Data Augmentation MCMC for Bayesian Inference from Privatized Data

Nianqiao P. Ju
Department of Statistics
Purdue University
West Lafayette, IN 47907
nianqiao@purdue.edu

Jordan A. Awan
Department of Statistics
Purdue University
West Lafayette, IN 47907
jawan@purdue.edu

Ruobin Gong
Department of Statistics
Rutgers University
Piscataway, NJ 08854
ruobin.gong@rutgers.edu

Vinayak A. Rao
Department of Statistics
Purdue University
West Lafayette, IN 47907
varao@purdue.edu

Abstract

Differentially private mechanisms protect privacy by introducing additional randomness into the data. Restricting access to only the privatized data makes it challenging to perform valid statistical inference on parameters underlying the confidential data. Specifically, the likelihood function of the privatized data requires integrating over the large space of confidential databases and is typically intractable. For Bayesian analysis, this results in a posterior distribution that is doubly intractable, rendering traditional MCMC techniques inapplicable. We propose an MCMC framework to perform Bayesian inference from the privatized data, which is applicable to a wide range of statistical models and privacy mechanisms. Our MCMC algorithm augments the model parameters with the unobserved confidential data, and alternately updates each one conditional on the other. For the potentially challenging step of updating the confidential data, we propose a generic approach that exploits the privacy guarantee of the mechanism to ensure efficiency. In particular, we give results on the computational complexity, acceptance rate, and mixing properties of our MCMC. We illustrate the efficacy and applicability of our methods on a naïve-Bayes log-linear model as well as on a linear regression model.

1 Introduction

Motivation. Differential privacy [Dwork et al., 2006] presents a formal mathematical framework to protect the confidentiality of individuals and businesses in aggregate data products. It is the state-of-the-art standard for statistical disclosure limitation (SDL), and has become widely adopted by curators of large-scale scientific, commercial, and official databases. Differentially private data products are produced by probabilistic mechanisms that carry proven privacy guarantees. Generally speaking, these mechanisms work by introducing carefully designed random noise into the query of interest, which is an otherwise deterministic function of the underlying database.

The privatization of data products through noise infusion poses a challenge to statistical analysis in the downstream. Statistical estimators are typically complex functions of the data. If instead of the confidential data, the analyst only has access to a probabilistically processed version of them, how can they maintain the statistical validity of the resulting inference?

Preprint. Under review.

A crucial statistical advantage of differentially private mechanisms over traditional SDL counterparts, such as swapping [Dalenius, 1977], is that their probabilistic design is publicly known. This knowledge allows the data analyst to, at least in theory, accurately account for the privatization mechanism and conduct reliable uncertainty quantification. Nevertheless, it remains a substantial computational challenge to incorporate the privacy procedure into the statistical analysis. The challenge is a wide-spread and varied one, as the extra layer of privacy protection calls for the revision of a wide range of existing statistical methodologies that previously operate on the original, non-privatized data, most of which are neither low-dimensional nor simply structured. This is the challenge we address in this work, in which we develop a general computational framework for practitioners to obtain valid statistical inference based on privatized data.

Related literature. Current inferential strategies for privatized data fall into two broad categories. One invokes traditional statistical asymptotics to approximate the sampling distribution of a differentially private statistic, on the grounds that the privacy noise is often asymptotically negligible compared to errors due to sampling [e.g. Smith, 2011, Cai et al., 2021]. These approximations are often inaccurate for finite sample sizes [Wang et al., 2018] and call for specific handling to incorporate the privacy mechanism [e.g. Gaboardi et al., 2016, Wang et al., 2015, Gaboardi and Rogers, 2018].

The second category recognizes (as Section 2 will explain) that the marginal likelihood of the model parameters in (3) is central to the problem of inference from privatized data. The marginal likelihood requires a potentially high-dimensional integral over the space of unobserved confidential databases, and one that is analytically tractable only in a few, simple settings [Awan and Slavković, 2018, 2020]. Typically, one must resort to either approximating it or sampling from it using Monte Carlo methods. Markov chain Monte Carlo (MCMC) techniques have been proposed for specific privacy mechanisms and data generating models. Karwa et al. [2017] propose an MCMC procedure for inference on exponential random graph models, and Bernstein and Sheldon [2018, 2019] devise MCMC methods designed to handle the low-dimensional latent sufficient statistics from exponential family models and linear regression. Gong [2019] shows that for certain differentially private statistics, approximate Bayesian computation (ABC) can give samples that are exact with respect to the marginal likelihood and the Bayesian posterior. In addition, when the statistical model for the confidential data is fully parametric, the parametric bootstrap may be used to produce inference accompanied by uncertainty quantification with better accuracy than asymptotic approximation [e.g. Gaboardi et al., 2016, Ferrando et al., 2020]. Variational Bayesian analysis [Karwa et al., 2015] is another alternative which invokes a non-asymptotic approximation to the posterior distribution.

Our contribution. We develop a general-purpose MCMC framework to perform Bayesian inference on the model parameters underlying the privatized data. Our framework allows us to overcome the intractable marginal likelihood resulting from privatization, and is applicable to a wide range of statistical models and privacy mechanisms. The resulting MCMC algorithms are *exact*, in that they target the posterior distribution precisely, without involving any approximation.

Our approach is general purpose, allowing data analysts to leverage existing inferential tools designed for non-private data. It can be viewed as a flexible, user-friendly wrapper that migrates existing MCMC algorithms for non-private data to the setting of privatized data access, requiring no further algorithm design or tuning. The sampler, formally a Metropolis-within-Gibbs sampler, is presented in Algorithm 1, and only further requires that the analyst can 1) sample from the statistical model for the confidential data and 2) can evaluate the probability density of the noise induced by the privacy mechanism. The algorithm augments the model parameters with the unobserved confidential data, and alternately updates each one conditioned on the other. While the imputation of an entire unobserved database might appear daunting, we demonstrate how knowledge of the privacy mechanism can be exploited to confer performance guarantees to the proposed MCMC algorithm. We provide theoretical results for the computational complexity, Metropolis-Hastings acceptance rate, and mixing properties. In particular, the higher the privacy, the more rapid is our algorithm's exploration of the parameter space. We illustrate the efficacy and applicability of our methods on a privatized naïve-Bayes log-linear model and a linear regression model with clamped and privatized input. Source code in R are available at `https://github.com/nianqiaoju/dataaugmentation-mcmc-differentialprivacy`.

2 Problem Setup

Let $x = (x_1, \ldots, x_n) \in \mathbb{X}^n$ denote the confidential database, containing n records. We assume these records are independent and identically distributed (i.i.d.) draws from a statistical model $f(\cdot \mid \theta)$, though this can be relaxed. The goal of the analyst is to conduct statistical inference on the unknown model parameter $\theta \in \Theta$. A Bayesian analyst represents *a priori* beliefs about θ with a prior probability distribution $p(\theta)$, and seeks to compute a posterior distribution $p(\theta \mid x) \propto p(\theta) f(x \mid \theta)$ that updates their beliefs in light of the observations x. In many modern applications, this posterior distribution is intractable, and it is common for analysts to represent it using samples drawn via some MCMC algorithm. In this work, we will assume access to such a posterior sampling method:

Assumption 1. *The analyst has available a Markov kernel that targets $p(\theta \mid x) \propto p(\theta) f(x \mid \theta)$, the posterior distribution over the model parameters given the confidential database x.*

Differential privacy. Our work here focuses on the following departure from the usual Bayesian setting: instead of observing the database x, we observe a privatized data product or query, denoted as s_{dp}. The quantity s_{dp} is probabilistically generated based on data x through a *privacy mechanism*, written as $\eta(\cdot \mid x)$. The privacy mechanism η is said to be ϵ-*differentially private* (ϵ-DP) [Dwork et al., 2006] if for all values of s_{dp}, and for all 'neighboring' databases $(x, x') \in \mathbb{X}^n \times \mathbb{X}^n$ differing by one record (denoted by $d(x, x') \leq 1$), the probability ratio is bounded:

$$\frac{\eta(s_{\mathrm{dp}} \mid x)}{\eta(s_{\mathrm{dp}} \mid x')} \leq \exp(\epsilon), \quad \epsilon > 0. \tag{1}$$

The parameter ϵ is called the *privacy loss budget*, and controls how informative s_{dp} is about x. Large values of ϵ guarantee less privacy, while $\epsilon = 0$ corresponds to perfect privacy. A simple and widely used ϵ-differentially private mechanism is the *Laplace mechanism*: for a deterministic query $s : \mathbb{X}^n \to \mathbb{R}^m$, the privatized query is defined as $s_{\mathrm{dp}} = s(x) + u$, where $u = (u_1, \ldots, u_m)$ are i.i.d. Laplace variables. The scale parameter of the Laplace distribution is inversely proportional to ϵ (more privacy requires more noise), and directly proportional to $\Delta(s) = \max_{(x,x') \in \mathbb{X}^n \times \mathbb{X}^n; d(x,x') \leq 1} \|s(x) - s(x')\|_1$, the ℓ_1 *(global) sensitivity* of s (the more sensitive the confidential query is to changes in one record of the database, the more noise we need).

