma(t, s) \Pr(Z + t \in S)\, ds\, dt.$$

Since each component of noise $Z_k \sim \mathrm{Lap}(W/\epsilon)$ is independently sampled, the Laplace Mechanism from differential privacy (Dwork et al., 2006) shows that

$$\int\!\!\int_{(t,s) \in R} \gamma(t, s) \Pr(Z + t \in S)\, ds\, dt$$
$$\leq \exp(\epsilon) \int\!\!\int_{(t,s) \in R} \gamma(t, s) \Pr(Z + s \in S)\, ds\, dt,$$

since $\|t - s\|_1 \leq W$ for all $(t, s) \in R$. We also have

$$\int\!\!\int_{(t,s) \notin R} \gamma(t, s) \Pr(Z + t \in S)\, ds\, dt$$
$$\leq \int\!\!\int_{(t,s) \notin R} \gamma(t, s)\, ds\, dt \leq \delta.$$

Thus,

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

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

We remark that (W, δ)-closeness can be viewed as an approximation of the ∞-Wasserstein distance between two distributions using a subset of their supports. More precisely, given distributions μ and ν, suppose that there exist subsets $R_1 \subseteq \mathrm{supp}(\mu)$ and $R_2 \subseteq \mathrm{supp}(\nu)$ 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

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

Comparing with the definition of ∞-Wasserstein distance, the above expression can be interpreted as only considering the movement of probability masses between the subsets $R_1 \subseteq \operatorname{supp}(\mu)$ and $R_2 \subseteq \operatorname{supp}(\nu)$. Thus, (W, δ)-closeness formalizes the notion that $W_\infty(\mu, \nu)$ can be approximated using subsets of the supports of μ and ν. In turn, this provides privacy guarantees when applying the Wasserstein Mechanism with common methods of computing ∞-Wasserstein distance, which proceed by approximating continuous distributions with bounded or discrete distributions (Panaretos and Zemel, 2019).

3.2 Functions Bounded with High Probability

As an example of using the Approximate Wasserstein Mechanism, consider settings in which the queries of interest are bounded with high probability, as is often the case for aggregated statistics such as average salary. Such settings allow the following efficient application of the Approximate Wasserstein Mechanism.

Theorem 3.3. *Suppose that $\|F(\mathfrak{Data}) - \mathbb{E}[F(\mathfrak{Data})]\|_1 \leq c$ with probability at least $1 - \delta/2$ for each $\mathfrak{Data} \sim \theta \in \Theta$. Let*

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

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 3.3 is provided in Appendix B.1.

Given Proposition 3.3, we can apply Theorem 3.2 to deduce that Laplace noise with scale $(\Delta_E(\Psi, F) + 2c)/\epsilon$ is sufficient for (ϵ, δ)-distribution privacy with respect to Ψ. Thus, if F is bounded by a small range of values with high probability, then c will be small and the Approximate Wasserstein Mechanism can be applied while requiring only a small amount of noise.

Observe also that $\Delta_E(\Psi, F) + 2c$ can be well-defined even if F is unbounded, as long as F is bounded with high probability. Moreover, it only requires expected values and bounds for f_θ to be computed. This is in contrast to the ∞-Wasserstein distances required by the Wasserstein Mechanism (Theorem 3.1), which can be expensive to compute (Song et al., 2017) and moreover may not be well-defined if F is unbounded.

4 EXPECTED VALUE MECHANISM

While theoretically applicable to general instantiations of distribution privacy, the Wasserstein Mechanism can be too computationally expensive to use in practice due to the need for ∞-Wasserstein distances (Song et al., 2017). The variation we proposed can accommodate more efficient approximations, however, in general, (W, δ)-closeness may still be difficult to compute.

We now propose the Expected Value Mechanism, an alternative, efficient mechanism that can achieve distribution privacy in specialized settings. In particular, it can be applied when the pairs of distributions to be protected are approximate translations of each other (e.g., Gaussian with similar variance). Our base mechanism adapts the Gaussian Mechanism from attribute privacy (Zhang et al., 2022). We then propose two variants that have smaller noise requirements in circumstances when directional and adversarial uncertainty assumptions can be made, allowing insights of a domain expert to be incorporated to improve utility. Finally, we formalize how data assumptions can be relaxed while still providing confidentiality guarantees. Note that theorem proofs are provided in the appendix.

4.1 Mechanism

As in previous sections, let f_θ denote the distribution of $F(\mathfrak{Data})$ given $\mathfrak{Data} \sim \theta$. We assume that f_{θ_i} and f_{θ_j} are translations of each other for each pair $(\theta_i, \theta_j) \in \Psi$. That is, for each such pair, there exists $c \in \mathbb{R}^m$ such that $f_{\theta_i}(t) = f_{\theta_j}(t + c)$ for all $t \in \mathbb{R}^m$. This occurs, for instance, when each distribution can be assumed to be Gaussian with the same variance but different means.

