The Differential Privacy of Bayesian Inference

The Differential Privacy of Bayesian Inference

Shijie (Joy) Zheng

An undergraduate thesis submitted in partial fulfillment for the degree of Bachelor of Arts in
Computer Science and Mathematics

Harvard College
Cambridge, Massachusetts
April 1, 2015

Abstract

Differential privacy is one recent framework for analyzing and quantifying the amount of privacy lost when data is released. Meanwhile, multiple imputation is an existing Bayesian-inference based technique from statistics that learns a model using real data, then releases synthetic data by drawing from that model. Because multiple imputation does not directly release any real data, it is generally believed to protect privacy.

In this thesis, we examine that claim. While there exist newer synthetic data algorithms specifically designed to provide differential privacy, we evaluate whether multiple imputation already includes differential privacy for free. Thus, we focus on several method variants for releasing the learned model and releasing the synthetic data, and how these methods perform for models taking on two common distributions: the Bernoulli and the Gaussian with known variance.

We prove a number of new or improved bounds on the amount of privacy afforded by multiple imputation for these distributions. We find that while differential privacy is ostensibly achievable for most of our method variants, the conditions needed for it to do so are often not realistic for practical usage. At least in theory, this is particularly true if we want absolute privacy (ε-differential privacy), but that the methods are more practically compatible with privacy when we allow a small probability of a catastrophic data leakage ((ε, δ)-differential privacy).

Acknowledgments

First and foremost, I would like to thank my thesis advisor Salil Vadhan for his encouragement, guidance, and feedback, without which this thesis would not have been possible. I would also like to thank my academic advisors Harry Lewis and Michael Mitzenmacher for their support.

I would like to thank Edo Airoldi and Kobbi Nissim for serving as thesis readers and for helpful discussions. I would also like to thank Or Sheffet, James Honaker, and Rich Zemel for helpful feedback and conversations, as well as the members of the Harvard Privacy Tools for Sharing Research Data Project.

As always, I would like to thank my parents for their love and support. I would like to thank my friends Spencer Kwon, Allen Yuan, and Aleksandar Makelov reading through and providing feedback on earlier drafts of this thesis. I would also like to thank my blockmates and friends who have defined so much of my college experience, and who have patiently listened to many random comments on probability injected into utterly unrelated conversations.

Contents

Chapter 1: Introduction

1.1 Motivation

Think about the amount of data now available about us online, both public and not. On the one hand, increasing digitalization facilitates both data collection and data sharing, opening new avenues for research and collaboration. On the other hand, as the quantity of available data increases, so do concerns about data privacy and security.

While data holders – researchers, governments, corporations, or individuals – might like to make their datasets available for others to examine, it is difficult for them to do so without leaking more information than was intended. Take one recent case from 2014:

Example 1.1 ([Toc14])**.** In response to a Freedom of Information Law request, the New York City Taxi and Limousine Commission released a database of all taxi rides in the city during 2013. This included information such as pickup and drop-off locations and times, fare amounts, and an (anonymized) identifier of the specific cabs involved.

Yet, researchers soon discovered that rides could be traced back to individuals in several ways, including:

- Time-stamped photos of celebrities with the cab number visible could be matched to rides using those two attributes,[1] thus releasing the destination of the ride and the amount tipped.
- Because the pickup data included very precise GPS coordinates, these could be linked to addresses. When the pickup address had only one inhabitant (as determined by a Google search), the identity of the taxi rider could thus be guessed with reasonably good probability.

□

As this shows, we have a difficult time understanding what constitutes a "safe" data release. Thus, the question:

When does a data release protect the privacy of the included individuals?

One recent framework to emerge for answering this question is the notion of *differential privacy* [DMNS06]. Differential privacy is a worst-case guarantee: it tries to ensure that no individual's data can be deduced from the information released, even by an adversary who knows the information of everybody else in the database. Differentially private algorithms are randomized algorithms: an algorithm is differentially private when its output distribution does not change too much when a single individual's data changes, making it difficult to identify the data for that individual.

