March 2020

Noise-Aware Inference for Differential Privacy

Garrett Bernstein

NOISE-AWARE INFERENCE FOR DIFFERENTIAL PRIVACY

A Dissertation Presented

by

GARRETT BERNSTEIN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2020

College of Information and Computer Sciences

NOISE-AWARE INFERENCE FOR DIFFERENTIAL PRIVACY

A Dissertation Presented

by

GARRETT BERNSTEIN

Approved as to style and content by:

Daniel Sheldon, Chair

Gerome Miklau, Member

Justin Domke, Member

Patrick Flaherty, Member

James Allan, Chair of the Faculty
College of Information and Computer Sciences

DEDICATION

To my family.

ACKNOWLEDGMENTS

First and foremost, I would like to thank Daniel Sheldon. I can confidently say that I owe to Dan the majority of my development as a researcher the past five years. From problem development to note-taking to mathematical exploration to writing papers to crafting presentations to mentoring others, all I needed to do was ask myself "How would Dan tackle this?" and the ingrained answer would quickly bubble up in my mind.

Thank you to my committee members for being extra accommodating and helpful: Gerome Miklau for being my privacy spirit guide, and Justin Domke and Patrick Flaherty for all the advice on inference procedures.

Thank you to Leeanne Leclerc, Michele Roberts, and Eileen Hamel, who shepherded me along all the way.

And to my friends and family for all the support. I couldn't have done a PhD if I was only doing the PhD.

ABSTRACT

NOISE-AWARE INFERENCE FOR DIFFERENTIAL PRIVACY

FEBRUARY 2020

GARRETT BERNSTEIN

B.Sc., CORNELL UNIVERSITY

M.Eng., CORNELL UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Daniel Sheldon

Domains involving sensitive human data, such as health care, human mobility, and online activity, are becoming increasingly dependent upon machine learning algorithms. This leads to scenarios in which data owners wish to protect the privacy of individuals comprising the sensitive data, while at the same time data modelers wish to analyze and draw conclusions from the data. Thus there is a growing demand to develop effective private inference methods that can marry the needs of both parties. For this we turn to differential privacy, which provides a framework for executing algorithms in a private fashion by injecting specifically-designed randomization at various points in the process. The majority of existing work proceeds by ignoring the injected randomization, potentially leading to pathologies in algorithmic performance. There is, however, a small body of existing work that performs inference over the

injected randomization in an attempt to design more principled algorithms. This thesis summarizes the subfield of noise-aware differentially private inference and contributes novel algorithms for important problems.

Differential privacy literature provides a multitude of privacy mechanisms. We opt for sufficient statistics perturbation (SSP), in which sufficient statistics, a quantity that captures all information about the model parameters, are corrupted with random noise and released to the public. This mechanism offers desirable efficiency properties in comparison to alternatives. In this thesis we develop methods in a principled manner that directly accounts for the injected noise in three settings: maximum likelihood estimation of undirected graphical models, Bayesian inference of exponential family models, and Bayesian inference of conditional regression models.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

Machine learning and probabilistic inference are pervasive aspects of our every day world. Specifically of interest are applications in which machine learning relies on data stemming from individual people. The curators or owners of sensitive data sets often find themselves responsible for protecting the data, which, when improperly handled, can violate the privacy of the individuals in the data. On the other hand, data modelers wish to perform analyses on the data in order to draw important population-level conclusions regarding the general behaviors and attributes of the individuals. This leads to seemingly opposed goals: data owners prioritize protecting individual data, whereas modelers prioritize leveraging individual data to perform analyses. Thus there is a pressing need for private machine learning techniques that can achieve appropriate tradeoffs.[1] How can we derive useful population-level outcomes without compromising the privacy of individuals?

This work relies on *differential privacy*, the dominant standard for private data analysis [Dwork et al., 2006]. Differential privacy provides a guarantee to individuals: The output of a differentially private algorithm is statistically nearly unchanged if any single individual's record is added to or removed from the input data set. Thus, subject to the setting of privacy parameters, there is negligible risk to the individual in allowing their data to be analyzed in this fashion. The general idea of many privacy mechanisms is to carefully randomize an algorithm's output by

[1] It may be the case that a single entity is both the data owner and modeler, in which case they wish to perform privatized analyses on their own data to be fit for public release.

calibrating the noise to the *sensitivity* of the function outputting a quantity dependent on sensitive data. Sensitivity captures how much the output of a function depends on any individual's data in the worst case [Dwork et al., 2006]; higher sensitivity requires more extreme randomization so that the (random) output does not depend too much on any individual's data. The randomization renders the noisy data safe for public release; all subsequent calculations using the noisy data, known as *post-processing*, are also safe [Dwork & Roth, 2014]. Perhaps surprisingly, public divulgence of the use of the release mechanism and its parameters do not impair the privacy guarantee; in fact, as we will see, specific knowledge of the release mechanism proves crucial in developing noise-aware inference techniques.

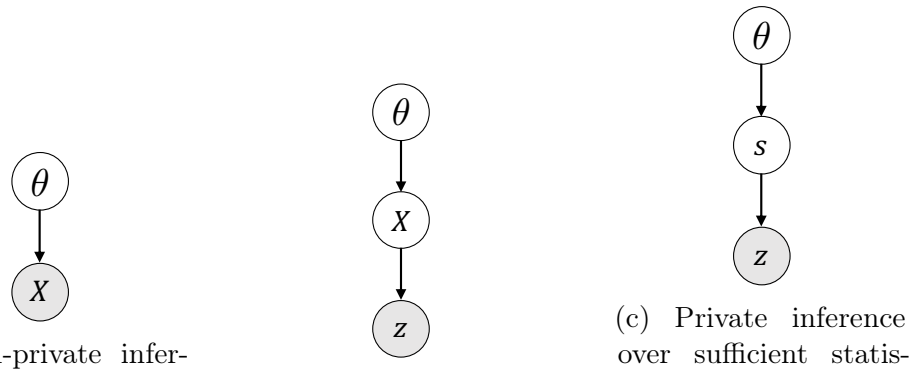
The decade and a half since the seminal differential privacy paper [Dwork et al., 2006] saw an early focus on developing privacy release mechanisms, and since the field has taken hold, there has been a growing focus on developing private machine learning methods for use on real world problems. Differential privacy has been applied to many areas of machine learning, including, as a small sample, learning specific models such as logistic regression [Chaudhuri & Monteleoni, 2009], support vector machines [Rubinstein et al., 2009], and deep neural networks [Abadi et al., 2016]; privacy in general frameworks such as empirical risk minimization (ERM; Bassily et al. [2014]; Chaudhuri et al. [2011]; Jain & Thakurta [2013]; Kifer et al. [2012]), gradient descent [Wu et al., 2016], and parameter estimation [Smith, 2011b]; and theoretical analysis of what can be learned privately, e.g., [Blum et al., 2005; Kasiviswanathan et al., 2011]. These methods can often be simple and efficient, but a major drawback of these works is the downstream analyses inherently ignore the randomization due to the release mechanism, i.e. they are "noise-naive", which potentially introduces pathologies (e.g. calculations leading to negative variance) and hampers results (e.g. poor point estimation accuracy). Ultimately, the output of the methods and mechanisms is used

directly in some downstream use case and thus is designed to be useful on its own, which brings along any associated pathologies.

There is, however, a small subfield of "noise-aware" works in which probabilistic inference methodologies are used in conjunction with the output of release mechanisms in order to account for the randomization introduced by the mechanism. There are two main advantages to this framework. First, noise-aware methods generally separate the release mechanism and the technical analysis methodology into two components, which can relieve the data owner of potentially unwanted analysis burdens and thus lead to greater uptake of private methods. Second, and perhaps most importantly, noise-aware methods are generally shown to outperform noise-naive methods and thus prove to be more useful in practical settings, such as with smaller data sizes or stricter privacy levels. *The goal of this thesis is to add to the subfield of noise-aware differentially private machine learning.*

In the rest of this chapter we will introduce the concept of noise-aware inference, review existing work in this field, and introduce the three technical chapters of this thesis.

1.1 Noise-Aware Differentially Private Machine Learning



(a) Non-private inference.

(b) Private inference.

(c) Private inference over sufficient statistics.

Figure 1.1: Inference models.

The noise-aware paradigm was first introduced by Williams & McSherry [2010]. In the non-private setup in Figure 1.1a, the collection of individual data X is generated from some model parameterized by θ and subsequently observed, with the goal of then drawing conclusions about θ given the data X. In the private setup in Figure 1.1b, the individual data X is protected and instead a perturbed quantity z is made publicly-available via a release mechanism. This latter setup can be thought of as a latent variable model in which a privately released quantity is observed and the true, unperturbed quantity of the original model is unobserved. In this way, the noise model due to the release mechanism can be directly accounted for in a probabilistic inference procedure, with the intent of producing more principled and higher-utility private analyses. The key insight is that exposing the specifications of the release mechanism does not harm the differential privacy guarantee. Knowledge of the mechanism defines the conditional distribution $p(z \mid X)$, which can be combined with the generative data model, $p(X \mid \theta)$, to form the marginal likelihood, $p(z \mid \theta) = \int p(X \mid \theta)\, p(z \mid X)\, dX$. With this likelihood we can perform analyses of interest regarding θ. The integral over all data sets, however, is generally intractable. Thus the main technical development required for novel noise-aware inferences is to form approximations of this integral and obtain either a closed form or a tractable sampling procedure.

The model in Figure 1.1b encapsulates a broad range of machine learning models. There are three main decisions to make in designing a private algorithm in this context. First, which statistic of X do we wish to privatize? Second, with which mechanism will we privately release that statistic as z? Third, what analyses or inferences do we perform given z? While Williams & McSherry [2010] introduce this broad research landscape of noise-aware inference, their technical contributions explored a small swath of problem space as a proof of concept. In this thesis we develop noise-aware methodology for *sufficient statistics perturbation* (SSP; Foulds et al. [2016]; McSherry & Mironov [2009]; Vu & Slavkovic [2009]; Zhang et al. [2016]), in which noise is added

to the sufficient statistics of a given model, which are a quantity that capture all information about the model parameters [Fisher, 1922]. This setup, shown in Figure 1.1c, allows us to work directly with the sufficient statistics instead of the individual data, which allows for the development of techniques leading to tractable and higher utility noise-aware methods for a broad range of problem settings falling into the exponential family.

1.2 Existing Noise-Aware Work

Here we review the limited existing noise-aware work before (and concurrent to) this thesis.

1.2.1 Differentially Private Exponential Random Graphs

Karwa et al. [2014] focuses on privately fitting and estimating a wide class of exponential random graph models (ERGMs), which model the structure of graphs with statistics of the network and node attributes. The work assumes the covariate data of each node is publicly available, e.g. personal characteristics, and the relationship data between individuals is what needs to be protected with *edge differential privacy*. Edge DP protects the addition or deletion of a single edge in a graph and guarantees the distribution of outputs on two neighboring graphs is nearly identical. The work uses randomized responses for edges as the release mechanism, which independently perturbs the values of the network adjacency matrix.

The goal is then to develop inference procedures to analyze the privatized network. The main challenge of MLE in ERGMs is the intractability of calculating the normalizing constant due to needing to sum over all possible network configurations. One solution is to approximate the normalizing constant using MCMC [Geyer & Thompson, 1992], and Handcock & Gile [2010] explore how to do so when only a sample of the network is observed. Karwa et al. [2014] adapts this idea to the case of

when a perturbation of the network is observed due to a privacy mechanism. They note a naive approach would be to use the perturbed network "as is" and thus ignore the privacy mechanism. Instead they develop a method to include the privacy mechanism in the model and perform MLE over the full likelihood. This leads to the need to approximate a new normalizing constant involving the noise mechanism, for which they introduce a second MCMC chain. They show quality of the model fit by noise-aware method degrades much more gracefully than the naive method as privacy level increases

1.2.2 Inference Using Noisy Degrees: Differentially Private β-Model and Synthetic Graphs

Karwa et al. [2016] again focuses on edge DP for graphs, but in this work turns to the β-model of random graphs. These are graphs with random edges whose distribution is an exponential family model, for which the sufficient statistics are the degree sequence of the graph. This work proves that releasing noisy sufficient statistics and using them "as is" will often result in methodological failures. There is asymptotically at best a 50% chance a valid graph can be sampled from a perturbed degree sequence. They also show the drastic rate of non-existence of MLE for the model when naively using the perturbed sufficient statistics. Further, even a naive projection to the nearest valid degree sequence results in hampered statistical performance. Those negative results motivate their development of a noise-aware method which instantiates an estimation of the latent true degree sequence. The proposed method is a modified Havel-Hakimi "certifying" algorithm [Hakimi, 1962; Havel, 1955] that performs an L_1 projection for the perturbed degree sequence onto the marginal polytope, which results in a valid degree sequence. They then show the resulting MLE is consistent and asymptotically normal.

1.2.3 Locally Private Bayesian Inference for Count Models

Concurrently to work done for this thesis, Schein et al. [2018] explores the problem of local privacy for Bayesian inference for Poisson factorization, which is motivated by problems such as topic modeling and community detection from email correspondences, e.g. the Enron data set [Klimt & Yang, 2004]. The work uses the geometric mechanism to achieve *local privacy*, which means only individuals have access to their unperturbed data. The generative process of the combined Poisson factorization and release mechanism can be written in multiple equivalent configurations. This enables careful application of probabilistic distribution properties, namely relationships between Poisson random variables with Skellam and Bessel random variables, allowing for tractable posterior sampling via MCMC. Experiments show the point estimates due to the noise-aware method outperforms the naive method in case studies on both topic modeling and community detection. Interestingly, the noise-aware method outperforms even the non-private method, indicating higher levels of robustness towards generalization [Dwork & Lei, 2009].

1.3 Thesis Contributions

This thesis develops noise-aware methods by leveraging techniques enabled by working directly with the sufficient statistics of a model. This release approach of *sufficient statistic perturbation* (SSP) is desirable from a privacy perspective for a number of reasons (see Section 2.1) but narrows the possible release mechanisms from which to choose, namely the Laplace and Gaussian mechanisms [Dwork et al., 2006]. The question then is, how much utility do we lose by restricting our methods to this choice of privately releasing sufficient statistics? SSP is most applicable in models which have compact sufficient statistics. In such cases, existing work and results in this thesis show that in fact even noise-naive SSP methods are very competitive or even state of the art for many classes of models in comparison to other release

approaches. For example, for the problem of point estimation of unconditional family models, Foulds et al. [2016] shows noise-naive SSP for exponential family models is a consistent estimator and out performs *one posterior sampling* (OPS), and Wang [2018] shows SSP for linear regression is competitive with OPS and significantly better than the mechanisms of *objective perturbation*, *subsample-and-aggregate*, and *noisy stochastic gradient descent*. In the Bayesian framework, Chapters 4 and 5 show that even noise-naive SSP outperforms other approaches in producing correct posteriors for unconditional and conditional exponential family models, respectively.

Another drawback of SSP is the need to "lock-in" a model for which to release the sufficient statistics. This potentially limits the flexibility of downstream analyses, since subsequently releasing sufficient statistics for a different model may require a larger privacy budget. This thesis does not explore the problem of initial model selection, nor the problem of efficiently releasing sufficient statistics to enable analyses of multiple models, though these are interesting avenues for future research.

1.3.1 Differentially Private Learning of Undirected Graphical Models using Collective Graphical Models

Chapter 3 addresses the problem of privately learning parameters in discrete, undirected graphical models with noise-aware inference. Graphical models are a central tool in probabilistic modeling and machine learning. They pair expressive probability models with algorithms that leverage the graphical structure for efficient inference and learning. These tools allow a practitioner to posit a model for observed data, and then fit parameters, assess model validity, and make predictions.

In this chapter we clarify the theory and practice of *noise-naive maximum likelihood estimation* for undirected graphical models. We show that it learns better models than existing state-of-the-art approaches, and in fact that it achieves the same asymptotic mean-squared error as the non-private method. This motivates the use of conducting

inference over noisy sufficient statistics. We do so using techniques from *collective graphical models* (CGMs; Sheldon & Dietterich [2011]), which allow for efficient inference over sufficient statistics of a graphical model given noisy observations thereof. We then show that this more principled noise-aware approach is superior to competing approaches in nearly all scenarios.

1.3.2 Differentially Private Bayesian Inference for Exponential Families

Chapter 4 develops the first fully noise-aware Bayesian method capable of producing the correct private posterior for exponential family models. Exponential family models include many of the most familiar parametric probability models, e.g. binomial, exponential, and Gaussian. Previous work has Bayesian inference that ignores noise due to the release mechanism [Dimitrakakis et al., 2014; Foulds et al., 2016; Geumlek et al., 2017; Wang et al., 2015; Zhang et al., 2016].

In this chapter we develop technical approximations that allow for tractable sampling from the correct posterior over noisy sufficient statistics. We also address the challenge of privately releasing unbounded sufficient statistics, e.g. those of the exponential distribution. We then show empirically that when compared with competing methods, ours is the only one that provides properly calibrated beliefs about model parameters in the non-asymptotic regime, and that it provides good utility compared with other private Bayesian inference approaches.

1.3.3 Differentially Private Bayesian Linear Regression

Chapter 5 develops the first differentially private method to produce a full publicly-available posterior for Bayesian linear regression. Linear regression is one of the most widely used statistical methods, especially in domains where data comes from humans. It is important to develop robust tools that can realize the benefits of regression analyses but maintain the privacy of individuals. Existing work on differentially private linear regression focuses on point estimation [Bassily et al., 2014; Dimitrakakis

et al., 2014; Dwork & Smith, 2010; Foulds et al., 2016; Geumlek et al., 2017; Kifer et al., 2012; Minami et al., 2016; Smith, 2008; Vu & Slavkovic, 2009; Wang, 2018; Wang et al., 2015; Zhang et al., 2016] and only a few recent works address uncertainty quantification of regression coefficients through confidence interval estimation [Sheffet, 2017] and hypothesis tests [Barrientos et al., 2019].

In this chapter we first show the noise-naive Bayesian method produces the correct posterior only in larger data regimes. This motivates our development of inference methods that properly account for the noise due to the privacy mechanism. We develop MCMC-based techniques to sample from posterior distributions, as done for exponential families in Chapter 4. A significant challenge relative to that chapter is the need to form some assumption about the distribution over covariate data, since it cannot be conditioned on as in non-private regression. The first noise-aware method instantiates individuals, which scales the runtime with population size and requires an explicit prior distribution for covariates. The second method marginalizes out individuals and approximates the distribution over the sufficient statistics; it requires weaker assumptions about the covariate distribution (only moments), and its running time does not scale with population size. We perform a range of experiments to show our noise-aware methods are as well or nearly as well calibrated as the non-private method, and have better utility than the naive method. We demonstrate using real data that our noise-aware methods quantify posterior predictive uncertainty significantly better than naive SSP. We then conclude with a case study drawn from the real problem of social mobility policy-making using sensitive census data.

1.4 Summary

The subfield of noise-aware inference has a limited number of existing works but has shown great potential to enable differentially private machine learning for practical problem settings in the real world. All previously discussed noise-aware works show,

in their respective problem settings and with a variety of release mechanisms, that the naive approach of ignoring noise due to the release mechanism leads to pathologies and hampers performance on desired tasks. Karwa et al. [2014], Karwa et al. [2016], and Chapter 3 do so for exponential random graph models, β-model random graphs, and undirected graphical models, respectively. Williams & McSherry [2010], Schein et al. [2018], Chapter 4, and Chapter 5 do so for Bayesian inference in logistic regression, Poisson factorization, exponential family models, and linear regression, respectively. The negative effects of ignoring noise range from lower accuracy, as in Chapter 3, to the extreme of MLE non-existence, as in Karwa et al. [2016] and Chapter 3. These effects are exacerbated in data settings where the randomization injected by the release mechanism overwhelms the signal in the data, namely when the privacy guarantee is strict or when the size of the population is small.

All works then go on to develop noise-aware methods that perform inference over the release mechanism, which is shown to improve upon the negative noise-naive results. By casting the privacy-preserving process as a latent variable problem and instantiating the latent variable, the original problem setting's model can be used with the perturbed statistics projected to be valid statistics, e.g. MLE for the β-model random graph with a valid degree sequence, or a Bayesian conjugate update with projected sufficient statistics that lead to a positive variance. The main hurdle to do so, as originally pointed out by Williams & McSherry [2010], is to make tractable the integral over all data sets found within the marginal likelihood of the publicly-released perturbed quantity. Each problem setting requires different insights to achieve this goal, e.g. Schein et al. [2018] rewriting the geometric mechanism as a Skellam distribution.

This thesis focuses on problems that fit into exponential family models, such that the dimensions of the sufficient statistics do not grow with respect to the population size. Chapter 3 uses the insight that the existing body of work on collective graphical

models allows for tractable inference over sufficient statistics of a graphical model given noisy observations. Chapters 4 and 5 turn to Bayesian inference as a natural framework for expressing uncertainty in the face of noisy observations. The insight that sufficient statistics of exponential family models and linear regression are sums over individuals allows for normal approximations via the central limit theorem, and an additional model augmentation of the Laplace mechanism into a scale mixture of normals allows for a tractable sampling procedure.

The framework of noise-aware inference is applicable to a wide range of problem settings and release mechanisms. All three chapters use the Laplace mechanism but are easily adapted to the Gaussian mechanism. As a first step, further more complicated mechanisms could potentially be tackled simply by the MCMC-based method introduced in Chapter 5, as long as the generative model distribution can be compatibly written with standard MCMC algorithms. As first observed by Williams & McSherry [2010] a decade ago, probabilistic inference is a powerful tool that, hand in hand with differential privacy, can be used to unlock a high level of utility in private machine learning settings that would be otherwise unattainable.

1.5 Published Work

Not included in this thesis is previous work done on consistently estimating Markov chains with noisy aggregate data. This work was motivated by the application of continent-wide bird migration, in which the actions of individual birds are unobservable, but noisy observations of the population counts of birds are available over time. The work done also attempted to fit models using noisy sufficient statistics and eventually led to the privacy work done in this thesis.

Below is a full list of publications by the author while in the PhD program at the University of Massachusetts Amherst.

- **Garrett Bernstein** & Daniel Sheldon. *Differentially Private Bayesian Linear Regression.* Advances in Neural Information Processing Systems (NeurIPS) 2019.

- Tsung-Yu Lin, Kevin Winner, **Garrett Bernstein**, et al. *MistNet: Measuring historical bird migration in the US using archived weather radar data and convolutional neural networks.* Methods in Ecology and Evolution, 2019.

- **Garrett Bernstein** & Daniel Sheldon. *Differentially Private Bayesian Inference for Exponential Families.* Advances in Neural Information Processing Systems (NeurIPS) 2018.

- **Garrett Bernstein**, Ryan McKenna, Tao Sun, Daniel Sheldon, Gerome Miklau, Michael Hay. *Differentially Private Learning of Undirected Graphical Models using Collective Graphical Models.* Proceedings of the 32nd International Conference on Machine Learning (ICML) 2017.

- Judy Shamoun-Baranes, Andrew Farnsworth, **Garrett Bernstein**, et al. *Innovative Visualizations Shed Light on Avian Nocturnal Migration.* PLOS ONE, 2016.

- **Garrett Bernstein** & Daniel Sheldon. *Consistently Estimating Markov Chains with Noisy Aggregate Data.* Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS) 2016.

- Kevin Winner, **Garrett Bernstein**, and Daniel Sheldon. *Inference in a partially observed queueing model with applications in ecology.* Proceedings of the 32nd International Conference on Machine Learning (ICML) 2015.

CHAPTER 2

DIFFERENTIAL PRIVACY BACKGROUND

Differential privacy requires that an individual's data has a limited effect on an algorithm's behavior. A data set $X = x_{1:n} := (x_1, \ldots, x_n)$ consists of records from n individuals, where x_i is the data of the ith individual. We will assume n is known. Differential privacy reasons about the hypothesis that one individual chooses to remove their data from the data set, and their record is replaced by another one.[1] Let $\text{nbrs}(X)$ denote the set of data sets that differ from X by exactly one record—i.e., if $X' \in \text{nbrs}(X)$, then $X' = (x_{1:i}, x'_i, x_{i+1:n})$ for some i.

Definition 1 (Differential Privacy; Dwork et al. [2006])**.** *A randomized algorithm $\mathcal{A}$ satisfies ϵ-differential privacy if for any input X, any $X' \in nbrs(X)$ and any subset of outputs $O \subseteq Range(\mathcal{A})$,*

$$\Pr[\mathcal{A}(X) \in O] \leq \exp(\epsilon)\Pr[\mathcal{A}(X') \in O].$$

The data owner can specify the level of privacy guarantee via a positive ϵ; smaller values satisfy stricter privacy levels.

We achieve differential privacy by injecting noise into statistics that are computed on the data. Let f be any function that maps datasets to $\mathbb{R}^d$. The amount of noise depends on the *sensitivity* of f. We drop the subscript f when it is clear from context.

[1] This variant assumes n remains fixed, which is sometimes called *bounded* differential privacy [Kifer & Machanavajjhala, 2011].

Definition 2 (Sensitivity)**.** *Let the* l_1-sensitivity *of a function* f *for two specific data sets,* X *and* X'*, be* $S_f(X, X') = \|f(X) - f(X')\|_1$*. Then the* l_1-sensitivity *of* f *is*

$$\Delta_f = \max_{X, X' \in nbrs(X)} S_f(X, X').$$

A *release mechanism* is an algorithm which has sensitive data as input, which only trusted parties are allowed to observe, and privatized data as output, which is allowed to be observed by any party. The ability of these release mechanisms to deliver privacy guarantees to individuals hinges on the fact they their outputs are immune to *post-processing* operations [Dwork & Roth, 2014]; if an algorithm $\mathcal{A}$ is ϵ-differentially private, then any algorithm that takes as input only the output of $\mathcal{A}$, and does not use the original data set X, is also ϵ-differentially private.

2.1 Release mechanisms

There are a multitude of available differentially private release mechanisms in the literature. All involve injecting randomization somewhere into the process such that the influence of any one individual on the output is negligible. There are several design choices to consider in order to learn accurate models under differential privacy. It is critical to randomize the mechanism "just enough" to achieve the desired privacy guarantee while minimizing the impact on the quality of the subsequently learned model. There are two general considerations to account for in achieving that goal. First, noise should be added at an "information bottleneck" so that as few quantities as possible require perturbation, e.g. adding noise to summary statistics instead of to individual data. Second, noise should be added at a location where sensitivity can be calculated exactly, or at least tightly bounded, in order to add only as much noise as needed to deliver the privacy guarantee. A consideration separate from model utility is to design release mechanisms that would be simple for data owners to implement, thus increasing potential uptake.

