Bayesian Pseudo Posterior Mechanism under Differential Privacy

Terrance D. Savitsky Savitsky.Terrance@bls.gov
Office of Survey Methods Research
U.S. Bureau of Labor Statistics
2 Massachusetts Ave NE
Washington, DC 20212, USA

Matthew R. Williams mrwillia@nsf.gov
National Center for Science and Engineering Statistics
National Science Foundation
Alexandria, VA 22314, USA

Jingchen Hu jihu@vassar.edu
Vassar College
124 Raymond Ave, Box 27
Poughkeepsie, NY 12604, USA

Editor:

Abstract

We propose a Bayesian pseudo posterior mechanism to generate record-level synthetic datasets equipped with a differential privacy (DP) guarantee from any proposed synthesis model. The pseudo posterior mechanism employs a data record-indexed, risk-based weight vector with weights $\in [0, 1]$ to surgically downweight high-risk records for the generation and release of record-level synthetic data. The differentially private pseudo posterior synthesizer constructs weights using Lipschitz bounds for a log-pseudo likelihood utility for each data record, which provides a practical, general formulation for using weights based on record-level sensitivities that we show achieves dramatic improvements in the DP expenditure as compared to the unweighted posterior mechanism. By selecting weights to remove likelihood contributions with non-finite log-likelihood values, we achieve a local privacy guarantee at every sample size. We compute a local sensitivity specific to our Consumer Expenditure Surveys dataset for family income, published by the U.S. Bureau of Labor Statistics, and reveal mild conditions that guarantee its contraction to a global sensitivity result over the space of databases. We further employ a censoring mechanism to lock-in a local result with desirable risk and utility performances to achieve a global privacy result as an alternative to relying on asymptotics. We show that utility is better preserved for our pseudo posterior mechanism as compared to the exponential mechanism (EM) estimated on the same non-private synthesizer due to the use of targeted downweighting. Our results may be applied to any synthesizing model envisioned by the data disseminator in a computationally tractable way that only involves estimation of a pseudo posterior distribution for parameter(s) θ, unlike recent approaches that use naturally-bounded utility functions under application of the EM.

Keywords: Differential privacy, Pseudo posterior, Pseudo posterior mechanism, Synthetic data

1. Introduction

Privacy protection is an important research topic, which attracts attention from government statistical agencies and private companies alike. A popular approach focuses on encoding privacy protection to a summary statistic composed from record-level data, through the addition of noise proportional to the "sensitivity", ϵ, of the statistic, defined as the supremum of the change in value of the statistic from the inclusion or exclusion of a single data record over the space of databases. Dwork et al. (2006) construct a mechanism that employs a Laplace-distributed perturbation of a target statistic. The mechanism, which produces the resultant statistic, achieves a privacy guarantee under the differential privacy (DP) framework. The guarantee is represented by a budget, some of which is expended for each query a user makes through the mechanism to the underlying, closely-held (by the statistical agency) database.

A related approach to privacy protection is the release of a synthetic record-level database. This approach replaces the closely-held (by the statistical agency) database with a synthetically generated record-level database. The synthetic database is released to the public who would use it to conduct any analyses of which they would conceive for the real, confidential record-level data. As a result of releasing a synthetic database encoded with privacy protection, the synthetic data approach replaces multiple queries performed on a summary statistic with the publication of the synthetic database, such that the synthetic data approach is independent of the specific queries performed by users or putative intruders.

Dimitrakakis et al. (2017) demonstrate theoretical results for the Bayesian posterior distribution, which may be employed as a mechanism for synthetic data generation; specifically, if the log-likelihood is Lipschitz continuous with bound Δ, then the posterior mechanism achieves an $\epsilon = 2\Delta-$DP guarantee for each posterior draw of θ, the model parameter(s); however, Dimitrakakis et al. (2017) acknowledge that computing a finite Δ, in practice, under the use of the log-likelihood is particularly difficult for an unbounded parameter space. They specify relatively simple Bayesian probability models where the Lipschitz bound is analytically available. Even in this simple model setting Dimitrakakis et al. (2017) require truncation of the support of the prior distribution to achieve a finite Δ. Relatively simply-constructed differentially private Bayesian synthesizers are similarly proposed by Machanavajjhala et al. (2008); Abowd and Vilhuber (2008); McClure and Reiter (2012); Bowen and Liu (2016). The utility performance to preserve the real data distribution in the published synthetic data of these simple posterior mechanisms under a truncated prior support may be severely compromised by truncation and over-smoothing (induced by simple, parametric prior distributions).

A common approach for generating parameter draws for θ is the exponential mechanism (EM) of McSherry and Talwar (2007), which inputs a non-private mechanism for θ and generates θ in such a way that induces a DP guarantee on the overall mechanism.

Definition 1 *The exponential mechanism releases values of θ from a distribution proportional to,*

$$\exp\left(\frac{\epsilon\, u(\mathbf{x},\theta)}{2\Delta_u}\right), \tag{1}$$

where $u(\mathbf{x},\theta)$ is a utility function, $\Delta_u = \sup_{\mathbf{x}\in\mathcal{X}^n}\ \sup_{\mathbf{x},\mathbf{y}:\delta(\mathbf{x},\mathbf{y})=1}\ \sup_{\theta\in\Theta} |u(\mathbf{x},\theta) - u(\mathbf{y},\theta)|$ is the sensitivity, defined globally over $\mathbf{x} = (x_1,\ldots,x_n) \in \mathcal{X}^n$, the $\sigma-$algebra of datasets,

$\mathbf{x}$, *governed by product measure,* P_{θ_0}*;* $\delta(\mathbf{x}, \mathbf{y}) = \#\{i : x_i \neq y_i\}$ *is the Hamming distance between* $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$*. Each draw of* θ *from the exponential mechanism satisfies* $\epsilon-DP$*, where* ϵ *is a budget target supplied by the publishing statistical agency.*

The EM inputs a global utility function and its sensitivity constructed as the supremum of the utility over the space of data, $\mathcal{X}^n$, and, simultaneously, the parameter space, Θ. Wasserman and Zhou (2010) and Snoke and Slavkovic (2018) construct utility functions based on the real and synthetic datasets (e.g., the Kolmogorov-Smirnov distance between the empirical distributions of the real and synthetic datasets) that are *naturally* bounded over all $\mathbf{x} \in \mathcal{X}^n$, resolving the challenge of using the potentially unbounded log-likelihood as the utility function. Although the use of a naturally bounded utility resolves the issue of truncating the data and parameter spaces, there is a *large*, and perhaps *intractable*, computational cost to the use of these naturally bounded utilities to draw samples of θ from the distribution constructed from the EM; for example, Snoke and Slavkovic (2018) must compute their $pMSE$ utility statistic multiple times for each proposed value, $\hat{\theta}_l$ $(l = 1, \cdots, L)$, under a Metropolis-Hastings algorithm used to draw samples under the EM. Furthermore, Snoke and Slavkovic (2018) assume the existence of some synthesizing distribution, $g(\hat{\theta})$, from which to draw synthetic data, needed to compute their $pMSE$. In practice, g will be defined as the posterior predictive distribution, $g(\mathbf{X} \mid \mathbf{x}, \hat{\theta}_l)$, which means the posterior distribution must be repeatedly estimated for *each* draw from of θ from the EM.

This paper focuses on formalizing and extending the pseudo posterior synthesizer in Hu and Savitsky (2019) as an alternative mechanism (to the EM) as a practical means of achieving a global Lipschitz without parameter truncation under richly-parameterized probability models, in a fashion that produces synthetic data that well-preserves the properties of the closely-held, real dataset distribution. Hu and Savitsky (2019) design a record-indexed weight $\alpha_i \in [0, 1]$, which is inversely proportional to their construction for the identification risk probability of record, i; a data record that expresses a relatively high probability of identification disclosure will receive a likelihood weight, α_i, that is closer to 0, while a data record with a low disclosure probability will receive a likelihood weight, α_i, that is closer to 1. The vector weights $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$ are subsequently applied to the likelihood function of all n records to form the pseudo posterior,

$$\xi^{\boldsymbol{\alpha}}(\theta \mid \mathbf{x}, \gamma) \propto \left[\prod_{i=1}^{n} p(x_i \mid \theta)^{\alpha_i}\right] \xi(\theta \mid \gamma), \tag{2}$$

where θ denotes the model parameters, γ denotes the model hyperparameters and $\xi(\cdot)$ denotes the prior distribution. This construction employs a data record-indexed, risk-based weight vector with weights $\in [0, 1]$ to surgically downweight high-risk records in estimation of a pseudo posterior distribution for θ, subsequently used to generate and release a synthetic record-level database. The authors show that this selective downweighting of records reduces the average of by-record risks as compared to an unweighted synthesis, while inducing only a minor reduction in utility. Hu and Savitsky (2019) base their risk measure on a calculated probability of identification for a record. They cast a radius around the true data value for each record and count the number of record values that lie outside of the radius, which directly measures the extent that the target record is isolated and, therefore, easier for an intruder to discover by random guessing. While their risk measure appeals to intuition, it

is based on an assumption about the behavior of a putative intruder. By contrast, the DP framework makes no explicit assumptions about the behavior or knowledge of an intruder.

In this work, we utilize the pseudo posterior synthesizer to formulate a private mechanism under the use of the log-pseudo likelihood to construct a global Lipschitz bound, which extends Dimitrakakis et al. (2017) and provides a practical, general approach that we show in the sequel achieves dramatic improvements in the DP guarantee as compared to the unweighted, non-private (posterior) synthesizer. We demonstrate that our pseudo posterior mechanism produces a finite global Lipschitz at every n that, in turn, provides a global DP guarantee. We compute a local sensitivity specific to our application to a Consumer Expenditure Surveys (CE) sample and reveal conditions that guarantee its contraction to a global Lipschitz, Δ, over all $\mathbf{x} \in \mathcal{X}^n$ (across all potential datasets) as n increases. An alternative non-asymptotic censoring formulation is developed that allows the "locking in" of a local result on a specific dataset to provide a formal DP guarantee over the space of datasets. Our results may be applied to any synthesizing mechanism envisioned by the data disseminator in a computationally tractable way that only involves a routine estimation of a pseudo posterior distribution for θ.

The remainder of the paper is organized as follows: Section 2 generalizes an unweighted Lipschitz assumption to a weighted Lipschitz assumption that guarantees a DP privacy result for our proposed pseudo posterior mechanism and we demonstrate how our pseudo posterior mechanism can be used to make a global DP guarantee. We present an asymptotic result on the contraction of a local Lipschitz to a global Lipschitz. In Section 3, we describe the computation details to produce a matrix of (absolute values for) log-likelihoods estimated for the n records and S parameter draws taken from the unweighted posterior mechanism and their subsequent use to formulate a vector of record-indexed weights, $\boldsymbol{\alpha}$. We then discuss the procedure to use the $\boldsymbol{\alpha}$ to estimate the pseudo posterior distribution and computation of the local Lipchitz bound for the pseudo posterior mechanism. This section additionally enumerates the connection between the scalar-weighted pseudo posterior mechanism and the EM. Section 4 formulates a non-asymptotic result that censors the log-pseudo likelihood at a threshold chosen based on a local Lipschitz bound for an observed database to lock-in that local bound as a global bound. This section presents a simulation of our non-asymptotic censoring approach under different values. Section 5 focuses on our application to synthesizing the family income in the CE sample, and presents the risk and utility profiles of differentially private synthetic data generated under the proposed pseudo posterior mechanism, compared to other competing methods. We conclude with a discussion in Section 6.

2. Differential Privacy for the Pseudo Posterior

In this section, we generalize the connection between achieving a global Lipschitz bound and a DP guarantee from the unweighted posterior distribution of Dimitrakakis et al. (2017), on the one hand, to the risk-weighted, pseudo posterior distribution, which defines our privacy mechanism, on the other hand. We further re-purpose a result from Wasserman and Zhou (2010) to extend a DP guarantee to the pseudo posterior predictive mechanism for generating synthetic data that is based on integrating with respect to the privately guaranteed pseudo posterior distribution mechanism (used to generate the model parameters). After

having shown that achievement of a global Lipschitz guarantees a DP result for a pseudo posterior mechanism, we discuss constructing by-record weights used in our pseudo posterior mechanism that are designed to be inversely proportional to the log-likelihood utilities computed over the parameter space. We demonstrate that our procedure for generating by-record weights guarantees a global Lipschitz bound over the space of databases and, hence, a global DP guarantee. Our result is non-asymptotic in that it applies to all sample sizes, n, but we do *not* know the global Lipschitz, Δ. As a result, our computation of a Lipschitz bound, $\Delta_{\mathbf{x}}$, is a local result (based on an observed dataset and sampled parameter space). We formally discuss the asymptotic behavior of the *local* Lipschitz bound as n increases, demonstrating its contraction on a *global* Lipschitz bound.

2.1 Preliminaries

We begin by constructing the probability space, (Θ, β_Θ), equipped with prior distribution, $\xi(\theta)$. Observe a database sequence, $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ under $x_1, \ldots, x_n \overset{\text{ind}}{\sim} P_{\theta_0}$, for some $\theta_0 \in \Theta$, we formulate the pseudo likelihood,

$$p_\theta^{\boldsymbol{\alpha}}(\mathbf{x}) = \prod_{i=1}^{n} p_{\theta_i}(x_i)^{\alpha_i(\mathbf{x})}, \tag{3}$$

for each $\theta \in \Theta$. The pseudo likelihood exponentiates likelihood contributions by $\boldsymbol{\alpha}(\mathbf{x}) = (\alpha_1(\mathbf{x}), \ldots, \alpha_n(\mathbf{x}))$, where $\alpha_i(\mathbf{x}) \in [0, 1]$ denote weights that are constructed to be inversely proportional to the local identification risk for each observed dataset record, and are used to selectively downweight the likelihood contributions for records that express relatively high identification disclosure risks. Under the DP paradigm for estimating risk, we formulate the $\boldsymbol{\alpha}-$weighted log-pseudo likelihood,

$$f_\theta^{\boldsymbol{\alpha}}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i(\mathbf{x}) \log p_\theta(x_i) \tag{4}$$

that we use to estimate the pseudo posterior mechanism. The vector of weights, $\boldsymbol{\alpha}$, is constructed to be inversely proportional to the local Lipschitz bound, $\sup_{\theta\in\Theta}|f_\theta(\mathbf{x}) - f_\theta(\mathbf{y})|$ for an observed $\mathbf{x}$ (where $f = \log p_\theta$) estimated from the unweighted posterior mechanism. Once we have computed the $\boldsymbol{\alpha}$ using the unweighted posterior mechanism and formed $f_\theta^\alpha(\mathbf{x})$, we evaluate the local Lipschitz bound for each $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n$ as, $\sup_{\theta\in\Theta}|f_\theta^\alpha(\mathbf{x}) - f_\theta^\alpha(\mathbf{y})|$ for all $\mathbf{y} \in \mathcal{X}^n$. We demonstrate in the sequel that we achieve a dramatic reduction in the local bound for the pseudo posterior mechanism than the unweighted posterior mechanism.