The Expected Value Mechanism achieves distribution privacy by applying Laplace or Gaussian noise proportional to the worst case distance between expected values

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

Theorem 4.1. *Suppose that f_{θ_i} is a translation of f_{θ_j} for every pair $(\theta_i, \theta_j) \in \Psi$. Then, the mechanism $\mathcal{M}(\mathfrak{Data}) = F(\mathfrak{Data}) + Z$, where $Z \sim \operatorname{Lap}(0, (\Delta_{E,1}(\Psi, F)/\epsilon)I)$, satisfies $(\epsilon, 0)$-distribution privacy with respect to Ψ.*

Theorem 4.2. *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 $\mathcal{M}(\mathfrak{Data}) = F(\mathfrak{Data}) + Z$, where $Z \sim \mathcal{N}(0, \sigma^2 I)$, satisfies (ϵ, δ)-distribution privacy with respect to Ψ.*

4.2 Variant Using Directional Assumptions

We now propose a variant based on the insight that distribution privacy may sometimes be achievable while applying noise only to particular directions or components for a

query. To illustrate, consider a company seeking to release customer statistics while protecting their proportion of female customers. Suppose that particular statistics, such as customer income and investments, tend 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 protection of the global property p. We formalize this with the Directional Expected Value Mechanism, described in Theorem 4.3 as follows.

Theorem 4.3. *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 $\mathcal{M}(\mathfrak{Data}) = F(\mathfrak{Data}) + Yv$, where $Y \sim \mathrm{Lap}(\Delta_{E,2}(\Psi, F)/\epsilon)$, satisfies $(\epsilon, 0)$-distribution privacy with respect to Ψ.*

If the differences of mean vectors $\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]$ are not all parallel to some vector v, the mechanism can be modified by first identifying 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. Noise proportional to $\Delta_{E,2}(\Psi, F)/\epsilon$ can then be applied in the directions of $v_1, v_2, \ldots, v_d$, reducing the overall noise requirement if the number of directions d is small in comparison to the total dimension m of the query. 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 4.4. Note also that the mechanism can readily be modified to use Gaussian noise instead of Laplace noise.

4.3 Variant Using Adversarial Uncertainty

We now discuss how adversarial uncertainty in the data distributions can be exploited by the mechanism. Using the scenario from the previous section as an example, variance in income may be enough to prevent accurate estimation of the proportion of female customers even if there is a correlation between these properties. The company may then be able to safely release their average customer income with little to no applied noise.

For this variant, 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 $(\theta_i, \theta_j) \in \Psi$. With these assumptions, we first identify a set of conditions under which a query $F(\mathfrak{Data})$ may be safely released without additional noise.

Theorem 4.4. *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 $\mathcal{M}(\mathfrak{Data}) = F(\mathfrak{Data})$ 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 4.4 is that every multivariate Gaussian variable can be expressed as a transformation $X = AZ + \mu$ of a standard normal vector Z, and $\Pr(X = x) = \frac{1}{|A|} \Pr(Z = A^{-1}x)$ for all $x \in \mathbb{R}^m$. The proof then proceeds similarly to that of the Gaussian Mechanism from differential privacy (Dwork and Roth, 2014).

From Theorem 4.4, it follows that privacy can be ensured when there is sufficient noise in the eigenvector directions of each covariance matrix Σ_θ. When these eigenvector directions are the same for each $\theta \in \Theta$, we thus arrive at the Eigenvector Gaussian Mechanism as follows.

Theorem 4.5. *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 $\theta \in \Theta$. Let $c = \sqrt{2\ln(1.25/\delta)}$. For each $\theta \in \Theta$, let*

$$\sigma_{\theta,k}^2 = \max\left(0, (c\Delta_{E,2}(\Psi, F)/\epsilon)^2 - \lambda_{\theta,k}^2\right)$$

where $\lambda_{\theta,k}^2 = v_k^{\mathrm{T}} \Sigma_\theta v_k$ are squared eigenvalues. Let $\sigma_k^2 = \max_{\theta \in \Theta} \sigma_{\theta,k}^2$ and $\Sigma = \sum_k \sigma_k^2 v_k v_k^{\mathrm{T}}$. Then, the mechanism $\mathcal{M}(\mathfrak{Data}) = F(\mathfrak{Data}) + Z$, where $Z \sim \mathcal{N}(0, \Sigma)$, satisfies (ϵ, δ)-distribution privacy with respect to Ψ.

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

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

If $\epsilon = 1$ and $\delta = 0.001$, then $(c\Delta_{E,2}(\Psi, F)/\epsilon)^2 \approx 28.52$. Since 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 corresponding eigenvalues 10 and 25, 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. In comparison, if adversarial uncertainty was not exploited, Gaussian noise with variance 28.52 would need to be added in both directions — significantly impacting the utility.

We note that the Gaussian Mechanism of Zhang et al. (2022) also exploits adversarial uncertainty, but only in the one-dimensional case. Thus, our work generalizes this idea to higher dimensions. We also remark that adversarial uncertainty can be exploited together with the directional assumptions of Section 4.2, as we detail in Appendix A.2.

4.4 Privacy Guarantees Under Relaxed Assumptions

In our analysis so far, we adopted various translation assumptions, directional assumptions, and adversarial uncertainty assumptions on the distributions involved. These assumptions provide an initial means of incorporating domain knowledge to improve utility, however, may be difficult to apply in practice. One approach to relax the assumption requirements is to use approximations $\tilde{f}_\theta$ in place of the exact distributions f_θ, producing a 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 $\tilde{f}_\theta$. The resultant privacy loss can then be quantified with max-divergence, similar to Zhang et al. (2022) for their setting.

Definition 4.1 (δ-Approximate Max-Divergence). *Let μ and ν be two distributions. The δ-approximate max-divergence $D_\infty^\delta(\mu \parallel \nu)$ between μ and ν is defined as*

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

Theorem 4.6. *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 $\delta' = (1 + \exp(\epsilon + \lambda))\eta + \exp(\lambda)\delta$.

