UNIVERSITY OF CALIFORNIA SAN DIEGO

Sampling from Distributions under Differential Privacy Notions

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science

by

Joseph Donald Geumlek

Committee in charge:

Professor Kamalika Chaudhuri, Chair
Professor Sanjoy Dasgupta
Professor Russell Impagliazzo
Professor Young-Han Kim
Professor Lawrence Saul

2020

The Dissertation of Joseph Donald Geumlek is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2020

DEDICATION

This work is dedicated to my family for their continual support of my academic goals, and to all the educators finding ways to maintain the learning landscape in the uncertain times of 2020.

EPIGRAPH

This was the path The Seer had found in her scrying game, a series of private dice rolls and inscrutable diagrams.

Kentucky Route Zero, Act V

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Kamalika Chaudhuri for serving as my advisor, encouraging me throughout this degree program, paying attention to my goals, and working with me on many of my publications.

I also extend my thanks to Tom Diethe and Borja Balle, who were instrumental in my internship at Amazon in Cambridge, UK, for the Summer of 2018. Along with this, I thank my collaborators Gilles Barthe and Marco Gaboardi who assisted during and after that internship on the NeurIPS 2019 publication.

Jimmy Foulds and Max Welling are also acknowledged for their input into my first publication at UAI 2016. I extend extra thanks to Jimmy, as he also provided guidance in navigating the details of graduate school and beyond.

I also thank the numerous bright graduate students I worked with in the AI group at UC San Diego, including but not limited to Chicheng Zhang, Shuang Song, Songbai Yan, Yizhen Wang, Julaiti Alafate, Akshay Balsubramani, Matt Der, Christopher Tosh, Sharad Vikram, Casey Meehan, Mary Anne Smart, Yaoyuan Yang, Jacob Imola, Robi Bhattacharjee, Zhifeng Kong, Zhi Wang, and Shuang Liu. Shuang Song also contributed significant additional results for the NeurIPS 2017 publication that do not appear in the corresponding chapter of this thesis.

I am grateful for the positive community I found throughout UC San Diego, both inside and outside the CSE department. The social hours and grad lounges were invaluable portions of this degree program.

Many thanks go to Erilynn Heinrichsen, Paul Hadjipieris, and the rest of the Engaged Teaching Hub, who instructed and supported me as I prepared for being an instructor myself. The Summer Graduate Teaching Scholarship was a fantastic opportunity and has tremendously helped me with my career goals.

Last, but not least, I acknowledge Sherry Guo, who has been a loving and supportive partner throughout my time in the doctoral program.

Chapter 4 is based on the material in Proceedings of the Thirty-Second Conference on

Uncertainty in Artificial Intelligence 2016 (James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri, ”On the theory and practice of privacy-preserving Bayesian data analysis”). The dissertation author was the primary investigator and author of this material.

Chapter 5 is based on the material in Advances in Neural Information Processing Systems 2017 (Joseph Geumlek, Shuang Song, and Kamalika Chaudhuri. ”Rényi differential privacy mechanisms for posterior sampling”). The dissertation author was the primary investigator and author of this material.

Chapter 6 is based on the material in Advances in Neural Information Processing Systems 2019 (Borja Balle, Gilles Barthe, Marco Gaboardi, and Joseph Geumlek. ”Privacy amplification by mixing and diffusion mechanisms”). The dissertation author was the primary investigator and author of this material.

Chapter 7 is based on the material in IEEE International Symposium on Information Theory 2019 (Joseph Geumlek, and Kamalika Chaudhuri. ”Profile-based privacy for locally private computations”). The dissertation author was the primary investigator and author of this material.

VITA

2014	B. S. in Mathematics, University of California, Irvine
2014	B. S. in Computer Science, University of California, Irvine
2016	M. S. in Computer Science, University of California San Diego
2019-2020	Associate-In, Instructor of Record, University of California San Diego
2020	Ph. D. in Computer Science, University of California San Diego

PUBLICATIONS

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

Joseph Geumlek, Shuang Song, and Kamalika Chaudhuri. ”Rényi differential privacy mechanisms for posterior sampling.” In Advances in Neural Information Processing Systems. 2017.

Joseph Geumlek, and Kamalika Chaudhuri. ”Profile-based privacy for locally private computations.” In 2019 IEEE International Symposium on Information Theory (ISIT). IEEE, 2019.

Borja Balle, Gilles Barthe, Marco Gaboardi, and Joseph Geumlek. ”Privacy amplification by mixing and diffusion mechanisms.” In Advances in Neural Information Processing Systems, pp. 13277-13287. 2019.

ABSTRACT OF THE DISSERTATION

Sampling from Distributions under Differential Privacy Notions

by

Joseph Donald Geumlek

Doctor of Philosophy in Computer Science

University of California San Diego, 2020

Professor Kamalika Chaudhuri, Chair

An individual's personal information is gathered by a multitude of different data collectors throughout the world today. In order to maintain the trust between the individual and the aggregator, there is a need for ensuring the methods that interact with that data collection are acting in a responsible manner. One such way for maintaining trust is to use methods that come with formal privacy guarantees. A well-established source of such guarantees can be found in the framework known as differential privacy, which places significant constraints on algorithms that operate on private data.

This thesis explores the challenges of releasing samples from distributions while satisfy-

ing the requirements of differential privacy or other closely related privacy notions. We present algorithms for releasing samples in a variety of settings that differ in their privacy aims. From one angle, we protect the data values directly that arise from exponential family distributions with methods attuned to differential privacy and further methods attuned to Rényi differential privacy. From another angle, we explore protecting the identity of a secret sensitive distribution while releasing what we can from the gathered data. Additionally, a coupling-based analysis is provided for reasoning about the impact of diffusing the samples from one distribution through another in order to achieve stronger privacy guarantees from sampling than either distribution separately. These proposed methods are proven to achieve formal privacy guarantees, and we also show empirical and theoretical results about their efficacy. These results empower numerous different styles of Bayesian privacy-preserving methods, and serve as useful primitives for further privacy analyses that move beyond frequentist probabilistic methods.

Chapter 1

Introduction

As enter another decade of the 21st century, it is almost impossible to ignore the growth of data collection and data analysis in the modern highly-interconnected world. Massive data sets, combined with machine learning methodologies, permit a great deal of insight, knowledge, and value to be extracted. However, this value is encumbered with sensitive privacy issues, since this data contains specific information collected from individuals with an innate claim to maintaining their privacy. When unsure about how the data will be used, individuals may be unwilling to share honest responses Warner [1965]. In some cases, such as medical or educational data, these privacy considerations also carry legal and regulatory weight.

The importance of preserving privacy during data analysis has not gone unnoticed by the research community. In a seminal work, Dwork et al. [2006b] introduced a privacy notion, *differential privacy* (DP), which carried a rigorous mathematical foundation of useful properties [Dwork and Roth, 2014]. This privacy notion, and closely related variants of it, has dominated the privacy research landscape since its introduction. In addition to the provable guarantees, differential privacy also provided a smooth parametrization of privacy levels, allowing for a richer discussion of the costs associated with the privatization of machine learning algorithms.

An important consequence of differential privacy's definition is that the algorithms employed must be randomized. It is this randomness, unknown to outside observers, that provides the uncertainty used mask sensitive information. The focus of a privacy analysis for

an algorithm therefore lies on studying the behavior of the probabilistic mapping of inputs to outputs, and the induced distributions arising from these randomized processes. Separate from other areas of machine learning and artificial intelligence, it has become customary to refer to the algorithms and methods as ”mechanisms”. These mechanisms provide means for achieving desired output behaviors subject to strict probabilistic constraints.

Differential privacy's appeal is not just limited to theorists, and has seen large-scale adoption and deployment. Major corporations have publicly commented on their implementation of differentially private methods around their collection of data from the general public. The United States census, the federal decennial counting of the nation's populations, has introduced differential privacy into the process for 2020. The wide use of differential privacy in today's world demonstrates a significant level of maturity this research area has achieved in the years since the publication of Dwork et al. [2006b].

Although differential privacy remains a central formal privacy guarantee in the literature, much work has been spent on exploring relaxations and variations of differential privacy. The original definition makes strong assumptions that need not always match reality, and maintaining these assumptions carries notable implementation costs. One common relaxation is approximate DP Dwork et al. [2006a], which introduces a simple secondary privacy parameter which allows arbitrarily bad events to occur with bounded probability. Several different privacy notions exist along these lines in which the precise mathematical formulations of the privacy guarantee is changedMironov [2017], Bun and Steinke [2016], Dwork and Rothblum [2016], Machanavajjhala et al. [2008], Chaudhuri and Mishra [2006], Wang et al. [2016]. Another variations change the dynamics of trust in the process, such as *Local Differential Privacy* Duchi et al. [2013] which places stricter requirements on mechanisms in exchange for not requiring a central trusted data curator.

This leaves open a rich and active research field continually exploring the frontiers of these privacy definitions.

1.1 Contributions

This thesis enhances the theoretical and practical domains of privacy-preserving machine learning.

The early chapters of this thesis are structured around the presentation of the contributions that follow. Chapter 2 provides a high-level discussion of works related to this thesis and how these contributions fit into the broader literature. Chapter 3 defines and discusses the preliminary information of differential privacy and its variants that will be used extensively in later chapters.

Chapter 4 explores the problem of privately returning samples from a posterior distribution is analyzed and improved for Bayesian data analysis for exponential family distributions. This chapter explores the impact of directly perturbing the statistics of observations and the asymptotic behaviors of the induced distortions raised by privatization. Compared to prior methods, this analysis formally demonstrates a statistical efficiency enjoyed by this proposed method. Numerical experiments also support the theoretical contributions.

Chapter 5 further builds off of the work seen in chapter 4 in also tackling the problem of private posterior sampling for exponential family distributions. In this chapter, Rényi differential privacy (a relaxation of differential privacy) is used. The change to this alternative privacy notion permits more of the structural properties of exponential family distributions to be exploited. Instead of perturbing the statistics, this chapter introduces methods that leverage the inherent randomness of posterior sampling. The work is based on altering the balance between the analyst's prior knowledge and the observed sensitive data. Multiple ways of affecting that balance are proposed in analyzed, allowing a more nuanced approach to controlling the statistical efficiency guarantees. The behaviors of this re-balancing strategies are explored theoretically and numerically.

Chapter 6 moves in a different direction, and instead studies the theoretical behavior of combining multiple Rényi differentially private operations by working with fundamental distributional measures. Multiple distributions sequenced together can be reasoned about in

a combined fashion, in which we can analyze the privacy arising from passing the output from one distribution through yet another sampling process. Unlike the common discussed privacy-degradation compositional properties of privacy mechanisms, this work presents a clean coupling-based analysis for privacy-amplifying composition. This formulation provides a simpler rederivation of known results, as well enabling the analysis of novel situations.

Chapter 7 continues the exploration in fundamental aspects of privacy-preservation definitions by proposing and analyzing the properties of a new rigorous privacy setting. Here, probabilistic models are introduced in which the sensitive information lies in the identity of distributions, not the collected data itself. Although other privacy notions have created such separations, this work identifies and explores a particular class of such settings and proposes mechanisms designed around its structural properties.

Chapter 2

Related Work

This thesis touches on a variety of topics related to the theory and practice of differential privacy. Specific mechanisms are proposed and presented, alternative privacy notions are defined and explored, and fundamental compositional bounds are proven. Broad surveys of privacy-preservation techniques can be found in [Dwork and Roth, 2014, Sarwate and Chaudhuri, 2013]. A full inventory of the field of privacy-preservation on both theoretical and practical grounds would be lengthy, so this section instead focuses on the works most closely related to each chapter.

Chapters 4 and 5 focus on privacy preservation in a particular Bayesian setting of posterior sampling. This work on exponential family distributions is related to Dimitrakakis et al. [2014] and Wang et al. [2015b]. Dimitrakakis et al. [2014] examines some common posterior distributions and presents conditions under which directly sampling from these posteriors is differentially private. Wang et al. [2015b] presents another criterion for private sampling, which requires certain modifications to be made for sampling. The contributions of Chapter 4 demonstrate how such modifications can be avoided while maintaining useful asymptotic behaviors, and Chapter 5 further improves the situations under which direct sampling can be employed by re-analyzing Bayesian learning for exponential family distributions under Rényi differential privacy Mironov [2017].

Chapter 6 extends the examination of Rényi differential privacy Mironov [2017] by pro-

viding a privacy-amplification result for Rényi differential privacy. This places it alongside other works in privacy amplification: by subsampling [Chaudhuri and Mishra, 2006, Kasiviswanathan et al., 2011, Li et al., 2012, Beimel et al., 2013, 2014, Bun et al., 2015, Balle et al., 2018, Wang et al., 2019], shuffling [Erlingsson et al., 2019, Cheu et al., 2019, Balle et al., 2019] and iteration [Feldman et al., 2018]. Chapter 6 is most closely related to the privacy amplification by iterations results of [Feldman et al., 2018], and can be viewed as a generalization of their arguments via the use of couplings.

Chapter 7 proposes a privacy framework and mechanisms that achieve those privacy guarantees in a variety settings. This places the alongside many in the literature that propose novel privacy notions Kifer and Machanavajjhala [2012], Song et al. [2017], Bassily and Freund [2016], He et al. [2014], Gehrke et al. [2011], Dwork and Rothblum [2016], Kawamoto and Murakami [2018]. It shares a a connection with [Kifer and Machanavajjhala, 2012] in that it can be viewed as a special case of their Pufferfish privacy framework, in which our results fill an understudied regime of instantiations. This chapter is also very complementary to Kawamoto and Murakami [2018], which also targets a distribution-level notion of privacy, but uses different methods and analyses for achieving that goal.

Chapter 3

Preliminaries

This section introduces the fundamental definitions that motivate the work in this thesis, along with an extended discussion of their properties.

3.1 Differential Privacy

Differential privacy, as proposed in Dwork et al. [2006b], provides a formal mathematical guarantee of privacy. It builds off of a notion of neighboring data sets, and defines a closeness requirement for the output distributions of a mechanism applied to each of those neighboring data sets.

We say two data sets $\mathbf{X}$ and $\mathbf{X}'$ are *neighboring* if they differ in the private record of a single *individual* or person. With this neighboring relation, we can then define a privacy guarantee in terms of a non-negative parameter ε in which values close to zero indicate stronger privacy guarantees.

Definition 1. Differential Privacy, ε-DP.

A randomized mechanism $\mathscr{A}(\mathbf{X})$ is said to be $(\varepsilon, \delta)-$differentially private if for any subset U of the output range of $\mathscr{A}$ and any neighboring data sets $\mathbf{X}$ and $\mathbf{X}'$, we have $\mathsf{Pr}(\mathscr{A}(\mathbf{X}) \in U) \leq \exp(\varepsilon)\mathsf{Pr}(\mathscr{A}(\mathbf{X}') \in U)$.

This single parameter definition is sometime called "pure differential privacy." A common relaxation of differential privacy introduces an additional parameter $\delta \in [0,1]$ into the guarantee,

which gives rise to approximate differential privacy Dwork et al. [2006a].

Definition 2. Approximate Differential Privacy, (ε, δ)-DP.

A randomized mechanism $\mathscr{A}(\mathbf{X})$ is said to be $(\varepsilon, \delta)-$differentially private if for any subset U of the output range of $\mathscr{A}$ and any neighboring data sets $\mathbf{X}$ and $\mathbf{X}'$, we have $\mathsf{Pr}(\mathscr{A}(\mathbf{X}) \in U) \leq \exp(\varepsilon)\mathsf{Pr}(\mathscr{A}(\mathbf{X}') \in U) + \delta$.

Differential privacy enjoys a number of useful properties, of which we briefly present a couple. When relevant, the properties of differential privacy will be explained and examined in the remaining chapters of this thesis.

Observation 1. *Robustness to Post-Processing for DP*

For any mechanism $\mathscr{A}$ that satisfies (ε, δ)-DP, and any randomized function f that operates on the outputs of $\mathscr{A}$, the mechanism formed by releasing $f(\mathscr{A}(\mathbf{X}))$ for a data set $\mathbf{X}$ also satisfies (ε, δ)-DP.

At a high-level, this means any additional computations done on the output released by $\mathscr{A}$ cannot possibly degrade the privacy guarantee offered by $\mathscr{A}$.

Observation 2. *Robustness to Composition for DP*

For any mechanism $\mathscr{A}$ that satisfies $(\varepsilon_{\mathscr{A}}, \delta_{\mathscr{A}})$-DP, and any other mechanism $\mathscr{B}$ that operates on the same data as $\mathscr{A}$ and satisfies $(\varepsilon_{\mathscr{B}}, \delta_{\mathscr{B}})$-DP, the mechanism formed by releasing the tuple $(\mathscr{A}(\mathbf{X}), \mathscr{B}(\mathbf{X}))$ for a data set $\mathbf{X}$ satisfies $(\varepsilon_{\mathscr{A}} + \varepsilon_{\mathscr{B}}, \delta_{\mathscr{A}} + \delta_{\mathscr{B}})$-DP.

At a high-level, this means that multiple differentially private releases result in a degraded (but still bounded) privacy guarantee. Since this property holds for any choice of private $\mathscr{B}$, this property also provides bounds for when the second mechanism depends on the output of the first. If we let o be the output from $\mathscr{A}(\mathbf{X})$, and $\mathscr{B}_o$ represent a second mechanism that has access to the output o, then the tuple $(o, \mathscr{B}_o(\mathbf{X}))$ satisfies $(\varepsilon_{\mathscr{A}} + \varepsilon_{\mathscr{B}}, \delta_{\mathscr{A}} + \delta_{\mathscr{B}})$-DP.

These two properties allow differential privacy mechanisms to be composed and combined in a variety of ways. The privacy analysis of primitives can therefore be used to give

privacy bounds for complex multi-stage algorithms. It should be noted that the composition result is not the strongest or tightest bound generally used in the privacy literature, but the technical details of such bounds are not relevant here.

Both versions of DP (pure and approximate) are concerned with the difference the participation of a individual might have on the output distribution of the mechanism. The requirements for DP can be phrased in terms of a privacy loss variable, a random variable that captures the effective privacy loss of the mechanism output.

Definition 3. Privacy Loss Variable.

We can define a random variable Z that measures the privacy loss of a given output of a mechanism across two neighboring data sets $\mathbf{X}$ and $\mathbf{X}'$.

$$Z = \log \frac{\Pr(\mathscr{A}(\mathbf{X}) = o)}{\Pr(\mathscr{A}(\mathbf{X}') = o)}\bigg|_{o \sim \mathscr{A}(\mathbf{X})} \tag{3.1}$$

This privacy loss variable measures the log of the probability ratio of the observed output across the two neighboring data sets $\mathbf{X}$ and $\mathbf{X}'$. From a mathematical point of view, ensuring that this loss variable Z is "small" under some probability gives a measure of closeness between the distribution arising from $\mathscr{A}(\mathbf{X})$ and $\mathscr{A}(\mathbf{X}')$. A more interpretative view can place the ratio can be seen from imagining an adversary trying to guess whether the observed output came from the data set X or X' with a simple application of Bayes' rule. In this setting, we can let $Data$ represent the latent data set identity, and Out represent the random variable arising from selecting a data set and passing it to the mechanism $\mathscr{A}$. Differential privacy can be interpreting as bounding the odds of $Data = \mathbf{X}$ vs $Data = \mathbf{X}'$ in the adversary's beliefs after observing the output $Out = o$.

$$\frac{\mathsf{Pr}(Data = \mathbf{X}|Out = o)}{\mathsf{Pr}(Data = \mathbf{X}'|Out = o)} = \frac{\mathsf{Pr}(Out = o|Data = \mathbf{X})\mathsf{Pr}(Data = \mathbf{X})/\mathsf{Pr}(Out = o)}{\mathsf{Pr}(Out = o|Data = \mathbf{X}')\mathsf{Pr}(Data = \mathbf{X}')/\mathsf{Pr}(Out = o)} \tag{3.2}$$

$$= \frac{\mathsf{Pr}(Out = o|Data = \mathbf{X})\mathsf{Pr}(Data = \mathbf{X})}{\mathsf{Pr}(Out = o|Data = \mathbf{X}')\mathsf{Pr}(Data = \mathbf{X}')} \tag{3.3}$$

$$= \frac{\mathsf{Pr}(Out = o|Data = \mathbf{X})}{\mathsf{Pr}(Out = o|Data = \mathbf{X}')} \cdot \frac{\mathsf{Pr}(Data = \mathbf{X})}{\mathsf{Pr}(Data = \mathbf{X}')} \tag{3.4}$$

$$\leq e^Z \cdot \frac{\mathsf{Pr}(Data = \mathbf{X})}{\mathsf{Pr}(Data = \mathbf{X}')} \tag{3.5}$$

In this way, the factor e^Z provides a multiplicative bound relative to the odds arising from the adversary's prior beliefs about the latent variable $Data$. When $Z = 0$, this bound implies no change whatsoever from the prior beliefs, and the framework of differential privacy views this as perfect privacy: the adversary has gained no information about $\mathbf{X}$ vs $\mathbf{X}'$ from the output. For larger values of Z, the odds are permitted to change more greatly, representing a greater leak of information. To make full use of this interpretation, it is important to note that differential privacy's guarantee is a bound over all pairs of neighboring data sets. This permits us to talk about differential privacy protecting individuals: if the adversary cannot effectively gain information about $\mathbf{X}$ vs $\mathbf{X}'$ for any possible choice for varying a single individual across these data sets, then they cannot effectively gain information about the single individual differing between the data sets. This statement holds regardless of the adversary's prior beliefs (potentially arising from arbitrary side information), and even under the extreme setting in which the adversary has full knowledge or control over all the remaining data in these data sets (the data which $\mathbf{X}$ and $\mathbf{X}'$ share).

A sufficient condition for (ε, δ)-DP is to have $Z \leq \varepsilon$ with probability at least $1 - \delta$ for any two neighboring data sets. The exact nature of the trade-off and semantics between ε and δ is subtle, and choosing them appropriately is difficult. For example, if n represents the number of individuals in a data set, a $(\varepsilon, \delta = 1/n)-$differentially private guarantee would also apply

to a mechanism that completely publishes the sensitive data of a single individual. It can be verified that for any pair of neighboring data set, such a mechanism would have the loss variable Z take on the value 0 with probability $\frac{n-1}{n}$. For any non-negative ε, this proposed mechanism has $\mathsf{Pr}(Z > \varepsilon) \leq 1 - \frac{n-1}{n} = \frac{1}{n}$, thus achieving a $(\varepsilon, \delta = 1/n)-$DP guarantee.

The previous example is intended to demonstrate just one facet of the nuances introduced by the second privacy parameter δ. When used haphazardly, these privacy guarantees can lose their meaning. Speaking briefly from a vague intuitive sense of designing a privacy-preserving mechanism, this simple mechanism that completely publishes one individual's data is in no way preserving privacy. Yet, that mechanism does satisfy the requirements of differential privacy for sufficiently large parameters. This work does not prescribe any particular appropriate choice for the parameters, it is the intent of this work to reinforce the importance of achieving privacy bounds with parameters as small as possible.

3.2 Rényi Differential Privacy

When faced with the risks and challenges of analyzing the impact of δ, other means of expressing parametric privacy guarantees can become attractive. The privacy definitions of 3.1 are not the only formulations for ensuring the privacy loss variable Z is "small" in a probabilistic sense. One alternative is to instead bound the Rényi divergence of $\mathscr{A}(\mathbf{X})$ and $\mathscr{A}(\mathbf{X}')$ Mironov [2017].

Definition 4. Rényi Divergence.

The Rényi divergence of order α between the two distributions P and Q is defined as

$$\mathsf{R}_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \int P(o)^\alpha Q(o)^{1-\alpha} do. \tag{3.6}$$

We note that a slightly different notation will be used in Chapter 5 which replaces α with λ, since much of the analysis will deal with distributions with their own α and β parameters.

The Rényi divergence can be viewed as a generalization of KL divergence. Examining the limiting behavior of the divergence with respect to its order offers useful insights into this divergence. When α approaches ∞, the integral places increasingly more importance on the maximal values of $P(o)/Q(o)$ achieved for any output o (in a way highly reminiscent of how the L_∞ norm of the L_p family places its focus on the maximal coordinate of its input). When $\alpha = \infty$, we can see $\mathsf{R}_\infty(P||Q)$ gives rise to the *max divergence* which equals $\sup_o \log(P(o)/Q(o))$. Turning our attention to smaller values for the order, we can examine the behavior as α approaches 1. In this regime, the divergence approaches a simple expected value of the ratio $\log P(o)/Q(o)$. In fact, in the limit, we can identify the case of $\alpha = 1$ with the KL divergence, $\mathsf{R}_1(P||Q) = KL(P||Q)$.

This leads us to Rényi Differential Privacy, a flexible privacy notion that uses its additional parameter to smoothly cover the range of intermediate behaviors between average-case KL divergences and worst-case max divergences.

Definition 5. Rényi Differential Privacy (RDP).

A randomized mechanism $\mathscr{A}(\mathbf{X})$ is said to be $(\alpha, \varepsilon)-$Rényi differentially private if for any neighboring data sets $\mathbf{X}$ and $\mathbf{X}'$ we have $\mathsf{R}_\alpha(\mathscr{A}(\mathbf{X})||\mathscr{A}(\mathbf{X}')) \leq \varepsilon$.

The choice of α in the bound required for RDP controls the balance between bounding extreme values Z versus bounding the average value of Z. One can consider a mechanism's privacy as being quantified by the entire curve of ε values associated with each order α, but the results of [Mironov, 2017] show that almost identical bounds can be achieved when this curve is known at only a finite collection of possible α values. Therefore RDP bounds can still be useful even when the relationship between α and ε is not easily written down.

RDP also enjoys useful properties, analogous to those we discussed about differential privacy. The composition result is mathematically convenient, and behaves much tighter than the simple composition result for (ε, δ)-DP presented earlier.

Observation 3. *Robustness to Post-Processing for RDP*

For any mechanism $\mathscr{A}$ that satisfies (α, ε)-RDP, and any randomized function f that operates on the outputs of $\mathscr{A}$, the mechanism formed by releasing $f(\mathscr{A}(\mathbf{X}))$ for a data set $\mathbf{X}$ also satisfies (α, ε)-RDP.

Observation 4. *Robustness to Composition for RDP*

For any mechanism $\mathscr{A}$ that satisfies $(\alpha, \varepsilon_{\mathscr{A}})$-RDP, and any other mechanism $\mathscr{B}$ that operates on the same data as $\mathscr{A}$ and satisfies $(\alpha, \varepsilon_{\mathscr{B}})$-DP, the mechanism formed by releasing the tuple $(\mathscr{A}(\mathbf{X}), \mathscr{B}(\mathbf{X}))$ for a data set $\mathbf{X}$ satisfies $(\alpha, \varepsilon_{\mathscr{A}} + \varepsilon_{\mathscr{B}})$-DP.

This bound behaves in a nicer fashion than the one for (ε, δ)-DP: the second privacy parameter α does not degrade under composition in RDP.

RDP behaves in a more mathematically convenient fashion than DP, especially under composition. The Rényi divergence is more naturally suited for tracking the behavior of the privacy loss variable. However, the parameter α is much harder to interpret than δ. It is a common practice to perform a privacy analysis in the RDP framework, get the overall privacy guarantee of a complicated proposed mechanism, and then as a final step convert the RDP guarantee into a (ε, δ)-DP guarantee. This final bound is a bit more digestible by human analysts, since the δ parameter has a natural connection to a bound on the probability of the privacy loss variable being large.

3.3 Exponential Family Distributions

A large part of the contributions of this thesis are focused on exponential family distributions, which we shall introduce here.

Exponential families form a broad class of probability distribution families, including many of the most commonly used distributions in machine learning. Each family contains a multitude of distributions, each sharing a similar structure and parameterizations. We begin with the most general description of exponential families, and present a specific concrete example later.

An exponential family's defining trait is how the family can be indexed by a parameter $\theta \in \mathbb{R}^d$. Given this parameter, each distribution in the family can be written in the following form for some choice of functions $h : \mathscr{X} \to \mathbb{R}$, $S : \mathscr{X} \to \mathbb{R}^d$, and $A : \Theta \to \mathbb{R}$ shared by all the distributions in the family:

$$\Pr(x_1, \ldots, x_n | \theta) = (\prod_{i=1}^{n} h(x_i)) \exp\left(\sum_{i=1}^{n} S(x_i)\right) \cdot \theta - n \cdot A(\theta) \; . \tag{3.7}$$

Of particular importance is S, the sufficient statistics function, and A, the log-partition function of this family. These two functions give raise to many of the useful behaviors and properties of these families. The work in this thesis will make use of the following definitions to ensure the families behave as intended.

Definition 6. The natural parameterization of an exponential family is the one that indexes the distributions of the family by the vector θ that appears in the inner product of equation (3.7).

Definition 7. An exponential family is minimal if the coordinates of the function S are not linearly dependent for all $x \in \mathscr{X}$.

When given a non-minimal exponential family, we can always find an alternative minimal natural parameterization for that family. When we have a minimal exponential family, we can then directly find a family of conjugate prior distributions. When the data arises from an exponential family distribution, and our prior beliefs distribution over the parameter of that distribution comes from the conjugate prior family, then our posterior beliefs over the parameter of the data distribution will remain within the conjugate prior family.

A minimal exponential family will always have a minimal conjugate prior family. This conjugate prior family is also an exponential family, and it satisfies the property that the posterior distribution formed after observing data is also within the same family. In fact, when we have the minimal natural representations, we can get the following formula for parameterizing our

prior and posterior beliefs in terms of η:

$$\Pr(\theta|\eta) = \exp T(\theta) \cdot \eta - C(\eta). \tag{3.8}$$

The sufficient statistics of θ can be written as $T(\theta) = (\theta, -A(\theta))$ and $\Pr(\theta|\eta_0, x_1, \ldots, x_n)$ $= \Pr(\theta|\eta')$ where $\eta' = \eta_0 + \sum_{i=1}^{n}(S(x_i), 1)$. This conjugate prior family is itself an exponential family of distributions.

Beta-Bernoulli System.

A specific example of an exponential family that we will be interested in is the Beta-Bernoulli system, where an individual's data is a single i.i.d. bit modeled as a Bernoulli variable with parameter ρ, along with a Beta conjugate prior. $\Pr(x|\rho) = \rho^x(1-\rho)^{1-x}$.

The Bernoulli distribution can be written in the form of equation (5.3) by letting $h(x) = 1$, $S(x) = x$, $\theta = \log(\frac{\rho}{1-\rho})$, and $A(\theta) = \log(1 + \exp\theta) = -\log(1-\rho)$. The Beta distribution with the usual parameters α_0, β_0 will be parameterized by $\eta_0 = (\eta_0^{(1)}, \eta_0^{(2)}) = (\alpha_0, \alpha_0 + \beta_0)$ in accordance equation (3.8). This system satisfies the properties we require, as this natural parameterization is minimal. In this system, $C(\eta) = \Gamma(\eta^{(1)}) + \Gamma(\eta^{(2)} - \eta^{(1)}) - \Gamma(\eta^{(2)})$.

Chapter 4

Private Posterior Sampling of Exponential Families via Noisy Statistics

4.1 Introduction

Probabilistic models trained via Bayesian inference are widely and successfully used in application domains where privacy is invaluable, from text analysis [Blei et al., 2003, Goldwater and Griffiths, 2007], to personalization [Salakhutdinov and Mnih, 2008], to medical informatics [Husmeier et al., 2006], to MOOCs [Piech et al., 2013]. In these applications, data scientists must carefully balance the benefits and potential insights from data analysis against the privacy concerns of the individuals whose data are being studied [Daries et al., 2014].

Dwork et al. [2006b] placed the notion of privacy-preserving data analysis on a solid foundation by introducing *differential privacy* [Dwork and Roth, 2014], an algorithmic formulation of privacy which is a gold standard for privacy-preserving data-driven algorithms. Differential privacy measures the privacy "cost" of an algorithm. When designing privacy-preserving methods, the goal is to achieve a good trade-off between privacy and utility, which ideally improves with the amount of available data.

As observed by Dimitrakakis et al. [2014] and Wang et al. [2015b], Bayesian posterior sampling behaves synergistically with differential privacy because it automatically provides a degree of differential privacy under certain conditions. However, there are substantial gaps between this elegant theory and the practical reality of Bayesian data analysis. Privacy-preserving

posterior sampling is hampered by data inefficiency, as measured by asymptotic relative efficiency (ARE). In practice, it generally requires artificially selected constraints on the spaces of parameters as well as data points. Its privacy properties are also not typically guaranteed for approximate inference.

This paper identifies these gaps between theory and practice, and begins to mend them via an extremely simple alternative technique based on the workhorse of differential privacy, the Laplace mechanism [Dwork et al., 2006b]. Our approach is equivalent to a generalization of Zhang et al. [2016]'s recently and independently proposed algorithm for beta-Bernoulli systems. We provide a theoretical analysis and empirical validation of the advantages of the proposed method. We extend both our method and Dimitrakakis et al. [2014], Wang et al. [2015b]'s *one posterior sample (OPS)* method to the case of approximate inference with privacy-preserving MCMC. Finally, we demonstrate the practical applicability of this technique by showing how to use a privacy-preserving HMM model to analyze sensitive military records from the Iraq and Afghanistan wars leaked by the Wikileaks organization. Our primary contributions are as follows:

- We analyze the privacy cost of posterior sampling for exponential family posteriors via OPS.
- We explore a simple Laplace mechanism alternative to OPS for exponential families.
- Under weak conditions we establish the consistency of the Laplace mechanism approach and its data efficiency advantages over OPS.
- We extend the OPS and Laplace mechanism methods to approximate inference via MCMC.
- We demonstrate the practical implications with a case study on sensitive military records.

4.2 Setup

We begin by discussing preliminaries on differential privacy and its application to Bayesian inference. Our novel contributions will begin in Section 4.3.1.

4.2.1 Differential Privacy

Differential privacy is a formal notion of the privacy of data-driven algorithms. For an algorithm to be differentially private the probabilities of the outputs of the algorithms may not change much when one individual's data point is modified, thereby revealing little information about any one individual's data. More precisely, a randomized algorithm $\mathscr{M}(\mathbf{X})$ is said to be (ε, δ)-differentially private if

$$\mathsf{Pr}(\mathscr{M}(\mathbf{X}) \in \mathscr{S}) \leq \exp(\varepsilon)\mathsf{Pr}(\mathscr{M}(\mathbf{X}') \in \mathscr{S}) + \delta \tag{4.1}$$

for all measurable subsets $\mathscr{S}$ of the range of $\mathscr{M}$ and for all datasets $\mathbf{X}$, $\mathbf{X}'$ differing by a single entry [Dwork and Roth, 2014]. If $\delta = 0$, the algorithm is said to be ε-differentially private.

The Laplace Mechanism

One straightforward method for obtaining ε-differential privacy, known as the *Laplace mechanism* [Dwork et al., 2006b], adds Laplace noise to the revealed information, where the amount of noise depends on ε, and a quantifiable notion of the sensitivity to changes in the database. Specifically, the $L1$ sensitivity $\triangle h$ for function h is defined as

$$\triangle h = \max_{\mathbf{X},\mathbf{X}'} \|h(\mathbf{X}) - h(\mathbf{X}')\|_1 \tag{4.2}$$

for all datasets $\mathbf{X}$, $\mathbf{X}'$ differing in at most one element. The Laplace mechanism adds noise via

$$\mathscr{M}_L(\mathbf{X}, h, \varepsilon) = h(\mathbf{X}) + (Y_1, Y_2, \ldots, Y_d) \,, \tag{4.3}$$
$$Y_j \sim \text{Laplace}(\triangle h/\varepsilon), \forall j \in \{1, 2, \ldots, d\} \,,$$

where d is the dimensionality of the range of h. The $\mathscr{M}_L(\mathbf{X}, h, \varepsilon)$ mechanism is ε-differentially private.

The Exponential Mechanism

The exponential mechanism [McSherry and Talwar, 2007] aims to output responses of high utility while maintaining privacy. Given a utility function $u(\mathbf{X}, \mathbf{r})$ that maps database $\mathbf{X}$/output $\mathbf{r}$ pairs to a real-valued score, the exponential mechanism $\mathscr{M}_E(\mathbf{X}, u, \varepsilon)$ produces random outputs via

$$\mathsf{Pr}(\mathscr{M}_E(\mathbf{X}, u, \varepsilon) = \mathbf{r}) \propto \exp\left(\frac{\varepsilon u(\mathbf{X}, \mathbf{r})}{2\triangle u}\right) , \tag{4.4}$$

where the sensitivity of the utility function is

$$\triangle u \triangleq \max_{r, (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})} \|u(\mathbf{X}^{(1)}, r) - u(\mathbf{X}^{(2)}, r)\|_1 \;\; , \tag{4.5}$$

in which $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ are pairs of databases that differ in only one element.

Composition Theorems

A key property of differential privacy is that it holds under composition, via an additive accumulation.

Theorem 1. *If $\mathscr{M}_1$ is $(\varepsilon_1, \delta_1)$-differentially private, and $\mathscr{M}_2$ is $(\varepsilon_2, \delta_2)$-differentially private, then $\mathscr{M}_{1,2}(\mathbf{X}) = (\mathscr{M}_1(\mathbf{X}), \mathscr{M}_2(\mathbf{X}))$ is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-differentially private.*

This allows us to view the total ε and δ of our procedure as a privacy "budget" that we spend across the operations of our analysis. There also exists an "advanced composition"

theorem which provides privacy guarantees in an adversarial adaptive scenario called k-fold composition, and also allows an analyst to trade an increased δ for a smaller ε in this scenario [Dwork et al., 2010]. Differential privacy is also immune to data-independent post-processing.

4.2.2 Privacy and Bayesian Inference

Suppose we would like a differentially private draw of parameters and latent variables of interest θ from the posterior $\Pr(\theta|\mathbf{X})$, where $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is the private dataset. We can accomplish this by interpreting posterior sampling as an instance of the exponential mechanism with utility function $u(\mathbf{X}, \theta) = \log \Pr(\theta, \mathbf{X})$, i.e. the log joint probability of the chosen θ assignment and the dataset $\mathbf{X}$ [Wang et al., 2015b]. We then draw θ via

$$f(\theta; \mathbf{X}, \varepsilon) \propto \exp\left(\frac{\varepsilon \log \Pr(\theta, \mathbf{X})}{2\triangle \log \Pr(\theta, \mathbf{X})}\right) = \Pr(\theta, \mathbf{X})^{\frac{\varepsilon}{2\triangle \log \Pr(\theta, \mathbf{X})}} \tag{4.6}$$

where the sensitivity is

$$\triangle \log \Pr(\theta, \mathbf{X}) \triangleq \max_{\theta, (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})} \| \log \Pr(\theta, \mathbf{X}^{(1)}) - \log \Pr(\theta, \mathbf{X}^{(2)}) \|_1 \tag{4.7}$$

in which $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ differ in one element. If the data points are conditionally independent given θ,

$$\log \Pr(\theta, \mathbf{X}) = \log \Pr(\theta) + \sum_{i=1}^{N} \log \Pr(\mathbf{x}_i|\theta) \,, \tag{4.8}$$

where $\Pr(\theta)$ is the prior and $\Pr(\mathbf{x}_i|\theta)$ is the likelihood term for data point $\mathbf{x}_i$. Since the prior does not depend on the data, and each data point is associated with a single log-likelihood term $\log \Pr(\mathbf{x}_i|\theta)$ in $\log \Pr(\theta, \mathbf{X})$, from the above two equations we have

$$\triangle \log \Pr(\theta, \mathbf{X}) = \max_{\mathbf{x}, \mathbf{x}', \theta} |\log \Pr(\mathbf{x}'|\theta) - \log \Pr(\mathbf{x}|\theta)| \,. \tag{4.9}$$

This gives us the privacy cost of posterior sampling:

Theorem 2. *If* $\max_{\mathbf{x},\mathbf{x}'\in\chi,\theta\in\Theta} |\log \mathsf{Pr}(\mathbf{x}'|\theta) - \log \mathsf{Pr}(\mathbf{x}|\theta)| \leq C$*, releasing one sample from the posterior distribution* $\mathsf{Pr}(\theta|\mathbf{X})$ *with any prior is* $2C$*-differentially private.*

Wang et al. [2015b] derived this form of the result from first principles, while noting that the exponential mechanism can be used, as we do here. Although they do not explicitly state the theorem, they implicitly use it to show two noteworthy special cases, referred to as the *One Posterior Sample (OPS)* procedure. We state the first of these cases:

Theorem 3. *If* $\max_{\mathbf{x}\in\chi,\theta\in\Theta} |\log \mathsf{Pr}(\mathbf{x}|\theta)| \leq B$*, releasing one sample from the posterior distribution* $\mathsf{Pr}(\theta|\mathbf{X})$ *with any prior is* $4B$*-differentially private.*

This follows directly from Theorem 2, since if $|\log \mathsf{Pr}(\mathbf{x}|\theta)| \leq B$, $C = \triangle \log \mathsf{Pr}(\theta, \mathbf{X}) = 2B$.

Under the exponential mechanism, ε provides an adjustable knob trading between privacy and fidelity. When $\varepsilon = 0$, the procedure samples from a uniform distribution, giving away no information about $\mathbf{X}$. When $\varepsilon = 2\triangle \log \mathsf{Pr}(\theta, \mathbf{X})$, the procedure reduces to sampling θ from the posterior $\mathsf{Pr}(\theta|\mathbf{X}) \propto \mathsf{Pr}(\theta, \mathbf{X})$. As ε approaches infinity the procedure becomes increasingly likely to sample the θ assignment with the highest posterior probability. Assuming that our goal is to sample rather than to find a mode, we would cap ε at $2\triangle \log \mathsf{Pr}(\theta, \mathbf{X})$ in the above procedure in order to correctly sample from the true posterior. More generally, if our privacy budget is ε', and $\varepsilon' \geq 2q\triangle \log \mathsf{Pr}(\theta, \mathbf{X})$, for integer q, we can draw q posterior samples within our budget.