We account for the dependence of α_i on $\mathbf{x}$, which generalizes Bhattacharya et al. (2019), in assessing the frequentist properties of our Bayesian estimator since under *frequentist* consistency, the $\mathbf{x}$ are random with respect to P_θ (for fixed θ), so taking probabilities and expectations with respect to P_θ requires us to address the dependence of α_i on $\mathbf{x}$ to construct the contraction rate for correctness and thoroughness. We drop the notation denoting the explicit dependence of $\alpha_i(\mathbf{x})$ for most of the paper and just use α_i for readability when the context is clear.

Given the prior and pseudo likelihood, we construct the pseudo posterior distribution,

$$\xi^{\boldsymbol{\alpha}}(B \mid \mathbf{x}) = \frac{\int_{\theta\in B} p_\theta^{\boldsymbol{\alpha}}(\mathbf{x}) d\xi(\theta)}{\phi^{\boldsymbol{\alpha}}(\mathbf{x})} = \frac{\int_{\theta\in B} e^{-r_{n,\boldsymbol{\alpha}(\theta,\theta^*)}} d\xi(\theta)}{\int_{\theta\in\Theta} e^{-r_{n,\boldsymbol{\alpha}(\theta,\theta^*)}} d\xi(\theta)} \tag{5}$$

where $\phi^{\boldsymbol{\alpha}}(\mathbf{x}) \stackrel{\Delta}{=} \int_{\theta\in\Theta} p_\theta^{\boldsymbol{\alpha}}(\mathbf{x}) d\xi(\theta)$ normalizes the pseudo posterior distribution and $r_{n,\boldsymbol{\alpha}}(\theta,\theta^*) = \sum_{i=1}^n \alpha_i \log\left\{p_{\theta_i^*}(x_i)/p_{\theta_i}(x_i)\right\}$, which is a generalization of the definition from Bhattacharya et al. (2019) to incorporate risk-adjusted weights, $(\alpha_i)_{i=1,\cdots,n}$.

Since our pseudo posterior formulation induces misspecification, we allow the true generating parameters, θ_0, to lie outside the parameter space, Θ. We will show in the sequel that our model contracts on $\theta^* \in \Theta$ in $P_{\theta_0}-$probability, where θ^* is the point that minimizes the Kullback-Liebler (KL) divergence from P_{θ_0}; that is,

$$\theta^* := \arg\min_{\theta\in\Theta} D\left(p_\theta, p_{\theta_0}\right), \tag{6}$$

where $D(p,q) = \int p \log(p/q) d\mu$ for dominating measure, μ.

Our asymptotic result on the contraction in $P_{\theta_0}-$probability relies on bounding the $\boldsymbol{\alpha}-$Rényi divergence measure,

$$D^{(n)}_{\theta_0,\boldsymbol{\alpha}}(\theta,\theta^*) = \sum_{i=1}^n D_{\theta_0,\alpha,i}(\theta,\theta^*) = \sum_{i=1}^n \frac{1}{\alpha_i - 1}\log\left\{A_{\theta_0,\alpha,i}(\theta,\theta^*)\right\}, \tag{7}$$

where $A_{\theta_0,\alpha,i}(\theta,\theta^*) = \int \left(\frac{p_{\theta_i}}{p_{\theta_i^*}}\right)^{\alpha_i} p_{\theta_0,i} d\mu_i$ under dominating measure μ_i is defined as the $\boldsymbol{\alpha}-$affinity for observation, x_i, such that $A^{(n)}_{\theta_0,\boldsymbol{\alpha}}(\theta,\theta^*) = \prod_{i=1}^n A_{\theta_0,\alpha,i}(\theta,\theta^*)$, the $\boldsymbol{\alpha}-$affinity for the product measure space.

The Hamming distance between databases that we use to estimate the local Lipschitz bound, $\Delta_{\mathbf{x}}$, is defined based on the number of data records excluded from a database, $\mathbf{x}$.

Definition 2 *(Hamming distance) Given databases* $\mathbf{x} = (x_1,\cdots,x_n)$ *and* $\mathbf{y} = (y_1,\cdots,y_n)$*, let* $\delta(\mathbf{x},\mathbf{y})$ *denote the Hamming distance between* $\mathbf{x}$ *and* $\mathbf{y}$*:*

$$\delta(\mathbf{x},\mathbf{y}) = \#\{i : x_i \neq y_i\} \tag{8}$$

2.2 Main Results

Our task is to specify assumptions that guarantee our pseudo posterior mechanism achieves an $\epsilon-$expenditure under the DP framework. We present a collection of related results in this section with all of the associated proofs in Appendix A.

2.2.1 Assuming a Global Lipschitz Bound

In this section and corresponding sections in Appendix A, we use the explicit notation $\boldsymbol{\alpha}(\mathbf{x})$. We begin by extending the definition of DP from Dimitrakakis et al. (2017) to our $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism.

Definition 3 *(Differential Privacy)*

$$\sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}^n:\delta(\mathbf{x},\mathbf{y})=1} \sup_{B\in\beta_\Theta} \frac{\xi^{\boldsymbol{\alpha}(\mathbf{x})}(B \mid \mathbf{x})}{\xi^{\boldsymbol{\alpha}(\mathbf{y})}(B \mid \mathbf{y})} \leq e^\epsilon,$$

which limits the change in the pseudo posterior distribution over all sets, $B \in \beta_\Theta$ (i.e. β_Θ is the $\sigma-$algebra of measurable sets on Θ), from the inclusion of a single record. Although the pseudo posterior distribution mass assigned to B depends on $\mathbf{x}$, the ϵ expenditure is defined as the supremum over all $\mathbf{x} \in \mathcal{X}^n$ and for all $\mathbf{y} \in \mathcal{X}^n : \delta(\mathbf{x}, \mathbf{y}) = 1$.

Our main assumption extends Dimitrakakis et al. (2017) to bound the *log-pseudo likelihood ratio*, uniformly, for all databases, $\mathbf{y} \in \mathcal{X}^n$ that are at a Hamming-1 distance (i.e. $\delta(\mathbf{x}, \mathbf{y}) = 1$) over all $\mathbf{x} \in \mathcal{X}^n$ and $\theta \in \Theta$. The uniform bound defines a maximum sensitivity in the log-pseudo likelihood from the inclusion of a record (at a Hamming-1 distance from each database in the space of databases). Our intuition that the magnitude of this sensitivity for the log-pseudo likelihood ratio is directly tied to the resulting sensitivity of the pseudo posterior, ϵ, that determines the DP expenditure is confirmed in two results below.

Assumption 1 *(Lipschitz continuity)*
Fix a $\theta \in \Theta$ and define a vector-valued mapping $\boldsymbol{\alpha}^*(\cdot)$: $\mathcal{X}^n \rightarrow [0,1]^n$ and construct the Lipschitz function of θ over the space of databases,

$$\ell^{\boldsymbol{\alpha}^*}(\theta) \triangleq \inf \left\{ w : \left| f_\theta^{\boldsymbol{\alpha}^*(\mathbf{x})}(\mathbf{x}) - f_\theta^{\boldsymbol{\alpha}^*(\mathbf{y})}(\mathbf{y}) \right| \leq w, \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}^n : \delta(\mathbf{x}, \mathbf{y}) = 1 \right\}.$$

Assumption 1 restricts Θ such that the Lipschitz function of θ is uniformly bounded from above,

$$\ell^{\boldsymbol{\alpha}}(\theta) \leq \Delta_{\boldsymbol{\alpha}^*}; \;\; \theta \in \Theta; \;\; \boldsymbol{\alpha}(\mathbf{z}) \leq \boldsymbol{\alpha}^*(\mathbf{z}), \forall \mathbf{z} \in \mathcal{X}^n$$

where the last inequality is elementwise bounding of all potential mappings $\boldsymbol{\alpha}(\mathbf{z})$: $\alpha_i(\mathbf{z}) \leq \alpha_i^*(\mathbf{z})$ for each coordinate $i = 1, \ldots, n$. We note that the subscripting of Δ with $\boldsymbol{\alpha}^*$ is a notational device that denotes the (scalar) Lipschitz bound computed using the log-pseudo likelihood, $f_\theta^{\boldsymbol{\alpha}^*(\mathbf{x})}(\mathbf{x})$ as contrasted with Δ computed using the unweighted posterior mechanism. We further note that the assumption from Dimitrakakis et al. (2017) is equivalent to using $\boldsymbol{\alpha}^* = \mathbf{1}$ and therefore $\ell^{\boldsymbol{\alpha}}(\theta) \leq \Delta_{\boldsymbol{\alpha}} \leq \Delta$.

Our next result connects the Lipschitz bound, $\Delta_{\boldsymbol{\alpha}}$, for the log-pseudo likelihood to the supremum over $\mathbf{x} \in \mathcal{X}^n$ and for each $\mathbf{x}$, those $\mathbf{y} \in \mathcal{X}^n : \delta(\mathbf{x}, \mathbf{y}) = 1$ of the KL divergence between the posterior densities (given $\mathbf{x}$ versus $\mathbf{y}$) from the inclusion of a database record.

Theorem 4 *$\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}^n : \delta(\mathbf{x}, \mathbf{y}) = 1$ and $\boldsymbol{\alpha}(\cdot)$ with $\Delta_{\boldsymbol{\alpha}} > 0$ satisfying Assumption 1,*

$$\sup_{\mathbf{x}, \mathbf{y} \in \mathcal{X}^n : \delta(\mathbf{x}, \mathbf{y}) = 1} D_{KL} \left[(\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\cdot \mid \mathbf{x}) \parallel \xi^{\boldsymbol{\alpha}(\mathbf{y})}(\cdot \mid \mathbf{y})) \right] \leq 2\Delta_{\boldsymbol{\alpha}}, \tag{9}$$

where $D_{KL}((P \parallel Q)) = \int_{\mathcal{X}^n} \ln \frac{dP}{dQ} dP$.

Our next result directly connects the Lipschitz bound, $\Delta_{\boldsymbol{\alpha}}$, for the log-pseudo likelihood of Assumption 1 to resulting DP expenditure, $\epsilon = 2\Delta_{\boldsymbol{\alpha}}$, for each draw of θ from the pseudo posterior distribution.

Theorem 5 *$\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}^n : \delta(\mathbf{x}, \mathbf{y}) = 1, B \in \beta_\Theta$ (where β_Θ is the $\sigma-$algebra of measurable sets on Θ) under $\boldsymbol{\alpha}(\cdot)$ with $\Delta_{\boldsymbol{\alpha}} > 0$ satisfying Assumption 1:*

$$\sup_{B \in \beta_\Theta} \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{X}^n : \delta(\mathbf{x}, \mathbf{y}) = 1} \frac{\xi^{\boldsymbol{\alpha}(\mathbf{x})}(B \mid \mathbf{x})}{\xi^{\boldsymbol{\alpha}(\mathbf{y})}(B \mid \mathbf{y})} \leq \exp(2\Delta_{\boldsymbol{\alpha}}), \tag{10}$$

i.e. the pseudo posterior $\xi^{\boldsymbol{\alpha}(\mathbf{z})}(\cdot \mid \mathbf{z})$ is $2\Delta_{\boldsymbol{\alpha}}-$DP.

Our next result extends our DP guarantee from posterior draws of θ for models that satisfy Assumption 1 to draws of synthetic data, $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_m)$, constructed from the model posterior predictive distribution, which is the focus of our pseudo posterior mechanism.

Lemma 6 *Define* $P^{\boldsymbol{\alpha}(\mathbf{x})}(\boldsymbol{\zeta} \in C \mid \mathbf{x}) = \int P(\boldsymbol{\zeta} \in C \mid \theta, \mathbf{x}) d\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\theta \mid \mathbf{x})$ *as the pseudo posterior predictive probability mass for* $\boldsymbol{\zeta}$ *in set* $C \in \mathcal{A}^n$ *(the* $\sigma-$*algebra of sets for* $\mathcal{X}^n$*), constructed from our pseudo posterior model for* θ *that satisfies DP with expenditure,* ϵ*. Let* $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_M)$ *be* M *independent draws from* $P^{\boldsymbol{\alpha}(\mathbf{x})}(\boldsymbol{\zeta} \in C \mid \mathbf{x})$*. This defines a mechanism for* $\boldsymbol{\zeta}$ *that satisfies DP with expenditure* ϵ *for any* $M \leq n$*.*

2.2.2 Achieving a Global Lipschitz Bound via Weighting

The above results, together, convey that if the pseudo log-likelihood, $f_\theta^{\boldsymbol{\alpha}}(\mathbf{x})$, is Lipschitz $\Delta_{\boldsymbol{\alpha}}$, our pseudo posterior mechanism provides a $2\Delta_{\boldsymbol{\alpha}}-$DP expenditure for *each* draw of a synthetic database. To satisfy Assumption 1 for *any* likelihood, the procedure for implementing our pseudo posterior mechanism sets weights, $(\alpha_i)_{i=1,\ldots,n}$, to be inversely proportional to the supremum of the likelihood values computed from the non-differentially private, unweighted mechanism over $\theta \in \Theta$ for each $x_1, \ldots, x_n$ for *all* possible observed $\mathbf{x} \in \mathcal{X}^n$. Evaluating the ratio, $\frac{p_\theta(\mathbf{x})}{p_\theta(\mathbf{y})}$ for all databases, $\mathbf{y} : \delta(\mathbf{x}, \mathbf{y}) = 1$ for our observed $\mathbf{x}$, we divide the full data likelihood by a leave-one-out likelihood for each $i \in (1, \ldots, n)$, such that the ratio simplifies to the individual likelihood values, $p_\theta^{\alpha_i}(x_i)$. Therefore, achieving the $\Delta_{\boldsymbol{\alpha}}$ bound for the log-pseudo likelihood ratios reduces to bounding the log-pseudo likelihood values for the individual data component contributions. Our weighting scheme is constructed such that for any log-likelihood contribution that is non-finite (violating Assumption 1), the associated weight is set to 0, which removes the log-likelihood contribution for these records from our pseudo posterior mechanism. We formalize the weighting scheme that characterizes our pseudo posterior mechanism in the assumption, below.

Assumption 2 *(Risk-based Weighting for Pseudo Posterior Mechanism)*
Fix an n*. Let* $m(\cdot)$ *be a monotonically decreasing scalar function* $m : [0, \infty) \rightarrow [0, 1]$ *such that* $m(0) = 1$*, and* $m(\infty) = 0$*. For every* $\mathbf{x} \in \mathcal{X}^n$ *choose a mapping* $\boldsymbol{\alpha}(\cdot)$ *such that*

$$\alpha_i = m\left(\sup_{\theta \in \Theta} |f_\theta(x_i)|\right), \tag{11}$$

where $f_\theta(x_i)$ *is computed from the unweighted, non-differentially private synthesizer. Under this procedure for selecting risk-based weights,* α_i*,* $i = 1, \ldots, n$*, if* $f_\theta(x_i)$ *is non-finite for any* x_i *and value of* $\theta \in \Theta$*,* α_i *is set to* $m(\infty) = 0$*, which removes the contribution of database record,* i*, from the pseudo likelihood of Equation (3) used to formulate the pseudo posterior mechanism of Equation (5).*

The mapping $m(\cdot)$ in Assumption 2 includes threshold ($m(z) = \mathbf{1}_{\{z < z^*\}}$) as well as smooth functions ($m(z) = (z+1)^{-1}$), providing the us flexibility for how to implement the weighting in practice. Since we remove the likelihood contributions for all database records with non-finite log-likelihoods by setting their associated weights in our pseudo posterior mechanism to $m(\infty) = 0$, our mechanism is guaranteed to satisfy Assumption 1 and thus be globally differentially private (i.e. a finite global budget ϵ exists). This is a non-asymptotic

result at every n; however we want to *estimate* the global $\Delta_{\boldsymbol{\alpha}}$ (and, therefore, ϵ), rather than simply knowing it exists.