In its definition, differential privacy uses two parameters – ε and δ – to describe, respectively, the degree of privacy obtained and the maximum probability of a catastrophic data leakage. When $\delta = 0$ – i.e. there is

[1] While the cab number was originally anonymized with an MD5 hash, this was quickly deanonymized using a combination of brute force given that cab numbers only take on a limited number of formats.

no possibility allowed for a catastrophic data release – we refer to this as ε-differential privacy; otherwise, we refer to it as (ε, δ)-differential privacy. As the values of these parameters increase, the likelihood that an adversary can guess an individual's data increases and the amount of privacy obtained decreases. As a result, differential privacy gives not only a binary yes-no answer to the question, "Does this algorithm preserve privacy?", but also a way to quantify the amount of privacy lost from a data release.

The natural next question is then:

> **How can we release data with differential privacy?**

In the differential privacy literature, a number of such algorithms have been developed, many of which add noise to the output of an otherwise non-private algorithm. In this vein, differentially private algorithms tend to take advantage of two properties:

1. The more random noise is added to the output of an algorithm, the more privacy is provided (but also the more accuracy is compromised).

2. The larger the dataset, the less noise is required to achieve the same values of ε and δ.

However, the use of randomness and dataset size in order to protect the privacy of individuals is hardly unique to differential privacy. Indeed, statisticians have been wrestling with the problem of how to privately release data for many years prior to the development of differential privacy; in statistics, this area is known as *statistical disclosure limitation* (SDL). Techniques in this area include data suppression (removing fields from a database), data perturbation (altering fields in a database), and synthetic data generation (replacing real data with data generated from a learned model).

Such techniques are believed to offer privacy, although rigorous analysis of the amount of privacy provided has been difficult without a good definition of privacy with which to work. Yet, because statistical disclosure algorithms using Bayesian inference often include randomness when handling probability distributions – mirroring the ways in which differentially private techniques do so – it is not unreasonable to believe that they may also offer differential privacy. This leads us to the primary focus of this thesis:

> **Do we obtain differential privacy for "free" when generating synthetic data using Bayesian inference?**

Roughly speaking, such synthetic data generation involves two steps:

1. Using the real dataset, learn (the parameters for) a model of the data.

2. Using the learned model, draw synthetic data points from its probability distribution and release these in place of the real data.

Compared to more recent algorithms specifically developed for differential privacy, Bayesian synthetic data generation has a number of benefits. First, in dealing with well-known probability distributions, it is generally well studied and understood; there consequently exist implementations of the distributions in statistical software. Additionally, there is a large swath of statistical research on both the accuracy of the synthetic data and how best to interpret it. This particular combination of the data generation and how to develop inferences from it is known as *multiple imputation.*

1.2 Outline

We will begin with a review of the relevant definitions and background from both the differential privacy and statistics literature.

In **Chapter 2**, we introduce differential privacy and its important properties. We also include several basic differentially private algorithms that are used as building blocks later in our analysis.

In **Chapter 3**, we flesh out the specific statistical techniques and distributions to be examined. Because we want to understand how both the model learning and data generation steps contribute to privacy, we separate them out into method variants that release model parameters ($\textsc{Param}_{\mathrm{MAP}}$ and $\textsc{Param}_{\mathrm{Post}}$) and those that release synthetic data ($\textsc{Synth}_{\mathrm{MAP}}$, $\textsc{Synth}_{\mathrm{Post\ (One)}}$, and $\textsc{Synth}_{\mathrm{Post\ (Many)}}$). We then detail the two distributions – the Bernoulli and Gaussian – to be analyzed in detail later. Finally, we situate multiple imputation in the context of preceding SDL techniques by briefly examining the benefits and shortcomings of both.

In **Chapter 4**, we survey the existing literature on the intersection of Bayesian inference and differential privacy. Here, we see that it is possible to release information similar to that of our five methods, but in a manner that is known to preserve differential privacy, using algorithms designed for differential privacy. Then, turning to multiple imputation, we look at prior bounds on the amount of differential privacy given by synthetic data generation.