We furthermore observe that the privacy loss can alternatively be controlled with a small amount of additional noise, using ∞-Wasserstein distance to quantify possible deviations between the assumed approximations and true distributions.

Theorem 4.7. *Suppose that the 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 $\delta' = \exp(\lambda)\delta$.

These results quantify how much protection can be ensured if modeling assumptions made by a domain expert do not exactly hold. Thus, they provide justification for the use of these assumptions when applying our mechanisms by providing fallback guarantees where necessary.

Table 1: Mechanisms Implemented in the Evaluation.

Variant of Expected Value Mechanism	Shorthand
Laplace (Section 4.1)	ExpM (L)
Laplace, Directional (Section 4.2)	DirM (L)
Gaussian (Section 4.1)	ExpM (G)
Gaussian, Eigenvector (Section 4.3)	EigM (G)
Gaussian, Directional with Adversarial Uncertainty (Section 4.3)	DauM (G)

5 EXPERIMENTS

We now evaluate our mechanisms on privacy-utility trade-offs and effectiveness against property inference attacks.[1]

5.1 Methodology

We use the Adult dataset from the UCI Machine Learning repository (Dua and Graff, 2017), which consists of records from the 1994 US Census database. We remove records with missing attributes, then for the remaining $45\,222$ records we randomly allocate $10\,000$ as auxiliary data to implement attacks, $10\,000$ as testing data to evaluate the accuracy of the attacks, and the remainder for modeling data distributions to implement our mechanisms.

Our task is to release several statistics on a subset of 100 records, namely average age, average number of years of education, number of individuals who have never married, number of female individuals, and average hours worked per week. We consider two possible sensitive properties global for a particular subset — the proportion p_I of individuals in the subset whose income exceeds \$50K, and the proportion p_W who are private sector workers. For each, we want to prevent an attacker from determining whether p has value $0.5 - \Delta p/2$ or $0.5 + \Delta p/2$, for $\Delta p \in [0.02, 0.4]$.

We evaluate five variants of the Expected Value Mechanism, summarized in Table 1. In order to apply these mechanisms, with the statistics of interest as the query F, we assume that the possible query distributions f_θ are approximately multivariate Gaussian with a shared covariance matrix $\Sigma_{\theta_i} = \Sigma_{\theta_j}$ for all pairs $(\theta_i, \theta_j) \in \Psi$. We compute appropriate mean vectors μ_θ and covariance matrices Σ_θ by randomly sampling 1000 subsets of 100 records from the Adult dataset with global sensitive property values as specified by θ, with details in Appendix C.1. We note that while these modeling steps may be difficult in practice, here they nonetheless allow initial evaluation of the utility and defense provided by our distribution privacy approach.

---

[1]Our code for these experiments can be found at `https://github.com/mgcsls/mechanisms-global-properties`

(a) Income, $\Delta p_I = 0.1$.

(b) Work class, $\Delta p_W = 0.04$.
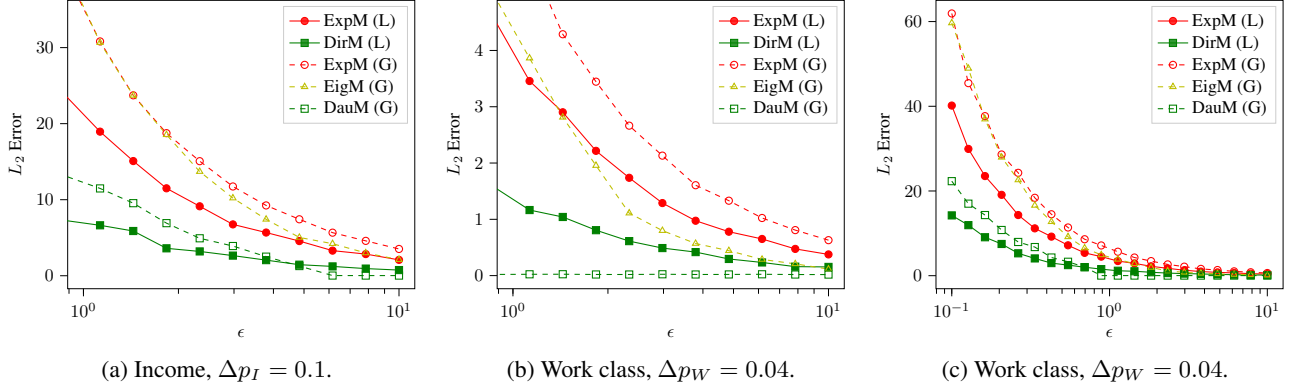
(c) Work class, $\Delta p_W = 0.04$.

Figure 2: Privacy-utility tradeoffs of our mechanisms measured by L_2 error when (a) protecting income, $\Delta p_I = 0.1$; and (b), (c) protecting work class, $\Delta p_W = 0.04$, for varying ϵ. For the Gaussian mechanisms, $\delta = 0.001$.