To the extent that a given local database, $\mathbf{x}^{'}$, contains relatively many records, i, such that Equation (11) is non-finite, more of the likelihood contributions for those records will be removed from the computation of the pseudo posterior in Equation (5), with the result that prior smoothing will induce more distortion in the resulting synthetic data. This greater degree of smoothing will, in turn, reduce the utility of the synthetic dataset and also increase the privacy protection, so the local $\Delta_{\boldsymbol{\alpha},\mathbf{x}'}$ will be relatively small. By contrast, to the extent a local database, $\mathbf{x}^{''}$, contains few-to-no non-finite likelihood log-likelihood values, the resulting local Lipschitz bound, $\Delta_{\boldsymbol{\alpha},\mathbf{x}''}$, will be larger since the risk and utility would be relatively higher such that the local $\Delta_{\boldsymbol{\alpha},\mathbf{x}''}$ would be closer to the global $\Delta_{\boldsymbol{\alpha}}$.

The following result allows us to simplify our estimation of $\Delta_{\boldsymbol{\alpha}}$ by using the leave-one-out method to estimate an alternative bound $\Delta_{\boldsymbol{\alpha}} \leq \Delta'_{\boldsymbol{\alpha}}$.

Lemma 7 *$\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}^n : \delta(\mathbf{x}, \mathbf{y}) = 1$ and $\boldsymbol{\alpha}(\cdot)$ with $\Delta_{\boldsymbol{\alpha}} > 0$ satisfying Assumption 1, denote the leave-one-out vector $\mathbf{x}_{-i} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_n\}$ and construct the corresponding function of θ over the space of databases,*

$$\begin{aligned} \ell^{\boldsymbol{\alpha}}_{loo}(\theta) &\stackrel{\Delta}{=} \inf\left\{w : \left|f_\theta^{\boldsymbol{\alpha}(\mathbf{x})}(\mathbf{x}) - f_\theta^{\boldsymbol{\alpha}(\mathbf{x}_{-i})}(\mathbf{x}_{-i})\right| \leq w, \forall \mathbf{x} \in \mathcal{X}^n, i \in 1, \ldots, n\right\} \\ &= \inf\left\{w : \left|f_\theta^{\alpha(x_i)}(x_i)\right| \leq w, \forall x_i \in \mathbf{x}\right\}. \end{aligned} \tag{12}$$

Then

$$\Delta_{\boldsymbol{\alpha}} = \sup_\theta\{\ell^{\boldsymbol{\alpha}}(\theta)\} \leq \sup_\theta\{\ell^{\boldsymbol{\alpha}}_{loo}(\theta)\} = \Delta'_{\boldsymbol{\alpha}}.$$

Since we bound the supremum of $\ell^{\boldsymbol{\alpha}}(\theta)$ from Assumption 1 from above by the supremum of $\ell^{\boldsymbol{\alpha}}_{loo}(\theta)$, we may thus simplify our search for a bound for $\Delta_{\boldsymbol{\alpha}}$ to the leave-one-out neighborhood and look at the magnitude of $\left|f_\theta^{\alpha(x_i)}(x_i)\right|$ rather than search across the Hamming-1 neighborhood and compare the magnitude of $\left|f_\theta^{\boldsymbol{\alpha}(\mathbf{x})}(\mathbf{x}) - f_\theta^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y})\right|$.

2.2.3 Asymptotic Convergence of Local Lipschitz to Global Lipschitz

Although our DP result is non-asymptotic for every n, we do not know the global $\Delta_{\boldsymbol{\alpha}}$. We employ asymptotics to identify the global Lipschitz bound, $\Delta_{\boldsymbol{\alpha}}$. We develop a contraction result for any $\boldsymbol{\alpha}-$weighted pseudo distribution to demonstrate under a set of conditions that convergence of the pseudo posterior distribution leads to asymptotic convergence of the local Lipschitz bound, $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$, to the global bound, $\Delta_{\boldsymbol{\alpha}}$ in $P_{\theta_0}-$probability for n sufficiently large; in particular, the posterior probability of the $\boldsymbol{\alpha}-$Rényi distance between $\theta \in \Theta$ and the point θ^* limits to 0 at a rate that is a function of n for any weighting scheme, $\boldsymbol{\alpha}(\mathbf{x})$, where the construction of $\boldsymbol{\alpha}$ depends on the observed data, $\mathbf{x}$, as does ours. We require the following two conditions to achieve contraction of the local $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$ to the global $\Delta_{\boldsymbol{\alpha}}$:

Assumption 3 *(Prior mass covering truth) We construct a KL neighborhood of θ^* with radius, η,with,*

$$B_n\left(\theta^*, \eta; \theta_0\right) = \Bigg\{ \theta \in \Theta : \sum_{i=1}^{n} \int p_{\theta_0,i} \log\left(p_{\theta_i^*}/p_{\theta_i}\right) d\mu_i \leq n\eta^2, \\ \sum_{i=1}^{n} \int p_{\theta_0,i} \log^2\left(p_{\theta_i^*}/p_{\theta_i}\right) d\mu_i \leq n\eta^2 \Bigg\} \quad (13)$$

Restrict the prior, ξ, to place positive probability on this KL neighborhood,

$$\xi\left(B_n\left(\theta^*, \eta; \theta_0\right)\right) \geq e^{-n\tau_n^2}. \quad (14)$$

Assumption 4 *(Control size of $\boldsymbol{\alpha}$) Let $A_n := \left\{i : \alpha_i < 1^-; i \in 1, \ldots, n\right\}$ and $n_A := |A_n|$, where $|A_n|$ denotes the number of elements in A_n. Let $Q_n := \left\{i : \alpha_i = \alpha^{(n)} \geq 1^-; i \in 1, \ldots, n\right\}$ for some constant $\alpha^{(n)}$ and $n_Q := |Q_n|$.*

$$\limsup_n |A_n| = \limsup_n n_A = \mathcal{O}\left(n^{\frac{1}{2}}\right), \text{ with } P_{\theta_0}-\text{probability } 1$$
$$\limsup_n (1 - \alpha^{(n)}) = \mathcal{O}\left(n_Q^{-\frac{1}{2}}\right), \text{ with } P_{\theta_0}-\text{probability } 1,$$

such that for constants $C_1, C_3 > 0$ and n sufficiently large,

$$\sup_n |A_n| \leq C_1 n^{\frac{1}{2}}$$
$$\sup_n (1 - \alpha^{(n)}) \leq C_3 \tau_n n_Q^{-\frac{1}{2}}$$

These two assumptions are required for consistency of our $\boldsymbol{\alpha}-$pseudo posterior mechanism at θ^*. The first assumption requires the prior to place some mass on a KL ball near θ^* as defined in Equation (6). The second assumption outlines a dyadic subgrouping of data records, where A_n contains those records whose likelihood contributions are downweighted to lessen the estimated identification disclosure risk (and improve privacy) for those records in the resulting synthetic data. The second subset of records, Q_n, contains those records that are minimally downweighted due to nearly zero values for identification disclosure risks. Since $\alpha_i < 1,\ \forall i \in (1, \ldots, n)$, the constant value, $\alpha^{(n)}$, for all units in Q_n approaches 1 from the left. We show that the consistency result to θ^* for the synthesizer is dominated by the likelihood weighting for records in the downweighted set, A_n. Assumption 4 restricts the number of downweighted records (where $\alpha_i < 1^-$) to grow at a slower rate than the sample size, n, such that the downweighting becomes relatively more sparse.

Theorem 8 *(Contraction of the $\boldsymbol{\alpha}-$pseudo posterior distribution).*
Let $\boldsymbol{\alpha} = (\alpha_1 \in (0,1), \ldots, \alpha_n \in (0,1))$. Define $\alpha_m := \max_{i \in A_n} \alpha_i \in (0,1)$ and $\alpha_l := \min_{i \in A_n} \alpha_i \in (0,1)$. Let $D_{\theta_0,\boldsymbol{\alpha}}^{(n_A)}(\theta, \theta^*) = \sum_{i \in A_n} D_{\theta_0,\alpha,i}$ and $D_{\theta_0,1^-}^{(n_Q)}(\theta, \theta^*) = \sum_{i \in Q_n} D_{\theta_0,1^-,i}$. Let θ^* be as

defined in Equation (6). Assume that τ_n *satisfies* $n\tau_n^2 \geq 2$ *and suppose Assumptions 3 and 4 hold. Let* $C_1^* = \sqrt{2 + C_1^2 + C_3^2} \geq \sqrt{2}$. *Then for any* $D \geq 2$ *and* $t > 0$,

$$\xi^{\boldsymbol{\alpha}}\left(\frac{1}{n}\left[(1-\alpha_m)D_{\theta_0,\boldsymbol{\alpha}}^{(n_A)}(\theta,\theta^*) + (1-\alpha^{(n)})D_{\theta_0,1^-}^{(n_Q)}(\theta,\theta^*)\right] \geq (D+3t)\tau_n^2 | \mathbf{x}\right) \leq e^{-tn\tau_n^2}, \quad (15)$$

hold with $P_{\theta_0}-$*probability at least* $1 - \left[(\alpha_l^2+2)(C_1^*)^2/\alpha_m^2 \times 2/\left\{(D+t-1)^2 n\tau_n^2\right\}\right]$.

Since $(1-\alpha^{(n)}) = \mathcal{O}(n_Q^{-1/2})$, while $n_A = \mathcal{O}(n^{1/2})$, the first term dominates with increasing n, so that the $(1-\alpha_m)^{-1}$ is the dominating penalty on the τ_n contraction rate of the $\boldsymbol{\alpha}-$pseudo posterior onto θ^*. Even though the downweighting becomes relatively more sparse due to Assumption 4, it is the maximum value of α_i for $i \in A_n$ on the set of downweighted records that penalizes the rate. We observe that the rate of contraction is injured by factor, $(1-\alpha_m)^{-1}$. Since $\alpha_i \leq 1^-$, $\forall i \in A_n$, our result generalizes Bhattacharya et al. (2019) to allow a tempering of a *portion* of the posterior distribution and there is a penalty to be paid in terms of contraction rate for the tempering. Since we induce the misspecification through the weights, $\boldsymbol{\alpha}$, the distance of the point of contraction, θ^* from the true generating parameters, θ_0, and the contraction rate on this point are *both* impacted by the induced misspecification. The requirement for increasing sparsity in the number of downweighted record likelihood contributions, however, ensures that θ^* will be relatively close to θ_0 that produces a high utility for our (pseudo posterior) estimator.

Asymptotically, then, the space $\theta \in \Theta$ collapses onto θ^* for n sufficiently large and the space of databases, $\mathbf{x} \in \mathcal{X}^n$ becomes $\mathbf{x} \sim P_{\theta^*}$, where the synthesized $\mathbf{x}$ derives from a "corrupted" or misspecified data generating process designed to encode privacy protection. Since the contraction of the pseudo posterior distribution induces the collapsing of the parameter space to a point and the space of databases to a single distribution (conditioned on θ^*) for large n, this result guarantees that the local Lipschitz bound, $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$ contracts on to the global bound $\Delta_{\boldsymbol{\alpha}}$ for n sufficiently large. Assumption 2 ensures a formal privacy guarantee since $\Delta_{\boldsymbol{\alpha}} < \infty$ and the asymptotic result provides assumptions under which $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$, computed on the observed database, contracts on the global $\Delta_{\boldsymbol{\alpha}}$ over the space of databases to reveal the associated ϵ.

3. Computing a Local Lipschitz Bound

In this section, we describe the implementation details to compute the pseudo likelihood weights, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ for a local database, $\mathbf{x}$, from the *unweighted* synthesizer and the subsequent computation of the Lipschitz bound, $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$, for the pseudo posterior mechanism. In Section 3.1, we lay out the connection between the scalar-weighted pseudo posterior mechanism and the EM, with a discussion of the implications on the data utility of differentially private synthetic data generated under the two mechanisms.

1. Compute weights $\boldsymbol{\alpha}$

    (a) Let $|f_{\theta_s,i}|$ denote the absolute value of the log-likelihood computed from the unweighted pseudo posterior synthesizer for database record, $i \in (1, \ldots, n)$ and MCMC draw, $s \in (1, \ldots, S)$ of θ.

(b) Compute the $S \times n$ matrix of by-record (absolute value of) log-likelihoods, $L = \{|f_{\theta_s,i}|\}_{i=1,\ldots,n,\ s=1,\ldots,S}$.

(c) Compute the maximum over each $S \times 1$ column of L to produce the $n \times 1$ (database record-indexed) vector, $\mathbf{f} = (f_1, \ldots, f_n)$. We use a linear transformation of each f_i to $\tilde{f}_i \in [0, 1]$ where values of $\tilde{f}_i$ closer to 1 indicates relatively higher identification disclosure risk: $\tilde{f}_i = \frac{f_i - \min_j f_j}{\max_j f_j - \min_j f_j}$.

(d) We formulate by-record weights, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$,

$$\alpha_i = c \times (1 - \tilde{f}_i) + g, \tag{16}$$

where c and g denote a scaling and a shift parameters, respectively, of the α_i used to tune the risk-utility trade-off. These $\boldsymbol{\alpha}$ satisfy Assumption 2.

As discussed in Hu and Savitsky (2019), the scaling parameter c compresses or expands the distribution by-record weights and induces a global affect on the risk-utility trade-off, while the shift parameter g shifts the distribution of by-record weights has a local effect on the risk-utility trade-off. We will show in Section 5 the effects of different configurations of c and g on the risk and utility profiles of the differentially private synthetic dataset for the CE sample, generated under our proposed $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism.

2. Compute Lipschitz bound, $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$

(a) Use $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ to construct the pseudo likelihood of Equation (3) from which the pseudo posterior of Equation (5) is estimated. Draw $(\theta_s)_{s=1,\ldots S}$ from the $\boldsymbol{\alpha}-$weighted pseudo posterior distribution.

(b) As earlier, compute the $S \times \ n$ matrix of log-pseudo likelihood values, $L^{\boldsymbol{\alpha}} = \left\{|f^{\boldsymbol{\alpha}}_{\theta_s,i}|\right\}_{i=1,\ldots,n,\ s=1,\ldots,S}$

(c) Compute $\Delta_{\boldsymbol{\alpha},\mathbf{x}} = \max_{s,i}|f^{\boldsymbol{\alpha}}_{\theta_s,i}|$.