Starting in Chapter 5, we present our own results.

In **Chapter 5**, we compare our parameter release and synthetic data release methods. We hypothesize that as the number of synthetic data points generated grows to infinity, some pairs of these methods (specifically $\textsc{Param}_{\mathrm{MAP}}/\textsc{Synth}_{\mathrm{MAP}}$ and $\textsc{Param}_{\mathrm{Post}}/\textsc{Synth}_{\mathrm{Post\ (One)}}$) are equivalent in the amount of privacy loss.

Afterwards, we move on to specific distributions, proving a number of new upper and lower bounds on amount of differential privacy achieved.

In **Chapter 6**, we look at the Bernoulli distribution. Here, the methods learn based on a combination of publicly known prior data and the real database, so noise comes in the form of prior data; more prior data helps camouflage the real data, but also means that we learn less from it. Our primary results are:

1. Neither of our parameter release methods ($\textsc{Param}_{\mathrm{MAP}}$ and $\textsc{Param}_{\mathrm{Post}}$) satisfies ε-differential privacy for any value of ε, but $\textsc{Param}_{\mathrm{Post}}$ satisfies (ε, δ)-differential privacy with a sublinear amount of prior data. We also show that our bound is the asymptotically minimal amount of prior data required to obtain (ε, δ)-differential privacy.

2. All of our synthetic data release methods ($\textsc{Synth}_{\mathrm{MAP}}$, $\textsc{Synth}_{\mathrm{Post\ (One)}}$, and $\textsc{Synth}_{\mathrm{Post\ (Many)}}$) can satisfy ε-differential privacy and (ε, δ)-differential privacy. In the case of ε-differential privacy, the amount of prior data required is linear in the number of synthetic data points, which makes this unsuitable for generating a large quantity of synthetic data. On the other hand, we can obtain (ε, δ)-differential privacy with a sublinear amount of prior data even when generating a synthetic database of size linear in the original database size.

3. Even looking at an average-case analog of differential privacy (which, as we mentioned, is a worst-case guarantee), we do not obtain privacy for methods in which it was previously unachievable.

4. Some of our synthetic data release methods can be interpreted as instantiations of the exponential mechanism, a basic differentially private algorithm. Indeed, while we may not have known how to efficiently compute the output of the exponential mechanism, knowing the statistical interpretation aids us in doing so for this specific case. Additionally, for $\textsc{Synth}_{\mathrm{MAP}}$ and $\textsc{Synth}_{\mathrm{Post\ (Many)}}$, this allows us to generalize our ε-differential privacy result from the Bernoulli distribution to a categorical distribution.

In **Chapter 7**, we turn to the Gaussian distribution with known variance. In this case, we show that if the data is allowed to be unbounded in $\mathbb{R}$, then none of our methods provide differential privacy. However, if we instead restrict the right combination of our data points, parameter choices, and synthetic data to a finite range $[-R, R]$, we find that:

1. $\textsc{Param}_{\mathrm{MAP}}$ is not differentially private in any sense.

2. It is possible to obtain both ε-differential privacy and (ε, δ)-differential privacy for $\textsc{Param}_{\text{Post}}$. However, obtaining ε-differential privacy requires restricting the data to an unrealistically small range, and hence is probably not applicable in practice.

3. Similarly, all of our synthetic generation methods are able to provide both ε-differential privacy and (ε, δ)-differential privacy. For a synthetic dataset the same size as the original, this again results in unrealistically small values of R in order to achieve ε-differential privacy; because we show that most our ε-differential privacy bounds are asymptotically optimal, this problem cannot be avoided. However, this is no longer a problem for (ε, δ)-differential privacy.

Finally, in **Chapter 8** we conclude with a few notes comparing the results obtained from previous chapters and commenting on possible future work. In general, for these two distributions examined, we find that synthetic data generation provides somewhat more privacy than does parameter generation, including privacy that is wholly unavailable from the parameter generation step alone. However, when we want to generate larger quantities of synthetic data – i.e. linear in the size of the original dataset – it becomes less clear which step is actually more essential for privacy protection.