(a) Income property p_I.
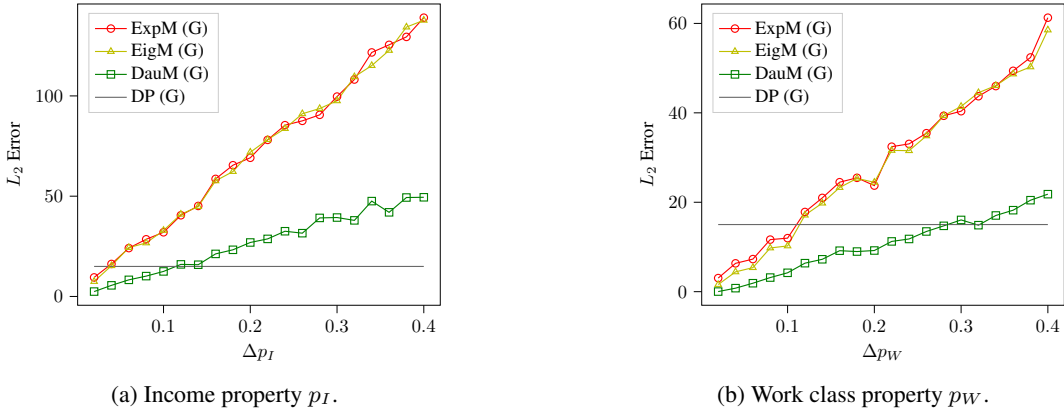
(b) Work class property p_W.

Figure 3: Error incurred by our mechanisms as Δp varies, in comparison to the Gaussian Mechanism from differential privacy (DP). Here, $\epsilon = 1$ and $\delta = 0.001$.

As a baseline, we use the Laplace and Gaussian Mechanisms from group differential privacy (Dwork and Roth, 2014) applied with group size $k = n = 100$, the size of the subsets queried. We note also that the Wasserstein Mechanism and Gaussian Mechanism of Zhang et al. (2022) are equivalent to ExpM (L) and ExpM (G) under our modeling assumptions. Thus, our experiments also provide insight into the practical impact of these mechanisms, which have not been evaluated on real datasets until now.

Table 2: L_2 Error Incurred Protecting the Income Property to an Additive Factor of $\Delta p_I = 0.1$. Here, $\delta = 0.001$.

Mechanism	L_2 **Error**		
	$\epsilon = 0.2$	$\epsilon = 1$	$\epsilon = 5$
ExpM (G)	177.28	34.98	7.11
EigM (G)	175.65	34.87	4.89
DauM (G)	69.85	13.40	1.24
GroupDP (G)	7394.67	1539.93	293.17

5.2 Privacy-Utility Tradeoffs

Figure 2 shows privacy-utility tradeoffs as our mechanisms are applied to protect the income and work class properties to additive factors $\Delta p_I = 0.1$ and $\Delta p_W = 0.04$, respectively. Group differential privacy baselines are not included in this figure since they incurred error values an order of magnitude larger than the Expected Value Mechanism, as shown in Table 2 for the Gaussian variants. Note that all error values are averaged over 50 repetitions of sampling

subsets of the Adult dataset, computing the statistics of interest, then applying each mechanism.

Figure 3 furthermore shows the error incurred by our mechanisms as Δp varies in comparison to the Gaussian Mechanism from differential privacy, noting that differential privacy does not provide distribution privacy guarantees. DauM incurs less error than differential privacy when protecting the income property to an additive factor of $\Delta p_I \leq 0.12$ (e.g., preventing an attacker from distinguishing be-

tween data with 44% and 56% of individuals with income $> \$50\text{K}$), or protecting the work class property to an additive factor of $\Delta p_W \leq 0.28$. This demonstrates an advantage of our mechanisms over differential privacy, as they are able to scale noise differently according to the specific sensitive property of interest and degree to which it needs to be protected. However, it also shows that these mechanisms may require considerable amounts of noise to protect sensitive properties to larger values of Δp, suggesting the need for alternative or additional domain specific analysis.

Nonetheless, Figure 2 shows how directional assumptions can reduce the error incurred by our mechanisms. This illustrates the potential to significantly reduce overall error by essentially only applying noise to query components with high correlation with the sensitive property being protected. On the other hand, our optimizations using adversarial uncertainty (i.e., EigM, DauM) were generally only noticeable for relatively large ϵ. These optimizations may be more applicable when there is less correlation between the query and sensitive properties, formally justifying a statistical release without additional noise if inherent adversarial uncertainty is sufficient to provide privacy.

5.3 Property Inference Attack

We now apply our mechanisms against a property inference attack aiming to determine whether the sensitive property has value $p_1 = 0.45$ or $p_2 = 0.55$, i.e., $\Delta p = 0.1$. Our attack is based on the standard meta-classifier technique (Ateniese et al., 2015; Zhang et al., 2021), with logistic regression as the meta-classifier, the query F as training features and the sensitive property p as labels. We generate training features and labels using the auxiliary data to sample 200 shadow datasets, each of size $n = 100$ records with half of the sampled datasets satisfying $p = p_1$ and half satisfying $p = p_2$. To evaluate meta-classifier accuracy, we use 200 sampled subsets of size $n = 100$ from the testing data, half with p_1 and half with p_2. We report the average meta-classifier accuracy over 50 repetitions.

For the income property p_I, the attack has 75% success rate when statistics are released with no defense, while Figure 4 shows that applying our mechanisms with $\epsilon = 0.1$ reduces the attack accuracy to near 50%. Note that varying δ between 0.0001 and 0.01 had a much smaller effect on the attack accuracy than varying ϵ from 0.1 to 10, as we summarise and discuss in Appendix C.2. Note also that the results for Laplace noise and for p_W are omitted as they show similar trends.

Now, group differential privacy appears to provide the strongest defense out of all of the implemented mechanisms. However, this is due to the excessive noise it applies, as was shown in Table 2. In contrast, our mechanisms are more calibrated for (ϵ, δ)-distribution privacy with a trend of high to low defense as ϵ increases.