Input perturbation, such as used by Schein et al. [2018] via the geometric mechanism, enables local differential privacy. In this setup each individual perturbs their own data before transmitting to a central data collector, e.g. cellphones sending a noisy location to the server. A very desirable result of this form of privacy is the guarantee that no entity but the individual has unfettered access to the individual's data. The drawback, however, is the amount of noise added to the "system" scales with the number of individuals, which can hamper utility.

Output perturbation adds noise to the final learned model parameters [Dwork et al., 2006]. This is appealing from the information bottleneck standpoint, but if the learning algorithm is complex then it may be difficult to analyze the sensitivity and thus coarse bounds are often relied upon. Indeed, general private learning frameworks bound the sensitivity using quantities such as Lipschitz, strong-convexity, and smoothness constants [Bassily et al., 2014; Wu et al., 2016] or diameter of the parameter space [Smith, 2008], which may be loose in practice.

The *exponential mechanism* randomly samples from all possible outputs weighted by an assigned utility function. One instantiation of this mechanism is *one posterior sampling* (OPS; Dimitrakakis et al. [2014]; Foulds et al. [2016]; Wang et al. [2015]; Zhang et al. [2016]), which leverages the Bayesian inference framework to release a limited number of samples from a perturbed posterior. This approach, however, provides samples that do not correctly quantify beliefs of the model parameters.

In this thesis we take the approach of *sufficient statistics perturbation* (SSP; Foulds et al. [2016]; McSherry & Mironov [2009]; Vu & Slavkovic [2009]; Zhang et al. [2016]), in which noise is added to the sufficient statistics of a given model, which are a quantity that capture all information about the model parameters [Fisher, 1922]. The two most prevalent release mechanisms used for SSP are the Laplace mechanism Dwork et al. [2006] and the Gaussian mechanism Dwork et al. [2006]. SSP has a number of advantages. First, sufficient statistics are, by definition, an information bottleneck.

Second, it is very easy to exactly analyze the sensitivity of sufficient statistics in many models of interest; for example, the sufficient statistics of discrete, undirected graphical models are contingency tables. Third, adding noise to sufficient statistics prior to release is very simple, so it is reasonable to imagine adoption in practice, say, by public agencies.

Definition 3 (Laplace Mechanism; Dwork et al. [2006])**.** *Given a function f that maps data sets to $\mathbb{R}^m$, the Laplace mechanism outputs the random variable $\mathcal{L}(X) \sim \mathrm{Lap}\left(f(X), \Delta_f/\epsilon\right)$ from the Laplace distribution, which has density $\mathrm{Lap}(z; u, b) = (2b)^{-m} \exp\left(-\|z-u\|_1/b\right)$. This corresponds to adding zero-mean independent noise $u_i \sim \mathrm{Lap}(0, \Delta_f/\epsilon)$ to each component of $f(X)$.*

CHAPTER 3

PRIVATE UNDIRECTED GRAPHICAL MODELS

3.1 Introduction

Graphical models are a central tool in probabilistic modeling and machine learning. They pair expressive probability models with algorithms that leverage the graphical structure for efficient inference and learning. These tools allow a practitioner to posit a model for observed data, and then fit parameters, assess model validity, and make predictions. This chapter addresses the problem of privately learning parameters in a widely used class of probabilistic models: discrete, undirected graphical models.

Previous work addresses private learning for *directed* graphical models [Zhang et al., 2014, 2016]. Our problem of learning in undirected models, which are not locally normalized, is more general and substantially harder computationally. Several OPS approaches show that a single sample drawn from a posterior distribution is differentially private [Dimitrakakis et al., 2014; Wang et al., 2015; Zhang et al., 2016]. This can be understood as applying the exponential mechanism to the log-likelihood function, and can provide a point estimate for graphical model parameters [Zhang et al., 2016]. To apply OPS, one must sample from the posterior over parameters, $p(\Theta|X)$, which is straightforward for directed graphical models with conjugate priors, but not in undirected models, where posteriors over parameters are usually intractable. Foulds et al. [2016] and Zhang et al. [2016] also developed Bayesian methods using Laplace noise-corrupted sufficient statistics to update posterior parameters. Similar considerations apply to this approach, which matches ours in that it uses the same data release mechanism, but, like OPS, requires conjugate priors and thus easily

applies only to directed graphical models. Wang et al. [2015] also describe MCMC approaches to draw many private samples from a posterior distribution; this is another general framework that could apply to our problem, but, it relies on loose sensitivity bounds and since we only request point estimates, it would waste privacy budget by drawing many samples.

Because sufficient statistics of discrete, undirected graphical models are contingency tables, our work connects to the well-studied problem of releasing differentially private contingency tables [Barak et al., 2007; Hardt et al., 2012; Yang et al., 2012]. It is not entirely clear, however, how to learn parameters of a graphical model with *noisy* sufficient statistics. One option, which we will refer to as *naive maximum likelihood estimation* (MLE), is to ignore the noise and conduct maximum-likelihood estimation as if we had true sufficient statistics. This works reasonably well in practice, and is competitive with or better than state-of-the-art general-purpose methods. In fact, we will show that naive MLE is consistent and achieves the same *asymptotic* mean-squared error as non-private MLE. However, at reasonable sample sizes the error due to privacy is significant, and the approach has several pathologies (see also Yang et al. [2012] and Karwa et al. [2014, 2016]), some of which make it difficult to apply in practice. We therefore adopt a more principled approach of performing *noise-aware inference* about the true sufficient statistics within an expectation–maximization (EM) learning framework.

Thus the problem is how to conduct inference over sufficient statistics of a graphical model given noisy observations thereof. This is exactly the goal of inference in *collective graphical models* (CGMs; Sheldon & Dietterich [2011]), and we will adapt CGM inference techniques to solve this problem. Put together, our results significantly advance the state-of-the-art for privately learning discrete, undirected graphical models. We clarify the theory and practice of naive MLE and show that it learns better models than existing state-of-the-art approaches in most scenarios across a broad range of

synthetic tasks, and in experiments modeling human mobility from wifi access point data. We then show the more principled approach of conducting inference with CGMs is superior to competing approaches in nearly all scenarios.

3.1.1 Problem statement

Our goal is to learn a probabilistic model $p(\mathbf{x})$ from the data set $\mathbf{X}$ while protecting the privacy of individuals. We will learn probability distributions $p(\mathbf{x})$ that are *undirected discrete graphical models* (also called Markov random fields [Koller & Friedman, 2009]). These are defined by a set of local *potential functions* of the form $\psi_C(\mathbf{x}_C)$, where $C \subseteq \{1, \ldots, T\}$ is an index set or *clique*, $\mathbf{x}_C$ is a subvector of $\mathbf{x}$ corresponding to C, and $\psi_C : \mathcal{X}^{|C|} \to \mathbb{R}^+$ assigns a *potential* value to each possible $\mathbf{x}_C$. The probability model is $p(\mathbf{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C)$ where $\mathcal{C}$ is the collection of cliques that appear in the model, and $Z = \sum_{\mathbf{x}} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C)$ is the normalizing constant or *partition function*. The graph G with node set $V = \{1, \ldots, T\}$ and edges between any two indices that co-occur in some $C \in \mathcal{C}$ is the *independence graph* of the model; therefore, each index set C is a clique in G.

For learning, it is most convenient to express the model in log-linear or exponential family form as:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \exp\left\{ \sum_{C \in \mathcal{C}} \sum_{i_C \in \mathcal{X}^{|C|}} \mathbb{I}\{\mathbf{x}_C = i_C\} \theta_C(i_C) - A(\boldsymbol{\theta}) \right\}. \tag{3.1}$$

In this expression: $\mathbb{I}\{\cdot\}$ is an indicator function; the variable $i_C \in \mathcal{X}^{|C|}$ denotes a particular setting of the variables $\mathbf{x}_C$; the *parameters* $\theta_C(i_C) = \log \psi_C(i_C)$ are log-potential values; the vector $\boldsymbol{\theta} \in \mathbb{R}^d$ is the concatenation of all parameters; and $A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$ is the log-partition function, with the dependence of Z on the parameters now made explicit. Note that, for any $\boldsymbol{\theta} \in \mathbb{R}^d$, the density is strictly positive: $p(\mathbf{x}; \boldsymbol{\theta}) > 0$ for all $\mathbf{x}$. This is true because the potential values $\psi_C(i_C)$ are strictly positive, so the log-potentials are finite.

The goal is to learn parameters $\hat{\boldsymbol{\theta}}$ from the data $\mathbf{X}$ in a way that is ϵ-differentially private and such that $p(\mathbf{x}; \hat{\boldsymbol{\theta}})$ is as accurate as possible. We will measure accuracy as Kullback-Leibler divergence from an appropriate reference distribution [Kullback & Leibler, 1951]. In synthetic experiments, we will measure the divergence $D\big(p(\cdot; \boldsymbol{\theta}) \| p(\cdot; \hat{\boldsymbol{\theta}})\big)$, where $p(\mathbf{x}; \boldsymbol{\theta})$ is the true density. For real data, we will measure the holdout log-likelihood $E_q\big[\log p(\mathbf{x}; \hat{\boldsymbol{\theta}})\big]$ where q is the empirical distribution of the holdout data, which is equal to a constant minus $D\big(q \| p(\cdot; \hat{\boldsymbol{\theta}})\big)$.

The problem of privately selecting which cliques to include in the model (i.e., *model selection* or *structure learning*) is interesting but not considered in this thesis; we assume the cliques $\mathcal{C}$ are fixed in advance by the modeler.

3.2 Approach

To develop our approach to privately learn graphical model parameters, we first discuss standard concepts related to maximum-likelihood estimation for graphical models.

Log-likelihood, sufficient statistics, marginals. From Eq. (3.1), the log-likelihood $\mathcal{L}(\boldsymbol{\theta}) = \log \prod_{i=1}^{N} p\big(\mathbf{x}^{(i)}; \boldsymbol{\theta}\big)$ of the entire data set can be written as

$$\mathcal{L}(\boldsymbol{\theta}) = \left[\sum_{C \in \mathcal{C}} \sum_{i_C \in \mathcal{X}^{|C|}} n_C(i_C) \theta_C(i_C) \right] - N A(\boldsymbol{\theta})$$

where $n_C(i_C) = \sum_{i=1}^{N} \mathbb{I}\{\mathbf{x}_C^{(i)} = i_C\}$ is a count of how many times the configuration i_C for the variables in clique C appears in the population. The collection of counts $\mathbf{n}_C = \big(n_C(i_C)\big)$ for all possible i_C is the (population) *contingency table* on clique C. Let $\mathbf{n}$ denote the vector concatenation of the contingency tables for all cliques. Then we can rewrite the log-likelihood more compactly as

$$\mathcal{L}(\boldsymbol{\theta}) = f(\mathbf{n}, \boldsymbol{\theta}) := \boldsymbol{\theta}^T \mathbf{n} - NA(\boldsymbol{\theta}) \tag{3.2}$$

The most common approach for parameter learning in graphical models is maximum likelihood estimation: find the parameters $\hat{\boldsymbol{\theta}}$ that maximize $\mathcal{L}(\boldsymbol{\theta})$. The resulting parameter vector $\hat{\boldsymbol{\theta}}$ is a *maximum-likelihood estimator* (MLE). It is clear from Eq. (3.2) that this problem depends on the data only through the contingency tables $\mathbf{n}$. Indeed, the clique contingency tables $\mathbf{n}$ are *sufficient statistics* of the model: they measure all of the information from the data set $\mathbf{X}$ that is relevant for estimating the parameter $\boldsymbol{\theta}$ [Fisher, 1922].

The algorithmic approach for maximum-likelihood estimation in graphical models is standard [Koller & Friedman, 2009], and we do not repeat the details here. However, there are a few concepts that are important for our development. The *marginals* of a graphical model are the marginal probabilities $\mu_C(i_C) = p(\mathbf{x}_C = i_C; \boldsymbol{\theta})$ for all cliques C and configurations i_C. Let $\boldsymbol{\mu}$ be the vector concatenation of all marginals, and note that $\boldsymbol{\mu} = \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{n}]/N$. Similarly, let $\hat{\boldsymbol{\mu}} = \mathbf{n}/N$ be the *data marginals*—these are marginal probabilities of the empirical distribution of the data.

Marginals play a fundamental role in estimation. First, note that we can divide Eq. (3.2) by N to see that the MLE only depends on the data through the data marginals $\hat{\boldsymbol{\mu}}$. However, we leave $\mathcal{L}(\boldsymbol{\theta})$ in the current form because it is more convenient for the CGM development in Section 3.2.4. Second, it is well known that $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = N(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$, so maximum likelihood estimation seeks to adjust $\boldsymbol{\theta}$ so that the data and model marginals match. Third, it can (almost) always succeed in doing so, even if the data marginals do not come from a graphical model. More formally, let $\mathcal{M}$ be the *marginal polytope*: the set of all vectors $\boldsymbol{\mu}$ such that there exists some distribution $q(\mathbf{x})$ with marginal probabilities $\boldsymbol{\mu}$.

Proposition 1 ([Wainwright & Jordan, 2008])**.** *For any $\boldsymbol{\mu}$ in the interior of $\mathcal{M}$, there is a unique distribution $p(\mathbf{x}; \boldsymbol{\theta})$ with marginals $\boldsymbol{\mu}$, i.e., such that $\boldsymbol{\mu} = E_{\boldsymbol{\theta}}[\mathbf{n}]/N$.*

Applying Proposition 1 to the data marginals $\hat{\boldsymbol{\mu}}$ shows that if these belong to the interior of $\mathcal{M}$, we may learn a distribution with marginals that match what we observe in the data. Note that, while the *distribution* $p(\mathbf{x};\boldsymbol{\theta})$ is unique, the parameters $\boldsymbol{\theta}$ are not, because our model is overcomplete. If $\boldsymbol{\mu}$ belongs to $\mathcal{M}$ but not the interior of $\mathcal{M}$, which occurs, for example, when some marginals are zero, the situation is more complex: there is no (finite) $\boldsymbol{\theta} \in \mathbb{R}^d$ such that $p(\mathbf{x};\boldsymbol{\theta})$ has marginals $\boldsymbol{\mu}$.[1] Similarly, the MLE does not exist, meaning that its maximum is not attained for any finite $\boldsymbol{\theta}$ [Fienberg & Rinaldo, 2012; Haberman, 1973]. This issue will end up being significant in our understanding of the naive MLE approach in the following section.

3.2.1 Noisy sufficient statistics

To use the Laplace mechanism to release noisy sufficient statistics we must first determine a bound on the sensitivity of $\mathbf{n}$, which is very easy to analyze and the analysis is tight: the local sensitivity (see Definition 2 in Chapter 2) is the same for all data sets.

Proposition 2. *Let* $\mathbf{n}(\mathbf{X})$ *be the sufficient statistics of a graphical model with clique set* $\mathcal{C}$ *on data set* $\mathbf{X}$*. The local sensitivity of* $\mathbf{n}$ *is* $|\mathcal{C}|$ *for all inputs* $\mathbf{X}$*. Therefore the sensitivity of* $\mathbf{n}$ *is* $|\mathcal{C}|$*.*

See Appendix A.1 for proof.

So, a simple approach to achieve privacy is to release noisy sufficient statistics $\mathbf{y}$ that are obtained after applying the Laplace mechanism:

$$y_C(i_C) = n_C(i_C) + \text{Laplace}\big(|\mathcal{C}|/\epsilon\big) \tag{3.3}$$

[1] However, there is a sequence $\{\boldsymbol{\theta}^k\}$ where $\boldsymbol{\theta}^k \in \mathbb{R}^d$ and $\lim_{k\to\infty} \mathbb{E}_{\boldsymbol{\theta}^k}[\mathbf{n}]/N = \boldsymbol{\mu}$.

3.2.2 Positive results for naive MLE

How can we learn with noisy sufficient statistics $\mathbf{y}$? A naive approach is to use $\mathbf{y}$ in place of $\mathbf{n}$ in maximum-likelihood estimation, i.e., to find $\hat{\boldsymbol{\theta}}$ to maximize $f(\mathbf{y}, \boldsymbol{\theta})$. The validity of this approach has been debated in the literature [Yang et al., 2012]. However, it is relatively easy to show that it behaves well asymptotically.

Proposition 3. *Assume $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$ are drawn iid from a probability distribution with marginals $\boldsymbol{\mu}$. The marginal estimate $\bar{\mu}_C(i_C) = \frac{1}{N} y_C(i_C)$ obtained from the noisy sufficient statistics is unbiased and consistent, with mean squared error:*

$$\mathrm{MSE}\left(\bar{\mu}_C(i_C)\right) = \frac{\mu_C(i_C)\big(1 - \mu_C(i_C)\big)}{N} + \frac{2|\mathcal{C}|^2}{N^2\epsilon^2} \tag{3.4}$$

Now let $\hat{\boldsymbol{\theta}} \in \mathrm{argmax}_{\boldsymbol{\theta}} f(\mathbf{y}, \boldsymbol{\theta})$ be parameters estimated using the noisy sufficient statistics $\mathbf{y}$. If the true distribution $p(\mathbf{x}; \boldsymbol{\theta})$ is a graphical model with cliques $\mathcal{C}$, then the estimated distribution $p(\mathbf{x}; \hat{\boldsymbol{\theta}})$ converges to $p(\mathbf{x}; \boldsymbol{\theta})$.

See Appendix A.1 for proof.

3.2.3 Pathologies in naive MLE

Asymptotically, the noisy sufficient statistics behave as desired in terms of MSE: the $O(1/N)$ term, which is due to sampling error and not privacy, dominates for large N. However, for practical settings of ϵ the $O(1/N^2)$ term, which is due to privacy, is dominant until N becomes very large, due to the large constant $2|\mathcal{C}|^2/\epsilon^2$. Figure 3.1 illustrates this issue. For large ϵ, the $O(1/N)$ sampling error is dominant; however, for smaller ϵ, the $O(1/N^2)$ privacy error term is dominant even for N approaching 10^7. [2]

A second pathology is that the noise added for privacy destroys some of the structure expected in the empirical marginals. The true data marginals $\hat{\boldsymbol{\mu}} = \mathbf{n}/N$

[2]Note that Proposition 1 suggests that the MSE results for the *estimated* marginals $\hat{\boldsymbol{\mu}}$ will carry over to marginals of the *learned* model $p(\mathbf{x}; \hat{\boldsymbol{\theta}})$. However the situation is complicated by the fact that $\hat{\boldsymbol{\mu}}$ does not belong to the marginal polytope. Despite this, we observe in practice that the MSE of the learned marginals follow the predictions of Proposition 3.

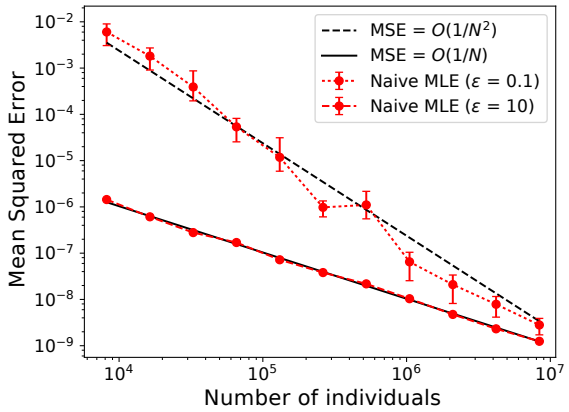


Figure 3.1: Sample results on synthetic data illustrating behavior of naive MLE (see Section 3.3.2 for experiment details). MSE of learned marginals vs population size N on a chain model with $T = 10$, $|\mathcal{X}| = 10$; reference lines indicate predicted slope for $O(1/N)$ and $O(1/N^2)$ error terms, respectively (the function c/N^d has slope $-d$ on a log-log plot)

belong to the marginal polytope: in particular, this means that each clique marginal $\hat{\boldsymbol{\mu}}_C$ is nonnegative and sums to one, and that clique marginals agree on common subsets of variables. After adding noise, the *pseudo*-marginals $\bar{\boldsymbol{\mu}} = \mathbf{y}/N$ do not belong to the marginal polytope: $\bar{\boldsymbol{\mu}}$ may have negative values, and does not satisfy consistency constraints. We find that a partial fix is very helpful empirically: project the pseudo-marginal $\bar{\boldsymbol{\mu}}_C$ for each clique onto the simplex prior to conducting MLE, which can be done via a standard procedure [Duchi et al., 2008]. Let $\tilde{\boldsymbol{\mu}}$ be the projected marginals. We now have that $\tilde{\boldsymbol{\mu}}_C$ is a valid marginal for each clique C, but consistency constraints are not satisfied among cliques, and it is still the case that $\tilde{\boldsymbol{\mu}} \notin \mathcal{M}$. Figure 3.2 illustrates the benefits of projection on the quality of the model learned by Naive MLE.

A more significant pathology has to do with zeros in the projected marginals $\tilde{\boldsymbol{\mu}}$, which are more prevalent than in true data marginals $\hat{\boldsymbol{\mu}}$. This is because the addition of Laplace noise creates negative values, which are then truncated to zero during projection. As discussed following Proposition 1, zero values in the marginals lead to non-existence of the MLE [Fienberg & Rinaldo, 2012; Haberman, 1973]. If

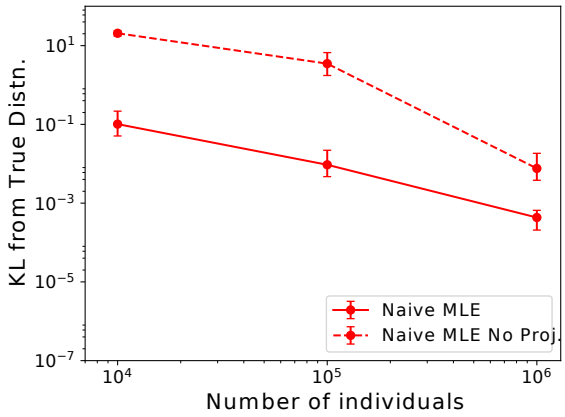


Figure 3.2: Sample results on synthetic data illustrating behavior of naive MLE (see Section 3.3.2 for experiment details). The effect of projecting marginals on performance of naive MLE for an Erdős-Réyni graph with $T = 10$, $|\mathcal{X}| = 20$, $\epsilon = 0.5$.

$\tilde{\mu}_C(i_C) = 0$, the likelihood increases monotonically as $\theta_C(i_C)$ goes to negative infinity; in other words, the model attempts to drive the learned marginal probability to zero. Numerically, we can address this by regularization, e.g., adding $\lambda\|\boldsymbol{\theta}\|^2$ to the objective function for arbitrarily small $\lambda > 0$. However, we may still learn vanishingly small marginal probabilities, which can lead to a very large KL-divergence between the true and learned models. Figure 3.3 illustrates the effect of λ on KL-divergence with both noisy sufficient statistics and true sufficient statistics. At high λ (strong regularization), both methods underfit and yield poor KL divergence. Learning with true sufficient statistics has no tendency to overfit; it achieves good performance for a broad range of λ approaching zero. Naive MLE with noisy sufficient statistics overfits badly (to zeros) for small λ, and must be tuned "just right" to achieve reasonable performance.

3.2.4 Collective Graphical Models

Since learning with noisy sufficient statistics "as-is" has several pathologies and is less robust than maximum-likelihood estimation in the absence of privacy, we investigate a more principled approach, which matches the data generating process: We treat the true sufficient statistics $\mathbf{n}$ as latent variables, and learn $\boldsymbol{\theta}$ to maximize

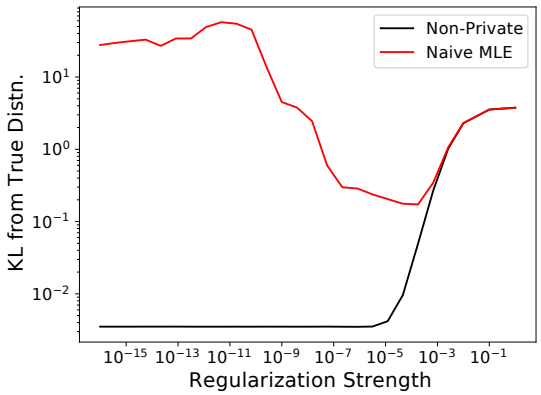


Figure 3.3: Sample results on synthetic data illustrating behavior of naive MLE (see Section 3.3.2 for experiment details). The effect of regularization on KL-divergence for learning with and without privacy; chain model with $T = 10$, $|\mathcal{X}| = 10$, $\epsilon = 0.1$.

the *marginal* likelihood $p(\mathbf{y}; \boldsymbol{\theta}) = \sum_{\mathbf{n}} p(\mathbf{n}, \mathbf{y}; \boldsymbol{\theta})$. In this section, we will develop an EM approach to accomplish this.

In EM, we need to conduct inference to compute $\mathbb{E}[\mathbf{n} \mid \mathbf{y}; \boldsymbol{\theta}]$ for a fixed value of $\boldsymbol{\theta}$. This is the central problem of *collective graphical models* (CGMs) [Sheldon & Dietterich, 2011]. Consider the joint distribution $p(\mathbf{n}, \mathbf{y}; \boldsymbol{\theta}) = p(\mathbf{n}; \boldsymbol{\theta}) p(\mathbf{y} \mid \mathbf{n})$, which we use to compute $\mathbb{E}[\mathbf{n} \mid \mathbf{y}; \boldsymbol{\theta}]$. The noise mechanism $p(\mathbf{y} \mid \mathbf{n})$ arises directly from the Laplace mechanism (see Eq. (3.3)). The distribution of the sufficient statistics, $p(\mathbf{n}; \boldsymbol{\theta})$, is known as the *CGM distribution*. It can be written in closed form when the model is *decomposable*, i.e., the cliques $\mathcal{C}$ correspond to the nodes of some junction tree $\mathcal{T}$. Although decomposability is a significant restriction, let us assume that such a tree $\mathcal{T}$ exists; we will use the exact results derived for this case to develop an approximation for the general case. Let $\mathcal{S}$ be the set of separators of $\mathcal{T}$, and let $\nu(S)$ be the multiplicity of $S \in \mathcal{S}$, i.e., the number of distinct edges $(C_i, C_j) \in \mathcal{T}$ for which $S = C_i \cap C_j$. Under these assumptions, the CGM distribution has the form [Liu et al., 2014]:

$$p(\mathbf{n};\boldsymbol{\theta}) = h(\mathbf{n}) \cdot \exp\big(f(\mathbf{n},\boldsymbol{\theta})\big),$$
$$h(\mathbf{n}) = N! \cdot \frac{\prod_{S\in\mathcal{S}} \prod_{i_S\in\mathcal{X}^{|S|}} \big(n_S(i_S)!\big)^{\nu(S)}}{\prod_{C\in\mathcal{C}} \prod_{i_C\in\mathcal{X}^{|C|}} n_C(i_C)!} \cdot \mathbb{I}\{\mathbf{n}\in\mathcal{M}_N^{\mathbb{Z}}\}$$

The term $\exp\big(f(\mathbf{n},\boldsymbol{\theta})\big)$ is the probability of an ordered data set $\mathbf{X}$ with sufficient statistics $\mathbf{n}$, as discussed previously. The term $h(\mathbf{n})$ is a base measure that counts the number of ordered data sets with sufficient statistics equal to $\mathbf{n}$, and enforces constraints on $\mathbf{n}$. The *integer-valued marginal polytope* $\mathcal{M}_N^{\mathbb{Z}}$ is the set of all vectors $\mathbf{n}$ that are sufficient statistics of some data set $\mathbf{X}$ of size N.