Chapter 2: Differential Privacy

Why Differential Privacy?

Former methods of privacy preservation have proved insufficient.

In practice, a common such method is *anonymization* – the removal or obfuscation of "identifying" fields – after which the remainder of the database is released. These fields vary from database to database, but generally encompass relatively unique identifiers such as name, address, or social security number; instead of being deleted wholesale, some of these fields may instead be replaced by rough categorizations – for example, address by zip code. While the suppression of these fields is clearly necessary to protect privacy, it is oftentimes treated instead as a sufficient step:

Example 2.1 ([oHS]). The Health Insurance Portability and Accountability Act (HIPAA) provides a number of regulations on the usage and disclosure of personal health information that can be traced back to an individual. One of the standards by which data can be considered deidentified (and thus available for broader use) is the Safe Harbor standard, which specifies the removal of 18 identifying fields, including names, license plate numbers, and photographs. □

However, removing the obviously identifying fields in a database often fails to prevent reidentification of those whose data is contained within it. For example, Sweeney found that 87% of the U.S. population can be uniquely identified by a trio of zip code, gender, and birthday – fields constituting sufficiently broad categories that they are often not deleted prior to database release [Swe00].

Alternatively, reidentification can be performed by correlating otherwise innocuous information in a database with publicly available information.

Example 2.2 ([NS08]). After Netflix released an anonymized database of movie rentals – including user IDs, rental dates, and movie ratings – for the Netflix Prize, researchers identified some of the users by matching movie rental times with movie review times on IMDB, a public movie-centric website that displays user reviews.[1] □

Due to the quantity of information available online, it is increasingly difficult to make any types of guarantees about what does and does not constitute "public" information. Thus, a framework for evaluating the privacy of data releases without relying on a sparsity of outside information is needed: this is where differential privacy comes in.

We need a way to quantify data loss, especially over *multiple* data releases.

The previous example shows that privacy problems may often arise not just from the release of one database, but by correlations between multiple data releases. As we will see, differential privacy's properties enable us

[1] See also [CKN$^+$11] for an example of how researchers were able to learn about individual purchasing patterns using Amazon recommendations, without even a specific "database" release.

to measure the values of its parameters ε and δ over multiple such releases.

In this chapter, we outline the definition of differential privacy, along with a few of its basic properties and differentially private mechanisms that will be relevant for later analysis; an in-depth treatment can be found in the monograph by Dwork and Roth [DR13]. Proofs for some of the propositions in this chapter may also be found in the Appendix.

2.1 Definitions

Let $\mathcal{X}$ be a universe of possible database rows, where each row corresponds to the data of a single individual. We consider databases $X \in \mathcal{X}^n$ of size n under the assumption that n is publicly known.

As we have mentioned, differentially private algorithm outputs should be relatively independent of changes to a single row in the database. Thus, we define:

Definition 2.3. Two databases X and X' are *neighbors* if they differ on at most one row.

In particular, because each row is assumed to belong to a specific individual, the order of the rows matters.

Example 2.4. In the table below, we have three databases X, X', and Y, with their rows listed in order. In this case X and X' are neighbors. However, X and Y are not neighbors, even though Y contains the same set of rows as X', because Y contains the rows in a different order.

X	X' (neighbors with X)	Y (not neighbors with X)
$\{0,0\}$	$\{1,1\}$	$\{1,1\}$
$\{0,0\}$	$\{0,0\}$	$\{1,1\}$
$\{1,1\}$	$\{1,1\}$	$\{0,0\}$

□

2.1.1 ε-differential privacy

Definition 2.5 ([DMNS06])**.** Let $\mathcal{A}$ be an algorithm and S be a subset of its possible outputs.[2] $\mathcal{A}$ is *ε-differentially private* if for all S and neighboring databases X, X',

$$\Pr[\mathcal{A}(X) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{A}(X') \in S].$$

Smaller values of ε correspond to more privacy, because they guarantee that the output distributions for neighboring databases lie closer together. However, it is not yet well-understood how to choose values of ε in practice, or what values of ε would be considered acceptable by those individuals whose information is contained in the database.