As observed by Huang and Kannan [2012], the exponential mechanism can be understood via statistical mechanics. We can write it as a Boltzmann distribution (a.k.a. a Gibbs measure)

$$f(\theta; \mathbf{x}, \varepsilon) \propto \exp\left(\frac{-E(\theta)}{T}\right), T = \frac{2\triangle u(\mathbf{X}, \theta)}{\varepsilon}, \tag{4.10}$$

where $E(\theta) = -u(\mathbf{X}, \theta) = -\log \mathsf{Pr}(\theta, \mathbf{X})$ is the energy of state θ in a physical system, and T is the temperature of the system (in units such that Boltzmann's constant is one). Reducing ε

Table 4.1. Comparison of the properties of the two methods for private Bayesian inference.

Mechanism	Sensitivity	$S(\mathbf{X})$ is	Release	ARE	Pay Gibbs cost
Laplace	$\sup_{\mathbf{X},\mathbf{X}'} \|\| \sum_{i=1}^{N} S(\mathbf{x}'^{(i)}) - \sum_{i=1}^{N} S(\mathbf{x}^{(i)}) \|\|_1$	Noised	Statistics	1	Once
Exponential (OPS)	$\sup_{\mathbf{x},\mathbf{x}'\in\chi,\theta\in\Theta} \|\theta^\intercal \Big(S(\mathbf{x}') - S(\mathbf{x})\Big) + \log h(\mathbf{x}') - \log h(\mathbf{x})\|$	Rescaled	One Sample	$1+T$	Per update (unless converged)

corresponds to increasing the temperature, which can be understood as altering the distribution such that a Markov chain moves through the state space more rapidly.

4.3 Privacy for Exponential Families: Exponential vs. Laplace Mechanisms

By analyzing the privacy cost of sampling from exponential family posteriors in the general case we can recover the privacy properties of many standard distributions. These results can be applied to full posterior sampling, when feasible, or to Gibbs sampling updates, as we discuss in Section 4.4. In this section we analyze the privacy cost of sampling from exponential family posterior distributions exactly (or at an appropriate temperature) via the exponential mechanism, following Dimitrakakis et al. [2014] and Wang et al. [2015b], and via a method based on the Laplace mechanism, which is a generalization of Zhang et al. [2016]. The properties of the two methods are compared in Table 4.1.

4.3.1 The Exponential Mechanism

Consider exponential family models with likelihood

$$\Pr(\mathbf{x}|\theta) = h(\mathbf{x})g(\theta)\exp\left(\theta^\intercal S(\mathbf{x})\right),$$

where $S(\mathbf{x})$ is a vector of sufficient statistics for data point $\mathbf{x}$, and θ is a vector of natural parameters. For N i.i.d. data points, we have

$$\Pr(\mathbf{X}|\theta) = \Big(\prod_{i=1}^{N} h(\mathbf{x}^{(i)})\Big) g(\theta)^N \exp\left(\theta^\intercal \sum_{i=1}^{N} S(\mathbf{x}^{(i)})\right).$$

Further suppose that we have a conjugate prior which is also an exponential family distribution,

$$\mathsf{Pr}(\theta|\chi,\alpha) = f(\chi,\alpha)g(\theta)^{\alpha}\exp\left(\alpha\theta^{\mathsf{T}}\chi\right),$$

where α is a scalar, the number of prior "pseudo-counts," and χ is a parameter vector. The posterior is proportional to the prior times the likelihood,

$$\mathsf{Pr}(\theta|\mathbf{X},\chi,\alpha) \propto g(\theta)^{N+\alpha}\exp\left(\theta^{\mathsf{T}}(\sum_{i=1}^{N} S(\mathbf{x}^{(i)}) + \alpha\chi)\right). \tag{4.11}$$

To compute the sensitivity of the posterior, we have

$$|\log Pr(\mathbf{x}'|\theta) - \log \mathsf{Pr}(\mathbf{x}|\theta)| = |\theta^{\mathsf{T}}\left(S(\mathbf{x}') - S(\mathbf{x})\right) + \log h(\mathbf{x}') - \log h(\mathbf{x})|\,.$$

From Equation 4.9, we obtain

$$\triangle \log \mathsf{Pr}(\theta,\mathbf{X}) = \sup_{\mathbf{x},\mathbf{x}'\in\chi,\theta\in\Theta} |\theta^{\mathsf{T}}\left(S(\mathbf{x}') - S(\mathbf{x})\right) + \log h(\mathbf{x}') - \log h(\mathbf{x})|\,. \tag{4.12}$$

A posterior sample at temperature T,

$$\mathsf{Pr}_T(\theta|\mathbf{X},\chi,\alpha) \propto g(\theta)^{\frac{N+\alpha}{T}}\exp\left(\theta^{\mathsf{T}}\frac{\sum_{i=1}^{N} S(\mathbf{x}^{(i)}) + \alpha\chi}{T}\right), T = \frac{2\triangle \log p(\theta,X)}{\varepsilon}, \tag{4.13}$$

has privacy cost ε, by the exponential mechanism. As an example, consider a beta-Bernoulli model,

$$\begin{aligned}
\mathsf{Pr}(p|\alpha,\beta) &= \frac{1}{B(\alpha,\beta)}p^{\alpha-1}(1-p)^{\beta-1} \\
&= \frac{1}{B(\alpha,\beta)}\exp((\alpha-1)\log p + (\beta-1)\log(1-p)) \\
\mathsf{Pr}(x|p) &= p^{x}(1-p)^{1-x} = \exp(x\log p + (1-x)\log(1-p))
\end{aligned}$$

where $B(\alpha,\beta)$ is the beta function. Given N binary-valued data points $\mathbf{X} = x^{(1)},\ldots,x^{(N)}$ from the Bernoulli distribution, the posterior is

$$Pr(p|\mathbf{X},\alpha,\beta) \propto \exp\Big(\big(n_+ + \alpha - 1\big)\log p + \big(n_- + \beta - 1\big)\log(1-p)\Big)$$
$$n_+ = \sum_{i=1}^{N} x^{(i)}, \qquad n_- = \sum_{i=1}^{N}(1-x^{(i)}) \; .$$

The sufficient statistics for each data point are $S(x) = [x, 1-x]^\intercal$. The natural parameters for the posterior are $\theta = [\log p, \log(1-p)]^\intercal$, and $h(x) = 0$. The exponential mechanism sensitivity for a *truncated* version of this model, where $a_0 \leq p \leq 1-a_0$, can be computed from Equation 4.12,

$$\begin{aligned}\triangle \log \mathsf{Pr}(\theta,\mathbf{X}) &= \sup_{x,x'\in\{0,1\},p\in[a_0,1-a_0]} \big|x\log p + (1-x)\log(1-p) - \big(x'\log p + (1-x')\log(1-p)\big)\big| \\ &= -\log a_0 + \log(1-a_0) \; . \end{aligned} \tag{4.14}$$

Note that if $a_0 = 0$, corresponding to a standard untruncated beta distribution, the sensitivity is unbounded. This makes intuitive sense because some datasets are impossible if $p = 0$ or $p = 1$, which violates differential privacy.

4.3.2 The Laplace Mechanism

One limitation of the exponential mechanism / OPS approach to private Bayesian inference is that the temperature T of the approximate posterior is fixed for any ε that we are willing to pay, regardless of the number of data points N (Equation 4.10). While the posterior becomes more accurate as N increases, and the OPS approximation becomes more accurate by proxy, the OPS approximation remains a factor of T flatter than the posterior at N data points. This is not simply a limitation of the analysis. An adversary can choose data such that the dataset-specific privacy cost of posterior sampling approaches the worst case given by the exponential mechanism as N increases, by causing the posterior to concentrate on the worst-case θ.

Here, we provide a simple Laplace mechanism alternative for exponential family posteriors, which becomes increasingly faithful to the true posterior with N data points, as N increases,

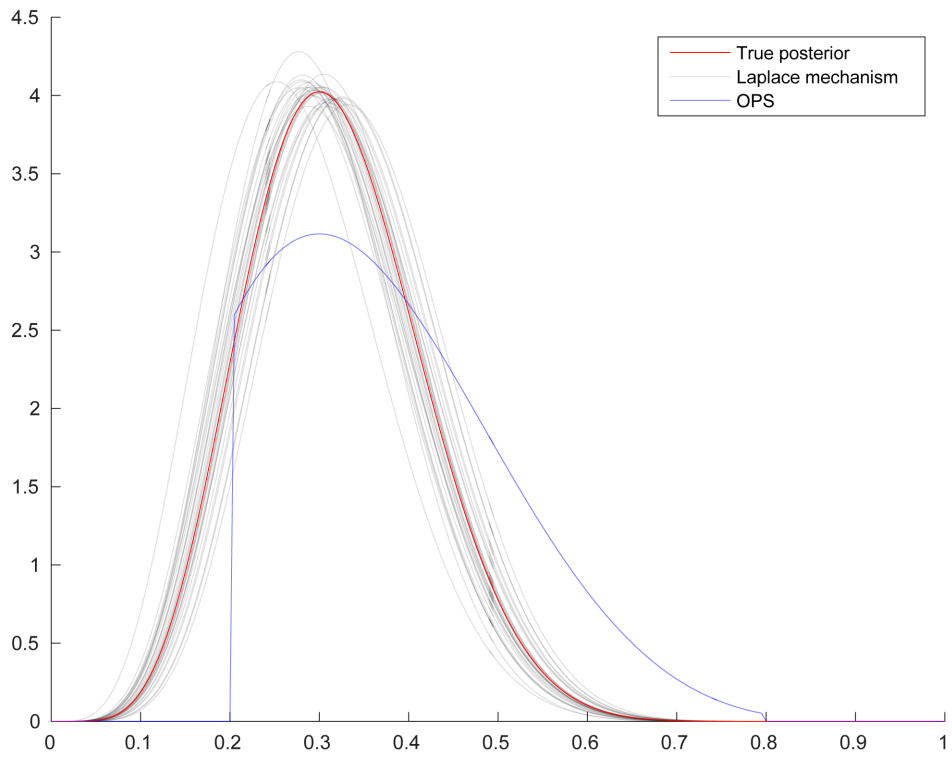

Figure 4.1. Privacy-preserving approximate posteriors for a truncated beta-Bernoulli model ($\varepsilon = 1$, the true parameter $p = 0.3$, truncation point $a_0 = 0.2$, and number of observations $N = 20$). For the Laplace mechanism, 30 privatizing draws are rendered.

for any fixed privacy cost ε, under general assumptions. The approach is based on the observation that for exponential family posteriors, as in Equation 4.11, the data interacts with the distribution only through the aggregate sufficient statistics, $S(\mathbf{X}) = \sum_{i=1}^{N} S(\mathbf{x}^{(i)})$. If we release privatized versions of these statistics we can use them to perform any further operations that we'd like, including drawing samples, computing moments and quantiles, and so on. This can straightforwardly be accomplished via the Laplace mechanism:

$$\hat{S}(\mathbf{X}) = \text{proj}(S(\mathbf{X}) + (Y_1, Y_2, \ldots, Y_d)) \,, \tag{4.15}$$
$$Y_j \sim \text{Laplace}(\triangle S(\mathbf{X})/\varepsilon), \forall j \in \{1, 2, \ldots, d\} \,,$$

where proj$(\cdot)$ is a projection onto the space of sufficient statistics, if the Laplace noise takes it out of this region. For example, if the statistics are counts, the projection ensures that they are non-negative. The L_1 sensitivity of the aggregate statistics is

$$\begin{aligned} \triangle S(\mathbf{X}) &= \sup_{\mathbf{X},\mathbf{X}'} \| \sum_{i=1}^{N} S(\mathbf{x}'^{(i)}) - \sum_{i=1}^{N} S(\mathbf{x}^{(i)}) \|_1 \\ &= \sup_{\mathbf{x},\mathbf{x}'} \|S(\mathbf{x}') - S(\mathbf{x})\|_1 \ , \end{aligned} \tag{4.16}$$

where $\mathbf{X}$, $\mathbf{X}'$ differ in at most one element. Note that perturbing the sufficient statistics is equivalent to perturbing the parameters, which was recently and independently proposed by Zhang et al. [2016] for beta-Bernoulli models such as Bernoulli naive Bayes.

A comparison of Equations 4.16 and 4.12 reveals that the L1 sensitivity and exponential mechanism sensitivities are closely related. The L1 sensitivity is generally easier to control as it does not involve θ or $h(\mathbf{x})$ but otherwise involves similar terms to the exponential mechanism sensitivity. For example, in the beta posterior case, where $S(\mathbf{x}) = [x, 1-x]$ is a binary indicator vector, the L1 sensitivity is 2. This should be contrasted to the exponential mechanism sensitivity of Equation 4.14, which depends heavily on the truncation point, and is unbounded for a standard untruncated beta distribution. The L1 sensitivity is fixed regardless of the number of data points N, and so the amount of Laplace noise to add becomes smaller relative to the total $S(\mathbf{X})$ as N increases.

Figure 4.1 illustrates the differences in behavior between the two privacy-preserving Bayesian inference algorithms for a beta distribution posterior with Bernoulli observations. The OPS estimator requires the distribution be truncated, here at $a_0 = 0.2$. This controls the exponential mechanism sensitivity, which determines the temperature T of the distribution, i.e. the extent to which the distribution is flattened, for a given ε. Here, $T = 2.7$. In contrast, the Laplace mechanism achieves privacy by adding noise to the sufficient statistics, which in this case are the pseudo-counts of successes and failures for the posterior distribution. In Figure 4.2

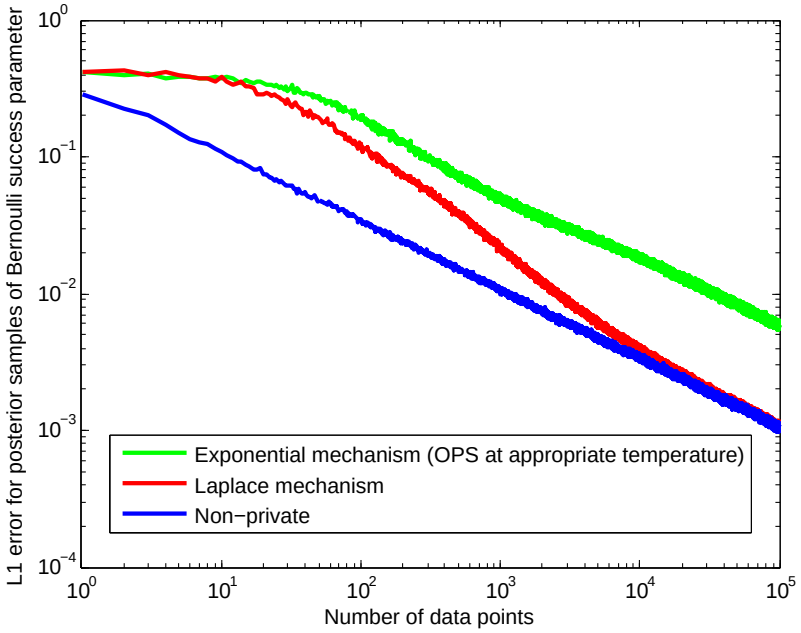

Figure 4.2. L1 error for private approximate samples from a beta posterior over a Bernoulli success parameter p, as a function of the number of Bernoulli(p) observations, averaged over 1000 repeats. The true parameter was $p = 0.1$, the exponential mechanism posterior was truncated at $a_0 = 0.05$, and $\varepsilon = 0.1$.

we illustrate the fidelity benefits of posterior sampling based on the Laplace mechanism instead of the exponential mechanism as the amount of data increases. In this case the exponential mechanism performs better than the Laplace mechanism only when the number of data points is very small (approximately $N = 10$), and is quickly overtaken by the Laplace mechanism sampling procedure. As N increases the accuracy of sampling from the Laplace mechanism's approximate posterior converges to the performance of samples from the true posterior at the current number of observations N, while the exponential mechanism behaves similarly to the posterior with fewer than N observations. We show this formally in the next subsection.

4.3.3 Theoretical Results

First, we show that the Laplace mechanism approximation of exponential family posteriors approaches the true posterior distribution *evaluated at N data points*. Proofs are given in Section 4.6.

Lemma 1. *For a minimal exponential family given a conjugate prior, where the posterior takes the form* $\mathsf{Pr}(\theta|\mathbf{X},\chi,\alpha) \propto g(\theta)^{n+\alpha}\exp\left(\theta^{\mathsf{T}}\left(\sum_{i=1}^{n} S(\mathbf{x}^{(i)}) + \alpha\chi\right)\right)$*, where* $\mathsf{Pr}(\theta|\eta)$ *denotes this posterior with a natural parameter vector* η*, if there exists a* $\delta > 0$ *such that these assumptions are met:*

1. *The data* $\mathbf{X}$ *comes i.i.d. from a minimal exponential family distribution with natural parameter* $\theta_0 \in \Theta$
2. θ_0 *is in the interior of* Θ
3. *The function* $A(\theta)$ *has all derivatives for* θ *in the interior of* Θ
4. $cov_{\mathsf{Pr}(\mathbf{x}|\theta)}(S(\mathbf{x})))$ *is finite for* $\theta \in \mathscr{B}(\theta_0,\delta)$
5. $\exists w > 0$ *s.t.* $\det(cov_{\mathsf{Pr}(\mathbf{x}|\theta)}(S(\mathbf{x})))) > w$ *for* $\theta \in \mathscr{B}(\theta_0,\delta)$
6. *The prior* $\mathsf{Pr}(\theta|\chi,\alpha)$ *is integrable and has support on a neighborhood of* θ^*

then for any mechanism generating a perturbed posterior $\tilde{p}_N = \mathsf{Pr}(\theta|\eta_N + \gamma)$ *against a noiseless posterior* $p_N = \mathsf{Pr}(\theta|\eta_N)$ *where* γ *comes from a distribution that does not depend on the number of data observations N and has finite covariance, this limit holds:*

$$\lim_{N\to\infty} E[KL(\tilde{p}_N||p_N)] = 0\,.$$

Corollary 2. *The Laplace mechanism on an exponential family satisfies the noise distribution requirements of Lemma 1 when the sensitivity of the sufficient statistics is finite and either the exponential family is minimal, or if the exponential family parameters* θ *are identifiable.*

These assumptions correspond to the data coming from a distribution where the Laplace regularity assumptions hold and the posterior satisfies the asymptotic normality given by the Bernstein-von Mises theorem. For example, in the beta-Bernoulli setting, these assumptions hold as long as the success parameter p is in the open interval $(0,1)$. For $p = 0$ or 1, the relevant parameter is not in the interior of Θ, and the result does not apply. In the setting of learning a

normal distribution's mean μ where the variance $\sigma^2 > 0$ is known, the assumptions of Lemma 1 always hold, as the natural parameter space is an open set. However, Corollary 2 does not apply in this setting because the sensitivity is infinite (unless bounds are placed on the data). Our efficiency result, in Theorem 4, follows from Lemma 1 and the Bernstein-von Mises theorem.

Theorem 4. *Under the assumptions of Lemma 1, the Laplace mechanism has an asymptotic posterior of* $\mathcal{N}(\theta_0, 2\mathbb{I}^{-1}/N)$ *from which drawing a single sample has an asymptotic relative efficiency of 2 in estimating* θ_0*, where* $\mathbb{I}$ *is the Fisher information at* θ_0*.*

Above, the asymptotic posterior refers to the normal distribution, whose variance depends on N, that the posterior distribution approaches as N increases. This ARE result should be contrasted to that of the exponential mechanism [Wang et al., 2015b].

Theorem 5. *The exponential mechanism applied to the exponential family with temperature parameter* $T \geq 1$ *has an asymptotic posterior of* $\mathcal{N}(\theta^*, (1+T)\mathbb{I}^{-1}/N)$ *and a single sample has an asymptotic relative efficiency of* $(1+T)$ *in estimating* θ^**, where* $\mathbb{I}$ *is the Fisher information at* θ^**.*

Here, the ARE represents the ratio between the variance of the estimator and the optimal variance $\mathbb{I}^{-1}/N$ achieved by the posterior mean in the limit. Sampling from the posterior itself has an ARE of 2, due to the stochasticity of sampling, which the Laplace mechanism approach matches. These theoretical results provide an explanation for the difference in the behavior of these two methods as N increases seen in Figure 4.2. The Laplace mechanism will eventually approach the true posterior and the impact of privacy on accuracy will diminish when the data size increases. However, for the exponential mechanism with $T > 1$, the ratio of variances between the sampled posterior and the true posterior given N data points approaches $(1+T)/2$, making the sampled posterior more spread out than the true posterior even as N grows large.

So far we have compared the ARE values for *sampling*, as an apples-to-apples comparison. In reality, the Laplace mechanism has a further advantage as it releases a full posterior with

privatized parameters, while the exponential mechanism can only release a finite number of samples with a finite ε, which we discuss in Remark 1.

Remark 1. *Under the the assumptions of Lemma 1, by using the full privatized posterior instead of just a sample from it, the Laplace mechanism can release the privatized posterior's mean, which has an asymptotic relative efficiency of 1 in estimating* θ^*.

4.4 Private Gibbs Sampling

We now shift our discussion to the case of approximate Bayesian inference. While the analysis of Dimitrakakis et al. [2014] and Wang et al. [2015b] shows that posterior sampling is differentially private under certain conditions, exact sampling is not in general tractable. It does not directly follow that approximate sampling algorithms such as MCMC are also differentially private, or private at the same privacy level. Wang et al. [2015b] give two results towards understanding the privacy properties of approximate sampling algorithms. First, they show that if the approximate sampler is "close" to the true distribution in a certain sense, then the privacy cost will be close to that of a true posterior sample:

Proposition 3. *If procedure* $\mathscr{A}$ *which produces samples from distribution* $P_{\mathbf{X}}$ *is* ε*-differentially private, then any approximate sampling procedures* $\mathscr{A}'$ *that produces a sample from* $P'_{\mathbf{X}}$ *such that* $\|P_{\mathbf{X}} - P'_{\mathbf{X}}\|_1 \leq \delta$ *for any* $\mathbf{X}$ *is* $(\varepsilon, (1+\exp(\varepsilon)\delta)$*-differentially private.*

Unfortunately, it is not in general feasible to verify the convergence of an MCMC algorithm, and so this criterion is not generally verifiable in practice. In their second result, Wang et al. study the privacy properties of stochastic gradient MCMC algorithms, including stochastic gradient Langevin dynamics (SGLD) [Welling and Teh, 2011] and its extensions. SGLD is a stochastic gradient method with noise injected in the gradient updates which converges in distribution to the target posterior.

In this section we study the privacy cost of MCMC, allowing us to quantify the privacy of many real-world MCMC-based Bayesian analyses. We focus on the case of Gibbs sampling,

under exponential mechanism and Laplace mechanism approaches. By reinterpreting Gibbs sampling as an instance of the exponential mechanism, we obtain the "privacy for free" cost of Gibbs sampling. Metropolis-Hastings and annealed importance sampling also have privacy guarantees.

4.4.1 Gibbs Sampling with the Exponential Mechanism

We consider the privacy cost of a Gibbs sampler, where data $\mathbf{X}$ are behind the privacy wall, current sampled values of parameters and latent variables $\theta = [\theta_1, \ldots, \theta_D]$ are publicly known, and a Gibbs update is a randomized algorithm which queries our private data in order to randomly select a new value θ_l' for the current variable θ_l. The transition kernel for a Gibbs update of θ_l is

$$T^{(Gibbs,l)}(\theta, \theta') = \mathsf{Pr}(\theta_l' \big| \theta_{\neg l}, \mathbf{X}) \;, \tag{4.17}$$

where $\theta_{\neg l}$ refers to all entries of θ except l, which are held fixed, i.e. $\theta_{\neg l}' = \theta_{\neg l}$. This update can be understood via the exponential mechanism:

$$T^{(Gibbs,l,\varepsilon)}(\theta, \theta') \propto \mathsf{Pr}(\theta_l', \theta_{\neg l}, \mathbf{X})^{\frac{\varepsilon}{2\triangle \log \mathsf{Pr}(\theta_l', \theta_{\neg l}, \mathbf{X})}} \;, \tag{4.18}$$

with utility function $u(\mathbf{X}, \theta_l'; \theta_{\neg l}) = \log \mathsf{Pr}(\theta_l', \theta_{\neg l}, \mathbf{X})$, over the space of possible assignments to θ_l, holding $\theta_{\neg l}$ fixed. A Gibbs update is therefore ε-differentially private, with $\varepsilon = 2\triangle \log \mathsf{Pr}(\theta_l', \theta_{\neg l}, \mathbf{X})$. This update corresponds to Equation 4.6 except that the set of responses for the exponential mechanism is restricted to those where $\theta_{\neg l}' = \theta_{\neg l}$. Note that

$$\triangle \log \mathsf{Pr}(\theta_l', \theta_{\neg l}, \mathbf{X}) \leq \triangle \log \mathsf{Pr}(\theta, \mathbf{X}) \tag{4.19}$$

as the worst case is computed over a strictly smaller set of outcomes. In many cases each parameter and latent variable θ_l is associated with only the lth data point $\mathbf{x}_l$, in which case the privacy cost of a Gibbs scan can be improved over simple additive composition. In this case a

random sequence scan Gibbs pass, which updates all N θ_l's exactly once, is $2\triangle \log \mathsf{Pr}(\theta, \mathbf{X})$-differentially private by parallel composition [Song et al., 2013]. Alternatively, a random scan Gibbs sampler, which updates a random Q out of N θ_l's, is $4\triangle \log \mathsf{Pr}(\theta, \mathbf{X})\frac{Q}{N}$-differentially private from the *privacy amplification* benefit of subsampling data [Li et al., 2012].

4.4.2 Gibbs Sampling with the Laplace Mechanism

Suppose that the conditional posterior distribution for a Gibbs update is in the exponential family. Having privatized the sufficient statistics arising from the data for the likelihoods involved in each update, via Equation 4.15, and publicly released them with privacy cost ε, we may now perform the update by drawing a sample from the approximate conditional posterior, i.e. Equation 4.11 but with $S(\mathbf{X}) = \sum_{i=1}^{N}(\mathbf{x}^{(i)})$ replaced by $\hat{S}(\mathbf{X})$. Since the privatized statistics can be made public, we can also subsequently draw from an approximate posterior based on $\hat{S}(\mathbf{X})$ with any other prior (selected based on public information only), without paying any further privacy cost. This is especially valuable in a Gibbs sampling context, where the "prior" for a Gibbs update often consists of factors from other variables and parameters to be sampled, which are updated during the course of the algorithm.

In particular, consider a Bayesian model where a Gibbs sampler interacts with data only via conditional posteriors and their corresponding likelihoods that are exponential family distributions. We can privatize the sufficient statistics of the likelihood just once at the beginning of the MCMC algorithm via the Laplace mechanism with privacy cost ε, and then approximately sample from the posterior by running the entire MCMC algorithm based on these privatized statistics without paying any further privacy cost. This is typically much cheaper in the privacy budget than exponential mechanism MCMC which pays a privacy cost for every Gibbs update, as we shall see in our case study in Section 4.5. The MCMC algorithm does not need to converge to obtain privacy guarantees, unlike the OPS method. This approach applies to a very broad class of models, including Bayesian parameter learning for fully-observed MRF and Bayesian network models. Of course, for this technique to be useful in practice, the aggregate sufficient

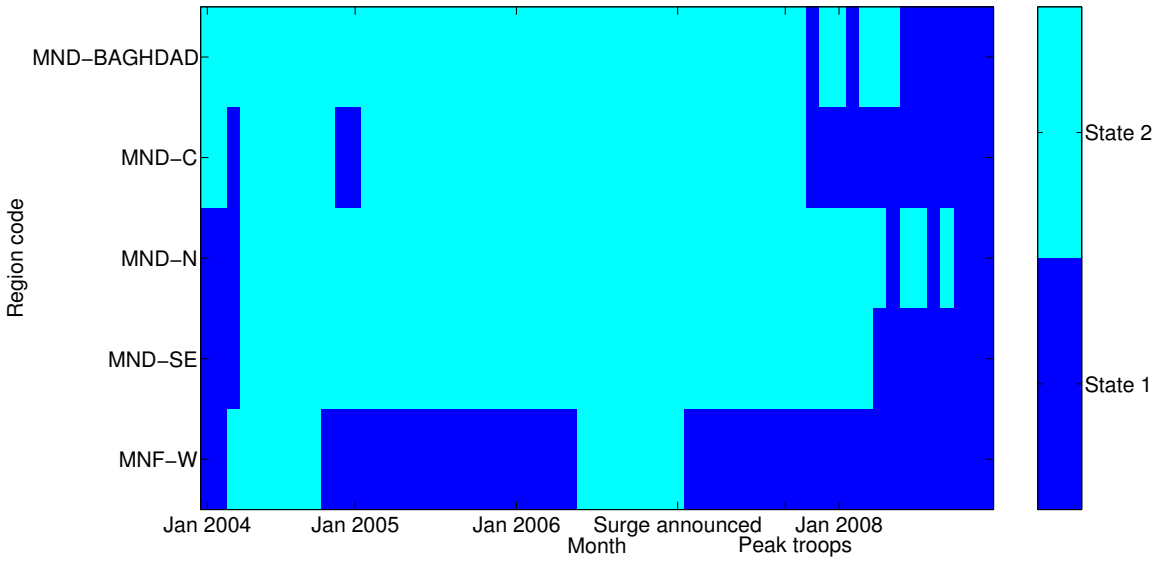

Figure 4.3. State assignments of privacy-preserving HMM on Iraq (Laplace mechanism, $\varepsilon = 5$).

statistics for each Gibbs update must be large relative to the Laplace noise. For latent variable models, this typically corresponds to a setting with many data points per latent variable, such as the HMM model with multiple emissions per timestep which we study in the next section.

4.5 Discussion

A primary goal of this work is to establish the practical feasibility of privacy-preserving Bayesian data analysis using complex models on real-world datasets. In this section we investigate the performance of the methods studied in this paper for the analysis of sensitive military data. In July and October 2010, the Wikileaks organization disclosed collections of internal U.S. military field reports from the wars in Afghanistan and Iraq, respectively. Both disclosures contained data from between January 2004 to December 2009, with ~75,000 entries from the war in Afghanistan, and ~390,000 entries from Iraq. Hillary Clinton, at that time the U.S. Secretary of State, criticized the disclosure, stating that it "puts the lives of United States and its partners' service members and civilians at risk."[1] These risks, and the motivations for the leak, could

[1]Fallon, Amy (2010). "Iraq war logs: disclosure condemned by Hillary Clinton and Nato." The Guardian. Retrieved on 2/22/2016.

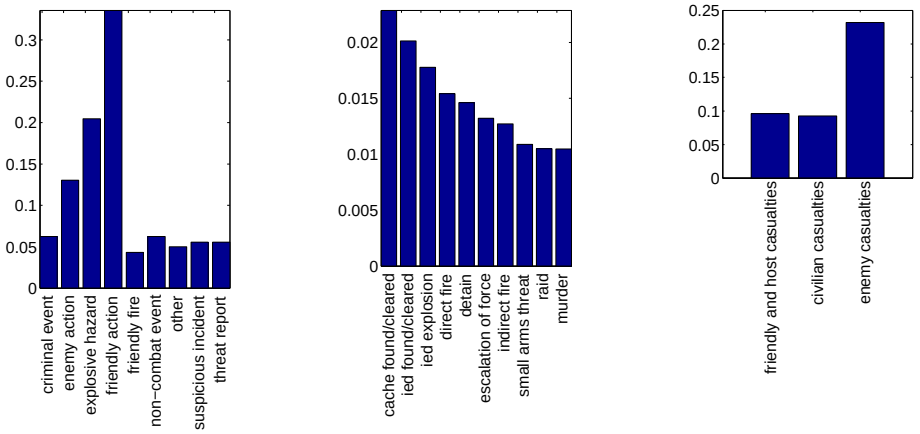

Figure 4.4. State 1 for Iraq (*type*, *category*, *casualties*).

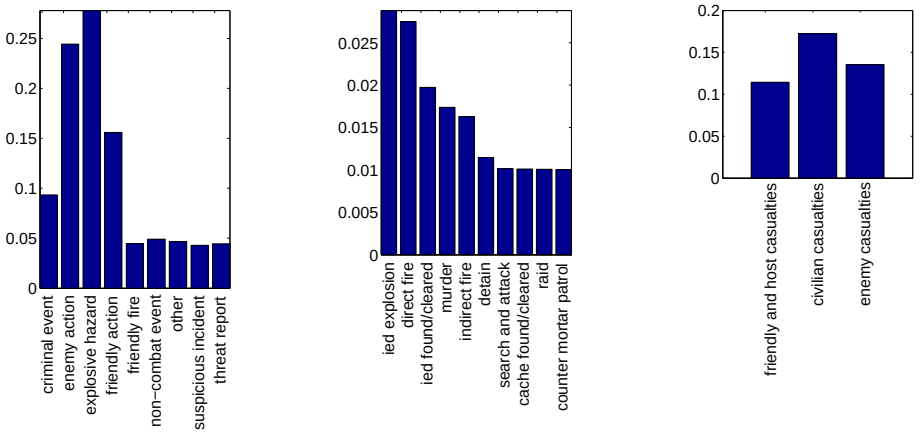

Figure 4.5. State 2 for Iraq (*type*, *category*, *casualties*).

potentially have been mitigated by releasing a differentially private analysis of the data, which protects the contents of each individual log entry while revealing high-level trends. Note that since the data are publicly available, although our *models* were differentially private, other aspects of this manuscript such as the evaluation may reveal certain information, as in other works such as Wang et al. [2015a,b].

The disclosed war logs each correspond to an individual event, and contain textual reports, as well as fields such as coarse-grained *types* (*friendly action*, *explosive hazard*, ...), fine-grained *categories* (*mine found/cleared*, *show of force*, ...), and casualty counts (*wounded/killed/detained*) for the different factions (*Friendly*, *HostNation* (i.e. Iraqi and Afghani forces), *Civilian*, and *Enemy*, where the names are relative to the U.S. military's perspective).

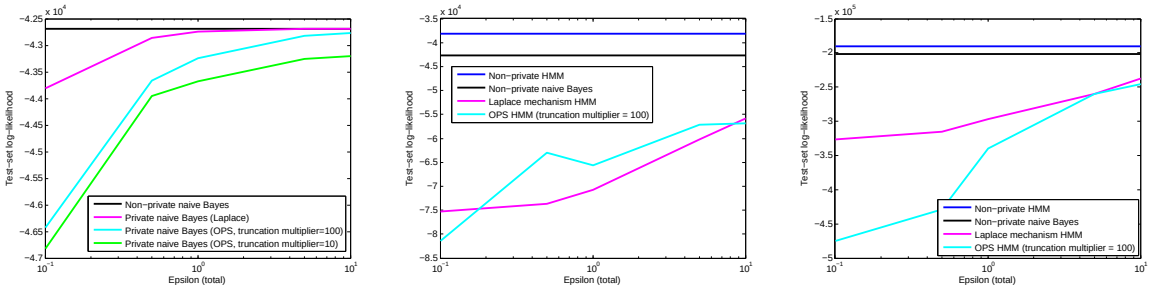

Figure 4.6. Log-likelihood results on HMMs. **Left:** Naive Bayes (Afghanistan). **Middle**: Afghanistan. **Right**: Iraq. For OPS, Dirichlets were truncated at $a_0 = \frac{1}{MK_d}$, $M = 10$ or 100, where $K_d =$ feature d's dimensionality.

We use the techniques discussed in this paper to privately infer a hidden Markov model on the log entries. The HMM was fit to the non-textual fields listed above, with one timestep per month, and one HMM chain per region code. A naive Bayes conditional independence assumption was used in the emission probabilities for simplicity and parameter-count parsimony. Each field was modeled via a discrete distribution per latent state, with casualty counts binarized (0 versus > 0), and with *wounded*/*killed*/*detained* and *Friendly*/*HostNation* features combined, respectively, via disjunction of the binary values. This decreased the number of features to privatize, while slightly increasing the size of the counts per field to protect and simplifying the model for visualization purposes. After preprocessing to remove empty timesteps and near-empty region codes, the median number of log entries per region/timestep pair was 972 for Iraq, and 58 for Afghanistan. The number of log entries per timestep was highly skewed for Afghanistan, due to an increase in density over time.

The models were trained via Gibbs sampling, with the transition probabilities collapsed out, following Goldwater and Griffiths [2007]. We did not collapse out the naive Bayes parameters in order to keep the conditional likelihood in the exponential family. The details of the model and inference algorithm are given in the supplementary material for Foulds et al. [2016]. We trained the models for 200 Gibbs iterations, with the first 100 used for burn-in. Both privatization methods have the same overall computational complexity as the non-private sampler. The Laplace mechanism's computational overhead is paid once up-front, and did not

greatly affect the runtime, while OPS roughly doubled the runtime. For visualization purposes we recovered parameter estimates via the posterior mean based on the latent variable assignments of the final iteration, and we reported the most frequent latent variable assignments over the non-burn-in iterations. We trained a 2-state model on the Iraq data, and a 3-state model for the Afghanistan data, using the Laplace approach with total $\varepsilon = 5$ ($\varepsilon = 1$ for each of 5 features).

Interestingly, when given 10 states, the privacy-preserving model only assigned substantial numbers of data points to these 2-3 states, while a non-private HMM happily fit a 10-state model to the data. The Laplace noise therefore appears to play the role of a regularizer, consistent with the noise being interpreted as a "random prior," and along the lines of noise-based regularization techniques such as [Srivastava et al., 2014, van der Maaten et al., 2013], although of course it may correspond to more regularization than we would typically like. This phenomenon potentially merits further study, beyond the scope of this paper.

We visualized the output of the Laplace HMM for Iraq in Figures 4.3–4.5. State 1 shows the U.S. military performing well, with the most frequent outcomes for each feature being *friendly action*, *cache found/cleared*, and *enemy casualties*, while the U.S. military performed poorly in State 2 (*explosive hazard*, *IED explosion*, *civilian casualties*). State 2 was prevalent in most regions until the situation improved to State 1 after the troop surge strategy of 2007. This transition typically occurred after troops peaked in Sept.–Nov. 2007.

We also evaluated the methods at prediction. A uniform random 10% of the timestep / region pairs were held out for 10 train/test splits, and we reported average test likelihoods over the splits. We estimated test log-likelihood for each split by averaging the test likelihood over the burned-in samples (Laplace mechanism), or using the final sample (OPS). All methods were given 10 latent states, and ε was varied between 0.1 and 10. We also considered a naive Bayes model, equivalent to a 1-state HMM. The Laplace mechanism was superior to OPS for the naive Bayes model, for which the statistics are corpus-wide counts, corresponding to a high-data regime in which our asymptotic analysis was applicable. OPS was competitive with the Laplace mechanism for the HMM on Afghanistan, where the amount of data was relatively low. For the

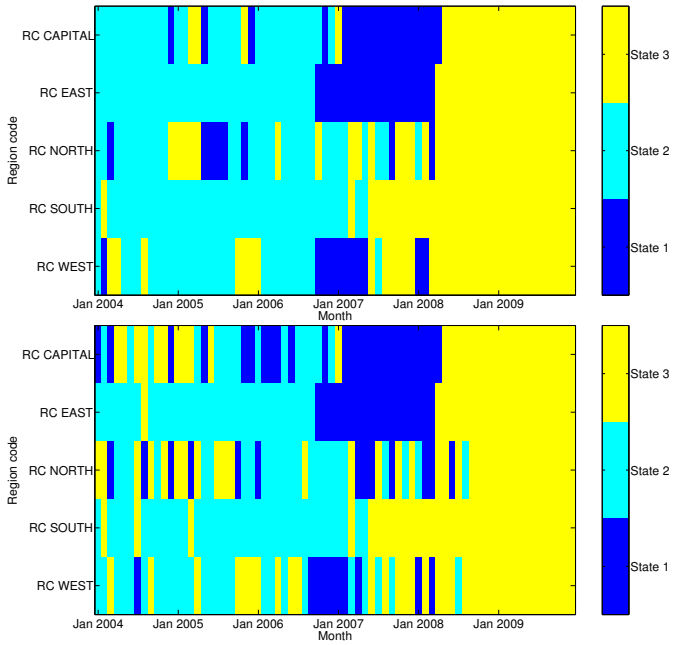

Figure 4.7. State assignments for OPS privacy-preserving HMM on Afghanistan. ($\varepsilon = 5$, truncation point $a_0 = \frac{1}{100K_d}$). **Top**: Estimate from last 100 samples. **Bottom**: Estimate from last one sample.

Iraq dataset, where there was more data per timestep, the Laplace mechanism outperformed OPS, particularly in the high-privacy regime. For OPS, privacy at ε is only guaranteed if MCMC has converged. Otherwise, from Section 4.4.1, the worst case is an impractical $\varepsilon^{(Gibbs)} \leq 400\varepsilon$ (200 iterations of latent variable and parameter updates with worst-case cost ε). OPS only releases one sample, which harmed the coherency of the visualization for Afghanistan, as latent states of the final sample were noisy relative to an estimate based on all 100 post burn-in samples (Figure 4.7). Privatizing the Gibbs chain at a privacy cost of $\varepsilon^{(Gibbs)}$ would avoid this.

4.6 Proofs of Theoretical Results

Here we provide proofs for the results presented in Section 4.3.3.

Our results hold specifically over the class of exponential families. A family of distribu-

tions parameterized by θ which has the form

$$\Pr(\mathbf{x}|\theta) = h(\mathbf{x}) \exp\left(\theta^\intercal S(\mathbf{x}) - A(\theta)\right) \tag{4.20}$$

is said to be an exponential family. Breaking down this structure into its parts, θ is a vector known as the natural parameters for the distribution and lies in some space Θ. $S(\mathbf{x})$ represents a vector of sufficient statistics that fully capture the information needed to determine how likely $\mathbf{x}$ is under this distribution. $A(\theta)$ represents the log-normalizer, a term used to make this a probability distribution sum to one over all possibilities of $\mathbf{x}$. $h(\mathbf{x})$ is a base measure for this family, independent of which distribution in the family is used.