Exact inference in CGMs is intractable [Sheldon et al., 2013]. Therefore, it is typical to relax the integrality constraint and apply Stirling's approximation: $\log n! \approx n\log n - n$. Let $\mathcal{M}_N$ be the feasible set with the integrality constraint removed, which is now just the standard marginal polytope scaled so that each marginal sums to N instead of one.

Proposition 4 (Nguyen et al. [2016]; Sun et al. [2015])**.** *For a decomposable CGM with junction tree $\mathcal{T}$, the following approximation of the CGM log-density for any $\mathbf{n}\in\mathcal{M}_N$ is obtained by applying Stirling's approximation:*

$$\log p(\mathbf{n},\mathbf{y};\boldsymbol{\theta}) \approx \boldsymbol{\theta}^T\mathbf{n} - NA(\boldsymbol{\theta}) + H(\mathbf{n}) + \log p(\mathbf{y}|\mathbf{n}). \tag{3.5}$$

Here, $H(\mathbf{n}) = -N\sum_{\mathbf{x}} q(\mathbf{x})\log q(\mathbf{x})$ is the entropy of the unique distribution $q(\mathbf{x}) = p(\mathbf{x};\boldsymbol{\theta})$ in the graphical model family with marginals equal to $\mathbf{n}/N$.

See Appendix A.1 for proof.

Proposition 4 is the basis for *approximate MAP inference problem in CGMs*: find $\mathbf{n}$ to maximize Eq. (3.5) and obtain an approximate mode of $p(\mathbf{n}\mid\mathbf{y};\boldsymbol{\theta})$. Even though our goal is to compute the *mean* $\mathbb{E}[\mathbf{n}\mid\mathbf{y};\boldsymbol{\theta}]$, it has been shown that the approximate mode, which is also a real-valued vector, is an excellent approximation to the mean for

use within the EM algorithm [Sheldon et al., 2013]. Note that for non-decomposable models, we will simply apply the same approximation as in Proposition 4, even though an exact expression for the counting measure $h(\mathbf{n})$, and therefore the correspondence of $\log h(\mathbf{n})$ to an entropy $H(\mathbf{n})$, is not known in this case. Then, after dropping the term $NA(\boldsymbol{\theta})$ from Proposition 4, which is constant with respect to $\mathbf{n}$, the approximate MAP problem can be rewritten as:

$$\mathbf{n}^* \in \operatorname*{argmax}_{\mathbf{n} \in \mathcal{M}_N} \boldsymbol{\theta}^T \mathbf{n} + H(\mathbf{n}) + \log p(\mathbf{y} \mid \mathbf{n}) \tag{3.6}$$

This equation reveals a close connection to variational principles for graphical models [Wainwright & Jordan, 2008]. It is identical to the variational optimization problem for marginal inference in standard graphical models, except the objective has an additional term $\log p(\mathbf{y}|\mathbf{n})$, which is non-linear in $\mathbf{n}$. Several message-passing based algorithms have been developed to efficiently solve the approximate MAP problem. For trees or junction trees, Problem (3.6) is convex as long as $\log p(\mathbf{y}|\mathbf{n})$ is concave in $\mathbf{n}$ (which is true in most cases of interest, such as Laplace noise) so it can be solved exactly [Sun et al., 2015; Vilnis et al., 2015]. For loopy models, both the entropy $H(\mathbf{n})$ and the feasible set $\mathcal{M}_N$ must be approximated [Nguyen et al., 2016].

Algorithm 1 shows pseudocode *non-linear belief propagation* (NLBP [Sun et al., 2015]), which we select as our primary inference approach due to its simplicity. It is a thin wrapper around standard BP, and can be applied to trees, in which case it exactly solves Problem (3.6), or it can be applied to loopy graphs by using loopy BP (LBP) as the subroutine, in which case it is approximate.

Our final EM learning procedure is shown in Algorithm 2. It alternates between inference steps that solve the approximate MAP problem to find $\mathbf{n}_t \approx \mathbb{E}[\mathbf{n} \mid \mathbf{y}; \theta_t]$, and optimization steps to re-estimate parameters given the inferred sufficient statistics $\mathbf{n}_t$. See also Sheldon et al. [2013], Liu et al. [2014], and Sun et al. [2015].

Algorithm 1 Non-Linear Belief Propagation (NLBP)

input: $\boldsymbol{\theta}$, $\mathbf{y}$, damping parameter $\alpha > 0$
while $\neg$ converged **do**
  $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \nabla_{\mathbf{n}} \log p(\mathbf{y} \mid \mathbf{n})$
  $\mathbf{n}' \leftarrow \text{STANDARD-BP}(\boldsymbol{\theta}')$ ▷ Normalized to sum to N
  $\mathbf{n} \leftarrow (1 - \alpha)\mathbf{n} + \alpha\mathbf{n}'$
return: $\mathbf{n}$

Algorithm 2 EM for CGMs

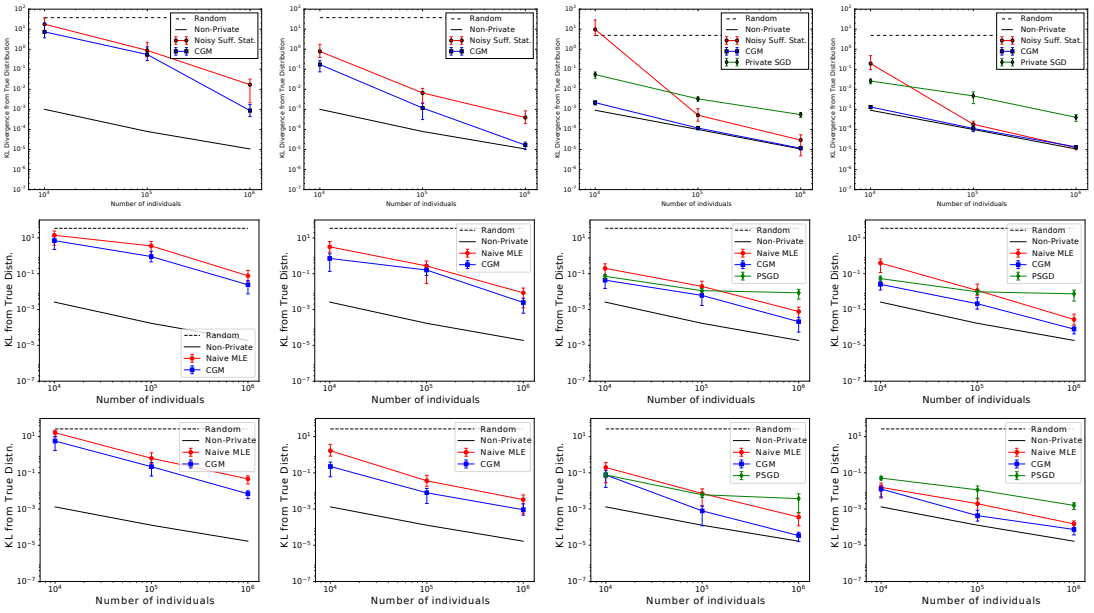
input: $\mathbf{y}$
while $\neg$ converged **do**
  $\mathbf{n}_t \leftarrow \text{NLBP}(\boldsymbol{\theta}_t, \mathbf{y})$
  $\boldsymbol{\theta}_{t+1} \leftarrow \text{argmax}_{\boldsymbol{\theta}}\, \boldsymbol{\theta}^T \mathbf{n}_t - NA(\boldsymbol{\theta})$
return: $\boldsymbol{\theta}_{t+1}$

3.3 Experiments

We conduct a number of experiments on synthetic and real data to evaluate the quality of models learned by both naive MLE and CGM.

3.3.1 Methods

We compare three algorithms: naive MLE, CGM, and a version of private stochastic gradient descent (PSGD) due to Abadi et al. [2016]. PSGD belongs to a class of general-purpose private learning algorithms that can be adapted to our problem, including gradient descent or stochastic gradient descent algorithms for empirical risk minimization [Abadi et al., 2016; Bassily et al., 2014; Chaudhuri et al., 2011; Jain & Thakurta, 2013; Kifer et al., 2012] and the subsample-and-aggregate approach for parameter estimation [Smith, 2011b]. We chose PSGD because it is a state-of-the-art method and it significantly outperformed other approaches in preliminary experiments. However, note that PSGD satisfies only (ϵ, δ)-*differential privacy* for $\delta > 0$, which is a weaker privacy guarantee than ϵ-differential privacy. We tune PSGD using a grid search over all relevant parameters to ensure it performs as well as possible.



(a) $\epsilon = 0.01$ (b) $\epsilon = 0.1$ (c) $\epsilon = 0.5$ (d) $\epsilon = 1.0$

Figure 3.4: Results on synthetic data generated from first-order chains (top-row), third-order chains (middle-row), and connected Erdős-Réyni random graphs (bottom-row). Each column represents a different privacy level. Lower ϵ signifies stricter privacy guarantees. The x-axis measures population size. The y-axis is KL divergence from the true distribution.

3.3.2 Synthetic data

We evaluate three types of pairwise graphical models: first order chains, third-order chains with edges between two nodes i and j if $1 \leq |i - j| \leq 3$, and (connected) Erdős-Réyni (ER) random graphs. We report results for graphs of 10 nodes, where potentials on each edge are drawn from a Dirichlet distribution with concentration parameter of one; results are similar for smaller and larger models, models with different structures, and for different types of potentials. We vary data size N and privacy parameter ϵ. For each setting of model type, N, and ϵ, we conduct 25 trials. The trials are nested, with five random populations and five replications per population, i.e.: $\mathbf{n}_i \sim p(\mathbf{n}), \mathbf{y}_{i,j} \sim p(\mathbf{y} \mid \mathbf{n}_i)$ for $i \in \{1, \ldots, 5\}, j \in \{1, \ldots, 5\}$. We measure the quality of learned models using KL divergence from the true distribution, and include

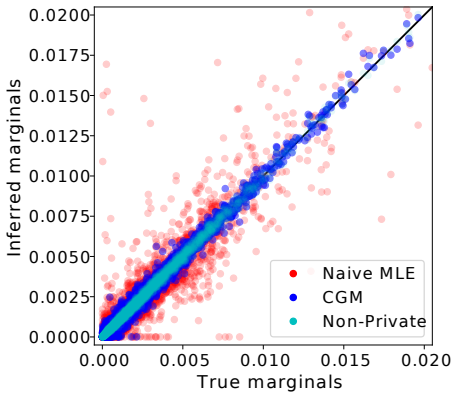


Figure 3.5: Scatter plots for true vs. inferred values of all edge marginals in an ER graph of 10 nodes with 20 states each.

for comparison two reference models: a random estimator and a non-private MLE estimator. The random estimator is obtained by randomly generating marginals $\bar{\boldsymbol{\mu}}$ and then learning potentials via MLE.

3.3.3 Results

Figure 3.4 shows the results for the two models (top: third-order chain, bottom: ER) for different values of N and ϵ. CGM improves upon naive MLE for all models, privacy levels, and population sizes. Recall that PSGD promises only (ϵ, δ)-differential privacy. While δ is often assumed to be "cryptographically small", e.g., $O(2^{-N})$, we set δ to a relatively large value of $\delta = 1/N$. Increasing δ weakens the privacy guarantee but enables PGSD to run on a wider range of ϵ. However, even with this setting for δ, some of the smaller values of ϵ are not attainable by PGSD and are omitted from those plots.

Figure 3.5 shows a qualitative comparison of edge marginals of a single graph learned by the different methods, compared with the true model marginals; it is evident that CGM learns marginals that are much closer to both the true marginals and those learned by the non-private estimator than naive MLE is able to learn. Naive

MLE is the fastest method; CGM is approximately 4x/8x slower on third-order chains and ER graphs, respectively, and PSGD is approximately 27x/40x slower.

3.3.4 Wifi data

We study human mobility data in the form of connections to wifi access points throughout a highly-trafficked academic building over a twenty-one day period. We treat each (user ID, day) combination as an "individual", leading to 124,399 unique individuals; with this data preparation scheme, the unit of protection is one day's worth of a user's data. We discretize time by recording the location every 10 minutes, and assign null if the user is not connected to the network. Our probability model $p(\mathbf{x})$ is a pairwise graphical model over hour-long segments. Therefore, we break each individual's data into 24 one-hour long segments.

An individual now contributes 24 records to each contingency table for the model $p(\mathbf{x})$. Therefore, the sensitivity is now 24 times the number of edges (cliques). However, real data is typically sparse—i.e., an individual is typically observed only a small number of times over the observation period. Therefore, to reduce the sensitivity, the data is *normalized* prior to calculating sufficient statistics, in a fashion similar to He et al. [2015]. Each user contributes a value of $1/K$ to each contingency table, where K is the number of edges (x_s, x_t) for which the user's values are not both null. With this pre-processing in place, the sensitivity equals the number of edges in the model. A trade-off of this technique is that we bias the model towards individuals with fewer transitions, but we reduce the amount of noise by limiting sensitivity caused by null–null transitions.

We reserve data from 25% of the individuals for testing. To compare different approaches, we apply naive MLE, CGM, and PSGD to privately learn parameters of a graphical model from the training set (75% of the data), with varying privacy levels. We then calculate holdout log-likelihood of the learned parameters on the test set.

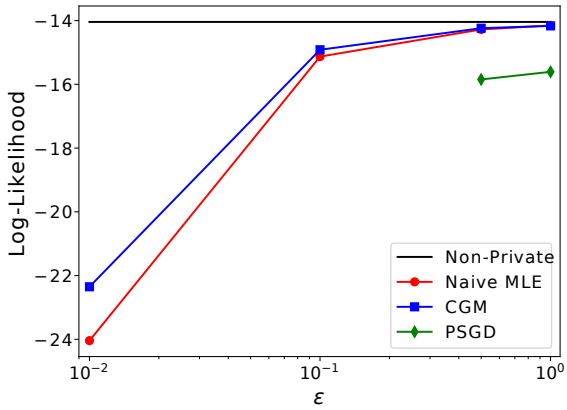


Figure 3.6: Results for fitting a first-order chain on wifi data. The x-axis is privacy level; lower ϵ signifies stronger privacy guarantees. The y-axis is holdout log-likelihood.

We again include a non-private method for reference, but in this case, all methods perform better than the random estimator, so we do not show it.

Figure 3.6 shows the results for fitting a time-homogeneous chain model (edges between adjacent time steps, every potential $\psi(x_t, x_{t+1})$ is the same, and the model includes a node potential $\phi(x_1)$ so it can learn a time-stationary model). As in the synthetic data experiments, CGM improves upon naive MLE across all parameter regimes, and performance improves with population size N and with weakening of privacy (larger ϵ). Both methods outperform PSGD. Naive MLE is the fastest method; CGM is approximately 15x slower, and PSGD is approximately 46x slower.

CHAPTER 4

PRIVATE BAYESIAN INFERENCE FOR EXPONENTIAL FAMILY MODELS

4.1 Introduction

There is a growing interest in private methods for Bayesian inference [Dimitrakakis et al., 2014; Foulds et al., 2016; Geumlek et al., 2017; Wang et al., 2015; Zhang et al., 2016]. In Bayesian inference, a modeler selects a prior distribution $p(\theta)$ over some parameter, observes data x that depends probabilistically on θ through a model $p(x \mid \theta)$, and then reasons about θ through the posterior distribution $p(\theta \mid x)$, which quantifies updated beliefs and uncertainty about θ after observing x. Bayesian inference is a core machine learning task and there is an obvious need to be able to conduct it in a way that protects privacy when x is sensitive. Additionally, recent work has identified surprising connections between sampling from posterior distributions and differential privacy—for example, a single perfect sample from $p(\theta \mid x)$ satisfies differential privacy for some setting of the privacy parameter [Dimitrakakis et al., 2014; Foulds et al., 2016; Wang et al., 2015; Zhang et al., 2016].

An "obvious" way to conduct private Bayesian inference is to privatize the computation of the posterior, that is, to design a differentially private algorithm $\mathcal{A}$ that outputs $y = \mathcal{A}(x)$ with the goal that $y \approx p(\theta \mid x)$ is a privatized representation of the posterior. However, using y directly as "the posterior" will not correctly quantify beliefs, because the Bayesian modeler never observes x, they observe y; their posterior beliefs are now quantified by $p(\theta \mid y)$.

We will take a different approach to private Bayesian inference by specifying a pairing of algorithms: The release mechanism $\mathcal{A}$ computes a private statistic $y = \mathcal{A}(x)$

of the input data; the inference algorithm $\mathcal{P}$ computes $p(\theta \mid y)$. These algorithms should satisfy the following criteria:

- **Privacy**. The release mechanism $\mathcal{A}$ is differentially private. By the post-processing property of differential privacy [Dwork & Roth, 2014], all further computations are also private.
- **Calibration**. The inference algorithm $\mathcal{P}$ can efficiently compute or approximate the correct posterior, $p(\theta \mid y)$ (see Section 4.4 for our process to measure calibration).
- **Utility**. Informally, the statistic y should capture "as much information as possible" about x so that $p(\theta \mid y)$ is "close" to $p(\theta \mid x)$ (see Section 4.4 for our process to measure utility).

Importantly, the release mechanism $\mathcal{A}$ is public, so the distribution $p(y \mid x)$ is known. Williams and McSherry first suggested conducting inference on the output of a differentially private algorithm and showed how to do this for the factored exponential mechanism Williams & McSherry [2010]; see also Karwa et al. [2014], Karwa et al. [2016], Bernstein et al. [2017], and Schein et al. [2018].

This chapter focuses specifically on Bayesian inference when the private data $X = x_{1:n}$ is an iid sample of (publicly known) size n from an exponential family model $p(x \mid \theta)$. Exponential families include many of the most familiar parametric probability models. We will adopt the straightforward Laplace mechanism (see Chapter 1.3), where the sufficient statistics are corrupted with a random Laplace draw and the subsequent noisy sufficient statistics y are released [Bernstein et al., 2017; Foulds et al., 2016].

The technical challenge is then to develop an efficient general-purpose inference algorithm $\mathcal{P}$. One challenge is computational efficiency. The exact posterior $p(\theta \mid y) \propto \int p(\theta)p(x_{1:n} \mid \theta)p(y|x_{1:n})dx_{1:n}$ integrates over all possible data sets [Williams & McSherry, 2010], which is intractable to do directly for large n. We integrate instead over the sufficient statistics s, which have fixed dimension and completely

characterize the posterior; furthermore, since they are a sum over individuals, $p(s \mid \theta)$ is asymptotically normal. We develop an efficient Gibbs sampler that uses a normal approximation for s together with variable augmentation to model the Laplace noise in a way that yields simple updates [Park & Casella, 2008].

A second challenge is that the sufficient statistics may be unbounded, which makes their release incompatible with the Laplace mechanism. We address this by imposing truncation bounds and only computing statistics from data that fall within the bounds. We show how to use automatic differentiation and a "random sum" central limit theorem to compute the parameters of the normal approximation $p(s \mid \theta)$ for a *truncated* exponential family when the number of individuals that fall within the truncation bounds is unknown.

Our overall contribution is the pairing of an existing simple release mechanism $\mathcal{A}$ with a novel, efficient, and general-purpose Gibbs sampler $\mathcal{P}$ that meets the criteria outlined above for private Bayesian inference in any univariate exponential family or multivariate exponential family with bounded sufficient statistics.[1] We show empirically that when compared with competing methods, ours is the only one that provides properly calibrated beliefs about θ in the non-asymptotic regime, and that it provides good utility compared with other private Bayesian inference approaches.

We consider the canonical setting of Bayesian inference in an exponential family. The modeler posits a prior distribution $p(\theta)$, assumes the data $x_{1:n}$ is an iid sample from an exponential family model $p(x \mid \theta)$, and wishes to compute the posterior $p(\theta \mid x_{1:n})$. An exponential family in natural parameterization has density

$$p(x \mid \eta) = h(x) \exp\left(\eta^T t(x) - A(\eta)\right),$$

[1]There are remaining technical challenges for multivariate models with unbounded sufficient statistics that we leave for future work.

where η are the natural parameters, $t(x)$ is the sufficient statistic, $A(\eta) = \int h(x) \exp\left(\eta^T t(x)\right) dx$ is the log-partition function, and $h(x)$ is the base measure. The density of the full data is

$$p(x_{1:n} \mid \eta) = h(x_{1:n}) \exp\left(\eta^T t(x_{1:n}) - nA(\eta)\right),$$

where $h(x_{1:n}) = \prod_{i=1}^n h(x_i)$ and $t(x_{1:n}) = \sum_{i=1}^n t(x_i)$. Notice that once normalizing constants are dropped, this density is dependent on the data only directly through the sufficient statistics, $s = t(x_{1:n})$.

We will write exponential families more generally as $p(x \mid \theta)$ to indicate the case when the natural parameters $\eta = \eta(\theta)$ depend on a different parameter vector θ.

Every exponential family distribution has a conjugate prior distribution $p(\theta; \lambda)$[Diaconis & Ylvisaker, 1979] with hyperparameters λ. A conjugate prior has the property that, if it is used as the prior, then the posterior belongs to the same family, i.e., $p(\theta \mid x_{1:n}; \lambda) = p(\theta; \lambda')$ for some λ' that depends only on λ, n, and the sufficient statistics s. We write this function as $\lambda' = \text{Conjugate-Update}(\lambda, s, n)$; our methods are not tied to the specific choice of conjugate prior, only that the posterior parameters can be calculated in this form. See Section B.1 for a general form of Conjugate-Update.

Release algorithm: noisy sufficient statistics If privacy were not a concern, the Bayesian modeler would simply compute the sufficient statistics $s = t(x_{1:n})$ and use them to update the posterior beliefs. However, to maintain privacy, the modeler must access the sensitive data only through a randomized release mechanism $\mathcal{A}$. As a result, in order to obtain proper posterior beliefs the modeler must account for the randomization of the release mechanism by performing inference.

We take the simple approach of releasing noisy sufficient statistics via the Laplace mechanism, as in [Bernstein et al., 2017; Foulds et al., 2016; Zhang et al., 2016]. Sufficient statistics are a natural quantity to release. They are an "information bottleneck"—a finite-dimensional quantity that captures all the relevant information

about θ. The released value is $y = \mathcal{A}(x_{1:n}) \sim \text{Lap}(s, \Delta_s/\epsilon)$. Because $s = t(x_{1:n}) = \sum_{i=1}^{n} t(x_i)$ is a sum over individuals, the sensitivity is $\Delta_s = \max_{x,x' \in \mathbb{R}^d} \|t(x) - t(x')\|_1$. When $t(\cdot)$ is unbounded this quantity becomes infinite; we will modify the release mechanism so the sensitivity is finite (Sec. 4.3).

4.2 Basic Inference Approach: Bounded Sufficient Statistics

The goal of the inference algorithm $\mathcal{P}$ is to compute $p(\theta \mid y)$. We first develop the basic approach for the simpler case when $t(x)$ is bounded, and then extend both $\mathcal{A}$ and $\mathcal{P}$ to handle the unbounded case. The full joint distribution of the probability model can be expressed as:

$$p(\theta, s, y) = p(\theta)\, p(s \mid \theta)\, p(y \mid s),$$

where $p(\theta) = p(\theta; \lambda)$ is a conjugate prior and the goal is to compute a representation of $p(\theta \mid y) \propto \int_s p(\theta, s, y) ds$ by integrating over the sufficient statistics.

We will develop a Gibbs sampler to sample from this distribution. There are two main challenges. First, the distribution $p(s \mid \theta)$ is obtained by marginalizing over the data sample $x_{1:n}$, and is usually not known in closed form. We will address this with an asymptotically correct normal approximation. Second, when resampling s within the Gibbs algorithm, we require the full conditional distribution of s given the other variables, which is proportional to $p(s|\theta)p(y \mid s)$. Care must be taken to make it easy to sample from this conditional distribution. We address this via variable augmentation. We discuss our approach to both challenges in detail below.

4.2.1 Normal approximation of $p(s \mid \theta)$

The exact form of the sufficient statistic distribution $p(s \mid \theta)$ is obtained by marginalizing over the data:

$$p(s \mid \theta) = \int_{t^{-1}(s)} p(x_{1:n} \mid \theta) dx_{1:n}, \qquad t^{-1}(s) := \big\{x_{1:n} : t(x_{1:n}) = s\big\}.$$

In general, the exact form of this distribution is not available. In some cases, it is—for example if $x \sim \text{Bernoulli}(\theta)$ then $s \sim \text{Binomial}(n, \theta)$—but even then it may not lead to a tractable full conditional for s.

Properties of exponential families pave the way toward a general approach that always leads to a tractable full conditional. By the central limit theorem (CLT), because $s = \sum_i t(x_i)$ is a sum of iid random variables, it is asymptotically normal. It can be approximated as $p(s \mid \theta) \approx \mathcal{N}(s; n\mu, n\Sigma)$, where $\mu = \mathbb{E}[t(x)]$ and $\Sigma = \text{Var}[t(x)]$ are the mean and variance of the sufficient statistic of a single individual. This approximation is asymptotically correct: $\frac{1}{\sqrt{n}}(s - n\mu) \xrightarrow{D} \mathcal{N}(0, \Sigma)$ [Bickel & Doksum, 2015]. The quantities μ and Σ can be computed using well-known properties of exponential families [Bickel & Doksum, 2015]:

$$\mu = \mathbb{E}[t(x)] = \frac{\partial}{\partial \eta^T} A(\eta), \qquad \Sigma = \text{Var}[t(x)] = \frac{\partial^2}{\partial \eta \partial \eta^T} A(\eta), \tag{4.1}$$

where $\eta = \eta(\theta)$ is the natural parameter.

Note that we will *not* use this approximation for Gibbs updates of θ. Instead, we will compute the conditional $p(\theta \mid s)$ using standard conjugacy formulas. In this sense, we maintain two views of the joint distribution $p(\theta, s)$—when updating θ, it is the standard exponential family model, which leads to conjugate updates; when updating s, it is approximated as $p(\theta)\mathcal{N}(s; n\mu, s\Sigma)$, which will lead to simple updates when combined with a variable augmentation technique.