Because we are primarily interested in how our bounds change as the amount of privacy required increases, we assume for the sake of asymptotic notation that ε is upper-bounded by some constant, say $\varepsilon < 1$.

[2] Where the output of A falls on a continuous range, we also require that $\Pr[\mathcal{A}(X) \in S] > 0$.

As desired, ε-differential privacy is a worst-case guarantee: it is difficult for an adversary to distinguish between X and X' based on the output of $\mathcal{A}$, regardless of the database X. Moreover, this is true even if an adversary knows all of the database except for the one row that differs between X and X' [KS08].

This is, however, a very strong definition of privacy: many algorithms that fulfill our intuitive understanding of privacy nevertheless do not fulfill differential privacy.

Example 2.6. An algorithm that deterministically releases the average salary from a sample of several thousand people is *not* ε-differentially private. □

Despite the lack of ε-differential privacy in this example, many people would reasonably agree that the information released does not result in much, if any, privacy loss.[3]

2.1.2 (ε, δ)-differential privacy

One way to weaken this definition is to allow for a small probability of some catastrophic privacy leakage. This leads to the following:

Definition 2.7. Two random variables A and A' are (ε, δ)-*indistinguishable* if for all subsets S of their outputs,

$$\Pr[A \in S] \leq e^{\varepsilon} \cdot \Pr[A' \in S] + \delta \quad \text{and} \quad \Pr[A' \in S] \leq e^{\varepsilon} \cdot \Pr[A \in S] + \delta.$$

We denote this by

$$A \approx_{\varepsilon,\delta} A'.$$

Definition 2.8 ([DMNS06])**.** Let $\mathcal{A}$ be an algorithm and S be a subset of its possible outputs. $\mathcal{A}$ is (ε, δ)-*differentially private* if for all S and neighboring databases X, X',

$$\mathcal{A}(X) \approx_{\varepsilon,\delta} \mathcal{A}(X').$$

Note that an algorithm is ε-differentially private if and only if it is $(\varepsilon, 0)$-differentially private.

Example 2.9. An algorithm that outputs the entire database with probability δ and some fixed value $\perp$ with probability $1 - \delta$ is still (ε, δ)-differentially private for any value of ε. □

As this example shows, in (ε, δ)-differential privacy there are no guarantees about how much leakage occurs in the case of a "bad" output. We can also think of (ε, δ)-differential privacy as requiring that the algorithm release a "good" output with probability at least $1 - \delta$. In particular:

Proposition 2.10. *Let $\mathcal{A}$ be an algorithm. Define*

$$S_{\varepsilon}(X, X') = \left\{x : \Pr[\mathcal{A}(X) = x] \leq e^{\varepsilon} \cdot \Pr[\mathcal{A}(X') = x]\right\}.$$

Then, if for all neighboring databases X, X',

$$\Pr[\mathcal{A}(X) \notin S_{\varepsilon}(X, X')] \leq \delta,$$

$\mathcal{A}$ is (ε, δ)-differentially private. This also holds if the output of $\mathcal{A}$ is a continuous random variable with probability density function p, i.e. if we define $S_{\varepsilon}(X, X') = \{x : p(\mathcal{A}(X) = x) \leq e^{\varepsilon} \cdot p(\mathcal{A}(X') = x)\}$.

[3] As it turns out, releasing these types of aggregate statistics is not necessarily so safe as the number of statistics released grows sufficiently large.

Proposition 2.11. *Define $S_\varepsilon(X, X')$ as before. If there exist neighboring databases X, X' for which*

$$\Pr\left[\mathcal{A}(X) \notin S_\varepsilon(X, X')\right] > \delta,$$

then $\mathcal{A}$ is not $(\varepsilon/2, \delta(1 - e^{-\varepsilon/2}))$-differentially private.

Additionally, when it comes to parameter choice, (ε, δ)-differential privacy is only meaningful for reasonably small δ.