As we are interested in learning θ, we are considering algorithms that generate a posterior distribution for θ. The exponential families always have a conjugate prior family which is itself an exponential family. When speaking of these prior and posterior distributions, θ becomes the random variable and we introduce a new vector of natural parameters η in a space M to parameterize these distributions. To ease notation, we will express this conjugate prior exponential family as $\Pr(\theta|\eta) = f(\theta)\exp\left(\eta^\intercal T(\theta) - B(\eta)\right)$, which is simply a relabelling of the exponential family structure. The posterior from this conjugate prior is often written in an equivalent form

$$\Pr(\theta|\mathbf{X},\chi,\alpha) \propto g(\theta)^{N+\alpha} \exp\left(\theta^\intercal \left(\textstyle\sum_{i=1}^N S(\mathbf{x}^{(i)}) + \alpha\chi\right)\right),$$

where the vector χ and the scalar α together specify the vector η of natural parameters for this distribution. From the interaction of χ, α, and $\mathbf{X}$ on the posterior, one can see that this prior acts like α observations with average sufficient statistics χ have already been observed. This parameterization with χ and α has many nice intuitive properties, but our proofs center around the natural parameter vector η for this prior.

These two forms for the posterior can be reconciled by letting $\eta = (\alpha\chi + \sum_{i=1}^N S(\mathbf{x}^{(i)}), N + \alpha)$ and $T(\theta) = (\theta, -A(\theta))$. This definition for the natural parameters η and sufficient statistics $T(\theta)$ fully specify the exponential family the posterior resides in, with $B(\eta)$ defined as the

appropriate log-normalizer for this distribution (and $f(\theta) = 1$ is merely a constant). We note that the space of $T(\Theta)$ is not the full space $\mathbb{R}^{d+1}$, as the last component of $T(\theta)$ is a function of the previous components. Plugging in these expressions for η and $T(\theta)$ we get the following form for the conjugate prior:

$$\mathsf{Pr}(\theta|\mathbf{X},\chi,\alpha) = \exp\Big(\theta^{\mathsf{T}}(\alpha\chi + \sum_{i=1}^{N} S(\mathbf{x}^{(i)})) - (N+\alpha)A(\theta) - B(\eta)\Big). \tag{4.21}$$

We begin by defining minimal exponential families, a special class of exponential families with nice properties. To be minimal, the sufficient statistics must be linearly independent. We will later relax the requirement that we consider only minimal exponential families.

Definition 8. An exponential family of distributions generating a random variable $\mathbf{x} \in \mathscr{X}$ with $S(\mathbf{x}) \in \mathbb{R}^d$ is said to be minimal if $\nexists \phi \in \mathbb{R}^d, \phi \neq 0$ s.t. $\exists c \in \mathbb{R}$ s.t. $\forall \mathbf{x} \in \mathscr{X}$ $\phi^{\mathsf{T}} S(\mathbf{x}) = c$.

Next we present a few simple algebraic results of minimal exponential families.

Lemma 4. *For two distributions p, q from the same minimal exponential family,*

$$KL(p||q) = A(\theta_q) - A(\theta_p) - (\theta_q - \theta_p)^{\mathsf{T}} \nabla A(\theta_p) \tag{4.22}$$

where θ_p, θ_q are the natural parameters of p and q, and $A(\theta)$ is the log-normalizer for the exponential family.

Lemma 5. *A minimal exponential family distribution satisfies these equalities:*

$$\nabla A(\theta) = E_{\mathsf{Pr}(\mathbf{x}|\theta)}[S(\mathbf{x})]$$

$$\nabla^2 A(\theta) = cov_{\mathsf{Pr}(\mathbf{x}|\theta)}(S(\mathbf{x}))\,.$$

Lemma 6. *For a minimal exponential family distribution, its log-normalizer $A(\theta)$ is a strictly convex function over the natural parameters. This implies a bijection between θ and $E_{\mathsf{Pr}(\mathbf{x}|\theta)}[S(\mathbf{x})]$.*

These are standard results coming from some algebraic manipulations as seen in [Brown, 1986], and we omit the proof of these lemmas. Lemma 6 immediately leads to a useful corollary about minimal families and their conjugate prior families.

Corollary 7. *For a minimal exponential family distribution, the conjugate prior family given in equation* (4.21) *is also minimal.*

PROOF:

$T(\theta) = (\theta, -A(\theta))$ forms the sufficient statistics for the conjugate prior. Since $A(\theta)$ is strictly convex, there can be no linear relationship between the components of θ and $A(\theta)$. Definition 8 applies. □

Our next result looks at sufficient conditions for getting a KL divergence of 0 in the limit when adding a finite perturbance vector γ to the natural parameters. The limit is taken over N, which will later be tied to the amount of data used in forming the posterior. As we now discuss posterior distributions also forming exponential families, our natural parameters will now be denoted by η and the random variables are now θ.

Lemma 8. *Let $\mathsf{Pr}(\theta|\eta)$ denote the distribution from an exponential family of natural parameter η, and let γ be a constant vector of the same dimensionality as η, and let η_N be a sequence of natural parameters. If for every ζ on the line segment connecting η and $\eta+\gamma$ we have the spectral norm $||\nabla^2 B(\zeta)|| < D_N$ for some constant D_N, then*

$$KL(\mathsf{Pr}(\theta|\eta_N+\gamma)||\mathsf{Pr}(\theta|\eta_N)) \leq D_N||\gamma||\,.$$

PROOF: This follows from noticing that equation (4.22) in Lemma 4 becomes the first-order Taylor approximation of $B(\eta_N)$ centered at $B(\eta_N+\gamma)$. From Taylor's theorem, there exists α

between η_N and $\eta_N + \gamma$ such that $\frac{1}{2}\gamma^\intercal \nabla^2 B(\alpha)\gamma$ is equal to the error of this approximation.

$$B(\eta_N) = B(\eta_N + \gamma) + (-\gamma)^\intercal \nabla B(\eta_N + \gamma) + \frac{1}{2}\gamma^\intercal \nabla^2 B(\alpha)\gamma \tag{4.23}$$

From rearranging equation (4.22),

$$\begin{aligned} B(\eta_N + \gamma) = B(\eta_N) - KL(\mathsf{Pr}(\theta|\eta_N + \gamma)||\mathsf{Pr}(\theta|\eta_N)) \\ +(\gamma)^\intercal \nabla B(\eta_n + \gamma) \end{aligned} \tag{4.24}$$

Using this substitution in (4.23) gives

$$B(\eta_N) = B(\eta_N) - KL(\mathsf{Pr}(\theta|\eta_N + \gamma)||\mathsf{Pr}(\theta|\eta_N)) + \frac{1}{2}\gamma^\intercal \nabla^2 B(\alpha)\gamma\,. \tag{4.25}$$

Solving for $KL(\mathsf{Pr}(\theta|\eta_N + \gamma)||\mathsf{Pr}(\theta|\eta_N))$ then gives the desired result:

$$KL(\mathsf{Pr}(\theta|\eta_N + \gamma)||\mathsf{Pr}(\theta|\eta_N)) = \frac{1}{2}\gamma^\intercal \nabla^2 B(\alpha)\gamma \leq D_N||\gamma||\,.$$

□

This provides the heart of our results: If $||\nabla^2 B(\zeta)||$ is small for all ζ connecting η and $\eta + \gamma$, then we can conclude that $KL(\mathsf{Pr}(\theta|\eta_N + \gamma)||\mathsf{Pr}(\theta|\eta_N))$ is small with respect to $||\gamma||$. We wish to show that for η_N arising from observing N data points we have D_N approaching 0 as N grows. To achieve this, we will analyze a relationship between the norm of the natural parameter η and the covariance of the distribution it parameterizes. This relationship shows that posteriors with plenty of observed data have low covariance over $T(\theta)$, which permits us to use Lemma 8 to bound the KL divergence of our perturbed posteriors. Before we reach this relationship, first we prove that our posteriors have a well-defined mode, as our later relationship will require this mode to be well-behaved.

Lemma 9. *Let* $\mathsf{Pr}(\mathbf{x}|\theta) = h(\mathbf{x})\exp\Big(\theta^\intercal S(\mathbf{x}) - A(\theta)\Big)$ *be a likelihood function for* θ *and let there be a conjugate prior* $\mathsf{Pr}(\theta|\eta) = f(\theta)\exp\Big(\eta^\intercal T(\theta) - B(\eta)\Big)$*, where both distributions are minimal exponential families. Let M be the space of natural parameters* η*, and* Θ *be the space of* θ*. Furthermore, assume* η *is the parameterization arising from the natural conjugate prior, such that* $\eta = (\alpha\chi, \alpha)$*. If the following conditions hold:*

1. η *is in the interior of M*
2. $\alpha > 0$
3. $A(\theta)$ *is a real, continuous, and differentiable*
4. $B(\eta)$ *exists, the distribution* $\mathsf{Pr}(\theta|\eta)$ *is normalizable.*

then

$$\text{argmax}_{\theta\in\Theta}\eta^\intercal T(\theta) = \theta^*_\eta$$

is a well-defined function of η*, and* θ^*_η *is in the interior of* Θ*.*

PROOF:

Using our structure for the conjugate prior from (4.21), we can expand the expression $\eta^\intercal T(\theta)$.

$$\eta^\intercal T(\theta) = \alpha\chi^\intercal\theta - \alpha A(\theta)$$

We note that the first term is linear in θ, and that by minimality and Lemma 6, $A(\theta)$ is strictly convex. This implies $\eta^\intercal T(\theta)$ is strictly concave over θ. Thus any interior local maximum must also be the unique global maximum.

The gradient of with $\eta^\intercal T(\theta)$ respect to θ is simple to compute.

$$\nabla(\eta^\intercal T(\theta)) = \alpha\chi^\intercal - \alpha\nabla A(\theta)$$

This expression can be set to zero, and solving for θ_η^* shows it must satisfy

$$\nabla A(\theta_\eta^*) = \chi \ . \tag{4.26}$$

We remark by Lemma 5 that $\nabla A(\theta_\eta^*)$ is equal to $E_{\mathsf{Pr}(\mathbf{x}|\theta_\eta^*)}[S(\mathbf{x})]$, and so this is the θ that generates a distribution with mean χ.

By the strict concavity, this is sufficient to prove θ_η^* is a unique local maximizer and thus the global maximum.

To see that θ_η^* must be in the interior of Θ, we use the fact that $A(\theta)$ is continuously differentiable. This means $\nabla A(\theta)$ is a continuous function of θ. Since η is in the interior of M, we can construct an open neighborhood around χ. The preimage of an open set under a continuous function is also an open set, so this implies an open neighborhood exists around θ_η^*.

□

Now that we know θ_η^* is well defined for η in the interior of M, we can express our relationship on high magnitude posterior parameters and the covariance of the distribution over $T(\theta)$ they generate.

Lemma 10. *Let* $\mathsf{Pr}(\mathbf{x}|\theta) = h(\mathbf{x})\exp\Big(\theta^\intercal S(\mathbf{x}) - A(\theta)\Big)$ *be a likelihood function for* θ *and let there be a conjugate prior* $\mathsf{Pr}(\theta|\eta) = f(\theta)\exp\Big(\eta^\intercal T(\theta) - B(\eta)\Big)$*, where both distributions are minimal exponential families. Let M be the space of natural parameters* η*, and* Θ *be the space of* θ*. Furthermore, assume* η *is the parameterization arising from the natural conjugate prior, such that* $\eta = (\alpha\chi, \alpha)$*.*

If $\exists \eta_0, \delta_1 > 0, \delta_2 > 0$ *such that the conditions of Lemma 9 hold for* $\eta \in \mathscr{B}(\eta_0, \delta_1)$*, and we have these additional assumptions,*

1. *the cone* $\{k\eta' | k > 1, \eta' \in \overline{\mathscr{B}(\eta_0, \delta_1)}\}$ *lies entirely in M*

2. $A(\theta)$ *is differentiable of all orders*

3. $\exists P$ *s.t.* $\forall \theta \in \cup_{\eta' \in \overline{\mathscr{B}(\eta_0,\delta_1)}} \mathscr{B}(\theta^*_{\eta'}, \delta_2)$ *all partial derivatives up to order 7 of* $A(\theta)$ *have magnitude bounded by* P

4. $\exists w > 0$ *such that* $\forall \theta \in \cup_{\eta' \in \overline{\mathscr{B}(\eta_0,\delta_1)}} \mathscr{B}(\theta^*_{\eta'}, \delta_2)$ *we have* $\det(\nabla^2 A(\theta)) > w$

then there exists C, K *such that for* $k > K$ *the following bound holds* $\forall \eta \in \mathscr{B}(\eta_0, \delta_1)$*:* $||cov(T(\theta)|k\eta)|| < \frac{C}{k}$.

PROOF:

This result follows from the Laplace approximation method for $B(\eta) = \int_\Theta e^{\eta^\intercal T(\theta)} d\theta$. The inner details of this approximation are show in Lemma 14. Here we show that our setting satisfies all the regularity assumptions for this approximation. First we define functions $s(\theta, \eta)$ and $F_k(\eta)$.

$$s(\theta, \eta) = \eta^\intercal T(\theta) = \alpha \chi^\intercal \theta - \alpha A(\theta) \tag{4.27}$$

$$\begin{aligned} F_k(\eta) &= B(k\eta) \\ &= \int_\Theta e^{k\eta^\intercal T(\theta)} d\theta \\ &= \int_\Theta e^{ks(\theta,\eta)} d\theta \end{aligned} \tag{4.28}$$

With these definitions, we may now begin to check the assumptions of Lemma 14 hold. We copy these assumptions below, with a substitution of θ for ϕ and η for Y. The full details of Lemma 14 can be found at the end of this section.

1. $\phi^*_Y = \text{argmax}_{\phi \in M} s(\phi, Y) = g(Y)$, a function of Y.

2. $\phi^*_{Y'}$ is in the interior of M for all $Y' \in \mathscr{B}(Y_0, \delta_1)$.

3. $g(Y)$ is continuously differentiable over the neighborhood $\mathscr{B}(Y_0, \delta_1)$.

4. $s(\phi, Y')$ has derivatives of all orders for $Y' \in \mathscr{B}(Y_0, \delta_1)$,$\phi \in \mathscr{B}(\phi^*_{Y'}, \delta_2)$ and all partial derivatives up to order 7 are bounded by some constant P on this neighborhood.

5. $\exists w > 0$ such that $\forall Y' \in \mathscr{B}(Y_0, \delta_1), \forall \phi \in \mathscr{B}(\phi^*_{Y'}, \delta_2)$ we have $det(\nabla^2_\phi s(\phi, Y)) > w$.

6. $F_1(Y')$ exists for $Y' \in \mathscr{B}(Y_0, \delta_1)$, the integral is finite.

We now show these conditions hold one-by-one. Let η denote an arbitrary element of $B(\eta_0, \delta)$.

1. θ^*_η is a well-defined function (Lemma 9).

2. θ^*_η is in the interior of Θ (Lemma 9).

3. $g(\eta)$ follows the inverse of $\nabla A(\theta) : \mathbb{R}^d \to \mathbb{R}^d$. This vector mapping has a Jacobian $\nabla^2 A(\theta)$ which assumption 4 guarantees has non-zero determinant on this neighborhood. This satisfies the Inverse Function Theorem to show $g(\eta)$ is continuously differentiable.

4. $s(\theta, \eta)$ has derivatives of all orders, and are suitably bounded as s is composed of a linear term and the differentiable function $A(\theta)$, where we have bounded the derivatives of $A(\theta)$.

5. Assumption 4 from this lemma translates directly.

6. $F_1(\eta) = B(\eta)$ which exists by virtue of η being in the space of valid natural parameters.

This completes all the requirements of Lemma 14, which guarantees the existence of C and K such that for any $k > K$ and any $\eta \in \mathscr{B}(\eta_0, \delta_1)$, if we let ψ denote $k\eta$, we have:

$||\nabla^2_\psi B(\psi)|| = ||\nabla^2_\psi \log F_k(\psi/k)|| < \frac{C}{k}$.

We conclude by noting that $\nabla^2_\psi B(\psi)$ is the covariance of the posterior with parameterization $\psi = k\eta$.

□

Now that all our machinery is in place, it remains to be seen under what conditions the posterior satisfies the conditions of the previous Lemmas, along with extending to the case where γ is a random variable, and not just a fixed finite vector.

Lemma 11. *Lemma 1 revisited. For a minimal exponential family given a conjugate prior, where the posterior takes the form* $\mathsf{Pr}(\theta|\mathbf{X},\chi,\alpha) \propto g(\theta)^{n+\alpha}\exp\left(\theta^{\mathsf{T}}\left(\sum_{i=1}^{n} S(\mathbf{x}^{(i)}) + \alpha\chi\right)\right)$, *where* $\mathsf{Pr}(\theta|\eta)$ *denotes this posterior with a natural parameter vector* η, *if there exists a* $\delta > 0$ *such that these assumptions are met:*

1. *the data* $\mathbf{X}$ *comes i.i.d. from a minimal exponential family distribution with natural parameter* $\theta_0 \in \Theta$
2. θ_0 *is in the interior of* Θ
3. *the function* $A(\theta)$ *has all derivatives for* θ *in the interior of* Θ
4. $cov_{\mathsf{Pr}(\mathbf{x}|\theta)}(S(\mathbf{x})))$ *is finite for* $\theta \in \mathscr{B}(\theta_0,\delta)$
5. $\exists w > 0$ *s.t.* $\det(cov_{\mathsf{Pr}(\mathbf{x}|\theta)}(S(\mathbf{x})))) > w$ *for* $\theta \in \mathscr{B}(\theta_0,\delta)$
6. *the prior* $\mathsf{Pr}(\theta|\chi,\alpha)$ *is integrable and has support on a neighborhood of* θ^*

then for any mechanism generating a perturbed posterior $\tilde{p}_N = \mathsf{Pr}(\theta|\eta_N+\gamma)$ *against a noiseless posterior* $p_N = \mathsf{Pr}(\theta|\eta_N)$ *where* γ *comes from a distribution that does not depend on the number of data observations N and has finite covariance, this limit holds:*

$\lim_{N\to\infty} E[KL(\tilde{p}_N||p_N)] = 0$.

PROOF:

We begin by fixing the randomness of the noise γ that the mechanism will add to the natural parameters of the posterior.

We wish to show that the KL divergence goes to zero in the limit, which we will achieve by showing that for large enough data sizes, both the perturbed and unperturbed posteriors lie w.h.p. in a region where we can use Lemmas 8 and 10 apply.

To compute the posterior, after drawing a collection $\mathbf{X}$ of N data observations, we compute the sum of the sufficient statistics and add them to the prior's parameters.

$$\eta_N = \Big(\alpha\chi + \sum S(\mathbf{x}^{(i)}), \alpha + N\Big)$$

η_N is a random variable depending on the data observations $\mathbf{X}$. To analyze how it behaves, a couple related random variables will be defined, all implicitly conditioned on the constant θ_0. Let $\mathbf{Y}$ denote a random variable matching the distribution of a single observation, and let $\mathbf{U}_N = \frac{1}{N}\sum S(\mathbf{x}^{(i)})$ which has covariance $\frac{1}{N}cov(S(\mathbf{Y}))$. The expected value for $\mathbf{U}_N$ is of course $E[S(\mathbf{Y})]$.

By a vector version of the Chebyshev inequality for a random vector $\mathbf{U}$, [Chen, 2007]

$$\begin{aligned} Pr\Big((\mathbf{U}-E[\mathbf{U}])^\intercal (cov(\mathbf{U}))^{-1}(\mathbf{U}-E[\mathbf{U}]) \geq \nu\Big), \\ \leq \frac{d}{\nu}, \end{aligned} \tag{4.29}$$

where d is the dimensionality of $\mathbf{U}$. Using the spectral norm $||(cov(\mathbf{U}_N))^{-1}||$ and the l_2 norm $||\mathbf{U}_N - E[\mathbf{U}_N]||$ with some some rearrangement, we can show the following inequalities. We note that the covariance of $\mathbf{U}_N$ must be invertible, since the covariance of $\mathbf{Y}$ is invertible by assumption (5).

$$Pr\Big(||\mathbf{U}_N - E[\mathbf{U}_N]|| \cdot ||(cov(\mathbf{U}_N))^{-1}|| \geq \nu\Big) \leq \frac{d}{\nu} \tag{4.30}$$

$$Pr\Big(||\mathbf{U}_N - E[\mathbf{U}_N]|| \geq \nu ||cov(\mathbf{U}_N)||\Big) \leq \frac{d}{\nu} \tag{4.31}$$

$$Pr\Big(||\mathbf{U}_N - E[S(\mathbf{Y})]|| \geq \frac{\nu}{N}||cov(\mathbf{Y})||\Big) \leq \frac{d}{\nu} \tag{4.32}$$

Thus for any $\varepsilon > 0$, $\tau > 0$, there exists $N_{\varepsilon,\tau}$ such that when the number of data observations N exceeds $N_{\varepsilon,\tau}$

$$Pr\big(||\mathbf{U}_N - E[\mathbf{Y}]|| \geq \varepsilon\big) \leq \tau. \tag{4.33}$$

We now define two modified vectors of natural parameters $\eta_a = \frac{\eta_N}{N} = (\mathbf{U}_N, 1) + \frac{1}{N}(\alpha\chi, \alpha)$ and $\eta_b = \frac{\eta_N + \gamma}{N} = (\mathbf{U}_n, 1) + \frac{1}{N}(\alpha\chi, \alpha) + \frac{1}{N}\gamma$. From these definitions, one can see

$$E[\eta_a] = (E[\mathbf{Y}], 1) + \frac{1}{N}(\alpha\chi, \alpha)$$

$$E[\eta_b] = E[\eta_a] + \frac{1}{N}\gamma$$

$$||\eta_a - (E[\mathbf{Y}], 1)|| \leq ||(\mathbf{U}_N, 1) - (E[\mathbf{Y}], 1)|| + \frac{1}{N}||\alpha\chi|| \tag{4.34}$$

$$\begin{aligned} ||\eta_b - (E[\mathbf{Y}], 1)|| \leq\ & ||(\mathbf{U}_N, 1) - (E[\mathbf{Y}], 1)|| \\ & + \frac{1}{N}\big(||\alpha\chi|| + ||\gamma||\big). \end{aligned} \tag{4.35}$$

From the concentration bound in (4.33), we know η_a and η_b can be made to lie w.h.p. in a region near their expectations with large N, and we wish to show this region satisfies all the regularity assumptions seen in Lemma 10. Lemma 9 states θ^*_η is a continuously differentiable function of η. Let it be denoted by the function $r(\eta)$. For $\eta_0 = (E[\mathbf{Y}], 1)$, we see from equation (4.26) that $r(\eta_0) = \theta_0$.

The preimage $r^{-1}\big(\mathscr{B}(\theta_0, \delta)\big)$ is an open set, since it is the continuous preimage of an open set. Thus there exists δ' such that $\mathscr{B}(\eta_0, \delta') \subset r^{-1}(\mathscr{B}(\theta_0, \delta/2))$.

We may now pick $\varepsilon \leq \delta'/2$ and let $N'_{\delta',\tau} = \max\big(\frac{2}{\delta'}(||\gamma|| + ||\alpha\chi||), N_{\varepsilon,\tau}\big)$. When $n > N'_{\delta',\tau}$,

we have $\frac{1}{N}||\alpha\chi|| + \frac{1}{N}||\gamma|| \leq \delta'/2$ and (4.33), (4.34),(4.35) together show the following:

$$\mathsf{Pr}(\eta_a \notin \mathscr{B}(\eta_0, \delta') \vee \eta_b \notin \mathscr{B}(\eta_0, \delta')) \leq \tau. \tag{4.36}$$

With high probability, η_a and η_b both lie in a neighborhood of η_0. Further, all η in this neighborhood have modes $\theta^*_\eta \in \mathscr{B}(\theta_0, \delta)$, a region that assumptions (4) and (5) tell us is well-behaved. The assignment $\delta_1 = \delta'$ and $\delta_2 = \delta/2$ satisfies the conditions for Lemma 10 with assumptions (2),(3),(4),(5),(6) serving to round out the rest of the regularity assumptions of Lemma 10 with trivial translations.

By the construction, we have $\eta_N = N\eta_a$ and $\eta_N + \gamma = N\eta_b$. For any ζ on the line segment connecting η_N and $\eta_N + \gamma$, we have $\zeta = N\eta_c$ for some η_c on the line segment connecting η_a and η_b.

Therefore by Lemma 10, there exists a K and a C such that if $N > K$ we have $||cov(T(\theta)|\zeta)|| < \frac{C}{N}$. This bound can be used in Lemma 8 with $D_N = O(1/N)$ to see

$$KL(\tilde{p}_N||p_N) = O(1/N)C||\gamma||$$

whenever $N > max(N'_{\delta',\tau}, K)$ with arbitrarily high probability $1 - \tau$. Letting τ approach 0, we can extend this to the expectation over the randomness of $\mathbf{X}$, as with probability 1 our random variables will lie in the region where this inequality holds.

$$\limsup_{N\to\infty} E_{\mathbf{X}}[KL(\tilde{p}_N||p_N)] = 0 \tag{4.37}$$

Equation (4.37) is w.r.t. to a fixed γ, but the desired result is an expectation over γ and $\mathbf{X}$. First, let us express this expectation in terms of γ and $\mathbf{X}$. Letting $D_N = O(1/N)$ denote the bound used in Lemma 8 and N being sufficiently large:

$$E[KL(\tilde{p}_N||p_N)] = \int E_{\mathbf{X}}\big[KL(\tilde{p}_N||p_N)|\gamma\big]d\mathsf{Pr}(\gamma)$$
$$\leq \int D_N||\gamma||d\mathsf{Pr}(\gamma). \tag{4.38}$$

The assumption that γ comes from a distribution of finite variance ensures the right side of (4.38) is integrable. By an application of Fatou's Lemma, the following inequality holds:

$$\int \limsup_{N\to\infty} E_{\mathbf{X}}\big[KL(\tilde{p}_N||p_N)|\gamma\big]d\mathsf{Pr}(\gamma)$$
$$\geq \limsup_{N\to\infty} \int E_{\mathbf{X}}\big[KL(\tilde{p}_N||p_N)|\gamma\big]d\mathsf{Pr}(\gamma). \tag{4.39}$$

The left hand side has been shown to be zero by equations (4.37) and (4.38), and the right hand side is bounded below by 0 since KL divergences are never negative. Thus this inequality suffices to show the limit is zero and prove the desired result.

□

Corollary 12. *The Laplace mechanism on an exponential family satisfies the noise distribution requirements of Lemma 1 when the sensitivity of the sufficient statistics is finite and either the exponential family is minimal, or if the exponential family parameters θ are identifiable.*

PROOF: If the exponential family is already minimal, this result is trivial. If it is not minimal, there exists a minimal parameterization. We wish to show adding noise to the non-minimal parameters is equivalent to adding differently distributed noise to the minimal parameterization, and this new noise distribution also satisfies the noise distribution requirements of Lemma 1: the noise distribution does not depend on N and it has finite covariance.

Let us explicitly construct a minimal parameterization for this family of distributions. If the exponential family is not minimal, this means the d dimensions of the sufficient statistics $S(\mathbf{x})$

of the data are not fully linearly independent. Let $S(x)_j$ be the j^{th} component of $S(\mathbf{x})$ and k be the maximal number of linearly independent sufficient statistics, and without loss of generality assume they are the first k components. Let $\tilde{S}(\mathbf{x})$ be the vector of these k linearly independent components.

For $\forall j > k$, $\forall x \exists \phi_j \in \mathbb{R}^k$ such that $S(\mathbf{x})_j = \phi_j \cdot \tilde{S}(\mathbf{x}) + z_j$. We wish to build a minimal exponential family distribution that is identical to the original one, but is parameterized only by $\tilde{S}(\mathbf{x})$ as the sufficient statistics and some $\tilde{\theta}$ as the natural parameters. For these two distributions to be equivalent for all x, it suffices to have equality on the exponents.

$$(\theta^\intercal S(\mathbf{x}) - A(\theta)) = (\tilde{\theta}^\intercal \tilde{S}(\mathbf{x}) - \tilde{A}(\tilde{\theta})) \tag{4.40}$$

Examining the difference of the two sides, we get

$$\begin{aligned} &\theta^\intercal S(x) - \tilde{\theta}^\intercal \tilde{S}(x) - A(\theta) + \tilde{A}(\tilde{\theta}) \\ &= \sum_{j=1}^{k} (\theta_j - \tilde{\theta}_j) S(x)_j + \sum_{j=k+1}^{d} \theta_j S(x)_j - A(\theta) + \tilde{A}(\tilde{\theta}). \end{aligned} \tag{4.41}$$

Using the known linear dependence for $j > k$, this can be rewritten as

$$\begin{aligned} &\sum_{j=1}^{k} (\theta_j - \tilde{\theta}_j) S(\mathbf{x})_j + \sum_{j=k+1}^{d} \theta_j (\phi_j \cdot \tilde{S}(\mathbf{x}) + z_j) \\ &\qquad - A(\theta) + \tilde{A}(\tilde{\theta}) \end{aligned} \tag{4.42}$$

$$\begin{aligned} &= \sum_{j=1}^{k} (\theta_j - \tilde{\theta}_j) S(\mathbf{x})_j + \sum_{j=k+1}^{d} \theta_j (\phi_j \cdot \tilde{S}(\mathbf{x})) \\ &\qquad + \sum_{j=k+1}^{d} \theta_j z_j - A(\theta) + \tilde{A}(\tilde{\theta}). \end{aligned} \tag{4.43}$$

Now since $\tilde{S}(\mathbf{x})$ is merely the first k components of $S(\mathbf{x})$, the first two sums of (4.43) are

each simply dot products of $\tilde{S}(\mathbf{x})$ and can be combined as $(\theta_{[k]} - \tilde{\theta} + \sum_{j=k+1}^{d} \theta_j \phi_j)^\intercal \tilde{S}(\mathbf{x})$ where $\theta_{[k]}$ is the vector of the first k components of θ. We can force equation (4.40) to hold by choosing $\tilde{\theta}$ and $\tilde{A}(\tilde{\theta})$ appropriately to set equation (4.43) to zero.

- $\tilde{\theta} = \theta_{[k]} + \sum_{j=k+1}^{d} \theta_j \phi_j$
- $\tilde{A}(\tilde{\theta}) = -\sum_{j=k+1}^{d} \theta_j z_j + A(\theta)$

We note that this requires $\tilde{A}(\tilde{\theta})$ to truly be a function depending only on $\tilde{\theta}$, but we have written it in terms of θ instead. This is justifiable by the assumption that the natural parameters θ are identifiable, that is each distribution over $\mathbf{x}$ is associated with just one $\theta \in \Theta$. This means there is a bijection from θ and $\tilde{\theta}$, which ensures $\tilde{A}(\tilde{\theta})$ is a well-defined function.

This suffices to characterize the way the additional natural parameters affect the parameters of the equivalent minimal system. Any additive noise to a component θ_j translates linearly to additive noise on the components $\tilde{\theta}_j$, meaning the Laplace mechanism's noise distribution on the non-minimal parameter space still corresponds to some noise distribution on the minimal parameters that does not depend on the data size N, and it still has a finite covariance. If the minimal exponential family tends towards a KL divergence of zero, the equivalent non-minimal exponential family must as well. □

Theorem 6. *Under the assumptions of Lemma 1, the Laplace mechanism has an asymptotic posterior of* $\mathcal{N}(\theta_0, 2\mathbb{I}^{-1}/N)$ *from which drawing a single sample has an asymptotic relative efficiency of 2 in estimating* θ_0*, where* $\mathbb{I}$ *is the Fisher information at* θ_0*.*

PROOF:

The assumptions of Lemma 1 match the Laplace regularity assumptions under which asymptotic normality holds, and we know that the unperturbed posterior p_N converges to $\mathcal{N}(\theta^*, 2\mathbb{I}^{-1}/N)$ under the Bernstein-von Mises theorem [Kass et al., 1990]. If $\tilde{p}_N$ is the posterior of the Laplace mechanism for a fixed randomness, then we have $\lim_{N\to\infty} KL(\tilde{p}_N || p_N) = 0$ and

$\tilde{p}_N$ must converge to the same distribution as p_N. From this it is clear that samples from p_N and from $\tilde{p}_N$ both have an asymptotic relative efficiency of 2. We once again argue that if this asymptotic behavior holds for any fixed randomness of the Laplace mechanism, it also holds for the Laplace mechanism as a whole. □

To show the previous results, we relied on some mathematical results involving the covariances of posteriors after observing a large amount of data. We still need to show these bounds on the covariances, which will be accomplished by adapting existing Laplace approximation methods. Before we get there, we will need one quick result about convex functions with a positive definite Hessian in order to perform the approximation:

Lemma 13. *Let $f(y) : \mathbb{R}^d \to \mathbb{R}$ be a strictly convex function with minimum at y^*. If $\nabla^2 f(y^*)$ is positive definite and $\nabla^3 f(y)$ exists everywhere, then for any $c > 0$ there exists $b > 0$ such that $|f(y) - f(y^*)| \leq b$ implies $||y - y^*|| \leq c$.*

PROOF:

By the existence of $\nabla^3 f(y)$ and thus the continuity of $\nabla^2 f(y)$, we know there exists a positive $\delta < c$ and a $w > 0$ such that $y \in B(y^*, \delta)$ implies $\nabla^2 f(y) - w\mathbb{I}$ is positive semi-definite, where $\mathbb{I}$ is the identity matrix. (i.e. the spectral norm $||\nabla^2 f(y)|| \geq w$)

As y^* is the global minimum, we know the gradient is 0 at y^*. Thus for $y \in B(y^*, \delta)$ this leads to a Taylor expansion of the form

$$\begin{aligned} f(y) &= f(y^*) + (y - y^*)\frac{1}{2}\nabla^2 f(y')(y - y^*)^\intercal \\ &\geq f(y^*) + \frac{w}{2}||y - y^*|| \end{aligned} \tag{4.44}$$

for some y' on the line segment connecting y and y^*. The inequality follows from the second derivative being positive definite on this neighborhood.

Consider the set $Q_\varepsilon = \{y \text{ s.t. } ||y - y^*|| = \varepsilon\}$. By equation (4.44) we know for $y \in Q_\varepsilon$ we

have $|f(y)-f(y^*)| \geq \frac{w\varepsilon}{2}$ if $\varepsilon \leq \delta$.

For any $y \notin B(y^*,\delta)$, there exists $t \in (0,1)$ such that $(1-t)y^*+ty \in Q_\delta$ by the continuity of the norm.

By strict convexity, we know

$$tf(y)+(1-t)f(y^*) > f(ty+(1-t)y^*)$$

$$f(y) > \frac{1}{t}f(ty+(1-t)y^*)+\frac{t-1}{t}f(y^*)$$

$$f(y)-f(y^*) > \frac{1}{t}f(ty+(1-t)y^*)-\frac{1}{t}f(y^*).$$

If we let t satisfy $(1-t)y^*+ty \in Q_\delta$ we know $t = \delta/||y-y^*|| \leq 1$. Substituting with (4.44) we get

$$f(y)-f(y^*) > \frac{(w/2)\delta+f(y^*)}{t}-\frac{1}{t}f(y^*) = \frac{w\delta}{2t} \geq \frac{w\delta}{2}.$$

Thus if we let $b = \frac{w\delta}{2}$, we see $||y-y^*|| > c$ implies $|f(y)-f(y^*)| > b$.

The desired result then follows as the contrapositive.

□

Lemma 13 will be used to demonstrate a regularity assumption required in the next lemma, which performs all the heavy lifting in using the Laplace approximation. Lemma 14 adapts a previous argument about Laplace approximations of a posterior. This adapted Laplace approximation argument forms the core of Lemma 10, which allows us to see the covariance of posteriors shrink as more data is observed.

Lemma 14. *Let $s(\phi,Y)$ be a function $M \times U \to \mathbb{R}$, where M is the space of ϕ and U is the space*

of Y.

For functions of the form $F_k(Y) = \int_{\phi \in M} e^{ks(\phi,Y)} d\phi$, if the following regularity assumptions hold for some $\delta_1 > 0$, $\delta_2 > 0$, $Y_0 \in M$:

1. *$\phi_Y^* = \text{argmax}_{\phi \in M} s(\phi, Y) = g(Y)$, a function of Y*
2. *$\phi_{Y'}^*$ is in the interior of M for all $Y' \in \mathscr{B}(Y_0, \delta_1)$*
3. *$g(Y)$ is continuously differentiable over the neighborhood $\mathscr{B}(Y_0, \delta_1)$*
4. *$s(\phi, Y')$ has derivatives of all orders for $Y' \in \mathscr{B}(Y_0, \delta_1)$, $\phi \in \mathscr{B}(\phi_{Y'}^*, \delta_2)$ and all partial derivatives up to order 7 are bounded by some constant P on this neighborhood*
5. *$\exists w > 0$ such that $\forall Y' \in \mathscr{B}(Y_0, \delta_1), \forall \phi \in \mathscr{B}(\phi_{Y'}^*, \delta_2)$ we have $det(\nabla_\phi^2 s(\phi, Y)) > w$*
6. *$F_1(Y')$ exists for $Y' \in \mathscr{B}(Y_0, \delta_1)$, the integral is finite*

then there exists C and K such that for any $k > K$ and any $Y' \in \mathscr{B}(Y_0, \delta_1)$, letting $\psi = kY'$, the spectral norm $||\nabla_\psi^2 \log F_k(\psi/k)|| < \frac{C}{k}$.

PROOF:

Our goal here is to bound $||\nabla_\psi^2 \log F_k(\psi/k)||$, which we will achieve by characterizing $F_k(\psi/k)$ and analyzing its derivatives.

We will be using standard Laplace approximation methods seen in [Kass et al., 1990] to explore $F_k(\psi)$. To begin, we must show our assumptions satisfy the regularity assumptions for the approximation.

For a fixed $Y' \in B(Y_0, \delta)$, from condition 5 we know there exists a neighborhood around ϕ_Y^* where $\nabla_\phi^2 s(\phi, Y)$ is positive definite. For $\delta' > 0$, let $Q_{\delta',Y} = \{\phi \in M \text{ s.t. } ||\phi - \phi_Y^*|| \leq \delta'\}$. By using Lemma 13 we can verify the following expression for any $\delta' \in (0, \delta)$:

$$\limsup_{k \to \infty} \sup_{\phi \notin Q_{\delta',Y}} s(\phi, Y) - s(\phi_Y^*, Y) < 0. \tag{4.45}$$

Note that the right hand side does not depend on k, and Lemma 13 guarantees a non-zero bound for the right hand side for any $\delta' \in (0, \delta)$. Equation (4.45) exactly matches condition $(iii)'$ of Kass, and its intuitive meaning is that for any δ', there exists sufficiently large k such that the integral F_k is negligible outside the region $Q_{\delta'}$.

Conditions (4),(5),(6) also match directly the conditions given by Kass, though we note we require even higher derivatives to be bounded or present. These extra derivatives will be used later to extend the argument given by Kass to suit our purposes and give a uniform bound across a neighborhood.

Theorem 1 of [Kass et al., 1990] gives the following result, when we set their b to the constant 1:

$$F_k(Y) = (2\pi)^{\frac{m}{2}} [\det(k\nabla^2 s(\phi_Y^*, Y)]^{-\frac{1}{2}} \exp(-ks(\phi_Y^*, Y)) Z(kY) \tag{4.46}$$

$$\begin{aligned} Z(kY) = 1 + \frac{1}{k}\Big(& \\ & \frac{1}{72} \sum (\nabla^3_\phi s(\phi_Y^*, Y))_{(pqr)} (\nabla^3 s(\phi_Y^*, Y))_{(def)} \mu^6_{pqrdef} \\ & - \frac{1}{24} \sum (\nabla^4 s(\phi_Y^*, Y))_{(defg)} \mu^4_{defg} \Big) + O(k^{-2}), \end{aligned} \tag{4.47}$$

where m is the dimensionality of Y, μ^6_{pqrdef} and μ^4_{defg} are the sixth and fourth central moments of a multivariate Gaussian with covariance matrix $(\nabla^2 s(\phi_Y^*, Y))^{-1}$. All sums are written in the Einstein summation notation. We remark that the $O(k^{-2})$ error term of this approximation also depends on kY.

What we are really interested in is the quantity $\nabla^2_\psi \log F_k(\psi)$ evaluated at $\psi = kY$. We take the logarithm of (4.46):

$$\begin{aligned}\log F_k(\psi/k) =& \log\Big((2\pi)^{\frac{m}{2}}[\det(k\nabla^2 s(\phi_Y^*,Y)]^{-\frac{1}{2}} \\ & \cdot \exp(-ks(\phi_Y^*,Y))Z(\psi)\Big) \\ =& \log\left((2\pi)^{\frac{m}{2}}\right) - \frac{1}{2}\log\left([\det(k\nabla^2 s(\phi_Y^*,Y))]\right) \\ & - ks(\phi_Y^*,Y) + \log(Z(\psi)). \end{aligned} \tag{4.48}$$

We define new functions $\tilde{s_0}, \tilde{s_1}, \tilde{s_2}$ to simplify the analysis.

$$\tilde{s_0}(Y) = s(\phi_Y^*,Y) = s(g(Y),Y) \tag{4.49}$$

$$\tilde{s_1}(Y) = \nabla_\phi s(\phi_Y^*,Y) = \nabla_\phi s(g(Y),Y) \tag{4.50}$$

$$\tilde{s_2}(Y) = \nabla_\phi^2 s(\phi_Y^*,Y) = \nabla_\phi^2 s(g(Y),Y) \tag{4.51}$$

By assumptions (3) and (4) we know these functions are continuously differentiable on $\overline{\mathscr{B}(Y_0,\delta_1)}$ as they are the composition of continuously differentiable functions on the compact set $\overline{\mathscr{B}(Y_0,\delta_1)}$.