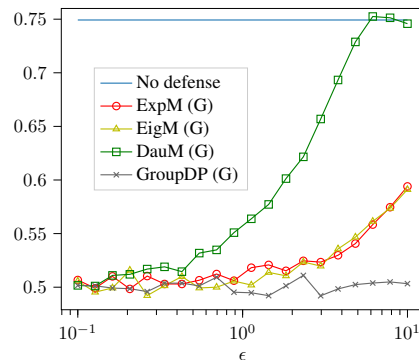
Figure 4: Accuracy of attack determining whether the income property is $p_I = 0.45$ or $p_I = 0.55$. Defense mechanisms are applied with varying ϵ and $\delta = 0.001$.

Finally, we remark that our mechanisms' empirical effectiveness against this attack does not necessarily guarantee defense capabilities against other attacks. Indeed, the need to defend against a range of attacks, including possibly not yet known attacks, highlights the importance of our mechanisms' formal guarantees. Nonetheless, these attack results provide an initial demonstration of our theoretical defenses translating to defense in practice.

6 CONCLUDING REMARKS

We presented an approach to protecting global properties of datasets using the distribution privacy framework and investigated two mechanisms for achieving distribution privacy — the Wasserstein Mechanism, adapted from Pufferfish privacy, and the Expected Value Mechanism, generalizing the Gaussian Mechanism from attribute privacy. Our experiments demonstrated that these mechanisms can reduce the accuracy of a property inference attack while providing significantly better utility than a group differential privacy baseline.

We believe that our work provides initial steps towards rigorously protecting global properties of datasets, leaving several open questions in this area. Future work could include generalizing our mechanisms to a wider range of data distributions, designing new mechanisms under different data modeling assumptions, and evaluating mechanisms in a broader range of application scenarios.

References

Ateniese, G., Mancini, L. V., Spognardi, A., Villani, A., Vitali, D., and Felici, G. (2015). Hacking smart machines with smarter ones: How to extract meaningful data from

machine learning classifiers. *International Journal of Security and Networks*, 10(3):137–150.

Bassily, R., Groce, A., Katz, J., and Smith, A. (2013). Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 439–448.

Bhaskar, R., Bhowmick, A., Goyal, V., Laxman, S., and Thakurta, A. (2011). Noiseless database privacy. In *Proceedings of the 17th International Conference on the Theory and Application of Cryptology and Information Security*, volume 7073, pages 215–232.

Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy. In *Proceedings of the 22nd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 202–210.

Domadiya, N. H. and Rao, U. P. (2013). Hiding sensitive association rules to maintain privacy and data quality in database. In *Proceedings of the 3rd IEEE International Advanced Computing Conference*, pages 1306–1310.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, page 265–284.

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

Forbes, C., Evans, M., Hastings, N., and Peacock, B. (2011). *Statistical Distributions*. John Wiley & Sons, 4th edition.

Fredrikson, M., Jha, S., and Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333.

Ganju, K., Wang, Q., Yang, W., Gunter, C. A., and Borisov, N. (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*, pages 619–633.

Kawamoto, Y. and Murakami, T. (2019). Local obfuscation mechanisms for hiding probability distributions. In *Proceedings of the 24th European Symposium on Research in Computer Security*, volume 11735, pages 128–148.

Kifer, D. and Machanavajjhala, A. (2014). Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, 39(1):3:1–3:36.

Machanavajjhala, A., Gehrke, J., and Götz, M. (2009). Data publishing against realistic adversaries. *Proceedings of the VLDB Endowment*, 2(1):790–801.

Mahloujifar, S., Ghosh, E., and Chase, M. (2022). Property inference from poisoning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1569–1569, Los Alamitos, CA, USA. IEEE Computer Society.

Maktabi, R. (1999). The lebanese census of 1932 revisited. who are the lebanese? *British Journal of Middle Eastern Studies*, 26(2):219–241.

Melis, L., Song, C., Cristofaro, E. D., and Shmatikov, V. (2019). Exploiting unintended feature leakage in collaborative learning. In *Proceedings of the 40th IEEE Symposium on Security and Privacy*, pages 691–706.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431.

Samarati, P. and Sweeney, L. (1998). Generalizing data to provide anonymity when disclosing information. In *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, page 188.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *Proceedings of the 38th IEEE Symposium on Security and Privacy*, pages 3–18.

Song, S., Wang, Y., and Chaudhuri, K. (2017). Pufferfish privacy mechanisms for correlated data. In *Proceedings of the 43rd ACM SIGMOD International Conference on Management of Data*, pages 1291–1306.

Suri, A. and Evans, D. (2021). Formalizing and estimating distribution inference risks. *CoRR*.

U.S. Department of State (2019). International religious freedom report: Lebanon. *2019 Report on International Religious Freedom*.

Verykios, V. S. and Gkoulalas-Divanis, A. (2008). A survey of association rule hiding methods for privacy. In *Privacy-Preserving Data Mining - Models and Algorithms*, volume 34, pages 267–289.

Zhang, W., Ohrimenko, O., and Cummings, R. (2022). Attribute privacy: Framework and mechanisms. *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*.

Zhang, W., Tople, S., and Ohrimenko, O. (2021). Leakage of dataset properties in multi-party machine learning. In *Proceedings of the 30th USENIX Security Symposium*, pages 2687–2704.

A ADVERSARIAL UNCERTAINTY RESULTS

In this section, we provide proofs and further discussion for the results stated in Section 4.3.