Example 2.12. An algorithm that randomly chooses and outputs a single row of the database is still $(\varepsilon, \frac{1}{n})$-differentially private for all ε. □

As we would generally agree that such an algorithm is not particularly private, we require that $\delta = o\left(\frac{1}{n}\right)$. In fact, it is common take δ to be "cryptographically small," and in particular subpolynomial in n – i.e. $\delta(n) = \frac{1}{n^{\omega(1)}}$.

Even with the allowance for δ, we note that (ε, δ)-differential privacy does not necessarily encompass many intuitive notions of privacy, such as those that we will see when examining statistical disclosure limitation techniques. In particular, it still excludes deterministic methods:

Proposition 2.13. *If an algorithm $\mathcal{A}$ is deterministic but not constant, then it is neither ε-differentially private nor (ε, δ)-differentially private.*

2.1.3 $(\varepsilon, \delta, \Delta)$-differential privacy

Moving back to randomized algorithms, the databases X that are most problematic for differential privacy with regards to a particular algorithm may be quite rare. For example, we will see that the worst case for privacy with Bayesian inference often occurs when everybody in a database has the same attribute. But with real databases drawn at random from a population, we would not actually expect this to happen.

Consequently, we examine one further relaxation of (ε, δ)-differential privacy: while the two formulations of differential privacy thus far hold for worst-case data, we can also look at "average"-case privacy, where "average" is defined over the sampling that produced the database X. To do this, we use a definition from Bassili et al.:[4]

Definition 2.14 ([BGKS13])**.** Let $\mathcal{A}$ be an algorithm, S be a subset of its possible outputs, and Δ be a set of probability distributions $\mathcal{D}$ on possible databases X. Let X_i denote the ith row of X, and let X_{-i} denote the database X with the ith row removed. We say that $\mathcal{A}$ is $(\varepsilon, \delta, \Delta)$-*differentially private* if there exists a simulator Sim such that for all $\mathcal{D} \in \Delta$ and $x \in \mathcal{X}$,

$$(\mathcal{A}(X) \mid X_i = x) \approx_{\varepsilon,\delta} (\mathrm{Sim}(X_{-i}) \mid X_i = x)$$

(where the probability is taken over both the randomness of $\mathcal{A}$ and of the draw $X \leftarrow \mathcal{D}$).

However, we will not really need to deal with the presence of a simulator: because we are only interested in databases X where each the rows are assumed to be generated independently from the same distribution, we can reformulate this to more resemble differential privacy.

[4] Bassili et al. also included the concept of outside information as a parameter in their definition, but we have omitted it since we do not use it during the analysis in later chapters.

Proposition 2.15. *Suppose Δ only contains distributions $\mathcal{D}$ such that each row of X is generated i.i.d. If $\mathcal{A}$ is $(\varepsilon, \delta, \Delta)$-differentially private, then for all $\mathcal{D} \in \Delta$ and $x, x' \in \mathcal{X}$,*

$$\mathcal{A}(X) \mid X_i = x \approx_{2\varepsilon, \delta(1+e^{\varepsilon})} \mathcal{A}(X) \mid X_i = x'.$$

Proof. We use $\Pr_{\mathcal{D}}$ to denote the fact that probabilities are taken over the distribution $\mathcal{D}$ in addition to the randomness of $\mathcal{A}$ and Sim. Then,

$$\begin{aligned} \Pr_{\mathcal{D}}[\mathcal{A}(X) \in S \mid X_i = x] &\leq e^{\varepsilon} \cdot \Pr_{\mathcal{D}}[\mathrm{Sim}(X_{-i}) \in S \mid X_i = x] + \delta \\ &= e^{\varepsilon} \cdot \Pr_{\mathcal{D}}[\mathrm{Sim}(X_{-i}) \in S \mid X_i = x'] + \delta \\ &\leq e^{\varepsilon}(e^{\varepsilon} \cdot \Pr_{\mathcal{D}}[\mathcal{A}(X) \in S \mid X_i = x'] + \delta) + \delta, \end{aligned}$$

where the two inequalities follow from $(\varepsilon, \delta, \Delta)$-differential privacy and the equality follows from the independence of the rows of X. □

2.2 Properties

Following directly from the definitions, differential privacy has a few key properties that are useful for analysis and mechanism design:

- Differential privacy cannot be undone; that is, no amount of processing on differentially private information alone can reduce the privacy involved.