3. Draw synthetic data, $\boldsymbol{\zeta}_\ell$, from the pseudo posterior distribution

(a) Using the $(\theta_s)_{s=1,\ldots S}$ drawn from the $\boldsymbol{\alpha}-$weighted pseudo posterior distribution estimated in the earlier step, randomly sample $\ell = 1, \ldots, (M = 20)$ parameter values and draw synthetic data value, $\zeta_{\ell,i} \overset{\text{ind}}{\sim} p_{\theta_\ell}(\cdot)$ for parameter draw $\ell \in (1, \ldots, M)$ and database record $i \in (1, \ldots, n)$. This step accomplishes a draw from the pseudo posterior predictive distribution.

(b) Release the synthetic data, $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1, \cdots, \boldsymbol{\zeta}_M)$, in place of the closely-held real data, $\mathbf{x}$.

Our pseudo posterior mechanism *indirectly* sets the DP expenditure level, ϵ, through the computation and subsequent scaling and shifting of the likelihood weights, $\boldsymbol{\alpha}$. This mechanism is guaranteed to be differentially private for any n because we set $\alpha_i = 0$ if the log-likelihood contribution for record i is non-finite. We showed in Section 2 that this local Lipschitz bound of our log-pseudo likelihood, $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$, contracts on the global bound, $\Delta_{\boldsymbol{\alpha}}$, for n sufficiently large which, in turn, determines the privacy expenditure, ϵ.

3.1 Exponential Mechanism Reduces to Scalar Weighting

Wasserman and Zhou (2010); Zhang et al. (2016); Snoke and Slavkovic (2018) use the EM to generate synthetic data with privacy guarantees from a non-private mechanism. Suppose we start with a non-private mechanism, such as an unweighted synthesizer in Equation (17),

$$\xi\left(\theta \mid \mathbf{x}, \gamma\right) \propto \left[\prod_{i=1}^{n} p\left(x_i \mid \theta\right)\right] \xi\left(\theta \mid \gamma\right). \tag{17}$$

Under our set-up that the log-likelihood function as the utility function, i.e. $u(\mathbf{x}, \theta) = \log\left(\prod_{i=1}^{n} p\left(x_i \mid \theta\right)\right)$, the EM generates private samples from

$$\hat{\theta} \propto \exp\left(\frac{\epsilon \, \log\left(\prod_{i=1}^{n} p\left(x_i \mid \theta\right)\right)}{2\Delta}\right) \xi\left(\theta \mid \gamma\right), \tag{18}$$

where the prior, $\xi(\theta \mid \gamma)$, is chosen as the "base" distribution as in Zhang et al. (2016) specified by McSherry and Talwar (2007) that ensures the EM produces a proper density function. Furthermore,

$$\begin{aligned} \exp\left(\frac{\epsilon \, \log\left(\prod_{i=1}^{n} p\left(x_i \mid \theta\right)\right)}{2\Delta}\right) \xi\left(\theta \mid \gamma\right) &= \exp(\log(\prod_{i=1}^{n} p\left(x_i \mid \theta\right))^{\frac{\epsilon}{2\Delta}}) \xi\left(\theta \mid \gamma\right) \\ &= \left(\prod_{i=1}^{n} p\left(x_i \mid \theta\right)^{\frac{\epsilon}{2\Delta}}\right) \xi\left(\theta \mid \gamma\right), \end{aligned} \tag{19}$$

which means that the EM is equivalent to a risk-adjusted, scalar-weighted pseudo posterior synthesizer with scalar weight $\frac{\epsilon}{2\Delta}$, where $\alpha_i = \frac{\epsilon}{2\Delta}$, $\forall i \in (1, \ldots, n)$.

There are important implications of the EM reducing to a scalar-weighted pseudo posterior under use of the log-likelihood as the utility function. First, we must assume a global Lipschitz bound exists via Assumption 1: we cannot appeal to Assumption 2 which depends on differential weights. Second, using a scalar weight, $\alpha_i = \frac{\epsilon}{2\Delta}$, $\forall i \in (1, \ldots, n)$, shown in Equation (19), we expect a resulting lower utility for synthetic data draws under this mechanism than we do under our $\boldsymbol{\alpha}-$weighted pseudo posterior shown in Equation (16). The $\boldsymbol{\alpha}-$weighted pseudo posterior is more surgical and concentrates the downweighting to higher risk records, whereas the EM must downweight all records the same amount. Downweighting all records the same amount will be conservative because the scalar weight is based on the *worst case* sensitivity over the entire database of records, which is required to achieve an $\epsilon-$privacy guarantee and parameter spaces and not tuned to the risk ($\tilde{f}_i$) of each record.

We illustrate in Section 5 the reduction in utility of the differentially private synthetic dataset generated under the EM, compared to that under our proposed $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism, for the CE sample for an equivalent privacy guarantee for both mechanisms.

4. Turning A Local Bound into A Global Bound

Assumption 4 restricts the *number* of records, $i \in (1, \ldots, n)$, that are downweighted by receiving an $\alpha_i < 1$ (as opposed to setting $\alpha_i = 1$) to grow at $\mathcal{O}(n^{\frac{1}{2}})$. This restriction

requires a progressively sparser downweighting of records as n increases, which generally accords with the more "surgical" nature of downweighting by using a vector of weights, $\boldsymbol{\alpha}$, with the weight for each record based on the disclosure risk measured by the sensitivity of its log-likelihood, such that the downweighting is confined to those records which express identification disclosure risk. These records mainly reside in the tails of the density, $p_{\theta^*}(\mathbf{x})$, under our weighting scheme. Yet, to assure contraction of the local Lipschitz bound to the global Lipschitz bound needed to compute DP expenditure, ϵ, the number of records that express non-finite log-likelihoods must be very small. This probability would be very small but bounded away from 0 for any arbitrarily large finite bound M and will not shrink with increasing sample sizes. In other words, it is possible that the number of x_i that need to set $\alpha_i = 0$ in the case of non-finite log-likelihoods *may* violate the requirement for an increasing sparsity in the number of records downweighted ($\mathcal{O}(n)$ records instead of $\mathcal{O}(n^{\frac{1}{2}})$), though the relatively small probability for a non-finite observation under the class of unimodal distributions mitigates this concern. In this section, we demonstrate how a finite global DP guarantee can still be made even when Assumption 1 or the sparsity assumptions in Section 2 do not hold, at the expense of further loss of utility.

4.1 Global DP via a Censored Pseudo Likelihood

A possible alternative to relying on the asymptotic contraction of $\theta \in \Theta$ (to θ^*) to achieve a global privacy guarantee at large samples, n, is to explore the use of censoring the log-likelihood at some threshold, $M_{\mathbf{x}}$, that is defined based on a local result characterized by attractive risk and utility performances; we provide examples of local results with attractive risk and utility performances on our real data application to the CE sample in the sequel. The use of censoring would "lock in" the finite Lipschitz bound and, hence, the finite global privacy guarantee such that for all subsequent samples of sizes $\geq n$, the privacy guarantee would be fixed without relying on asymptotics. This method constructs the pseudo likelihood as,

$$p_c^{\boldsymbol{\alpha}}(x_i \mid \theta) = \begin{cases} \exp(M_{\mathbf{x}}), & p(x_i \mid \theta)^{\boldsymbol{\alpha}} > \exp(M_{\mathbf{x}}), \\ \exp(-M_{\mathbf{x}}), & p(x_i \mid \theta)^{\boldsymbol{\alpha}} < \exp(-M_{\mathbf{x}}), \\ p(x_i \mid \theta)^{\boldsymbol{\alpha}}, & \text{otherwise}, \end{cases}$$

for use in

$$\xi_c^{\boldsymbol{\alpha}}(\theta \mid X) \propto \prod_{i=1}^{n} p_c^{\boldsymbol{\alpha}}(x_i \mid \theta)\xi(\theta). \tag{20}$$

While this formulation of $f_c^{\boldsymbol{\alpha}}(x_i \mid \theta) = \log p_c^{\boldsymbol{\alpha}}(x_i \mid \theta)$ is simple, it can lead to serious computational issues. For many combinations of $\{x_i, \theta\}$, the censored likelihood cannot discriminate between better and worse fit. These can lead optimization and rejection sampling algorithms to drift. One alternative is to use a *strictly monotonic* transformation such as the arc tangent,

$$f_{ac}^{\boldsymbol{\alpha}}(\mathbf{x} \mid \theta) = \left(\frac{2M_{\mathbf{x}}}{\pi}\right) \arctan\left(f(\mathbf{x} \mid \theta)^{\boldsymbol{\alpha}}\left(\frac{\pi}{2M_{\mathbf{x}}}\right) - \mu_x\right) + \mu_x. \tag{21}$$

The transformation $g(f \mid M_{\mathbf{x}}, \mu_x)$ has the property that $g(\mu_x \mid M_{\mathbf{x}}, \mu_x) = \mu_x$ with local slope $g'(\mu_x \mid M_{\mathbf{x}}, \mu_x) = 1$ and that $\lim_{M_{\mathbf{x}} \to \infty} g(f \mid M_{\mathbf{x}}, \mu_x) = f$. Together these properties

mean that $g(f)$ is a good local approximation to f in the neighborhood of μ_x and converges to the original f as the bounds increase towards ∞. It is also clear that $g(f)$ is strictly monotonic, so optimization and rejection sampling approaches can still be viable. Figure 1 displays the mapping $g(f \mid M_{\mathbf{x}} = 100, \mu_x = 0)$ which respects the $M_{\mathbf{x}}$ bounds globally while still being reasonably close to f for small and moderate values of f.

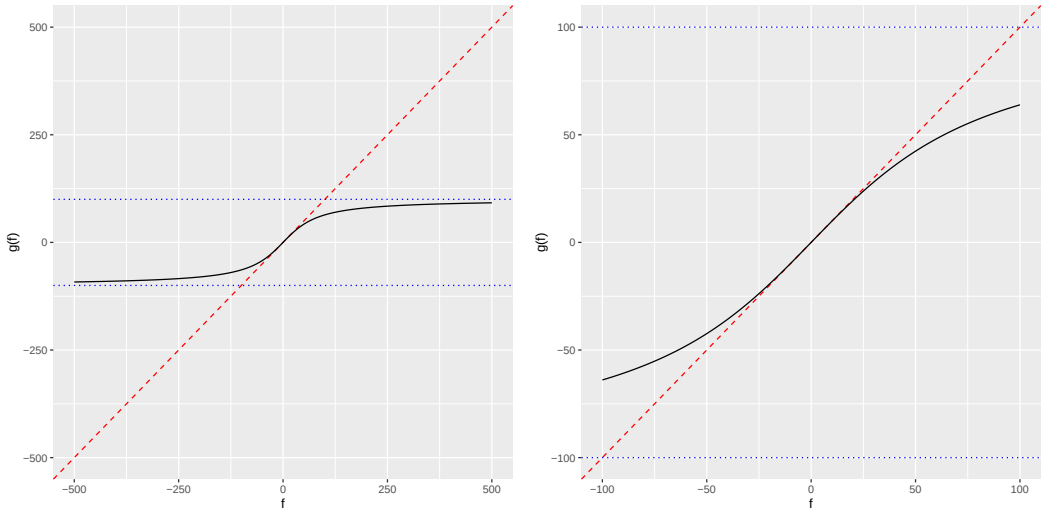

Figure 1: The bounded tranformation $g(f \mid M_{\mathbf{x}} = 100, \mu_x = 0)$ (solid) and one-to-one line (dashed red) for bounds $\pm M_{\mathbf{x}} = 100$ (dotted blue). Zoom out (left), zoom in (right).

To demonstrate the risk-utility trade-off for a non-asymptotic global bound we use Equation (21) to estimate the posterior of the mean parameter from a Poisson data generating distribution across different fixed bounds $M = \{1, 2, 5, 10, 100\}$ and sample sizes $n = \{100, 1000\}$. We provide a simple, Monte Carlo simulation study to demonstrate the process of taking a result based on a *local* estimate of a Lipschitz bound on a *single*, observed dataset and applying it to repeated generation of new datasets such that analysis of these subsequent datasets have Lipschitz bounds at or below that of the original dataset under use of the arctan censoring mechanism of Equation (21). We compare the distributions of Lipschitz bounds for the unweighted posterior, the $\boldsymbol{\alpha}-$weighted pseudo posterior and the arctan censoring (that embeds the $\boldsymbol{\alpha}-$weighted pseudo posterior) mechanisms. We also compare their utility performances to preserve key characteristics of the true data distribution. As earlier discussed, *guaranteeing* a finite non-asymptotic global result under the arctan censoring mechanism leads to additional loss in utility, although this loss is mitigated by larger sample sizes.

Using the simple means model for Poisson distributed data, $y \sim Pois(\mu)$ (with $\mu = 50$) our procedure for locking in a local result, globally, is the following:

1. Generate one base set of $\mathbf{y}_0 = \{y_1, \ldots y_{100}\}$ under sample size $n = 100$ (or $n = 1000$) as our *observed* dataset.

2. Compute the maximum local Lipschitz, $\Delta_{\boldsymbol{\alpha},\mathbf{y}_0}$, for the local database under the $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism for estimating μ.

3. For $j = 1, \ldots, 100$:

   - Generate $\mathbf{y}_j \sim \text{Pois}(\mu)$, each of size $n = 100$ (or $n = 1000$).
   - Compute the *local* Lipschitz bound for the unweighted and $\boldsymbol{\alpha}-$weighted pseudo posterior mechanisms *without* the arctan transformation.
   - Use the arctan transformation of Equation (21) that embeds the $\boldsymbol{\alpha}-$weighted pseudo posterior as our mechanism under each of $M = \{\Delta_{\boldsymbol{\alpha},\mathbf{y}_0}, 1.5\Delta_{\boldsymbol{\alpha},\mathbf{y}_0}, 2\Delta_{\boldsymbol{\alpha},\mathbf{y}_0}\}$.

Figure 2 compares the distributions across the $J = 100$ replications. The $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism (without embedding in the arctan censoring), labeled "Weighted", produces a marked decrease in local maximum Lipschitz compared to the unweighted (labeled "Unweighted"). The local $\Delta_{\boldsymbol{\alpha},\mathbf{y}_0}$ bound for our observed dataset, $\mathbf{y}_0$, is close to the mode of the 100 Monte Carlo simulated weighted Lipschitz bounds. So, while there is a contraction of the local Lipschitz bounds for the Weighted result under $\boldsymbol{\alpha}$ vectorized weighting, a given local $\Delta_{\boldsymbol{\alpha},\mathbf{y}_0}$ is not strictly an upper bound. In contrast, using $M = \Delta_{\boldsymbol{\alpha},\mathbf{y}_0}$ with censoring (labeled with prefix, "Wt") does ensure all realized local Lipschitz bounds are below the local Lipschitz bound, $\Delta_{\boldsymbol{\alpha},\mathbf{y}_0}$, but at the cost of reduced utility (90th quantile is shifted). Using slightly larger bounds ($1.5\Delta_{\boldsymbol{\alpha},\mathbf{y}_0}$ and $2\Delta_{\boldsymbol{\alpha},\mathbf{y}_0}$) recovers some of the utility at the expense of looser global guaranteed bounds.

To compare asymptotic versus censored bounds, we repeat the simulation above using sample size $n = 1000$. Figure 3 demonstrates that the local Lipschitz bounds for the unweighted likelihood increases (or drifts) with larger sample sizes. The $\boldsymbol{\alpha}-$weighted log-pseudo likelihood shows a pronounced decrease in drift, with only a slightly larger bound for $n = 1000$, indicating an asymptotic contraction of the local result towards a global result.