4.2.2 Variable augmentation for $p(y \mid s)$

We seek a tractable form for the full conditional of s under the normal approximation, which is the product of a normal density and a Laplace density:

Algorithm 3 Gibbs Sampler, Bounded Δ_s

1: Initialize θ, s, σ^2
2: **repeat**
3: $\theta \sim p(\theta; \lambda')$ where $\lambda' =$ Conjugate-Update(λ, s, n)
4: Calculate $\mu = \mathbb{E}[s]$ and $\Sigma = \mathrm{Var}[s]$ (e.g., use Eq. (4.1))
5: $s \sim \text{NormProduct}\left(n\mu, n\Sigma, y, \mathrm{diag}(\sigma^2)\right)$
6: $1/\sigma_j^2 \sim \text{InverseGaussian}\Big(\frac{\epsilon}{\Delta_s |y - s|}, \frac{\epsilon^2}{\Delta_s^2}\Big)$

Subroutine NormProduct

1: **input:** $\mu_1, \Sigma_1, \mu_2, \Sigma_2$
2: $\Sigma_3 = \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1}$
3: $\mu_3 = \Sigma_3 \left(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2\right)$
4: **return:** $\mathcal{N}(\mu_3, \Sigma_3)$

$$p(s \mid \theta, y) \propto \mathcal{N}(s; n\mu, n\Sigma)\ \mathrm{Lap}(y; s, \Delta_s/\epsilon).$$

A similar situation arises in the Bayesian Lasso [Park & Casella, 2008], and we will employ the same variable augmentation trick. A Laplace random variable $z \sim \mathrm{Lap}(u, b)$ can be written as a scale mixture of normals by introducing a latent variable $\sigma^2 \sim \mathrm{Exp}(1/(2b^2))$, i.e., the distribution with density $1/(2b^2)\exp\left(-\sigma^2/(2b^2)\right)$, and letting $z \sim \mathcal{N}(u, \sigma^2)$. We apply this separately to each dimension of the vector y so that:

$$\sigma_j^2 \sim \mathrm{Exp}\left(\frac{\epsilon^2}{2\Delta_s^2}\right), \quad y \sim \mathcal{N}\big(s, \mathrm{diag}(\sigma^2)\big).$$

4.2.3 The Gibbs sampler

After the normal approximation and variable augmentation, the generative process is as shown in Figure 4.1. The final Gibbs sampling algorithm is shown in Algorithm 3. Note that the update for θ is based on conjugacy in the exact distribution $p(\theta, s)$, while the update for s uses the density of the generative process to the right, so that $p(s \mid \theta, \sigma^2, y) \propto p(s \mid \theta)\, p(y \mid \sigma^2, s)$, which is a product of two normal densities

$$\mathcal{N}(s; n\mu, n\Sigma)\,\mathcal{N}\big(y; s, \mathrm{diag}(\sigma^2)\big) \propto \mathcal{N}(s; \mu_s, \Sigma_s),$$

where μ_s and Σ_s are are defined in Algorithm 3 [Petersen & Pedersen, 2008].

The update for σ^2 follows Park & Casella [2008]; the inverse Gaussian density is $\mathrm{InverseGaussian}(x; m, v) = \sqrt{v/(2\pi x^3)}\exp\left(-v(x-m)^2/(2m^2x)\right)$. Full derivations are given in Section B.2.

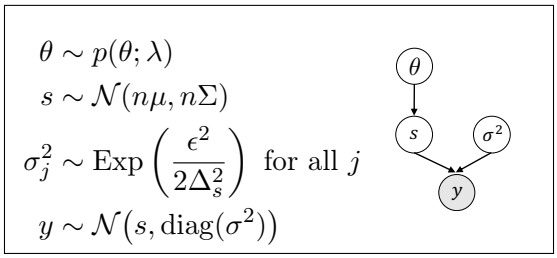


Figure 4.1: Full generative model

4.3 Unbounded Sufficient Statistics and Truncated Exponential Families

The Laplace mechanism does not apply when the sufficient statistics are unbounded, because $\Delta_s = \max_{x,y} \|t(x) - t(y)\|_1 = \infty$. Thus, we need a new release mechanism $\mathcal{A}$ and inference algorithm $\mathcal{P}$. We present a solution for the case when x is univariate. All elements of the solution can generalize to higher dimensions, except that one step will have running time that is exponential in d; we leave improvement of this to future work and focus on the simpler univariate case.

4.3.1 Release mechanism

Our solution is to truncate the support of the (now univariate) $p(x \mid \theta)$ to $x \in [a, b]$, where a and b are finite bounds provided by the modeler. If the modeler cannot select

bounds *a priori*, they may be selected privately as a preliminary step using a variant of the exponential mechanism (see `PrivateQuantile` in Smith [2011a]).[2] Then, given truncation bounds, the data owner redacts individuals where $x_i \notin [a, b]$ and reports the truncated sufficient statistics $\hat{s} = \sum_{i=1}^{n} \mathbf{1}_{[a,b]}(x_i) \cdot t(x_i)$ where $\mathbf{1}_S(x)$ is the indicator function of the set S. The sensitivity of $\hat{s}$ is now $\Delta_{\hat{s}} = \max_{x,y \in \mathbb{R}} \|\hat{t}(x) - \hat{t}(y)\|_1$ where $\hat{t}(x) = \mathbf{1}_{[a,b]}(x)\, t(x)$. An easy upper bound for this quantity is:

$$\Delta_{\hat{s}} \leq \sum_{j=1}^{d} \max \Big\{ \max_{x \in [a,b]} |t_j(x)|, \max_{x,y \in [a,b]} \big|t_j(x) - t_j(y)\big| \Big\},$$

where $t_j(x)$ is the jth component of the sufficient statistics. See Section B.3 for derivation. The bounds $[a, b]$ will be selected so this quantity is bounded. The released value is $y \sim \text{Lap}(\hat{s}, \Delta_{\hat{s}}/\epsilon)$.

4.3.2 Inference: truncated exponential family

Several new challenges arise for inference. The quantity $\hat{s}$ is no longer a sufficient statistic for the model $p(x \mid \theta)$, and we will need new insights to understand $p(\hat{s} \mid \theta)$ and $p(\theta \mid \hat{s})$. Since $\hat{s}$ is a sum over individuals where $x_i \in [a, b]$, it will be useful to examine the probability of the event $x \in [a, b]$ as well as the conditional distribution of x given this event. To facilitate a general development, assume a generic truncation interval $[v, w]$, not necessarily equal to $[a, b]$. Let $F(x; \theta) = \int_{-\infty}^{x} p(x \mid \theta) dx$ be the CDF of the original (univariate) exponential family model. It is clear that $\Pr(x \in [v, w]) = F(w; \theta) - F(v; \theta)$. The conditional distribution of x given $x \in [v, w]$ is a *truncated* exponential family, which, in its natural parameterization is:

[2] Selecting truncation bounds will consume some of the privacy budget and modify the release mechanism $\mathcal{A}$. We do not consider inference with respect to this part of the release mechanism.

$$\hat{p}(x \mid \eta) = \mathbf{1}_{[v,w]}(x)\, h(x)\, \exp\left(\eta^T t(x) - \hat{A}(\eta)\right), \quad \hat{A} = \int_v^w h(x) \exp\left(\eta^T t(x)\right) dx. \tag{4.2}$$

Note that this is still an exponential family model (with a modified base measure), and all of the standard results apply, such as the existence of a conjugate prior and the formulas in Eq. (4.1) for the mean and variance of $t(x)$ under the truncated distribution.

4.3.3 Random sum CLT for $p(\hat{s} \mid \theta)$

We would like to again apply an asymptotic normal approximation for $\hat{s}$, but we do not know how many individuals fall within the truncation bounds. The "random sum CLT" of Robbins [1948] applies to the setting where the number of terms in the sum is itself a random variable. The sum can be rewritten as $\hat{s} = \sum_{k=1}^{N} t(x_{i_k})$, where $\{i_1, \ldots, i_N\}$ is the set of indices of individuals with data inside the truncation bounds, i.e., the indices such that $x_{i_k} \in [v, w]$. The number N is now a random variable distributed as $N \sim \text{Binom}(n, q)$, where $q = F(w; \theta) - F(v; \theta)$.

Proposition 5. *Let $\hat{\mu} = \mathbb{E}_{\hat{p}}[t(x)]$ and $\hat{\Sigma} = \text{Var}_{\hat{p}}[t(x)]$ be the mean and variance of $t(x)$ in the truncated exponential family. Then $\hat{s} = \sum_{k=1}^{N} t(x_{i_k})$ is asymptotically normal with mean and variance:*

$$\mathbf{m} := \mathbb{E}[\hat{s}] = \mathbb{E}[N]\hat{\mu} = nq\hat{\mu},$$
$$\mathbf{V} := \text{Var}(\hat{s}) = \mathbb{E}[N]\hat{\Sigma} + \text{Var}[N]\hat{\mu}\hat{\mu}^T = nq\hat{\Sigma} + nq(1-q)\hat{\mu}\hat{\mu}^T.$$

Specifically, $\frac{1}{\sqrt{n}}\left(\hat{s} - \mathbf{m}\right) \xrightarrow{D} \mathcal{N}(0, \bar{\Sigma})$ as $n \to \infty$, where $\bar{\Sigma} = \mathbf{V}/n = q\hat{\Sigma} + q(1-q)\hat{\mu}\hat{\mu}^T$.

Proof. Each term of the sum has mean $\hat{\mu}$ and variance $\hat{\Sigma}$, and the number of terms is $N \sim \text{Binom}(n, q)$. The result follows from Robbins [1948]. □

4.3.4 Computing $\hat{\mu}$ and $\hat{\Sigma}$ by automatic differentiation (autodiff)

To use the normal approximation we need to compute $\hat{\mu}$ and $\hat{\Sigma}$.

Lemma 1. *Let $p(x \mid \theta)$ be a univariate exponential family model and let $\hat{p}(x \mid \theta)$ be the corresponding exponential family model truncated to generic interval $[v, w]$. Then*

$$\hat{\mu} = \mathbb{E}_{\hat{p}}[t(x)] = \mathbb{E}_p[t(x)] + \frac{\partial}{\partial \eta^T} \log\big(F(w;\eta) - F(v;\eta)\big) \tag{4.3}$$

$$\hat{\Sigma} = \mathrm{Var}_{\hat{p}}[t(x)] = \mathrm{Var}_p[t(x)] + \frac{\partial^2}{\partial \eta \partial \eta^T} \log\big(F(w;\eta) - F(v;\eta)\big) \tag{4.4}$$

Proof. It is straightforward to derive from Eq. (4.2) that $\hat{A}(\eta) = A(\eta) + \log\big(F(w;\eta) - F(v;\eta)\big)$. The result follows from applying Eq. (4.1) to this expression for $\hat{A}(\eta)$. See Section B.1 for derivation of $\hat{A}(\eta)$ and proof of this lemma. □

We will use Equations (4.3) and (4.4) to compute $\hat{\mu}$ and $\hat{\Sigma}$ by using autodiff to compute the desired derivatives. If the mean and variance $\mathbb{E}_p[t(x)]$ and $\mathrm{Var}_p[t(x)]$ of the untruncated distribution are not known, we can apply autodiff to compute them as well using Eq. (4.1).

When x is multivariate, analogous expressions can be derived for $\hat{\mu}$ and $\hat{\Sigma}$. The adjustment factors will include multivariate CDFs, with a number of terms that grow exponentially in d. This is currently the main limitation in applying our methods to multivariate models with unbounded sufficient statistics.

4.3.5 Conjugate updates for $p(\theta \mid \hat{s})$

The final issue is the distribution $p(\theta \mid \hat{s})$, which is no longer characterized by conjugacy because $\hat{s}$ are not the full sufficient statistics. We again turn to variable augmentation. Let $\hat{s}_\ell = \sum_{i=1}^n \mathbf{1}_{[-\infty,a]} t(x_i)$ and $\hat{s}_u = \sum_{i=1}^n \mathbf{1}_{[b,\infty]} t(x_i)$ be the sufficient statistics for the individuals that fall in the lower portion $[-\infty, a]$ and upper portion $[b, \infty]$ of the support of x, respectively. We will instantiate $\hat{s}_\ell$ and $\hat{s}_u$ as latent variables and model their distributions using the random sum CLT approximation from Prop. 5 and Lemma 1 (but with different truncation bounds). Let $\hat{s}_c = \hat{s}$ be the sufficient statistics for the "center" portion, and define the three truncation intervals as

Algorithm 2 Gibbs Sampler, Unbounded Δ_s

Initialize $\theta, \hat{s}, \sigma^2, a, b$
$[v_\ell, w_\ell] \leftarrow [-\infty, a]$
$[v_c, w_c] \leftarrow [a, b]$
$[v_u, w_u] \leftarrow [b, \infty]$
repeat
    $\mathbf{m}_r, \mathbf{V}_r \leftarrow$ RS-CLT(θ, v_r, w_r) for $r \in \{\ell, c, u\}$
    $\mathbf{m}'_c, \mathbf{V}'_c \leftarrow$ NormProduct $\left(\mathbf{m}_c, \mathbf{V}_c, y, \text{diag}\left(\sigma^2\right)\right)$
    $s \sim \mathcal{N}(\mathbf{m}_\ell + \mathbf{m}'_c + \mathbf{m}_u, \mathbf{V}_\ell + \mathbf{V}'_c + \mathbf{V}_u)$
    $\theta \sim p(\theta; \lambda')$ where $\lambda' =$ Conjugate-Update(λ, s, n)
    Recalculate $\mathbf{m}_c$ and $\mathbf{V}_c$, then draw $\hat{s}_c \sim \mathcal{N}(\mathbf{m}_c, \mathbf{V}_c)$
    $1/\sigma_j^2 \sim \text{InverseGaussian}\left(\frac{\epsilon}{\Delta_{\hat{s}}|y - \hat{s}_c|}, \frac{\epsilon^2}{\Delta_{\hat{s}}^2}\right)$
until

Algorithm 3 RS-CLT

input: θ, v, w
$q \leftarrow F(b; w) - F(a; v)$
$\hat{\mu}, \hat{\Sigma} \leftarrow$ autodiff of Eqns. 4.3, 4.4
$\mathbf{m} \leftarrow nq$
$\mathbf{V} \leftarrow nq\hat{\Sigma} + nq(1-q)\hat{\mu}\hat{\mu}^T$
return: $\mathbf{m}, \mathbf{V}$

$$[v_\ell, w_\ell] = [-\infty, a] \tag{4.5}$$

$$[v_c, w_c] = [a, b] \tag{4.6}$$

$$[v_u, w_u] = [b, \infty]. \tag{4.7}$$

The full sufficient statistics are equal to $s = \hat{s}_\ell + \hat{s}_c + \hat{s}_u$. Conditioned on all other variables, *each* component is multivariate normal, so the sum s is also multivariate normal. We can therefore sample s and then sample from $p(\theta \mid s)$ using conjugacy. We will also need to draw $\hat{s}_c$ separately to be used to update σ^2.

4.3.5.1 The Gibbs sampler

The (approximate) generative process in the unbounded case is:

$$
\begin{aligned}
\theta &\sim p(\theta; \lambda), \\
\hat{s}_r &\sim \mathcal{N}\big(\mathbf{m}_r,\ \mathbf{V}_r\big),\ \text{for } r \in \{\ell, c, u\} \text{ where } \mathbf{m}_r, \mathbf{V}_r = \text{RS-CLT}(\theta, v_r, w_r) \\
\sigma_j^2 &\sim \text{Exp}\left(\frac{\epsilon^2}{2\Delta_{\hat{s}}^2}\right) \text{ for all } j\ , \\
y &\sim \mathcal{N}\big(\hat{s}_c, \text{diag}(\sigma^2)\big).
\end{aligned}
$$

The Gibbs sampler to sample from this distribution is given in Algorithm 2. Note that in Line 8 we employ rejection sampling in which sufficient statistics are sampled until the values drawn are valid for the given data model, e.g., s must be positive for the binomial distribution. The RS-CLT algorithm to compute parameters of the random sum CLT is shown in Algorithm 3.

4.4 Experiments

We design experiments to measure the *calibration* and *utility* of our method for posterior inference. We conduct experiments for the binomial model with beta prior, the multinomial model with Dirichlet prior, and the exponential model with gamma prior. The last model is unbounded and requires truncation; we set the bounds to keep the middle 95% of individuals, which is reasonable to assume known a priori for some cases, such as modeling human height.

4.4.1 Methods

We run our Gibbs sampler for 5000 iterations after 2000 burnin iterations (see Section B.5 for convergence results), which we compare to two baselines. The first method uses the same release mechanism as our Gibbs sampler and performs conjugate updates using the noisy sufficient statistics [Foulds et al., 2016; Zhang et al., 2016]. This method converges to the true posterior as $n \to \infty$ because the Laplace noise will eventually become negligible compared to sampling variability [Foulds et al., 2016]. However, the noise is not negligible for moderate n; we refer to this method as "naive".

For truncated models we allow the naive method to "cheat" by accessing the noisy *untruncated* sufficient statistics s. Thus the method is not private, and receives strictly more information than our Gibbs sampler, but with the same magnitude noise. This allows us to demonstrate miscalibration without highly technical modifications to the baseline method to be able to deal with truncated sufficient statistics.

The second baseline is a version of the one-posterior sampling (OPS) mechanism [Foulds et al., 2016; Wang et al., 2015; Zhang et al., 2016], which employs the exponential mechanism [McSherry & Talwar, 2007] to release samples from a privatized posterior. We release 100 samples using the method of [Foulds et al., 2016], each with $\epsilon_{ops} = \epsilon/100$, such that the entire algorithm achieves ϵ-differential privacy. Private MCMC sampling [Wang et al., 2015] is a more sophisticated method to release multiple samples from a privatized posterior and could potentially make better use of the privacy budget; however, private MCMC will also necessarily be miscalibrated, and only achieves the weaker privacy guarantee of (ϵ, δ)-differential privacy for $\delta > 0$, so would not be directly comparable to our method. OPS serves as a suitable baseline that achieves ϵ-differential privacy. We include OPS only for experiments on the binomial model, for which it requires the support of θ to be truncated to $[a_0, 1 - a_0]$ where $a_0 > 0$. We set $a_0 = 0.1$.

We also include a non-private posterior for comparison, which performs conjugate updates using the non-noisy sufficient statistics.

4.4.2 Evaluation

We evaluate both the *calibration* and *utility* of the posterior. For calibration we adapt a method of Cook et al. [2006]: the idea is to draw iid samples (θ_i, x_i) from the joint model $p(\theta)p(x \mid \theta)$, and conduct posterior inference in each trial. Let $F_i(\theta)$ be the CDF of the *true* posterior $p(\theta \mid x_i)$ in trial i. Then we know that $U_i = F_i(\theta_i)$ is uniformly distributed, because $\theta_i \sim p(\theta \mid x_i)$ (see Section B.4). In other words, the

actual parameter θ_i is equally likely to land at any quantile of the posterior. To test the posterior inference procedure, we instead compute U_i as the quantile at which θ_i lands within a set of samples from the *approximate* posterior. After M trials of the whole procedure we test for uniformity of $U_{1:M}$ using the Kolmogorov-Smirnov goodness-of-fit test [Massey Jr., 1951], which measures the maximum distance between the empirical CDF of $U_{1:M}$ and the uniform CDF; lower values are better and zero corresponds to perfect uniformity. We also visualize the empirical CDFs to assess calibration qualitatively.

Higher utility of a private posterior is indicated by closeness to the non-private posterior, which we measure with *maximum mean discrepancy* (MMD), a kernel-based statistical test to determine if two sets of samples are drawn from different distributions [Gretton et al., 2012]. Given m i.i.d. samples $(p, q) \sim P \times Q$, an unbiased estimate of the MMD is

$$\mathrm{MMD}^2(P, Q) = \frac{1}{m(m-1)} \sum_{i \neq j}^{m} \left(k(p_i, p_j) + k(q_i, q_j) - k(p_i, q_j) - k(p_j, q_i) \right),$$

where k is a continuous kernel function; we use a standard normal kernel. The higher the value the more likely the two samples are drawn from different distributions.

4.4.3 Results

Figure 4.2 shows the results for three models and varying n and ϵ. Our method (Gibbs) achieves the same calibration level as non-private posterior inference for all settings. The naive method ignores noise and is too confident about parameter values implied by treating the noisy sufficient statistics as true ones; it is only well-calibrated with increasing n and ϵ when noise becomes negligible relative to population size. OPS is not calibrated because it samples from an over-dispersed version of $p(\theta \mid x)$.

Figure 4.3 shows the empirical CDF plots for $n = 1000$ and $\epsilon = 0.01$. Our method and the non-private method are both perfectly calibrated. The naive method's over-

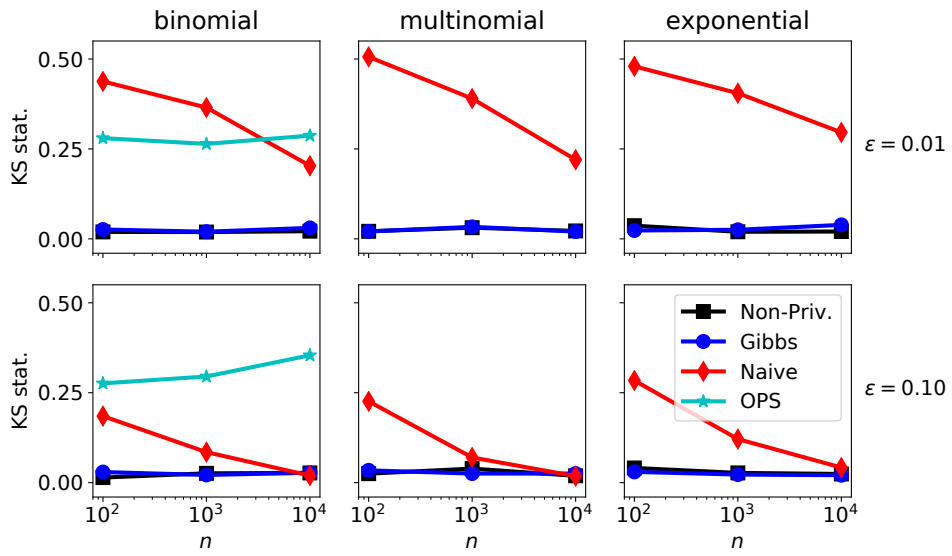


Figure 4.2: Calibration as Kolmogorov-Smirnov statistic vs. number of individuals at $\epsilon = [0.01, 0.10]$ for binomial, multinomial, and exponential models.

confidence in the wrong sufficient statistics causes its posterior to usually be too tight at the wrong value; thus the true parameter always lies in a tail of the approximate posterior, so too much mass is placed near 0 and 1. OPS shows the opposite behavior: its posterior is always too diffuse, so the true parameter lies close to the middle. For multinomial we show measures only for the parameter of the first category, but results hold for all categories.

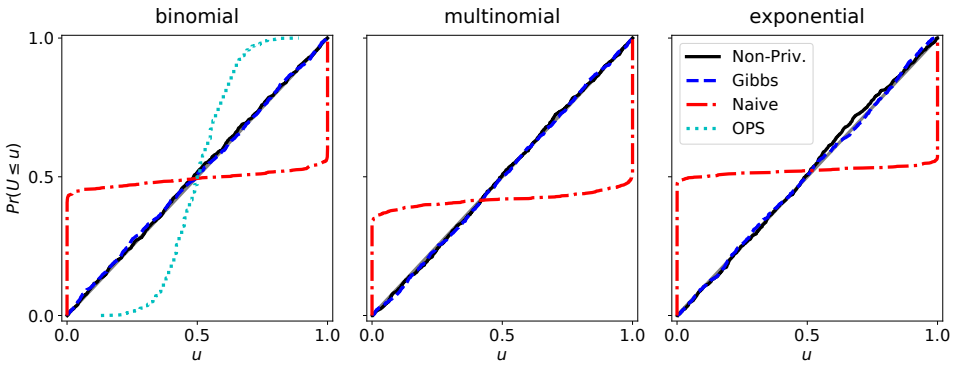


Figure 4.3: Empirical CDF plots at $(n = 1000; \epsilon = 0.01)$ for binomial, multinomial, and exponential models.

Figure 4.4 shows the MMD test statistic between each method and the non-private posterior, used as a measure of utility. Our method consistently achieves utility at least as good as the naive method for binomial and multinomial models. We omit OPS, which is never calibrated. For the exponential model (not shown) we did not obtain conclusive utility comparisons due to the lack of a naive baseline that properly handles truncation; the "cheating" naive method from our calibration experiments sometimes attains higher utility than our method, and sometimes lower, but this comparison is not meaningful because it receives strictly more information.

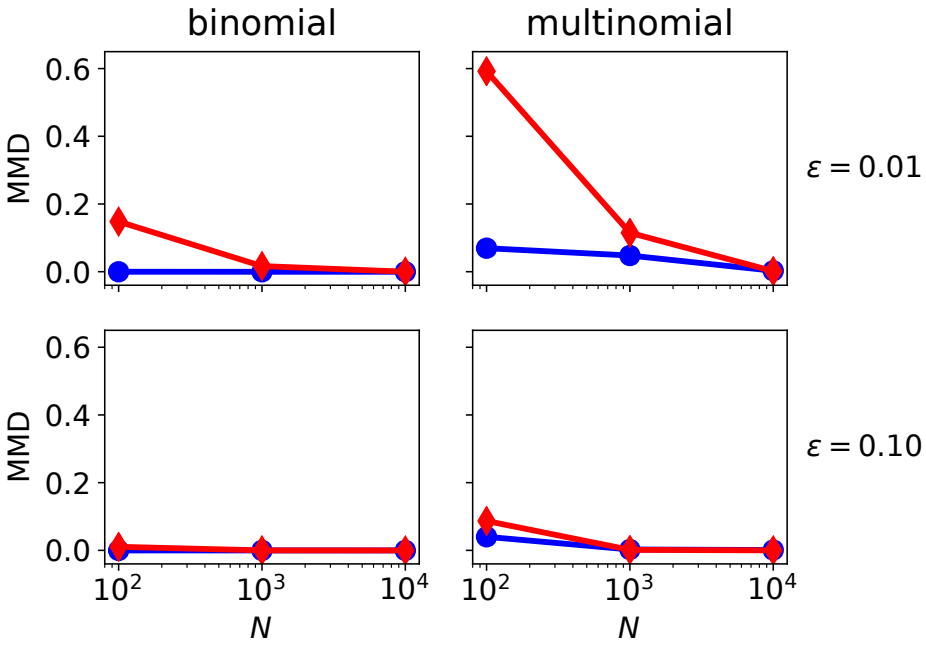


Figure 4.4: Utility as MMD with non-private posterior vs. number of individuals at $\epsilon = [0.01, 0.10]$ for binomial and multinomial models.