Our methodology requires that the privacy mechanism η is known and can be evaluated. This is true of ϵ- (or *pure*) DP, as well as common variants such as (ϵ, δ)- (or *approximate*) DP, *zero-concentrated* DP (zCDP) [Dwork and Rothblum, 2016, Bun and Steinke, 2016], and *Gaussian*-DP [Dong et al., 2021]. To ensure computational efficiency, we make the following additional assumption.

Assumption 2 (Record Additivity)**.** *The privacy mechanism can be written in the form $\eta(s_{\mathrm{dp}} \mid x) = g(s_{\mathrm{dp}}, \sum_{i=1}^n t_i(x_i, s_{\mathrm{dp}}))$ for some known and tractable functions $g, t_1, \ldots, t_n$.*

We refer to privacy mechanisms that satisfy Assumption 2 as *record-additive*. An implication of record additivity is that after changing one record in x, we do not have to scan the entire database to reevaluate η . This is satisfied by many commonly used mechanisms, two important examples being: 1) mechanisms that add data-independent noise to a query of the form $s = \sum_{i=1}^n s_i(x_i)$, such as the sample mean, sample variance-covariance, and sufficient statistics of an exponential family distribution (see Sections 4 and 5 for examples), and 2) mechanisms designed to optimize empirical risk functions of the form $u(x, s_{dp}) = \sum_{i=1}^n u_i(x_i, s_{dp})$, such as the exponential mechanism [McSherry and Talwar, 2007], K-norm gradient mechanism [Reimherr and Awan, 2019], objective perturbation [Chaudhuri et al., 2011, Kifer et al., 2012], and functional mechanism [Zhang et al., 2012].

Doubly intractable Bayesian inference from privatized data. Without access to the confidential database x, and given only the privatized query s_{dp}, the Bayesian analyst is now concerned with the following posterior distribution:

$$p(\theta \mid s_{\mathrm{dp}}) \propto p(\theta) p(s_{\mathrm{dp}} \mid \theta). \tag{2}$$

Here, $p(s_{\mathrm{dp}}|\theta)$ is the *marginal likelihood* of θ, integrating over all possible confidential databases:

$$p(s_{\mathrm{dp}} \mid \theta) = \int_{\mathbb{X}^n} \eta(s_{\mathrm{dp}} \mid x) f(x \mid \theta) dx. \tag{3}$$

The marginal likelihood contributes all the information that is available in the privatized observation s_{dp} about the parameter θ, and is the foundation to statistical inference using privatized statistics

[Williams and McSherry, 2010]. The posterior distribution (2) reveals that the inferential uncertainty about the parameter θ consists of three contributing sources: 1) prior uncertainty as encoded in $p(\theta)$, 2) sampling (or modeling) uncertainty of the confidential database as reflected in f, and 3) uncertainty due to privacy as induced by the probabilistic mechanism η.

We now come to the core challenge to address in this work: the marginal likelihood in (3) calls for an integral over the entire space of possible input databases $x \in \mathbb{X}^n$. This is usually computationally challenging, especially if the privacy mechanism is not a function of a low-dimensional sufficient statistic. If the integral underlying the marginal likelihood is intractable, then $p(s_{\text{dp}} \mid \theta)$ cannot be analytically evaluated. This makes the corresponding posterior distribution $p(\theta \mid s_{\text{dp}})$ of (2) *doubly* intractable [Murray et al., 2012] in the sense that it cannot be analytically evaluated even up to a normalizing constant. Thus, traditional MCMC techniques are inapplicable and inference strategies devised for privatized statistics must tame this possibly high-dimensional integration problem.

3 Data Augmentation MCMC for Inference from Privatized Data

In this paper, we present a simple, efficient, and general *data augmentation* MCMC [Tanner and Wong, 1987, Van Dyk and Meng, 2001] framework, allowing practitioners to perform valid Bayesian inference on a wide-range of data models and privacy mechanisms. Our approach is to augment the MCMC state space with the latent confidential database x, so that the stationary distribution is the *joint* posterior distribution

$$p(\theta, x \mid s_{\text{dp}}) \propto p(\theta) f(x \mid \theta) \eta(s_{\text{dp}} \mid x). \tag{4}$$

Marginally, the θ samples produced by such an algorithm follow the posterior $p(\theta \mid s_{\text{dp}})$ in (2). Our sampler is *exact*, targeting the marginal posterior distribution $p(\theta|s_{\text{dp}})$ without any approximation error, despite the fact that the marginal likelihood (3) is intractable.

Our approach of imputing the latent confidential database x is motivated by two factors: 1) we wish our algorithm to be *general-purpose*, applicable to a wide range of models and privacy mechanisms, and 2) we wish our algorithm to inherit guarantees on mixing performance from guarantees of the privacy mechanism. Towards these ends, we do not assume any specific form of the underlying model of x and the privacy mechanism beyond Assumptions 1 and 2 respectively. In this light, our contribution can be viewed as a flexible wrapper that allows existing MCMC algorithms for models of the confidential data to be extended to settings where the data is now protected by some privacy mechanism. Though imputing the confidential dataset might appear to present a significant challenge, we show that properties of the mechanism can be exploited to give performance guarantees on our sampling scheme, and show that it has a runtime of the same order as the non-private sampler.

In what follows, we outline our proposed MCMC algorithm, derive guarantees on the runtime and acceptance rate of the algorithm, and provide mild conditions for the proposed samplers to be ergodic, as well as additional conditions for our sampler to achieve geometric rates of convergence.

3.1 A Privacy-Aware Metropolis-within-Gibbs Sampler

Our approach to sample from the joint posterior distribution $p(\theta, x \mid s_{\text{dp}})$ is through a sequence of alternating Gibbs updates. Let $(x^{(t)}, \theta^{(t)})$ denote the state of the Gibbs sampler at the t-th iterations. Each iteration of the Gibbs sampler entails two steps:
(**Step 1**) sample $\theta^{(t+1)}$ from $p(\cdot \mid x^{(t)}, s_{\text{dp}})$, and (**Step 2**) sample $x^{(t+1)}$ from $p(\cdot \mid \theta^{(t+1)}, s_{\text{dp}})$.

The conditional distribution in Step 1 simplifies as $p(\theta|x^{(t)}, s_{\text{dp}}) = p(\theta|x^{(t)})$, highlighting why data-augmentation is useful: this conditional distribution is independent of the privacy mechanism, and we can use existing sampling algorithms (Assumption 1) for the confidential data. We note that with the exception of a few models, such as simple models with conjugate priors, it is usually not possible to directly sample from $p(\theta|x)$. Assumption 1 however only requires that we can *conditionally* simulate a new value of θ from a Markov kernel that has $p(\theta|x^{(t)})$ as its stationary distribution. Our overall Gibbs sampler then becomes a Metropolis-within-Gibbs sampler [Gilks et al., 1995], that nevertheless targets the joint posterior $p(\theta, x|s_{\text{dp}})$.