5. Application to the CE Sample

We introduce the CE sample of consumer units (CU) or households in Section 5.1, where our goal is to synthesize a highly-skewed continuous variable, family income, under a local DP guarantee provided by our $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism. In Section 5.2, we present risk and utility profiles of synthetic data drawn from our $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism, along with comparisons to the EM, the risked-weighted synthesizer of Hu and Savitsky (2019) and the unweighted posterior mechanism. Section 5.3 presents privacy and utility results with different scaling and shifting, (c, g), configurations for vector weights in Equation (16) to sketch out a risk-utility curve for our $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism that we compare to that of the EM. A risk-utility curve provides the Bureau of Labor Statistics (BLS) options for selecting a risk-utility setting that matches their policy objectives.

5.1 The CE Sample and Unweighted Synthesizer

Our application of the $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism focuses on providing privacy protection for a family income variable published by the CE. The CE is administered by the BLS with the purpose of providing income and expenditure patterns indexed by

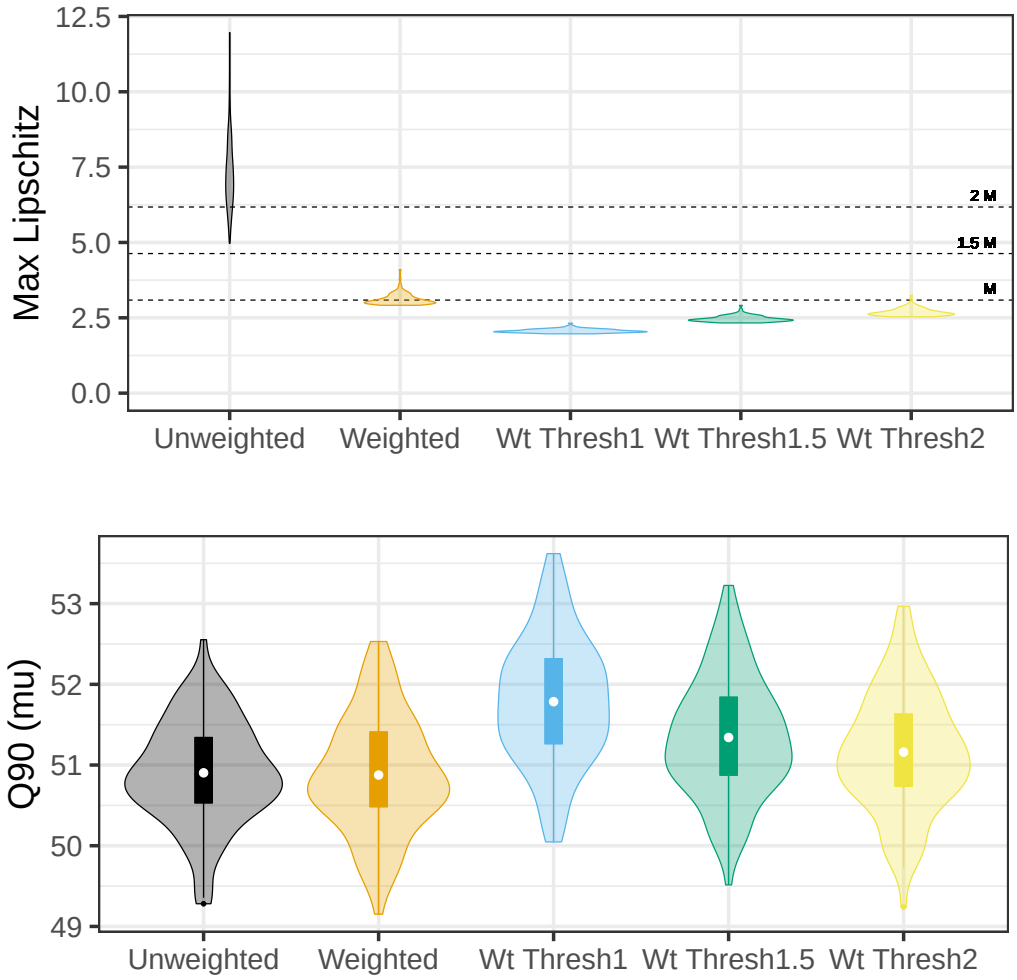

Figure 2: Distribution of the maximum observed Lipschitz bound $\Delta_{\mathbf{y}}$ (top) to threshold bounds M (dashed) and estimates of the 90th quantile of the mean parameter μ (bottom) from posterior samples of (left to right) unweighted, weighted, and weighted with global bounds. Based on 100 realizations of size 100.

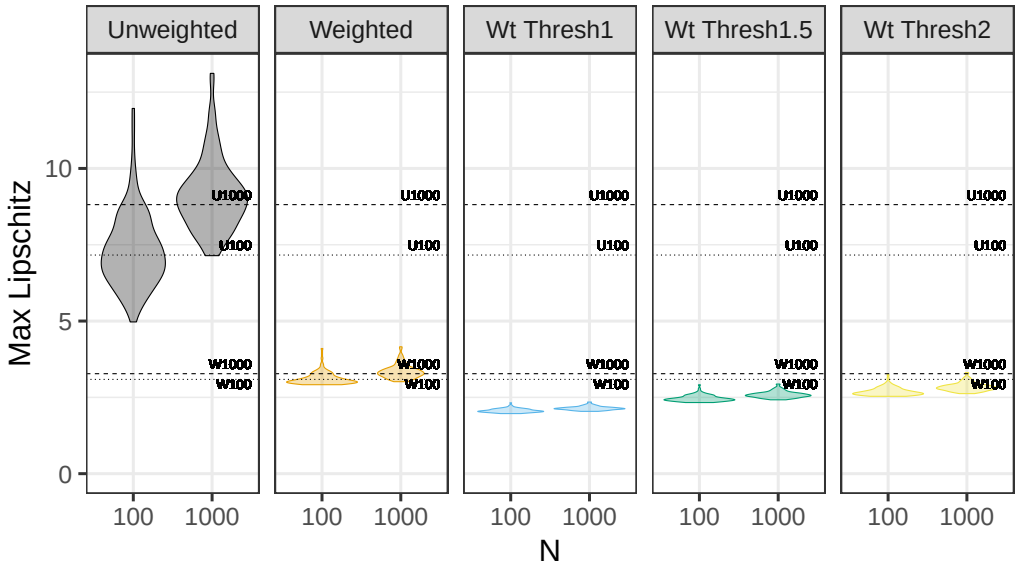

Figure 3: Distribution of the maximum observed Lipschitz bound $\Delta_{\mathbf{y}}$ by sample size (100, 1000) from 100 realizations of posterior samples of (left to right) unweighted, weighted, and weighted with global bounds. Dashed lines use $n = 1000$ for baseline M, dotted lines use $n = 100$. Top lines (U) are for unweighted local M, bottom lines (W) are weighted local M.

geographic domains to support policy-making by State and Federal governments. The description of the CE sample included here closely follows that in Hu and Savitsky (2019). The CE contain data on expenditures, income, and tax statistics about CUs across the U.S. The CE public-use microdata (PUMD)[1] is publicly available record-level data, published by the CE. The CE PUMD has undergone masking procedures to provide privacy protection of survey respondents. Notably, the family income variable has undergone top-coding, a popular Statistical Disclosure Limitation (SDL) procedure that may result in reduced utility and insufficient privacy protection (An and Little, 2007; Hu and Savitsky, 2019).

The CE sample in our application contains $n = 6208$ CUs, coming from the 2017 1st quarter CE Interview Survey. It includes the family income variable, which is highly right-skewed and deemed sensitive; see Figure 4 for its density plot. The CE sample also contains 10 categorical variables, listed in Table 1. These categorical variables are deemed insensitive and used as predictors in building a flexible synthesizer for the synthesis of the sensitive family income variable.

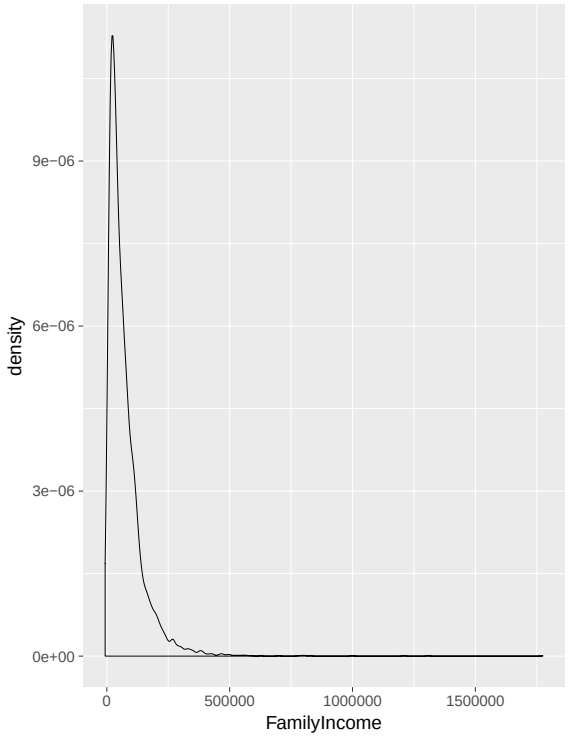

Figure 4: Density plot of Family Income in the CE sample.

1. For for information about CE PUMD, visit `https://www.bls.gov/cex/pumd.htm`.

Table 1: Variables used in the CE sample. Data taken from the 2017 Q1 Consumer Expenditure Surveys.

Variable	Description
Gender	Gender of the reference person; 2 categories
Age	Age of the reference person; 5 categories
Education Level	Education level of the reference person; 8 categories
Region	Region of the CU; 4 categories
Urban	Urban status of the CU; 2 categories
Marital Status	Marital status of the reference person; 5 categories
Urban Type	Urban area type of the CU; 3 categories
CBSA	2010 core-based statistical area (CBSA) status; 3 categories
Family Size	Size of the CU; 11 categories
Earner	Earner status of the reference person; 2 categories
Family Income	Imputed and reported income before tax of the CU; approximate range: (-7K, 1,800K)

To generate partially synthetic datasets for the CE sample with synthetic family income, we use an unweighted, non-private synthesizer: a flexible, parametric finite mixture synthesizer. This finite mixture synthesizer has been shown to produce synthetic data characterized by a high utility, but also with an unacceptable level of disclosure risk in previous work (Hu and Savitsky, 2019). We leave the details of the synthesizer in the Appendix B for brevity and direct interested readers to the aforementioned work for further information.

5.2 Risk and Utility Comparisons

To generate synthetic data and compare results, we apply four synthesizers: 1) the unweighted, non- (locally) private synthesizer, labeled "Unweigthed"; 2) the locally private synthesizer under the $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism, labeled "DPweighted", with configuration $(c, g) = (0.7, 0.0)$; 3) the locally private synthesizer under the EM, labeled "EMweighted", which is designed to privacy target, ϵ, achieved by "DPweighted"; 4) and the weighted, though non- (locally) private pseudo posterior synthesizer proposed by Hu and Savitsky (2019), labeled "Countweighted", that utilizes their method for measuring the by-record disclosure risk (based on an assumption about the behavior of an intruder). The labels are used throughout the remainder of this paper when presenting various risk and utility results.

We first look at the risk profiles of the four synthesizers. Figure 5 plots the distributions of the Lipschitz bounds, Δ_{x_i}'s, for each of the four synthesizers computed by taking the maximum of the S log-likelihood ratios for each record, $i = 1, \ldots, (n = 6208)$ over the S draws of θ from it's posterior distribution. The maximum value of the (Δ_{x_i}) over all of the records is denoted as $\Delta_{\mathbf{x}}$, the Lipschitz bound for the mechanism.

The Unweighted, non-private synthesizer clearly has the highest maximum $\Delta_{\mathbf{x}}$ with $\Delta_{Unweighted} = 78.7$. The other non-private, Countweighted synthesizer achieves a much lower maximum $\Delta_{\mathbf{x}}$ with $\Delta_{\alpha_c, Countweighted} = 11.17$. The large reduction in the Countweighted synthesizer owes to the positive correlation between by-record weights, $\propto 1/\alpha_c$,

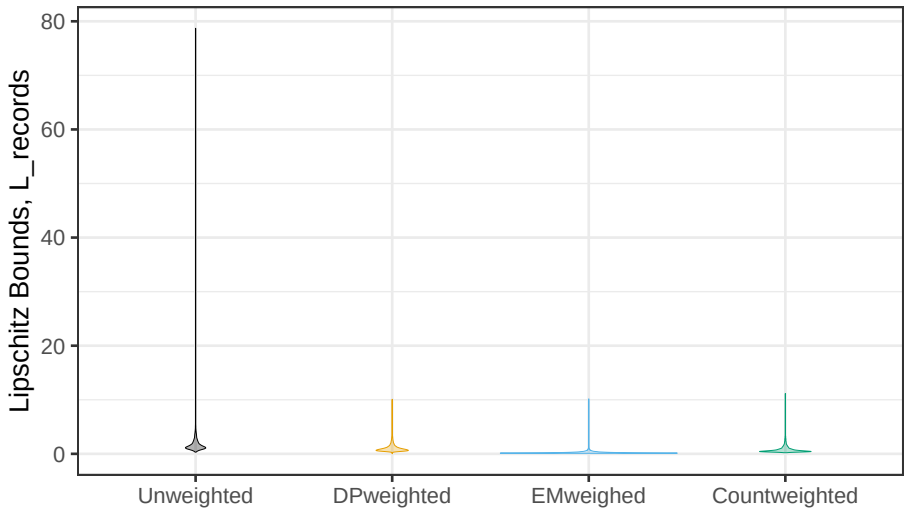

Figure 5: Violin plots of the distribution of the Lipschitz bounds, $\Delta_{\mathbf{x}}$'s, for synthetic data generated the four synthesizers. The corresponding maximum $\Delta_{\mathbf{x}}$ values are: $\Delta_{Unweighted} = 78.7, \Delta_{\boldsymbol{\alpha},DPweighted} = 10.1, \Delta_{EMweighted} = 10.2, \Delta_{\boldsymbol{\alpha}_c,Countweighted} = 11.17$.

where each is computed as the probability that the value for each target record is relatively isolated from that of other records used in the Countweighted synthesizer, on the one hand, with the by-record log-pseudo likelihood ratio bounds used for the DPweighted mechanism, on the other hand. The two locally private synthesizers both achieve even lower maximum $\Delta_{\mathbf{x}}$: $\Delta_{\boldsymbol{\alpha},DPweighted} = 10.1, \Delta_{EMweighted} = 10.2$, indicating the best risk profiles. The EMweighted mechanism was estimated by setting the scalar with a target $\epsilon = 2\Delta_{\boldsymbol{\alpha},\mathbf{x}}$, the local privacy guarantee (expenditure) achieved by our DPweighted mechanism with Lipschitz $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$. Our intent is to compare the utility performances between the two private mechanisms (DPweighted and EMweighted) where each achieves an equivalent privacy guarantee. It bears mention that while the DPweighted under the pseudo posterior mechanism and the EMweighted under the EM achieve similar maximum local Lipschitz bounds, which governs the local DP guarantee, the EM tends to produce notably lower risk for most records than the DPweighted mechanism, evident in the flattened shape of the violin plot. The EM sets the scalar weight based on the risk of the worst case records because the same level of downweighting must be applied to all records in contrast with the by-record weighting under of our $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism.