CHAPTER 5

PRIVATE BAYESIAN LINEAR REGRESSION

5.1 Introduction

Linear regression is one of the most widely used statistical methods, especially in the social sciences [Agresti & Finlay, 2009] and other domains where data comes from humans. It is important to develop robust tools that can realize the benefits of regression analyses but maintain the privacy of individuals. Existing work on differentially private linear regression focuses on frequentist approaches. A variety of privacy mechanisms have been applied to point estimation of regression coefficients, including sufficient statistic perturbation (SSP) [Foulds et al., 2016; Vu & Slavkovic, 2009; Wang, 2018; Zhang et al., 2016], posterior sampling (OPS) [Dimitrakakis et al., 2014; Geumlek et al., 2017; Minami et al., 2016; Wang, 2018; Wang et al., 2015; Zhang et al., 2016], subsample and aggregate [Dwork & Smith, 2010; Smith, 2008], objective perturbation [Kifer et al., 2012], and noisy stochastic gradient descent [Bassily et al., 2014]. Only a few recent works address uncertainty quantification through confidence interval estimation [Sheffet, 2017] and hypothesis tests [Barrientos et al., 2019] for regression coefficients.

We develop a differentially private method for *Bayesian* linear regression. A Bayesian approach naturally quantifies parameter uncertainty through a full posterior distribution and provides other Bayesian capabilities such as the ability to incorporate prior knowledge and compute posterior predictive distributions. Existing approaches to private Bayesian inference include OPS (see above), MCMC [Wang et al., 2015], and SSP [Bernstein & Sheldon, 2018; Foulds et al., 2016], but none provide a fully

satisfactory approach for Bayesian regression modeling. OPS does not naturally produce a representation of a full posterior distribution. MCMC approaches incur per-iteration privacy costs and satisfy only approximate (ϵ, δ)-differential privacy. SSP is more promising, since perturbed sufficient statistics can be used in conjugate updates to obtain parameters of full posterior distributions. However, Chapter 4 demonstrated (for unconditional exponential family models) that naive SSP, which ignores noise introduced by the privacy mechanism, systematically underestimates uncertainty at small to moderate sample sizes. We show that the same phenomenon holds for Bayesian linear regression: naive SSP produces private posteriors that are properly calibrated asymptotically in the sample size, but for realistic data sets and privacy levels may need very large population sizes to reach the asymptotic regime.

This motivates our development of Bayesian inference methods for linear regression that properly account for the noise due to the privacy mechanism [Bernstein & Sheldon, 2018; Bernstein et al., 2017; Karwa et al., 2014, 2016; Schein et al., 2018; Williams & McSherry, 2010]. We leverage a model in which the data and model parameters are latent variables, and noisy sufficient statistics are observed, and then develop MCMC-based techniques to sample from posterior distributions, as done for exponential families in [Bernstein & Sheldon, 2018]. A significant challenge relative to prior work is the handling of covariate data. Typical regression modeling treats only response variables and parameters as random, and conditions on covariates. This is not possible in the private setting, where covariates must be kept private and therefore treated as latent variables. We therefore require some form of assumption about the distribution over covariates. We develop two inference methods. The first includes latent variables for each individual; it requires an explicit prior distribution for covariates and its runtime scales with population size. The second marginalizes out individuals and approximates the distribution over the sufficient statistics; it requires weaker assumptions about the covariate distribution (only moments), and its

running time does not scale with population size. We perform a range of experiments to measure the calibration and utility of these methods. Our noise-aware methods are as well or nearly as well calibrated as the non-private method, and have better utility than the naive method. We demonstrate using real data that our noise-aware methods quantify posterior predictive uncertainty significantly better than naive SSP. We then conclude with a case study drawn from the real problem of state-wide budget allocation using sensitive census data predictions.

We start with a standard (non-private) linear regression problem. An individual's *covariate* or *regressor* data is $\mathbf{x} \in \mathbb{R}^d$ and the dependent *response* data is $y \in \mathbb{R}$. We will assume a conditionally Gaussian model $y \sim \mathcal{N}(\boldsymbol{\theta}^T \mathbf{x}, \sigma^2)$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ are the regression coefficients and σ^2 is the error variance. An intercept or bias term may be included in the model by appending a unit-valued feature to $\mathbf{x}$. The goal, given an observed population of n individuals, is to obtain a point estimate of $\boldsymbol{\theta}$. The population data can be written as $X \in \mathbb{R}^{n \times d}$, where each row corresponds to an individual $\mathbf{x}$, and $\mathbf{y} \in \mathbb{R}^n$. The ordinary least squares (OLS) solution is $\hat{\boldsymbol{\theta}} = \left(X^T X\right)^{-1} X^T \mathbf{y}$ [Rencher, 2003].

$$
\begin{aligned}
p(\boldsymbol{\theta}, \sigma^2 \mid X, \mathbf{y}) &= \mathrm{NIG}(\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, a_n, b_n) \\
\boldsymbol{\mu}_n &= \left(X^T X + \boldsymbol{\Lambda}_0\right)^{-1} \left(X^T \mathbf{y} + \boldsymbol{\mu}_0^T \boldsymbol{\Lambda}_0\right) \\
\boldsymbol{\Lambda}_n &= X^T X + \boldsymbol{\Lambda}_0 \\
a_n &= a_0 + \frac{1}{2} n \\
b_n &= b_0 + \frac{1}{2} \left(\mathbf{y}^T \mathbf{y} + \boldsymbol{\mu}_0^T \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 - \boldsymbol{\mu}_n^T \boldsymbol{\Lambda}_n \boldsymbol{\mu}_n\right)
\end{aligned}
\tag{5.1}
$$

In Bayesian linear regression the parameters $\boldsymbol{\theta}$ and σ^2 are random variables with a specified prior distribution. The conjugate priors are $p(\sigma^2) = \mathrm{InverseGamma}(a_0, b_0)$ and $p(\boldsymbol{\theta} \mid \sigma^2) = \mathcal{N}(\boldsymbol{\mu}_0, \sigma^2 \boldsymbol{\Lambda}_0^{-1})$, which defines a normal-inverse gamma prior distri-

bution: $p(\boldsymbol{\theta}, \sigma^2) = \text{NIG}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, a_0, b_0)$. Due to conjugacy of the prior distribution with the likelihood model, the posterior distribution, shown in Equation (5.1), is also normal-inverse gamma [O'Hagan & Forster, 1994].

Let $t(\mathbf{x}, y) := [\text{vec}(\mathbf{x}\mathbf{x}^T),\ \mathbf{x}y,\ y^2]$ for an arbitrary individual. Then the sufficient statistics of the above model are $\mathbf{s} := t(X, \mathbf{y}) = \sum_i t\big(\mathbf{x}^{(i)}, y^{(i)}\big) = \big[X^T X,\ X^T\mathbf{y},\ \mathbf{y}^T\mathbf{y}\big]$. These capture all information about the model parameters contained in the sample and are the only quantities needed for the conjugate posterior updates above [Casella & Berger, 2002].

5.2 Private Bayesian Linear Regression

The goal is to perform Bayesian linear regression in an ϵ-differentially private manner. We ensure privacy by employing sufficient statistic perturbation (SSP) [Foulds et al., 2016; Vu & Slavkovic, 2009; Zhang et al., 2016], in which the Laplace mechanism is used to inject noise into the sufficient statistics of the model, making them fit for public release. The question is then how to compute the posterior over the model parameters $\boldsymbol{\theta}$ and σ^2 given the noisy sufficient statistics. We first consider a *naive* method that ignores the noise in the noisy sufficient statistics. We then consider more principled *noise-aware* inference approaches that account for the noise due to the privacy mechanism.

5.2.1 Privacy mechanism

Using the Laplace mechanism to release the noisy sufficient statistics $\mathbf{z}$ results in the model shown in Figure 5.1. This is the same model used in non-private linear regression except for the introduction of $\mathbf{z}$, which requires the exact sufficient statistics $\mathbf{s}$ to have finite sensitivity. A standard assumption in literature [Awan & Slavkovic, 2018; Sheffet, 2017; Wang, 2018; Zhang et al., 2012] is to assume $\mathbf{x}$ and y have known a priori lower and upper bounds, $(a_\mathbf{x}, b_\mathbf{x})$ and (a_y, b_y), with bound widths $w_\mathbf{x} = b_\mathbf{x} - a_\mathbf{x}$

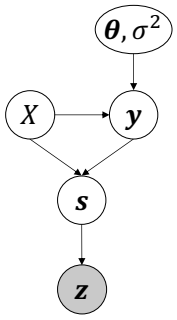


Figure 5.1: Private regression model.

(assuming, for simplicity, equal bounds for all covariate dimensions) and $w_y = b_y - a_y$, respectively. We can then reason about the worst case influence of an individual on each component of $\mathbf{s} = \left[X^T X, X^T \mathbf{y}, \mathbf{y}^T \mathbf{y}\right]$, recalling that $\mathbf{s} = \sum_i t(\mathbf{x}^{(i)}, y^{(i)})$, so that $\left[\Delta_{(X^T X)_{jk}}, \Delta_{(Xy)_j}, \Delta_{y^2}\right] = \left[w_{\mathbf{x}}^2, w_{\mathbf{x}} w_y, w_y^2\right]$. The number of unique elements[1] in $\mathbf{s}$ is $[d(d+1)/2, d, 1]$, so $\Delta_{\mathbf{s}} = w_{\mathbf{x}}^2 d(d+1)/2 + w_{\mathbf{x}} w_y d + w_y^2$. The noisy sufficient statistics fit for public release are $\mathbf{z} = \left[z_i \sim \text{Lap}(s_i, \Delta_{\mathbf{s}}/\epsilon) : s_i \in \mathbf{s}\right]$.

5.2.2 Noise-naive method

Previous work developed methods to obtain OLS solutions via SSP by ignoring the noise injected into the sufficient statistics [Awan & Slavkovic, 2018; Sheffet, 2017; Wang, 2018]. One corresponding approach for Bayesian regression is to naively replace $\mathbf{s}$ in Figure 5.1 with the noisy version $\mathbf{z}$ and then perform the conjugate update in Equation (5.1). This noise-naive method (`Naive`) is simple and fast, and we empirically show in Section 4.4 that it produces an asymptotically correct posterior.

5.2.3 Noise-aware inference

Instead of ignoring the noise introduced by the privacy mechanism, we propose to perform inference over the noise in the model in Figure 5.1 in order to produce correct posteriors regardless of the data size. The biggest change from

[1]Note that $X^T X$ is symmetric.

non-private to private Bayesian linear regression is that due to privacy constraints we can no longer condition on the covariate data X. The non-private posterior is $p(\boldsymbol{\theta}, \sigma^2|X, \mathbf{y}) \propto p(\boldsymbol{\theta}, \sigma^2)\, p(\mathbf{y}|X, \boldsymbol{\theta}, \sigma^2)$ while the private posterior is $p(\boldsymbol{\theta}, \sigma^2|\mathbf{z}) \propto \int p(X)p(\boldsymbol{\theta}, \sigma^2)\, p(\mathbf{y}|X, \boldsymbol{\theta}, \sigma^2)\, p(\mathbf{z}|X, \mathbf{y})\, dX\, d\mathbf{y}$ (see derivations in supplementary material). The private posterior contains the term $p(X)$, which means that in order to calculate it *we need to know something about the distribution of* X!

Given an explicitly specified prior $p(X)$, we can perform inference over the model in Figure 5.1 using general-purpose MCMC algorithms. We use the No-U-Turn Sampler [Hoffman & Gelman, 2014] from the PyMC3 package [Salvatier et al., 2016], and call this method *noise-aware individual-based inference* (`MCMC-Ind`). This approach is simple to implement using existing tools but places a substantial burden on the modeler relative to the non-private case by requiring an explicit prior distribution $p(X)$, with poor choices potentially leading to incorrect inferences. Additionally, because `MCMC-Ind` instantiates latent variables for each individual, its runtime scales with population size and it may be slow for large populations.

5.2.4 Sufficient statistics-based inference

An appealing possibility is to marginalize out the variables X and $\mathbf{y}$ representing individuals and instead perform inference directly over the latent sufficient statistics $\mathbf{s}$. The joint distribution is $p(\boldsymbol{\theta}, \sigma^2, \mathbf{s}, \mathbf{z}) = p(\boldsymbol{\theta}, \sigma^2)\, p(\mathbf{s} \mid \boldsymbol{\theta}, \sigma^2)\, p(\mathbf{z} \mid \mathbf{s})$. The goal is to compute a representation of $p(\boldsymbol{\theta}, \sigma^2 \mid \mathbf{z}) \propto \int_{\mathbf{s}} p(\boldsymbol{\theta}, \sigma^2, \mathbf{s}, \mathbf{z})\, d\mathbf{s}$ by integrating over the sufficient statistics. Because this distribution cannot be written in closed form we develop a Gibbs sampler to sample from the posterior as done by Bernstein & Sheldon [2018] for unconditional exponential family models. This requires methods to sample from the conditional distributions for both the parameters $(\boldsymbol{\theta}, \sigma^2)$ and the sufficient statistics $\mathbf{s}$ given all other variables. The full conditional $p(\boldsymbol{\theta}, \sigma^2 \mid \mathbf{s})$ for the model parameters can be computed and sampled using conjugacy, exactly as in the non-

private case. The full conditional for $\mathbf{s}$ factors into two terms: $p(\mathbf{s} \mid \boldsymbol{\theta}, \sigma^2, \mathbf{z}) \propto p(\mathbf{s} \mid \boldsymbol{\theta}, \sigma^2)\, p(\mathbf{z} \mid \mathbf{s})$. The first is the distribution over sufficient statistics of the regression model, for which we develop an asymptotically correct normal approximation. The second is the noise model due to the privacy mechanism, for which we use variable augmentation to ensure it is possible to sample from the full conditional distribution of $\mathbf{s}$.

5.2.4.1 Normal approximation of s

The conditional distribution over the sufficient statistics given the model parameters is

$$p(\mathbf{s} \mid \boldsymbol{\theta}, \sigma^2) = \int_{t^{-1}(\mathbf{s})} p\left(X, \mathbf{y} \mid \boldsymbol{\theta}, \sigma^2\right)\, dX\, d\mathbf{y}, \qquad t^{-1}(\mathbf{s}) := \left\{X, \mathbf{y} : t(X, \mathbf{y}) = \mathbf{s}\right\}.$$

The integral over $t^{-1}(\mathbf{s})$, all possible populations which have sufficient statistics $\mathbf{s}$, is intractable to compute. Instead we observe that the components of $\mathbf{s} = \sum_i t(\mathbf{x}^{(i)}, y^{(i)})$ are sums over individuals. Therefore, using the central limit theorem (CLT), we approximate their distribution as $p(\mathbf{s} \mid \boldsymbol{\theta}, \sigma^2) \approx \mathcal{N}(\mathbf{s}; n\boldsymbol{\mu}_t, n\Sigma_t)$, where $\boldsymbol{\mu}_t = \mathbb{E}[t(\mathbf{x}, y)]$ and $\Sigma_t = \operatorname{Cov}(t(\mathbf{x}, y))$ are the mean and covariance of the function $t(\mathbf{x}, y)$ on a single individual, This approximation is asymptotically correct, i.e., $\frac{1}{\sqrt{n}}(\mathbf{s} - n\boldsymbol{\mu}_t) \xrightarrow{D} \mathcal{N}(0, \Sigma_t)$ [Bickel & Doksum, 2015]. We write the conditional distribution as

$$\begin{aligned}
&\mathbf{s} \mid \cdot \sim \mathcal{N}(n\boldsymbol{\mu}_t, n\Sigma_t), \\
&\boldsymbol{\mu}_t = \left[\mathbb{E}\left[\operatorname{vec}(\mathbf{x}\mathbf{x}^T)\right], \mathbb{E}\left[\mathbf{x}y\right], \mathbb{E}\left[y^2\right]\right], \qquad (5.2) \\
&\Sigma_t = \begin{bmatrix} \operatorname{Cov}\left(\operatorname{vec}(\mathbf{x}\mathbf{x}^T)\right) & \operatorname{Cov}\left(\operatorname{vec}(\mathbf{x}\mathbf{x}^T), \mathbf{x}^T y\right) & \operatorname{Cov}\left(\operatorname{vec}(\mathbf{x}\mathbf{x}^T), y^2\right) \\ \operatorname{Cov}\left(\mathbf{x}y, \operatorname{vec}(\mathbf{x}\mathbf{x}^T)\right) & \operatorname{Cov}\left(\mathbf{x}y\right) & \operatorname{Cov}\left(\mathbf{x}y, y^2\right) \\ \operatorname{Cov}\left(y^2, \operatorname{vec}(\mathbf{x}\mathbf{x}^T)\right) & \operatorname{Cov}\left(y^2, \mathbf{x}y\right) & \operatorname{Var}\left(y^2\right) \end{bmatrix}. \qquad (5.3)
\end{aligned}$$

The components of $\boldsymbol{\mu}_t$ and Σ_t can be written in terms of the model parameters $(\boldsymbol{\theta}, \sigma^2)$ and the second and fourth non-central moments of $\mathbf{x}$ as shown below, where we have defined $\eta_{ij} := \mathbb{E}[x_i x_j]$, $\eta_{ijkl} := \mathbb{E}[x_i x_j x_k x_l]$, and $\xi_{ij,kl} := \mathrm{Cov}(x_i x_j, x_k x_l) = \eta_{ijkl} - \eta_{ij}\eta_{kl}$. Full derivations can be found in the supplementary material. We call this family of methods `Gibbs-SS`. To use this normal distribution for sampling, we need the parameters $(\boldsymbol{\theta}, \sigma^2)$ and the moments η_{ij}, η_{ijkl}, and $\xi_{ij,kl}$. The current parameter values are available within the sampler, but the modeler must provide estimates for the moments of $\mathbf{x}$, either using prior knowledge or by (privately) estimating the moments from the data. We discuss three specific possibilities in Section 5.2.4.4.

$$
\begin{aligned}
\mathbb{E}[x_i y] &= \sum_j \theta_j \eta_{ij} \\
\mathbb{E}\left[y^2\right] &= \sigma^2 + \sum_{i,j} \theta_i \theta_j \eta_{ij} \\
\mathrm{Cov}(x_i x_j, x_k y) &= \sum_l \theta_l \xi_{ij,kl} \\
\mathrm{Cov}\left(x_i x_j, y^2\right) &= \sum_{k,l} \theta_k \theta_l \xi_{ij,kl} \\
\mathrm{Cov}(x_i y, x_j y) &= \sigma^2 \eta_{ij} + \sum_{k,l} \theta_k \theta_l \xi_{ij,kl} \\
\mathrm{Cov}\left(x_i y, y^2\right) &= \sum_{j,k,l} \theta_j \theta_k \theta_l \xi_{ij,kl} + 2\sigma^2 \sum_j \theta_j \eta_{ij} \\
\mathrm{Var}\left(y^2\right) &= 2\sigma^4 + \sum_{i,j,k,l} \theta_i \theta_j \theta_k \theta_l \xi_{ij,kl} \\
&\quad + 4\sigma^2 \sum_{i,j} \theta_i \theta_j \eta_{ij}
\end{aligned}
$$

Once again, more modeling assumptions are needed than in the non-private case, where it is possible to condition on $\mathbf{x}$. `Gibbs-SS` requires milder assumptions (second and fourth moments), however, than `MCMC-Ind` (a full prior distribution).

$$
\begin{aligned}
\boldsymbol{\theta}, \sigma^2 &\sim \mathrm{NIG}(\boldsymbol{\theta}, \sigma^2; \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, a_0, b_0) \\
\mathbf{s} &\sim \mathcal{N}(n\boldsymbol{\mu}_t, n\Sigma_t) \\
\omega_j^2 &\sim \mathrm{Exp}\left(\frac{\epsilon^2}{2\Delta_{\mathbf{s}}^2}\right) \text{ for all } j \\
\mathbf{z} &\sim \mathcal{N}\big(\mathbf{s}, \mathrm{diag}(\omega^2)\big)
\end{aligned}
$$

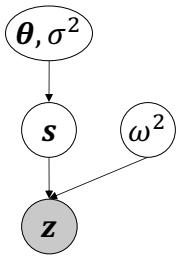


Figure 5.2: Full generative model.

5.2.4.2 Variable augmentation for $p(\mathbf{z} \mid \mathbf{s})$

The above approximation for the distribution over sufficient statistics means the full conditional distribution involves the product of a normal and a Laplace distribution,

$$p(\mathbf{s} \mid \theta, \mathbf{z}) \propto \mathcal{N}(\mathbf{s}; n\boldsymbol{\mu}_t, n\Sigma_t) \cdot \mathrm{Lap}(\mathbf{z}; \mathbf{s}, \Delta_{\mathbf{s}}/\epsilon).$$

It is unclear how to sample from this distribution directly. A similar situation arises in the Bayesian Lasso, where it is solved by variable augmentation [Park & Casella, 2008]. Bernstein & Sheldon [2018] adapted the variable augmentation scheme to private inference in exponential family models. We take the same approach here, and represent a Laplace random variable as a scale mixture of normals. Specifically, $l \sim \mathrm{Lap}(u, b)$ is identically distributed to $l \sim \mathcal{N}(u, \omega^2)$ where the variance $\omega^2 \sim \mathrm{Exp}\left(1/(2b^2)\right)$ is drawn from the exponential distribution (with density $1/(2b^2)\exp\left(-\omega^2/(2b^2)\right)$). We augment separately for each component of the vector $\mathbf{z}$ so that $\mathbf{z} \sim \mathcal{N}\big(\mathbf{s}, \mathrm{diag}(\omega^2)\big)$, where $\omega_j^2 \sim \mathrm{Exp}\left(\epsilon^2/(2\Delta_{\mathbf{s}}^2)\right)$. The augmented full conditional $p(\mathbf{s} \mid \theta, \mathbf{z}, \omega)$ is a product of two multivariate normal distributions, which is itself a multivariate normal distribution.

5.2.4.3 The Gibbs sampler

The full generative process is shown in Figure 5.2, and the corresponding Gibbs sampler is shown in Algorithm 7. The update for ω^2 follows Park & Casella [2008]; the inverse Gaussian density is $\mathrm{InverseGaussian}(w; m, v) = \sqrt{v/(2\pi w^3)}\exp\left(-v(w-m)^2/(2m^2 w)\right)$.

Algorithm 7 Gibbs Sampler

Initialize $\boldsymbol{\theta}, \sigma^2, \omega^2$
repeat
  Calculate $\boldsymbol{\mu}_t$ and Σ_t via Eqs. (5.2) and (5.3)
  $\mathbf{s} \sim \text{NormProduct}\left(n\boldsymbol{\mu}_t, n\Sigma_t, \mathbf{z}, \text{diag}(\omega^2)\right)$
  $\boldsymbol{\theta}, \sigma^2 \sim \text{NIG}(\boldsymbol{\theta}, \sigma^2; \boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n, a_n, b_n)$ via Eqn. (5.1)
  $1/\omega_j^2 \sim \text{InverseGaussian}\left(\frac{\epsilon}{\Delta_{\mathbf{s}}|\mathbf{z}-\mathbf{s}|}, \frac{\epsilon^2}{\Delta_{\mathbf{s}}^2}\right)$ for all j

Subroutine NormProduct

input: $\boldsymbol{\mu}_1, \Sigma_1, \boldsymbol{\mu}_2, \Sigma_2$
$\Sigma_3 = \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1}$
$\boldsymbol{\mu}_3 = \Sigma_3\left(\Sigma_1^{-1}\boldsymbol{\mu}_1 + \Sigma_2^{-1}\boldsymbol{\mu}_2\right)$
return: $\mathcal{N}(\boldsymbol{\mu}_3, \Sigma_3)$

Note that the resulting $\mathbf{s}$ drawn from $p(\mathbf{s} \mid \boldsymbol{\mu}_t, \Sigma_t, \omega^2)$ may require projection onto the space of valid sufficient statistics. This can be done by observing that if $A = [X, \mathbf{y}]$ then the sufficient statistics are contained in the positive-semidefinite (PSD) matrix $B = A^T A$. For a randomly drawn $\mathbf{s}$, we project if necessary so the corresponding B matrix is PSD.

5.2.4.4 Distribution over X

As discussed above, `Gibbs-SS` requires the second and fourth population moments of $\mathbf{x}$ to calculate $\boldsymbol{\mu}_t$ and Σ_t. We propose three different options for the modeler to provide these and discuss the algorithmic considerations for each. Because we include the unit feature in $\mathbf{x}$ we can restrict our attention to the fourth moment $\mathbb{E}\left[\mathbf{x}^{\otimes 4}\right]$, which includes the second moment as a subcomponent.

- Private sample moments (`Gibbs-SS-Noisy`)

  The first option is to estimate population moments privately by computing the fourth sample moments from X and privately releasing them via the Laplace mechanism. The sensitivity of the estimate for η_{ijkl} is w_x^4, and for $d = 2$

there are $D = 5$ unique entries, for a total sensitivity of Dw_x^4. This approach requires splitting the privacy budget between the release mechanisms for sufficient statistics and moments, which we do evenly. We do not perform inference over the noisy sample moments, which may introduce some miscalibration of uncertainty. Pursuing this additional layer of inference is an interesting avenue for future work.

- Moments from generic prior (`Gibbs-SS-Prior`)

  A second option is to propose a prior distribution $p(\mathbf{x})$ and obtain population moments directly from the prior, either through known formulas or from Monte Carlo estimation. This approach does not access the individual data and does not consume any privacy budget, but requires proposing a prior distribution and computing the fourth moments of $\mathbf{x}$ (once) for that distribution.

- Hierarchical normal prior (`Gibbs-SS-Update`)

  A final option is to perform inference over the data moments by specifying an individual-level prior $p(\mathbf{x})$ and then marginalizing away individuals, as we did for the regression model sufficient statistics. We propose a hierarchical normal prior, as shown in Figure 5.3a, which is more dispersed than a normal distribution and allows the modeler to propose vague priors, but still permits attainable conditional updates. The data $\mathbf{x}$ is normally distributed: $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \tau^2)$, with parameters drawn from the normal-inverse Wishart (NIW) conjugate prior distribution, $\boldsymbol{\mu}_x, \tau^2 \sim \text{NIW}(\boldsymbol{\mu}_0', \Lambda_0', \Psi_0', \nu_0')$. After marginalizing individuals, the latent quantities are the sufficient statistics XX^T (which includes the sample mean and covariance because of the unit feature). For fixed parameters $(\boldsymbol{\mu}_x, \tau^2)$ the distribution $p(\mathbf{x})$ is multivariate normal, and we calculate its fourth moments as the fourth derivative (via automatic differentiation) of its moment generating function.