Step 2 is the data-augmentation step, and connects the statistical model and the privacy mechanism on x. Again, we cannot expect to produce a conditionally independent sample of the latent database from

$p(x \mid \theta, s_{\text{dp}})$, as this is model and mechanism dependent. Instead, we take the much more tractable approach of cycling through the elements of latent database x, sequentially updating x one element of a time. Writing $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ to denote the vector x excluding the ith element, step 2 then consists of the sequence of updates $p(x_1 \mid \theta, x_{-1}, s_{\text{dp}}), p(x_2 \mid \theta, x_{-2}, s_{\text{dp}}), \ldots, p(x_n \mid \theta, x_{-n}, s_{\text{dp}})$. The complete sweep can be viewed as a dependent update of the latent database x that targets the conditional distribution $p(x \mid \theta, s_{\text{dp}})$.

Before we specify our complete sampler in Algorithm 1, we first address the following questions: (Q1) *What is the performance loss from updating x one element at a time, rather than jointly?*, and (Q2) *How can we efficiently carry out the conditional updates $p(x_i \mid \theta, x_{-i}, s_{\text{dp}}),\ i = 1, \ldots, n$?*

Q1 concerns whether the dependence of x_i given (x_{-i}, s_{dp}) is so strong as to impede efficient exploration of the $\mathbb{X}^n$-space and cause poor mixing. Here we note that the privacy mechanism limits the change in the likelihood $\eta(s_{\text{dp}} \mid x)$ when one element of x is changed, and therefore limits the coupling between x_i and x_{-i}. This suggests a Gibbs sweep through the latent database x will not suffer from poor mixing.

Algorithm 1 One iteration of the privacy-aware Metropolis-within-Gibbs sampler

1. Conditional update of $p(\theta \mid x)$ using the kernel from Assumption 1.
2. For each $i = 1, 2, \ldots$, sequentially update $x_i \mid x_{-i}, \theta, s_{\text{dp}}$.
   (a) Propose $x_i^\star \sim f(\cdot \mid \theta)$.
   (b) Update $t(x^\star, s_{dp}) = t(x, s_{dp}) - t_i(x_i, s_{dp}) + t_i(x_i^\star, s_{dp})$ according to Assumption 2.
   (c) Accept the proposed state with probability $\alpha(x_i^\star \mid x_i, x_{-i}, \theta)$ given by:

$$\alpha(x_i^\star \mid x_i, x_{-i}, \theta) = \min\left\{\frac{\eta(s_{dp} \mid x_i^\star, x_{-i})}{\eta(s_{dp} \mid x_i, x_{-i})}, 1\right\} = \min\left\{\frac{g(s_{\text{dp}}, t(x^\star, s_{\text{dp}}))}{g(s_{\text{dp}}, t(x, s_{\text{dp}}))}, 1\right\}. \quad (5)$$

Q2 recognizes that the conditionals $p(x_i|\theta, s_{\text{dp}}, x_{-i})$ are model- and mechanism-specific, and simulating from these is challenging in most settings. For this, we take the following simple approach in Algorithm 1: at each step, we propose x_i from the model $f(x|\theta)$, and accept it with the appropriate MH probability (5). Observe that our choice of proposal distribution is independent of the privacy mechanism, and ignores the privatized data s_{dp} as well as all other elements x_{-i}. Despite being simple and general purpose, we show in Proposition 3.1 that for $\epsilon-$DP, we can lower-bound the acceptance probability of proposals produced this way by $\exp(-\epsilon)$. This lower bound is key to efficiency: despite the unconstrained nature of the proposal distribution, we can guarantee a minimum acceptance probability. These two facts suggest our sampler will explore the space of databases relatively quickly.

Proposition 3.1. *For a pure ϵ-DP privacy mechanism η, the acceptance probability α from Equation (5) satisfies $\alpha(x_i^\star \mid x_i, x_{-i}, \theta) \geq \exp(-\epsilon)$, for all $\theta, x_{-i}, x_i, x_i^*$.*

The privacy loss budget ϵ is usually understood to be a small constant, which privacy experts recommend be between .01 and 1 [Dwork, 2011]. When $\epsilon = 1$, Proposition 3.1 ensures that the acceptance rate in Algorithm 1 is no less than 36.7%, and as ϵ approaches zero, the bound on the acceptance rate approaches one. Intuitively, this is because as ϵ decreases, the distribution of the privatized data s_{dp} depends less and less on any individual element of the database.

The simplicity of our approach arises through a decoupling of the data model from the privacy mechanism: the former is used to update θ and propose x_i's, while the latter is used to calculate the acceptance probabilities. The next result formalizes the computational efficiency of our approach. Specifically, for any record-additive mechanism, one iteration of our algorithm requires $O(n)$ operations, where n is the size of the latent database. Essentially, this arises because of Assumption 2, which allows the acceptance probability in (5) to be calculated in $O(1)$ time.

Proposition 3.2. *The Gibbs sampler described in Algorithm 1 requires $O(n)$ number of operations to update the full latent database according to $p(x \mid \theta, s_{\text{dp}})$.*

Note that even without privacy, one round of an MCMC procedure typically takes $O(n)$ time. This is because updating θ given the confidential data requires computing the data likelihood $f(x \mid \theta) = \prod_{i=1}^n f(x_i \mid \theta)$, an $O(n)$ operation in general. Thus, as a result of the mild and typical condition that

η is record-additive, our MCMC procedure enjoys the *same order* of runtime as the original MCMC algorithm for confidential data.

The previous two results are key to understanding the efficiency of our approach. In the next section we formally establish geometric ergodicity of the sampler in Theorem 3.3 and 3.4.

Computational complexity. The i.i.d. assumption on the records ensures that step 2a of Algorithm 1 take $O(1)$ time, though this assumption can easily be weakened. Assumption 2 allows steps 2b and 2c to also takes $O(1)$ time. Overall, step 2 of our algorithm then takes $O(n)$ (rather than $O(n^2)$) time, as stated in Proposition 3.2. This matches the typical per-iteration cost of samplers for the non-private posterior distribution required in step 1. Thus, the overall cost of an iteration of our MCMC sampler is $O(n)$, which is typical when dealing with datasets of size n.

3.2 Ergodicity of the Privacy-Aware Sampler

Ergodicity ensures the MCMC chain converges to the posterior distribution in total variation distance. [Tierney, 1994], which is essential for an MCMC sampler to consistently estimate functionals of the posterior distribution. In Theorem 3.3, we provide mild and sufficient conditions for our proposed Metropolis-within-Gibbs sampler to be ergodic.

Theorem 3.3. *Under conditions A1 - A3 below, the Metropolis-within-Gibbs sampler of Algorithm 1 on the joint space $(\mathbb{X}^n \times \Theta)$ is ergodic and it admits $p(x, \theta \mid s_{\text{dp}})$ as the unique limiting distribution.*

A1. The prior distribution is proper and $p(\theta) > 0$ for all θ in $\Theta = \{\theta \mid f_\theta(x) > 0 \text{ for some } x\}$.
A2. The model is such that the set $\{x : f(x \mid \theta) > 0\}$ does not depend on θ.
A3. The privacy mechanism satisfies $\eta(s_{\text{dp}} \mid x) > 0$ for all $x \in \mathbb{X}^n$.

We prove this in the supplementary material by verifying invariance, aperiodicity, and irreducibility. Conditions A1-A3 concern model specification, prior specification, and privacy noise. These mild assumptions are typically true and are easy to verify. While there are some mechanisms, such as the release-one-at-random mechanism, which satisfy approximate-DP but which violate A3 [Barber and Duchi, 2014], most privacy mechanisms of interest satisfy A3. It is easy to verify that if η satisfies ϵ-DP, zero-concentrated DP, or Gaussian-DP, then property A3 is guaranteed.