Figure 6 and Figure 7 show a collection of violin plots of the distribution (obtained from re-sampling) for each of the mean and the 90th quantile statistics, respectively, estimated on the synthetic data generated under each of our synthesizers and also on the closely-held confidential (real) data for comparison, labeled "Data". These figures allow us to compare the utility performances across our synthesizers by the examination of how well the real data distribution for each statistic is reproduced by the synthetic dataset for each of our

synthesizers. For the synthesizers, a set of $M = 20$ synthetic datasets were generated and the distribution for each statistic was estimated on each dataset (under re-sampling). The resulting barycenter of the individual distributions in the Wasserstein space of measures was computed by averaging the quantiles over the M datasets (Srivastava et al., 2015). Our privacy guarantees apply to *each* synthetic draw from our mechanism, so the total privacy expenditure is that for each dataset shown in Figure 5 multiplied by M. We compute utilities over $M = 20$ synthetic datasets for thoroughness, though the distribution of each statistic for a single synthetic dataset is very similar.

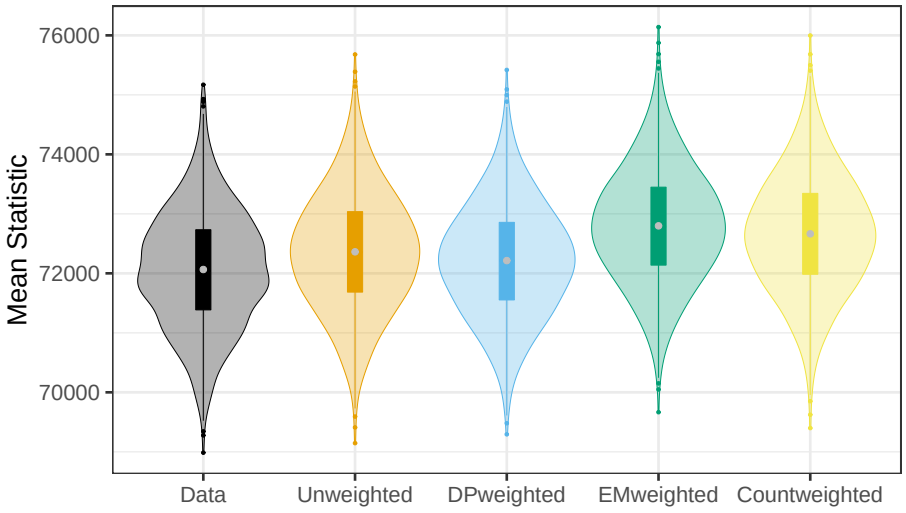

Figure 6: Violin plots of the mean estimation of the confidential CE sample and the four synthesizers.

The DPweighted synthesizer under the pseudo posterior mechanism outperforms the EMweighted and Countweighted mechanisms in utility preservation. First, especially evident in Figure 7, DPweighted (the $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism) provides better estimates than EMweighted (the scalar-weighted EM). The notably deteriorated utility preservation of the EM derives from the setting that scalar weight applied to all records based on the highest risk records as earlier discussed. Since both mechanisms achieve the same maximum Lipschitz bound $\Delta_{\mathbf{x}}$, which governs the local DP guarantee, these results indicate that the EM has to compromise a large amount of the utility to achieve a similar local DP guarantee compared to the $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism.

Second, while the non-private Unweighted synthesizer and the locally private DPweighted synthesizer provide equally good estimates for both the mean and the 90th quantile, the much greater Lipschitz bound of the Unweighted synthesizer shown in Figure 5 indicates a much worse balance for the utility-risk trade-off as compared to DPweighted. The third minor point is that the Countweighted synthesizer, albeit non-locally private, achieves only a slightly higher maximum Lipschitz bound compared to our private DPweighted synthe-

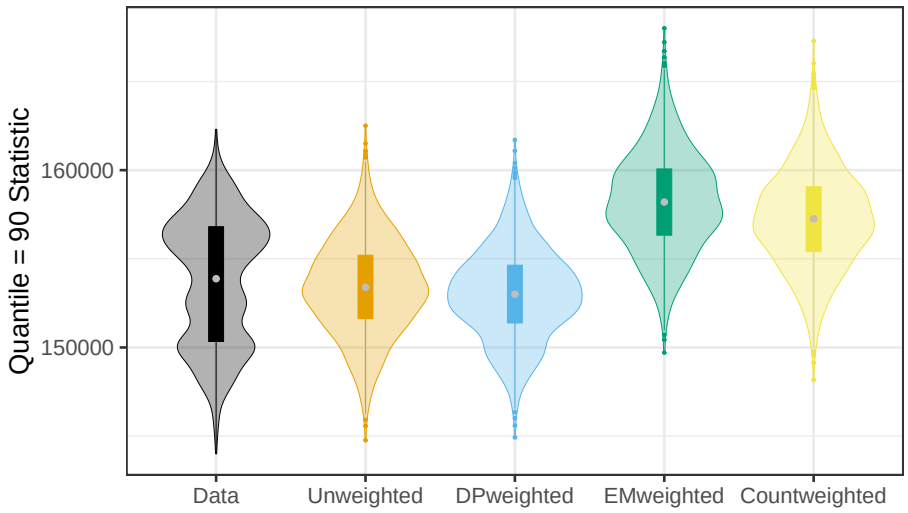

Figure 7: Violin plots of the 90th quantile estimation of the confidential CE sample and the four synthesizers.

sizer; however, its utility preservation is worse, especially evident in Figure 7 for the 90th quantile estimation.

In summary, our private DPweighted mechanism outperforms the other three synthesizers to achieve a highly satisfactory risk-utility trade-off balance. We next explore different scaling and shift configurations of (c, g) to sketch out the risk-utility curves for DPweighted and EMweighted.

5.3 Mapping DP Risk and Utility Curves

We conclude by applying a scaling parameter, c, and a shift parameter, g, to the distribution of weights, $\boldsymbol{\alpha}$, used in our $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism in order to enumerate the risk-utility settings for the purpose of allowing the BLS (or, more generally, the owner of the closely-held private database) to discover the setting configuration that best represents their policy goal. We compare the risk-utility mapping produced by the $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism to that of the EM, which we recall reduces to a scalar-weighted pseudo posterior under use of the log-likelihood as the utility measure. As discussed in Hu and Savitsky (2019), applying a scaling constant, $c < 1$, will induce a compression in the distribution of the weights while apply a scaling $g < 0$ will induce a downward shift in the distribution of the record-indexed weights. We apply the scaling and shifting in a manner that uses truncation to ensure each of the resulting weights are restricted to lie in $[0, 1]$.

Each violin plot in Figure 8 presents a distribution of the 90th quantile for a synthetic dataset generated under a particular (scale c, shift g) configuration. The sequence of plots from left-to-right are ordered from less scaling and shifting (with a relatively higher privacy expenditure) to more scaling and shifting (with a relatively lower privacy expenditure).

The specific sensitivity values, $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$, associated with each configuration are shown in Table 2, where we recall that the associated local privacy expenditure is $\epsilon = 2\Delta_{\boldsymbol{\alpha},\mathbf{x}}$. Table 2 demonstrates a nearly 80% reduction in the local DP expenditure over the range of configurations (though all are much less than the non-locally private, unweighted synthesizer). Figure 8 demonstrates a much flatter or reduced deterioration of utility for DPweighted, the $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism as compared to EMweighted. Such is not surprising due to the greater flexibility of DPweighted to concentrate downweighting to high risk records versus the application of a scalar weight based on the highest risk record to all records under EMweighted.

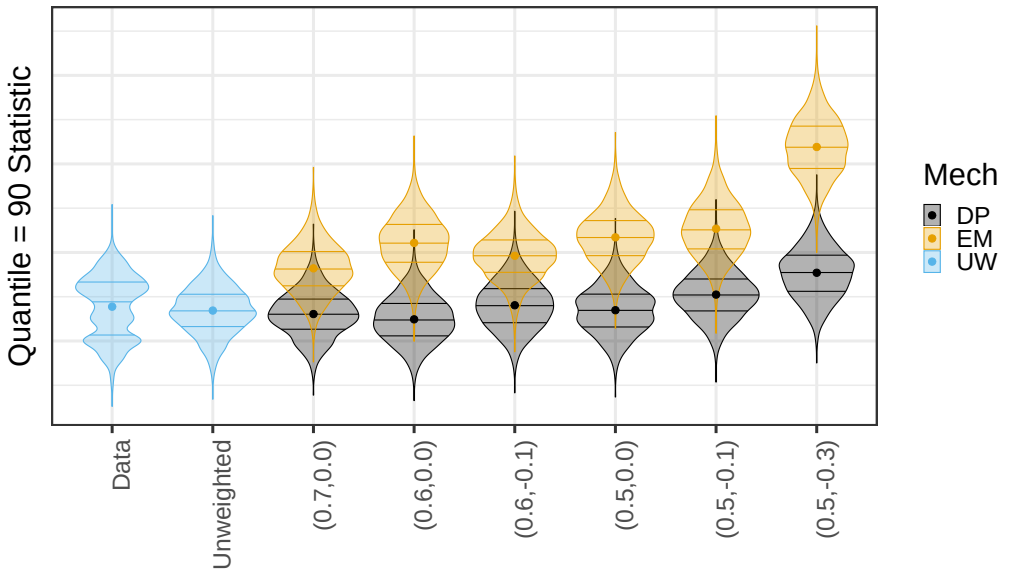

Figure 8: Violin plots of the 90th quantile estimation of: 1) the confidential CE sample; 2) the unweighted, non-private synthesizer; and overlapping violin plots of the 90th quantile estimation of the synthesizer under the pseudo posterior mechanism compared to the synthesizer under the EM with equivalent $\Delta_{\mathbf{x}}$ values, for the following (c, g) configurations: 3) $(c, g) = (0.7, 0.0)$; 4) $(c, g) = (0.6, 0.0)$; 5) $(c, g) = (0.6, -0.1)$; 6) $(c, g) = (0.5, 0.0)$; 7) $(c, g) = (0.5, -0.1)$; 8) $(c, g) = (0.5, -0.3)$.

6. Conclusion

This paper adapts the $\boldsymbol{\alpha}-$weighted pseudo posterior synthesizer as a mechanism that achieves markedly lower DP expenditures for a synthetic dataset in comparison to the non-private, unweighted synthesizer. Our pseudo posterior mechanism provides a much higher utility than the EM for equivalent risk due to a surgical downweighting of high risk records (as opposed to the scalar downweighting imposed by the EM). The construction for

Table 2: Table of values of the Lipchitz bound $\Delta_{\boldsymbol{\alpha},\mathbf{x}}$, of the synthesizer under the $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism, for a series of (c, g) configurations. $\Delta_{Unweighted} = 78.7$.

(c, g)	$\Delta_{\boldsymbol{\alpha},\mathbf{x}}$ value
(0.7, 0.0)	10.10
(0.6, 0.0)	8.16
(0.6, -0.1)	7.30
(0.5, 0.0)	6.09
(0.5, -0.1)	5.71
(0.5, -0.3)	2.25

the $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism utilizes the log-pseudo likelihood to develop the Lipschitz bound. We demonstrate in Section 2 a weighting scheme that guarantees an ϵ privacy expenditure at every sample size, n, as our weighting procedure selectively removes likelihood contributions for dataset records that express unbounded log-likelihood values, such that we do not have to explicitly truncate the parameter space or space of datasets under an unbounded support. We provide an asymptotic result on the contraction of a local Lipschitz to a global bound (tied to a global privacy guarantee) in the case that our vector weighting scheme becomes sparser in the number of records downweighted with increasing n. Finally, we incorporate a log-pseudo likelihood censoring step into our $\boldsymbol{\alpha}-$weighted pseudo posterior mechanism with the threshold set to lock-in a local result obtained for a large n that expresses desired risk and utility properties, making the local result a global one as an alternative to a reliance on an asymptotic contraction and the existence of a finite global bound. Our pseudo posterior mechanism has the feature that it accommodates any synthesizer model formulated by the statistical agency and offers a simple weighting scheme that guarantees a DP result. The simple weighting allows the posterior sampling scheme devised for the non-private synthesizer to be utilized for synthesis with minor modification for the differentially private pseudo posterior mechanism.

Appendix

Appendix A. Proofs for Theoretical Results in Section 2

A.1 Proof for Theorem 4

$$\begin{aligned} D_{KL}\left[(\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\cdot \mid \mathbf{x}) \parallel \xi^{\boldsymbol{\alpha}(\mathbf{y})}(\cdot \mid \mathbf{y}))\right] &= \int_\Theta \ln \frac{d\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\theta \mid \mathbf{x})}{d\xi^{\boldsymbol{\alpha}(\mathbf{y})}(\theta \mid \mathbf{y})} d\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\theta \mid \mathbf{x}) \\ &= \int_\Theta \ln \frac{p_\theta^{\boldsymbol{\alpha}(\mathbf{x})}(\mathbf{x})}{p_\theta^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y})} d\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\theta \mid \mathbf{x}) + \int_\Theta \ln \frac{\phi^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y})}{\phi^{\boldsymbol{\alpha}(\mathbf{x})}(\mathbf{x})} d\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\theta \mid \mathbf{x}) \\ &\leq \int_\Theta \left| \ln \frac{p_\theta^{\boldsymbol{\alpha}(\mathbf{x})}(\mathbf{x})}{p_\theta^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y})} \right| d\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\theta \mid \mathbf{x}) + \int_\Theta \ln \frac{\phi^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y})}{\phi^{\boldsymbol{\alpha}(\mathbf{x})}(\mathbf{x})} d\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\theta \mid \mathbf{x}) \\ &\leq \Delta_{\boldsymbol{\alpha}} + \left| \ln \frac{\phi^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y})}{\phi^{\boldsymbol{\alpha}(\mathbf{x})}(\mathbf{x})} \right| \end{aligned} \tag{22}$$

From Assumption 1, $p_\theta^{\boldsymbol{\alpha}(\mathbf{x})}(\mathbf{x}) \leq \exp(\Delta_{\boldsymbol{\alpha}}) p_\theta^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y}), \forall \theta \in \Theta$, so

$$\phi^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y}) = \int_\Theta p_\theta^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y}) d\xi(\theta) \leq \exp(\Delta_{\boldsymbol{\alpha}}) \int_\Theta p_\theta^{\boldsymbol{\alpha}(\mathbf{x})}(\mathbf{x}) d\xi(\theta) = \exp(\Delta_{\boldsymbol{\alpha}}) \phi^{\boldsymbol{\alpha}}(\mathbf{x}), \tag{23}$$

which gives

$$\sup_{\mathbf{x},\mathbf{y} \in \mathcal{X}^n : \delta(\mathbf{x},\mathbf{y})=1} D_{KL}\left[(\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\cdot \mid \mathbf{x}) \parallel \xi^{\boldsymbol{\alpha}(\mathbf{y})}(\cdot \mid \mathbf{y}))\right] \leq 2\Delta_{\boldsymbol{\alpha}}. \tag{24}$$