  Proposition 2.16 (Post-processing [DMNS06]). *If $\mathcal{A}$ is (ε, δ)-differentially private, then for every algorithm $\mathcal{B}$, $\mathcal{B} \circ \mathcal{A}$ is also (ε, δ)-differentially private.*

- The amount of cumulative privacy loss from running multiple algorithms is at most linear.

  Proposition 2.17 (Basic Composition [DMNS06]). *If $\mathcal{A}_1$ and $\mathcal{A}_2$ are algorithms that are, respectively, $(\varepsilon_1, \delta_1)$ and $(\varepsilon_2, \delta_2)$ differentially private, then the algorithm outputting $(\mathcal{A}_1(X), \mathcal{A}_2(X))$, using independent randomness for A_1 and A_2, is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-differentially private.*

  This composition theorem holds for both ε-differential privacy and (ε, δ)-differential privacy. In fact, it is the optimal bound possible for the composition of two ε-differentially private algorithms into another ε-differentially private algorithm.

- However, if we are willing to accept an increase in the probability of a catastrophic event, then the cumulative ε value can be made sublinear in the number of algorithms.

  Theorem 2.18 (Advanced Composition [DRV10]). *If $\mathcal{A}_1, \ldots, \mathcal{A}_k$ are all (ε, δ)-differentially private algorithms with independent randomness, then the algorithm outputting $(\mathcal{A}_1(X), \ldots, \mathcal{A}_k(X))$ is $(\varepsilon_k, \delta_k)$-differentially private for*

  $$\varepsilon_k = \varepsilon\sqrt{2k \ln \frac{1}{\delta'}} + k\varepsilon(e^{\varepsilon} - 1) \text{ and } \delta_k = k\delta + \delta'$$

  for any $\delta' > 0$.

2.3 Mechanisms

Armed with the composition theorems of the previous section, it is possible to build differentially private algorithms either from repeated trials or by a combination of basic mechanisms.

The first of these basic mechanisms – the Laplace mechanism – can make any algorithm with real-numbered output ε-differentially private by adding random noise. For this to work, the amount of noise added needs to be sufficiently large to mask any change in a single row in the data. Consequently, we need to know how much the output can change when a single row of the database does.

Definition 2.19. Let $\mathcal{A}$ be an algorithm whose output is in the real numbers. The *global sensitivity*,[5] or just sensitivity, of $\mathcal{A}$ is defined as

$$\text{Sen}(\mathcal{A}) = \max_{X,X' \text{ neighboring}} |\mathcal{A}(X) - \mathcal{A}(X')|.$$

Theorem 2.20 (Laplace mechanism [DMNS06])**.** *Let $\mathcal{A}$ be an algorithm whose output is in $\mathbb{R}$. Then, the algorithm*

$$\mathcal{A}' : X \to \mathcal{A}(X) + \text{Lap}\left(\frac{\text{Sen}(\mathcal{A})}{\varepsilon}\right)$$

is ε-differentially private, where Lap *is the Laplace distribution*[6] *over the real numbers.*

However, the Laplace mechanism does not obviously adapt to more structured outputs. In the context of this thesis, this means that while the Laplace mechanism can be used to output parameters, it does not offer a way to output synthetic data that is structured similarly to the original database. The following mechanism, instead, provides a more flexibility when it comes to the output format.