Next, we establish conditions for Algorithm 1 to be geometrically ergodic [Rosenthal, 1995, Roberts and Rosenthal, 1998]. A chain is said to be *geometrically ergodic* if its total variation distance to the target has a geometrically decaying upper bound. Geometric ergodicity is a desirable property since it provides a rate on convergence to the stationary distribution, guaranteeing central limit theorems, and allowing for the computation of asymptotically valid standard errors.

For simplicity, we focus on the situation where one can directly sample from the conditional posterior $p(\theta \mid x)$. This is an important and common case, relevant when either θ is low-dimensional or where one can place conjugate priors on θ. Both applications we present in this work, a log-linear model in Section 4, and a linear regression model in Section 5, fall under this setting.

Theorem 3.4. *Assume that in step 1 of Algorithm 1, one can directly sample from $p(\theta \mid x)$. Under A1-A3 of Theorem 3.3, the resulting (x, θ) chain, as well as the marginal chains, are geometrically ergodic if η satisfies ϵ-DP and there exists $0 < a \leq b < \infty$ such that $a \leq f(x \mid \theta) \leq b \quad \forall \theta, x$.*

To prove Theorem 3.4, we verify the drift and minorization conditions for component-wise Gibbs samplers in Theorem 8 of [Johnson et al., 2013]. See the supplementary material for details.

Unsurprisingly, geometric ergodicity requires stronger assumptions than just ergodicity. The first assumption concerning the ability to sample directly from $p(\theta|x)$ can be avoided, but for the sake of clarity we do not try to relax it, since we are mostly concerned with the interface with the privacy mechanism. The second assumption on the boundedness of the likelihood is stronger, but also typical. A common way to achieve bounded likelihoods is to require the sample space $\mathbb{X}$ (and typically also Θ) to be bounded. In many real-world settings, such bounds exist, even if they may be very loose.

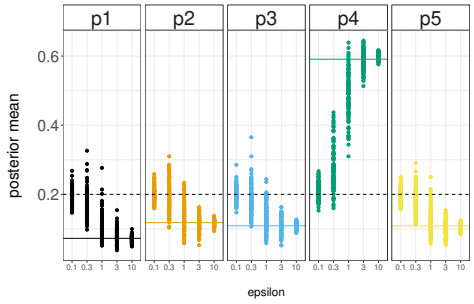
4 Naïve Bayes Log-Linear Model

Log-linear models are often used to model categorical data, a popular instance being the naïve Bayes classifier. Following Karwa et al. [2016], we consider the following model: $x = (x_1, \ldots, x_K)$

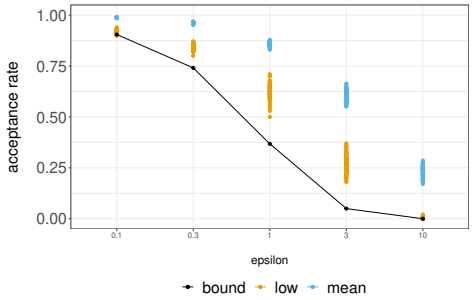
is the input *feature-vector*, with each x_k taking values in $\{1, 2, \ldots, J_k\}$, and y is the output *class* taking values in $\{1, 2, \ldots, I\}$. Each input-output pair (x, y) forms one record in our confidential database, the entire database consisting of N i.i.d. copies of (x, y). The naïve Bayes classifier assumes that $P(x \mid y) = \prod_{k=1}^{K} P(x_k \mid y)$, the model parameters being $p_{ij}^k = P(x_k = j | y = i)$ and $p_i = P(y = i)$. The sufficient statistics of this model are $n_{ij}^k = \#(y = i, x_k = j)$, which count the number of class-feature co-occurrences. This will form our confidential query s, which we privatize by adding Laplace noise to each of the n_{ij}^k. The resulting quantity s_{dp}, consisting of the noisy counts $m_{ij}^k = n_{ij}^k + L_{ijk}$, is what we release. When $L_{ijk} \overset{\text{i.i.d.}}{\sim} \text{Laplace}(0, 2K/\epsilon)$, the output s_{dp} satisfies ϵ-DP. Placing a $\text{Dirichlet}(2, \ldots, 2)$ on all parameter vectors, our goal is to obtain the marginal posterior distribution of $p, p_{i-}^k \mid s_{dp}$. While Karwa et al. [2016] approximate this private posterior distribution using variational Bayes methods, our MCMC procedure is able to target the exact private posterior distribution.

Simulation setup. We perform several simulation experiments where we apply our MCMC samplers to the log-linear model described above. For the simulation, we set $N = 100$ (number of records), $I = 5$ (number of classes), $K = 5$ (number of features), and $J_k = 3$ for all $k = 1, \ldots, K$ (possible values for each feature). We evaluate our sampler for privacy levels corresponding to $\epsilon \in \{.1, .3, 1, 3, 10\}$.

Posterior mean. We generate one non-private dataset from the model, and hold it fixed. We then create 100 private queries s_{dp} at each ϵ value, and for each s_{dp} we run Algorithm 1 for 10000 iterations. We discard the first 5000 iterations as burn-in. Finally, for each chain, we calculate the posterior mean. Figure 1a plots the 100 different posterior means for each ϵ-value for the parameters $p_i = P(Y = i)$ for $i = 1, \ldots, 5$. In this plot, the solid horizontal lines indicate the non-private posterior means, and the dashed horizonal lines indicate the prior means for each parameter. We see that as ϵ approaches zero, the posterior mean approaches the prior mean, reflecting the intuition that we learn less from the data as the privacy budget gets smaller. On the other hand, as ϵ increases, we see that the private posterior mean approaches the non-private posterior mean, which reflects the fact that as ϵ grows, we learn approximately the same from the data as if there were no privacy mechanism.



(a) Posterior means for the log-linear model. The solid horizonal lines indicate the non-private posterior means, and the dashed lines at .2 indicate the prior means.



(b) Observed acceptance rates for the log-linear model. The blue (above) point clouds indicate the average acceptance rate, and the orange (below) points indicate the observed minimum acceptance of each chain. The solid black line is the lower bound of Proposition 3.1.

Acceptance rate. Using the same simulation setup as for the posterior mean, we calculate the average and minimum acceptance rate of Step 2 in Algorithm 1. Since the privacy mechanism satisfies ϵ-DP, we know that $\exp(-\epsilon)$ is a lower bound on these acceptance rates. In Figure 1b, we confirm this bound, and see that the average acceptance rate is significantly higher than this lower bound. This suggests that the chain mixes even faster than indicated by Proposition 3.1.

Coverage of credible intervals. For the next experiment, we sample a set of parameters from the prior, and hold this fixed. Then for each ϵ value, we produce 100 non-private datasets (n_{ij}^k), one private dataset (m_{ij}^k) for each non-private one, and then run a chain for 10,000 iterations discarding

the first 5000 iterations for burn-in. From each chain, we produce a 90% credible interval for each $p_i = P(Y = i)$, and calculate the empirical coverage which is reported in Table 1.

ϵ	$p_1 = .097$	$p_2 = .148$	$p_3 = .145$	$p_4 = .446$	$p_5 = .163$
.1	1	1	1	**.36**	1
.3	.97	1	1	**.59**	1
1	.94	.99	.97	**.83**	.98
3	.95	.91	.97	.89	.93
10	.92	.88	.94	.92	.9

Table 1: Coverage of $p_i = P(Y = i)$ for the log-linear model at different ϵ. Top row is the true data generating parameter values. Coverage is based on 100 replicates.

At a sample size of only $N = 100$, we do not expect the coverage of the credible intervals to match the nominal level of 90%, but we see in Table 1 that most of the coverage values are above .9. Notable exceptions are the coverage of p_4 when ϵ is small. This may be because when ϵ is small, the private posterior is approximately equal to the prior, which is centered at .2; however p_4 is significantly further from .2 than the other parameters, which may explain why the coverage is low in this case.