A.2 Proof for Theorem 5

From Assumption 1, $\frac{p_\theta^{\boldsymbol{\alpha}(\mathbf{x})}(\mathbf{x})}{p_\theta^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y})} \leq \exp(\Delta_{\boldsymbol{\alpha}})$. From Theorem 4, we show $\phi^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y}) \leq \exp(\Delta_{\boldsymbol{\alpha}}) \phi^{\boldsymbol{\alpha}(\mathbf{x})}(\mathbf{x})$. Then, $\forall \mathbf{x} \in \mathcal{X}^n$ and for each $\mathbf{x}$, $\forall \mathbf{y} \in \mathcal{X}^n : \delta(\mathbf{x}, \mathbf{y} = 1)$,

$$\begin{aligned} \xi^{\boldsymbol{\alpha}(\mathbf{x})}(B \mid \mathbf{x}) &= \frac{\int_B \frac{p_\theta^{\boldsymbol{\alpha}(\mathbf{x})}(\mathbf{x})}{p_\theta^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y})} p_\theta^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y}) d\xi(\theta)}{\phi^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y})} \cdot \frac{\phi^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y})}{\phi^{\boldsymbol{\alpha}(\mathbf{x})}(\mathbf{x})} \\ &\leq \exp(2\Delta_{\boldsymbol{\alpha}}) \xi^{\boldsymbol{\alpha}(\mathbf{y})}(B \mid \mathbf{y}). \end{aligned} \tag{25}$$

A.3 Proof for Lemma 6

$$\begin{aligned} P^{\boldsymbol{\alpha}(\mathbf{x})}(\boldsymbol{\zeta} \in C \mid \mathbf{x}) &= \int P(\boldsymbol{\zeta} \in C \mid \mathbf{x}, \theta) d\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\theta \mid \mathbf{x}) \\ &= \int P(\boldsymbol{\zeta} \in C \mid \theta) d\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\theta \mid \mathbf{x}) \\ &= \int P(\boldsymbol{\zeta} \in C \mid \theta) \frac{d\xi^{\boldsymbol{\alpha}(\mathbf{x})}(\theta \mid \mathbf{x})}{d\xi^{\boldsymbol{\alpha}(\mathbf{y})}(\theta \mid \mathbf{y})} d\xi^{\boldsymbol{\alpha}(\mathbf{y})}(\theta \mid \mathbf{y}) \\ &\leq e^\epsilon \int P(\boldsymbol{\zeta} \in C \mid \theta) d\xi^{\boldsymbol{\alpha}(\mathbf{y})}(\theta \mid \mathbf{y}) \\ &= e^\epsilon P^{\boldsymbol{\alpha}(\mathbf{y})}(\boldsymbol{\zeta} \in C \mid \mathbf{y}). \end{aligned} \tag{26}$$

A.4 Proof for Lemma 7

Choose a data set $\mathbf{x} \in \mathcal{X}^n$. Then let $\mathbf{y}$ be a data set such that $\mathbf{y} \in \mathcal{X}^n$ and the Hamming distance $\delta(\mathbf{x}, \mathbf{y}) = 1$ .We can express every such $\mathbf{y}$ as $\mathbf{y} = (x_1, x_2, \ldots, x_{j-1}, y_j, x_{j+1}, \ldots, x_n)$, where all but the j^{th} element matches between $\mathbf{x}$ and $\mathbf{y}$: $x_i = y_i$ for $i \neq j$ and $x_j \neq y_j$. Then for each $\theta \in \Theta$:

$$\begin{aligned} \ell^{\boldsymbol{\alpha}}(\theta) &= \sup_{\mathbf{x},\mathbf{y} \in \mathcal{X}^n : \delta(\mathbf{x},\mathbf{y})=1} \left| f_\theta^{\boldsymbol{\alpha}(\mathbf{x})}(\mathbf{x}) - f_\theta^{\boldsymbol{\alpha}(\mathbf{y})}(\mathbf{y}) \right| \\ &= \sup_{x_j, y_j \in \mathcal{X}} \left| f_\theta^{\alpha(x_j)}(x_j) - f_\theta^{\alpha(y_j)}(y_j) \right| \\ &\leq \left| \sup_{z \in \mathcal{X}} f_\theta^{\alpha(z)}(z) - \inf_{z \in \mathcal{X}} f_\theta^{\alpha(z)}(z) \right| \end{aligned} \tag{27}$$

Under the assumption $p(x) \leq 1$, the last quantity becomes

$$\begin{aligned} \left| \sup_{z \in \mathcal{X}} f_\theta^{\alpha(z)}(z) - \inf_{z \in \mathcal{X}} f_\theta^{\alpha(z)}(z) \right| &\leq \left| 0 - \inf_{z \in \mathcal{X}} f_\theta^{\alpha(z)}(z) \right| \\ &= \sup_{z \in \mathcal{X}} \left| f_\theta^{\alpha(z)}(z) \right| \\ &= \ell_{loo}^{\boldsymbol{\alpha}}(\theta). \end{aligned} \tag{28}$$

If we allow for the unusual case that a density $p(x) > 1$ for some $x \in \mathcal{X}$, then the bound is larger but still related to the leave-one-out formulation:

$$\begin{aligned} \left| \sup_{z \in \mathcal{X}} f_\theta^{\alpha(z)}(z) - \inf_{z \in \mathcal{X}} f_\theta^{\alpha(z)}(z) \right| &\leq \left| \sup_{z \in \mathcal{X}} f_\theta^{\alpha(z)}(z) \right| + \left| \inf_{z \in \mathcal{X}} f_\theta^{\alpha(z)}(z) \right| \\ &= \sup_{z \in \mathcal{X}} \left| f_\theta^{\alpha(z)}(z) \right| + \sup_{z \in \mathcal{X}} \left| f_\theta^{\alpha(z)}(z) \right| \\ &= 2\ell_{loo}^{\boldsymbol{\alpha}}(\theta). \end{aligned} \tag{29}$$

where $\sup_{z \in \mathcal{X}} \left| f_\theta^{\alpha(z)}(z) \right| = \max \left\{ \left| \sup_{z \in \mathcal{X}} f_\theta^{\alpha(z)}(z) \right|, \left| \inf_{z \in \mathcal{X}} f_\theta^{\alpha(z)}(z) \right| \right\}$.

A.5 Proof of Theorem 8

Let us the define the following subset of $\theta \in \Theta$,

$$U_n = \left\{ \theta \in \Theta : \left[(1 - \alpha_m) D_{\theta_0, \boldsymbol{\alpha}}^{(n_A)}(\theta, \theta^*) + (1 - \alpha^{(n)}) D_{\theta_0, 1^-}^{(n_Q)}(\theta, \theta^*) \right] \geq (D + 3t) n \tau_n^2 \right\},$$

which is the restricted set for which we will bound the pseudo posterior distribution, $\xi^{\boldsymbol{\alpha}}(U_n \mid \mathbf{x})$, from above to achieve the result of Theorem 8. We begin with the statement and proof of Lemma 9 that extends Lemma 8.1 of Ghosal et al. (2000) to our $\alpha-$pseudo posterior in order to provide a concentration inequality to probabilistically (in $P_{\theta_0}-$probability) bound the denominator of the $\boldsymbol{\alpha}-$pseudo posterior distribution, $\xi^{\boldsymbol{\alpha}}(U_n \mid \mathbf{x})$, from below.

A.5.1 Enabling Lemma

Lemma 9 *(Concentration Inequality) Suppose Assumption 3 holds. Define* $\alpha_m = \max_{i \in A_n} \alpha_i$ *and* $\alpha_l = \min_{i \in A_n} \alpha_i$. *For every* $\tau_n > 0$ *and measure* Π *on the set* $B_n(\theta^*, \xi; \theta_0)$, *we have for every* $C_1^* = \sqrt{2 + C_1^2 + C_3^2}$, *and* n *sufficiently large,*

$$P_{\theta_0}\left\{ \int\limits_{\theta \in B_n} e^{-r_{n,\boldsymbol{\alpha}}(\theta,\theta^*)} \xi(d\theta) \leq e^{-\alpha_m (D+t) n \tau_n^2} \right\} \leq \frac{(1+\alpha_l^2)(C_1^*)^2}{\alpha_m^2} \times \frac{1}{(D+t-1)^2 n \tau_n^2}, \quad (30)$$

where the above probability is taken with the respect to P_{θ_0}.

Proof The proof follows that of Savitsky and Toth (2016) by bounding the probability expression on left-hand size of Equation (30). We construct an $\boldsymbol{\alpha}-$weighted empirical distribution that we will need for the proof with,

$$\mathbb{P}_{n,\boldsymbol{\alpha}} = \frac{1}{n} \sum_{i=1}^{n} \alpha_i \delta(x_i), \quad (31)$$

where $\delta(x_i)$ denotes the Dirac delta function with probability mass 1 at x_i. We construct the associated scaled and centered empirical process, $\mathbb{G}_{n,\boldsymbol{\alpha}} = \sqrt{n}(\mathbb{P}_{n,\boldsymbol{\alpha}} - P_{\theta_0})$. The usual equally-weighted empirical distribution, $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta(x_i)$ and associated, $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_{\theta_0})$ may be viewed as special cases. We may define the associated expectation functionals with respect to the $\boldsymbol{\alpha}-$weighted empirical distribution by $\mathbb{P}_{n,\boldsymbol{\alpha}} g = \frac{1}{n}\sum_{i=1}^n \alpha_i g(x_i)$.

Using Jensen's inequality,

$$\begin{aligned} &\log \int\limits_{\theta \in B_n} \prod_{i=1}^{n} \left[\frac{p_{\theta_i}}{p_{\theta_i^*}}(X_i) \right]^{\alpha_i} \xi(d\theta) \\ &\geq \sum_{i=1}^{n} \int\limits_{\theta \in B_n} \alpha_i \log \frac{p_{\theta_i}}{p_{\theta_i^*}} \xi(d\theta) \quad (32) \\ &= n \mathbb{P}_{n,\boldsymbol{\alpha}} \int\limits_{\theta \in B_n} \log \frac{p_\theta}{p_{\theta^*}} \xi(d\theta) \end{aligned}$$

We may use the above to now bound the left-hand size of Equation (30)

$$P_{\theta_0}\left\{\int_{\theta\in B_n} e^{-r_{n,\boldsymbol{\alpha}}(\theta,\theta^*)}\xi\left(d\theta\right) \leq e^{-\alpha_m(D+t)n\tau_n^2}\right\} \tag{33a}$$

$$\leq P_{\theta_0}\left\{n\mathbb{P}_{n,\boldsymbol{\alpha}}\int_{\theta\in B_n}\log\frac{p_\theta}{p_{\theta^*}}\xi\left(d\theta\right) \leq -\alpha_m(D+t)n\tau_n^2\right\} \tag{33b}$$

$$= P_{\theta_0}\left\{\mathbb{G}_{n,\boldsymbol{\alpha}}\int_{\theta\in B_n}\log\frac{p_\theta}{p_{\theta^*}}\xi\left(d\theta\right) \leq -\alpha_m(D+t)n\tau_n^2 - \sqrt{n}P_{\theta_0}\log\frac{p_\theta}{p_{\theta^*}}\xi\left(d\theta\right)\right\} \tag{33c}$$

$$\leq P_{\theta_0}\left\{\mathbb{G}_{n,\boldsymbol{\alpha}}\int_{\theta\in B_n}\log\frac{p_\theta}{p_{\theta^*}}\xi\left(d\theta\right) \leq -\alpha_m(D+t)\sqrt{n}\tau_n^2 - \sqrt{n}\tau_n^2\right\} \tag{33d}$$

$$= P_{\theta_0}\left\{\mathbb{G}_{n,\boldsymbol{\alpha}}\int_{\theta\in B_n}\log\frac{p_\theta}{p_{\theta^*}}\xi\left(d\theta\right) \leq -\alpha_m(D+t-1)\sqrt{n}\tau_n^2\right\}, \tag{33e}$$

where the bound in Equation (33d) uses the prior mass result from Assumption 3. We proceed to use Chebyshev to bound the resultant probability, as follows:

$$\begin{aligned} &P_{\theta_0}\left\{\mathbb{G}_{n,\boldsymbol{\alpha}}\int_{\theta\in B_n}\log\frac{p_\theta}{p_{\theta^*}}\xi\left(d\theta\right) \leq -\alpha_m(D+t-1)\sqrt{n}\tau_n^2\right\} \\ &\leq \frac{\int_{\theta\in B_n}\left[\mathbb{E}_{P_{\theta_0}}\left(\mathbb{G}_{n,\boldsymbol{\alpha}}\log\frac{p_\theta}{p_{\theta^*}}\right)^2\right]\xi\left(d\theta\right)}{\alpha_m^2(D+t-1)^2n\tau_n^4}, \end{aligned} \tag{34}$$

where we have applied Fubini to the right side of Equation (34) to move the expectation through the integral. We now proceed to further bound the expression in brackets on the right-hand side of Equation (34) from above. We may decompose the expectation, as follows

$$\mathbb{E}_{P_{\theta_0}}\left(\mathbb{G}_{n,\boldsymbol{\alpha}}\log\frac{p_\theta}{p_{\theta^*}}\right)^2 \leq n\mathbb{E}_{P_{\theta_0}}\left(\mathbb{P}_{n,\boldsymbol{\alpha}}\log\frac{p_\theta}{p_{\theta^*}} - \mathbb{P}_n\log\frac{p_\theta}{p_{\theta^*}}\right)^2 + \mathbb{E}_{P_{\theta_0}}\left(\mathbb{G}_{n,\boldsymbol{\alpha}}\log\frac{p_\theta}{p_{\theta^*}}\right)^2 \tag{35}$$

We first bound the second term on the right,

$$\mathbb{E}_{P_{\theta_0}}\left(\mathbb{G}_{n,\boldsymbol{\alpha}}\log\frac{p_\theta}{p_{\theta^*}}\right)^2 \tag{36a}$$

$$\leq \mathbb{E}_{P_{\theta_0}}\left(\sqrt{n}\mathbb{P}_{n,\boldsymbol{\alpha}}\log\frac{p_\theta}{p_{\theta^*}}\right)^2 \tag{36b}$$

$$\leq \mathbb{E}_{P_{\theta_0}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n\log\frac{p_\theta}{p_{\theta^*}}\right)^2 \tag{36c}$$

$$\leq \frac{1}{n}\sum_{i=1}^n\mathbb{E}_{P_{\theta_0}}\left(\log\frac{p_\theta}{p_{\theta^*}}\right)^2 \tag{36d}$$

$$\leq \frac{1}{n}\times n\tau_n^2=\tau_n^2, \tag{36e}$$

where we use independence of the X_i to establish the fourth equation and Assumption 3 to achieve the fifth equation.