Theorem 4.4. *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 $\mathcal{M}(\mathfrak{Data}) = F(\mathfrak{Data})$ 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 $x \in \mathbb{R}^m$,

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

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,

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

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

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

The conditions of the theorem moreover give

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

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

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

The above quantity is bounded by $\exp(\epsilon)$ whenever $|\alpha_1| \leq c - \epsilon/2c$. 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

$$\frac{\sqrt{2}\exp(-t^2/2)}{t\sqrt{\pi}} < \delta$$

$$\iff t\exp\left(\frac{t^2}{2}\right) > \frac{\sqrt{2}}{\sqrt{\pi}\delta}$$

$$\iff \ln(t) + \frac{t^2}{2} > \ln\left(\frac{\sqrt{2}}{\sqrt{\pi}\delta}\right).$$

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

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

Thus, it suffices to have

$$\ln\left(\frac{7}{6}\right) + \frac{c^2 - 1}{2} \geq \ln\left(\frac{\sqrt{2}}{\sqrt{\pi}\delta}\right)$$

$$\iff c^2 \geq 2\ln\left(\frac{6\sqrt{2}\exp(1/2)}{7\sqrt{\pi}\delta}\right)$$

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

$$R_1 = \{x : |\alpha_1| \leq c - \epsilon/2c\}$$
$$R_2 = \{x : |\alpha_1| > c - \epsilon/2c\}.$$

Then, $\Pr(F(\mathfrak{Data}) = x \mid \theta_i) \leq \exp(\epsilon)\Pr(F(\mathfrak{Data}) = x \mid \theta_j)$ 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

$$\Pr(F(\mathfrak{Data}) \in S \mid \theta_i) = \Pr(F(\mathfrak{Data}) \in S_1 \mid \theta_i) + \Pr(F(\mathfrak{Data}) \in S_2 \mid \theta_i)$$
$$\leq \exp(\epsilon)\Pr(F(\mathfrak{Data}) \in S_1 \mid \theta_j) + \delta$$
$$\leq \exp(\epsilon)\Pr(F(\mathfrak{Data}) \in S \mid \theta_j) + \delta.$$

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

Theorem 4.4 gives a means of estimating the amount of privacy that can be guaranteed only using adversarial uncertainty 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 A.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*

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

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 4.4 and the result follows. $\square$

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 4.1.

Corollary A.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

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

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,

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

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

A.1 Eigenvector Gaussian Mechanism

Theorem 4.5. *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 $\theta \in \Theta$. Let $c = \sqrt{2 \ln(1.25/\delta)}$. For each $\theta \in \Theta$, let*

$$\sigma_{\theta,k}^2 = \max\left(0, (c\Delta_{E,2}(\Psi, F)/\epsilon)^2 - \lambda_{\theta,k}^2\right)$$

where $\lambda_{\theta,k}^2 = v_k^{\mathrm{T}} \Sigma_\theta v_k$ are squared eigenvalues. Let $\sigma_k^2 = \max_{\theta \in \Theta} \sigma_{\theta,k}^2$ and $\Sigma = \sum_k \sigma_k^2 v_k v_k^{\mathrm{T}}$. Then, the mechanism $\mathcal{M}(\mathfrak{Data}) = F(\mathfrak{Data}) + Z$, where $Z \sim \mathcal{N}(0, \Sigma)$, 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

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

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 A.2 are thus satisfied, and it follows that $\mathcal{M}$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ. $\square$

A.2 Adversarial Uncertainty with Directional Assumptions

We briefly discuss how directional assumptions (Section 4.2) can be used in conjunction with adversarial uncertainty (Section 4.3). 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 1. 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 4.2) and Eigenvector Gaussian Mechanism (Section 4.3), this mechanism can also be modified to account for cases where this assumption may not exactly hold.

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

---

1 Set $c = \sqrt{2\ln(1.25/\delta)}$.
2 **for** *each $\theta \in \Theta$* **do**
3   Set $\mu_\theta = \mathbb{E}[F(\mathfrak{Data}) \mid \mathfrak{Data} \sim \theta]$, $\Sigma_\theta = \mathrm{Cov}(F(\mathfrak{Data}) \mid \mathfrak{Data} \sim \theta)$.
4 **end**
5 **for** *each $(\theta_i, \theta_j) \in \Psi$* **do**
6   Set $\alpha_{i,j} = (\mu_{\theta_i} - \mu_{\theta_j})^{\mathrm{T}} v$.
7   Find $\sigma_{i,j}^2 > 0$ such that $\Sigma_{\theta_i} + (\sigma_{i,j}^2 - (\alpha_{i,j} c/\epsilon)^2) v v^{\mathrm{T}} > 0$.
8 **end**
9 Set $\sigma^2 = \max_{(\theta_i, \theta_j) \in \Psi} \sigma_{i,j}^2$.
10 **return** $F(\mathfrak{Data}) + Z$, where $Z \sim \mathcal{N}(0, \sigma^2 v v^{\mathrm{T}})$.

To prove that Algorithm 1 satisfies distribution privacy, we will use the following lemma.

Lemma A.1. *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

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

$\square$

Theorem A.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 $(\theta_i, \theta_j) \in \Psi$. 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 1 satisfies (ϵ, δ)-distribution privacy with respect to Ψ.*

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

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

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,

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

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

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

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

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

Now, since $((\alpha_{i,j}c/\epsilon)^2 vv^{\mathrm{T}} + \eta I)v = ((\alpha_{i,j}c/\epsilon)^2 + \eta)v$, Lemma A.1 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 (1) holds. The conditions of Corollary A.1 are thus satisfied, and it follows that $\mathcal{M}$ satisfies (ϵ, δ)-distribution privacy with respect to Ψ. $\qquad\square$

Note that the condition on line 7 of Algorithm 1 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 1 to the Gaussian variant of the Directional Expected Value Mechanism described in Section 4.2. 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 1 in comparison to the Directional Expected Value Mechanism is then proportional to $(\alpha_{i,j}c/\epsilon)^2 - \sigma_{i,j}^2$.