5 Linear Regression

Next, we consider ordinary linear regression with n subjects and p predictors. We write x_0 for the matrix of predictors excluding the intercept columns, $x = (\underline{1}, x_0)$ for the matrix including the intercept, and y for the vector of outcomes. We model the explanatory variables x_0 as $x_0^i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(m, \Sigma)$ for $i = 1, \ldots, n$, with $y|x$ given by $\mathcal{N}_n(x\beta, \sigma^2 I_n)$. Here I_n is the $n \times n$ identity matrix and $\mathcal{N}_n$ denotes the n-dimentional multivariate Normal distribution. The parameters of interest are β, the $(p+1)-$dimension vector of regression coefficients, with σ, m and Σ assumed known. We use independent $\mathcal{N}(0, 2^2)$ priors for the components of β.

To achieve ϵ-DP via the Laplace mechanism, we require a finite global sensitivity. To achieve this, standard practice in the DP literature is to bound each predictor and response variable in a data-independent fashion. The bounds chosen by the privacy expert are $[a_i, b_i]$ for each instance of x_0^i and $[a_y, b_y]$ for the entries of y, and these values are shared with the analyst.

Definition 5.1. For a real value z, and $a \leq b$, define the *clamp function* $[z]_a^b := \min\{\max\{z, a\}, b\}$. If z is a vector of length d, we use the same notation to apply an entry-wise clamp: $[z]_a^b := ([z_1]_a^b, [z_2]_a^b, \ldots, [z_d]_a^b)^\top$.

Before adding noise for privacy, we first clamp the predictors and response, and then normalize them to take values in $[-1, 1]$: $\widetilde{x}_0^i := (b_i - a_i)^{-1} 2([x_0^i]_{a_i}^{b_i} - a_i) - 1$ and $\widetilde{y} := (b_y - a_y)^{-1} 2([y]_{a_y}^{b_y} - a_y) - 1$. Call $\widetilde{x} := [\underline{1}, \widetilde{x}_0^1, \widetilde{x}_0^2, \ldots, \widetilde{x}_0^p]$ and $x := (\widetilde{x}^\top \widetilde{y}, \widetilde{y}^\top \widetilde{y}, \widetilde{x}^\top \widetilde{x})$. The s is the summary statistic to which we will add noise for privacy. The ℓ_1 sensitivity of s (ignoring duplicate entries of $\widetilde{x}^\top \widetilde{x}$, and the constant entry $(\widetilde{x}^\top \widetilde{x})_{1,1}$) is $\Delta = p^2 + 3p + 3$. To satisfy ϵ-DP, we add independent $\text{Laplace}(0, \Delta/\epsilon)$ noise to each of the $d = \frac{1}{2}(p+1)(p+2) + (p+1)$ unique entries of x, which gives our final private summary s_{dp}. We notice that s is an additive function and each individual's contribution to s is $t(x_i, y_i) = ((\tilde{x}^i)^\top \tilde{y}_i, \tilde{y}_i^2, (\tilde{x}_i)^\top \tilde{x}_i)$. This mechanism producing s_{dp} is record-additive.

Simulation setup. Our experiments focus on posterior inference about β based on S_{dp}. For simplicity, we fix other parameters σ^2, m, Σ at the true data generating parameters (reported in the supplementary materials). When they are unknown, the posterior distributions of these parameters can be estimated by our Gibbs sampler as well. Confidential predictors and responses are clamped with bounds $b = 10$ and $a = -10$. Given a confidential database (x, y), the posterior distribution of β is multivariate Normal and can be sampled directly with a runtime linear in n.

Posterior mean. We generate one confidential dataset (x, y) and hold it fixed. At each ϵ value, we create 100 private outputs s_{dp} and run Gibbs samplers for 10,000 iterations targeting the posterior $\beta \mid s_{\text{dp}}$, discarding the first 5000 iterations. We plot the 100 different posterior means of β in Figure 2. In this plot, the solid horizontal lines indicate posterior means given confidential data (x, y), which

we do not expect to fully recover due to clamping. The posterior quantities display the same trend with respect to change in privacy level as observed in Figure 1b. The other experiments from Section 4 were also run on this linear regression model, and produced similar results. Simulation details, plots and discussion are in the supplementary materials.

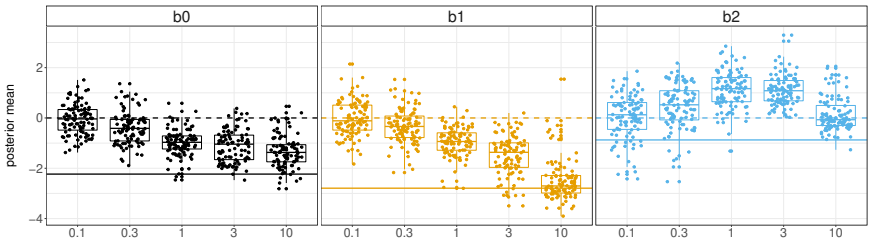


Figure 2: Posterior mean for private linear regression $\beta \mid s_{\text{dp}}$ with fixed confidential data. The solid horizonal lines indicate the confidential data posterior means, and the dashed lines indicate 0.

6 Discussion

We proposed a novel, but simple sampling procedure for parameter inference on $p(\theta \mid s_{\text{dp}})$, which leverage existing samplers for the non-private posterior, as well as the structure of the privacy mechanism. The result is a simple wrapper for practitioners to obtain valid statistical inference from privatized data using the same models for the unobserved confidential data. As a side product, our algorithm also produces multiple copies of the confidential database from the posterior $p(x \mid s_{\text{dp}})$, which could be useful when one is interested in inferring properties of x as well. Although we did not discuss this, our data augmentation scheme can also enable frequentist analysis through the Monte Carlo expectation-maximization algorithm.

We acknowledge the limitations of present work. First, we point out that strong assumptions such as bounded parameter space Θ and sample space $\mathbb{X}$ are required to establish geometric ergodicity of the Gibbs sampler in Theorem 3.4. They can likely be relaxed, at the cost of a more complex theorem statement and proof. Second, our current proposal for updating $x_i \mid x_{-i}, \theta, s_{\text{dp}}$ only tailors to the model $f(\cdot \mid \theta)$ and it is not customized for s_{dp} yet. In the future, we might be able to design algorithms that also incorporate the privatized output s_{dp} in these proposals. Third, we point out that strong correlations in the posterior $p(\theta, x \mid s_{\text{dp}})$ can potentially cause poor mixing in practice, despite geometric convergence rate. While in simple problems, this can be fixed by reparameterization, we plan to develop MCMC algorithms for this setting in follow-up studies. Finally, while the proposed algorithm converges so long as the privacy mechanism η is known, for alternative versions of DP (such as zCDP), the acceptance probability results in Proposition 3.1 may no longer hold. Developing alternatives to Proposition 3.1 for different privacy definitions is another goal of future work.

Acknowledgments and Disclosure of Funding

R. Gong is supported in part by the National Science Foundation grant DMS-1916002, and V. Rao by the National Science Foundation grants RI-1816499 and DMS-1812197.

References

Jordan Alexander Awan and Aleksandra Slavković. Differentially private uniformly most powerful tests for binomial data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4208–4218. Curran Associates, Inc., 2018.

Jordan Alexander Awan and Aleksandra Slavković. Differentially private inference for binomial data. *Journal of Privacy and Confidentiality*, 10(1), 2020.

R. F. Barber and J. C. Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *ArXiv e-prints*, December 2014.

Garrett Bernstein and Daniel R Sheldon. Differentially private bayesian inference for exponential families. *Advances in Neural Information Processing Systems*, 31:2919–2929, 2018.

Garrett Bernstein and Daniel R Sheldon. Differentially private bayesian linear regression. *Advances in Neural Information Processing Systems*, 32:525–535, 2019.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

Kamalika Chaudhuri, Claire Monteleoni, and D. Sarwate. Differentially private empirical risk minimization. In *Journal of Machine Learning Research*, volume 12, pages 1069–1109, 2011.

T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15:429–444, 1977.

Jinshuo Dong, Aaron Roth, and Weijie Su. Gaussian differential privacy. *Journal of the Royal Statistical Society, Series B*, 2021.

Cynthia Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1): 86–95, 2011.

Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Cecilia Ferrando, Shufan Wang, and Daniel Sheldon. General-purpose differentially-private confidence intervals. *arXiv preprint arXiv:2006.07749*, 2020.

Marco Gaboardi and Ryan Rogers. Local private hypothesis testing: Chi-square tests. In *International Conference on Machine Learning*, pages 1626–1635. PMLR, 2018.

Marco Gaboardi, Hyun Lim, Ryan Rogers, and Salil Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International conference on machine learning*, pages 2111–2120. PMLR, 2016.

Wally R Gilks, Nicky G Best, and Keith KC Tan. Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 44(4): 455–472, 1995.

Ruobin Gong. Exact inference with approximate computation for differentially private data via perturbations. *arXiv preprint arXiv:1909.12237*, 2019.

Alicia A Johnson, Galin L Jones, and Ronald C Neath. Component-wise Markov chain Monte Carlo: Uniform and geometric ergodicity under mixing and composition. *Statistical Science*, 28 (3):360–375, 2013.

Vishesh Karwa, Dan Kifer, and Aleksandra B Slavković. Private posterior distributions from variational approximations. *arXiv preprint arXiv:1511.07896*, 2015.

Vishesh Karwa, Dan Kifer, and Aleksandra Slavković. Private posterior distributions from variational approximations. *NIPS 2015 Workshop on Learning and Privacy with Incomplete Data and Weak Supervision*, 2016.

Vishesh Karwa, Pavel N Krivitsky, and Aleksandra B Slavković. Sharing social network data: differentially private estimation of exponential family random-graph models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):481–500, 2017.

D Kifer, A Smith, and A Thakurta. Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research*, 1:1–41, 01 2012.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

Iain Murray, Zoubin Ghahramani, and David MacKay. MCMC for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*, 2012.

Matthew Reimherr and Jordan Awan. KNG: the k-norm gradient mechanism. *Advances in Neural Information Processing Systems*, 32, 2019.

Gareth O Roberts and Jeffrey S Rosenthal. Two convergence properties of hybrid samplers. *The Annals of Applied Probability*, 8(2):397–407, 1998.

Jeffrey S Rosenthal. Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995.

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 813–822, 2011.

Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987.

Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

David A Van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

Yue Wang, Jaewoo Lee, and Daniel Kifer. Revisiting differentially private hypothesis tests for categorical data. *arXiv preprint arXiv:1511.03376*, 2015.

Yue Wang, Daniel Kifer, Jaewoo Lee, and Vishesh Karwa. Statistical approximating distributions under differential privacy. *Journal of Privacy and Confidentiality*, 8(1), 2018.

Oliver Williams and Frank McSherry. Probabilistic inference and differential privacy. *Advances in Neural Information Processing Systems*, 23:2451–2459, 2010.

Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. Functional mechanism: Regression analysis under differential privacy. *Proc. VLDB Endow.*, 5(11):1364–1375, July 2012. ISSN 2150-8097. doi: 10.14778/2350229.2350253.

Supplemental Materials: Data Augmentation MCMC for Bayesian Inference from Privatized Data

S-1 Statement on Societal Impacts

We do not foresee direct negative societal impact from the current work. Admittedly, our method is based on imputing the confidential database which privacy mechanisms seek to protect. We can assure the reader that such imputations are based on formally differentially private data products and hence do not violate differential privacy. Also, one may argue that our work is catalytic to enhancing the 'disclosure risk' of individuals, i.e. an adversary might be able to make accurate posterior inference about an individual if the adversary has highly informative and correct prior and modeling information to begin with. Granted, no existing privacy frameworks can guard against this.

S-2 Proofs in Section 3.1

Proposition 3.1. *For a pure ϵ-DP privacy mechanism η, the acceptance probability α from Equation (5) satisfies $\alpha(x_i^\star \mid x_i, x_{-i}, \theta) \geq \exp(-\epsilon)$, for all $\theta, x_{-i}, x_i, x_i^*$.*

Proof. Step 2a of Algorithm 1 proposes a new state $x_i^\star$ for the i-th record x_i according to the model $f(\cdot \mid \theta)$. Notice that the proposed latent database $x^\star = (x_i^\star, x_{-i})$ and the current latent database $x = (x_i, x_{-i})$ differ in only one entry. Then, the probability of accepting a proposed state $x_i^\star$ is $\alpha(x_i^\star \mid x_i, x_{-i}, \theta) = \min\left(\eta_\epsilon(s_{dp} \mid x^\star)/\eta_\epsilon(s_{dp} \mid x), 1\right)$. This ratio compares two adjacent databases $x^\star$ and x. ϵ-DP guarantees that the probability ratio of any output is within $\exp(\pm\epsilon)$ for adjacent databases by Equation (1). □

Proposition 3.2. *The Gibbs sampler described in Algorithm 1 requires $O(n)$ number of operations to update the full latent database according to $p(x \mid \theta, s_{\text{dp}})$.*

Proof. We prove that each update for $x_i \mid x_{-i}, \theta, s_{\text{dp}}$ is $O(1)$ and hence the full sweep for the latent database $x \mid \theta, s_{\text{dp}}$ is $O(n)$. Given current state (x, θ), in Step 2a, the method proposes from $x_i^\star \sim f(\cdot \mid \theta)$ independent of other entries x_{-i} and the current state x_i; the runtime of this local proposal step does not depend on n. Since $\eta(s_{\text{dp}} \mid x)$ is record-additive (Assumption 2), then $t(x^\star, s_{\text{dp}})$ can be computed in $O(1)$ time by $t(x^\star, s_{\text{dp}}) = t(x, s_{\text{dp}}) - t_i(x_i, s_{\text{dp}}) + t_i(x_i^\star, s_{\text{dp}})$ of Step 2b. The density evaluations in Step 2c are also $O(1)$. Overall, to update all $x_i,\ i = 1, 2, \ldots, n$, the runtime is $O(n)$. □

S-3 Proofs in Section 3.2

S-3.1 Ergodicity

In Algorithm 2, we first present a Metropolis-within-Gibbs sampler that is more general than Algorithm 1. We prove its ergodicity in Theorem S-3.1, which implies Theorem 3.3.

The Metropolis-within-Gibbs sampler in Algorithm 2 consists of alternating Metropolis-Hastings steps targeting $p(\theta \mid x, s_{\text{dp}}) = p(\theta \mid x)$ and $p(x \mid \theta, s_{\text{dp}})$. In Assumption 1 we have assumed that a Markov kernel for $p(\theta \mid x)$ exists. A typical kernel involves first proposing from some distribution $q_\theta(\theta \mid x)$ and then accepting or rejecting the proposed state an appropriate probability. The data-augmentation steps consist of the sequence of updates $p(x_i | x_{-i}, \theta, s_{\text{dp}})$, for $i = 1, 2, \ldots, n$. Algorithm 1 suggests using the proposal $x_i^\star \sim f(\cdot \mid \theta)$ independent of current state x. In this more general sampler, described in algorithm 2, we use proposals $q_x(x_i^\star \mid x_i, x_{-i}, \theta, s_{\text{dp}})$ that can depend on current states of x and θ, as well as the private query s_{dp}. Notice that since latent records are exchangeable in both $f(x \mid \theta)$ and $\eta(s_{\text{dp}} \mid x)$, respectively by the i.i.d. model assumption and by record-additivity, it is sufficient to use the same kernel q_x for all x_i.

Theorem S-3.1. *Under conditions A1 - A4 below, the Gibbs sampler of Algorithm 2 on the joint space $(\mathbb{X}^n \times \mathbb{R}^p)$ is ergodic and it admits $\pi(x, \theta)$ as the unique limiting distribution.*

A1. The prior distribution is proper and $\pi_0(\theta) > 0$ for all θ in $\Theta = \{\theta \mid f_\theta(x) > 0 \text{ for some } x\}$.