We proceed to further simplify the bound in the first term on the right in Equation (35):

$$n\mathbb{E}_{P_{\theta_0}}\left(\mathbb{P}_{n,\boldsymbol{\alpha}}\log\frac{p_\theta}{p_{\theta^*}}-\mathbb{P}_n\log\frac{p_\theta}{p_{\theta^*}}\right)^2 \tag{37a}$$

$$=n\mathbb{E}_{P_{\theta_0}}\left(\frac{1}{n}\sum_{i=1}^n(\alpha_i-1)\log\frac{p_{\theta_i}}{p_{\theta_i^*}}\right)^2 \tag{37b}$$

$$=\frac{1}{n}\sum_{i,j=1}^n\mathbb{E}_{P_{\theta_0}}\left[(\alpha_i-1)(\alpha_j-1)\log\frac{p_{\theta_i}}{p_{\theta_i^*}}(X_i)\log\frac{p_{\theta,j}}{p_{\theta^*,j}}(X_j)\right] \tag{37c}$$

$$\begin{aligned}&=\frac{1}{n}\sum_{i=j=1}^n\mathbb{E}_{P_{\theta_0}}\left[(\alpha_i-1)^2\log\frac{p_{\theta_i}}{p_{\theta_i^*}}(X_i)^2\right]\\&+\frac{1}{n}\sum_{i\neq j=1}^n\mathbb{E}_{P_{\theta_0}}\left|\left[(\alpha_i-1)(\alpha_j-1)\log\frac{p_{\theta_i}}{p_{\theta_i^*}}(X_i)\log\frac{p_{\theta,j}}{p_{\theta^*,j}}(X_j)\right]\right|\end{aligned} \tag{37d}$$

$$\leq\frac{1}{n}\left\{(1-\alpha_l)^2\sum_{i\neq j=1}^n\mathbb{E}_{P_{\theta_0}}\left[\log\frac{p_{\theta_i}}{p_{\theta_i^*}}(X_i)^2\right]\right\}$$

$$+\frac{1}{n}(1-\alpha_l)^2\sum_{i\neq j\in A_n}\left|\mathbb{E}_{P_{\theta_0}}\log\frac{p_{\theta_i}}{p_{\theta_i^*}}(X_i)\log\frac{p_{\theta,j}}{p_{\theta^*,j}}(X_j)\right| \tag{37e}$$

$$+\frac{1}{n}\left(1-\alpha^{(n)}\right)^2\sum_{i\neq j\in Q_n}\left|\mathbb{E}_{P_{\theta_0}}\log\frac{p_{\theta_i}}{p_{\theta_i^*}}(X_i)\log\frac{p_{\theta,j}}{p_{\theta^*,j}}(X_j)\right|$$

$$\leq\frac{1}{n}\left\{(1-\alpha_l)^2n\tau_n^2\right\}+\frac{1}{n}(1-\alpha_l)^2\left(C_1^2n-C_1\sqrt{n}\right)\tau_n^2+n_Q\frac{C_3^2\tau_n^2}{n_Q} \tag{37f}$$

$$=\left\{(1-\alpha_l)^2\tau_n^2\right\}+(1-\alpha_l)^2C_1^2\tau_n^2+C_3^2\tau_n^2, \tag{37g}$$

for sufficiently large n. The bound in Equation (37f) results from the restriction of θ to $B_n(\theta^*, \eta; \theta_0)$ and also from Assumption 4 that regulates the growth of the number of $\alpha_i < 1^-$ and the magnitude of $(1 - \alpha^{(n)})$.

We may now bound the expectation on the right-hand size of Equation (34),

$$\mathbb{E}_{P_{\theta_0}}\left(\mathbb{G}_{n,\boldsymbol{\alpha}} \log \frac{p_\theta}{p_{\theta^*}}\right)^2 \leq \left\{(1-\alpha_l)^2 \tau_n^2\right\}(1-\alpha_l)^2 C_1^2 \tau_n^2 + \tau_n^2 \tag{38a}$$

$$\leq \left\{\left(1 - 2\alpha_l + \alpha_l^2\right)\tau_n^2 + \left(1 - 2\alpha_l + \alpha_l^2\right) C_1^2 \tau_n^2 + C_3^2 \eta n^2 + \tau_n^2\right\} \tag{38b}$$

$$\leq (2 + C_1^2 + C_3^2)\tau_n^2 + (1 + C_1^2)\alpha_l^2 \tau_n^2 \leq (1 + \alpha_l)^2 (C_1^*)^2 \tau_n^2 \tag{38c}$$

for n sufficiently large, where we set $C_1^* := \sqrt{C_1^2 + C_3^2 + 2}$. This concludes the proof. ■

A.5.2 Proof of Theorem 8

We begin by constructing the $\boldsymbol{\alpha}-$pseudo posterior distribution on the set, U_n,

$$\xi^{\boldsymbol{\alpha}}(U_n \mid \mathbf{x}) = \frac{\int_{U_n} e^{-r_{n,\boldsymbol{\alpha}}(\theta,\theta^*)}\xi(d\theta)}{\int_\Theta e^{-r_{n,\boldsymbol{\alpha}}(\theta,\theta^*)}\xi(d\theta)}. \tag{39}$$

We next bound the numerator from above in $P_{\theta_0}-$ probability.

$$\mathbb{E}_{P_{\theta_0}} \int_{U_n} e^{-r_{n,\boldsymbol{\alpha}}(\theta,\theta^*)}\xi(d\theta) \tag{40a}$$

$$= \int_{U_n} A^{(n)}_{\theta_0,\boldsymbol{\alpha}}(\theta, \theta^*)\,\xi(d\theta) \tag{40b}$$

$$= \int_{U_n} e^{-\sum_{i=1}^n (1-\alpha_i) D_{\theta_0,\boldsymbol{\alpha},i}}\xi(d\theta) \tag{40c}$$

$$\leq \int_{U_n} e^{-(1-\alpha_m)\sum_{i\in A_n} D_{\theta_0,\boldsymbol{\alpha},i} - \left(1-\alpha^{(n)}\right)\sum_{i\in Q_n} D_{\theta_0,1^-,i}}\,\xi(d\theta) \tag{40d}$$

$$\leq e^{-(D+3t)n\tau_n^2}, \tag{40e}$$

where we use Fubini to switch the order of expectation and integration in Equation (40b). We achieve the bound in Equation (40d) since $D_{\theta_0,\boldsymbol{\alpha},i} > 0,\ \forall i \in (1,\ldots,n)$ and Bhattacharya et al. (2019) shows that $D^{(n)}_{\theta_0,1^-}(\theta,\theta^*)$ is finite and contracts on the KL divergence. The final bound uses the definition of U_n.

We proceed to use the Markov inequality and the definition for U_n to achieve the numerator bound with respect to $P_{\theta_0}-$probability,

$$P_{\theta_0}\left\{\int_{U_n} e^{-r_{n,\boldsymbol{\alpha}}(\theta,\theta^*)}\xi(d\theta) \geq e^{-(D+2t)n\tau_n^2}\right\} \tag{41a}$$

$$\leq \frac{e^{-(D+3t)n\tau_n^2}}{e^{-(D+2t)n\tau_n^2}} = e^{-tn\tau_n^2} \leq \frac{(1+\alpha_l^2)(C_1^*)^2}{\alpha_m^2 (D-1+t)^2 n\tau_n^2}. \tag{41b}$$

We, next, turn to bounding the denominator of Equation (39), from below. Since,

$$\int\limits_{\theta\in\Theta} e^{-r_{n,\boldsymbol{\alpha}}(\theta,\theta^*)}\xi(d\theta) \geq \int\limits_{\theta\in B_n} e^{-r_{n,\boldsymbol{\alpha}}(\theta,\theta^*)}\xi(d\theta),$$

we may use the result of Lemma 9 in,

$$P_{\theta_0}\left\{\int\limits_{\theta\in\Theta} e^{-r_{n,\boldsymbol{\alpha}}(\theta,\theta^*)}\xi(d\theta) \geq e^{-\alpha_m(D+t)n\tau_n^2}\right\} > 1 - \frac{(1+\alpha_l^2)(C_1^*)^2}{\alpha_m^2(D-1+t)^2 n\tau_n^2}. \tag{42}$$

Finally, combining the results of Equations (39), (41) and (42): With probability at least $1 - \left[2/(D+t-1)^2 n\tau_n^2 \times (1+\alpha_l^2(C_1^*)^2)/\alpha_m^2\right]$,

$$\begin{aligned}\xi^{\boldsymbol{\alpha}}\left(\left[(1-\alpha_m)D^{(n_A)}_{\theta_0,\boldsymbol{\alpha}}(\theta,\theta^*) + (1-\alpha^{(n)})D^{(n_Q)}_{\theta_0,1^-}(\theta,\theta^*)\right] \geq (D+3t)n\tau_n^2 \big| \mathbf{x}\right) \leq \\ e^{-(D+2t)n\tau_n^2} e^{\alpha_m(D+t)n\tau_n^2} \\ \leq e^{-tn\tau_n^2}\end{aligned}$$

Appendix B. Unweighted, Non-private Synthesizer

Our description of the unweighted, non-private synthesizer follows closely of that in Hu and Savitsky (2019). To simulate partially synthetic data for the CE sample, where only the sensitive, continuous family income variable is synthesized, we propose using a flexible, parametric finite mixture synthesizer.

Equation (43) and Equation (44) present the first two levels of the hierarchical parametric finite mixture synthesizer: y_i is the logarithm of the family income for CU i, and $\boldsymbol{x}_i$ is the $R \times 1$ predictor vector for CU i. The finite mixture utilizes a hyperparameter for the maximum number of mixture components (i.e., clusters), K, that is to set to be over-determined to permit the flexible clustering of CUs. A subset of CUs that are assigned to cluster, k, employ the same generating parameters for y, (β_k^*, σ_k^*), that we term a "location". Locations, $(\boldsymbol{\beta}^*, \boldsymbol{\sigma}^*)$, and the $n \times 1$ vector of cluster indicators, $z_i \in (1, \cdots, K)$, are all sampled for each CU, $i \in (1, \ldots, n)$.

$$\begin{aligned} y_i \mid \mathbf{X}_i, z_i, \mathbf{B}^*, \boldsymbol{\sigma}^* &\sim \text{Normal}(y_i \mid \mathbf{x}_i^{'}\boldsymbol{\beta}^*_{z_i}, \sigma^*_{z_i}), &(43)\\ z_i \mid \pi &\sim \text{Multinomial}(1; \pi_1, \cdots, \pi_K), &(44)\end{aligned}$$

where the $K \times R$ matrix of regression locations, $\mathbf{B}^* = (\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*)^{'}$, denote cluster-indexed regression coefficients for R predictors. The $(\pi_1, \ldots, \pi_K)$ are, in turn, assigned a sparsity inducing Dirichlet distribution with hyperparameters specified as α/K for $\alpha \in \mathbb{R}^+$. We next describe our prior specification.

We induce sparsity in the number of clusters with,

$$\begin{aligned}(\pi_1, \ldots, \pi_K) &\sim \text{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right), &(45)\\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha). &(46)\end{aligned}$$

We specify multivariate Normal priors for each regression coefficient vector of coefficient locations, $\boldsymbol{\beta}_k^*$,

$$\boldsymbol{\beta}_k^* \overset{\text{iid}}{\sim} \text{MVN}_R(\mathbf{0}, \text{diag}(\boldsymbol{\sigma}_\beta) \times \overset{R\times R}{\Omega_\beta} \times \text{diag}(\boldsymbol{\sigma}_\beta)), \tag{47}$$

where the $R \times R$ correlation matrix, Ω_β, receives a uniform prior over the space of $R \times R$ correlation matrices, and each component of $\boldsymbol{\sigma}_\beta$ receives a Student-t prior with 3 degrees of freedom,

$$\sigma_k^* \overset{\text{iid}}{\sim} \text{t}(3, 0, 1). \tag{48}$$

We proceed to describe how to generate partially synthetic data for the CE sample. To implement the finite mixture synthesizer, we first generate sample values of $(\boldsymbol{\pi}^{(l)}, \boldsymbol{\beta}^{*,(l)}, \boldsymbol{\sigma}^{*,(l)})$ from the posterior distribution at MCMC iteration l. Second, for CU i, we generate cluster assignments, $z_i^{(l)}$, from its full conditional posterior distribution given in Hu and Savitsky (2019) using the posterior samples of $\boldsymbol{\pi}^{(l)}$. Lastly, we generate synthetic family income for CU i, $y_i^{*,(l)}$, from Equation (43) given $\boldsymbol{x}_i$, and samples of $z_i^{(l)}, \boldsymbol{\beta}^{*,(l)}$ and $\boldsymbol{\sigma}^{*,(l)}$. We perform these draws for all n CUs, and obtain a partially synthetic dataset, $\boldsymbol{Z}^{(l)}$ at MCMC iteration l. We repeat this process for m times, creating m independent partially synthetic datasets $\boldsymbol{Z} = (\boldsymbol{Z}^{(1)}, \cdots, \boldsymbol{Z}^{(m)})$.

References

J. Abowd and L. Vilhuber. How protective are synthetic data? In J. Domingo-Ferrer and Y. Saygin, editors, *Privacy in Statistical Databases*, volume 5262 of *Lecture Notes in Computer Science*, pages 239–246. Springer, 2008.

D. An and R. J. A. Little. Multiple imputation: an alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society, Series A*, 170:923–940, 2007.

A. Bhattacharya, D. Pati, and Y. Yang. Bayesian fractional posteriors. *The Annals of Statistics*, 47(1):39–66, 2019.

C. M. Bowen and F. Liu. Comparative study of differentially private data synthesis methods. page arXiv:1602.01063, 2016.

C. Dimitrakakis, B. Nelson, Z. Zhang, A. Mitrokotsa, and B. I. P. Rubinstein. Differential privacy for bayesian inference through posterior sampling. *J. Mach. Learn. Res.*, 18(1): 343–381, January 2017. ISSN 1532-4435.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-32731-2, 978-3-540-32731-8. doi: 10.1007/11681878_14.

S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart. Convergence rates of posterior distributions. *Ann. Statist*, pages 500–531, 2000.

J. Hu and T. D. Savitsky. Risk-efficient Bayesian data synthesis for privacy protection. page arXiv:1908.07639, 2019.

A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 277–286. IEEE Computer Society, 2008.

D. McClure and J. P. Reiter. Differential privacy and statistical disclosure risk measures: An illustration with binary synthetic data. *Transactions on Data Privacy*, 5:535–552, 2012.

M. McSherry and K. Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, pages 94–103. 2007.

T. D. Savitsky and D. Toth. Bayesian Estimation Under Informative Sampling. *Electronic Journal of Statistics*, 10(1):1677–1708, 2016.

J. Snoke and A. Slavkovic. pMSE mechanism: Differentially private synthetic data with maximal distributional similarity. In J. Domingo-Ferrer and F. Montes, editors, *Privacy in Statistical Databases*, volume 11126 of *Lecture Notes in Computer Science*, pages 138–159. Springer, 2018.

S. Srivastava, V. Cevher, Q. Dinh, and D. Dunson. WASP: Scalable Bayes via barycenters of subset posteriors. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 912–920, 2015.

L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105:375–389, 2010.

Z. Zhang, B. I. P. Rubinstein, and C. Dimitrakakis. On the differential privacy of Bayesian inference. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 2365–2371. AAAI, 2016.