We next look at the first derivative of (4.48). We remark that the partial derivatives of $\log\det(X)$ are given by $X^{-\intercal}$.

$$\begin{aligned}\nabla_\psi \log F_k(\psi/k) =& \nabla_\psi[-\frac{1}{2}\log\left([\det(k\tilde{s_2}(\psi/k)]\right)] \\ & - \nabla_\psi[k\tilde{s_0}(\psi/k)] + \nabla_\psi \log(Z(\psi)) \\ =& -\frac{1}{2}(k\tilde{s_2}(\psi/k)))^{-\intercal}\frac{1}{k} \\ & + \tilde{s_1}(\psi/k) + \frac{\nabla_\psi(Z(\psi))}{Z(\psi)} \end{aligned} \tag{4.52}$$

Now that we have an expression for $\nabla_\psi \log F_k(\psi/k)$, we take yet another derivative w.r.t. to ψ to get our desired ∇^2_ψ.

$$\begin{aligned}\nabla^2_\psi \log F_k(\psi/k) =& \nabla_\psi[-\frac{1}{2}(k\tilde{s_2}(\psi/k)))^{-\intercal}\frac{1}{k}] \\ & + \nabla_\psi[\tilde{s_1}(\psi/k)] \\ & + \nabla_\psi[\frac{\nabla_\psi(Z(\psi))}{Z(\psi)}] \end{aligned} \tag{4.53}$$

Let us consider each of the three terms on the right side of (4.53) in isolation. For the first term, we introduce yet another function $\tilde{s}_{-2}(Y)$, the composition of $\tilde{s_2}$ with the matrix inversion.

$$\tilde{s}_{-2}(Y) = (\tilde{s_2}(Y))^{-1}$$

With this new function in hand, we further condense the first term of (4.53).

$$\begin{aligned}\nabla_\psi[-\frac{1}{2}(k\tilde{s_2}(\psi/k)))^{-\intercal}\frac{1}{k}] =& \nabla_\psi[-\frac{1}{2k}(\tilde{s}_{-2}(\psi/k)))\frac{1}{k}] \\ =& -\frac{1}{2k^3}\nabla_Y \tilde{s}_{-2}(\psi/k) \\ =& O(k^{-3}) \end{aligned} \tag{4.54}$$

We previously remarked that $\tilde{s_2}$ is continuously differentiable on the compact set $\overline{\mathscr{B}(Y_0,\delta_1)}$. Condition (5) informs us that $\tilde{s_2}(Y)$ is bounded away from being a singular matrix on $\overline{\mathscr{B}(Y_0,\delta_1)}$, so the matrix inversion is also uniformly continuous on this compact set. This means $\nabla_Y \tilde{s}_{-2}(\psi/k)$ has a finite supremum over $\overline{\mathscr{B}(Y_0,\delta_1)}$ and thus we can say this term is $O(k^{-3})$ uniformly on this neighborhood.

Next we consider the second term of (4.53).

$$\nabla_\psi[\tilde{s_1}(\psi/k)] = \frac{1}{k}\tilde{s_2}(\psi/k) = O(k^{-1}) \tag{4.55}$$

From the continuity of $\tilde{s_2}(\psi/k)$ on our compact neighborhood, we know $\tilde{s_2}(Y)$ has a finite supremum over the compact set $\overline{\mathscr{B}(Y_0,\delta_1)}$, which gives the uniform $O(k^{-1})$ bound.

Finally, we must consider the third term of (4.53).

$$\nabla_\psi[\frac{\nabla_\psi(Z(\psi))}{Z(\psi)}] = \frac{\nabla^2(Z(\psi))}{Z(\psi)} - \frac{\nabla(Z(\psi))(\nabla(Z(\psi)))^\intercal}{Z(\psi)^2} \tag{4.56}$$

Recall that $Z(\psi)$ had a local $O(k^{-2})$ error term as given by [Kass et al., 1990]. We wish to bound the derivatives of $\log F_k(\psi)$, but the local bound on this error term given by Kass does not bound its derivatives. However, a slight modification of the argument of [Kass et al., 1990] shows that our added assumptions about the higher order derivatives are sufficient to control the behavior of this error term. The following expression is their equation (2.2), translated to our setting:

$$\exp(-ks(\phi,Y)) =$$

$$\exp(-ks(\phi_Y^*,Y))\exp(\frac{1}{2}\nabla^2 s(\phi_Y^*,Y)u^2)W(\phi,Y) \tag{4.57}$$

$$\begin{aligned} W(\phi,Y) = 1 &- \frac{1}{6}k^{-1/2}\nabla^3 s(\phi_Y^*,Y)u^3 \\ &+ \frac{1}{72}k^{-1}(\nabla^3 s(\phi_Y^*,Y))^2 u^6 \\ &- \frac{1}{24}k^{-1}\nabla^4 s(\phi_Y^*,Y)u^4 \\ &- \frac{1}{120}k^{-3/2}\nabla^5 s(\phi_Y^*,Y)u^5 \\ &+ \frac{1}{72}k^{-3/2}\nabla^3 A(s(\phi_Y^*,Y))\nabla^4 s(\phi_Y^*,Y)u^7 \\ &+ G(\phi,\phi_Y^*,Y), \end{aligned} \tag{4.58}$$

where $G(\phi,\phi_Y^*,Y)$ is the fifth-order Taylor expansion error term (i.e. it depends on the sixth-order partial derivatives at some ϕ' between ϕ and ϕ_Y^*).

We may continue this Taylor expansion another degree further to bound the variation of $G(\phi,\phi_Y^*,Y)$ for $\phi \in \mathscr{B}(\phi_Y^*,\delta_2)$. We will consider $Z(\psi)$, $\nabla_\psi Z(\psi)$, and $\nabla_\psi^2 Z(\psi)$ as three separate functions, each permitting a higher order Taylor expansion. Each will have their own respective error term depending on the seventh-order partial derivatives at some ϕ', but we note that ϕ' is not necessarily the same for each of them.

The argument of [Kass et al., 1990] already shows how the terms composing their $O(k^{-2})$ error term can be bounded in terms of $\nabla_\phi^6 S(\phi_Y^*,Y)$. It is trivial to show an analogous result for our higher order approximations. This allows us to extend our approximation of $Z(\psi)$ and its derivatives uniformly to the neighborhood $\mathscr{B}(\phi_Y^*,\delta_2)$. The newly introduced extra approximation terms are $O(k^{-v})$ with $v \geq 2$, and so our uniform bounds are still simply $O(k^{-2})$, though with a larger constant now.

Let k be sufficiently large, and let Q,R,S be positive constants satisfying $0 < Q < ||Z(\psi)||$,

$R > k||\nabla_\psi Z(\psi)||$, $S > k||\nabla^2_\psi Z(\psi)||$ for all ψ in $\{\psi|\psi/k \in B(Y_0, \delta)$. We remark that Q exists by virtue of $Z = 1 + O(k^{-1}) + O(k^{-2})$. R and S similarly exist by $||\nabla_\psi Z(\psi)||$ and $||\nabla^2_\psi Z(\psi)||$ both being $O(k^{-1})$ with no constant term in front.

$$\nabla_\psi[\frac{\nabla_\psi(Z(\psi))}{Z(\psi)}] \leq \frac{S}{kQ} - \frac{R^2}{k^2Q^2} \text{for all } Y' \in B(Y_0, \delta)$$

This right hand side is clearly $O(k^{-1})$, and we have uniform bounds across our neighborhood.

$$\nabla_\psi[\frac{\nabla_\psi(Z(\psi))}{Z(\psi)}] = O(k^{-1}) \tag{4.59}$$

Combining the results of (4.54), (4.55), (4.59) with their sum in (4.53), we get this result:

$$||\nabla^2_\psi \log F_k(\psi/k)|| = O(k^{-1}). \tag{4.60}$$

This uniform asymptotic bound then ensures we have the intended result: $\exists C, K$ such that $\forall Y \in \mathscr{B}(Y_0, \delta_1)$ when $k > K$ and $\psi = kY$ we have $||\nabla^2_\psi \log F_k(\psi/k)|| \leq C/k$

□

4.7 Conclusion

This paper studied the practical limitations of using posterior sampling to obtain privacy "for free." We explored an alternative based on the Laplace mechanism, and analyzed it both theoretically and empirically. We illustrated the benefits of the Laplace mechanism for privacy-preserving Bayesian inference to analyze sensitive war records. The study of privacy-preserving Bayesian inference is only just beginning. We envision extensions of these techniques to other approximate inference algorithms, as well as their practical application to sensitive real-world data sets. Finally, we have argued that asymptotic efficiency is important in a privacy context, leading to an open question: how large is the class of private methods that are asymptotically

efficient?

4.8 Acknowledgements

The work of K. Chaudhuri and J. Geumlek was supported in part by NSF under IIS 1253942, and the work of M. Welling was supported in part by Qualcomm, Google and Facebook. We also thank Mijung Park, Eric Nalisnick, and Babak Shahbaba for helpful discussions.

This chapter is based on the material in Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence 2016 (James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri, ”On the theory and practice of privacy-preserving Bayesian data analysis”). The dissertation author was the primary investigator and author of this material.

Chapter 5

Private Posterior Sampling of Exponential Families via Prior Rebalancing

5.1 Introduction

As data analysis continues to expand and permeate ever more facets of life, the concerns over the privacy of one's data grow too. Many results have arrived in recent years to tackle the inherent conflict of extracting usable knowledge from a data set without over-extracting or leaking the private data of individuals. Before one can strike a balance between these competing goals, one needs a framework by which to quantify what it means to preserve an individual's privacy.

Since 2006, Differential Privacy (DP) has reigned as the privacy framework of choice. It quantifies privacy by measuring how indistinguishable the mechanism is across whether or not any one individual is in or out of the data set. This gave not just privacy semantics, but also robust mathematical guarantees. However, the requirements have been cumbersome for utility, leading to many proposed relaxations. One common relaxation is approximate DP, which allows arbitrarily bad events to occur with probability at most δ. A more recent relaxation is Rényi Differential Privacy (RDP) proposed in [Mironov, 2017], which uses the measure of Rényi divergences to smoothly vary between bounding the average and maximum privacy loss. However, RDP has very few mechanisms compared to the more established approximate DP. We expand the RDP repertoire with novel mechanisms inspired by Rényi divergences, as well as

re-analyzing an existing method in this new light.

Inherent to DP and RDP is that there must be some uncertainty in the mechanism; they can not be deterministic. Many privacy methods have been motivated by exploiting pre-existing sources of randomness in machine learning algorithms. One promising area has been Bayesian data analysis, which focuses on maintaining and tracking the uncertainty within probabilistic models. Posterior sampling is prevalent in many Bayesian methods, serving to introduce randomness that matches the currently held uncertainty.

We analyze the privacy arising from posterior sampling as applied to two domains: sampling from exponential family and Bayesian logistic regression. Along with these analyses, we offer tunable mechanisms that can achieve stronger privacy guarantees than directly sampling from the posterior. These mechanisms work via controlling the relative strength of the prior in determining the posterior, building off the common intuition that concentrated prior distributions can prevent overfitting in Bayesian data analysis. We experimentally validate our new methods on synthetic and real data.

5.2 Setup

5.2.1 Privacy Model.

We say two data sets $\mathbf{X}$ and $\mathbf{X}'$ are *neighboring* if they differ in the private record of a single *individual* or person. We use n to refer to the number of records in the data set.

Definition 9. Differential Privacy (DP). A randomized mechanism $\mathscr{A}(\mathbf{X})$ is said to be (ε, δ)-differentially private if for any subset U of the output range of $\mathscr{A}$ and any neighboring data sets $\mathbf{X}$ and $\mathbf{X}'$, we have $\mathsf{Pr}(\mathscr{A}(\mathbf{X}) \in U) \leq \exp(\varepsilon)\mathsf{Pr}(\mathscr{A}(\mathbf{X}') \in U) + \delta$.

DP is concerned with the difference the participation of a individual might have on the output distribution of the mechanism. When $\delta > 0$, it is known as approximate DP while the $\delta = 0$ case is known as pure DP. The requirements for DP can be phrased in terms of a privacy loss variable, a random variable that captures the effective privacy loss of the mechanism output.

Definition 10. Privacy Loss Variable. We can define a random variable Z that measures the privacy loss of a given output of a mechanism across two neighboring data sets $\mathbf{X}$ and $\mathbf{X}'$.

$$Z = \log \frac{\Pr(\mathscr{A}(\mathbf{X}) = o)}{\Pr(\mathscr{A}(\mathbf{X}') = o)}\bigg|_{o \sim \mathscr{A}(\mathbf{X})} \tag{5.1}$$

(ε, δ)-DP is the requirement that for any two neighboring data sets $Z \leq \varepsilon$ with probability at least $1 - \delta$. The exact nature of the trade-off and semantics between ε and δ is subtle, and choosing them appropriately is difficult. For example, setting $\delta = 1/n$ permits $(\varepsilon, \delta)-$DP mechanisms that always violate the privacy of a random individual. However, there are other ways to specify that a random variable is mostly small. One such way is to bound the Rényi divergence of $\mathscr{A}(\mathbf{X})$ and $\mathscr{A}(\mathbf{X}')$.

Definition 11. Rényi Divergence. The Rényi divergence of order λ between the two distributions P and Q is defined as

$$\mathsf{R}_\lambda(P||Q) = \frac{1}{\lambda - 1} \log \int P(o)^\lambda Q(o)^{1-\lambda} do. \tag{5.2}$$

We remark that this notation differs from that seen in 3.2, namely that the parameter α has been renamed to λ. This is to avoid a conflict with the parameters α, β that arise in our discussion of exponential family distributions. This alternative notation will be used consistently for the rest of this chapter.

As $\lambda \to \infty$, Rényi divergence becomes the *max divergence*; moreover, setting $P = \mathscr{A}(\mathbf{X})$ and $Q = \mathscr{A}(\mathbf{X}')$ ensures that $\mathsf{R}_\lambda(P||Q) = \frac{1}{\lambda-1} \log \mathbb{E}_Z[e^{(\lambda-1)Z}]$, where Z is the privacy loss variable. Thus, a bound on the Rényi divergence over all orders $\lambda \in (0, \infty)$ is equivalent to $(\varepsilon, 0)$-DP, and as $\lambda \to 1$, this approaches the expected value of Z equal to $KL(\mathscr{A}(\mathbf{X})||\mathscr{A}(\mathbf{X}'))$. This leads us to Rényi Differential Privacy, a flexible privacy notion that covers this intermediate behavior.

Definition 12. Rényi Differential Privacy (RDP). A randomized mechanism $\mathscr{A}(\mathbf{X})$ is said

to be $(\lambda, \varepsilon)-$Rényi differentially private if for any neighboring data sets $\mathbf{X}$ and $\mathbf{X}'$ we have $\mathsf{R}_\lambda(\mathscr{A}(\mathbf{X})||\mathscr{A}(\mathbf{X}')) \leq \varepsilon$.

The choice of λ in RDP is used to tune how much concern is placed on unlikely large values of Z versus the average value of Z. One can consider a mechanism's privacy as being quantified by the entire curve of ε values associated with each order λ, but the results of [Mironov, 2017] show that almost identical results can be achieved when this curve is known at only a finite collection of possible λ values.

5.2.2 Posterior Sampling.

In Bayesian inference, we have a model class Θ, and are given observations $x_1, \ldots, x_n$ assumed to be drawn from a $\theta \in \Theta$. Our goal is to maintain our beliefs about θ given the observational data in the form of the posterior distribution $\mathsf{Pr}(\theta|x_1, \ldots, x_n)$. This is often done in the form of drawing samples from the posterior.

Our goal in this paper is to develop privacy preserving mechanisms for sampling from the exponential family posterior, which we address in Section 5.3.

5.2.3 Related Work.

Differential privacy has emerged as the gold standard for privacy in a number of data analysis applications – see Dwork and Roth [2014], Sarwate and Chaudhuri [2013] for surveys. Since enforcing pure DP sometimes requires the addition of high noise, a number of relaxations have been proposed in the literature. The most popular relaxation is approximate DP Dwork et al. [2006a], and a number of uniquely approximate DP mechanisms have been designed by Dwork and Lei [2009], Thakurta and Smith [2013], Chaudhuri et al. [2014], Bun et al. [2015] among others. However, while this relaxation has some nice properties, recent work Mironov [2017], McSherry [2017] has argued that it can also lead privacy pitfalls in some cases. Approximate differential privacy is also related to, but is weaker than, the closely related δ-probabilistic

privacy Machanavajjhala et al. [2008] and $(1, \varepsilon, \delta)$-indistinguishability Chaudhuri and Mishra [2006].

Our privacy definition of choice is Rényi differential privacy Mironov [2017], which is motivated by two recent relaxations – concentrated DP Dwork and Rothblum [2016] and z-CDP Bun and Steinke [2016]. Concentrated DP has two parameters, μ and τ, controlling the mean and concentration of the privacy loss variable. Given a privacy parameter α, z-CDP essentially requires $(\lambda, \alpha\lambda)$-RDP for all λ. While Bun and Steinke [2016], Dwork and Rothblum [2016], Mironov [2017] establish tighter bounds on the privacy of existing differentially private and approximate DP mechanisms, we provide mechanisms based on posterior sampling from exponential families that are uniquely RDP. RDP is also a generalization of the notion of KL-privacy Wang et al. [2016], which has been shown to be related to generalization in machine learning.

There has also been some recent work on privacy properties of Bayesian posterior sampling; however most of the work has focused on establishing pure or approximate DP. Dimitrakakis et al. [2014] establishes conditions under which some popular Bayesian posterior sampling procedures directly satisfy pure or approximate DP. Wang et al. [2015b] provides a pure DP way to sample from a posterior that satisfies certain mild conditions by raising the temperature. Chapter 4 and Zhang et al. [2016] provide a simple statistically efficient algorithm for sampling from exponential family posteriors. Minami et al. [2016] shows that directly sampling from the posterior of certain GLMs, such as logistic regression, with the right parameters provides approximate differential privacy. While our work draws inspiration from all Dimitrakakis et al. [2014], Wang et al. [2015b], Minami et al. [2016], the main difference between their and our work is that we provide RDP guarantees.

5.2.4 Background: Exponential Families

This section will give a in-depth explanation of exponential families and the properties of them we exploit in our analysis.

An exponential family of distributions takes the following form, indexed by a parameter $\theta \in \Theta$:

$$\Pr(x_1,\ldots,x_n|\theta) = (\prod_{i=1}^{n} h(x_i)) \exp((\sum_{i=1}^{n} S(x_i)) \cdot \theta - n \cdot A(\theta)) \ . \tag{5.3}$$

We call h the base measure, S the sufficient statistics of x, and A as the log-partition function of this family. Note that the data $\{x_1,\ldots,x_n\}$ interact with the parameter θ solely through the dot product of θ and the sum of their sufficient statistics. When the parameter θ is used in this dot product unmodified (as in (5.3)), we call this a natural parameterization. Our analysis will be restricted to the families that satisfy the following two properties:

Definition 13. An exponential family is minimal if the coordinates of the function S are not almost surely linearly dependent, and the interior of Θ is non-empty.

Definition 14. For any for $\Delta \in \mathbb{R}$, an exponential family is $\Delta-$bounded if
$\Delta \geq \sup_{x,y\in\mathscr{X}} ||S(x) - S(y)||$.

When a family is minimal, the log-partition function A has many interesting characteristics. It can be defined as $A(\theta) = \log \int_{\mathscr{X}} h(x) \exp(S(x)\cdot\theta)dx$, and serves to normalize the distribution. Its derivatives form the cumulants of the distribution, that is to say $\nabla A(\theta) = \kappa_1 = \mathbb{E}_{x|\theta}[S(x)]$ and $\nabla^2 A(\theta) = \kappa_2 = \mathbb{E}_{x|\theta}[(S(x)-\kappa_1)(S(x)-\kappa_1)^\intercal]$. This second cumulant is also the covariance of $S(x)$, which demonstrates that $A(\theta)$ must be a convex function since covariances must be positive semidefinite.

In Bayesian data analysis, we are interested in finding our posterior distribution over the parameter θ that generated the data. We must introduce a prior distribution $\Pr(\theta|\eta)$ to describe our initial beliefs on θ, where η is a parameterization of our family of priors.

$$\Pr(\theta|x_1,\ldots,x_n,\eta) \propto \Pr(x_1,\ldots,x_n|\theta)\Pr(\theta|\eta) \tag{5.4}$$

$$\propto (\prod_{i=1}^{n} h(x_i)) \exp((\sum_{i=1}^{n} S(x_i)) \cdot \theta - n \cdot A(\theta))\Pr(\theta|\eta) \tag{5.5}$$

$$\propto \exp((\sum_{i=1}^{n} S(x_i), n) \cdot (\theta, -A(\theta)))\Pr(\theta|\eta) \tag{5.6}$$

$$\tag{5.7}$$

Notice that we can ignore the $(\prod_{i=1}^{n} h(x_i))$ as it is a constant that will be normalized out. If we let our prior take the form of another exponential family $\Pr(\theta|\eta) = \exp(T(\theta)\cdot\eta - B(\eta))$ where $T(\theta) = (\theta, -A(\theta))$ and $B(\eta) = \log \int_\Theta \exp(T(\theta)\cdot\eta) d\theta$, the we can perform these manipulations,

$$\Pr(\theta|x_1,\ldots,x_n,\eta) \propto \exp((\sum_{i=1}^{n} S(x_i), n) \cdot T(\theta) + \eta \cdot T(\theta) - B(\eta)) \tag{5.8}$$

$$\propto \exp(\left(\eta + (\sum_{i=1}^{n} S(x_i), n)\right) \cdot T(\theta) - B(\eta)) \tag{5.9}$$

and see that expression (5.9) can be written as

$$\Pr(\theta|\eta') = \exp(T(\theta)\cdot\eta' - C(\eta')) \tag{5.10}$$

where $\eta' = \eta + \sum_{i=1}^{n}(S(x_i), 1)$ and $C(\eta')$ is chosen such that the distribution is normalized.

This family of posteriors is precisely the same exponential family that we chose for our prior. We call this a conjugate prior, and it offers us an efficient way of finding the parameter of our posterior: $\eta_{posterior} = \eta_{prior} + \sum_{i=1}^{n}(S(x_i), 1)$. Within this family, $T(\theta)$ forms the sufficient

statistics of θ, and the derivatives of $C(\eta)$ give the cumulants of these sufficient statistics.

Beta-Bernoulli System.

A specific example of an exponential family that we will be interested in is the Beta-Bernoulli system, where an individual's data is a single i.i.d. bit modeled as a Bernoulli variable with parameter ρ, along with a Beta conjugate prior.

$$\Pr(x_1,\ldots,x_n|\rho) = \prod_{i=1}^{n} \rho^{x_i}(1-\rho)^{1-x_i} \tag{5.11}$$

Letting $\theta = \log(\frac{\rho}{1-\rho})$ and $A(\theta) = \log(1+\exp(\theta)) = -\log(1-\rho)$, we can rewrite the equation as follows:

$$\Pr(x_1,\ldots,x_n|\rho) = \prod_{i=1}^{n} (\frac{\rho}{1-\rho})^{x_i}(1-\rho) \tag{5.12}$$

$$= \exp(\sum_{i=1}^{n} x_i \log(\frac{\rho}{1-\rho}) + \log(1-\rho)) \tag{5.13}$$

$$= \exp((\sum_{i=1}^{n} x_i)\cdot\theta - A(\theta)). \tag{5.14}$$

This system satisfies the properties we require, as this natural parameterization with θ is both minimal and $\Delta-$bounded for $\Delta = 1$.

As our mechanisms are interested mainly in the posterior, the rest of this section will be written with respect the family specified by equation (5.10).

Now that we have the notation for our distributions, we can write out the expression for the Rényi divergence of two posterior distributions P and Q (parameterized by η_P and η_Q) from the same exponential family. This expression allows us to directly compute the Rényi divergences of posterior sampling methods, and forms the crux of the analysis of our exponential family mechanisms.

Observation 5. *Let P and Q be two posterior distributions from the same exponential family*

that are parameterized by η_P and η_Q. Then,

$$\mathsf{R}_\lambda(P||Q) = \frac{1}{\lambda-1}\log\left(\int_\Theta P(\theta)^\lambda Q(\theta)^{1-\lambda} d\theta\right) = \frac{C(\lambda\eta_P + (1-\lambda)\eta_Q) - \lambda C(\eta_P)}{\lambda-1} + C(\eta_Q). \tag{5.15}$$

To help analyze the implication of equation (5.15) for Rényi Differential Privacy, we introduce a some more helpful definitions.

Definition 15. We say a posterior parameter η is normalizable if $C(\eta) = \log\int_\Theta \exp(T(\theta)\cdot\eta))d\theta$ is finite.

Let E denote the set of all normalizable η for the conjugate prior family.

Definition 16. Let $pset(\eta_0, n)$ be the convex hull of all parameters η of the form $\eta_0 + n(S(x),1)$ for $x \in \mathscr{X}$. When n is an integer this represents the hull of possible posterior parameters after observing n data points starting with the prior η_0.

Definition 17. Let $Diff$ be the difference set for the family, where $Diff$ is the convex hull of all vectors of the form $(S(x) - S(y), 0)$ for $x, y \in \mathscr{X}$.

Definition 18. Two posterior parameters η_1 and η_2 are neighboring iff $\eta_1 - \eta_2 \in Diff$.
They are $r-$neighboring iff $(\eta_1 - \eta_2)/r \in Diff$.

5.3 Mechanisms and Privacy Guarantees

We begin with our simplest mechanism, Direct Sampling, which samples according to the true posterior. This mechanism is presented as Algorithm 1.

Algorithm 1: Direct Posterior

Input: Dataset $D = (x_1, \ldots, x_n)$, prior parameter η_0
Sample $\theta \sim \mathsf{Pr}(\theta|\eta')$ where $\eta' = \eta_0 + \sum_{i=1}^n (S(x_i), 1)$
return θ

Even though Algorithm 1 is generally not differentially private Dimitrakakis et al. [2014], Theorem 7 suggests that it offers RDP for $\Delta-$bounded exponential families and certain orders λ.

Theorem 7. *For a $\Delta-$bounded minimal exponential family of distributions $\Pr(x|\theta)$ with continuous log-partition function $A(\theta)$, there exists $\lambda^* \in (1,\infty]$ such Algorithm 1 achieves $(\lambda, \varepsilon(\eta_0, n, \lambda))$-RDP for $\lambda < \lambda^*$.*

λ^* is the supremum over all λ such that all η in the set $\eta_0 + (\lambda - 1)Diff$ are normalizable.

Corollary 15. *For the Beta-Bernoulli system with a prior $Beta(\alpha_0, \beta_0)$, Algorithm 1 achieves $(\lambda, \varepsilon)-$RDP iff $\lambda > 1$ and $\lambda < 1 + min(\alpha_0, \beta_0)$.*

Notice the implication of Corollary 15: for any η_0 and $n > 0$, there exists finite λ such that direct posterior sampling does not guarantee (λ, ε)-RDP for any finite ε. This also prevents $(\varepsilon, 0)$-DP as an achievable goal as well. Algorithm 1 is inflexible; it offers us no way to change the privacy guarantee.

This motivates us to propose two different modifications to Algorithm 1 that are capable of achieving arbitrary privacy parameters. Algorithm 2 modifies the contribution of the data $\mathbf{X}$ to the posterior, while Algorithm 3 modifies the contribution of the prior η_0.

Algorithm 2: Diffused Posterior

Input: Dataset $D = (x_1, \ldots, x_n)$, prior parameter η_0, privacy parameters (λ, ε)
Find $r \in (0, 1]$ such that $\forall r$-neighboring $\eta_P, \eta_Q \in pset(\eta_0, rn)$ we have
$\mathsf{R}_\lambda(\Pr(\theta|\eta_P)||\Pr(\theta|\eta_Q)) \leq \varepsilon$
Sample $\theta \sim \Pr(\theta|\eta')$ where $\eta' = \eta_0 + r\sum_{i=1}^{n}(S(x_i), 1)$
return θ

Theorem 8. *For any $\Delta-$bounded minimal exponential family with prior η_0 in the interior of E, any $\lambda > 1$, and any $\varepsilon > 0$, there exists $r^* \in (0, 1]$ such that using $r \in (0, r^*]$ in Algorithm 2 will achieve $(\lambda, \varepsilon)-$RDP.*

Algorithm 3: Concentrated Posterior

Input: Dataset $D = (x_1, \ldots, x_n)$, prior parameter η_0, privacy parameters (λ, ε)
Find $m \in (0, 1]$ such that $\forall$neighboring $\eta_P, \eta_Q \in pset(\eta_0/m, n)$,
$\mathsf{R}_\lambda(\mathsf{Pr}(\theta|\eta_P)||\mathsf{Pr}(\theta|\eta_Q)) \leq \varepsilon$
Sample $\theta \sim \mathsf{Pr}(\theta|\eta')$ where $\eta' = \eta_0/m + \sum_{i=1}^{n}(S(x_i), 1)$
return θ

Theorem 9. *For any $\Delta-$bounded minimal exponential family with prior η_0 in the interior of E, any $\lambda > 1$, and any $\varepsilon > 0$, there exists $m^* \in (0, 1]$ such that using $m \in (0, m^*]$ in Algorithm 3 will achieve $(\lambda, \varepsilon)-$RDP.*

We have not yet specified how to find the appropriate values of r or m, and the condition requires checking the supremum of divergences across the appropriate *pset* range of parameters. However, with an additional assumption this supremum of divergences can be efficiently computed.

Theorem 10. *Let $e(\eta_P, \eta_Q, \lambda) = \mathsf{R}_\lambda\left(\mathsf{Pr}(\theta|\eta_P)||\mathsf{Pr}(\theta|\eta_Q)\right)$. For a fixed λ and fixed η_P, the function e is a convex function over η_Q.*

If for any direction $v \in Diff$, the function $g_v(\eta) = v^\intercal \nabla^2 C(\eta) v$ is convex over η, then for a fixed λ, the function $f_\lambda(\eta_P) = \sup_{\eta_Q r-neighboring\ \eta_P} e(\eta_P, \eta_Q, \lambda)$ is convex over η_P in the directions spanned by $Diff$.

Corollary 16. *The Beta-Bernoulli system satisfies the conditions of Theorem 10 since the functions $g_v(\eta)$ have the form $(v^{(1)})^2(\psi_1(\eta^{(1)}) + \psi_1(\eta^{(2)} - \eta^{(1)}))$, and ψ_1 is the digamma function. Both pset and $Diff$ are defined as convex sets. The expression $\sup_{r-neighboring\ \eta_P, \eta_Q \in pset(\eta_0, n)}$ $\mathsf{R}_\lambda(\mathsf{Pr}(\theta|\eta_P)||\mathsf{Pr}(\theta|\eta_Q))$ is therefore equivalent to the maximum of $\mathsf{R}_\lambda(\mathsf{Pr}(\theta|\eta_P)||\mathsf{Pr}(\theta|\eta_Q))$ where $\eta_P \in \eta_0 + \{(0, n), (n, n)\}$ and $\eta_Q \in \eta_P \pm (r, 0)$.*

We can do a binary search over $(0, 1]$ to find an appropriate value of r or m. At each candidate value, we only need to consider the boundary situations to evaluate the supremum and check the RDP guarantee. These boundary situations depend on the choice of model, and not

the data size n. For example, in the Beta-Bernoulli system, evaluating the supremum involves calculating the Rényi divergence across at most 4 pairs of distributions, as in Corollary 16.

Eventually, the search process will find a valid non-zero choice for r or m. If stopped early and none of the tested candidate values satisfy the privacy constraint, the analyst can either continue to iterate or decide not to release anything.

5.3.1 Extension: Public Data for Exponential Families

The use of a conjugate prior makes the interaction of observed data versus the prior easy to see. The prior η_0 can be expressed as $(\alpha\chi, \alpha)$, where χ is a vector expressing the average sufficient statistics of pseudo-observations and α represents a count of these pseudo-observations. After witnessing the n data points, the posterior becomes a prior that has averaged the data sufficient statistics into a new χ and added n to α.

If the data analyst had some data in addition to $\mathbf{X}$ that was not privacy sensitive, perhaps from a stale data set for which privacy requirements have lapsed, then this data can be used to form a better prior for the analysis.

Not only would this improve utility by adding information that can be fully exploited, it would also in most cases improve the privacy guarantees as well. A stronger prior, especially a prior farther from the boundaries where $C(\eta)$ becomes infinite, will lead to smaller Rényi divergences. This is effectively the same behavior as the Concentrated Sampling mechanism, which scales the prior to imagine more pseudo-observations had been seen. This also could apply to settings in which the analyst can adaptively pay to receive non-private data, since this method will inform us once our prior formed from this data becomes strong enough to sample directly at our desired RDP level.

This also carries another privacy implication for partial data breaches. If an adversary learns the data of some individuals in the data set, the Direct Sampling mechanism's privacy guarantee for the remaining individuals can actually improve. Any contributions of the affected individuals to the posterior become in effect yet more public data placed in the prior. The privacy

analysis and subsequent guarantees will match the setting in which this strengthened prior was used.

5.3.2 Extension: Releasing the result of a Statistical Query

Here we are given a sensitive database $\mathbf{X} = \{x_1, \ldots, x_n\}$ and a predicate $\phi(\cdot)$ which maps each x_i into the interval $[0,1]$. Our goal is to release a Rényi DP approximation to the quantity: $F(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$.

Observe that directly releasing $F(\mathbf{X})$ is neither DP nor Rényi DP, since this is a deterministic algorithm; our goal is to release a random sample from a suitable distribution so that the output is as close to $F(\mathbf{X})$ as possible.

The task of releasing a privatized result of a statistical query can be embedded into our Beta-Bernoulli system. This allows the privatized statistical query release to be done using either Algorithm 2 or Algorithm 3.

We can extend the Beta-Bernoulli model to allow the sufficient statistics $S(x)$ to range over the interval $[0,1]$ instead of just the discrete set $\{0,1\}$. This alteration still results in a Δ-bounded exponential family, and the privacy results hold.

The sampled posterior will be a Beta distribution that will concentrate around the mean of the data observations and the pseudo-observations of the prior. The process is described in the Beta-Sampled Statistical Query algorithm. The final transformation maps the natural parameter $\theta \in (-\infty, \infty)$ onto the mean of the distribution $\rho \in (0,1)$.

Algorithm 4: Beta-Sampled Statistical Query

Input: Dataset $D = (x_1, \ldots, x_n)$, prior parameter η_0, query function f, privacy parameters (λ, ε)

Compute $\mathbf{X}_f = \{f(x_1), \ldots, f(x_n)\}$

Sample θ via Algorithm 2 or Algorithm 3 applied to $\mathbf{X}_f$ with η_0, ε, and λ.

Compute $\rho = \frac{\exp(\theta)}{1+\exp(\theta)}$

return ρ

5.4 Experiments

In this section, we present the experimental results for our proposed algorithms for both exponential family and GLMs. Our experimental design focuses on two goals – first, analyzing the relationship between λ and ε in our privacy guarantees and second, exploring the privacy-utility trade-off of our proposed methods in relation to existing methods.

5.4.1 Synthetic Data: Beta-Bernoulli Sampling Experiments

In this section, we consider posterior sampling in the Beta-Bernoulli system. We compare three algorithms. As a baseline, we select a modified version of the algorithm in Chapter 4, which privatizes the sufficient statistic of the data to create a privatized posterior. Instead of Laplace noise that is used in Chapter 4, we use Gaussian noise to do the privatization; Mironov [2017] shows that if Gaussian noise with variance σ^2 is added, then this offers an RDP guarantee of $(\lambda, \lambda\frac{\Delta^2}{\sigma^2})$ for Δ-bounded exponential families. We also consider the two algorithms presented in Section 5.3 – Algorithm 2 and 3; observe that Algorithm 1 is a special case of both. 500 iterations of binary search were used to select r and m when needed.

Achievable Privacy Levels.

We plot the $(\lambda, \varepsilon)-$RDP parameters achievable by the Algorithm 2 and Algorithm 3. These parameters are plotted for a prior $\eta_0 = (6, 18)$ and the data size $n = 100$ which are selected arbitrarily for illustrative purposes. We plot over six values $\{0.1, 0.3, 0.5, 0.7, 0.9, 1\}$ of the scaling constants r and m. The results are presented in Figures 5.1 and 5.2. Our primary observation is the presence of the vertical asymptotes for our proposed methods. As r and m decrease, the ε guarantees improve and become finite at larger orders λ, but a vertical asymptote still exists. The results of the baseline are not plotted: it simply achieves RDP along any line of positive slope passing through the origin.

Privacy-Utility Tradeoff.

We next evaluate the privacy-utility tradeoff of the algorithms by plotting $KL(P||\mathscr{A})$ as a function of ε with λ fixed, where P is the true posterior and $\mathscr{A}$ is the output distribution of a mechanism. For Algorithms 2 and 3, the KL divergence can be evaluated in closed form. For the Gaussian mechanism, numerical integration was used to evaluate the KL divergence integral. We have arbitrarily chosen $\eta_0 = (6, 18)$ and data set $\mathbf{X}$ with 100 total trials and 38 successful trials. We have plotted the resulting divergences over a range of ε for $\lambda = 2$ in Figure 5.3 and for $\lambda = 15$ in Figure 5.4. When $\lambda = 2 < \lambda^*$, both Algorithms 2 and 3 reach zero KL divergence once direct sampling is possible. The Gaussian mechanism must always add nonzero noise. As $\varepsilon \to 0$, Algorithm 3 approaches a point mass distribution heavily penalized by the KL divergence. Due to its projection step, the Gaussian Mechanism follows a bimodal distribution as $\varepsilon \to 0$. Algorithm 2 degrades to the prior, with modest KL divergence. When $\lambda = 15 > \lambda^*$, the divergences for Algorithms 2 and 3 are bounded away from 0, while the Gaussian mechanism still approaches the truth as $\varepsilon \to \infty$.

Finally, we plot $\log \mathsf{Pr}(\mathbf{X}_H|\theta)$ as a function of ε, where θ comes from one of the mechanisms applied to $\mathbf{X}$. Both $\mathbf{X}$ and $\mathbf{X}_H$ consist of 100 Bernoulli trials with proportion parameter $\rho = 0.5$. This experiment was run 10000 times, and we report the mean and standard deviation. Similar to the previous section, we have a fixed prior of $\eta_0 = (6, 18)$. The results are shown for $\lambda = 2$ in Figure 5.5 and for $\lambda = 15$ in Figure 5.6. These results agree with the limit behaviors in the KL test. This experiment is more favorable for Algorithm 3, as it degrades only to the log likelihood under the mode of the prior.

5.5 Proofs of Exponential Family Sampling Theorems

Our proofs will make extensive use of the definitions laid out in Section 5.2.4. We will however need an additional definition for a modified version of *pset*, and as well the set of possible updates to the posterior parameter that might arise from the data.

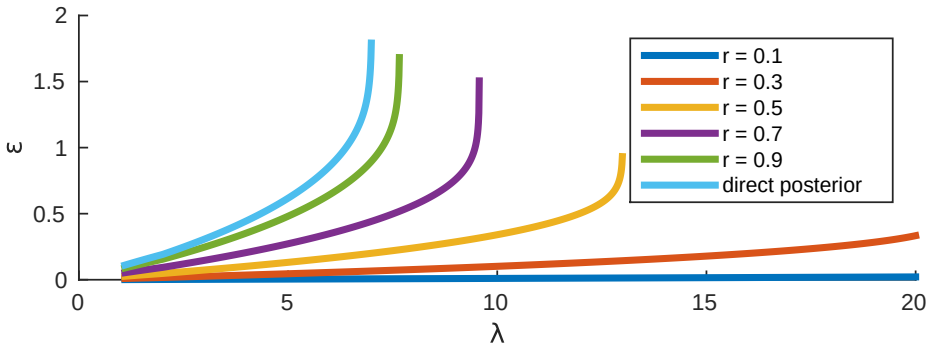

Figure 5.1. Achievable $(\lambda, \varepsilon)-$RDP Levels for Algorithm 2

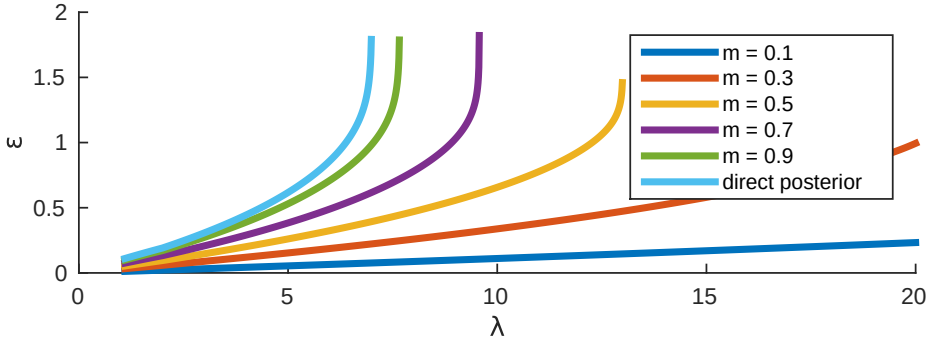

Figure 5.2. Achievable $(\lambda, \varepsilon)-$RDP Levels for Algorithm 3

Definition 19. Let $lpset(\eta_0, n, b) = pset(\eta_0, n) + bDiff$. This is the set of posterior parameters that are $b-$neighboring at least one of the elements of $pset(\eta_0, n)$

Definition 20. Let U be the set of posterior updates for an exponential family, where U is the convex hull of all vectors of the form $(S(x), 1)$ for $x, y \in \mathscr{X}$.

We begin by noting that observing a data set when starting at a normalizable prior η_0 must result in a normalizable posterior parameter η'.