Algorithm 2 A general Metropolis-within-Gibbs sampler for $p(\theta, x \mid s_{\text{dp}})$

1. Conditional update of $p(\theta \mid x)$:
   (a) Propose $\theta^\star \sim q_\theta(\theta^\star \mid \theta, x)$.
   (b) Accept $\theta^\star$ with probability
$$\alpha(\theta^\star \mid \theta, x) = \min\left\{\frac{q_\theta(\theta \mid \theta^\star, x)p(\theta^\star)\prod_{i=1}^n f(x_i \mid \theta^\star)}{q_\theta(\theta^\star \mid \theta, x)p(\theta)\prod_{i=1}^n f(x_i \mid \theta)}, 1\right\}$$
2. For each $i = 1, \ldots, n$, update $p(x_i \mid x_{-i}, \theta, s_{\text{dp}})$ by:
   (a) Propose $x_i' \sim q_x(x_i^\star \mid x_i, x_{-i}, \theta, s_{\text{dp}})$,
   (b) Accept the proposed state $x_i^\star$ with probability
$$\min\left\{\frac{q_x(x_i \mid x_i^\star, x_{-i}, \theta, s_{\text{dp}})\eta(s_{dp} \mid x_i^\star, x_{-i})f(x_i^\star \mid \theta)}{q_x(x_i^\star \mid x_i, x_{-i}, \theta, s_{\text{dp}})\eta(s_{dp} \mid x_i, x_{-i})f(x_i \mid \theta)}, 1\right\}.$$

A2. The model is such that the set $\{x : f(x \mid \theta) > 0\}$ does not depend on θ.

A3. The privacy mechanism satisfies $\eta(s_{\text{dp}} \mid x) > 0$ for all $x \in \mathbb{X}^n$.

A4. From a valid current state, the proposal kernels satisfies (a) $q_\theta(\theta^\star \mid x, \theta) > 0$ for all $\theta^\star \in \Theta$, and (b) $q_x(x_i^\star \mid x_i, x_{-i}, \theta, s_{\text{dp}}) > 0$ for all $x_i^\star$ with $f(x_i^\star, x_{-i} \mid \theta) > 0$.

Proof. It is sufficient to show that the chain is π-invariant, aperiodic, and π-irreducible [Tierney, 1994]. The Metropolis-within-Gibbs sampler is aperiodic by construction, since some proposals can be rejected. It is also π-invariant because it is composed of kernels that satisfy detailed balance with respect to π.

Irreducibility means that, informally, every set A with $\pi(A) > 0$ can be reached by the Gibbs sampler from any starting point within finitely many steps. We first prove irreducibility for $n = 1$ and generalize this to a sample size of $n \geq 2$. Suppose $A \subset \mathbb{X}^1 \times \Theta$ with $\pi(A) > 0$ and suppose the current state of the Gibbs chain is $(x^{(0)}, \theta^{(0)})$. For any state $(x, \theta) \in A$ we have $q(\theta \mid x^{(0)}, x^{(0)})q(x \mid x^{(0)}, \theta, s_{\text{dp}}) > 0$ by A4. The acceptance ratios are also positive by A1-A4. As a result

$$\begin{aligned}&P(A \mid x^{(0)}, \theta^{(0)})\\ &\geq \int\int_A q(\theta \mid x^{(0)}, x^{(0)})q(x \mid x^{(0)}, \theta, s_{\text{dp}})\alpha(\theta \mid x^{(0)}, x^{(0)})\alpha(x \mid x^{(0)}, \theta, s_{\text{dp}})dxd\theta > 0.\end{aligned}$$

So when $n = 1$, we can reach A from any starting point in one iteration of the Gibbs sampler. For $n \geq 2$, we can reach the set A in at most n steps: the first iteration moves x_1 and θ into A, and subsequent steps moves other x_i's into A while keeping all previous x_j's inside A by rejecting proposals leaving A. □

A4 details conditions on the proposal distributions to ensure ergodicity of Algorithm 2. It can be relaxed so long as π-irreducibility is satisfied. Also, A4a should be viewed as a condition implied by the validility of a kernel targeting $p(\theta \mid x)$ from Assumption 1 and, therefore, is not an additional assumption. Importantly, conditions in A4 are mild because they cover common proposal distributions; Gaussian random walk on θ for A4a and the independent Metropolis proposals $f(\cdot \mid \theta)$ for A4b are such examples. In Algorithm 1, we use the kernel $q_x(x_i^\star \mid x_i, x_{-i}, \theta, s_{\text{dp}}) = f(x \mid \theta)$, which satisfies $f(x^\star \mid \theta) > 0$ by A2. Hence Theorem S-3.1 implies Theorem 3.3.

S-3.2 Geometric ergodicity of Algorithm 1

Theorem 3.4. *Assume that in step 1 of Algorithm 1, one can directly sample from $p(\theta \mid x)$. Under A1-A3 of Theorem 3.3, the resulting (x, θ) chain, as well as the marginal chains, are geometrically ergodic if η satisfies ϵ-DP and there exists $0 < a \leq b < \infty$ such that $a \leq f(x \mid \theta) \leq b \quad \forall \theta, x$.*

Proof of Theorem 3.4. The assumption of $a \leq f(x \mid \theta) \leq b$ leads to the inequality

$$p(\theta \mid x) = \frac{p(\theta) f(x \mid \theta)}{\int p(\theta') f(x \mid \theta') d\theta'} \geq \frac{a}{b} p(\theta),$$

since $p(\theta)$ is a proper prior by A1 of algorithm 1.

This proof proceeds by verifying the drift and minorization conditions of the marginal Markov transition kernel on X according to Theorem 8 of Johnson et al. [2013]. We first present a full proof for $n = 1$ and then generalize the arguments to $n \geq 2$. In this proof, we abbreviate $\eta(s_{\text{dp}} \mid x)$ as $\eta(x)$.

Recall that the probability of accepting proposed state $x^\star$ is $\alpha(x^\star \mid x, \theta) = \min\left(1, \frac{\eta(x^\star)}{\eta(x)}\right)$. The probability of accepting any proposal from the current state is $\alpha(x, \theta) = \int \alpha(x^\star \mid x, \theta) f(x^\star \mid \theta) dx^\star$. Let $K(x' \mid x, \theta)$ denote the Markov transition kernel with respect to the proposal $x^\star \sim f(\cdot \mid \theta)$, and let $K(x' \mid x) = \int K(x' \mid x, \theta) p(\theta \mid x) d\theta$ be the marginal kernel, which integrates out the exact θ update from $p(\theta \mid x)$. We have

$$K(x' \mid x, \theta) = f(x' \mid \theta) \alpha(x' \mid x, \theta) + (1 - \alpha(x, \theta)) \delta_x(x'),$$

where $\delta_x(x')$ is the Dirac-delta function. Then the marginal transition kernel satisfies

$$K(x' \mid x, \theta) \geq f(x' \mid \theta) \alpha(x' \mid x, \theta) \geq a \exp(-\epsilon)$$

since $a(x' \mid x, \theta) \geq \exp(-ep)$ according to Proposition 3.1. As a result, we have

$$p(\theta \mid x) K(x' \mid x, \theta) \geq \frac{a}{b} p(\theta) \cdot a \exp(-\epsilon) \tag{S1}$$

Equation (S1) is sufficient for a minorization condition $K(x' \mid x) \geq a^2 b^{-1} \exp(-\epsilon)$ to hold on $x' \in \mathbb{X}$ since $p(\theta)$ is proper.

To establish a drift condition, let $w : \mathbb{X} \to \mathbb{R}_{>0}$ be integrable with $v = \int w(x) dx < \infty$. Then we have the conditional expectation

$$\begin{aligned}
K_X[w(x)] &= \mathbb{E}\left[w(X^{(t+1)}) \mid X^{(t)} = x\right] \\
&= \int w(x') K(x' \mid x, \theta) dx' \\
&= \int \int w(x') K(x' \mid x, \theta) p(\theta \mid x) d\theta dx' \\
&= \int \int w(x') f(x' \mid \theta) \alpha(x' \mid x, \theta) p(\theta \mid x) d\theta dx' + w(x) \int (1 - \alpha(x, \theta)) p(\theta \mid x) d\theta \\
&\leq \int \int w(x') f(x' \mid \theta) p(\theta \mid x) d\theta dx' + w(x) \int p(\theta \mid x) d\theta
\end{aligned}$$

Using $f(x \mid \theta) \leq b$, we can show that

$$K_X[w(x)] \leq bv + w(x), \tag{S2}$$

which is the drift condition. Combining Equations (S1) and (S2), we invoke Theorem 8 of Johnson et al. [2013] to establish geometric ergodicity of the Gibbs sampler.