B ADDITIONAL PROOFS

B.1 Approximate Wasserstein Mechanism

Theorem 3.3. *Suppose that* $\|F(\mathfrak{Data}) - \mathbb{E}[F(\mathfrak{Data})]\|_1 \le c$ *with probability at least* $1 - \delta/2$ *for each* $\mathfrak{Data} \sim \theta \in \Theta$. *Let*

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

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 $R_\theta = \{t : \|t - \mathbb{E}[f_\theta]\|_1 \le c\}$. Then, $\Pr(F(\mathfrak{Data}) \in R_\theta \mid \theta) \ge 1 - \delta/2$ 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

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

By the union bound, we also have

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

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

B.2 Expected Value Mechanism

The proofs of Theorem 4.1 and 4.2 are straight-forward applications of the Laplace and Gaussian Mechanisms from differential privacy (Dwork and Roth, 2014) using the condition that each pair of distributions f_{θ_i}, f_{θ_j} are translations of each other. We provide only the proof of Theorem 4.2, as Theorem 4.1 can similarly be derived by replacing references to the Gaussian Mechanism from differential privacy with the Laplace Mechanism from differential privacy.

Theorem 4.2. *Suppose that* f_{θ_i} *is a translation of* f_{θ_j} *for every pair* $(\theta_i, \theta_j) \in \Psi$. *Let* $\sigma \ge c\Delta_{E,2}(\Psi, F)/\epsilon$, *where* $\epsilon \in (0, 1)$ *and* $c = \sqrt{2\ln(1.25/\delta)}$ *for some* $\delta > 0$. *Then, the mechanism* $\mathcal{M}(\mathfrak{Data}) = F(\mathfrak{Data}) + Z$, *where* $Z \sim \mathcal{N}(0, \sigma^2 I)$, *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 $S \subseteq \mathrm{Range}(\mathcal{M})$,

$$\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_i) = \int_t f_{\theta_i}(t) \Pr(Z + t \in S)\, dt$$

$$= \int_s f_{\theta_j}(s) \Pr(Z + s - b \in S)\, ds.$$

Now, note that $\|b\|_2 = \|\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]\|_2 \le \Delta_{E,2}(\Psi, F)$, and that $\sigma \ge c\Delta_{E,2}(\Psi, F)/\epsilon$. Then, since each component of noise $Z_k \sim \mathcal{N}(0, \sigma^2)$ is independently sampled, the Gaussian Mechanism from differential privacy (Dwork and Roth, 2014) implies that

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

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

$$\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_i) \le \int_s f_{\theta_j}(s) \left(\exp(\epsilon) \Pr(Z + s \in S) + \delta \right)\, ds$$

$$= \exp(\epsilon) \int_s f_{\theta_j}(s) \Pr(Z + s \in S)\, ds + \delta$$

$$= \exp(\epsilon) \Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_j) + \delta.$$

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

B.3 Directional Expected Value Mechanism

The proof of Theorem 4.3 applies ideas from the mechanism described in Section 4.1 while also taking into account the possible directions of the differences of mean vectors $\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]$. Note also that it can readily be modified to use Gaussian noise instead of Laplace noise.

Theorem 4.3. *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 $\mathcal{M}(\mathfrak{Data}) = F(\mathfrak{Data}) + Yv$, where $Y \sim \mathrm{Lap}(\Delta_{E,2}(\Psi, F)/\epsilon)$, 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 $S \subseteq \mathrm{Range}(\mathcal{M})$,

$$\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_i) = \int_t f_{\theta_i}(t) \Pr(Yv + t \in S)\, dt$$

$$= \int_s f_{\theta_j}(s) \Pr(Yv + s - b \in S)\, ds. \tag{2}$$

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

$$\Pr(Yv + s - b \in S) = \Pr(Yv + s - b \in S')$$

$$= \Pr(Yv - b \in S' - s)$$

$$= \Pr(Y - \|b\|_2 \in v^{\mathrm{T}}(S' - s)),$$

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

Now, observe that $\|\|b\|_2\|_1 = \|b\|_2 = \|\mathbb{E}[f_{\theta_i}] - \mathbb{E}[f_{\theta_j}]\|_2 \le \Delta_{E,2}(\Psi, F)$. Since $Y \sim \mathrm{Lap}(\Delta_{E,2}(\Psi, F)/\epsilon)$, the Laplace Mechanism from differential privacy (Dwork et al., 2006) then gives that

$$\Pr(Y - \|b\|_2 \in v^{\mathrm{T}}(S' - s)) \le \exp(\epsilon) \Pr(Y \in v^{\mathrm{T}}(S' - s))$$

$$= \exp(\epsilon) \Pr(Yv + s \in S).$$

Thus, $\Pr(Yv + s - b \in S) \le \exp(\epsilon) \Pr(Yv + s \in S)$ for all $s \in \mathbb{R}^m$. Together with (2), this implies

$$\Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_i) \le \exp(\epsilon) \int_s f_{\theta_j}(s) \Pr(Yv + s \in S)\, ds$$

$$= \exp(\epsilon) \Pr(\mathcal{M}(\mathfrak{Data}) \in S \mid \theta_j).$$

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

B.4 Approximations with Max-Divergence

Theorem 4.6. *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 $S \subseteq \operatorname{Range}(\widetilde{\mathcal{M}})$,

$$\Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}} \in S) \leq \exp(\epsilon) \Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_j}} \in S) + \delta. \tag{3}$$

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 (3), 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. $\qquad\square$

B.5 Approximations with Wasserstein Distance

To prove Theorem 4.7, we first observe that the Wasserstein Mechanism of Section 3 can be viewed as a result on the effect of Laplace noise on the max-divergence between distributions. In particular, Theorem 3.1 may be reinterpreted as follows.