Observation 6. *In a minimal exponential family, for any prior parameter η_0, any $n > 0$, and any posterior update, every possible posterior parameter in the set $\eta_0 + nU$ is also normalizable. As $C(\eta)$ must be a convex function for minimal families, this must apply to positive non-integer values of n as well.*

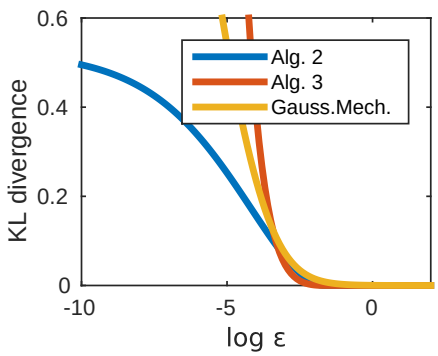

Figure 5.3. KL divergences, where $\lambda = 2 < \lambda^*$

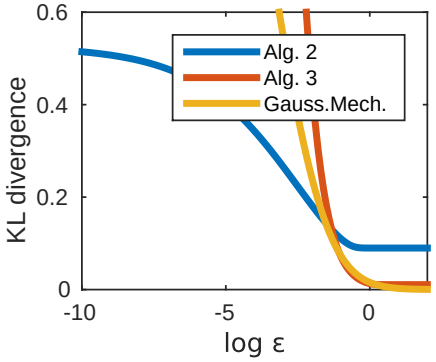

Figure 5.4. KL divergences, where $\lambda = 15 > \lambda^*$

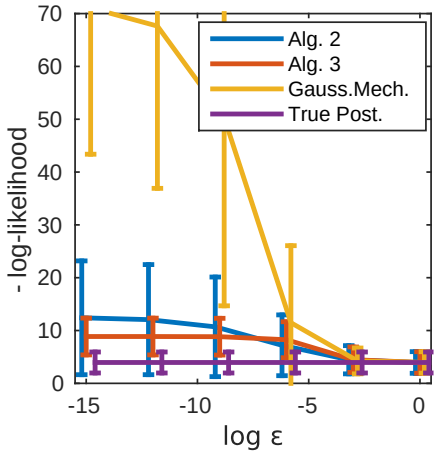

Figure 5.5. $-\log \Pr(\mathbf{X}_H)$, where $\lambda = 2 < \lambda^*$

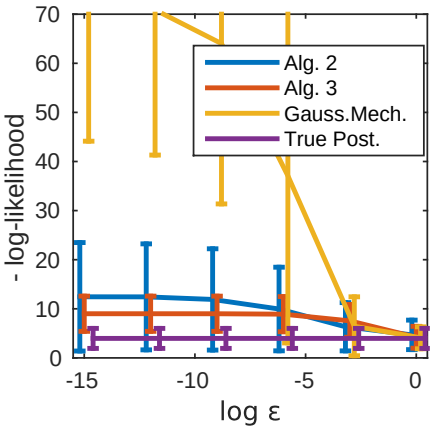

Figure 5.6. $-\log \Pr(\mathbf{X}_H)$, where $\lambda = 15 > \lambda^*$

With this observation, we are ready to prove our result on the conditions under which sampling from our posterior gives a finite $(\lambda, \varepsilon)-$RDP guarantee.

Theorem 11. *Theorem 7 revisited.*

For a $\Delta-$bounded minimal exponential family of distributions $\Pr(x|\theta)$ with continuous log-partition function $A(\theta)$, there exists $\lambda^* \in (1, \infty]$ such Algorithm 1 achieves $(\lambda, \varepsilon(\eta_0, n, \lambda))$-RDP for $\lambda < \lambda^*$.

λ^* is the supremum over all λ such that all η in the set $\eta_0 + (\lambda - 1)Diff$ are normalizable.

PROOF:

Algorithm 1 samples directly from the posterior $\eta_{post} = \eta_0 + \sum_i (S(x_i), 1)$. When applied to neighboring data sets $\mathbf{X}$ and $\mathbf{X}'$, it selects posterior parameters that are neighboring.

The theorem can be reinterpreted as saying there exists λ^* such that for $\lambda < \lambda^*$ we have

$$\sup_{\text{neighboring } \eta_P, \eta_Q \in pset(\eta_0, n)} \mathsf{R}_\lambda(p(\theta|\eta_P)||\mathsf{Pr}(\theta|\eta_Q)) < \infty. \tag{5.16}$$

For these two posteriors from the same exponential family, we can write out the Rényi divergence in terms of the log-partition function $C(\eta)$.

$$\mathsf{R}_\lambda(p(\theta|\eta_P)||\mathsf{Pr}(\theta|\eta_Q)) = \frac{C(\lambda\eta_P + (1-\lambda)\eta_Q) - \lambda C(\eta_P)}{\lambda - 1} + C(\eta_Q) \tag{5.17}$$

We wish to show that this is bounded above over all neighboring η_P and η_Q our mechanism might generate, and will do so by showing that $|C(\eta)|$ must be bounded every where it is applied in equation (5.17) if $\lambda < \lambda^*$. To find this bound, we will ultimately show each potential application of $C(\eta)$ lies within a closed subset of E, from which the continuity of C will imply an upper bound.

Let's begin by observing that η_P and η_Q must lie within $pset(\eta_0, n)$ as they arise as posteriors for neighboring data sets $\mathbf{X}$ and $\mathbf{X}'$. The point $\eta_L = \lambda\eta_P + (1-\lambda)\eta_Q = \eta_P + (\lambda - 1)(\eta_P - \eta_Q)$ might not lie within $pset(\eta_0, n)$. However, we know $\eta_P - \eta_Q$ lies within $Diff$ and that $\eta_L - \eta_P$ is within $(\lambda - 1)Diff$. This means for any neighboring data sets, η_P, η_Q, and η_L lie inside $lpset(\eta_0, n, \lambda - 1)$.

If $\lambda < \lambda^*$, then $\eta_0 + (\lambda - 1)Diff \subseteq E$. The set $\eta_0 + (\lambda - 1)Diff$ is potentially an open set, but the closure of this set must be within E as well, since we can always construct $\lambda' \in (\lambda, \lambda^*)$ where $\eta_0 + (\lambda' - 1)Diff \subseteq E$, and the points inside $\eta_0 + (\lambda - 1)Diff$ can't converge to any point outside of $\eta_0 + (\lambda' - 1)Diff$.

Any point in $\eta \in lpset(\eta_0, n, \lambda - 1)$ can be broken down into three components using the definition of $lpset$: $\eta = \eta_0 + u + d$, where $u \in nU$ and $d \in (\lambda - 1)Diff$. For any point in this

$lpset$, we can therefore subtract off the component u to reach a point in the set $\eta_0 + (\lambda - 1)Diff$. With Observation 6, we can conclude that η is normalizable if $\eta - u$ is normalizable, and therefore the closure of $lpset(\eta_0, n, \lambda - 1)$ is a subset of E if $\eta_0 + (\lambda - 1)Diff$ is a subset of E, which we have shown for $\lambda < \lambda^*$.

As $C(\eta)$ is a continuous function, we know that the supremum of $|C(\eta)|$ over the closure of $lpset(\eta_0, n, \lambda - 1)$ must be finite. Remember that for any neighboring data sets, η_P,η_Q, and η_L are inside $lpset(\eta_0, n, \lambda - 1)$. Since $|C(\eta)|$ is bounded over this $lpset$, so too must our expression for $\mathsf{R}_\lambda(\mathsf{Pr}(\theta|\eta_P)||\mathsf{Pr}(\theta|\eta_Q))$ in equation (5.17). Therefore there exists an upper-bound for the order λ Rényi divergence across all pairs of posterior parameters selected by Algorithm 1 on neighboring data sets. This finite upper-bound provides a finite value for $\varepsilon(\eta_0, n, \lambda)$ for which Algorithm 1 offers $(\lambda, \varepsilon(\eta_0, n, \lambda))$-RDP .

□

To prove our results for Algorithm 2 and Algorithm 3, we'll need an additional result that bounds the Rényi divergence in terms of the Hessian of the log-partition function and the distance between the two distribution parameters.

Lemma 17. *For $\lambda > 1$, if $||\nabla^2 C(\eta)|| < H$ over the set $\{\eta_P + x(\eta_P - \eta_Q)|x \in [-\lambda + 1, \lambda - 1]\}$, then*

$$\mathsf{R}_\lambda(\mathsf{Pr}(\theta|\eta_P)||\mathsf{Pr}(\theta|\eta_Q)) \leq ||\eta_P - \eta_Q||^2 H\lambda \tag{5.18}$$

PROOF:

Define the function $g(x) = C(\eta_P + xv)$ where $x \in \mathbb{R}$ and $v = \eta_P - \eta_Q$. This allows us to rewrite the Rényi divergence as

$$\mathsf{R}_\lambda(P||Q) = \frac{g(1-\lambda) - \lambda g(0)}{\lambda - 1} + g(1) \tag{5.19}$$

Now we will replace g with its first order Taylor expansion

$$g(x) = g(0) + xg'(0) + e(x) \tag{5.20}$$

where $e(x)$ is the approximation error term, satisfying $|e(x)| \leq x^2 \max_{y\in[-x,x]} g''(y)/2$. This results in

$$\mathsf{R}_\lambda(\mathsf{Pr}(\theta|\eta_P)||\mathsf{Pr}(\theta|\eta_Q)) = \frac{g(0) + (1-\lambda)g'(0) + e(1-\lambda) - \lambda g(0)}{\lambda - 1} + g(0) + g'(0) + e(1) \tag{5.21}$$

$$= -\frac{e(1-\lambda)}{\lambda - 1} + e(1) \tag{5.22}$$

$$\leq \frac{(\lambda-1)^2}{\lambda-1} \max_{y\in[-\lambda+1,\lambda-1]} g''(y)/2 + \max_{y\in[-1,1]} g''(y)/2\,. \tag{5.23}$$

Further, we can express g'' in terms of C and v.

$$g''(y) = v^\intercal \nabla^2 C(\eta_P + yv)v \tag{5.24}$$

$$\leq ||\eta_P - \eta_Q||^2 ||\nabla^2 C(\eta_P + yv)|| \tag{5.25}$$

$$\leq ||\eta_P - \eta_Q||^2 H \tag{5.26}$$

Plugging in this bound on g'' gives the desired result.

$$\mathsf{R}_\lambda(\mathsf{Pr}(\theta|\eta_P)||\mathsf{Pr}(\theta|\eta_Q)) \leq \frac{(\lambda-1)^2}{\lambda-1} \max_{y\in[-\lambda+1,\lambda-1]} g''(y)/2 + \max_{y\in[-1,1]} g''(y)/2 \tag{5.27}$$

$$\leq (\lambda-1)||\eta_P-\eta_Q||^2 H/2 + ||\eta_P-\eta_Q||^2 H/2 \tag{5.28}$$

$$\leq ||\eta_P-\eta_Q||^2 H\lambda/2 \tag{5.29}$$

$$\leq ||\eta_P-\eta_Q||^2 H\lambda \tag{5.30}$$

□

We will also make use of the following standard results about the Hessian of the log-partition function of minimal exponential families, given in [Liese and Miescke, 2007] as Theorem 1.17 and Corollary 1.19 and rephrased for our purposes.

Theorem 18. *(Theorem 1.17 from [Liese and Miescke, 2007] The log-partition function $C(\eta)$ of a minimal exponential family is infinitely often differentiable at parameters η in the interior of the normalizable set E.*

Theorem 19. *(Corollary 1.19 from [Liese and Miescke, 2007] For minimal exponential family, the Hessian of the log-partition function $\nabla^2 C(\eta)$ is nonsingular for every parameter η in the interior of the normalizable set E.*

These results imply that the Hessian $\nabla^2 C(\eta)$ must exist and be continuous over η in the interior of E, as well as having non-zero determinant.

Theorem 12. *For any $\Delta-$bounded minimal exponential family with prior η_0 in the interior of E, any $\lambda > 1$, and any $\varepsilon > 0$, there exists $r^* \in (0,1]$ such that using $r \in (0,r^*]$ in Algorithm 2 will achieve (λ,ε)-RDP.*

PROOF:

Recall that Algorithm 2 uses the posterior parameter $\eta' = \eta_0 + r\sum_i^n (S(x),1)$ where the data contribution has been scaled by r. Our first step of this proof is to show that there exists

$r_0 \in (0,1]$ such that the order λ Rényi divergences of the generated parameters are finite for $r < r_0$.

Similar to the proof of Theorem 7, we will do so by creating a closed set where $C(\eta)$ is finite and that must contain η_P, η_Q, and η_L for any choice of neighboring data sets. On neighboring data sets, this generates $r-$neighboring parameters η_P and η_Q. The point $\eta_L = \lambda\eta_P + (1-\lambda)\eta_Q$ is therefore $r(\lambda-1)-$neighboring η_P. These points must be contained in the set $lpset(\eta_0, rn, r(\lambda-1)) = \eta_0 + rnU + r(\lambda-1)Diff$. For any point in this set, we can subtract off the component in rnU to get to a modified prior that is $r(\lambda-1)-$neighboring η_0.

By the assumption that η_0 is in the interior of E, there exists $\delta > 0$ such that the ball $\mathscr{B}(\eta_0, \delta) \subseteq E$. For the choice $r_0 = \frac{\delta}{2(\lambda-1)\Delta}$, for any $r \in (0, r_0)$, the modified prior we constructed for each point in $lpset(\eta_0, rn, r(\lambda-1))$ is within distance $r(\lambda-1)\Delta$ of η_0 and therefore within $\mathscr{B}(\eta_0, \delta/2) \subset \mathscr{B}(\eta_0, \delta) \subseteq E$. Observation 6 then allows us to conclude that every point η in $lpset(\eta_0, rn, r(\lambda-1))$ has an open neighborhood of radius δ_2 where $C(\eta)$ is finite. This is enough to conclude that the closure of this $lpset$ must also lie entirely within E, and $C(\eta)$ is finite and continuous over this closed set. As in Theorem 7, this suffices to show that the supremum of order λ Rényi divergences on neighboring data sets is bounded above.

We have thus shown there exists r_0 where the ε of our $(\lambda, \varepsilon)-$RDP guarantee is finite for $r < r_0$. However, our goal was to achieve a specific ε guarantee. Our proof of the existence of r^* centers around the claim that there must exist a bound H for the Hessian of $C(\eta)$ over all choices of $r \in [0, r_0)$.

We can construct the set $D = \cup_{r\in[0,r_0]} lpset(\eta_0, rn, r(\lambda-1))$, which will contain every possible η_P, η_Q, and η_L that might arise from any pair neighboring data sets and any choice of r in that interval. The previous argument still applies: each point in this union must have an open neighborhood of radius $\delta/2$ that is a subset of E. This is enough to conclude that closure of D is also a subset of E. Theorem 18 implies $\nabla^2 C(\eta)$ exists and is continuous on the interior of E, and this further implies that there must exist H such that for all η in this closure we have $||\nabla^2 C(\eta)|| \leq H$.

For any value r, we know that η_P and η_Q are $r-$neighboring, so we know $||\eta_P - \eta_Q|| \leq r\Delta$. Since D contains $lpset(\eta_0, rn, r(\lambda-1))$, the bound H must apply for all η in the set $\{\eta_P + x(\eta_P - \eta_Q) | x \in [-\lambda+1, \lambda-1]\}$. This allows us to use Lemma 17 to get the following expression:

$$\mathsf{R}_\lambda(\mathsf{Pr}(\theta|\eta_P)||\mathsf{Pr}(\theta|\eta_Q)) \leq ||\eta_P - \eta_Q||^2 H\lambda \tag{5.31}$$

$$\leq r\Delta^2 H\lambda. \tag{5.32}$$

If we set $r^* = \frac{\varepsilon}{\Delta^2 H\lambda}$, then for $r < r^*$ the order λ Rényi divergence of Algorithm 2 is bounded above by ε, which gives us the desired result.

□

The concentrated mechanism is a bit more subtle in how it reduces the influence of the data, and so we need this result modified from Lemmas 9 and 10 in Chapter 4. These results are presented here in a way that matches our notation. It effectively states that if we start at a prior η_0 satisfy mild but technical regularity assumptions, then the Hessians $C(k\eta_0)$ must converge to zero as k grows. In practical terms, this implies the covariance of our prior distribution must shrink as we increase the number of pseudo-observations.

Definition 21. Let $T^*_\eta = T(\text{argmax}_{\theta\in\Theta}\eta \cdot T(\theta))$. This represents the mode of the sufficient statistics under the distribution $\mathsf{Pr}(T(\theta)|\eta)$.

Lemma 20. *If $A(\theta)$ is continuously differentiable and η_0 is in the interior of E, then* $\text{argmax}_{\theta\in\Theta}\eta \cdot T(\theta)$ *must be in the interior of* Θ.

Lemma 21. *If we have a minimal exponential family in which $A(\theta)$ is differentiable of all orders, there exists $\delta_1 > 0$ such that the ball $\mathscr{B}(\eta_0, \delta_1)$ is a subset of E, there exists $\delta_2 > 0$ and a bound L such that all the seventh order partial derivatives of $A(\theta)$ on the set $D_{\eta_0,\delta_1,\delta_2} = \{\theta | \min_{\eta\in\mathscr{B}(\eta_0,\delta_1)} ||T(\theta) - T^*_\eta|| < \delta_2\}$ are bounded by P, and the determinant of $\nabla^2 A(\theta)$ is*

bounded away from zero on $D_{\eta_0,\delta_1,\delta_2}$, then there exists real number V,K such that for $k > K$ we have

$$\forall \eta \in \mathscr{B}(\eta_0, \delta_1) \ ||\nabla^2 C(k\eta)|| < \frac{V}{k}. \tag{5.33}$$

Theorem 13. *For any $\Delta-$bounded minimal exponential family with prior η_0 in the interior of E, for any $\lambda > 1$, and any $\varepsilon > 0$, there exists $m^* \in (0,1]$ such that using $m \in (0, m^*]$ in Algorithm 3 will achieve $(\lambda, \varepsilon)-$RDP.*

PROOF:

For a fixed value of m, recall that Algorithm 3 selects the posterior parameter $\eta' = m^{-1}\eta_0 + \sum_{i=1}^{n}(S(x_i), 1)$. For neighboring data sets $\mathbf{X}$ and $\mathbf{X}'$, the selected posterior parameters η_P, η_Q, and $\eta_L = \lambda\eta_P + (1-\lambda)\eta_Q$ lie within $lpset(m^{-1}\eta_0, n, \lambda - 1) = m^{-1}\eta_0 + nU + (\lambda - 1)Diff$.

We start by showing that the conditions of Lemma 21 are met. As we assumed η_0 is in the interior of E, there exists $\delta_1 > 0$ such that we have the ball $\mathscr{B}(\eta_0, \delta_1) \subseteq E$. By Theorem 18, the log-partition function of the data likelihood $A(\theta)$ is differentiable of all orders, and Theorem 19 tells us that the Hessian $\nabla^2 A(\theta)$ is non-singular with non-zero determinant on the interior of Θ. This permits the application of Lemma 20, offering a mapping from η in the interior of E to their mode T^*_η corresponding to a parameter θ in the interior of Θ. Knowing that $A(\theta)$ is infinitely differentiable on the interior of Θ further implies that the seventh order derivatives are well-behaved in a neighborhood around each mode resulting from this mapping. This provides the rest of the requirements for Lemma 21.

Therefore there exists V and K such that the following holds

$$\forall \eta \in \mathscr{B}(\eta_0, \delta_1) : ||\nabla^2 C(k\eta)|| \leq \frac{V}{k}. \tag{5.34}$$

We wish to show that $||\nabla^2 C(\eta)||$ must be bounded on the expanded set $lpset(m^{-1}\eta_0, n, \lambda - 1) = m^{-1}\eta_0 + nU + (\lambda - 1)Diff$, and will do so by showing that for small

enough m we can use equation (5.34) to bound the Hessians.

Let $\alpha(\eta)$ denote the last coordinate of η. This represents the pseudo-observation count of this parameter, and notice that $\forall u \in U : \alpha(u) = 1$ and $\forall v \in Diff : \alpha(v) = 0$. We are going to analyze the scaled set $c_m \cdot lpset(m^{-1}\eta_0, n, \lambda - 1)$ where c_m is a positive scaling constant that will depend on m.

$$c_m \cdot lpset(m^{-1}\eta_0, n, \lambda - 1) = c_m m^{-1}\eta_0 + c_m nU + c_m(\lambda - 1)Diff \tag{5.35}$$

.

For each η in this $c_m \cdot lpset$, we have

$$\alpha(\eta) = c_m m^{-1}\alpha(\eta_0) + c_m n \cdot 1 + c_m(\lambda - 1) \cdot 0 = c_m(m^{-1}\alpha(\eta_0) + n)\ . \tag{5.36}$$

Setting $c_m = \frac{\alpha(\eta_0)}{m^{-1}\alpha(\eta_0)+n}$ thus guarantees that for all η in $c_m \cdot lpset(m^{-1}\eta_0, n, \lambda - 1)$ we have $\alpha(\eta) = \alpha(\eta_0)$. We want to know how far the points in this $c_m \cdot lpset$ are from η_0, so we simply subtract η_0 to get a set D_m of vectors. These offset vectors have the form $c_m \cdot lpset(m^{-1}\eta_0, n, \lambda - 1) - \eta_0$ and therefore lie in the set

$$D_m = (c_m m^{-1} - 1)\eta_0 + c_m nU + c_m(\lambda - 1)Diff. \tag{5.37}$$

Using our expression of c_m as a function of m, we can see the following limiting behavior:

$$\lim_{m\to 0} c_m = \lim_{m\to 0} \frac{\alpha(\eta_0)}{m^{-1}\alpha(\eta_0) + n} = 0 \tag{5.38}$$

$$\lim_{m\to 0} c_m m^{-1} - 1 = \lim_{m\to 0} \frac{m^{-1}\alpha(\eta_0)}{m^{-1}\alpha(\eta_0) + n} - 1 = 1 - 1 = 0. \tag{5.39}$$

These limits lets us take the limit of the size of the vectors in D_m as $m \to 0$:

$$\lim_{m\to 0} \sup_{v\in D_m} ||v|| \leq \lim_{m\to 0} (c_m m^{-1} - 1)||\eta_0|| + c_m n \sup_{u_1\in U} ||u_1|| + c_m(\lambda - 1) \sup_{u_2\in Diff} ||u_2|| \tag{5.40}$$

$$\leq 0 \cdot ||\eta_0|| + 0 \cdot \sup_{u_1\in U} ||u_1|| + 0 \cdot \sup_{u_2\in Diff} ||u_2|| \tag{5.41}$$

$$\leq 0. \tag{5.42}$$

This limit supremum on D_m tells us that as $m \to 0$, the maximum distance between points in the scaled set $c_m \cdot lpset(m^{-1}\eta_0, n, \lambda - 1)$ and η_0 gets arbitrarily small. This means there exists some m_0 such that for $m < m_0$ the scaled set $c_m \cdot lpset(m^{-1}\eta_0, n, \lambda - 1)$ lies within $\mathscr{B}(\eta_0, \delta_1)$. This scaling mapping can be inverted, and it implies $lpset(m^{-1}\eta_0, n, \lambda - 1)$ is contained within $\frac{1}{c_m}\mathscr{B}(\eta_0, \delta_1)$. Being contained within this scaled ball is precisely what we need to use equation (5.34) with $\frac{1}{k} = c_m$.

Equation (5.34) bounds $||\nabla^2 C(\eta)|| \leq H_m = V c_m$ for all η in $lpset(m^{-1}\eta_0, n, \lambda - 1)$, which in turn lets us use Lemma 17 to bound our Rényi divergences.

$$\mathsf{R}_\lambda(\mathsf{Pr}(\theta|\eta_P)||\mathsf{Pr}(\theta|\eta_Q)) \leq ||\eta_P - \eta_Q||^2 H_m \lambda \tag{5.43}$$

$$\leq \Delta^2 V c_m \lambda. \tag{5.44}$$

As we have $c_m \to 0$ as $m \to 0$, we know there must exist m^* such that for $m < m^*$ we have $c_m \leq \frac{\varepsilon}{\Delta^2 V \lambda}$. This means the order λ Rényi divergences of Algorithm 3 on neighboring data sets is bounded above by ε, which gives us the desired result.

□

We have one last theorem to prove, the result claiming the Rényi divergences of order λ between η_P and its neighbors is convex, which greatly simplifies finding the supremum of these divergences over the convex sets being considered.

Theorem 14. *Let* $e(\eta_P, \eta_Q, \lambda) = \mathsf{R}_\lambda\left(\mathsf{Pr}(\theta|\eta_P)||\mathsf{Pr}(\theta|\eta_Q)\right)$.

For a fixed λ *and fixed* η_P*, the function e is a convex function over* η_Q.

If for any direction $v \in Diff$*, the function* $g_v(\eta) = v^\intercal \nabla^2 C(\eta) v$ *is convex over* η*, then for a fixed* λ*, the function*

$$f_\lambda(\eta_P) = \sup_{\eta_Q \; r-neighboring \; \eta_P} e(\eta_P, \eta_Q, \lambda) \tag{5.45}$$

is convex over η_P *in the directions spanned by* $Diff$.

PROOF:

First, we can show that for a fixed η_P and fixed λ, the choice of η_Q in the supremum must lie on the boundary of possible neighbors. This is derived from showing that $\mathsf{R}_\lambda(P||Q)$ is convex over the choice of η_Q.

Consider once again the expression for our Rényi divergence, expressed now as the function $e(\eta_P, \eta_Q, \lambda)$:

$$e(\eta_P, \eta_Q, \lambda) = \mathsf{R}_\lambda(P||Q) = \frac{C(\lambda\eta_P + (1-\lambda)\eta_Q) - \lambda C(\eta_P)}{\lambda - 1} + C(\eta_Q). \tag{5.46}$$

Let $\nabla_{\eta_Q} e(\eta_P, \eta_Q, \lambda)$ denote the gradient of the divergence with respect to η_Q.

$$\nabla_{\eta_Q} e(\eta_P, \eta_Q, \lambda) = \nabla C(\eta_Q) + \frac{1-\lambda}{\lambda - 1} \nabla C(\lambda\eta_P + (1-\lambda)\eta_Q) \tag{5.47}$$

$$= \nabla C(\eta_Q) - \nabla C(\lambda\eta_P + (1-\lambda)\eta_Q). \tag{5.48}$$

We can further find the Hessian with respect to η_Q:

$$\nabla^2_{\eta_Q} e(\eta_P, \eta_Q, \lambda) = \nabla^2 C(\eta_Q) - (1-\lambda)\nabla^2 C(\lambda\eta_P + (1-\lambda)\eta_Q). \tag{5.49}$$

By virtue of being a minimal exponential family, we know C is convex and thus $\nabla^2 C$ is PSD everywhere. Combined with the fact that $\lambda > 1$, this is enough to conclude that $\nabla^2_{\eta_Q} e(\eta_P, \eta_Q, \lambda)$ is also PSD for everywhere with $\lambda > 1$. This means $e(\eta_P, \eta_Q, \lambda)$ is a convex function with respect to η_Q for any fixed η_P and λ.

We now wish to characterize the function $f_\lambda(\eta_P)$, which takes a supremum over $\eta_Q \in \eta_P + rDiff$ of $e(\eta_P, \eta_Q, \lambda)$.

$$f_\lambda(\eta_P) = \sup_{\eta_Q r-\text{neighboring } \eta_P} e(\eta_P, \eta_Q, \lambda) \tag{5.50}$$

We re-parameterize this supremum in terms of the offset $b = \eta_Q - \eta_P$.

$$f_\lambda(\eta_P) = \sup_{b \in rDiff} e(\eta_P, \eta_P + b, \lambda) \tag{5.51}$$

Now for any fixed offset b, x we can find the expression for the Hessian of $\nabla^2_{\eta_P} e(\eta_P, \eta_P + b, \lambda)$.

$$\nabla^2_{\eta_P} e(\eta_P, \eta_P + b, \lambda) = \nabla^2 C(\eta_P + b) - \frac{\lambda}{\lambda - 1}\nabla^2 C(\eta_P) + \frac{1}{\lambda - 1}\nabla^2 C(\eta_p + (1-\lambda)b) \tag{5.52}$$

We wish to show this Hessian is PSD, i.e. for any vector v we have $v^\intercal \nabla^2_{\eta_P} e(\eta_P, \eta_P + b, \lambda) v$ is non-negative. We can rewrite this in terms of the function $g_v(\eta)$ introduced in the theorem statement.

$$v^\intercal \nabla^2_{\eta_P} e(\eta_P, \eta_P + b, \lambda) v = g_v(\eta_P + b) - \frac{\lambda}{\lambda - 1} g_v(\eta_P) + \frac{1}{\lambda - 1} g_v(\eta_p + (1-\lambda)b) \tag{5.53}$$

$$= \frac{\lambda}{\lambda - 1} \left(\frac{\lambda - 1}{\lambda} g_v(\eta_P + b) - g_v(\eta_P) + \frac{1}{\lambda} g_v(\eta_p + (1-\lambda)b) \right) \tag{5.54}$$

We know $\frac{\lambda}{\lambda - 1} > 0$ and that η_P must lie between $\eta_P + b$ and $\eta_P - (\lambda - 1)b$. Our assumption that $g_v(\eta)$ is convex over η for all directions v then lets us use Jensen's inequality to see that the expression (5.54) must be non-negative.

This lets us conclude that $v^\intercal(\nabla^2_{\eta_P} e(\eta_P, \eta_P + b, \lambda) v \geq 0$ for all v, and thus this Hessian is PSD for any η_P. This in turn means our divergence $e(\eta_P, \eta_P + b, \lambda)$ is convex over η_P assuming a fixed offset b.

We return to $f_\lambda(\eta_P)$, and observe that it is a supremum of functions that are convex, and therefore it is convex as well.

□

5.6 Additional Beta-Bernoulli Experiments

The utility of the prior-based methods (Algorithms 2 and 3) depends on how well the prior matches the observed data. Figure 5.7 shows several additional situations for the experimental procedure of measuring the log-likelihood of the data.

In each case, the prior $\eta_0 = (6, 18)$ was used, and both $\mathbf{X}$ and $\mathbf{X}_H$ had 100 data points. $\lambda = 15$ was fixed in these additional experiments. The only thing that varies is the true population parameter ρ. In (a), $\rho = 1/3$ closely matches the predictions of the prior η_0. In (b), $\rho = 0.5$, presented as an intermediate case where the prior is misleading. Finally, in (c), $\rho = 2/3$, which is biased in the opposite direction as the prior. In all cases, the proposed methods act conservatively in the face of high privacy, but in (a) this worst case limiting behavior still has high utility. Having a strong informative prior helps these mechanisms. The setting in which the prior is based off of

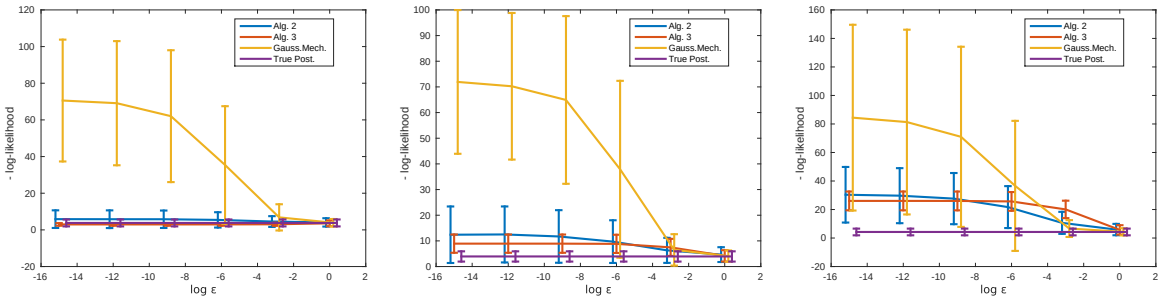

Figure 5.7. Utility Comparison for a fixed η_0 but varying true population parameter. **Left:** $\rho = 1/3$ (high match with η_0). **Middle:** $\rho = 1/2$. **Right:** $\rho = 2/3$ (low match with η_0).

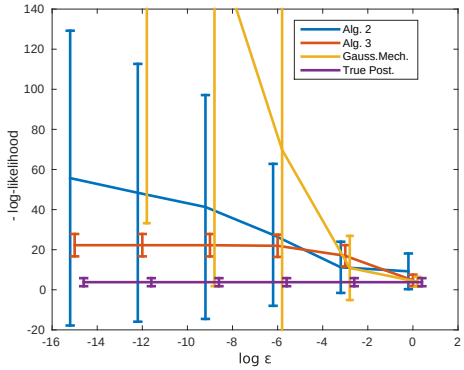

Figure 5.8. Utility Experiment for the non-informative uniform prior

a representative sample of non-private data from the same population as the private data is likely to be beneficial for Algorithms 2 and 3.

One other case is presented in Figure 5.8, where $\rho = 0.2$ but the prior has been changed $\eta_0 = (1,2)$. λ is still 15, and the number of data points is still 100. This prior corresponds to the uniform prior, as it assigns equal probability to all estimated data means on $(0,1)$. It represents an attractive case on a non-informative prior, but also represents a situation in which privacy is difficult. In particular, $\lambda^* = 2$ in this setting. When Algorithm 3 scales up this prior, it becomes concentrated around $\rho = 0.2$, so this setting also corresponds to a case where the true population parameter does not match well with the predictions from the prior.

5.7 Conclusion

The inherent randomness of posterior sampling and the mitigating influence of a prior can be made to offer a wide range of privacy guarantees. Our proposed methods outperform existing methods in specific situations. The privacy analyses of the mechanisms fit nicely into the recently introduced RDP framework, which continues to present itself as a relaxation of DP worthy of further investigation.

5.8 Acknowledgements

This work was partially supported by NSF under IIS 1253942, ONR under N00014-16-1-2616, and a Google Faculty Research Award.

This chapter is based on the material in Advances in Neural Information Processing Systems 2017 (Joseph Geumlek, Shuang Song, and Kamalika Chaudhuri. ”Rényi differential privacy mechanisms for posterior sampling”). The dissertation author was the primary investigator and author of this material.

Chapter 6

Privacy Amplification of Diffusion

6.1 Introduction

Differential privacy (DP) [Dwork et al., 2006b] has arisen in the last decade into a strong de-facto standard for privacy-preserving computation in the context of statistical analysis. The success of DP is based, at least in part, on the availability of robust building blocks (e.g., the Laplace, exponential and Gaussian mechanisms) together with relatively simple rules for analyzing complex mechanisms built out of these blocks (e.g., composition and robustness to post-processing). The inherent tension between privacy and utility in practical applications has sparked a renewed interest into the development of further rules leading to tighter privacy bounds. A trend in this direction is to find ways to measure the privacy introduced by sources of randomness that are not accounted for by standard composition rules. Generally speaking, these are referred to as *privacy amplification* rules, with prominent examples being amplification by *subsampling* [Chaudhuri and Mishra, 2006, Kasiviswanathan et al., 2011, Li et al., 2012, Beimel et al., 2013, 2014, Bun et al., 2015, Balle et al., 2018, Wang et al., 2019], *shuffling* [Erlingsson et al., 2019, Cheu et al., 2019, Balle et al., 2019] and *iteration* [Feldman et al., 2018].

Motivated by these considerations, in this paper we initiate a systematic study of privacy amplification by *stochastic post-processing*. Specifically, given a DP mechanism M producing (probabilistic) outputs in $\mathbb{X}$ and a Markov operator K defining a stochastic transition between $\mathbb{X}$ and $\mathbb{Y}$, we are interested in measuring the privacy of the post-processed mechanism $K \circ M$

producing outputs in $\mathbb{Y}$. The standard post-processing property of DP states that $K \circ M$ is at least as private as M. Our goal is to understand under what conditions the post-processed mechanism $K \circ M$ is *strictly* more private than M. Roughly speaking, this amplification should be non-trivial when the operator K "forgets" information about the distribution of its input $M(D)$. Our main insight is that, at least when $\mathbb{Y} = \mathbb{X}$, the forgetfulness of K from the point of view of DP can be measured using similar tools to the ones developed to analyze the speed of convergence, i.e. *mixing*, of the Markov process associated with K.

In this setting, we provide three types of results, each associated with a standard method used in the study of convergence for Markov processes. In the first place, Section 6.3 provides DP amplification results for the case where the operator K satisfies a uniform mixing condition. These include standard conditions used in the analysis of Markov chains on discrete spaces, including the well-known Dobrushin coefficent and Doeblin's minorization condition [Levin and Peres, 2017]. Although in principle uniform mixing conditions can also be defined in more general non-discrete spaces [Del Moral et al., 2003], most Markov operators of interest in $\mathbb{R}^d$ do not exhibit uniform mixing since the speed of convergence depends on how far apart the initial inputs are. Convergence analyses in this case rely on more sophisticated tools, including Lyapunov functions [Meyn and Tweedie, 2012], coupling methods [Lindvall, 2002] and functional inequalities [Bakry et al., 2013].

Following these ideas, Section 6.4 investigates the use of coupling methods to quantify privacy amplification by post-processing under Rényi DP [Mironov, 2017]. These methods apply to operators given by, e.g., Gaussian and Laplace distributions, for which uniform mixing does not hold. Results in this section are intimately related to the privacy amplification by iteration phenomenon studied in [Feldman et al., 2018] and can be interpreted as extensions of their main results to more general settings. In particular, our analysis unpacks the *shifted* Rényi divergence used in the proofs from [Feldman et al., 2018] and allows us to easily track the effect of iterating arbitrary noisy Lipschitz maps. As a consequence, we show an exponential improvement on the privacy amplification by iteration of Noisy SGD in the strongly convex case which follows from

applying this generalized analysis to *strict* contractions.

6.2 Setup

We start by introducing notation and concepts that will be used throughout the paper. We write $[n] = \{1, \dots, n\}$, $a \wedge b = \min\{a, b\}$ and $[a]_+ = \max\{a, 0\}$.

Probability.

Let $\mathbb{X} = (\mathbb{X}, \Sigma, \lambda)$ be a measurable space with sigma-algebra Σ and base measure λ. We write $\mathscr{P}(\mathbb{X})$ to denote the set of probability distributions on $\mathbb{X}$. Given a probability distribution $\mu \in \mathscr{P}(\mathbb{X})$ and a measurable event $E \subseteq \mathbb{X}$ we write $\mu(E) = \mathsf{Pr}[X \in E]$ for a random variable $X \sim \mu$, denote its expectation under $f : \mathbb{X} \to \mathbb{R}^d$ by $\mathsf{E}[f(X)]$, and can get back its distribution as $\mu = \mathsf{Law}(X)$. Given two distributions μ, ν (or, in general, arbitrary measures) we write $\mu \ll \nu$ to denote that μ is absolutely continuous with respect to ν, in which case there exists a Radon-Nikodym derivative $\frac{d\mu}{d\nu}$. We shall reserve the notation $p_\mu = \frac{d\mu}{d\lambda}$ to denote the density of μ with respect to the base measure. We also write $\mathscr{C}(\mu, \nu)$ to denote the set of couplings between μ and ν; i.e. $\pi \in \mathscr{C}(\mu, \nu)$ is a distribution on $\mathscr{P}(\mathbb{X} \times \mathbb{X})$ with marginals μ and ν. The support of a distribution is $\mathrm{supp}(\mu)$.

Markov Operators.

We will use $\mathscr{K}(\mathbb{X}, \mathbb{Y})$ to denote the set of Markov operators $K : \mathbb{X} \to \mathscr{P}(\mathbb{Y})$ defining a stochastic transition map between $\mathbb{X}$ and $\mathbb{Y}$ and satisfying that $x \mapsto K(x)(E)$ is measurable for every measurable $E \subseteq \mathbb{Y}$. Markov operators act on distributions $\mu \in \mathscr{P}(\mathbb{X})$ on the left through $(\mu K)(E) = \int K(x)(E)\mu(dx)$, and on functions $f : \mathbb{Y} \to \mathbb{R}$ on the right through $(Kf)(x) = \int f(y)K(x, dy)$, which can also be written as $(Kf)(x) = \mathsf{E}[f(X)]$ with $X \sim K(x)$. The kernel of a Markov operator K (with respect to λ) is the function $k(x, \cdot) = \frac{dK(x)}{d\lambda}$ associating with x the density of $K(x)$ with respect to a fixed measure.

Divergences.

A popular way to measure dissimilarity between distributions is to use Csiszár divergences $\mathsf{D}_\phi(\mu \| \nu) = \int \phi(\frac{d\mu}{d\nu}) d\nu$, where $\phi : \mathbb{R}_+ \to \mathbb{R}$ is convex with $\phi(1) = 0$. Taking $\phi(u) = \frac{1}{2}|u - 1|$

yields the total variation distance $\mathsf{TV}(\mu,\nu)$, and the choice $\phi(u) = [u - e^{\varepsilon}]_+$ with $\varepsilon \geq 0$ gives the hockey-stick divergence $\mathsf{D}_{e^{\varepsilon}}$, which satisfies

$$\mathsf{D}_{e^{\varepsilon}}(\mu\|\nu) = \int \left[\frac{d\mu}{d\nu} - e^{\varepsilon}\right]_+ d\nu = \int [p_\mu - e^{\varepsilon} p_\nu]_+ d\lambda = \sup_{E\subseteq\mathbb{X}} (\mu(E) - e^{\varepsilon}\nu(E)) \ .$$

It is easy to check that $\varepsilon \mapsto \mathsf{D}_{e^{\varepsilon}}(\mu\|\nu)$ is monotonically decreasing and $\mathsf{D}_1 = \mathsf{TV}$. All Csiszár divergences satisfy joint convexity $\mathsf{D}((1-\gamma)\mu_1 + \gamma\mu_2\|(1-\gamma)\nu_1 + \gamma\nu_2) \leq (1-\gamma)\mathsf{D}(\mu_1\|\nu_1) + \gamma\mathsf{D}(\mu_2\|\nu_2)$ and the data processing inequality $\mathsf{D}(\mu K\|\nu K) \leq \mathsf{D}(\mu\|\nu)$ for any Markov operator K. Rényi divergences, which do not belong to the family of Csiszár divergences, are another way to compare distributions. For $\alpha > 1$ the Rényi divergence of order α is defined as $\mathsf{R}_\alpha(\mu\|\nu) = \frac{1}{\alpha-1}\log\int(\frac{d\mu}{d\nu})^\alpha d\nu$, and also satisfies the data processing inequality. We note that this notation for the Rényi divergence differs from that seen in Chapter 3, but this notation is chosen to help differentiate it from the other general divergences discussed in this chapter. Finally, to measure similarity between $\mu,\nu \in \mathscr{P}(\mathbb{R}^d)$ we sometimes use the ∞-Wasserstein distance:

$$\mathsf{W}_\infty(\mu,\nu) = \inf_{\pi\in\mathscr{C}(\mu,\nu)} \inf\{w \geq 0 : \|X - Y\| \leq w \text{ holds almost surely for } (X,Y) \sim \pi\} \ .$$

Differential Privacy.