When $n \geq 2$, the proof shall proceed by denoting $K(x' \mid x, \theta)$ as the Markov transition kernel on $x, x' \in \mathbb{X}^n$ and similarly for $K(x' \mid x)$. The drift condition becomes $K_X[w(x)] \leq b^n v + w(x)$ and minorization condition becomes $K(x' \mid x) \leq (a^2 b^{-1} \exp(-\epsilon))^n$. □

S-4 Log-linear Model: More Details

Our full model, along with conjugate priors is given in the following equation array:

$$
\begin{aligned}
&\text{prior} && p \sim \text{Dirichlet}(\alpha), && \text{(S3)}\\
& && p^k_{i-} \sim \text{Dirichlet}(\alpha^k_i) \quad \forall i, && \text{(S4)}\\
&\text{data model} && n_- \sim \text{Multinomial}(N, p_-), && \text{(S5)}\\
& && n^k_{i-} \sim \text{Multinomial}(n_i, p^k_{i-}) \quad \forall i, && \text{(S6)}\\
&\text{privacy noise} && L_{ijk} \overset{\text{i.i.d.}}{\sim} \text{Laplace}(0, 2K/\epsilon), && \text{(S7)}\\
& && m^k_{ij} = n^k_{ij} + L_{ijk} \quad \forall i, j, k, && \text{(S8)}\\
&\text{privatized output} && s_{dp} = (m^k_{ij}). && \text{(S9)}
\end{aligned}
$$

S-5 Linear Regression: More Details and Results

Data generating parameters. Our experiments use continuous predictors X_0, which we model as $X_0^i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(m, \Sigma)$. We choose $\Sigma = I_n$. We simulate $m_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and hold it fixed at $m = (0.9, -1.17)$.

Conjugate prior distribution. Our experiments fix σ^2 at the data generating value of $\sigma^2 = 2$. Given prior $\beta \sim \mathcal{N}_{p+1}(0, \tau^2 I_{p+1})$, the posterior distribution $\beta \mid \sigma^2, x, y$ is multivariate Normal with covariance matrix $\Sigma_n = (x^\top x/\sigma^2 + I_{p+1}/\tau^2)^{-1}$ and mean vector $\mu_n = \Sigma_n(x^\top y)/\sigma^2$. The prior for β is $\beta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau^2 = 2^2)$.

The effect of clamping. We view clamping as part of the privacy mechanism. The clamping step first truncates x and y values into a fixed range, and then performs data-independent location-scale transformation so that all values of $\tilde{x}$ and $\tilde{y}$ are in the range $[-1, 1]$. Although with conjugate priors the confidential data posterior $p(\theta \mid x, y)$ is tractable, the clamped data posterior $p(\theta \mid \tilde{x}, \tilde{y})$ no longer enjoys conjugacy and is now intractable. Since the clamping parameters are known, to sample from the clamped data posterior, one can design data-augmentation MCMC algorithms to impute truncated values. Such an imputation algorithm might take $O(n)$ per iteration. We also highlight that as $\epsilon \to \infty$, in which case privacy noise approaches zero, the posterior $p(\theta \mid s_{\text{dp}})$ approaches $p(\theta \mid \tilde{x}, \tilde{y})$.

Acceptance rate. In Section 5, we report the posterior means of β, β_1 and β_2 given s_{dp} produced from the same fixed latent database (x, y), with different privacy levels. We also report the acceptance rate of $p(x_i \mid x_{-\theta}, \theta, s_{\text{dp}})$ updates in each iteration of the Gibbs samplers. Recall that for each s_{dp}, we run the Gibbs sampler for 10000 iterations and discard the first half for burn-in. From Figure S1, we can see that the empirical acceptance rate of the IM proposals is much higher than the lower bound of Proposition 3.1.

Posterior credible intervals. We repeat the credible interval experiment on log-linear models. First we sample one β parameter from the prior, and hold this fixed. Then for each ϵ value, we produce 100 confidential databases (x, y) and one private s_{dp} for each non-private one, and then run a chain for 10,000 iterations targeting $\beta \mid s_{\text{dp}}$. After burn-in, from each chain, we produce a 90% credible interval for each β_0, β_1 and β_2. We then calculate the empirical coverage which is reported in Table 1.

While at $n = 100$, we do not expect the frequentist coverage of the credible intervals to exactly match the nominal level of .9, note that most of the values are close to or above .9. The coverage on β_1 is lower than 90%, which might be due to the true parameter being furthest from the prior mean of 0. Another explanation is that data quality loss from truncation and location-scale transformations during the clamping procedure can not be fully recovered by our inference procedure.

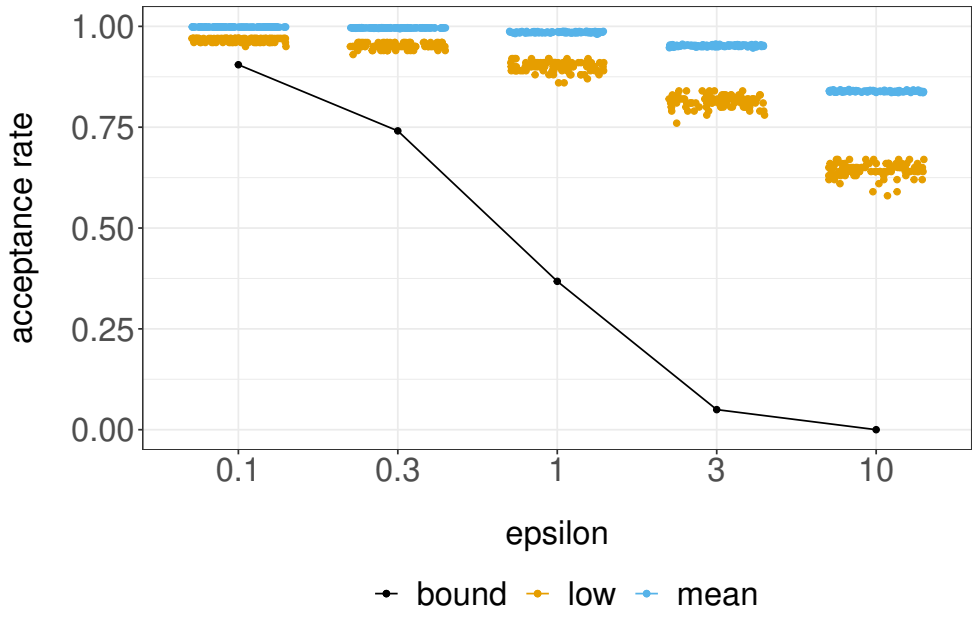


Figure S1: Observed acceptance rates for the log-linear model. The blue (above) point clouds indicate the average acceptance rate, and the orange (below) points indicate the observed minimum acceptance rate of each chain. The solid black line is the lower bound of Proposition 3.1.

ϵ	$\beta_0 = -1.79$	$\beta_1 = -2.89$	$\beta_2 = -0.66$
0.1	.99	**.60**	.99
0.3	1	**.66**	.94
1	1	**.84**	**.80**
3	1	**.84**	**.75**
10	.93	.87	.85

Table 1: Coverage of $\beta_0, \beta_1, \beta_2$ in linear regression. Coverage is based on 100 replicates.

S-6 Statement on Computing Resources

We ran the experiments on an internal cluster. We used a server with a pair of 64-core AMD Epyc 7662 'Rome' processors and with 256GB of RAM. We ran each MCMC chain for 10000 iterations and a typical chain takes approximately 330 seconds for linear regression and approximately 404 seconds for the log-linear model.