However, we introduced the new latent variables $\boldsymbol{\mu}_x$ and τ^2 into the full model (see Figure 5.3a) and must now derive conditional updates for them within the Gibbs sampler. Naively marginalizing X and $\mathbf{y}$ from the full model in Figure 5.3a would cause both $(\boldsymbol{\mu}_x, \tau^2)$ and $(\boldsymbol{\theta}, \sigma^2)$ to be parents of $\mathbf{s}$ and thus *not* conditionally independent given $\mathbf{s}$—this would require their updates to be coupled and we could no longer use simple conjugacy formulas for each component of the model. To avoid this issue, we reformulate the joint distribution represented as in Figure 5.3b. The justification for this is as follows. Because X^TX is a sufficient statistic for $p(X)$ under a normal model, we can encode the generative process *either* as $(\boldsymbol{\mu}_x, \tau^2) \to X \to X^TX$ *or* as $(\boldsymbol{\mu}_x, \tau^2) \to X^TX \to X$. In general, the latter formulation would require an arrow from $(\boldsymbol{\mu}_x, \tau^2)$ to X; this drops precisely because X^TX is a sufficient statistic [Casella & Berger, 2002]. Then, upon marginalizing X and $\mathbf{y}$, we obtain the model in Figure 5.3c. The two sets of parameters are now conditionally independent given the sufficient statistics $\mathbf{s}$, and can be updated independently as standard conjugate updates.

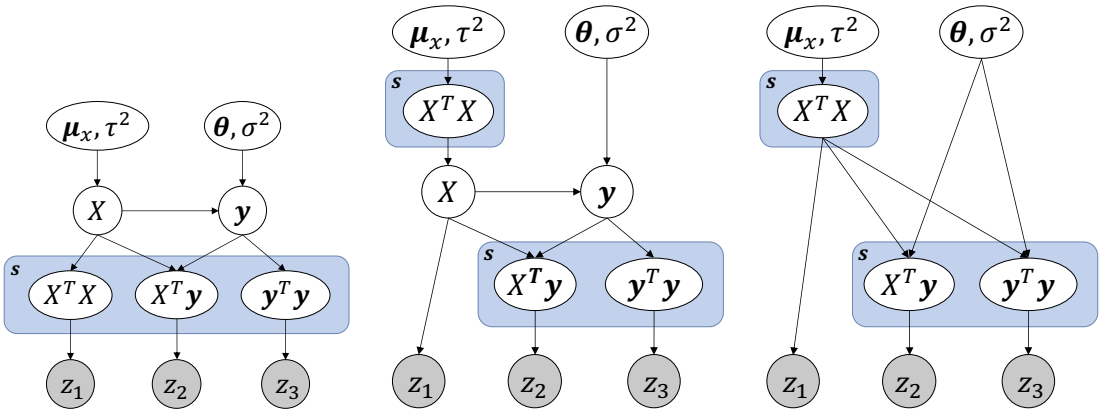


Figure 5.3: (a) Private Bayesian linear regression model with hierarchical normal data prior. (b) Alternative data model configuration and (c) with individual variables marginalized out.

5.3 Experiments

We design experiments to measure the *calibration* and *utility* of the private methods. Calibration measures how close the computed posterior is to $p(\boldsymbol{\theta}, \sigma^2|\mathbf{z})$, the correct posterior given noisy statistics. *Utility* measures how close the computed posterior is to the non-private posterior $p(\boldsymbol{\theta}, \sigma^2|\mathbf{s})$.

5.3.1 Methods

The noise-aware individual-based method (`MCMC-Ind`) is implemented using PyMC3 [Salvatier et al., 2016]; it runs with 500 burnin iterations and collects 2000 posterior samples. The three flavors of noise-aware sufficient statistic-based methods use noisy sample moments (`Gibbs-SS-Noisy`), use moments sampled from a data prior (`Gibbs-SS-Prior`), and use an updated hierarchical normal prior (`Gibbs-SS-Update`); all three collect 20000 posterior samples after 5000 and 20000 burnin iterations for $n \in [10, 100]$ and $n = 1000$, respectively. We compare against the baseline noise-naive method (`Naive`) and the non-private posterior (`Non-Private`); both collect 2000 posterior samples.

5.3.2 Evaluation on synthetic data

Evaluation measures. We adapt a method of Cook et al. [2006] to measure calibration. Consider a model $p(\boldsymbol{\beta}, \mathbf{w}) = p(\boldsymbol{\beta})p(\mathbf{w}|\boldsymbol{\beta})$. If $(\boldsymbol{\beta}', \mathbf{w}') \sim p(\boldsymbol{\beta}, \mathbf{w})$, then, for any j, the quantile of β'_j in the true posterior $p(\beta_j|\mathbf{w}')$ is a uniform random variable. We can check our approximate posterior $\hat{p}$ by computing the quantile u_j of β'_j in $\hat{p}(\beta_j|\mathbf{w}')$ and testing for uniformity of u_j over M trials. We test for uniformity using the Kolmogorov-Smirnov (KS) goodness-of-fit test [Massey Jr., 1951]. The KS-statistic is the maximum distance between the empirical CDF of u_j and the uniform CDF; lower values are better and zero corresponds to perfect uniformity, meaning $\hat{p}$ is exact.

While this test is elegant, it requires that parameters and data are drawn from the model used by the method. We use $\boldsymbol{\theta}, \sigma^2 \sim \text{NIG}\left([0,0], \text{diag}\left(\left[\frac{.5}{20-1}, \frac{.5}{20-1}\right]\right), 20, .5\right)$. In addition, for `Gibbs-SS-Prior` and `Gibbs-SS-Update`, the test requires the covariate

data be drawn from the data prior used by the methods. We specify $\boldsymbol{\mu}_x, \tau^2 \sim \text{NIW}(0, 1, 1, 50)$ and $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \tau^2)$. These ensure at least 95% of $\mathbf{x}$ and y values are within $[-1, 1]$. We compute sensitivity assuming data bounded in this range but do not actually truncate data outside the bounds in order to avoid changing the generative process (a limitation of the evaluation method, not the inference routine).

For each combination of n and ϵ we run $M = 300$ trials. We qualitatively assess calibration with the empirical CDFs, which is also the *quantile-quantile* (QQ) plot between the empirical distribution of u_j and the uniform distribution. A diagonal line indicates thats u_j is perfectly uniform.

Between two calibrated posteriors, the tighter posterior will provide higher utility.[2] We evaluate utility as *closeness to the non-private posterior*, which we measure with *maximum mean discrepancy* (MMD), a kernel-based statistical test to determine if two sets of samples are drawn from different distributions [Gretton et al., 2012]. Given m i.i.d. samples $(p, q) \sim P \times Q$, an unbiased estimate of the MMD is

$$\text{MMD}^2(P, Q) = \frac{1}{m(m-1)} \sum_{i \neq j}^{m} \left(k(p_i, p_j) + k(q_i, q_j) - k(p_i, q_j) - k(p_j, q_i) \right),$$

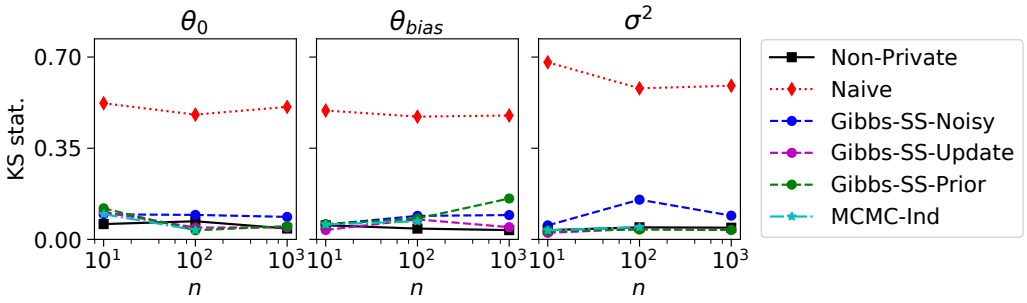
where k is a continuous kernel function; we use a standard normal kernel. The higher the value the more likely the two samples are drawn from different distributions, therefore lower MMD between `Non-Private` and the method indicates higher utility.

We measure method runtime as the average process time over the 300 trials. Note that PyMC3 provides parallelization; we report total process time across all chains for `MCMC-Ind`.

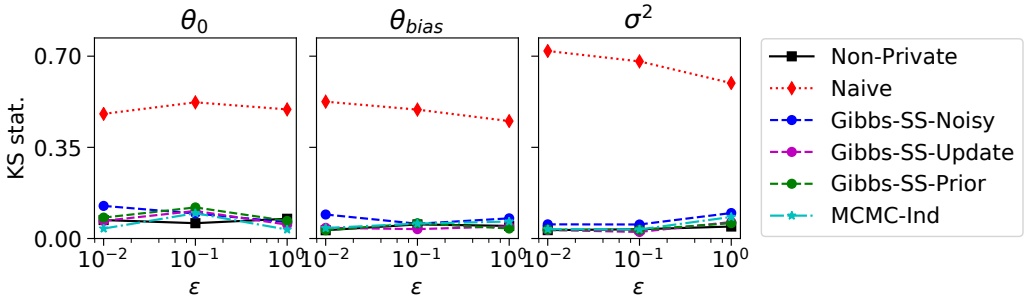
Results. Calibration results are shown in Figure 5.4. The QQ plot for $n = 10$ and $\epsilon = 0.1$ is shown in Figure 5.5. Coverage results for 95% credible intervals are shown in Figure 5.6. All four noise-aware methods are at or near the calibration-level

[2]Note that the prior itself is a calibrated distribution.

of the non-private method, and better than `Naive`'s calibration, regardless of data size. As expected, `Gibbs-SS-Noisy` suffers slight miscalibration from not accounting for the noise injected into the privately released fourth data moment. There is slight miscalibration in certain settings and parameters for `Gibbs-SS-Prior` due to approximations in the calculation of multivariate normal distribution fourth moments from a data prior. Utility results are shown in Figure 5.7; the noise-aware methods provide at least as good utility as `Naive`. Run time results are shown in Figure 5.8; `MCMC-Ind` scales with increasing population size while the `Gibbs-SS` methods, `Naive`, and `Non-Private` remain constant. Accordingly, we do not include results for `MCMC-Ind` for $n = 1000$ as its run time is prohibitive in those settings.



(a)



(b)

Figure 5.4: Calibration vs. n (for $\epsilon = 0.1$) and vs. ϵ (for $n = 10$).

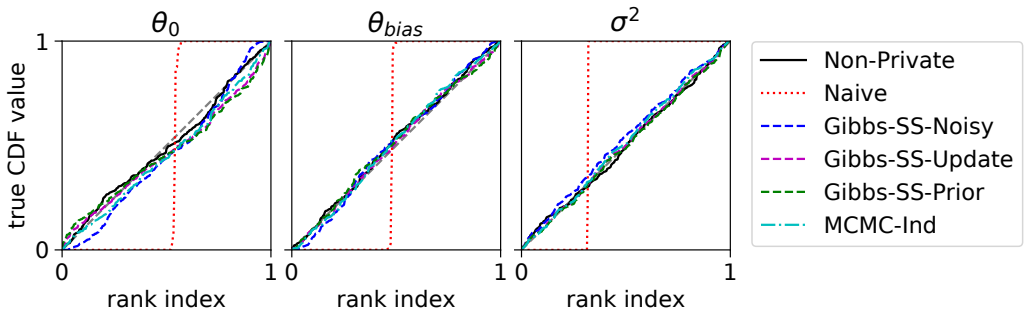


Figure 5.5: QQ plot for $n = 10$ and $\epsilon = 0.1$.

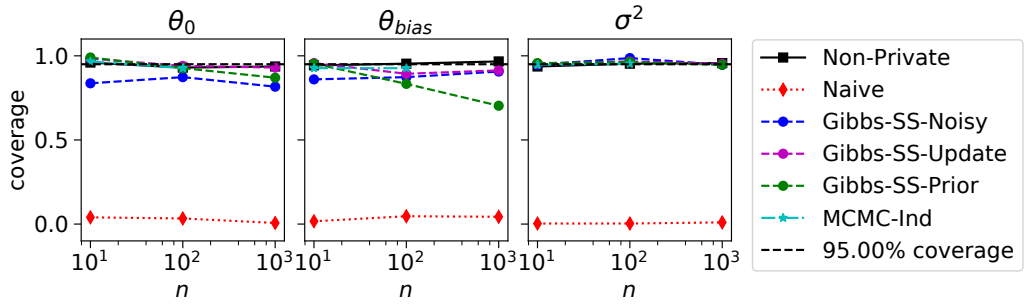


Figure 5.6: 95% credible interval coverage.

5.3.3 Predictive posteriors on real data

We evaluate the predictive posteriors of the methods on a real world data set measuring the effect of drinking rate on cirrhosis rate.[3] We scale both covariate and response data to $[0, 1]$. There are 46 total points, which we randomly split into 36 training examples and 10 test points for each trial. After preliminary exploration to gain domain knowledge, we set a reasonable model prior of $\boldsymbol{\theta}, \sigma^2 \sim \text{NIG}\big([1, 0], \text{diag}([.25, .25]), 20, .5\big)$. We draw samples $\boldsymbol{\theta}^{(k)}, \sigma_k^2$ from the posterior given training data, and then form the posterior predictive distribution for each test point y_i from these samples. Figure 5.9 shows coverage of 50% and 90% credible intervals on 1000 test points collected over 100 random train-test splits. `Non-Private` achieves nearly correct coverage, with the discrepancy due to the fact that the data is not actually drawn from the

[3] `http://people.sc.fsu.edu/~jburkardt/datasets/regression/x20.txt`

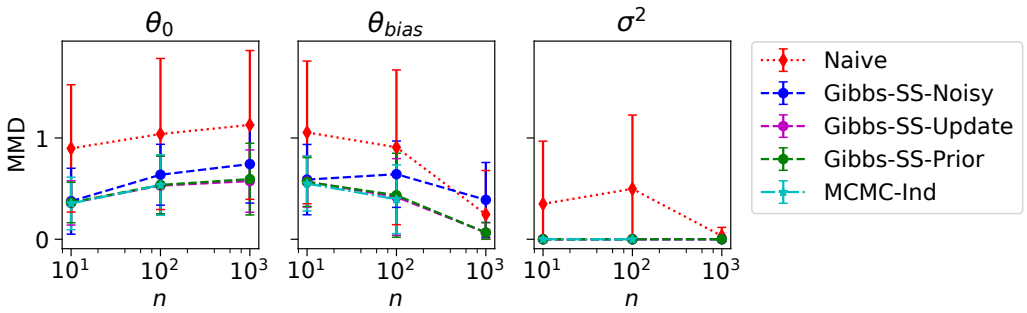


Figure 5.7: Utility as MMD to non-private posterior.

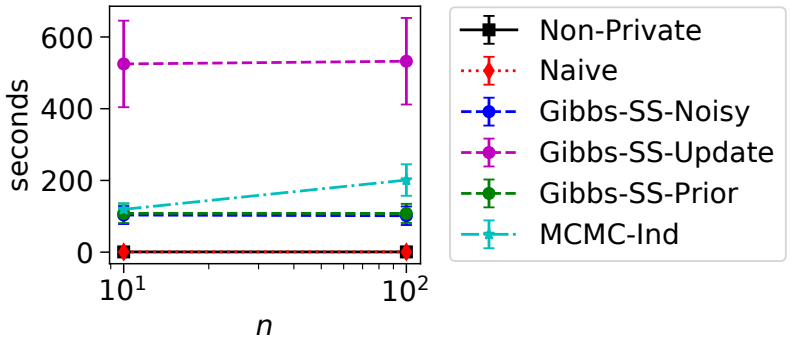


Figure 5.8: Method runtimes for $\epsilon = 0.1$.

prior. `Gibbs-SS-Noisy` achieves nearly the coverage of `Non-Private`, while `Naive` is drastically worse in this regime. We note that this experiment emphasizes the advantage of `Gibbs-SS-Noisy` not needing an explicitly defined data prior, as it only requires the same parameter prior that is needed in non-private analysis.

5.4 Social Mobility Case Study

In this section we conduct a case study in order to explore the application of our regression methods to a real world problem. We hope this will serve as useful documentation for future researchers attempting to transition privacy-based algorithms from synthetic to real problem settings.

We follow the work done by Opportunity Insights as part of their Opportunity Atlas project to analyze census data in a private fashion in order to aid policy-

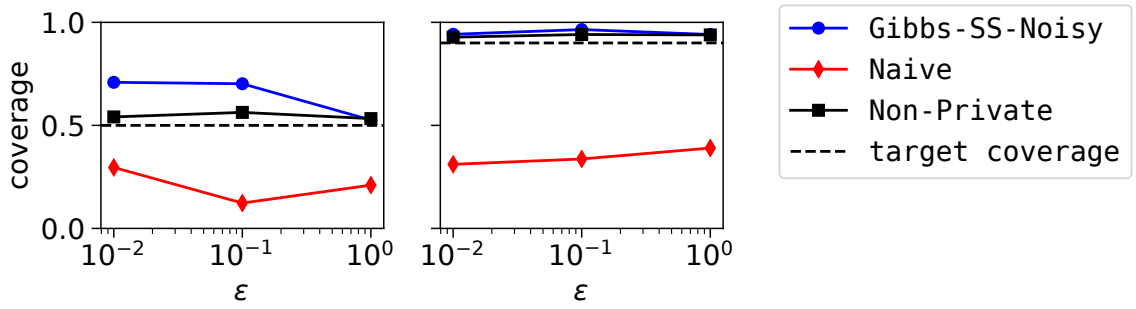


Figure 5.9: Coverage for predictive posterior 50% and 90% credible intervals.

making in regards to social mobility [Chetty & Friedman, 2019]. Research has shown that the census tract (neighborhood) in which a child grows up has a large effect on the adulthood outcomes such as college attendance, income, and incarceration rates [Peter Bergman, 2019]. Identifying "high opportunity" vs. "low opportunity" tracts is an important step in social mobility research and policy intervention, such as in programs to allow low-income families to fully benefit from housing voucher programs [Peter Bergman, 2019].

Opportunity Insights identifies the opportunity level of a tract in the following (simplified) manner. From census and tax record data, individual adults are grouped into the tract in which they grew up. [4] Their adulthood income percentile rank (`kir`) is then paired with the individuals' family income level during their childhood (`pir`). Example scatter plots of this data are shown for two tracts in Figure 5.10. Despite both tracts having a wide spread of parental income, children growing up in the tract on the right have much higher adulthood income. To quantify this effect for use in policy making, a regression line on the data is formed and the value of `kir` is used at `pir` $= 0.25$ [5]. Tracts can then be ranked by these `kir` values. To ensure individual

[4]The general assumption is made that the opportunity level of a tract does not significantly change over time, although the potential for this is accounted for in Opportunity Insight's analyses.

[5]`pir` $= 0.25$ is denoted to be a low-income percentile

privacy, Opportunity Insights wishes to develop private regression methods such that the private rankings are as close as possible to the non-private rankings, with higher emphasis placed on maintaining rankings in the top and bottom tenth percentiles.

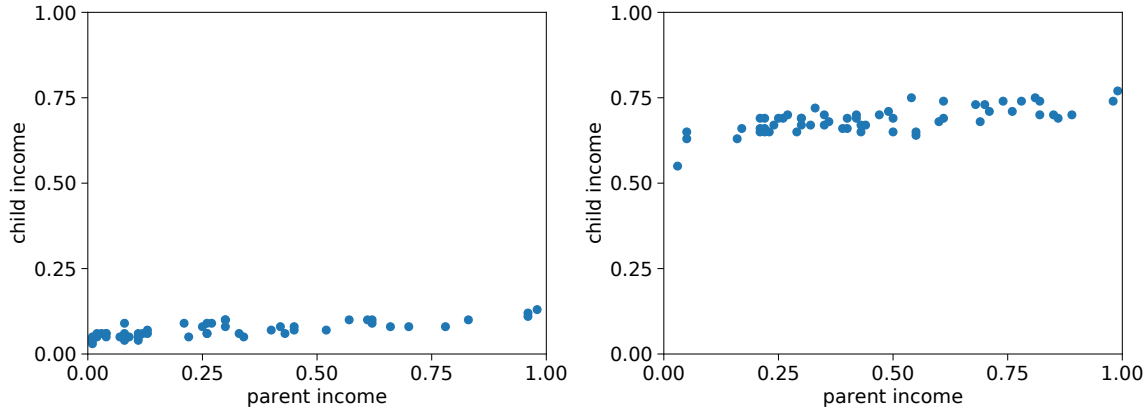


Figure 5.10: Scatter plots of adulthood income rank vs. family income rank as a child for two tracts.

We first run experiments to examine the effectiveness of our methods to quantify uncertainty in estimating the value of `kir` at `pir` $= 0.25$. We also study point estimates extracted from the posteriors, along with point estimates from the method due to Chetty & Friedman [2019] (`OI`), which uses output perturbation. We note that `OI` is not fully differentially private, goes to extreme measures to bound sensitivity, and produces a limited number of point estimates, whereas we seek a general purpose method that can make predictions at all `pir`.

5.4.1 Data

We note the data is in fact synthetic, but stress that it was simulated by Opportunity Insights to match the real data with high fidelity and allow for external analyses on protected internal data. We focus on the data for a single state, Illinois. There are 3108 tracts (neighborhoods) with a long-tailed distribution of size from 20 to 446 individuals. Both `pir` and `kir` are pre-scaled to the range $[0, 1]$. In line with current

best practice in regards to this data set, we throw out 363 tracts in which the 50th percentile of the tract's `kir` is below 0.1 or above 0.9.

5.4.2 Methods

All of the included Bayesian linear regression methods require a model prior. After preliminary exploration to gain domain knowledge, we set a model prior that reasonably matches the general trend of tract data: $\boldsymbol{\theta}, \sigma^2 \sim \text{NIG}\big([.1, .5], \text{diag}([.001, .001]), 1000, .5\big)$. `Non-Private`, `Naive`, and `Gibbs-SS-Noisy` need no further setup. `Gibbs-SS-Update` requires a data prior, which we specify to be centered and generally weak as $\boldsymbol{\mu}_x, \tau^2 \sim \text{NIW}(.5, 1, 1, 50)$. The `Gibbs-SS` methods collect 20000 posterior samples after 5000 burnin iterations; `Non-Private` and `Naive` generate 2000 samples. Given the model parameter posterior samples we then draw samples of `kir` at `pir` $= 0.25$ to form the predictive posterior. We take the predictive posterior mean as the method's point estimate.

We also include the method due to Chetty & Friedman [2019], denoted as `OI`, which uses output perturbation: it calculates the ordinary least squares (OLS) coefficients via non-private regression of a tract and then releases a single noisy point estimate of the predicted `kir` at `pir` $= 0.25$. The method tightly bounds local sensitivity of the released coefficient by using state-wide statistics.

5.4.3 Experiments

We run experiments to investigate the uncertainty quantification of the private Bayesian methods in addition to the point estimation in relation to the use case of ranking tracts. Because this is "real" data we do not have ground truth, therefore we compare the private method point estimates to `Non-Private`. After an analysis following Abowd & Schmutte [2019], the analyses in Chetty & Friedman [2019] are performed at $\epsilon = 4$, which we center around in these experiments. Figure 5.11 shows 90% credible intervals at `pir` $= 0.25$ overlaid on a scatter plot of the tract data for

three tracts of varying sizes. Figure 5.12 shows a scatter plot of the `Gibbs-SS-Update` 2745 tract point estimates vs. the `Non-Private` point estimate, overlaid on 90% and 50% credible intervals. Figure 5.13 shows coverage for private 90% and 50% credible intervals with respect to the `Non-Private` point estimate. The next experiments all focus solely on private point estimate error with respect to the `Non-Private` point estimate. Figure 5.14 shows the average mean absolute error vs. ϵ, and Figure 5.15 shows a scatter of residuals vs. tract size. Figure 5.16 shows a confusion matrix depicting the use case of state-wide budgeting based on ranking of tract `kir` point estimates at `pir` = .25; we turn this problem into a classification task by labeling tracts with their decile membership.

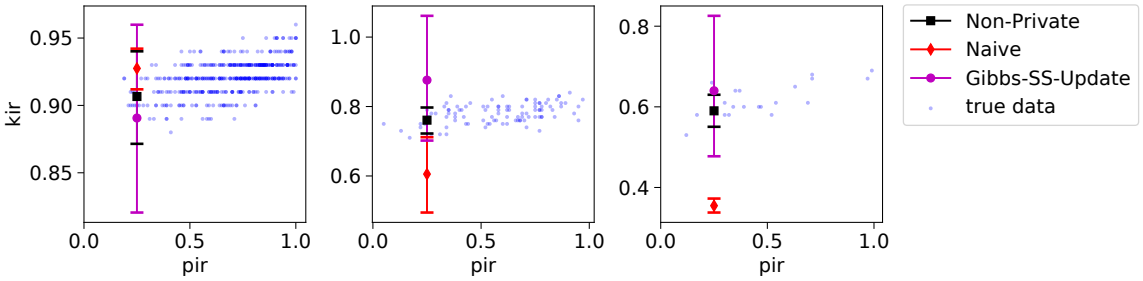


Figure 5.11: Predictive posterior 90% credible intervals and point estimate at `pir` = .25 for $\epsilon = 4$ overlaid on scatter plot of `kir` vs. `pir` for county-tract combos 167-2000, 201-3705, and 31-816100.

5.4.4 Discussion

Here we discuss the experimental results, as well as considerations and lessons learned when applying our methods to a real world problem. We first focus on the credible interval results. As expected, in Figure 5.11 the `Non-Private` 90% credible intervals nicely encapsulate the full height of the true scatter data while not being overly loose, while the `Naive` credible interval is drastically biased and overly tight so that very little if any of the scatter data lies within the range. The `Gibbs-SS-Update` credible intervals are significantly looser than those of `Non-Private` and are less biased

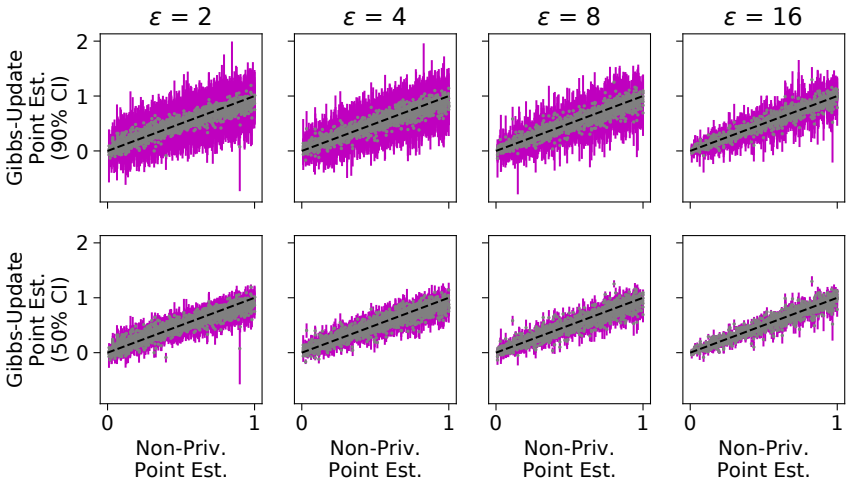


Figure 5.12: `Gibbs-SS-Update` predictive point estimate and credible intervals vs. `Non-Private` predictive point estimate.