A mechanism $M : \mathbb{D}^n \to \mathscr{P}(\mathbb{X})$ is a randomized function that takes a dataset $D \in \mathbb{D}^n$ over some universe of records $\mathbb{D}$ and returns a (sample from) distribution $M(D)$. We write $D \simeq D'$ to denote two databases differing in a single record. We say that M satisfies[1] (ε,δ)-DP if $\sup_{D\simeq D'} \mathsf{D}_{e^{\varepsilon}}(M(D)\|M(D')) \leq \delta$ [Dwork et al., 2006b]. Furthermore, we say that M satisfies (α,ε)-RDP if $\sup_{D\simeq D'} \mathsf{R}_\alpha(M(D)\|M(D')) \leq \varepsilon$ [Mironov, 2017].

[1] This divergence characterization of DP is due to [Barthe and Olmedo, 2013].

6.3 Amplification From Uniform Mixing

We start our analysis of privacy amplification by stochastic post-processing by considering settings where the Markov operator K satisfies one of the following uniform mixing conditions.

Definition 22. Let $K \in \mathscr{K}(\mathbb{X}, \mathbb{Y})$ be a Markov operator, $\gamma \in [0,1]$ and $\varepsilon \geq 0$. We say that K is:

(1) γ-*Dobrushin* if $\sup_{x,x'} \mathsf{TV}(K(x), K(x')) \leq \gamma$,

(2) (γ, ε)-*Dobrushin* if $\sup_{x,x'} \mathsf{D}_{e^\varepsilon}(K(x) \| K(x')) \leq \gamma$,

(3) γ-*Doeblin* if there exists a distribution $\omega \in \mathscr{P}(\mathbb{Y})$ such that $K(x) \geq (1-\gamma)\omega$ for all $x \in \mathbb{X}$,

(4) γ-*ultra-mixing* if for all $x, x' \in \mathbb{X}$ we have $K(x) \ll K(x')$ and $\frac{dK(x)}{dK(x')} \geq 1-\gamma$.

Most of these conditions arise in the context of mixing analyses in Markov chains. In particular, the Dobrushin condition can be tracked back to [Dobrushin, 1956], while Doeblin's condition was introduced earlier [Doeblin, 1937] (see also [Nummelin, 2004]). Ultra-mixing is a strengthening of Doeblin's condition used in [Del Moral et al., 2003]. The (γ, ε)-Dobrushin is, on the other hand, new and is designed to be a generalization of Dobrushin tailored for amplification under the hockey-stick divergence.

It is not hard to see that Dobrushin's is the weakest among these conditions, and in fact we have the implications summarized in Figure 6.1 (see Lemma 1). This explains why the amplification bounds in the following result are increasingly stronger, and in particular why the first two only provide amplification in δ, while the last two also amplify the ε parameter.

Lemma 1. *The implications in Figure 6.1 hold.*

Proof. That (γ, ε)-Dobrushin implies γ-Dobrushin follows directly from $\mathsf{D}_{e^\varepsilon}(K(x) \| K(x')) \leq \mathsf{TV}(K(x), K(x'))$.

To see that γ-Doeblin implies γ-Dobrushin we observe that the kernel of a γ-Doeblin operator must satisfy $\inf_x k(x,y) \geq (1-\gamma) p_\omega(y)$ for any y. Thus, we can use the characterization

of TV in terms of a minimum to get

$$\mathsf{TV}(K(x),K(x')) = 1 - \int (k(x,y) \wedge k(x',y))\lambda(dy) \leq 1 - (1-\gamma)\int p_\omega(y)\lambda(dy) = \gamma \ .$$

Finally, to get the γ-Doeblin condition for an operator K satisfying γ-ultra-mixing we recall from [Del Moral et al., 2003, Lemma 4.1] that for such an operator we have that $K(x) \geq (1-\gamma)\tilde{\omega}K$ is satisfied for any probability distribution $\tilde{\omega}$ and $x \in \mathrm{supp}(\tilde{\omega})$. Thus, taking $\tilde{\omega}$ to have full support we obtain Doeblin's condition with $\omega = \tilde{\omega}K$. □

Theorem 15. *Let M be an (ε,δ)-DP mechanism. For a given Markov operator K, the post-processed mechanism $K \circ M$ satisfies:*

(1) (ε,δ')-DP with $\delta' = \gamma\delta$ if K is γ-Dobrushin,

(2) (ε,δ')-DP with $\delta' = \gamma\delta$ if K is $(\gamma,\tilde{\varepsilon})$-Dobrushin with[2] $\tilde{\varepsilon} = \log(1+\frac{e^\varepsilon-1}{\delta})$,

(3) (ε',δ')-DP with $\varepsilon' = \log(1+\gamma(e^\varepsilon-1))$ and $\delta' = \gamma(1-e^{\varepsilon'-\varepsilon}(1-\delta))$ if K is γ-Doeblin,

(4) (ε',δ')-DP with $\varepsilon' = \log(1+\gamma(e^\varepsilon-1))$ and $\delta' = \gamma\delta e^{\varepsilon'-\varepsilon}$ if K is γ-ultra-mixing.

A few remarks about this result are in order. First we note that (2) is stronger than (1) since the monotonicity of hockey-stick divergences implies $\mathsf{TV} = \mathsf{D}_1 \geq \mathsf{D}_{e^{\tilde{\varepsilon}}}$. Also note how in the results above we always have $\varepsilon' \leq \varepsilon$, and in fact the form of ε' is the same as obtained under amplification by subsampling when, e.g., a γ-fraction of the original dataset is kept. This is not a coincidence since the proofs of (3) and (4) leverage the *overlapping mixtures* technique used to analyze amplification by subsampling in [Balle et al., 2018]. However, we note that for (3) we can have $\delta' > 0$ even with $\delta = 0$. In fact the Doeblin condition only leads to an amplification in δ if $\gamma \leq \frac{\delta e^\varepsilon}{(1-\delta)(e^\varepsilon-1)}$.

For convenience, we split the proof of Theorem 15 into four separate statements, each corresponding to one of the claims in the theorem.

[2] We take the convention $\tilde{\varepsilon} = \infty$ whenever $\delta = 0$, in which case the (γ,∞)-Dobrushin condition is obtained with respect to the divergence $\mathsf{D}_\infty(\mu\|\nu) = \mu(\mathrm{supp}(\mu) \setminus \mathrm{supp}(\nu))$.

Recall that a Markov operator $K \in \mathscr{K}(\mathbb{X}, \mathbb{Y})$ is *γ-Dobrushin* if $\sup_{x,x'} \mathsf{TV}(K(x), K(x')) \leq \gamma$.

Theorem 16. *Let M be an (ε, δ)-DP mechanism. If K is a γ-Dobrushin Markov operator, then the composition $K \circ M$ is $(\varepsilon, \gamma\delta)$-DP.*

Proof. This follows directly from the *strong Markov contraction lemma* established by Cohen et al. [1993] in the discrete case and by Del Moral et al. [2003] in the general case (see also [Raginsky, 2016]). In particular, this lemma states that for any divergence D in the sense of Csiszár we have $\mathsf{D}(\mu K \| \nu K) \leq \gamma \mathsf{D}(\mu \| \nu)$. Letting $\mu = M(D)$ and $\nu = M(D')$ for some $D \simeq D'$ and applying this inequality to $\mathsf{D}_{e^\varepsilon}(\mu K \| \nu K)$ yields the result. □

Next we prove amplification when K is a (γ, ε)-Dobrushin operator. Recall that a Markov operator $K \in \mathscr{K}(\mathbb{X}, \mathbb{Y})$ is *(γ, ε)-Dobrushin* if $\sup_{x,x'} \mathsf{D}_{e^\varepsilon}(K(x) \| K(x')) \leq \gamma$. We will require the following technical lemmas in the proof of Theorem 17.

Lemma 2. *Let $\mu \perp \nu$ denote the fact* $\mathrm{supp}(\mu) \cap \mathrm{supp}(\nu) = \emptyset$. *If K is (γ, ε)-Dobrushin, then we have*

$$\sup_{\mu \perp \nu} \mathsf{D}_{e^\varepsilon}(\mu K \| \nu K) \leq \gamma \ .$$

Proof. Note that the condition on γ can be written as $\sup_{x,x'} \mathsf{D}_{e^\varepsilon}(\delta_x K \| \delta_{x'} K) \leq \gamma$. This shows that by hypothesis the condition already holds for the distributions $\delta_x \perp \delta_{x'}$ with $x \neq x'$. Thus, all we need to do is prove that these distributions are extremal for $\mathsf{D}_{e^\varepsilon}(\mu K \| \nu K)$ among all distributions with $\mu \perp \nu$. Let $\mu \perp \nu$ and define $U = \mathrm{supp}(\mu)$ and $V = \mathrm{supp}(\nu)$. Working in the discrete setting for simplicity, we can write $\mu = \sum_{x \in U} \mu(x) \delta_x$, with an equivalent expression for ν. Now we use

the joint convexity of $\mathsf{D}_{e^\varepsilon}$ to write

$$\begin{aligned}\mathsf{D}_{e^\varepsilon}(\mu K\|\nu K) &\le \sum_{x\in U}\mu(x)\mathsf{D}_{e^\varepsilon}(\delta_x K\|\nu K) \le \sum_{x\in U}\sum_{x'\in V}\mu(x)\nu(x')\mathsf{D}_{e^\varepsilon}(\delta_x K\|\delta_{x'}K)\\ &\le \sup_{x\neq x'} D(\delta_x K\|\delta_x' K) \le \gamma \ .\end{aligned}$$

□

Lemma 3. *Let $a\wedge b \triangleq \min\{a,b\}$. Then we have*

$$\mathsf{D}_{e^\varepsilon}(\mu\|\nu) = 1 - \int \left(p_\mu(x)\wedge e^\varepsilon p_\nu(x)\right)\lambda(dx) \ .$$

Proof. Define $A=\{x : p_\mu(x)\le e^\varepsilon p_\nu(x)\}$ to be set of points where μ is dominated by $e^\varepsilon\nu$, and let A^c denote its complementary. Then we have the identities

$$\begin{aligned}\int (p_\mu\wedge e^\varepsilon p_\nu)d\lambda &= \int_A d\mu + e^\varepsilon\int_{A^c} d\nu \ ,\\ \int [p_\mu - e^\varepsilon p_\nu]_+ d\lambda &= \int_{A^c} d\mu - e^\varepsilon\int_{A^c} d\nu \ .\end{aligned}$$

Thus we obtain the desired result since

$$\begin{aligned}\mathsf{D}_{e^\varepsilon}(\mu\|\nu) + \int (p_\mu\wedge e^\varepsilon p_\nu)d\lambda &= \int [p_\mu - e^\varepsilon p_\nu]_+ d\lambda + \int (p_\mu\wedge e^\varepsilon p_\nu)d\lambda\\ &= \int_{A^c} d\mu + \int_A d\mu = 1 \ .\end{aligned}$$

□

Theorem 17. *Let M be an (ε,δ)-DP mechanism and let $\varepsilon' = \log\left(1+\frac{e^\varepsilon-1}{\delta}\right)$. If K is a (γ,ε')-Dobrushin Markov operator, then the composition $K\circ M$ is $(\varepsilon,\gamma\delta)$-DP.*

Proof. Fix $\mu = M(D)$ and $\nu = M(D')$ for some $D\simeq D'$ and let $\theta = \mathsf{D}_{e^\varepsilon}(\mu\|\nu)\le\delta$. We start

by constructing overlapping mixture decompositions for μ and ν as follows. First, define the function $f = p_\mu \wedge e^\varepsilon p_\nu$ and let ω be the probability distribution with density $p_\omega = \frac{f}{\int f d\lambda} = \frac{f}{1-\theta}$, where we used Lemma 3. Now note that by construction we have the inequalities

$$\begin{aligned} p_\mu - (1-\theta)p_\omega &= p_\mu - p_\mu \wedge e^\varepsilon p_\nu \geq 0 \ , \\ p_\nu - \frac{1-\theta}{e^\varepsilon} p_\omega &= p_\nu - p_\nu \wedge e^{-\varepsilon} p_\mu \geq 0 \ . \end{aligned}$$

Assuming without loss of generality that $\mu \neq \nu$, these inequalities imply that we can construct probability distributions μ' and ν' such that

$$\begin{aligned} \mu &= (1-\theta)\omega + \theta\mu' \ , \\ \nu &= \frac{1-\theta}{e^\varepsilon}\omega + \left(1 - \frac{1-\theta}{e^\varepsilon}\right)\nu' \ . \end{aligned}$$

Now we observe that the distributions μ' and ν' defined in this way have disjoint support. To see this we first use the identity $p_\mu = (1-\theta)p_\omega + \theta p_{\mu'}$ to see that

$$p_{\mu'}(x) > 0 \equiv p_\mu(x) - (1-\theta)p_\omega(x) > 0 \equiv p_\mu(x) - p_\mu(x) \wedge e^\varepsilon p_\nu(x) > 0 \equiv p_\mu(x) > e^\varepsilon p_\nu(x) \ .$$

Thus we have $\mathrm{supp}(\mu') = \{x : p_\mu(x) > e^\varepsilon p_\nu(x)\}$. A similar argument applied to p_ν shows that on the other hand $\mathrm{supp}(\nu') = \{x : p_\mu(x) < e^\varepsilon p_\nu(x)\}$, and thus $\mu' \perp \nu'$.

Finally, we proceed to use the mixture decomposition of μ and ν and the condition $\mu' \perp \nu'$ to bound $\mathsf{D}_{e^\varepsilon}(\mu K \| \nu K)$ as follows. By using the mixture decompositions we get

$$\mu - e^\varepsilon \nu = \theta\mu' - e^\varepsilon \left(1 - \frac{1-\theta}{e^\varepsilon}\right)\nu' = \theta(\mu' - e^{\tilde{\varepsilon}}\nu') \ ,$$

where $\tilde{\varepsilon} = \log\left(1 + \frac{e^\varepsilon - 1}{\theta}\right) \geq \varepsilon'$. Thus, applying the definition of $\mathsf{D}_{e^\varepsilon}$, using the linearity of

Markov operators, and the monotonicity $\mathsf{D}_{e^{\tilde{\varepsilon}}} \leq \mathsf{D}_{e^{\varepsilon'}}$ we obtain the bound:

$$\mathsf{D}_{e^{\varepsilon}}(\mu K \| \nu K) = \theta \mathsf{D}_{e^{\tilde{\varepsilon}}}(\mu' K \| \nu' K) \leq \theta \mathsf{D}_{e^{\varepsilon'}}(\mu' K \| \nu' K) \leq \gamma\theta = \gamma \mathsf{D}_{e^{\varepsilon'}}(\mu \| \nu) \ ,$$

where the last inequality follows from Lemma 2. □

Recall that a Markov operator $K \in \mathscr{K}(\mathbb{X}, \mathbb{Y})$ is *γ-Doeblin* if there exists a distribution $\omega \in \mathscr{P}(\mathbb{Y})$ such that $K(x) \geq (1-\gamma)\omega$ for all $x \in \mathbb{X}$. The proof of amplification for γ-Doeblin operators further leverages overlapping mixture decompositions like the one used in Theorem 17, but this time the mixture arises at the level of the kernel itself.

Theorem 18. *Let M be an (ε, δ)-DP mechanism. If K is a γ-Doeblin Markov operator, then the composition $K \circ M$ is (ε', δ')-DP with $\varepsilon' = \log(1 + \gamma(e^{\varepsilon} - 1))$ and $\delta' = \gamma\Big(1 - e^{\varepsilon' - \varepsilon}(1-\delta)\Big)$.*

Proof. Fix $\mu = M(D)$ and $\nu = M(D')$ for some $D \simeq D'$. Let ω be a witness that K is γ-Doeblin and let K_ω be the constant Markov operator given by $K_\omega(x) = \omega$ for all x. Doeblin's condition $K(x) \geq (1-\gamma)\omega = (1-\gamma)K_\omega(x)$ implies that the following is again a Markov operator:

$$\tilde{K} = \frac{K - (1-\gamma)K_\omega}{\gamma} \ .$$

Thus, we can write K as the mixture $K = (1-\gamma)K_\omega + \gamma\tilde{K}$ and then use the *advanced joint convexity* property of $\mathsf{D}_{e^{\varepsilon'}}$ [Balle et al., 2018, Theorem 2] with $\varepsilon' = \log(1 + \gamma(e^{\varepsilon} - 1))$ to obtain the following:

$$\begin{aligned}
\mathsf{D}_{e^{\varepsilon'}}(\mu K \| \nu K) &= \mathsf{D}_{e^{\varepsilon'}}((1-\gamma)\omega + \gamma\mu\tilde{K} \| (1-\gamma)\omega + \gamma\nu\tilde{K}) \\
&= \gamma \mathsf{D}_{e^{\varepsilon}}(\mu\tilde{K} \| (1-\beta)\omega + \beta\nu\tilde{K}) \\
&\leq \gamma\big((1-\beta)\mathsf{D}_{e^{\varepsilon}}(\mu\tilde{K} \| \omega) + \beta \mathsf{D}_{e^{\varepsilon}}(\mu\tilde{K} \| \nu\tilde{K})\big) \ ,
\end{aligned}$$

where $\beta = e^{\varepsilon' - \varepsilon}$. Finally, using the immediate bounds $\mathsf{D}_{e^{\varepsilon}}(\mu\tilde{K} \| \nu\tilde{K}) \leq \mathsf{D}_{e^{\varepsilon}}(\mu \| \nu)$ and

$\mathsf{D}_{e^\varepsilon}(\mu\tilde{K}\|\omega) \leq 1$, we get

$$\mathsf{D}_{e^{\varepsilon'}}(\mu K\|\nu K) \leq \gamma(1-e^{\varepsilon'-\varepsilon}+e^{\varepsilon'-\varepsilon}\delta) \ .$$

□

Our last amplification result applies to operators satisfying the ultra-mixing condition of Del Moral et al. [2003]. We say that a Markov operator $K \in \mathscr{K}(\mathbb{X},\mathbb{Y})$ is *γ-ultra-mixing* if for all $x,x' \in \mathbb{X}$ we have $K(x) \ll K(x')$ and $\frac{dK(x)}{dK(x')} \geq 1-\gamma$. The proof strategy is based on the ideas from the previous proof, although in this case the argument is slightly more technical as it involves a strengthening of the Doeblin condition implied by ultra-mixing that only holds under a specific support.

Theorem 19. *Let M be an (ε,δ)-DP mechanism. If K is a γ-ultra-mixing Markov operator, then the composition $K \circ M$ is (ε',δ')-DP with $\varepsilon' = \log(1+\gamma(e^\varepsilon-1))$ and $\delta' = \gamma\delta e^{\varepsilon'-\varepsilon}$.*

Proof. Fix $\mu = M(D)$ and $\nu = M(D')$ for some $D \simeq D'$. The proof follows a similar strategy as the one used in Theorem 18, but coupled with the following consequence of the ultra-mixing property: for any probability distribution ω and $x \in \mathrm{supp}(\omega)$ we have $K(x) \geq (1-\gamma)\omega K$ [Del Moral et al., 2003, Lemma 4.1]. We use this property to construct a collection of mixture decompositions for K as follows. Let $\alpha \in (0,1)$ and take $\tilde{\omega} = (1-\alpha)\mu + \alpha\nu$ and $\omega = \tilde{\omega}K$. By the ultra-mixing condition and the argument used in the proof of Theorem 18, we can show that

$$\tilde{K} = \frac{K-(1-\gamma)K_\omega}{\gamma}$$

is a Markov operator from $\mathrm{supp}(\mu)\cup\mathrm{supp}(\nu)$ into $\mathbb{X}$. Here K_ω is the constant Markov operator $K_\omega(x) = \omega$. Furthermore, the expression for $\tilde{K}$ and the definition of ω imply that

$$\tilde{\omega}\tilde{K} = \frac{\tilde{\omega}K-(1-\gamma)\tilde{\omega}K_\omega}{\gamma} = \omega \ . \tag{6.1}$$

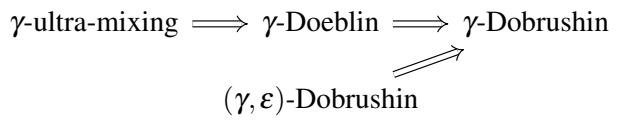

Figure 6.1. Implications between mixing conditions

Mixing Condition	Local DP Condition
γ-Dobrushin	$(0,\gamma)$-LDP
(γ,ε)-Dobrushin	(ε,γ)-LDP
γ-Doeblin	Blanket condition[3]
γ-ultra-mixing	$(\log\frac{1}{1-\gamma},0)$-LDP

Figure 6.2. Relation between mixing conditions and local DP

Now note that the mixture decompositions $\mu K = (1-\gamma)\omega + \gamma\mu\tilde{K}$ and $\nu K = (1-\gamma)\omega + \gamma\nu\tilde{K}$ and the *advanced joint convexity* property of $\mathsf{D}_{e^{\varepsilon'}}$ [Balle et al., 2018, Theorem 2] with $\varepsilon' = \log(1+\gamma(e^{\varepsilon}-1))$ yield

$$\begin{aligned}\mathsf{D}_{e^{\varepsilon'}}(\mu K\|\nu K) &\le \gamma\big((1-\beta)\mathsf{D}_{e^{\varepsilon}}(\mu\tilde{K}\|\omega)+\beta\mathsf{D}_{e^{\varepsilon}}(\mu\tilde{K}\|\nu\tilde{K})\big)\\ &\le \gamma\big((1-\beta)\mathsf{D}_{e^{\varepsilon}}(\mu\tilde{K}\|\omega)+\beta\mathsf{D}_{e^{\varepsilon}}(\mu\|\nu)\big)\\ &\le \gamma\big((1-\beta)\mathsf{D}_{e^{\varepsilon}}(\mu\tilde{K}\|\omega)+\beta\delta\big)\ ,\end{aligned}$$

where $\beta = e^{\varepsilon'-\varepsilon}$. Using (6.1) we can expand the remaining divergence above as follows:

$$\mathsf{D}_{e^{\varepsilon}}(\mu\tilde{K}\|\omega) = \mathsf{D}_{e^{\varepsilon}}(\mu\tilde{K}\|\tilde{\omega}\tilde{K}) \le \mathsf{D}_{e^{\varepsilon}}(\mu\|\tilde{\omega}) \le \alpha\mathsf{D}_{e^{\varepsilon}}(\mu\|\nu) \le \alpha\delta\ ,$$

where we used the definition of $\tilde{\omega}$ and joint convexity. Since α was arbitrary, we can now take the limit $\alpha\to 0$ to obtain the bound $\mathsf{D}_{e^{\varepsilon'}}(\mu K\|\nu K)\le\gamma\delta e^{\varepsilon'-\varepsilon}$. □

Proof of Theorem 15. It follows from Theorems 16, 17, 18 and 19. □

We conclude this section by noting that the conditions in Definition 22, despite being quite natural, might be too stringent for proving amplification for DP mechanisms on, say, $\mathbb{R}^d$.

One way to see this is to interpret the operator $K : \mathbb{X} \to \mathscr{P}(\mathbb{Y})$ as a mechanism and to note that the uniform mixing conditions on K can be rephrased in terms of *local DP* (LDP) [Kasiviswanathan et al., 2011] properties (see Table 6.2 for property[3] translations)where the supremum is taken over *any pair* of inputs (instead of neighboring ones). This motivates the results on next section, where we look for finer conditions to prove amplification by stochastic post-processing.

6.4 Amplification From Couplings

In this section we turn to coupling-based proofs of amplification by post-processing under the Rényi DP framework. Our first result is a measure-theoretic generalization of the shift-reduction lemma in [Feldman et al., 2018] which does not require the underlying space to be a normed vector space. The main differences in our proof are to use explicit couplings instead of the shifted Rényi divergence which implicitly relies on the existence of a norm (through the use of W_∞), and replace the identity $U + W - W = U$ between random variables which depends on the vector-space structure with a transport operators H_π and $H_{\pi'}$ which satisfy $\mu H_{\pi'} H_\pi = \mu$ in a general measure-theoretic setting.

Given a coupling $\pi \in \mathscr{C}(\mu, \nu)$ with $\mu, \nu \in \mathscr{P}(\mathbb{X})$, we construct a *transport* Markov operator $H_\pi : \mathbb{X} \to \mathscr{P}(\mathbb{X})$ with kernel[4] $h_\pi(x,y) = \frac{p_\pi(x,y)}{p_\mu(x)}$, where $p_\pi = \frac{d\pi}{d\lambda \otimes \lambda}$ and $p_\mu = \frac{d\mu}{d\lambda}$. It is immediate to verify from the definition that H_π is a Markov operator satisfying the transport property $\mu H_\pi = \nu$ (see Lemma 4).

Lemma 4. *The transport operator H_π with $\pi \in \mathscr{C}(\mu, \nu)$ satisfies $\mu H_\pi = \nu$.*

Proof. Take an arbitrary event E and note that:

$$
\begin{aligned}
(\mu H_\pi)(E) &= \int_{\mathbb{X}} H_\pi(x)(E)\mu(dx) = \int_{\mathbb{X}} \int_E h_\pi(x,y)\mu(dx)\lambda(dy) = \int_{\mathbb{X}} \int_E \frac{p_\pi(x,y)}{p_\mu(x)} \mu(dx)\lambda(dy) \\
&= \int_{\mathbb{X}} \int_E p_\pi(x,y)\lambda(dx)\lambda(dy) = \int_E p_\nu(y)\lambda(dy) = \nu(E) \ ,
\end{aligned}
$$

[3] The *blanket condition* is a necessary condition for LDP introduced in [Balle et al., 2019] to analyze privacy amplification by shuffling.

[4] Here we use the convention $\frac{0}{0} = 0$.

where we used the coupling property $\int_{\mathbb{X}} p_\pi(x,y)\lambda(dx) = p_\nu(y)$. □

Theorem 20. *Let $\alpha \geq 1$, $\mu, \nu \in \mathscr{P}(\mathbb{X})$ and $K \in \mathscr{K}(\mathbb{X}, \mathbb{Y})$. For any distribution $\omega \in \mathscr{P}(\mathbb{X})$ and coupling $\pi \in \mathscr{C}(\omega, \mu)$ we have*

$$\mathsf{R}_\alpha(\mu K \| \nu K) \leq \mathsf{R}_\alpha(\omega \| \nu) + \sup_{x \in \mathrm{supp}(\nu)} \mathsf{R}_\alpha((H_\pi K)(x) \| K(x)) \ . \tag{6.2}$$

Proof. Let $\omega \in \mathscr{P}(\mathbb{X})$ and $\pi \in \mathscr{C}(\omega, \mu)$ be as in the statement, and let $\pi' = C(\mu, \omega)$. Note that taking H_π and $H_{\pi'}$ to be the corresponding transport operators we have $\mu = \mu H_{\pi'} H_\pi = \omega H_\pi$. Now, given a $\lambda \in \mathscr{P}(\mathbb{X} \times \mathbb{X})$ let $\Pi_2(\lambda) = \int \lambda(dx, \cdot)$ denote the marginal of λ on the second coordinate. In particular, if $\mu \otimes K$ denotes the joint distribution of μ and μK, then we have $\Pi_2(\mu \otimes K) = \mu K$. Thus, by the data processing inequality we have

$$\mathsf{R}_\alpha(\mu K \| \nu K) = \mathsf{R}_\alpha(\omega H_\pi K \| \nu K) = \mathsf{R}_\alpha(\Pi_2(\omega \otimes H_\pi K) \| \Pi_2(\nu \otimes K)) \leq \mathsf{R}_\alpha(\omega \otimes H_\pi K \| \nu \otimes K) \ .$$

The final step is to expand the RHS of the derivation above as follows:

$$\begin{aligned}
e^{(\alpha-1)\mathsf{R}_\alpha(\omega \otimes H_\pi K \| \nu \otimes K)} &= \iint \left(\frac{d(\omega \otimes H_\pi K)}{d(\nu \otimes K)} \right)^\alpha \nu(dx) K(x, dy) \\
&= \iint \left(\frac{p_\omega(x) \int h_\pi(x, dz) k(z, y)}{p_\nu(x) k(x, y)} \right)^\alpha \nu(dx) K(x, dy) \\
&= \iint \left(\frac{p_\omega(x)}{p_\nu(x)} \right)^\alpha \left(\frac{\int h_\pi(x, dz) k(z, y)}{k(x, y)} \right)^\alpha \nu(dx) K(x, dy) \\
&\leq \left(\int \left(\frac{p_\omega(x)}{p_\nu(x)} \right)^\alpha \nu(dx) \right) \left(\sup_x \int \left(\frac{\int h_\pi(x, dz) k(z, y)}{k(x, y)} \right)^\alpha K(x, dy) \right) \\
&= e^{(\alpha-1)\mathsf{R}_\alpha(\omega \| \nu)} \cdot e^{(\alpha-1) \sup_x \mathsf{R}_\alpha((H_\pi K)(x) \| K(x))} \ ,
\end{aligned}$$

where the supremums are taken with respect to $x \in \mathrm{supp}(\nu)$. □

Note that this result captures the data-processing inequality for Rényi divergences since taking $\omega = \mu$ and the identity coupling yields $\mathsf{R}_\alpha(\mu K \| \nu K) \leq \mathsf{R}_\alpha(\mu \| \nu)$. The next examples

illustrate the use of this theorem to obtain amplification by operators corresponding to the addition of Gaussian and Laplace noise.

Example 1 (Iterated Gaussian)**.** *We can show that* (6.2) *is tight and equivalent to the shift-reduction lemma [Feldman et al., 2018] on* $\mathbb{R}^d$ *by considering the simple scenario of adding Gaussian noise to the output of a Gaussian mechanism. In particular, suppose* $M(D) = \mathscr{N}(f(D), \sigma_1^2 I)$ *for some function* f *with global* L_2*-sensitivity* Δ *and the Markov operator* K *is given by* $K(x) = \mathscr{N}(x, \sigma_2^2 I)$*. The post-processed mechanism is given by* $(K \circ M)(D) = \mathscr{N}(f(D), (\sigma_1^2 + \sigma_2^2)I)$*, which satisfies* $(\alpha, \frac{\alpha\Delta^2}{2(\sigma_1^2+\sigma_2^2)})$*-RDP. We now show how this result also follows from Theorem 20. Given two datasets* $D \simeq D'$ *we write* $\mu = M(D) = \mathscr{N}(u, \sigma_1^2 I)$ *and* $\nu = M(D') = \mathscr{N}(v, \sigma_1^2 I)$ *with* $\|u - v\| \leq \Delta$*. We take* $\omega = \mathscr{N}(w, \sigma_1^2 I)$ *for some* w *to be determined later, and couple* ω *and* μ *through a translation* $\tau = u - w$*, yielding a coupling* π *with* $p_\pi(x,y) \propto \exp(-\frac{\|x-w\|^2}{2\sigma_1^2})\mathbb{I}[y = x + \tau]$ *and a transport operator* H_π *with kernel* $h_\pi(x,y) = \mathbb{I}[y = x + \tau]$*. Plugging these into* (6.2) *we get*

$$\mathsf{R}_\alpha(\mu K \| \nu K) \leq \frac{\alpha\|w-v\|^2}{2\sigma_1^2} + \sup_{x \in \mathbb{R}^d} \mathsf{R}_\alpha(K(x+\tau)\|K(x)) = \frac{\alpha}{2}\left(\frac{\|w-v\|^2}{\sigma_1^2} + \frac{\|u-w\|^2}{\sigma_2^2}\right) .$$

Finally, taking $w = \theta u + (1-\theta)v$ *with* $\theta = (1 + \frac{\sigma_2^2}{\sigma_1^2})^{-1}$ *yields* $\mathsf{R}_\alpha(\mu K \| \nu K) \leq \frac{\alpha\Delta^2}{2(\sigma_1^2+\sigma_2^2)}$.

Example 2 (Iterated Laplace)**.** *To illustrate the flexibility of this technique, we also apply it to get an amplification result for iterated Laplace noise, in which Laplace noise is added to the output of a Laplace mechanism. We begin by noting a negative result that there is no amplification in the* $(\varepsilon, 0)$*-DP regime.*

Lemma 5. *Let* $M(D) = \mathbf{Lap}(f(D), \lambda_1)$ *for some function* $f : \mathbb{D} \to \mathbb{R}$ *with global* L_1*-sensitivity* Δ *and let the Markov operator* K *be given by* $K(x) = \mathbf{Lap}(x, \lambda_2)$*. The post-processed mechanism* $(K \circ M)$ *does not achieve* $(\varepsilon, 0)$*-DP for any* $\varepsilon < \frac{\Delta}{\max\{\lambda_1, \lambda_2\}}$*. Note that* M *achieves* $(\frac{\Delta}{\lambda_1}, 0)$*-DP and* $K(f(D))$ *achieves* $(\frac{\Delta}{\lambda_2}, 0)$*-DP.*

Proof. This can be shown by directly analyzing the distribution arising from the sum of two independent Laplace variables. Let $Lap2(\lambda_1, \lambda_2)$ denote this distribution. In the following

equations, we assume $x > 0$. Due to symmetry around the origin, densities at negative values can be found by looking instead at the corresponding positive location.

$$
\begin{aligned}
Lap2(x;\lambda_1,\lambda_2) &= \int_{-\infty}^{\infty} \frac{1}{2\lambda_1} \exp\left(-\frac{|x-t|}{\lambda_1}\right) \frac{1}{2\lambda_2} \exp\left(-\frac{|t|}{\lambda_2}\right) dt \\
&= \frac{1}{4\lambda_1\lambda_2} \int_{-\infty}^{\infty} \exp\left(-\frac{\lambda_2|x-t| + \lambda_1|t|}{\lambda_1\lambda_2}\right) dt \\
&= \frac{1}{4\lambda_1\lambda_2} \left(\int_{-\infty}^{0} e^{-\frac{\lambda_2(x-t)-\lambda_1 t}{\lambda_1\lambda_2}} dt + \int_{0}^{x} e^{-\frac{\lambda_2(x-t)+\lambda_1 t}{\lambda_1\lambda_2}} dt + \int_{x}^{\infty} e^{-\frac{-\lambda_2(x-t)+\lambda_1 t}{\lambda_1\lambda_2}} dt \right) \\
&= \frac{1}{4\lambda_1\lambda_2} \left(\int_{-\infty}^{0} e^{-\frac{\lambda_2 x-(\lambda_1+\lambda_2)t}{\lambda_1\lambda_2}} dt + \int_{0}^{x} e^{-\frac{\lambda_2 x+(\lambda_1-\lambda_2)t}{\lambda_1\lambda_2}} dt + \int_{x}^{\infty} e^{-\frac{-\lambda_2 x+(\lambda_1+\lambda_2)t}{\lambda_1\lambda_2}} dt \right) \\
&= \frac{1}{4\lambda_1\lambda_2} \left(\frac{e^{-\frac{\lambda_2 x-(\lambda_1+\lambda_2)t}{\lambda_1\lambda_2}}}{(\lambda_1+\lambda_2)/\lambda_1\lambda_2} \Big|_{t=-\infty}^{t=0} + \int_{0}^{x} e^{-\frac{\lambda_2 x+(\lambda_1-\lambda_2)t}{\lambda_1\lambda_2}} dt + \frac{e^{-\frac{-\lambda_2 x+(\lambda_1+\lambda_2)t}{\lambda_1\lambda_2}}}{(\lambda_1+\lambda_2)/\lambda_1\lambda_2} \Big|_{t=x}^{t=\infty} \right)
\end{aligned}
$$

The integration on the middle term varies between the cases $\lambda_1 = \lambda_2$ and $\lambda_1 \neq \lambda_2$. Finishing this derivation and replacing x with $|x|$ to account for both positive and negative values, we get a complete expression for our $Lap2(\lambda_1,\lambda_2)$ density.

$$
Lap2(x;\lambda_1,\lambda_2) = \begin{cases} \frac{1}{4}\left((\frac{1}{\lambda_1+\lambda_2} + \frac{1}{\lambda_1-\lambda_2}) e^{-\frac{|x|}{\lambda_1}} + (\frac{1}{\lambda_1+\lambda_2} - \frac{1}{\lambda_1-\lambda_2}) e^{-\frac{|x|}{\lambda_2}} \right) & \text{if } \lambda_1 \neq \lambda_2 \ , \\ \frac{1}{4\lambda_1^2} e^{-\frac{|x|}{\lambda_1}} (\lambda_1 + |x|) & \text{if } \lambda_1 = \lambda_2 \ . \end{cases} \tag{6.3}
$$

To finish this lemma, we need to derive the best $(\varepsilon,0)$-DP guarantee offered by adding noise from $Lap2(\lambda_1,\lambda_2)$. From the post-processing property of DP and the commutativity of additive mechanisms, we know this guarantee is upper-bounded by $\Delta/\max\{\lambda_1,\lambda_2\}$. A direct computation of $\lim_{x\to\infty} \log(Lap2(x;\lambda_1,\lambda_2)/Lap2(x+\Delta;\lambda_1,\lambda_2))$ results in $\Delta/\max\{\lambda_1,\lambda_2\}$ in both cases of equation (6.3). This arises from the limit depending entirely on the dominating term with the largest exponent. Therefore, this lower-bounds the privacy guarantee by the same value. Thus we can conclude this is the exact level of $(\varepsilon,0)$-DP offered by this mechanism.

□

However, the iterated Laplace mechanism $K \circ M$ above still offers additional privacy in the relaxed RDP setting. An application of (6.2) *allows us to identify some of this improvement. Recall from [Mironov, 2017, Corollary 2] that M satisfies $(\alpha, \frac{1}{\alpha-1}\log g_\alpha(\frac{\Delta}{\lambda_1}))$-RDP with $g_\alpha(z) = \frac{\alpha}{2\alpha-1}\exp(z(\alpha-1)) + \frac{\alpha-1}{2\alpha-1}\exp(-z\alpha)$. As in Example 1, we take $\omega = \mathbf{Lap}(w, \lambda_1)$ for some w to be determined later, and couple ω and μ through a translation $\tau = u - w$. Through* (6.2) *we obtain*

$$\begin{aligned}\mathsf{R}_\alpha(\mu K \| \nu K) &\leq \frac{1}{\alpha-1}\log\left(g_\alpha\left(\frac{|w-v|}{\lambda_1}\right)\right) + \sup_{x\in\mathbb{R}} \mathsf{R}_\alpha(K(x+\tau)\|K(x)) \\ &= \frac{1}{\alpha-1}\log\left(g_\alpha\left(\frac{|w-v|}{\lambda_1}\right) g_\alpha\left(\frac{|u-w|}{\lambda_2}\right)\right) \quad .\end{aligned}$$

In the simple case where $\lambda_1 = \lambda_2$, an amplification result is observed from the log-convexity of g_α, since $g_\alpha(a)g_\alpha(b) \leq g_\alpha(a+b)$. When $\lambda_1 \neq \lambda_2$, certain values of w still result in amplification, but they depend nontrivially on α. However, we also observe that this improvement vanishes as $\alpha \to \infty$, since the necessary convexity also vanishes. In the limit, the lowest upper bound offered by (6.2) *for R_∞ (which reduces to $(\varepsilon, 0)$-DP) matches the $\frac{\Delta}{\max\{\lambda_1,\lambda_2\}}$ result of Lemma 5.*

Example 3 (Lipschitz Kernel)**.** *As a warm-up for the results in Section 6.4.1, we now re-work Example 1 with a slightly more complex Markov operator. Suppose ψ is an L-Lipschitz map*[5] *and let $K(x) = \mathcal{N}(\psi(x), \sigma_2^2 I)$. Taking M to be the Gaussian mechanism from Example 1, we will show that the post-processed mechanism $K \circ M$ satisfies $(\alpha, \frac{\alpha\Delta^2}{2\sigma_*^2})$-RDP with $\sigma_*^2 = \sigma_1^2 + \frac{\sigma_2^2}{L^2}$. To prove this bound, we instantiate the notation from Example 1, and use the same coupling strategy to obtain*

$$\mathsf{R}_\alpha(\mu K\|\nu K) \leq \frac{\alpha}{2}\left(\frac{\|w-v\|^2}{\sigma_1^2} + \sup_{x\in\mathbb{R}^d}\frac{\|\psi(x+\tau)-\psi(x)\|^2}{\sigma_2^2}\right) \leq \frac{\alpha}{2}\left(\frac{\|w-v\|^2}{\sigma_1^2} + \frac{L^2\|u-w\|^2}{\sigma_2^2}\right) \quad ,$$

where the second inequality uses the Lipschitz property. As before, the result follows from

[5] That is, $\|\psi(x) - \psi(y)\| \leq L\|x-y\|$ for any pair x, y.

taking $w = \theta u + (1-\theta)v$ *with* $\theta = (1 + \frac{\sigma_2^2}{L^2\sigma_1^2})^{-1}$. *This example shows that we get amplification (i.e.* $\sigma_*^2 > \sigma_1^2$*) for any* $L < \infty$ *and* $\sigma_2 > 0$*, although the amount of amplification decreases as* L *grows. On the other hand, for* $L < 1$ *the amplification is stronger than just adding Gaussian noise (Example 1).*

6.4.1 Amplification by Iteration in Noisy Projected SGD with Strongly Convex Losses

Now we use Theorem 20 and the computations above to show that the proof of privacy amplification by iteration [Feldman et al., 2018, Theorem 22] can be extended to explicitly track the Lipschitz coefficients in a "noisy iteration" algorithm. In particular, this allows us to show an exponential improvement on the rate of privacy amplification by iteration in noisy SGD when the loss is strongly convex. To obtain this result we first provide an *iterated* version of Theorem 20 in $\mathbb{R}^d$ with Lipschitz Gaussian kernels. This version of the analysis introduces an explicit dependence on the W_∞ distances along an "interpolating" path between the initial distributions $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$ which could later be optimized for different applications. In our view, this helps to clarify the intuition behind the previous analysis of amplification by iteration.