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

We can then use Lemma B.1 to prove Theorem 4.7.

Theorem 4.7. *Suppose that the 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 $S \subseteq \text{Range}(\widetilde{\mathcal{M}})$,

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

In turn, this implies

$$\begin{aligned}
\Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}} + Z' \in S) &= \int_z \Pr(Z' = z) \Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}} \in S - z)\, dz \\
&\le \int_z \Pr(Z' = z)(\exp(\epsilon) \Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}} \in S - z) + \delta)\, dz \\
&\le \exp(\epsilon) \Pr(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_j}} + Z' \in S) + \delta. \quad\quad (4)
\end{aligned}$$

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

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

Hence, by Lemma B.1, we have $D_\infty(\widetilde{\mathcal{M}}_{f_{\theta_i}} + Z' \parallel \widetilde{\mathcal{M}}_{\tilde{f}_{\theta_i}} + Z') \le \lambda$. Similarly, $D_\infty(\widetilde{\mathcal{M}}_{\tilde{f}_{\theta_j}} + Z' \parallel \widetilde{\mathcal{M}}_{f_{\theta_j}} + Z') \le \lambda$. Together with (4), this gives

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

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

We remark that the privacy guarantees in Theorem 4.7 may sometimes be attainable even without applying the additional noise Z'. Observe that the proof of Theorem 4.7 only requires Z' to satisfy $D_\infty^\eta(\mathcal{M}_{f_\theta} + Z' \parallel \mathcal{M}_{\tilde{f}_\theta} + Z') \le \lambda$ and $D_\infty^\eta(\mathcal{M}_{\tilde{f}_\theta} + Z' \parallel \mathcal{M}_{f_\theta} + Z') \le \lambda$. 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.
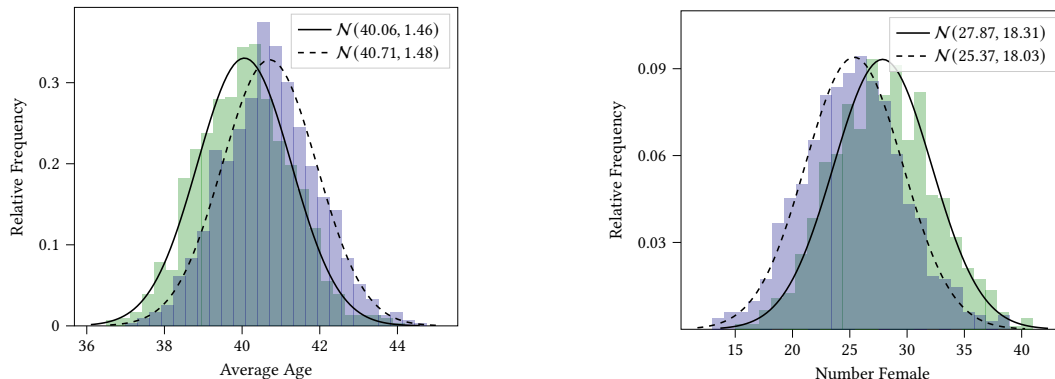
Figure 5: Gaussian approximations for selected statistics of interest over randomly sampled subsets of the Adult dataset. Here, the subsets are sampled to have $n = 100$ records with fixed proportions p_I of individuals having income $> \$50K$, where green and purple histograms are used for $p_I = 0.45$ and $p_I = 0.55$, respectively.

C ADDITIONAL EXPERIMENTAL DETAILS

C.1 Gaussian Modeling

To approximate each distribution f_θ with a Gaussian distribution, we used the standard parameter estimation approach Forbes et al. (2011) of randomly sampling from f_θ, then computing the sample mean μ_θ and covariance matrix Σ_θ to yield an approximation $\mathcal{N}(\mu_\theta, \Sigma_\theta)$. Specifically, for each distribution $\theta \in \Theta$, we randomly sampled a subset of 100 records from the Adult dataset having the value of the global sensitive property p_I or p_W as specified by θ. We then computed the query F over this subset, producing one sample value for f_θ. Repeating this process 1000 times yielded 1000 samples from which we computed a mean vector μ_θ and covariance matrix Σ_θ.

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 4.4.

C.2 Further Attack Results

Table 3 shows how the accuracy of the attack in Section 5.3 changes as we apply Gaussian variants of the Expected Value Mechanism with varying values of ϵ and δ. We note that varying δ between 0.0001 and 0.01 does not substantially impact the measured attack accuracies in comparison to varying ϵ. This is expected as our Gaussian variants of the Expected Value Mechanism roughly apply noise scaled to $c\Delta_{E,2}(\Psi, F)/\epsilon$, where $c = \sqrt{2\ln(1.25/\delta)}$ grows sub-logarithmically as δ decreases. Nonetheless, we remark that there is still an important theoretical distinction between (ϵ, δ)-distribution privacy for different values of δ, even if not evident against this specific property inference attack.

Table 3: 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