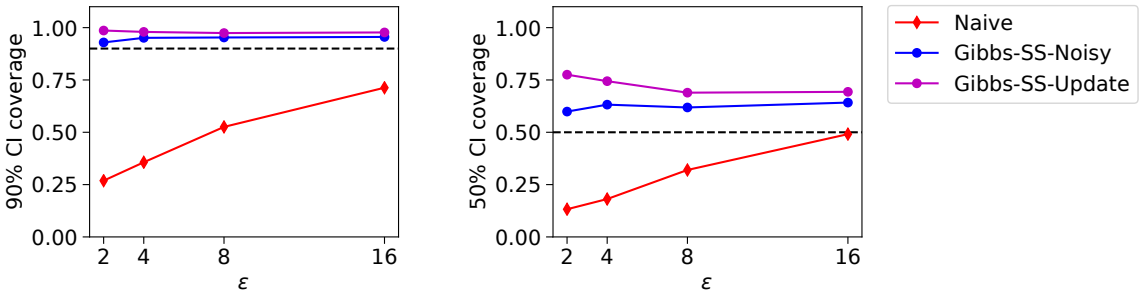


Figure 5.13: Predictive posterior credible interval coverage for `Non-Private` mean vs. ϵ.

than those of `Naive`, and thus while they are not exactly centered on the data they still encapsulate the full height of the scatter data. This trend can be seen at the state-wide scale in Figure 5.12 where the `Gibbs-SS-Update` credible intervals generally follow the diagonal line and stay in step with the `Non-Private` point estimate. As expected they become tighter with increasing ϵ. We quantitatively summarize the credible interval performance with coverage analysis with respect to the `Non-Private` point estimate. Figure 5.13 shows that the `Gibbs-SS-Update` and `Gibbs-SS-Noisy` coverage are reasonably close to the target coverage rate regardless of ϵ, whereas the `Naive` coverage only improves with higher ϵ. We note that the model parameters

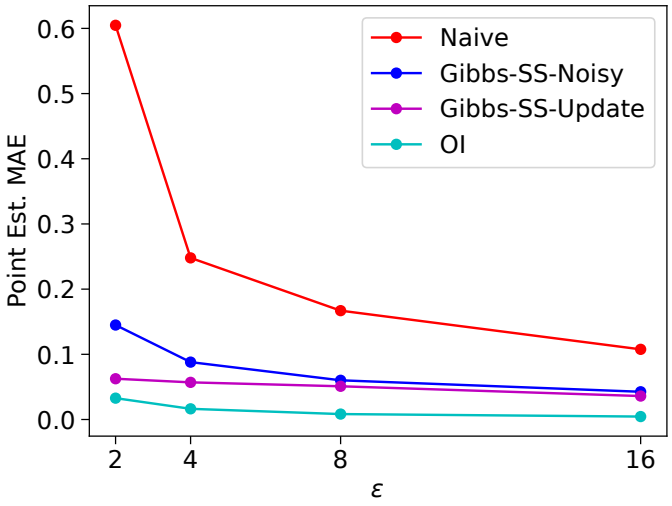


Figure 5.14: Predictive point estimate mean absolute error against `Non-Private` point estimate.

and data are not necessarily drawn from our specified prior, and the `Non-Private` point estimate is not necessarily the truth, so we do not expect the private methods' coverages to exactly achieve the specified level.

Visual inspection of the `Non-Private` credible intervals encapsulating the data qualitatively confirms the model prior is reasonably chosen. It is then clear that, as in synthetic experiments, ignoring the noise due to the privacy mechanism drastically hurts uncertainty quantification at stricter privacy levels. Due to the nature of these experiments without ground truth, we cannot definitively say that performing inference over the noise mechanism delivers perfect uncertainty quantification, but we can confidently say the noise-aware methods perform much better.

The second aspect to examine is the methods' point estimation capabilities, which is more in line with the existing regression work and the analyses done in Chetty & Friedman [2019]. We can qualitatively examine the point estimate scatter plot in Figure 5.12 and see the points are clustered on the diagonal, becoming tighter as ϵ increases, which is quantitatively confirmed in Figure 5.14. It is interesting to note that while we would not necessarily expect the noise-aware `Gibbs-SS` methods to have

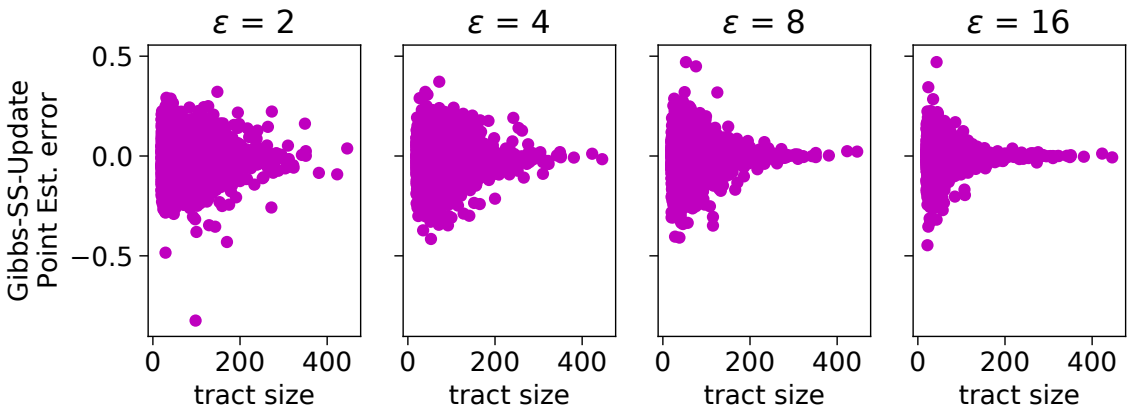


Figure 5.15: `Gibbs-SS-Update` predictive point estimate residual against `Non-Private` point estimate vs. tract size.

better point estimates than `Naive`, they do in fact perform significantly better. One conjecture is that noise-naive regression leads to biased estimates, whereas noise-aware regression better handles perturbed sufficient statistics that do not satisfy model requirements Figure 5.15 shows the same decreasing point estimation error vs. sample size trend as in Figure 3.4; interestingly the rate changes favorably as ϵ increases. The final use case is to produce a ranking of tracts, to be used in budget allocation. The goal is for the private methods to produce a ranking as close to the non-private ranking as possible, with emphasis placed on the top and bottom deciles. Figure 5.16 shows the confusion matrix for `Gibbs-SS-Update`, which is diagonal heavy, as desired. Even more promising is that the top and bottom deciles are almost fully contained with the two top and bottom deciles, respectively, indicating relatively minimal loss due to privacy in the ranking problem of interest.

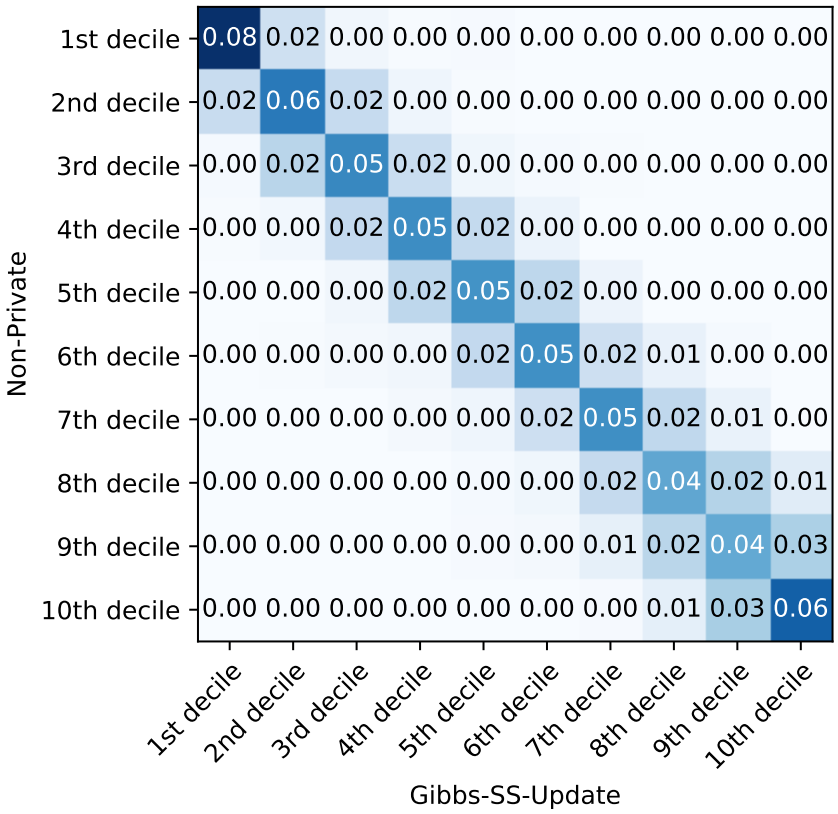


Figure 5.16: Confusion matrix depicting the tract point estimate decile for `Gibbs-SS-Update` vs. `Non-Private` at $\epsilon = 4$.

CHAPTER 6

CONCLUSION AND FUTURE DIRECTIONS

This dissertation coalesces and expands the field of private noise-aware inference. Differential privacy, the dominant privacy framework over the past decade and a half, allows for the collection and analysis of sensitive human data while at the same time balancing the need to protect individual privacy. There has been significant work on mechanisms and methods to release privatized statistics in this setting. The majority of these methods, however, are noise-naive and do not account for the randomization required to deliver the privacy guarantee and thus produce lower quality analyses that may not be useful in practical data settings. Noise-aware inference, first introduced by Williams & McSherry [2010], addresses this problem by performing inference over both the data and noise models in a principled fashion. This approach has been shown to produce higher quality and more practical results in comparison to noise-naive methods. We develop several methods as contributions to the field of private noise-aware inference. We specifically use sufficient statistics perturbation (SSP), allowing us to leverage properties of sufficient statistics in exponential family models in order to achieve tractable and effective approximations.

6.1 Review of contributions

Chapter 3 focuses on point estimation in undirected graphical models, for which we show the existing noise-naive SSP method is asymptotically consistent at the same rate as the non-private method, but that ignoring the noise due to the release mechanism requires ad hoc data patching and sensitive tuning of regularization parameters. We

then develop a noise-aware expectation-maximization (EM) algorithm based on the collective graphical models (CGM; Sheldon & Dietterich [2011]) framework to account for the noise and show that it produces higher quality point estimates than competing methods. We also find similar results in a case-study on learning human mobility patterns based on wifi device traces.

Chapter 4 focuses on Bayesian inference in unconditional exponential family models. We show the existing noise-naive methods produce the correct private posterior only in asymptotically large data settings. We then develop a Gibbs sampler-based method that is able to produce the correct private posterior regardless of data regime. The main technical contribution is to develop a tractable conditional distribution over sufficient statistics, which are a sum over iid individuals, by leveraging the central limit theorem (CLT) approximation. This further requires a model augmentation technique to enable closed-form sampling from the product of the subsequent sufficient statistics and noise distributions.

Chapter 5 focuses on Bayesian linear regression. There is significant existing work for frequentist linear regression, but we are the first to show results for a fully Bayesian noise-naive method that produces a publicly-available posterior. That posterior is only asymptotically correct, however, which motivates the development of noise-aware methods. This requires overcoming non-trivial technical hurdles, the largest being that while the non-private regression problem can condition on the individual covariate and response data, private regression cannot. This means that in order to do inference over the noise mechanism we need to introduce *some* assumptions about the covariate data. We first introduce an MCMC-based sampling method that, while it produces the correct posterior, requires an explicit data prior and instantiates individuals so that its runtime scales with the population size. We then extend the Gibbs sampling framework from the previous chapter to the regression setting. This allows for more efficient inference over the sufficient statistics as well as more flexibility in making assumptions

about the data. We develop three flavors of the Gibbs sampler that obtain assumptions from different sources, each with their own advantages and drawbacks. The produced posteriors are as or nearly as correct as the non-private method regardless of data regime. We conclude with a case study on the problem of social mobility with real world data from an economic non-profit organization.

6.2 Future directions

We see a number of potential directions for future work.

6.2.1 Model selection

In order to perform sufficient statistics perturbation, the choice of model must be in hand so that we know what sufficient statistics need to be released. The work in this thesis assumes we have already chosen a model. But if one were to use these methods in a real world setting, one would have to go through a model selection process. This would presumably require some initial exploration of the sensitive data, which would in and of itself use some of the privacy budget. What approaches would choose the best model while still effectively using the privacy budget?

6.2.2 Point estimation

Chapters 4 and 5 focus on Bayesian inference, but social scientists tend to focus first on point estimation for use in real world problems. While we briefly touched upon point estimation for the social mobility case study in Section 5.4, a more-in depth study is needed to assess which algorithms are most practical in terms of point estimates for unconditional exponential family models and for linear regression. Point estimates can be obtained from a posterior, and there is also potential to develop noise-aware MLE methods (as in Chapter 3).

- **Noise-aware SSP vs. noise-naive SSP** The effectiveness of noise-aware point estimates must be more thoroughly compared to the corresponding noise-naive methods. All noise-aware methods developed in this work can employ the corresponding noise-naive method as an initialization, i.e. using the noisy sufficient statistics for MLE and conjugate updates in Chapter 3 and in Chapters 4 and 5, respectively. With each subsequent iteration, the noise-aware inferences potentially improve upon the noise-naive initialization, but as long as close approximations are used to enable noise-aware inference and as long as priors are relatively informative then there should be no deterioration from the initialization. Therefore we conjecture that noise-aware inference should produce point estimates that are at least as good as those produced by noise-naive methods using the corresponding release mechanism, especially in the smaller data regimes where noise-naive methods have been shown to not perform well.

- **SSP vs. other release mechanisms** In this thesis we have focused specifically on developing inference algorithms for noise-aware inference when SSP is the release algorithm. Since SSP has been observed to give state-of-the-art point estimation performance in a number of models (Chapter 3, Chapter 4, Foulds et al. [2016]; Wang [2018]), we expect that noise-aware Bayesian inference via SSP is competitive or better in terms of point estimation for these problems against other release mechanisms, but this requires further empirical exploration. Could SSP-based methods be outperformed by methods leveraging more suitable release mechanisms? Initial work on this question leads us to conjecture that in general, the more the release mechanism is specialized towards the specific problem at hand, the better the resulting point estimates. In Chapter 3 our SSP-based method is designed specifically with point estimation in undirected graphical models in mind, and it outperforms the state-of-the-art general-purpose private stochastic gradient descent method (PSGD; Abadi et al.

[2016]). Likewise, in Section 5.4 where the problem is to make a regression prediction at a specific x value, the output perturbation-based `OI` method goes to great lengths to minimize sensitivity of that one released point estimate and subsequently outperforms the more general Bayesian methods which are designed to produce posteriors and thus point estimates at any x value. The corollary to this conjecture is that in designing and choosing methods, one must balance the trade-offs between point estimation quality and applicability to more generalized problems.

- **Noise-aware inference for other release mechanisms** If SSP may not be the most suitable release mechanism for a given problem, can we develop noise-aware inference algorithms for a broader class of release mechanisms, or even very general purpose inference routines, so that we have the option of adding noise awareness to the best existing algorithms for individual problems? SSP importantly allows us to leverage approximations that lead to tractable yet effective methods, e.g. the CLT approximation for sufficient statistics as sums over iid individuals, so equivalent tricks would need to be developed for other release mechanisms.

6.2.3 More complex models

For what other models can we develop noise-aware inference methods? More specifically, in what settings does noise-aware inference provide state-of-the-art performance, and in what settings are noise-naive methods as good as possible? Of specific interest would be unbounded multivariate exponential family models, generalized linear models, and logistic regression.

APPENDIX A

CHAPTER 3: UNDIRECTED GRAPHICAL MODELS

A.1 Extra Proofs

A.1.1 Proof of Proposition 2

Proof. It is well known that the local sensitivity of any contingency table with respect to our definition of nbrs($\mathbf{X}$) is one. This is easy to see from the definition of $\mathbf{n}_C$ following Eq. (3.2): each individual contributes a count of exactly one to each clique contingency table. Since there are $|\mathcal{C}|$ tables, the local sensitivity is exactly $|\mathcal{C}|$ for all data sets, and, therefore, the sensitivity is the same. □

A.1.2 Proof of Proposition 3

Proof. Note that $n_C(i_C)$ is a sum of N iid indicator variables, so $n_C(i_C) \sim \text{Binomial}\big(N, \mu_C(i_C)\big)$, and $\text{Var}\left(n_c(i_C)\right) = N\mu_C(i_C)\big(1 - \mu_C(i_C)\big)$. Now let $z \sim \text{Laplace}(|\mathcal{C}|/\epsilon)$ and write:

$$\bar{\mu}_C(i_C) = \frac{1}{N}\big(n_C(i_C) + z\big)$$

Recall that $\mathbb{E}[z] = 0$ and $\text{Var}(z) = 2|\mathcal{C}|^2/\epsilon^2$. We see immediately that $\mathbb{E}[\bar{\mu}_C(i_C)] = \mathbb{E}\big[n_C(i_C)/N\big] = \mu_C(i_C)$. Therefore, the estimator is unbiased and its mean-squared error is equal to its variance. Since $n_C(i_C)$ and z are independent, we have:

$$\begin{aligned}\text{Var}\left(\bar{\mu}_C(i_C)\right) &= \frac{\text{Var}\left(n_C(i_C)\right)}{N^2} + \frac{\text{Var}(z)}{N^2} \\ &= \frac{\mu_C(i_C)\big(1 - \mu_C(i_C)\big)}{N} + \frac{2|\mathcal{C}|^2}{N^2\epsilon^2}\end{aligned}$$

The fact that $p(\mathbf{x}; \hat{\boldsymbol{\theta}})$ converges to $p(\mathbf{x}; \boldsymbol{\theta})$ follows from Proposition 1 and the consistency of the marginals, as long as the true marginals $\boldsymbol{\mu}$ lie in the interior of the marginal polytope $\mathcal{M}$. However, this is guaranteed because the true distribution $p(\mathbf{x}; \boldsymbol{\theta})$ is strictly positive. □

A.1.3 Proof of Proposition 4

Proof. After applying Stirling's approximation to $\log p(\mathbf{n}; \boldsymbol{\theta})$ we obtain [Nguyen et al., 2016]:

$$\log h(\mathbf{n}) \approx H(\mathbf{n}) = N \log N + \sum_{C \in \mathcal{C}} \hat{H}_C - \sum_{S \in \mathcal{S}} \nu(S) \hat{H}_S \tag{A.1}$$

where we define $\hat{H}_A = -\sum_{i_A \in \mathcal{X}^{|A|}} n_A(i_A) \log n_A(i_A)$ for any $A \in \mathcal{C} \cup \mathcal{S}$. The term $\hat{H}_A$ is a scaled entropy. We can rewrite it as:

$$\begin{aligned}
\hat{H}_A &= -N \sum_{i_A} \frac{n_A(i_A)}{N} \log \Big(\frac{n_A(i_A)}{N} \cdot N \Big) \\
&= -N \sum_{i_A} \hat{\mu}_A(i_A) \log \hat{\mu}_A(i_A) - N \sum_{i_A} \hat{\mu}_A(i_A) \log N \\
&= N H_A - N \log N
\end{aligned}$$

where H_A is now the entropy of the empirical marginal distribution $\hat{\boldsymbol{\mu}}_A = \mathbf{n}_A / N$. Since the total multiplicity of the separators is one less than the number of cliques, when we substitute back into Eq. (A.1), all of the $N \log N$ terms cancel, and we are left only with

$$H(\mathbf{n}) = N \cdot \Big(\sum_{C \in \mathcal{C}(\mathcal{T})} H_A - \sum_{S \in \mathcal{S}(\mathcal{T})} \nu(S) H_A \Big)$$

But, from standard arguments about the decomposition of entropy on junction trees, the term in parentheses is exactly the entropy of distribution q defined as:

$$q(\mathbf{x}) = \frac{\prod_{C \in \mathcal{C}} \prod_{i_C \in \mathcal{X}^{|C|}} \hat{\mu}_C(\mathbf{x}_C)}{\prod_{S \in \mathcal{S}} \prod_{i_S \in \mathcal{X}^{|S|}} \hat{\mu}_S(\mathbf{x}_S)^{\nu(S)}},$$

which factors according to $\mathcal{C}$ and can be written as $p(\mathbf{x}; \boldsymbol{\theta})$ for parameters $\boldsymbol{\theta}$ derived from the marginal probabilities. Although the mapping from parameters to distributions is many-to-one, for any marginals $\hat{\boldsymbol{\mu}}$, there is a unique distribution $p(\mathbf{x}; \boldsymbol{\theta})$ in the model family that has marginals $\hat{\boldsymbol{\mu}}$ [Wainwright & Jordan, 2008], so this uniquely defines $q(\mathbf{x})$ as stated in the Proposition.

□

APPENDIX B

CHAPTER 4: EXPONENTIAL FAMILY MODELS

B.1 Properties of Exponential Families

B.1.1 Form of Conjugate-Update$(\lambda, x_{1:n})$

Following Diaconis & Ylvisaker [1979], the prior is

$$p(\eta \mid \lambda) = h(\lambda) \exp\left(\lambda_1^\top \eta - \lambda_2 A(\eta) - B(\lambda)\right),$$

where the parameters are $\lambda = [\lambda_1, \lambda_2]$ and sufficient statistics are $[\eta, -A(\eta)]$

The posterior after observing $x_{1:n}$ is

$$p(\eta \mid \lambda, x_{1:n}) = h(\lambda') \exp\left(\lambda_1'^\top x - \lambda_2' A(\eta) - B(\lambda')\right)$$

$$\lambda_1' = \lambda_1 + \sum_i t(x_i)$$

$$\lambda_2' = \lambda_2 + n$$

Define above updates as $\lambda' = \text{Conjugate-Update}(\lambda, x_{1:n})$

B.1.1.1 Proof of Log-Partition Function of Truncated Distribution used in Lemma 1

Claim:

$$\hat{A}(\eta) = A(\eta) + \log\left(F(w; \eta) - F(v; \eta)\right)$$

Proof:

$$\begin{aligned}\exp\left(\hat{A}(\eta)\right) &= \int_v^w h(x)\exp\left(\eta^T t(x)\right)dx \\ &= \exp\left(A(\eta)\right)\int_v^w h(x)\exp\left(\eta^T t(x) - A(\eta)\right)\,dx \\ &= \exp\left(A(\eta)\right)\left(F(w;\theta) - F(v;\theta)\right)\end{aligned}$$

B.1.1.2 Proof of Lemma 1: Mean and Variance of $t(x)$ in truncated distribution

Claim

$$\begin{aligned}\mathbb{E}_{\hat{p}}[t(x)] &= \mathbb{E}_p[t(x)] + \frac{\partial}{\partial\eta^T}\log\left(F(w;\eta) - F(v;\eta)\right) \\ \mathrm{Var}_{\hat{p}}[t(x)] &= \mathrm{Var}_p[t(x)] + \frac{\partial^2}{\partial\eta\partial\eta^T}\log\left(F(w;\eta) - F(v;\eta)\right)\end{aligned}$$

Proof:

$$\begin{aligned}\mathbb{E}_{\hat{p}}[t(x)] &= \frac{\partial}{\partial\eta^T}\hat{A}(\eta) \\ &= \frac{\partial}{\partial\eta^T}\Big(A(\eta) + \log\left(F(w;\eta) - F(v;\eta)\right)\Big) \\ &= \frac{\partial}{\partial\eta^T}A(\eta) + \frac{\partial}{\partial\eta^T}\log\left(F(w;\eta) - F(v;\eta)\right) \\ &= \mathbb{E}_p[t(x)] + \frac{\partial}{\partial\eta^T}\log\left(F(w;\eta) - F(v;\eta)\right)\end{aligned}$$

The proof for $\mathrm{Var}_{\hat{p}}[t(x)]$ is similar.

B.2 Derivation of σ^2 Gibbs update

We fully derive the Gibbs update for the noise variance σ^2 of the augmented model as stated in Park & Casella [2008]. We represent the Laplace distribution with scale $b = \Delta_s/\epsilon$ as a scale mixture of normals, i.e. a zero-mean normal with an exponential prior on the variance:

$$p(z \mid b) = \frac{1}{2b}\exp\left(-\frac{|z|}{b}\right) = \int_0^\infty \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{z^2}{2\sigma^2}\right)}_{p(z|\sigma^2)} \cdot \underbrace{\ell\exp\left(-\ell\sigma^2\right)}_{p(\sigma^2|b)} d\sigma^2, \quad \ell = 1/2b^2$$

For clarity we have written the exponential rate as $\ell = 1/2b^2$. Also recall that the noise z corresponds to the difference $y - s$ between the noisy and non-noisy sufficient statistics in our model. As per Park & Casella [2008] we can write the conditional update for σ^2 as a Wald distribution (inverse-Gaussian) with the change of variable $t = 1/\sigma^2$:

$$\begin{aligned}
p_t\left(t \mid z, \ell\right) &= \left|\frac{d}{dt}\frac{1}{t}\right| \cdot p_{\sigma^2}\left(\frac{1}{t} \mid z, \ell\right) \\
&= \frac{1}{t^2} \cdot p_{\sigma^2}\left(\frac{1}{t} \mid z, \ell\right) \\
&= \frac{1}{t^2} \cdot \frac{1}{\sqrt{2\pi\frac{1}{t}}}\exp\left(-\frac{z^2}{2\frac{1}{t}}\right) \cdot \ell\exp\left(-\frac{\ell}{t}\right) \\
&\propto \frac{1}{\sqrt{t^3}}\exp\left(-\frac{z^2}{2}t - \frac{\ell}{t}\right)
\end{aligned}$$

`numpy.random.Wald` is a two-parameter (mean and scale) implementation of inverse-Gaussian. Its pdf is

$$
\begin{aligned}
\text{Wald}(t; \mu, \gamma) &= \frac{\gamma}{\sqrt{2\pi t^3}} \exp\left(-\frac{\gamma(t-\mu)^2}{2\mu^2 t}\right) \\
&\propto \frac{1}{\sqrt{t^3}} \exp\left(-\frac{\gamma(t-\mu)^2}{2\mu^2 t}\right) \\
&= \frac{1}{\sqrt{t^3}} \exp\left(-\frac{\gamma t^2 - 2\gamma\mu t + \gamma\mu^2}{2\mu^2 t}\right) \\
&= \frac{1}{\sqrt{t^3}} \exp\left(-\frac{\gamma}{2\mu^2} t + \frac{\gamma}{\mu} - \frac{\gamma}{2t}\right) \\
&\propto \frac{1}{\sqrt{t^3}} \exp\left(-\frac{\gamma}{2\mu^2} t - \frac{\gamma}{2t}\right)
\end{aligned}
$$

Then matching parameters we have

$$
\begin{aligned}
\gamma &= 2\ell \\
&= \frac{1}{b^2}
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\gamma}{\mu^2} &= z^2 \\
\mu &= \sqrt{\frac{\gamma}{z^2}} = \frac{1}{bz}
\end{aligned}
$$

So we draw t from

$$
p\left(t \mid z, b\right) = \text{Wald}\left(t;\, \frac{1}{bz}, \frac{1}{b^2}\right)
$$

and set $\sigma^2 = 1/t$.