Theorem 21. *Let* $\alpha \geq 1$*,* $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$ *and let* $\mathbb{K} \subseteq \mathbb{R}^d$ *be a convex set. Suppose* $K_1, \ldots, K_r \in \mathscr{K}(\mathbb{R}^d, \mathbb{R}^d)$ *are Markov operators where* $Y_i \sim K_i(x)$ *is obtained as*[6] $Y_i = \Pi_{\mathbb{K}}(\psi_i(x) + Z_i)$ *with* $Z_i \sim \mathscr{N}(0, \sigma^2 I)$*, where the maps* $\psi_i : \mathbb{K} \to \mathbb{R}^d$ *are* L*-Lipschitz for all* $i \in [r]$*. For any* $\mu_0, \mu_1, \ldots, \mu_r \in \mathscr{P}(\mathbb{R}^d)$ *with* $\mu_0 = \mu$ *and* $\mu_r = \nu$ *we have*

$$\mathsf{R}_\alpha(\mu K_1 \cdots K_r \| \nu K_1 \cdots K_r) \leq \frac{\alpha L^2}{2\sigma^2} \sum_{i=1}^{r} L^{2(r-i)} \mathsf{W}_\infty(\mu_i, \mu_{i-1})^2 \ . \tag{6.4}$$

Furthermore, if $L \leq 1$ *and* $\mathsf{W}_\infty(\mu, \nu) = \Delta$*, then*

$$\mathsf{R}_\alpha(\mu K_1 \cdots K_r \| \nu K_1 \cdots K_r) \leq \frac{\alpha \Delta^2 L^{r+1}}{2r\sigma^2} \ . \tag{6.5}$$

[6] Here $\Pi_{\mathbb{K}}(x) = \operatorname{argmin}_{y \in \mathbb{K}} \|x - y\|$ denotes the projection operator onto the convex set $\mathbb{K} \subseteq \mathbb{R}^d$.

The proof of Theorem 21 relies on the following technical lemma about the effect of a projected Lipschitz Gaussian operator on the ∞-Wasserstein distance between two distributions.

Lemma 6. *Let $\mathbb{K} \subseteq \mathbb{R}^d$ be a convex set and $\psi : \mathbb{K} \to \mathbb{R}^d$ be L-Lipschitz. Suppose $K \in \mathscr{K}(\mathbb{R}^d, \mathbb{R}^d)$ is a Markov operator where $Y \sim K(x)$ is obtained as $Y = \Pi_{\mathbb{K}}(\psi(x) + Z)$ with $Z \sim \mathscr{N}(0, \sigma^2 I)$. Then, for any $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$ we have $\mathsf{W}_\infty(\mu K, \nu K) \leq L\mathsf{W}_\infty(\mu, \nu)$.*

Proof. Let $\pi \in \mathscr{C}(\mu, \nu)$ be a witness of $\mathsf{W}_\infty(\mu, \nu) = \Delta$. We construct a witness of $\mathsf{W}_\infty(\mu K, \nu K) \leq L\Delta$ as follows: sample $(X, X') \sim \pi$ and $Z \sim \mathscr{N}(0, \sigma^2 I)$ and then let $Y = \Pi_{\mathbb{K}}(\psi(X) + Z)$ and $Y' = \Pi_{\mathbb{K}}(\psi(X') + Z)$. It is clear from the construction that $\mathsf{Law}((Y, Y')) \in \mathscr{C}(\mu K, \nu K)$. Furthermore, by the Lipschitz assumption on ψ and that fact that the map $\Pi_{\mathbb{K}}$ is contractive, the following holds almost surely:

$$\|Y - Y'\| \leq \|\psi(X) - \psi(X')\| \leq L\|X - X'\| \leq L\Delta \ .$$

□

Proof of Theorem 21. We prove (6.4) by induction on r. For the base case $r = 1$ we apply Theorem 20 with $\omega = \nu$ and a coupling $\pi \in \mathscr{C}(\nu, \mu)$ witnessing that $\mathsf{W}_\infty(\mu, \nu) = \Delta$. This choice of coupling guarantees that for any $x \in \mathrm{supp}(\nu)$ we have $\mathrm{supp}(H_\pi(x)) \subseteq B_\Delta(x)$, where $B_\Delta(x)$ is the ball of radius Δ around x. Note also that $(H_\pi K_1)(x) = H_\pi(x) K_1$. Thus, from (6.2) we obtain, using Hölder's inequality and the monotonicity of the logarithm, that:

$$\begin{aligned}\mathsf{R}_\alpha(\mu K_1 \| \nu K_1) &\leq \sup_{x \in \mathrm{supp}(\nu)} \mathsf{R}_\alpha((H_\pi K_1)(x) \| K_1(x)) \leq \sup_{x \in \mathrm{supp}(\nu)} \sup_{y \in \mathrm{supp}(H_\pi(x))} \mathsf{R}_\alpha(K_1(y) \| K_1(x)) \\ &\leq \sup_{\|x-y\| \leq \Delta} \mathsf{R}_\alpha(K_1(y) \| K_1(x)) \ .\end{aligned}$$

Now note that the Markov operator K_1 can be obtained by post-processing $\tilde{K}_1(x) = \mathscr{N}(\psi_1(x), \sigma^2 I)$ with the projection $\Pi_{\mathbb{K}}$. Thus, by the data processing inequality we

obtain

$$\sup_{\|x-y\|\leq\Delta} \mathsf{R}_\alpha(K_1(y)\|K_1(x)) \leq \sup_{\|x-y\|\leq\Delta} \mathsf{R}_\alpha(\tilde{K}_1(y)\|\tilde{K}_1(x)) \\ = \sup_{\|x-y\|\leq\Delta} \frac{\alpha\|\psi_1(x)-\psi_1(y)\|^2}{2\sigma^2} \leq \frac{\alpha\Delta^2L^2}{2\sigma^2} \ .$$

For the inductive case we suppose that (6.4) holds for some $r \geq 1$ and consider the case $r+1$, in which we need to bound $\mathsf{R}_\alpha(\mu K_1 \cdots K_{r+1}\|\nu K_1 \cdots K_{r+1})$. Let $\mu_0, \mu_1, \ldots, \mu_{r+1}$ be a sequence of distributions with $\mu_0 = \mu$ and $\mu_{r+1} = \nu$. Applying (6.2) with $\omega = \mu_1 K_1 \cdots K_r$ and some coupling $\pi \in \mathscr{C}(\mu_1 K_1 \cdots K_r, \mu K_1 \cdots K_r)$ we have

$$\mathsf{R}_\alpha(\mu K_1 \cdots K_{r+1}\|\nu K_1 \cdots K_{r+1}) \leq \mathsf{R}_\alpha(\mu_1 K_1 \cdots K_r\|\nu K_1 \cdots K_r) \\ + \sup_{x\in\mathrm{supp}(\nu K_1\cdots K_r)} \mathsf{R}_\alpha((H_\pi K_{r+1})(x)\|K_{r+1}(x)) \ .$$

By the inductive hypothesis, the first term in the RHS above can be bounded as follows:

$$\mathsf{R}_\alpha(\mu_1 K_1 \cdots K_r\|\nu K_1 \cdots K_r) \leq \frac{\alpha L^2}{2\sigma^2} \sum_{i=1}^{r} L^{2(r-i)} \mathsf{W}_\infty(\mu_{i+1}, \mu_i)^2 \\ = \frac{\alpha L^2}{2\sigma^2} \sum_{i=2}^{r+1} L^{2(r+1-i)} \mathsf{W}_\infty(\mu_i, \mu_{i-1})^2 \ .$$

To bound the second term we assume the coupling π is a witness of $\mathsf{W}_\infty(\mu_1 K_1 \cdots K_r, \mu K_1 \cdots K_r) = \Delta'$, in which case a similar argument to the one we used in the base case yields:

$$\sup_x \mathsf{R}_\alpha((H_\pi K_{r+1})(x)\|K_{r+1}(x)) \leq \sup_x \sup_{y\in\mathrm{supp}(H_\pi(x))} \mathsf{R}_\alpha(K_{r+1}(y)\|K_{r+1}(x)) \\ \leq \sup_{\|x-y\|\leq\Delta'} \mathsf{R}_\alpha(K_{r+1}(y)\|K_{r+1}(x)) \\ \leq \frac{\alpha\Delta'^2L^2}{2\sigma^2} \leq \frac{\alpha L^{2r+2}\mathsf{W}_\infty(\mu_1,\mu)^2}{2\sigma^2} \ ,$$

where the last inequality follows from Lemma 6. Plugging the last three inequalities together we

finally obtain

$$\mathsf{R}_\alpha(\mu K_1 \cdots K_{r+1} \| \nu K_1 \cdots K_{r+1}) \leq \frac{\alpha L^{2r+2} \mathsf{W}_\infty(\mu_1, \mu_0)^2}{2\sigma^2} + \frac{\alpha L^2}{2\sigma^2} \sum_{i=2}^{r+1} L^{2(r+1-i)} \mathsf{W}_\infty(\mu_i, \mu_{i-1})^2$$
$$= \frac{\alpha L^2}{2\sigma^2} \sum_{i=1}^{r+1} L^{2(r+1-i)} \mathsf{W}_\infty(\mu_i, \mu_{i-1})^2 \quad .$$

When $L \leq 1$, we can obtain (6.5) from (6.4) as follows. First, construct a sequence of distributions $\mu_0, \ldots, \mu_r$ such that $\Delta_i \triangleq \mathsf{W}_\infty(\mu_i, \mu_{i-1}) = \Delta_0 L^i$ for $i \in [r]$, where $\Delta_0 = \frac{\Delta}{L} \frac{1-L}{1-L^r}$ is a normalization constant chosen such that $\sum_{i \in [r]} \Delta_i = \Delta$. With this choice plugged into (6.4) we obtain

$$\mathsf{R}_\alpha(\mu K_1 \cdots K_r \| \nu K_1 \cdots K_r) \leq \frac{\alpha L^2}{2\sigma^2} r \Delta_0^2 L^{2r} = \frac{\alpha \Delta^2 L^{r+1} r}{2\sigma^2} \left(\frac{L^{-\frac{1}{2}} - L^{\frac{1}{2}}}{L^{-\frac{r}{2}} - L^{\frac{r}{2}}} \right)^2 = \frac{\alpha \Delta^2 L^{r+1} r}{2\sigma^2} \phi(L)^2 \quad .$$

Now we note the function $\phi(L)$ defined above is increasing in $[0,1]$ and furthermore $\lim_{L \to 1} \phi(L) = \frac{1}{r}$, which can be checked by applying L'Hôpital's rule twice. Thus, we can plug the inequality $\phi(L) \leq \frac{1}{r}$ above to obtain (6.5).

But we still need to show that a sequence $\mu_0, \ldots, \mu_r$ with Δ_i as above exists. To construct such a sequence we let $\pi \in \mathscr{C}(\mu, \nu)$ be a witness of $\mathsf{W}_\infty(\mu, \nu) = \Delta$, take random variables $(X, X') \sim \pi$, and define $\mu_i = \mathsf{Law}((1-\theta_i)X + \theta_i X')$ with $\theta_i = \frac{\Delta_0}{\Delta} \sum_{j=1}^i L^j = \frac{1-L^i}{1-L^r}$. Clearly we get $\mu_0 = \mathsf{Law}(X) = \mu$ and $\mu_r = \mathsf{Law}(X') = \nu$.

To see that $\mathsf{W}_\infty(\mu_i, \mu_{i-1}) \leq \Delta_0 L^i$ we construct a coupling between μ_i and μ_{i-1} as follows: sample $(X, X') \sim \pi$ and let $Y = (1-\theta_i)X + \theta_i X'$ and $Y' = (1-\theta_{i-1})X + \theta_{i-1} X'$. Clearly we have $\mathsf{Law}((Y, Y')) \in \mathscr{C}(\mu_i, \mu_{i-1})$. Furthermore, with probability one the following holds:

$$\|Y - Y'\| = \|(\theta_{i-1} - \theta_i)X - (\theta_{i-1} - \theta_i)X'\| = \frac{\Delta_0}{\Delta} L^i \|X - X'\| \leq \Delta_0 L^i \quad ,$$

where the last inequality uses that π is a witness of $\mathsf{W}_\infty(\mu, \nu) \leq \Delta$. This concludes the proof. □

Note how taking $L = 1$ in the bound above we obtain $\frac{\alpha \Delta^2}{2r\sigma^2} = O(1/r)$, which matches

[Feldman et al., 2018, Theorem 1]. On the other hand, for L strictly smaller than 1, the analysis above shows that the amplification rate is $O(L^{r+1}/r)$ as a consequence of the maps ψ_i being strict contractions, i.e. $\|\psi_i(x) - \psi_i(y)\| < \|x-y\|$. For $L > 1$ this result is not useful since the sum will diverge; however, the proof could easily be adapted to handle the case where each ψ_i is L_i-Lipschitz with some $L_i > 1$ and some $L_i < 1$. We now apply this result to improve the per-person privacy guarantees of noisy projected SGD (Algorithm 5) in the case where the loss function is smooth and strongly convex.

Algorithm 5: Noisy Projected Stochastic Gradient Descent — NoisyProjSGD$(D, \ell, \eta, \sigma, \xi_0)$

Input: Dataset $D = (z_1, \ldots, z_n)$, loss function $\ell : \mathbb{K} \times \mathbb{D} \to \mathbb{R}$, learning rate η, noise parameter σ, initial distribution $\xi_0 \in \mathscr{P}(\mathbb{K})$

Sample $x_0 \sim \xi_0$

for $i \in [n]$ **do**

  $x_i \leftarrow \Pi_{\mathbb{K}}\left(x_{i-1} - \eta(\nabla_x \ell(x_{i-1}, z_i) + Z)\right)$ with $Z \sim \mathscr{N}(0, \sigma^2 I)$

return x_n

A function $f : \mathbb{K} \subseteq \mathbb{R}^d \to \mathbb{R}$ defined on a convex set is *β-smooth* if it is continuously differentiable and ∇f is β-Lipschitz, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x-y\|$, and is ρ-strongly convex if the function $g(x) = f(x) - \frac{\rho}{2}\|x\|^2$ is convex. When we say that a loss function $\ell : \mathbb{K} \times \mathbb{D} \to \mathbb{R}$ satisfies a property (e.g. smoothness) we mean the property is satisfied by $\ell(\cdot, z)$ for all $z \in \mathbb{D}$. Furthermore, we recall from [Feldman et al., 2018] that a mechanism $M : \mathbb{D}^n \to \mathbb{X}$ satisfies *(α, ε)-RDP at index i* if $\mathsf{R}_\alpha(M(D) \| M(D')) \leq \varepsilon$ holds for any pair of databases D and D' *differing on the ith coordinate*.

Theorem 22. *Let $\ell : \mathbb{K} \times \mathbb{D} \to \mathbb{R}$ be a C-Lipschitz, β-smooth, ρ-strongly convex loss function. If $\eta \leq \frac{2}{\beta+\rho}$, then* NoisyProjSGD$(D, \ell, \eta, \sigma, \xi_0)$ *satisfies $(\alpha, \alpha\varepsilon_i)$-RDP at index i, where $\varepsilon_n = \frac{2C^2}{\sigma^2}$ and $\varepsilon_i = \frac{2C^2}{(n-i)\sigma^2}\left(1 - \frac{2\eta\beta\rho}{\beta+\rho}\right)^{\frac{n-i+1}{2}}$ for $1 \leq i \leq n-1$.*

Since [Feldman et al., 2018, Theorem 23] shows that for smooth Lipschitz loss functions the guarantee at index i of NoisyProjSGD is given by $\varepsilon_i = O(\frac{C^2}{(n-i)\sigma^2})$, our result provides an exponential improvement in the strongly convex case. This implies, for example, that using

the technique in [Feldman et al., 2018, Corollary 31] one can show that, in the strongly convex setting, running $\Theta(\log(d))$ additional iterations of NoisyProjSGD on *public* data is enough to attain (up to constant factors) the same optimization error as non-private SGD while providing privacy for all individuals.

To prove Theorem 22 we will use the following well-known fact about convex optimization: gradient iterations on a strongly convex function are strict contractions. The lemma below provides an expression for the contraction coefficient.

Lemma 7. *Let $\mathbb{K} \subseteq \mathbb{R}^d$ be a convex set and suppose the function $f : \mathbb{K} \to \mathbb{R}$ is β-smooth and ρ-strongly convex. If $\eta \leq \frac{2}{\beta+\rho}$, then the map $\psi(x) = x - \eta\nabla f(x)$ is L-Lipschitz on $\mathbb{K}$ with $L = \sqrt{1 - \frac{2\eta\beta\rho}{\beta+\rho}} < 1$.*

Proof. This follows from a standard calculation in convex optimization; see e.g. [Bubeck, 2015, Theorem 3.12]. We reproduce the proof here for completeness. Recall from [Bubeck, 2015, Lemma 3.11] that if a function f is β-smooth and ρ-strongly convex, then for any $x, y \in \mathbb{K}$ we have

$$\frac{\beta\rho}{\beta+\rho}\|x-y\|^2 + \frac{1}{\beta+\rho}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x-y\rangle \ .$$

Using this inequality, one can show the following:

$$\begin{aligned}
\|\psi(x) - \psi(y)\|^2 &= \|(x - \eta\nabla f(x)) - (y - \eta\nabla f(y))\|^2 \\
&= \|x-y\|^2 + \eta^2\|\nabla f(x) - \nabla f(y)\|^2 - 2\eta\langle \nabla f(x) - \nabla f(y), x-y\rangle \\
&\leq \left(1 - \frac{2\eta\beta\rho}{\beta+\rho}\right)\|x-y\|^2 + \eta\left(\eta - \frac{2}{\beta+\rho}\right)\|\nabla f(x) - \nabla f(y)\|^2 \\
&\leq \left(1 - \frac{2\eta\beta\rho}{\beta+\rho}\right)\|x-y\|^2 \ ,
\end{aligned}$$

where the last inequality uses our assumption on η. □

Proof of Theorem 22. Fix $1 \leq i \leq n-1$ and let $D \simeq D'$ be two datasets differing on the ith

coordinate. Let $\xi \triangleq \xi_{i-1} \in \mathscr{P}(\mathbb{R}^d)$ represent the distribution of x_{i-1} in the execution of Algorithm 5 with input D. Since D and D' differ only on the ith coordinate, the distribution of x_{i-1} on input D' is also ξ. Now let $\psi_0(x) = x - \eta\nabla_x\ell(x,z_i)$, $\psi_0'(x) = x - \eta\nabla_x\ell(x,z_i')$, and $\psi_j(x) = x - \eta\nabla_x\ell(x,z_{i+j})$ for $j \in [r]$ with $r = n-i$. Defining the Markov operators K_j, $j \in \{0,\dots,r\}$, where $Y_j \sim K_j(x)$ is given by $K_j(x) = \Pi_{\mathbb{K}}(\psi_j(x) + Z)$ with $Z \sim \mathscr{N}(0,\eta^2\sigma^2 I)$, we immediately obtain that the distribution of the output x_n of NoisyProjSGD(D,ℓ,η,σ) can be written as $\xi K_0 K_1 \cdots K_r$. Similarly, the distribution of the output of NoisyProjSGD(D',ℓ,η,σ) can be written as $\xi K_0' K_1 \cdots K_r$, where $K_0'(x) = \mathscr{N}(\psi_0'(x),\eta^2\sigma^2 I)$. Therefore, to obtain the Rényi differential privacy of NoisyProjSGD(D,ℓ,η,σ) at index i we need to bound $\mathsf{R}_\alpha(\xi K_0 K_1 \cdots K_r \| \xi K_0' K_1 \cdots K_r)$.

With the goal to apply Theorem 21, we first define $\mu = \xi K_0$ and $\nu = \xi K_0'$ and use the Lipschitz assumption on ℓ to conclude that $\mathsf{W}_\infty(\mu,\nu) \le 2\eta C$. Indeed, consider the coupling $\pi \in \mathscr{C}(\mu,\nu)$ obtained by sampling $(Y,Y') \sim \pi$ as follows: sample $X \sim \xi$ and $Z \sim \mathscr{N}(0,\eta^2\sigma^2 I)$, and then let $Y = \Pi_{\mathbb{K}}(\psi_0(X) + Z)$ and $Y' = \Pi_{\mathbb{K}}(\psi_0'(X) + Z)$. Now, since $\ell(\cdot,z_i)$ and $\ell(\cdot,z_i')$ are both C-Lipschitz and $\Pi_{\mathbb{K}}$ is contractive, we see that the following holds almost surely under π:

$$\begin{aligned}\|Y - Y'\| &\le \|\psi_0(X) - \psi_0'(X)\| = \eta\|\nabla_x\ell(X,z_i) - \nabla_x\ell(X,z_i')\| \\ &\le \eta\left(\|\nabla_x\ell(X,z_i)\| + \|\nabla_x\ell(X,z_i)\|\right) \le 2\eta C \ .\end{aligned}$$

Thus, $\mathsf{W}_\infty(\mu,\nu) \le 2\eta C$ as claimed.

Next we note that the assumption $\eta \le \frac{2}{\beta+\rho}$ together with Lemma 7 imply that ψ_j, $j \in [r]$, are all L-Lipschitz with $L = \sqrt{1 - \frac{2\eta\beta\rho}{\beta+\rho}} < 1$. Thus we can apply Theorem 21 with $\Delta = 2\eta C$ to obtain

$$\mathsf{R}_\alpha(\xi K_0 K_1 \cdots K_r \| \xi K_0' K_1 \cdots K_r) \le \frac{2\alpha\eta^2 C^2 L^{n-i+1}}{(n-i)\eta^2\sigma^2} = \frac{2\alpha C^2}{(n-i)\sigma^2}\left(1 - \frac{2\eta\beta\rho}{\beta+\rho}\right)^{\frac{n-i+1}{2}} \ .$$

This concludes the analysis of the case $i < n$.

For the case $i = n$ we need to bound $\mathsf{R}_\alpha(\xi K_0 \| \xi K_0')$, where now ξ is the distribution of x_{n-1}, and the operators K_0 and K_0' are defined as above. By Hölder's inequality, monotonicity of the logarithm, the contractiveness of $\Pi_{\mathbb{K}}$ and the Lipschitz assumption on ℓ we have

$$\begin{aligned}\mathsf{R}_\alpha(\xi K_0 \| \xi K_0') &\leq \sup_{x \in \mathrm{supp}(\xi)} \mathsf{R}_\alpha(K_0(x) \| K_0'(x)) \leq \sup_{x \in \mathbb{R}^d} \mathsf{R}_\alpha(K_0(x) \| K_0'(x)) \\ &\leq \sup_{x \in \mathbb{R}^d} \frac{\alpha\eta^2 \|\nabla_x \ell(x, z_n) - \nabla_x \ell(x, z_n')\|^2}{2\eta^2\sigma^2} \leq \frac{2\alpha C^2}{\sigma^2} \quad .\end{aligned}$$

□

6.5 Conclusion

We have undertaken a systematic study of amplification by post-processing. Our results yield improvements over recent work on amplification by iteration, and introduce a new Ornstein-Uhlenbeck mechanism which is more accurate than the Gaussian mechanism. In the future it would be interesting to study applications of amplification by post-processing. One promising application is *Hierarchical Differential Privacy*, where information is released under increasingly strong privacy constraints (e.g. to a restricted group within a company, globally within a company, and finally to outside parties).

6.6 Acknowledgements

This work was partially supported by NSF grant CCF-1718220.

This chapter is based on the material in Advances in Neural Information Processing Systems 2019 (Borja Balle, Gilles Barthe, Marco Gaboardi, and Joseph Geumlek. "Privacy amplification by mixing and diffusion mechanisms"). The dissertation author was the primary investigator and author of this material.

Chapter 7

Profile-based Privacy Preservation

7.1 Introduction

A great deal of machine learning in the 21st century is carried out on sensitive data, and hence the field of privacy preserving data analysis is of increasing importance. Differential privacy Dwork et al. [2006b], introduced in 2006, has become the dominant paradigm for specifying data privacy. A body of compelling results Chaudhuri et al. [2011, 2012], Kifer et al. [2012], Foulds et al. [2016], Wang et al. [2015b] have been achieved in the ”centralized” model, in which a trusted data curator has raw access to the data while performing the privacy-preserving operations. However, such trust is not always easy to achieve, especially when the trust must also extend to all future uses of the data.

An implementation of differential privacy that has been particularly popular in industrial applications makes each user into their own trusted curator. Commonly referred to as *Local Differential Privacy* Duchi et al. [2013], this model consists of users locally privatizing their own data before submission to an aggregate data curator. Due to the strong robustness of differential privacy under further computations, this model preserves privacy regardless of the trust in the aggregate curator, now or in the future. Two popular industrial systems implementing local differential privacy include Google’s RAPPOR and Apple’s iOS data collection systems.

However, a major barrier for the local model is the undesirable utility sacrifices of the submitted data. A local differential privacy implementation achieves much lower utility than a

similar method that assumes trusts in the curator. Strong lower bounds have been found for the local framework Duchi et al. [2013], leading to pessimistic results requiring massive amounts of data to achieve both privacy and utility.

In this work, we address this challenge by proposing a new restricted privacy definition, called Profile-based privacy. The central idea relies on specifying a graph G of data generating distributions, where edges encode sensitive pairs of distributions that should be made indistinguishable. Our framework does not require that all features of the observed data be obscured; instead only the information connected to identifying the distributions must be perturbed. This side-steps the utility costs of local differential privacy, where every possible pair of observations must be indistinguishable.

7.2 Setup

We begin with defining local differential privacy – a prior privacy framework that is related to our definition.

Definition 23. A mechanism $\mathscr{A} : \mathscr{X} \to \mathscr{Y}$ achieves ε-local differential privacy if for every pair (X, X') of individuals' private records in $\mathscr{X}$ we have:

$$\mathsf{Pr}(\mathscr{A}(X) = Y) \leq e^{\varepsilon} \mathsf{Pr}(\mathscr{A}(X') = Y). \tag{7.1}$$

Concretely, local differential privacy limits the ability of an adversary to increase their confidence in whether an individual's private value is X versus X' even with arbitrary prior knowledge. These protections are robust to any further computation performed on the mechanism output.

Pufferfish Privacy

Pufferfish privacy (Definition 24) is an inferential privacy framework that introduces explicit secret pairs and limits the ability of an adversary to infer secrets across each protected

pair of secrets. Typically, it is defined in its global form.

Definition 24. Given a set of data generation scenarios Θ, and a set $\mathbb{S}_{pairs}$ of pairs (s_i, s_j) of secrets, a mechanism $\mathscr{A} : \mathscr{X} \to \mathscr{Y}$ is achieves $(\varepsilon, \mathbb{S}_{pairs}, \Theta)$-Pufferfish privacy if for every pair (s_i, s_j) in $\mathbb{S}_{pairs}$, and for every data generating distribution $\theta \in \Theta$ that assigns non-zero probability to s_i and s_j, and every output $Y \in \mathscr{Y}$:

$$\mathsf{Pr}(\mathscr{A}(X) = Y | s_i, X \sim \theta) \leq e^{\varepsilon} \mathsf{Pr}(\mathscr{A}(X) = Y | s_j, X \sim \theta). \tag{7.2}$$

The secrets (s_i, s_j) protected by Pufferfish privacy can be any mutually disjoint events. With the flexibility of this framework, literature on this framework typically focuses on specific instantiations Kifer and Machanavajjhala [2012], Song et al. [2017]. The most popular instances either make the secrets more granular and value-dependent, or change the data generation scenario θ to account for correlations not addressed by differential privacy. While our framework can also be massaged into this framework, it is different in two ways; first, it is a *local* framework - the output is a distorted version of a single person's value, and second, it behaves differently from prior Pufferfish instantiations and requires its own mechanisms.

To find an instance matching differential privacy, these secrets take the form of "this individual contributed data with value t" versus "this individual did not contribute", and the set of data generation scenarios consist of all distributions where each individual's data is independently generated.

This secret pair encodes a limit on inferring the individual's contribution.

7.3 Profile-based Privacy Definition

Before we present the definition and discuss its implications, it is helpful to have a specific problem in mind. We present one possible setting in which our profiles have a clear interpretation.

7.3.1 Example: Resource Usage Problem Setting

Imagine a shared workstation with access to several resources, such as network bandwidth, specialized hardware, or electricity usage. Different users might use this workstation, coming from a diverse pool of job titles and roles. An analyst wishes to collect and analyze the metrics of resource usage, but also wishes to respect the privacy of the workstation users. One choice of privacy framework is local differentially privacy, in which every value of a resource usage metric is considered sensitive and privatized. Under our alternative profile-based framework, a choice exists to select only the user identities as the sensitive information protected. This shifts the goal away from hiding all features of the resource usages, and permits measurements to be released more faithfully when not indicative of a user's identity.

7.3.2 Definition of Local Profile-based Differential Privacy

Our privacy definition revolves around a notion of profiles, which represent distinct potential data-generating distributions. To preserve privacy, the mechanism's release must not give too much of an advantage in guessing the release's underlying profile. However, other facets of the observed data can (and should) be preserved, permitting greater utility than local differential privacy.

Definition 25. Given a graph $G = (P,E)$ consisting of a collection P of data generating profiles over the space $\mathscr{X}$ and collection of edges E, a mechanism $\mathscr{A} : \mathscr{X} \to \mathscr{Y}$ achieves (G,ε)-profile-based differential privacy if for every edge $e \in E$ connecting profiles P_i and P_j, and for all outputs $Y \in \mathscr{Y}$ we have:

$$\frac{\mathsf{Pr}(\mathscr{A}(X,P_i) = Y | X \sim P_i)}{\mathsf{Pr}(\mathscr{A}(X,P_j) = Y | X \sim P_j)} \leq e^{\varepsilon}. \tag{7.3}$$

Inherent in this definition is an assumption on adversarial prior knowledge: the adversary knows each profile distribution, but has no further auxiliary information about X. The protected secrets are the identities of the source distributions, and are not directly related to particular

features of the data X. These additional assumptions in the problem setting, however, open up avenues for increased performance. By not attempting to completely privatize the raw observations, information that is less relevant for guessing the sensitive profile identity can be preserved for downstream tasks.

The flexible specification of sensitive pairs via edges in the graph permits privacy design decisions that also impact the privacy-utility trade-off. When particular profile pairs can be declared less sensitive, the perturbations required to blur those profiles can be avoided. Such design decisions would be impractical in the data-oriented local differential privacy setting, where the space of pairs of data sets is intractably large.

The local profile-based differential privacy framework exists as an inverse to the goals seen in maximal-leakage-constrained hypothesis testing Liao et al. [2017], where their hypotheses act similarly to our profiles as data distributions. While they focus on protecting observation-privacy and maintaining distribution-utility, we focus on maintaining observation-utility and protecting distribution-privacy. Both settings are interesting and situational.

7.3.3 Discussion of the Resource Usage Problem

This privacy framework is quite general, and as such it helps to discuss its meaning in more concrete terms. Let us return to the resource usage setting. We'll assume that each user has a personal resource usage profile known prior to the data collection process. The choice of edges in the graph G opens up flexibility over what inferences are sensitive. If the graph has many edges, the broad identity of the workstation user will be hidden by forbidding many potential inferences. However, even with this protection not all the information about resource must be perturbed. For example, if all users require roughly the same amount of electricity at the workstation, then electrical usage metrics will not require much obfuscation. Contrast this with the standard local differential privacy scheme, in which every pair of distinct observed values must be obscured.

A more sparse graph might only connect profiles with the same job title or role. These

sensitive pairs will prevent inferences about particular identities within each role. However, without connections across job titles, no protection is enforced against inferring the job title of the current workstation user. Thus such a graph declares user identities sensitive, while a user's role is not sensitive. This permits the released data to be more faithful to the raw observations, since only the peculiarities of resource usage among users of the same role must be obscured, rather than the peculiarities of the different roles.

One important caveat of this definition is that the profile distributions must be known and are assumed to be released a priori, i.e. they are not considered privacy sensitive. If the user profiles cannot all be released, this can be mitigated somewhat by reducing the granularity of the graph. A graph consisting only of profiles for each distinct job role can still encode meaningful protections, since limiting inferences on job role can also limit inferences on highly correlated information like the user's identity.

The trade-off in profile granularity is subtle. More profiles permit more structure and opportunities for our definition to achieve better utility than local differential privacy, but also require a greater level of a priori knowledge.

7.4 Properties

Our privacy definition enjoys several similar properties to other differential-privacy-inspired frameworks. The post-processing and composition properties are recognized as highly desired traits for privacy definitions Kifer and Machanavajjhala [2012].

Post-Processing

The post-processing property specifies that any additional computation (without access to the private information) on the released output cannot result in worse privacy. Following standard techniques, our definition also shares this data processing inequality.

Observation 7. *If a data sample X_i is drawn from profile P_i, and $\mathscr{A}$ preserves (G,ε)-profile-based privacy, then for any (potentially randomized) function F, the release $F(\mathscr{A}(X_i,P_i))$*

preserves (G,ε)*-profile-based privacy.*

Composition

The composition property allows for multiple privatized releases to still offer privacy even when witnessed together. Our definition also gets a compositional property, although not all possible compositional settings behave nicely. We mitigate the need for composition by focusing on a local model where the data mainly undergoes one privatization.

Profile-based differential privacy enjoys additive composition if truly independent samples X are drawn from the same profile. The proof of this follows the same reasoning as the additive composition of differential privacy.

Observation 8. *If two independent samples* X_1 *and* X_2 *are drawn from profile* P_i*, and* $\mathscr{A}_1$ *preserves* (G,ε_1)*-profile-based privacy and* $\mathscr{A}_2$ *preserves* (G,ε_2)*-profile-based privacy, then the combined release* $(\mathscr{A}_1(X_1,P_i),\mathscr{A}_2(X_2,P_i))$ *preserves* $(G,\varepsilon_1+\varepsilon_2)$*-profile-based privacy.*

A notion of parallel composition can also be applied if two data sets come from two independent processes of selecting a profile. In this setting, information about one instance has no bearing on the other. This matches the parallel composition of differential privacy when applied to multiple independent individuals, and would be the analogous setting to local differential privacy where each individual applies their own mechanism.

Observation 9. *If two profiles* P_i *and* P_j *are independently selected, and two observations* $X_i \sim P_i$ *and* $X_j \sim P_j$ *are drawn, and* $\mathscr{A}_1$ *preserves* (G,ε_1)*-profile-based privacy and* $\mathscr{A}_2$ *preserves* (G,ε_2)*-profile-based privacy, then the combined release* $(\mathscr{A}_1(X_i,P_i),\mathscr{A}_2(X_j,P_j))$ *preserves* $(G,max\{\varepsilon_1,\varepsilon_2\})$*-profile-based privacy.*

The parallel composition result assumes that the choice of $\mathscr{A}_2$ does depend on the first release, or in other words that is non-adaptive. It should also be noted that the privacy guarantee is about how much protection a single edge receives in just one profile selection process. With two releases, clearly more information is being released, but the key idea in this result is that the

information released in the one round has no impact on the secret profile identity of the other round.

However, this framework cannot offer meaningful protections against adversaries that know about correlations in the profile selection process. For example, consider an adversary with knowledge that profile P_k is always selected immediately after either P_i or P_j are selected. An edge obscuring P_i versus P_j will not prevent the adversary from deducing P_k in the next round. This matches the failure of differential privacy to handle correlations across individuals. The definition also does not compose if the same observation X is reprocessed, as it adds correlations unaccounted for in the privacy analysis. Although such compositions would be valuable, it is less important when the privatization occurs locally at the time of data collection.

Placing these results in the context of reporting resource usage, we can bound the total privacy loss across multiple releases in two cases. Additive composition applies if a single user emits multiple independent measurements and each measurement is separately privatized. When two users independently release measurements, each has no bearing on the other and parallel composition applies. If correlations exist across measurements (or across the selection of users), no compositional result is provided.

7.5 Mechanisms

We now provide mechanisms to implement the profile-based privacy definition. Before getting into specifics, let us first consider the kind of utility goals that we can hope to achieve. We have two primary aspects of the graph G we wish to exploit. First, we wish to preserve any information in the input that does not significantly indicate profile identities. Second, we wish to use the structure of the graph and recognize that some regions of the graph might require less perturbations than others.

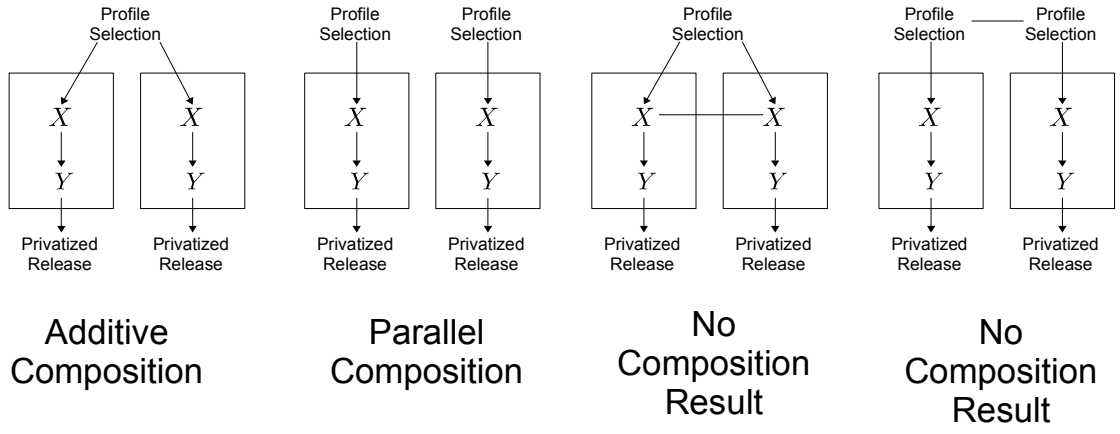

Figure 7.1. Summary of Composition Results. From left to right: independent observation samples with independent mechanism applications compose additively, independent profile selections compose in parallel, dependent observation samples from the same profile do not compose nicely, and dependent profile selections do not compose nicely.

7.5.1 The One-Bit Setting

We begin with a one-bit setting – where the input to the mechanism is a single private bit – and build up to the more general discrete setting.

The simplest case is when we have two profiles i and j represented by Bernoulli distributions P_i and P_j with parameters p_i and p_j respectively. Here, we aim to design a mechanism $\mathscr{A}$ that makes a bit b drawn from P_i or P_j indistinguishable; that is, for any $t \in \{0,1\}$, with $b_i \sim P_i$ and $b_j \sim P_j$,

$$\frac{\Pr(\mathscr{A}(b_i,P_i)=t)}{\Pr(\mathscr{A}(b_j,P_j)=t)} \leq e^{\varepsilon}. \tag{7.4}$$

A plausible mechanism is to draw a bit b' from a Bernoulli distribution that is independent of the original bit b. However, this is not desirable as the output bit would lose any correlation with the input, and any and all information in the bit b would be discarded.

We instead use a mechanism that flips the input bit with some probability $\alpha \leq 1/2$.

Lower values of α improve the correlation between the output and the input. The flip-probability α is obtained by solving the following optimization problem:

$$\begin{aligned} \min \quad & \alpha \\ \text{subject to} \quad & \alpha \geq 0 \\ & \frac{p_i(1-\alpha)+(1-p_i)\alpha}{p_j(1-\alpha)+(1-p_j)\alpha} \in [e^{-\varepsilon}, e^{\varepsilon}] \\ & \frac{(1-p_i)(1-\alpha)+p_i\alpha}{(1-p_j)(1-\alpha)+p_j\alpha} \in [e^{-\varepsilon}, e^{\varepsilon}]. \end{aligned} \tag{7.5}$$

When $p_i = 0$ and $p_j = 1$ (or vice versa), this reduces to the standard randomized response mechanism Warner [1965]; however, α may be lower if p_i and p_j are closer – a situation where our utility is better than local differential privacy's.

Algorithm 6: Single-bit Two-profile Mechanism

Input: Two Bernoulli profiles parameterized by p_1 and p_2, privacy level ε, input profile P_i, input bit x.

Solve the linearly constrained optimization (7.5) to get a flipping probability α.

Sample r as $\begin{cases} x & \text{w.p. } 1-\alpha \\ \neg x & \text{w.p. } \alpha \end{cases}$

return r

The mechanism described above only addresses two profiles. If we have a cluster of profiles representing a connected component of the profile graph, we can compute the necessary flipping probabilities across all edges in the cluster. To satisfy all the privacy constraints, it suffices to always use a flipping probability equal to the largest value required by an edge in the cluster. This results in a naive method we will call the One Bit Cluster mechanism, directly achieves profile-based privacy.

Theorem 23. *The One Bit Cluster mechanism achieves (G, ε)-profile-based privacy.*

The One Bit Cluster mechanism has two limitations. First, it applies only to single bit

Algorithm 7: One Bit Cluster Mechanism

Input: Graph $(\mathscr{P}, E)$ of Bernoulli profiles, privacy level ε, input profile P_i, input bit x.

for *each edge e in the connected component of the graph containing* P_i **do**

  Solve the linearly constrained optimization (7.5) to get a flipping probability α_e for this edge.

Compute $\alpha = \max_e \alpha_e$.

Sample r as $\begin{cases} x & \text{w.p. } 1-\alpha \\ \neg x & \text{w.p. } \alpha \end{cases}$

return r

settings and Bernoulli profiles, and not categorical distributions. Second, by treating all pairs of path-connected profiles similarly, it is overly conservative; when profiles are distant in the graph from a costly edge, it is generally possible to provide privacy with lesser perturbations for these distant profiles.

We address the second drawback while remaining in the one bit setting with the Smooth One Bit mechanism, which uses ideas inspired by the smoothed sensitivity mechanism in differential privacy Nissim et al. [2007]. However, rather than smoothly calibrating the perturbations across the entire space of data sets, a profile-based privacy mechanism needs only to smoothly calibrate over the specified profile graph. This presents a far more tractable task than smoothly handling all possible data sets in differential privacy.