B.3 Sensitivity of Sufficient Statistics in Truncated Model

Recall that $\hat{t}(x) = \mathbf{1}_{[v,w]}(x)\, t(x)$. Then

$$
\begin{aligned}
\Delta_{\hat{s}} &= \max_{x,y\in\mathbb{R}} \|\hat{t}(x) - \hat{t}(y)\|_1 \\
&= \max_{x,y\in\mathbb{R}} \sum_j |\hat{t}_j(x) - \hat{t}_j(y)| \\
&\leq \sum_j \max_{x,y\in\mathbb{R}} |\hat{t}_j(x) - \hat{t}_j(y)| \\
&= \sum_j \max \Big\{ \max_{x\in[v,w], y\notin[v,w]} |\hat{t}_j(x) - \hat{t}_j(y)|, \max_{x,y\in[v,w]} |\hat{t}_j(x) - \hat{t}_j(y)| \Big\} \\
&= \sum_j \max \Big\{ \max_{x\in[v,w]} |t_j(x)|, \max_{x,y\in[v,w]} |t_j(x) - t_j(y)| \Big\}
\end{aligned}
$$

B.4 Proof of uniformity of CDF transform used by Cook et al. [2006]

Claim: Let X be a random variable with CDF F. The random variable $U = F(X)$ is uniformly distributed.

Proof:

$$
\begin{aligned}
\Pr(U \leq u) &= \Pr(F(X) \leq u) \\
&= \Pr\big(F^{-1}(F(X)) \leq F^{-1}(u)\big) \\
&= \Pr(X \leq F^{-1}(u)) \\
&= F\big(F^{-1}(u)\big) \\
&= u
\end{aligned}
$$

B.5 Convergence of Gibbs Sampler

Figure B.1 shows the progress of sampled model parameters over the course of 500 iterations for both binomial and exponential models. For both models the samples quickly converge to the vicinity of the true parameter.

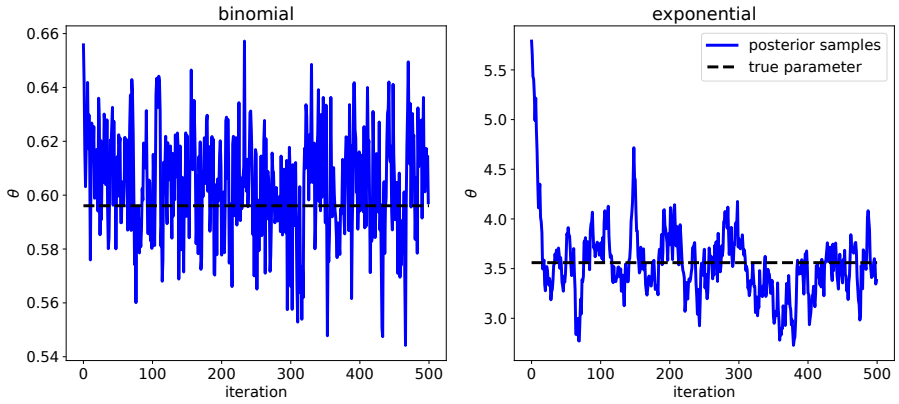


Figure B.1: Progress of Gibbs sampler parameters over iterations at ($n = 1000$; $\epsilon = 0.1$) for binomial and exponential models.

APPENDIX C

CHAPTER 5: LINEAR REGRESSION

C.1 Appendix

C.1.1 Derivation of non-private and private posteriors in Section 5.2.3

See corresponding models in Figure C.1.

$$
\begin{aligned}
p(\boldsymbol{\theta}, \sigma^2 \mid X, \mathbf{y}) &= \frac{p(\boldsymbol{\theta}, \sigma^2, X, \mathbf{y})}{p(X, \mathbf{y})} \\
&= \frac{p(X)p(\boldsymbol{\theta}, \sigma^2)p(\mathbf{y}|X, \boldsymbol{\theta}, \sigma^2)}{p(X)p(\mathbf{y} \mid X)} \\
&= \frac{p(\boldsymbol{\theta}, \sigma^2)p(\mathbf{y}|X, \boldsymbol{\theta}, \sigma^2)}{p(\mathbf{y} \mid X)} \\
&= \frac{p(\boldsymbol{\theta}, \sigma^2)p(\mathbf{y}|X, \boldsymbol{\theta}, \sigma^2)}{\int p(\mathbf{y}, \boldsymbol{\theta}, \sigma^2 \mid X)\, d\boldsymbol{\theta}, \sigma^2} \\
&= \frac{p(\boldsymbol{\theta}, \sigma^2)p(\mathbf{y}|X, \boldsymbol{\theta}, \sigma^2)}{\int p(\boldsymbol{\theta}, \sigma^2)p(\mathbf{y} \mid X, \boldsymbol{\theta}, \sigma^2)\, d\boldsymbol{\theta}, \sigma^2}
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{\theta}, \sigma^2 \mid z) &= \int p(X, \mathbf{y}, \boldsymbol{\theta}, \sigma^2, z)\, dX\, d\mathbf{y} \\
&= \int \frac{p(X, \mathbf{y}, \boldsymbol{\theta}, \sigma^2)p(z \mid X, \mathbf{y}, \boldsymbol{\theta}, \sigma^2)}{p(z)}\, dX\, d\mathbf{y} \\
&= p(z) \int p(X, \mathbf{y}, \boldsymbol{\theta}, \sigma^2)p(z \mid X, \mathbf{y}, \boldsymbol{\theta}, \sigma^2)\, dX\, d\mathbf{y}
\end{aligned}
$$

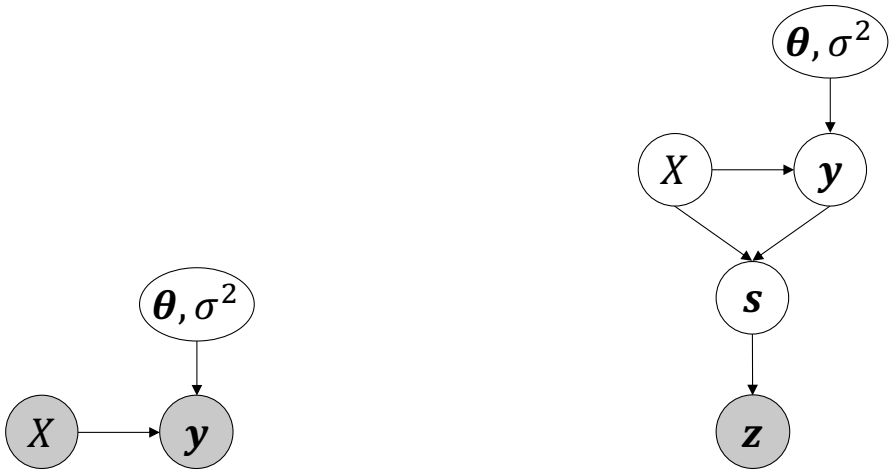


Figure C.1: (a) Non-private and (b) private regression models.

C.1.2 Gibbs Sufficient Statistic Update

C.1.2.1 Derivations of Equation 5.2: Components of $\boldsymbol{\mu}_t$

$$\begin{aligned}
\mathbb{E}\left[x_i y\right] &= \mathbb{E}_x\left[x_i \mathbb{E}_{y|x}\left[y\right]\right] \\
&= \mathbb{E}_x\left[x_i \boldsymbol{\theta}^T \mathbf{x}\right] \\
&= \mathbb{E}_x\left[x_i \sum_j \theta_j x_j\right] \\
&= \sum_j \theta_j \mathbb{E}\left[x_i x_j\right]
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}\left[y^2\right] &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{y|\mathbf{x}}\left[y^2\right]\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\sigma^2 + \left(\boldsymbol{\theta}^T\mathbf{x}\right)^2\right] \\
&= \sigma^2 + \mathbb{E}\left[\left(\sum_i \theta_i x_i\right)^2\right] \\
&= \sigma^2 + \mathbb{E}\left[\sum_{i,j} \theta_i\theta_j x_i x_j\right] \\
&= \sigma^2 + \sum_{i,j} \theta_i\theta_j \mathbb{E}\left[x_i x_j\right]
\end{aligned}
$$

C.1.2.2 Derivations of Equation 5.3: Components of Σ_t

$$
\begin{aligned}
\operatorname{Cov}\left(x_i x_j, x_k y\right) &= \mathbb{E}\left[x_i x_j x_k y\right] - \mathbb{E}\left[x_i x_j\right]\mathbb{E}\left[x_k y\right] \\
&= \mathbb{E}_x\left[x_i x_j x_k \mathbb{E}_{y|x}\left[y\right]\right] - \mathbb{E}\left[x_i x_j\right]\mathbb{E}\left[x_k y\right] \\
&= \mathbb{E}_x\left[x_i x_j x_k \sum_l \theta_l x_l\right] - \mathbb{E}\left[x_i x_j\right]\sum_l \theta_l \mathbb{E}\left[x_k x_l\right] \\
&= \sum_l \theta_l \mathbb{E}\left[x_i x_j x_k x_l\right] - \sum_l \theta_l \mathbb{E}\left[x_i x_j\right]\mathbb{E}\left[x_k x_l\right] \\
&= \sum_l \theta_l \operatorname{Cov}\left(x_i x_j, x_k x_l\right)
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{Cov}\left(x_i x_j, y^2\right) &= \mathbb{E}\left[x_i x_j y^2\right] - \mathbb{E}\left[x_i x_j\right]\mathbb{E}\left[y^2\right] \\
&= \mathbb{E}_x\left[x_i x_j \mathbb{E}_{y|x}\left[y^2\right]\right] - \mathbb{E}\left[x_i x_j\right]\mathbb{E}\left[y^2\right] \\
&= \mathbb{E}_x\left[x_i x_j\left(\sigma^2 + \sum_{k,l}\theta_k\theta_l x_k x_l\right)\right] - \mathbb{E}\left[x_i x_j\right]\left(\sigma^2 + \sum_{k,l}\theta_k\theta_l\mathbb{E}\left[x_k x_l\right]\right) \\
&= \sigma^2\mathbb{E}\left[x_i x_j\right] + \sum_{k,l}\theta_k\theta_l\mathbb{E}\left[x_i x_j x_k x_l\right] - \sigma^2\mathbb{E}\left[x_i x_j\right] - \sum_{k,l}\theta_k\theta_l\mathbb{E}\left[x_i x_j\right]\mathbb{E}\left[x_k x_l\right] \\
&= \sum_{k,l}\theta_k\theta_l \operatorname{Cov}\left(x_i x_j, x_k x_l\right)
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{Cov}\left(x_i y, x_j y\right) &= \mathbb{E}\left[x_i x_j y^2\right] - \mathbb{E}\left[x_i y\right]\mathbb{E}\left[x_j y\right] \\
&= \mathbb{E}_x\left[x_i x_j \mathbb{E}_{y|x}\left[y^2\right]\right] - \left(\sum_k \theta_k \mathbb{E}\left[x_i x_k\right]\right)\left(\sum_l \theta_l \mathbb{E}\left[x_j x_l\right]\right) \\
&= \mathbb{E}\left[x_i x_j \left(\sigma^2 + \sum_{k,l} \theta_k \theta_l x_k x_l\right)\right] - \sum_{k,l} \theta_k \theta_l \mathbb{E}\left[x_i x_k\right]\mathbb{E}\left[x_j x_l\right] \\
&= \sigma^2 \mathbb{E}\left[x_i x_j\right] + \sum_{k,l} \theta_k \theta_l \left(\mathbb{E}\left[x_i x_j x_k x_l\right] - \mathbb{E}\left[x_i x_k\right]\mathbb{E}\left[x_j x_l\right]\right) \\
&= \sigma^2 \mathbb{E}\left[x_i x_j\right] + \sum_{k,l} \theta_k \theta_l \operatorname{Cov}\left(x_i x_k, x_j x_l\right)
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{Cov}\left(x_i y, y^2\right) &= \mathbb{E}\left[x_i y^3\right] - \mathbb{E}\left[x_i y\right]\mathbb{E}\left[y^2\right] \\
&= \mathbb{E}_x\left[x_i \mathbb{E}_{y|x}\left[y^3\right]\right] - \mathbb{E}\left[x_i y\right]\mathbb{E}\left[y^2\right] \\
&= \mathbb{E}_x\left[x_i \left(\sum_{j,k,l} \theta_j \theta_k \theta_l x_j x_k x_l + 3\sigma^2 \sum_j \theta_j x_j\right)\right] \\
&\qquad - \sum_j \theta_j \mathbb{E}\left[x_i x_j\right]\left(\sigma^2 + \sum_{k,l} \theta_k \theta_l x_k x_l\right) \\
&= \sum_{j,k,l} \theta_j \theta_k \theta_l \mathbb{E}\left[x_i x_j x_k x_l\right] + 3\sigma^2 \sum_j \theta_j \mathbb{E}\left[x_i x_j\right] \\
&\qquad - \sigma^2 \sum_j \theta_j \mathbb{E}\left[x_i x_j\right] + \sum_{j,k,l} \theta_j \theta_k \theta_l \mathbb{E}\left[x_i x_j\right]\mathbb{E}\left[x_k x_l\right] \\
&= \sum_{j,k,l} \theta_j \theta_k \theta_l \operatorname{Cov}\left(x_i x_j, x_k x_l\right) + 2\sigma^2 \sum_j \theta_j \mathbb{E}\left[x_i x_j\right]
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{Var}\left(y^2\right) &= \mathbb{E}\left[y^4\right] - \mathbb{E}\left[y^2\right]^2 \\
&= 3\sigma^4 + \sum_{j,k,l,m} \theta_j\theta_k\theta_l\theta_m \mathbb{E}\left[x_j x_k x_l x_m\right] + 6\sigma^2 \sum_{j,k} \theta_j\theta_k \mathbb{E}\left[x_j x_k\right] - \left(\sigma^2 + \sum_{j,k} \theta_j\theta_k \mathbb{E}\left[x_j x_k\right]\right)^2 \\
&= 3\sigma^4 + \sum_{j,k,l,m} \theta_j\theta_k\theta_l\theta_m \mathbb{E}\left[x_j x_k x_l x_m\right] + 6\sigma^2 \sum_{j,k} \theta_j\theta_k \mathbb{E}\left[x_j x_k\right] \\
&\quad - \sigma^4 - 2\sigma^2 \sum_{j,k} \theta_j\theta_k \mathbb{E}\left[x_j x_k\right] - \sum_{j,k,l,m} \theta_j\theta_k\theta_l\theta_m \mathbb{E}\left[x_j x_k\right] \mathbb{E}\left[x_l x_m\right] \\
&= 2\sigma^4 + \sum_{j,k,l,m} \theta_j\theta_k\theta_l\theta_m \left(\mathbb{E}\left[x_j x_k x_l x_m\right] - \mathbb{E}\left[x_j x_k\right] \mathbb{E}\left[x_l x_m\right]\right) + 4\sigma^2 \sum_{j,k} \theta_j\theta_k \mathbb{E}\left[x_j x_k\right] \\
&= 2\sigma^4 + \sum_{j,k,l,m} \theta_j\theta_k\theta_l\theta_m \operatorname{Cov}\left(x_j x_k, x_l x_m\right) + 4\sigma^2 \sum_{j,k} \theta_j\theta_k \mathbb{E}\left[x_j x_k\right]
\end{aligned}
$$

BIBLIOGRAPHY

Abadi, Martín, Chu, Andy, Goodfellow, Ian, McMahan, H. Brendan, Mironov, Ilya, Talwar, Kunal, and Zhang, Li. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318. ACM, 2016.

Abowd, John M and Schmutte, Ian M. An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1):171–202, 2019.

Agresti, Alan and Finlay, Barbaracoaut. *Statistical methods for the social sciences*. Number 300.72 A3. 2009.

Awan, Jordan and Slavkovic, Aleksandra. Structure and sensitivity in differential privacy: Comparing k-norm mechanisms. *arXiv preprint arXiv:1801.09236*, 2018.

Barak, Boaz, Chaudhuri, Kamalika, Dwork, Cynthia, Kale, Satyen, McSherry, Frank, and Talwar, Kunal. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 273–282. ACM, 2007.

Barrientos, Andrés F., Reiter, Jerome P., Machanavajjhala, Ashwin, and Chen, Yan. Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics*, 0(0):1–24, 2019.

Bassily, Raef, Smith, Adam, and Thakurta, Abhradeep. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pp. 464–473. IEEE, 2014.

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

Bernstein, Garrett, McKenna, Ryan, Sun, Tao, Sheldon, Daniel, Hay, Michael, and Miklau, Gerome. Differentially private learning of undirected graphical models using collective graphical models. In *International Conference on Machine Learning*, pp. 478–487, 2017.

Bickel, Peter J. and Doksum, Kjell A. *Mathematical statistics: basic ideas and selected topics, volume I*, volume 117. CRC Press, 2015.

Blum, Avrim, Dwork, Cynthia, McSherry, Frank, and Nissim, Kobbi. Practical privacy: the SuLQ framework. In *Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pp. 128–138. ACM, 2005.

Casella, George and Berger, Roger Lee. *Statistical Inference*, chapter 6. Thomson Learning, 2002.

Chaudhuri, Kamalika and Monteleoni, Claire. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems*, pp. 289–296, 2009.

Chaudhuri, Kamalika, Monteleoni, Claire, and Sarwate, Anand D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12 (Mar):1069–1109, 2011.

Chetty, Raj and Friedman, John N. A practical method to reduce privacy loss when disclosing statistics based on small samples. In *AEA Papers and Proceedings*, volume 109, pp. 414–20, 2019.

Cook, Samantha R., Gelman, Andrew, and Rubin, Donald B. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692, 2006.

Diaconis, Persi and Ylvisaker, Donald. Conjugate priors for exponential families. *The Annals of statistics*, pp. 269–281, 1979.

Dimitrakakis, Christos, Nelson, Blaine, Mitrokotsa, Aikaterini, and Rubinstein, Benjamin I.P. Robust and private Bayesian inference. In *International Conference on Algorithmic Learning Theory*, pp. 291–305. Springer, 2014.

Duchi, John, Shalev-Shwartz, Shai, Singer, Yoram, and Chandra, Tushar. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pp. 272–279. ACM, 2008.

Dwork, Cynthia and Lei, Jing. Differential privacy and robust statistics. In *STOC*, volume 9, pp. 371–380, 2009.

Dwork, Cynthia and Roth, Aaron. *The Algorithmic Foundations of Differential Privacy*. Found. and Trends in Theoretical Computer Science, 2014.

Dwork, Cynthia and Smith, Adam. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2), 2010.

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

Fienberg, Stephen E. and Rinaldo, Alessandro. Maximum likelihood estimation in log-linear models. *The Annals of Statistics*, pp. 996–1023, 2012.

Fisher, Ronald Aylmer. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368, 1922.

Foulds, James, Geumlek, Joseph, Welling, Max, and Chaudhuri, Kamalika. On the theory and practice of privacy-preserving Bayesian data analysis. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pp. 192–201, 2016.

Geumlek, Joseph, Song, Shuang, and Chaudhuri, Kamalika. Renyi differential privacy mechanisms for posterior sampling. In *Advances in Neural Information Processing Systems*, pp. 5295–5304, 2017.

Geyer, Charles J and Thompson, Elizabeth A. Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):657–683, 1992.

Gretton, Arthur, Borgwardt, Karsten M., Rasch, Malte J., Schölkopf, Bernhard, and Smola, Alexander. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Haberman, Shelby J. Log-linear models for frequency data: Sufficient statistics and likelihood equations. *The Annals of Statistics*, 1(4):617–632, 1973. ISSN 00905364. URL `http://www.jstor.org/stable/2958307`.

Hakimi, S Louis. On realizability of a set of integers as degrees of the vertices of a linear graph. i. *Journal of the Society for Industrial and Applied Mathematics*, 10 (3):496–506, 1962.

Handcock, Mark S and Gile, Krista J. Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5, 2010.

Hardt, Moritz, Ligett, Katrina, and McSherry, Frank. A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems*, pp. 2339–2347, 2012.

Havel, Václav. A remark on the existence of finite graphs. *Casopis Pest. Mat.*, 80: 477–480, 1955.

He, Xi, Cormode, Graham, Machanavajjhala, Ashwin, Procopiuc, Cecilia M., and Srivastava, Divesh. DPT: differentially private trajectory synthesis using hierarchical reference systems. *Proceedings of the VLDB Endowment*, 8(11):1154–1165, 2015.

Hoffman, Matthew D. and Gelman, Andrew. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

Jain, Prateek and Thakurta, Abhradeep. Differentially private learning with kernels. *ICML (3)*, 28:118–126, 2013.

Karwa, Vishesh, Slavković, Aleksandra B., and Krivitsky, Pavel. Differentially private exponential random graphs. In *International Conference on Privacy in Statistical Databases*, pp. 143–155. Springer, 2014.

Karwa, Vishesh, Slavković, Aleksandra, et al. Inference using noisy degrees: Differentially private *beta*-model and synthetic graphs. *The Annals of Statistics*, 44(1): 87–112, 2016.

Kasiviswanathan, Shiva Prasad, Lee, Homin K., Nissim, Kobbi, Raskhodnikova, Sofya, and Smith, Adam. What can we learn privately? *SIAM Journal on Computing*, 40 (3):793–826, 2011.

Kifer, Daniel and Machanavajjhala, Ashwin. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pp. 193–204. ACM, 2011.

Kifer, Daniel, Smith, Adam, and Thakurta, Abhradeep. Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research*, 1(41):3–1, 2012.

Klimt, Bryan and Yang, Yiming. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pp. 217–226. Springer, 2004.

Koller, Daphne and Friedman, Nir. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Kullback, Solomon and Leibler, Richard A. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Liu, Li-Ping, Sheldon, Daniel R., and Dietterich, Thomas G. Gaussian approximation of collective graphical models. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

Massey Jr., Frank J. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

McSherry, Frank and Mironov, Ilya. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 627–636. ACM, 2009.

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

Minami, Kentaro, Arai, Hitomi, Sato, Issei, and Nakagawa, Hiroshi. Differential privacy without sensitivity. In *Advances in Neural Information Processing Systems*, pp. 956–964, 2016.

Nguyen, Duc Thien, Kumar, Akshat, Lau, Hoong Chuin, and Sheldon, Daniel. Approximate inference using DC programming for collective graphical models. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 685–693, 2016.

O'Hagan, Anthony and Forster, Jonathan. Kendall's advanced theory of statistics, volume 2b: Bayesian inference. 1994.

Park, Trevor and Casella, George. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Peter Bergman, Raj Chetty, Stefanie DeLuca Nathaniel Hendren Lawrence Katz Christopher Palmer. Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice. 2019.

Petersen, Kaare Brandt and Pedersen, Michael Syskind. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Rencher, Alvin C. *Methods of multivariate analysis*, volume 492. John Wiley & Sons, 2003.

Robbins, Herbert. The asymptotic distribution of the sum of a random number of random variables. *Bulletin of the American Mathematical Society*, 54(12):1151–1161, 1948.

Rubinstein, Benjamin I.P., Bartlett, Peter L., Huang, Ling, and Taft, Nina. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. *arXiv preprint arXiv:0911.5708*, 2009.

Salvatier, John, Wiecki, Thomas V., and Fonnesbeck, Christopher. Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2:e55, apr 2016. doi: 10.7717/peerj-cs.55. URL `https://doi.org/10.7717/peerj-cs.55`.

Schein, Aaron, Wu, Zhiwei Steven, Zhou, Mingyuan, and Wallach, Hanna. Locally private Bayesian inference for count models. *NIPS 2017 Workshop: Advances in Approximate Bayesian Inference*, 2018.

Sheffet, Or. Differentially private ordinary least squares. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Sheldon, Daniel R. and Dietterich, Thomas G. Collective graphical models. *Neural Information Processing Systems (NIPS)*, 2011.

Sheldon, Daniel R., Sun, Tao, Kumar, Akshat, and Dietterich, Thomas G. Approximate inference in collective graphical models. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

Smith, Adam. Efficient, differentially private point estimators. *arXiv preprint arXiv:0809.4794*, 2008.

Smith, Adam. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, pp. 813–822, 2011a.

Smith, Adam. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing (STOC)*, pp. 813–822. ACM, 2011b.

Sun, Tao, Sheldon, Daniel R., and Kumar, Akshat. Message passing for collective graphical models. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Vilnis, Luke, Belanger, David, Sheldon, Daniel, and McCallum, Andrew. Bethe projections for non-local inference. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 892–901. AUAI Press, 2015.

Vu, Duy and Slavkovic, Aleksandra. Differential privacy for clinical trial data: Preliminary evaluations. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pp. 138–143. IEEE, 2009.

Wainwright, Martin J. and Jordan, Michael I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2): 1–305, 2008.

Wang, Yu-Xiang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

Wang, Yu-Xiang, Fienberg, Stephen, and Smola, Alex. Privacy for free: Posterior sampling and stochastic gradient Monte Carlo. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 2493–2502, 2015.

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

Wu, Xi, Kumar, Arun, Chaudhuri, Kamalika, Jha, Somesh, and Naughton, Jeffrey F. Differentially private stochastic gradient descent for in-RDBMS analytics. *CoRR*, abs/1606.04722, 2016. URL `http://arxiv.org/abs/1606.04722`.

Yang, Xiaolin, Fienberg, Stephen E., and Rinaldo, Alessandro. Differential privacy for protecting multi-dimensional contingency table data: Extensions and applications. *Journal of Privacy and Confidentiality*, 4(1):5, 2012.

Zhang, Jun, Zhang, Zhenjie, Xiao, Xiaokui, Yang, Yin, and Winslett, Marianne. Functional mechanism: regression analysis under differential privacy. *Proceedings of the VLDB Endowment*, 5(11):1364–1375, 2012.

Zhang, Jun, Cormode, Graham, Procopiuc, Cecilia M., Srivastava, Divesh, and Xiao, Xiaokui. Privbayes: Private data release via Bayesian networks. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pp. 1423–1434, 2014.

Zhang, Zuhe, Rubinstein, Benjamin I.P., and Dimitrakakis, Christos. On the differential privacy of Bayesian inference. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.