This involves additional optimization variables, $\alpha_1, \ldots, \alpha_k$, for each of the k profiles in G. Thus each profile is permitted its own chance of inverting the released bit. Here, p_i once again refers to the parameter of the Bernoulli distribution P_i. We select our objective function as $\max(\alpha_1, \ldots, \alpha_k)$ in order to uniformly bound the mechanism's chances of inverting or corrupting the input bit. This task remains convex as before, and is still tractably optimized.

$$\begin{aligned} \min_{\alpha_1,\ldots,\alpha_k} \quad & \max(\alpha_1,\ldots,\alpha_k) \\ \text{subject to} \quad & \forall i \in [k]\colon\ \alpha_i \geq 0 \\ & \frac{p_i(1-\alpha_i)+(1-p_i)\alpha_i}{p_j(1-\alpha_j)+(1-p_j)\alpha_j} \in [e^{-\varepsilon}, e^{\varepsilon}] \\ & \frac{(1-p_i)(1-\alpha_i)+p_i\alpha_i}{(1-p_j)(1-\alpha_j)+p_j\alpha_j} \in [e^{-\varepsilon}, e^{\varepsilon}]. \end{aligned} \tag{7.6}$$

Algorithm 8: Smooth One Bit Mechanism

Input: Graph $(\mathscr{P}, E)$ of k Bernoulli profiles, privacy level ε, input profile P_i, input bit x.

Solve the linearly constrained optimization (7.6) to get flipping probabilities $\alpha_1, \ldots, \alpha_k$.

Sample r as $\begin{cases} x & \text{w.p. } 1-\alpha_i \\ \neg x & \text{w.p. } \alpha_i \end{cases}$

return r

Theorem 24. *The Smooth One Bit mechanism achieves (G, ε)-profile-based privacy.*

7.5.2 The Categorical Setting

We now show how to generalize this model into the categorical setting. This involves additional constraints, as well as a (possibly) domain specific objective that maximizes some measure of fidelity between the input and the output.

Specifically, suppose we have k categorical profiles each with d categories; we introduce kd^2 variables to optimize, with each profile receiving a $d \times d$ transition matrix. To keep track of these variables, we introduce the following notation:

- $P_i, \ldots, P_k$: a set of k categorical profiles in d dimensions encoded as a vector.
- $A^1, \ldots, A^k$: A set of d-by-d transition matrix that represents the mechanism's release probabilities for profile i. $A^i_{j,k}$ represents the (j,k)-th element of the matrix A^i.

- $P_i A^i$ represents the d dimensional categorical distribution induced by the transition matrix A^i applied to the distribution P_i.
- In an abuse of notation, $P_i A^i \leq e^{\varepsilon} P_j A^j$ is a constraint that applies element-wise to all components of the resulting vectors on each side.

With this notation, we can express our optimization task:

$$\min_{A^1,\ldots,A^k} \max(\text{off-diagonal entries of } A^1,\ldots,A^k) \tag{7.7}$$

$$\text{subject to } \forall i \in [n] \forall j \in [d] \forall k \in [d]: \quad 0 \leq A^i_{j,k} \leq 1$$

$$\forall i \in [n] \forall j \in [d]: \quad \sum_{k=1}^{d} A^i_{j,k} = 1$$

$$\forall (P_i, P_j) \in E: P_i A^i \leq e^{\varepsilon} P_j A^j, \quad P_j A^j \leq e^{\varepsilon} P_i A^i.$$

Algorithm 9: Smooth Categorical Mechanism

Input: Graph $(\mathscr{P}, E)$ of Categorical profiles, privacy level ε, input profile P_i, input x

Solve the linearly constrained optimization (7.7) to get the transition matrices $A^1,\ldots,A^k$

Sample r according to the discrete distribution given by the xth row of A^i.

return r

To address the tractability of the optimization, we note that each of the privacy constraints are linear constraints over our optimization variables. We further know the feasible solution set is nonempty, as trivial non-informative mechanisms achieve privacy. All that is left is to choose a suitable objective function to make this a readily solved convex problem.

To settle onto an objective will require some domain-specific knowledge of the trade-offs between choosing which profiles and which categories to report more faithfully. Our general choice is a maximum across the off-diagonal elements, which attempts to uniformly minimize

the probability of any data corruptions. This can be further refined in the presence of a prior distribution over profiles, to give more importance to the profiles more likely to be used.

We define the Smooth Categorical mechanism as the process that solves the optimization (7.7) and applies the appropriate transition probabilities on the observed input.

Theorem 25. *The Smooth Categorical mechanism achieves (G,ε)-profile-based privacy.*

7.5.3 Utility Results

The following results present utility bounds which illustrate potential improvements upon local differential privacy; a more detailed numerical simulation is presented in Section 7.6.

Theorem 26. *If $\mathscr{A}$ is a mechanism that preserves ε-local differential privacy, then for any graph G of sensitive profiles, $\mathscr{A}$ also preserves (G,ε)-profile-based differential privacy.*

An immediate result of Theorem 26 is that, in general and for any measure of utility on mechanisms, the profile-based differential privacy framework will never require worse utility than a local differential privacy approach. However, in specific cases, stronger results can be shown.

Observation 10. *Suppose we are in the single-bit setting with two Bernoulli profiles P_i and P_j with parameters p_i and p_j respectively. If $p_i \leq p_j \leq e^{\varepsilon} p_j$, then the solution α to* (7.5) *satisfies*

$$\alpha \leq \max\{0, \frac{p_j - e^{\varepsilon} p_i}{2(p_j - e^{\varepsilon} p_i) - (1 - e^{\varepsilon})}, \frac{p_i - e^{\varepsilon} p_j + e^{\varepsilon} - 1}{2(p_i - e^{\varepsilon} p_j) + e^{\varepsilon} - 1}\}. \tag{7.8}$$

Observe that to attain local differential privacy with parameter ε by a similar bit-flipping mechanism, we need a flipping probability of $\frac{1}{1+e^{\varepsilon}} = \frac{1}{1+(1+e^{\varepsilon}-1)}$, while we get bounds of the form $\frac{1}{1+(1+\frac{e^{\varepsilon}-1}{p_j - e^{\varepsilon} p_i})}$. Thus, profile based privacy does improve over local differential privacy in this simple case. The proof of Observation 10 follows from observing that this value of α satisfies all constraints in the optimization problem (7.5).

7.6 Evaluation

We next evaluate our privacy mechanisms and compare them against each other and the corresponding local differential privacy alternatives. In order to understand the privacy-utility trade-off unconfounded by model specification issues, we consider synthetic data in this paper.

7.6.1 Experimental Setup

We look at three experimental settings – Bernoulli-Couplet, Bernoulli-Chain and Categorical-Chain-3.

Settings.

In Bernoulli-Couplet, the profile graph consists of two nodes connected by a single edge $G = (\mathscr{P} = \{a,b\}, E = \{(a,b)\})$. Additionally, each profile is a Bernoulli distribution with a parameter p.

In Bernoulli-Chain, the profile graph consists of a *chain* of nodes, where successive nodes in the chain are connected by an edge. Each profile is still a Bernoulli distribution with parameter p. We consider two experiments in this category – Bernoulli-Chain-6, where there are six profiles corresponding to six values of p that are uniformly spaced across the interval $[0,1]$, and Bernoulli-Chain-21, where there are 21 profiles corresponding to p uniformly spaced on $[0,1]$.

Finally, in Categorical-Chain, the profile graph comprises of three nodes connected into a chain $P_1 - P_2 - P_3$. Each profile however, corresponds to a 4-dimensional categorical distribution, instead of Bernoulli.

Table 7.1. Categorical-Chain profiles used in our experiments

P_1	0.2	0.3	0.4	0.1
P_2	0.3	0.3	0.3	0.1
P_3	0.4	0.4	0.1	0.1

Baselines.

For Bernoulli-Couplet and Bernoulli-Chain, we use Warner's Randomized Response mechanism Warner [1965] as a local differentially private baseline. For Categorical-Chain, the corresponding baseline is the K-ary version of randomized response.

For Bernoulli-Couplet, we use our Smooth One Bit mechanism to evaluate our framework. For Categorical-Chain, we use the Smooth Categorical mechanism.

7.6.2 Results

Figure 7.2 plots the flipping probability for Bernoulli-Couplet as a function of the difference between profile parameters p. We find that as expected, as the difference between the profile parameters grows, so does the flipping probability and hence the noise added. However, in all cases, this probability stays below the corresponding value for local differential privacy – the horizontal black line – thus showing that profile-based privacy is an improvement.

Figures 7.5-7.8 plot the probability that the output is 1 as a function of ε for each profile in Bernoulli-Chain-6 and Bernoulli-Chain-21. The spread of the profiles for a given ε provides a glimpse into the distortions caused by these methods. The true profiles are uniformly spread on the unit interval, so a spread close to covering the entire interval represents smaller distortion. On the other hand, a small spread, with all profiles having almost identical distributions, represents a heavy distortion of the profiles and poor performance. We find that as expected for low ε, the probability that the output is 1 is close to $1/2$ for both the local differential privacy baseline and our method, whereas for higher ε, it is spread out more evenly, (which indicates higher correlation with the input and higher utility). Additionally, we find that our Smooth One Bit mechanism performs better than the baseline in both cases.

Figures 7.3-7.4 plot the utility across different outputs in the Categorical-Chain setting. We illustrate its behavior through a small setting with 3 profiles, each with 4 categories. We can no longer plot the entirety of these profiles, so at each privacy level we measure the maximum absolute error for each output. Thus, in this setting, each privacy level is associated with 4 costs

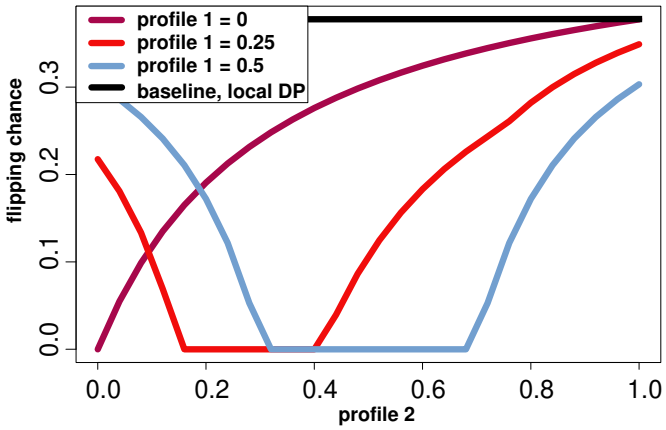

Figure 7.2. Bernoulli-Couplet, Our Method and Baseline.

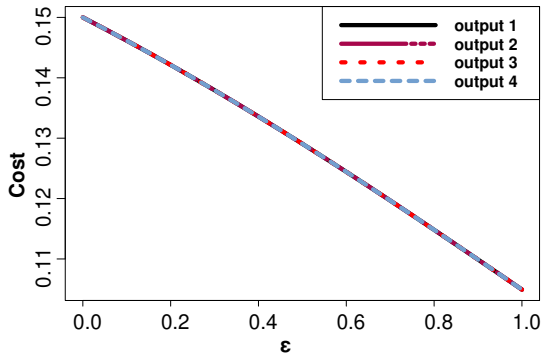

Figure 7.3. Categorical-Chain, Baseline (Local differential privacy). All 4 curves overlap.

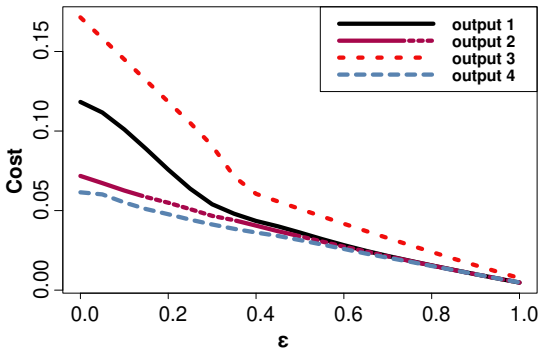

Figure 7.4. Categorical-Chain, Our Method.

of the form given in (7.9). This permits the higher fidelity of profile-irrelevant information to be seen.

$$cost_j = \max_{i\in[n]} |P^i_j - (P^i A^i)_j| \tag{7.9}$$

Our experiments show the categories less associated with the profile identity have lower associated costs than the more informative ones. However, the local differential privacy baseline fails to exploit any of this structure and performs worse.

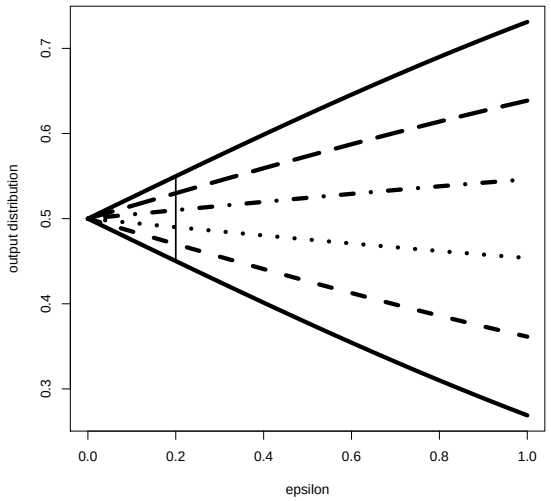

Figure 7.5. Probability that the output is 1 as a function of ε for each profile for Bernoulli-Chain-6, Baseline. A vertical line has been drawn at $\varepsilon = 0.2$ to illustrate the spread.

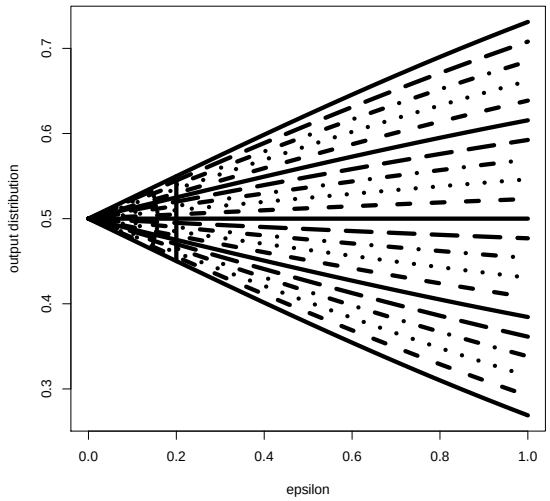

Figure 7.6. Probability that the output is 1 as a function of ε for each profile for Bernoulli-Chain-21, Baseline. A vertical line has been drawn at $\varepsilon = 0.2$ to illustrate the spread.

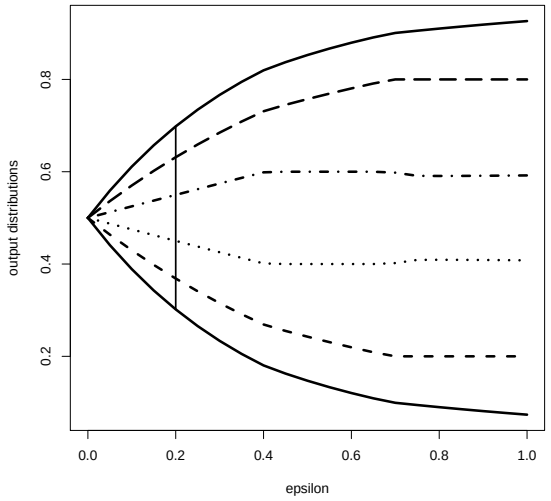

Figure 7.7. Probability that the output is 1 as a function of ε for each profile for Bernoulli-Chain-6, Our Method. A vertical line has been drawn at $\varepsilon = 0.2$ to illustrate the spread.

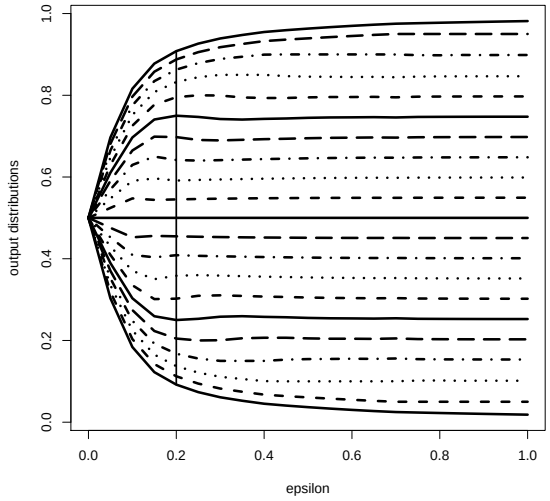

Figure 7.8. Probability that the output is 1 as a function of ε for each profile for Bernoulli-Chain-21, Our Method. A vertical line has been drawn at $\varepsilon = 0.2$ to illustrate the spread.

7.7 Proof of Theorems and Observations

Observation 11. *If a data sample X_i is drawn from profile P_i, and $\mathscr{A}$ preserves (G,ε)-profile-based privacy, then for any (potentially randomized) function F, the release $F(\mathscr{A}(X_i,P_i))$ preserves (G,ε)-profile-based privacy.*

Proof. Let $X_i \sim P_i$ and $X_j \sim P_j$ represent two random variables drawn from the profiles P_i and P_j. We define additional random variables as $Y_i = \mathscr{A}(X_i,P_i)$ and $Z_i = F(\mathscr{A}(X_i,P_i))$, along with the corresponding Y_j and Z_j that use X_j and P_j. This is a result following from a standard data processing inequality.

$$\frac{\mathsf{Pr}(Z_i=z)}{\mathsf{Pr}(Z_j=z)} = \frac{\int_{\mathscr{Y}} \mathsf{Pr}(Z_i=z,Y_i=y)dy}{\int_{\mathscr{Y}} \mathsf{Pr}(Z_j=z,Y_j=y)dy} \tag{7.10}$$

$$= \frac{\int_{\mathscr{Y}} \mathsf{Pr}(Z_i=z|Y_i=y)\mathsf{Pr}(Y_i=y)dY}{\int_{\mathscr{Y}} \mathsf{Pr}(Z_j=z|Y_j=y)\mathsf{Pr}(Y_j=y)dY} \tag{7.11}$$

$$\leq \max_Y \frac{\mathsf{Pr}(Y_i=y)}{\mathsf{Pr}(Y_j=y)} \tag{7.12}$$

$$\leq e^{\varepsilon} \tag{7.13}$$

□

Observation 12. *If two independent samples X_1 and X_2 are drawn from profile P_i, and $\mathscr{A}_1$ preserves (G,ε_1)-profile-based privacy and $\mathscr{A}_2$ preserves (G,ε_2)-profile-based privacy, then the combined release $(\mathscr{A}_1(X_1,P_i),\mathscr{A}_2(X_2,P_i))$ preserves $(G,\varepsilon_1+\varepsilon_2)$-profile-based privacy.*

Proof. The proof of this statement relies on the independence of the two releases $Y_1 = \mathscr{A}_1(X_1,P_i)$ and $Y_2 = \mathscr{A}_2(X_2,P_i)$, given the independence of X_1 and X_2 from the same same profile P_i. Let P_j be another profile such that there is an edge (P_i,P_j) in G. We will introduce X_1' and X_2' as independent samples from the profile P_j, and define $Y_1' = \mathscr{A}_1(X_1',P_j)$ and $Y_2' = \mathscr{A}_2(X_2',P_j)$ By marginalizing over the two independent variables X_1, X_2, we may bound the combined privacy loss. For brevity, we will use $\mathsf{Pr}(X)$ as a shorthand for the density at an arbitrary point $\mathsf{Pr}(X=x)$.

$$\frac{\mathsf{Pr}(Y_1,Y_2)}{\mathsf{Pr}(Y_1',Y_2')} = \frac{\int_{\mathscr{X}}\int_{\mathscr{X}}\mathsf{Pr}(X_1,Y_1,X_2,Y_2)dX_1dX_2}{\int_{\mathscr{X}}\int_{\mathscr{X}}\mathsf{Pr}(X_1',Y_1',X_2',Y_2'|P_j)dX_1dX_2} \tag{7.14}$$

$$= \frac{\int_{\mathscr{X}}\int_{\mathscr{X}}\mathsf{Pr}(X_1,Y_1)\mathsf{Pr}(X_2,Y_2)dX_1dX_2}{\int_{\mathscr{X}}\int_{\mathscr{X}}\mathsf{Pr}(X_1',Y_1')\mathsf{Pr}(X_2',Y_2')dX_1dX_2} \tag{7.15}$$

$$= \frac{\int_{\mathscr{X}}\mathsf{Pr}(X_1,Y_1)dX_1\int_{\mathscr{X}}\mathsf{Pr}(X_2,Y_2)dX_2}{\int_{\mathscr{X}}\mathsf{Pr}(X_1',Y_1')dX_1\int_{\mathscr{X}}\mathsf{Pr}(X_2',Y_2')dX_2} \tag{7.16}$$

$$= \frac{\mathsf{Pr}(Y_1)\mathsf{Pr}(Y_2)}{\mathsf{Pr}(Y_1')\mathsf{Pr}(Y_2')} \tag{7.17}$$

$$\leq e^{\varepsilon_1}e^{\varepsilon_2} \tag{7.18}$$

□

Observation 13. *This proof does not hold if X_1 and X_2 are not independent samples from P_i. This may occur if the same observational data X is privatized twice, or if other correlations exist between X_1 and X_2. We do not provide a composition result for this case.*

Observation 14. *If two profiles P_i and P_j are independently selected, and two observations $X_i \sim P_i$ and $X_j \sim P_j$ are drawn, and $\mathscr{A}_1$ preserves (G,ε_1)-profile-based privacy and $\mathscr{A}_2$ preserves (G,ε_2)-profile-based privacy, then the combined release $(\mathscr{A}_1(X_i,P_i),\mathscr{A}_2(X_j,P_j))$ preserves $(G,max\{\varepsilon_1,\varepsilon_2\})$-profile-based privacy.*

Proof. For the purposes of this setting, let Q_1 and Q_2 be two random variables representing the choice of profile in the first and second selections, with the random variables $X_1 \sim Q_1$ and $X_2 \sim Q_2$.

Since the two profiles and their observations are independent, the two releases $Y_1 = \mathscr{A}_1(X_1,Q_1)$ and $Y_2 = \mathscr{A}_2(X_2,Q_2)$ contain no information about each other. That is, $\mathsf{Pr}(Y_1 = y_1|Q_1 = P_i,Q_2 = P_j,Y_2 = y_2) = \mathsf{Pr}(Y_i|Q_1 = P_i)$. Similarly we have $\mathsf{Pr}(Y_2 = y_2|Q_1 = P_i,Y_1 = y_1,Q_2 = P_j) = \mathsf{Pr}(Y_2 = y_2)$.

Let P_h and P_k be profiles such that the edges (P_h,P_i) and (P_j,P_k) are in G.

$$\frac{\Pr(Y_1,Y_2|Q_1=P_i)}{\Pr(Y_1,Y_2|Q_1=P_h)} = \frac{\sum_{Q_2}\Pr(Y_1,Y_2,Q_2|Q_1=P_i)}{\sum_{Q_2}\Pr(Y_1,Y_2,Q_2|Q_1=P_h)} \tag{7.19}$$

$$= \frac{\sum_{Q_2}\Pr(Y_1|Q_1=P_i,Y_2,Q_2)\Pr(Y_2,Q_2|Q_1=P_i)}{\sum_{Q_2}\Pr(Y_1|Q_1=P_h,Y_2,Q_2)\Pr(Y_2,Q_2|Q_1=P_h)} \tag{7.20}$$

$$= \frac{\sum_{Q_2}\Pr(Y_1|Q_1=P_i)\Pr(Y_2,Q_2)}{\sum_{Q_2}\Pr(Y_1|Q_1=P_h)\Pr(Y_2,Q_2)} \tag{7.21}$$

$$= \frac{\Pr(Y_1|Q_1=P_i)}{\Pr(Y_1|Q_1=P_h)} \cdot \frac{\sum_{Q_2}\Pr(Y_2,Q_2)}{\sum_{Q_2}\Pr(Y_2,Q_2)} \tag{7.22}$$

$$= \frac{\Pr(Y_1|Q_1=P_i)}{\Pr(Y_1|Q_1=P_h)} \tag{7.23}$$

$$\leq e^{\varepsilon_1} \tag{7.24}$$

A similar derivation conditioning on $P_2 = P_j$ results in a ratio bounded by e^{ε_2} over the edge (P_j, P_k). Thus to get a single bound for the combined release (Y_1, Y_2) over the edges of the graph G, we take the maximum $e^{\max\{\varepsilon_1,\varepsilon_2\}}$.

□

Observation 15. *This proof does not hold if the profile selection process is not independent. We do not provide a composition result for this case.*

Theorem 22. *The One Bit Cluster mechanism achieves ε-profile based privacy.*

Proof. By direct construction, it is known that the flipping probabilities generated for single edges α_e will satisfy the privacy constraints. What remains to be shown for the privacy analysis is that taking $\alpha = \max_e \alpha_e$ will satisfy the privacy constraints for all the edges simultaneously.

To show this, we will demonstrate a monotonicity property: if a flipping probability $\alpha < 0.5$ guarantees a certain privacy level, then so too do all the probabilities in the interval $(\alpha, 0.5)$. By taking the maximum across all edges, this algorithm exploits the monotonicity to ensure all the constraints are met simultaneously.

Let $x_1 \sim P_1$ and $x_2 \sim P_2$, and let p_i and p_j denote the parameters of these two Bernoulli

distributions. When computing the privacy level, we have two output values and thus two output ratios to consider:

$$\left|\log \frac{\mathsf{Pr}(\mathscr{A}(x_1,P_1)=1)}{\mathsf{Pr}(\mathscr{A}(x_2,P_2)=1)}\right| = \left|\log \frac{p_1\cdot(1-\alpha)+(1-p_1)\cdot\alpha}{p_2\cdot(1-\alpha)+(1-p_2)\cdot\alpha}\right| \tag{7.25}$$

$$\left|\log \frac{\mathsf{Pr}(\mathscr{A}(x_1,P_1)=0)}{\mathsf{Pr}(\mathscr{A}(x_2,P_2)=0)}\right| = \left|\log \frac{p_1\cdot\alpha+(1-p_1)\cdot(1-\alpha)}{p_2\cdot\alpha+(1-p_2)\cdot(1-\alpha)}\right| \tag{7.26}$$

Without loss of generality, assume $p_1 > p_2$. (If they are equal, then all possible privacy levels are achieved trivially.) This means following two quantities are positive and equal to the absolute values above when $\alpha < 0.5$.

$$\log \frac{\mathsf{Pr}(\mathscr{A}(x_1,P_1)=1)}{\mathsf{Pr}(\mathscr{A}(x_2,P_2)=1)} = \log \frac{p_1\cdot(1-\alpha)+(1-p_1)\cdot\alpha}{p_2\cdot(1-\alpha)+(1-p_2)\cdot\alpha} \tag{7.27}$$

$$\log \frac{\mathsf{Pr}(\mathscr{A}(x_1,P_1)=0)}{\mathsf{Pr}(\mathscr{A}(x_2,P_2)=0)} = \log \frac{p_2\cdot\alpha+(1-p_2)\cdot(1-\alpha)}{p_1\cdot\alpha+(1-p_1)\cdot(1-\alpha)} \tag{7.28}$$

Our next task is to show that these quantities reveal monotonic increases in privacy levels as α increases up to 0.5. Taking just the $\mathscr{A}(x_1,P_1)=1$ term for now, we compute the derivatives.

$$\frac{\partial}{\partial\alpha}\left[\log \frac{p_1\cdot(1-\alpha)+(1-p_1)\cdot\alpha}{p_2\cdot(1-\alpha)+(1-p_2)\cdot\alpha}\right] = \frac{1-2p_1}{p_1\cdot(1-\alpha)+(1-p_1)\cdot\alpha} - \frac{1-2p_2}{p_2\cdot(1-\alpha)+(1-p_2)\cdot\alpha} \tag{7.29}$$

$$= \frac{1-2p_1}{p_1+(1-2p_1)\alpha} - \frac{1-2p_2}{p_2+(1-2p_2)\alpha} \tag{7.30}$$

$$= \frac{(1-2p_1)(p_2\cdot(1-\alpha)+(1-p_2)\cdot\alpha)-(1-2p_2)(p_1\cdot(1-\alpha)+(1-p_1)\cdot\alpha)}{(p_1\cdot(1-\alpha)+(1-p_1)\cdot\alpha)(p_2\cdot(1-\alpha)+(1-p_2)\cdot\alpha)} \tag{7.31}$$

$$= \frac{p_2-p_1}{\mathsf{Pr}(\mathscr{A}(x_1,P_1)=1)\mathsf{Pr}(\mathscr{A}(x_2,P_2)=1)} \leq 0 \tag{7.32}$$

The final inequality arises from our assumption that $p_1 > p_2$. A similar computation on the $\mathscr{A}(x_1,P_1)=0$ term also finds that the derivative is always negative.

This monotonicity implies that increasing α (up to 0.5 at most) only decreases the probability ratios. In the limit when $\alpha = 0.5$, the ratios are precisely 1 (and the logarithm is 0). Thus if $\alpha < 0.5$ achieves a certain privacy level, all α' satisfying $\alpha < \alpha' < 0.5$ achieve a privacy level at least as strong.

The One Bit Cluster mechanism takes the maximum across all edges, ensuring the final flipping probability is no less than the value needed by each edge to achieve probability ratios within $e^{\pm\varepsilon}$. Therefore each edge constraint is satisfied by the final choice of flipping probability, and the mechanism satisfies the privacy requirements.

□

Theorem 27. *The Smooth One Bit mechanism achieves (G,ε)-profile-based privacy.*

Theorem 28. *The Smooth Categorical mechanism achieves (G,ε)-profile-based privacy.*

The Smooth One Bit mechanism and Smooth Categorical mechanism satisfy a privacy analysis directly from the constraints of the optimization problem. These optimizations are done without needing any access to a sensitive observation, and as such pose no privacy risk. Implicitly, the solution to the optimization problem is verified to satisfy the constraints before being used.

Theorem 29. *If $\mathscr{A}$ is a mechanism that preserves ε-local differential privacy, then for any graph G of sensitive profiles, $\mathscr{A}$ also preserves (G,ε)-profile-based differential privacy.*

Proof. The proof of this theorem lies in that the strong protections given by local differential differential privacy to the observed data also extend to protecting the profile identities. Let $Y_i = \mathscr{A}(X_i, P_i)$, the output of a locally differentially private algorithm $\mathscr{A}$ that protects any two distinct data observations x and x'. As local differential privacy mechanisms do not use profile information, the distribution of Y_i depends only on X_i and ignores P_i. To prove the generality of this analysis over any graph G, we will show the privacy constraint is satisfied for any possible edge (P_i, P_j) of two arbitrary profiles.

$$\frac{\mathsf{Pr}(Y_i = y)}{\mathsf{Pr}(Y_j = y)} = \frac{\int_{\mathscr{X}} \mathsf{Pr}(Y_i = y | X_i = x)\mathsf{Pr}(X_i = x)dx}{\int_{\mathscr{X}} \mathsf{Pr}(Y_i = y | X_j = x)\mathsf{Pr}(X_j = x)dx} \tag{7.33}$$

$$\leq \frac{\sup_x \mathsf{Pr}(Y_i = y | X_i = x)}{\inf_X \mathsf{Pr}(Y_j = y | X_j = x)} \tag{7.34}$$

$$\leq e^{\varepsilon} \tag{7.35}$$

If the final inequality did not hold, one would be able to find two values X and X' such that the output Y violates the local differential privacy constraint, which contradicts our assumption on $\mathscr{A}$.

□

Observation 16. *Suppose we are in the single-bit setting with two Bernoulli profiles P_i and P_j with parameters p_i and p_j respectively. If $p_i \leq p_j \leq e^{\varepsilon} p_j$, then the solution α to* (7.5) *satisfies*

$$\alpha \leq \max\{0, \frac{p_j - e^{\varepsilon} p_i}{2(p_j - e^{\varepsilon} p_i) - (1 - e^{\varepsilon})}, \frac{p_i - e^{\varepsilon} p_j + e^{\varepsilon} - 1}{2(p_i - e^{\varepsilon} p_j) + e^{\varepsilon} - 1}\}. \tag{7.36}$$

Proof. Direct computation shows the desired constraints are met with this value for α.

$$\begin{aligned} \min \quad & \alpha \\ \text{subject to} \quad & \alpha \geq 0 \\ & \frac{p_i(1-\alpha) + (1-p_i)\alpha}{p_j(1-\alpha) + (1-p_j)\alpha} \in [e^{-\varepsilon}, e^{\varepsilon}] \\ & \frac{(1-p_i)(1-\alpha) + p_i\alpha}{(1-p_j)(1-\alpha) + p_j\alpha} \in [e^{-\varepsilon}, e^{\varepsilon}]. \end{aligned} \tag{7.37}$$

First, we note that by our assumption $p_i \leq p_j$ and $\varepsilon \geq 0$, we immediately have two of our constraints trivially satisfied given $\alpha \leq 0.5$, since $p_i(1-\alpha) + (1-p_i)\alpha \leq p_j(1-\alpha) + (1-p_j)\alpha$ and $(1-p_i)(1-\alpha) + p_i\alpha \geq (1-p_j)(1-\alpha) + p_j\alpha$.

Two constraints of interest remain:

$$\frac{p_i(1-\alpha)+(1-p_i)\alpha}{p_j(1-\alpha)+(1-p_j)\alpha} \geq e^{-\varepsilon} \tag{7.38}$$

$$\frac{(1-p_i)(1-\alpha)+p_i\alpha}{(1-p_j)(1-\alpha)+p_j\alpha} \leq e^{\varepsilon}. \tag{7.39}$$

We know that these ratios are monotonic in α, so to solve these inequalities, it suffices to find the values of α where we have equality on these two constraints. Any values of α larger than this (and less than $1/2$) will therefore satisfy the inequality.

For (7.38), we get $\alpha = \frac{p_j - e^{\varepsilon} p_i}{2(p_j - e^{\varepsilon} p_i) - (1 - e^{\varepsilon})}$. Solving (7.39) instead, we get $\alpha = \frac{p_i - e^{\varepsilon} p_j + e^{\varepsilon} - 1}{2(p_i - e^{\varepsilon} p_j) + e^{\varepsilon} - 1}$.

Since both constraints must be satisfied simultaneously, we can complete our statement by taking the maximum of the two points given by our constraints, along with knowing $\alpha \geq 0$.

□

7.8 Conclusion

In conclusion, we provide a novel definition of local privacy – profile based privacy – that can achieve better utility than local differential privacy. We prove properties of this privacy definition, and provide mechanisms for two discrete settings. Simulations show that our mechanisms offer superior privacy-utility trade-offs than standard local differential privacy.

7.9 Acknowledgements

We thank ONR under N00014-16-1-261, UC Lab Fees under LFR 18-548554 and NSF under 1804829 for research support. We also thank Takao Murakami for pointing us to Kawamoto and Murakami [2018] and discussions about Observation 9.

This chapter is based on the material in IEEE International Symposium on Information Theory 2019 (Joseph Geumlek, and Kamalika Chaudhuri. ”Profile-based privacy for locally private computations”). The dissertation author was the primary investigator and author of this material.

Bibliography

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 6280–6290, 2018.

Borja Balle, James Bell, Adrià Gascón, and Kobbi Nissim. The privacy blanket of the shuffle model. pages 638–667, 2019.

Gilles Barthe and Federico Olmedo. Beyond differential privacy: Composition theorems and relational logic for f-divergences between probabilistic programs. In *International Colloquium on Automata, Languages, and Programming*, pages 49–60. Springer, 2013.

Raef Bassily and Yoav Freund. Typical stability. *arXiv preprint arXiv:1604.03336*, 2016.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of private learners. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 97–110. ACM, 2013.

Amos Beimel, Hai Brenner, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. *Machine learning*, 94(3): 401–437, 2014.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Lawrence D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-Monograph Series*, 9:i–279, 1986. ISSN 07492170. URL http://www.jstor.org/stable/4355554.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

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

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 634–649. IEEE, 2015.

Kamalika Chaudhuri and Nina Mishra. When random sampling preserves privacy. In *Annual International Cryptology Conference*, pages 198–213. Springer, 2006.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12:1069–1109, July 2011. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1953048.2021036.

Kamalika Chaudhuri, Anand Sarwate, and Kaushik Sinha. Near-optimal differentially private principal components. In *Advances in Neural Information Processing Systems*, pages 989–997, 2012.

Kamalika Chaudhuri, Daniel Hsu, and Shuang Song. The large margin mechanism for differentially private maximization. In *Neural Inf. Processing Systems*, 2014.

Xinjia Chen. A new generalization of Chebyshev inequality for random vectors. *arXiv preprint arXiv:0707.0805*, 2007.

Albert Cheu, Adam D. Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In *Advances in Cryptology - EUROCRYPT 2019 - 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, May 19-23, 2019, Proceedings, Part I*, pages 375–403, 2019.

Joel E Cohen, Yoh Iwasa, Gh Rautu, Mary Beth Ruskai, Eugene Seneta, and Gh Zbaganu. Relative entropy under mappings by stochastic matrices. *Linear algebra and its applications*, 179:211–235, 1993.

Jon P Daries, Justin Reich, Jim Waldo, Elise M Young, Jonathan Whittinghill, Andrew Dean Ho, Daniel Thomas Seaton, and Isaac Chuang. Privacy, anonymity, and big data in the social sciences. *Communications of the ACM*, 57(9):56–63, 2014.

P Del Moral, M Ledoux, and L Miclo. On contraction properties of Markov kernels. *Probability theory and related fields*, 126(3):395–420, 2003.

Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin IP Rubinstein. Robust and private Bayesian inference. In *Algorithmic Learning Theory (ALT)*, pages 291–305. Springer, 2014.

Roland L Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability & Its Applications*, 1(1):65–80, 1956.

W. Doeblin. Sur les proprietes asymptotiques de mouvements rÉgis par certains types de chaÎnes simples (suite et fin). *Bulletin mathématique de la Société Roumaine des Sciences*, 39(2):3–61, 1937. ISSN 12203858. URL http://www.jstor.org/stable/43769812.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 429–438. IEEE, 2013.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380. ACM, 2009.

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

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

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006a.

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

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *The 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 51–60, 2010.

Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019.

Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 521–532. IEEE, 2018.

James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. In *UAI*, 2016.

Johannes Gehrke, Edward Lui, and Rafael Pass. Towards privacy for social networks: A zero-

knowledge based definition of privacy. In *Theory of cryptography conference*, pages 432–449. Springer, 2011.

Sharon Goldwater and Tom Griffiths. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 744–751, 2007.

Xi He, Ashwin Machanavajjhala, and Bolin Ding. Blowfish privacy: Tuning privacy-utility trade-offs using policies. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1447–1458, 2014.

Zhiyi Huang and Sampath Kannan. The exponential mechanism for social welfare: Private, truthful, and nearly optimal. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 140–149. IEEE, 2012.

Dirk Husmeier, Richard Dybowski, and Stephen Roberts. *Probabilistic modeling in bioinformatics and medical informatics*. Springer Science & Business Media, 2006.

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

R.E. Kass, L. Tierney, and J.B. Kadane. The validity of posterior expansions based on laplace's method. *Bayesian and likelihood methods in statistics and econometrics*, 7:473, 1990.

Yusuke Kawamoto and Takao Murakami. Differentially private obfuscation mechanisms for hiding probability distributions. *arXiv preprint arXiv:1812.00939*, 2018.

Daniel Kifer and Ashwin Machanavajjhala. A rigorous and customizable framework for privacy. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*, pages 77–88. ACM, 2012.

Daniel Kifer, Adam Smith, Abhradeep Thakurta, Shie Mannor, Nathan Srebro, and Robert C. Williamson. Private convex empirical risk minimization and high-dimensional regression. In *In COLT*, pages 94–103, 2012.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Ninghui Li, Wahbeh Qardaji, and Dong Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, pages 32–33. ACM, 2012.

Jiachun Liao, Lalitha Sankar, Flavio P Calmon, and Vincent YF Tan. Hypothesis testing under maximal leakage privacy constraints. In *Information Theory (ISIT), 2017 IEEE International*

Symposium on, pages 779–783. IEEE, 2017.

Friedrich Liese and Klaus-J Miescke. Statistical decision theory. In *Statistical Decision Theory*, pages 1–52. Springer, 2007.

Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 277–286. IEEE, 2008.

Frank McSherry. How many secrets do you have? https://github.com/frankmcsherry/blog/blob/master/posts/2017-02-08.md, 2017.

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

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

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

Ilya Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium, CSF 2017, Santa Barbara, CA, USA, August 21-25, 2017*, pages 263–275, 2017.

Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*, pages 75–84, 2007.

Esa Nummelin. *General irreducible Markov chains and non-negative operators*, volume 83. Cambridge University Press, 2004.

C Piech, J Huang, Z Chen, C Do, A Ng, and D Koller. Tuned models of peer assessment in MOOCs. In *Proceedings of the 6th International Conference on Educational Data Mining*, pages 153–160, 2013.

Maxim Raginsky. Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016.

Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 880–887, 2008.

Anand D Sarwate and Kamalika Chaudhuri. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *IEEE signal processing magazine*, 30(5):86–94, 2013.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 245–248. IEEE, 2013.

Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. Pufferfish privacy mechanisms for correlated data. In *SIGMOD*, 2017.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Abhradeep Guha Thakurta and Adam Smith. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, pages 819–850, 2013.

Laurens van der Maaten, Minmin Chen, Stephen Tyree, and Kilian Q. Weinberger. Learning with marginalized corrupted features. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 410–418, 2013.

Yining Wang, Yu-Xiang Wang, and Aarti Singh. Differentially private subspace clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1000–1008, 2015a.

Yu-Xiang Wang, Stephen E Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient Monte Carlo. *Proceedings of The 32nd International Conference on Machine Learning (ICML)*, pages 2493—2502, 2015b.

Yu-Xiang Wang, Jing Lei, and Stephen E Fienberg. On-average kl-privacy and its equivalence to generalization for max-entropy mechanisms. In *International Conference on Privacy in Statistical Databases*, pages 121–134. Springer, 2016.

Yu-Xiang Wang, Borja Balle, and Shiva Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 681–688, 2011.

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