Theorem 2.21 (Exponential mechanism [MT07])**.** *Let $\mathcal{R}$ be a set with measure $\mu : \mathcal{P}(\mathcal{R}) \to [0, 1]$, and let $q : \mathcal{X}^n \times \mathcal{R} \to \mathbb{R}$ be a utility function on (database, ouput) pairs. Then, so long as $\int_R e^{\frac{\varepsilon q(X,r)}{2\,\text{Sen}(q)}} \cdot \mu(r)dr < \infty$, the algorithm outputting each $r \in \mathcal{R}$ with probability density*

$$p(r) \propto e^{\frac{\varepsilon q(X,r)}{2\,\text{Sen}(q)}} \cdot \mu(r)$$

is ε-differentially private.

While this algorithm chooses higher-utility outputs with exponentially greater probability, it may be required to evaluate exponentially many such outputs. For example, there are $|\mathcal{X}|^m$ output options for a synthetic database of size m. Consequently, naïvely computing the full probability distribution function in this case requires exponential time.

Indeed, we know that the exponential mechanism can be used to privately release synthetic databases, such that the databases are very close to the original on many counting queries;[7] at the same time, doing so efficiently is not achievable and can lead to reconstruction attacks [BLR11, Ull12, KRS12].

[5] Some algorithms also reference the *local sensitivity* at X, which is defined as

$$\text{Sen}(\mathcal{A}, X) = \max_{X' \text{ neighboring } X} |\mathcal{A}(X) - \mathcal{A}(X')|.$$

For examples, see [NRS07] or [DL09].

[6] The Laplace distribution Lap(λ) has probability density function $p\,(x) \propto e^{-|x|/\lambda}$. It has mean 0 and standard deviation $\lambda\sqrt{2}$.

[7] A counting query $q : \mathcal{X}^n \to \mathbb{Z}$ is determined by some $q_{row} : \mathcal{X} \to \{0, 1\}$, with $q(X) = \sum_{x \in X} q_{row}(x)$.

As we will see, some of the Bayesian inference methods we examine can be interpreted as instantiations of the exponential mechanism for the appropriate choice of measure and utility functions. Thus, the statistical origins of the methods provide us with some help in computing them.

Finally, we conclude with one mechanism for amplifying differential privacy.

Theorem 2.22 (Subsampling [KLN+08])**.** *Suppose that $\mathcal{A} : \mathcal{X}^m \to \mathcal{S}$ is an (ε, δ)-differentially private algorithm. Then, the algorithm $A' : \mathcal{X}^n \to \mathcal{S}$ that first selects a random sample of m distinct rows from X and then runs $\mathcal{A}$ on the sample is (ε', δ')-differentially private for*

$$\varepsilon' = \frac{m}{n}(e^{\varepsilon} - 1), \quad \delta' = \frac{m}{n}\delta.$$

Because $\frac{m}{n}(e^{\varepsilon} - 1) \approx \frac{m\varepsilon}{n}$, this implies that we can convert any 1-differentially private algorithm into an ε-differentially private algorithm by increasing the database size by a factor of $\frac{1}{\varepsilon}$. However, privacy comes at a cost of greater variance or lesser accuracy compared to running the equivalent algorithm on the full database: if we think of the database as being, for instance, a random sample from the population, then we are effectively running the algorithm on a sample that is smaller by a factor of ε.

Chapter 3: Statistical Disclosure Limitation

For some time, statistics has also been concerned with privacy, in an area known as *statistical disclosure limitation* or *statistical disclosure control*. However, statistical research has often focused on examining the accuracy of techniques – for example, the amount of bias or variance introduced – rather than with making formal privacy guarantees; one of the exceptions to this is the Duncan-Lambert framework [DL86, Rei05].

3.1 Bayesian Inference and Multiple Imputation

We will evaluate the differential privacy of *multiple imputation*, a specific technique for synthetic data generation and analysis [Rub76, Rub93]. While multiple imputation (to be explained more formally in the next section) was originally developed to handle missing data by filling in the absent rows or values, it can equally be used to generate entirely synthetic data.

To do so, multiple imputation assumes that rows in the database are drawn independently and identically distributed (i.i.d.) from some underlying model. Then, via Bayesian inference, it learns the probability distribution for such a model from the existing data, and draws from the learned distribution to generate new rows. When filling in a partial row, it instead draws from the learned distribution while conditioning on the known values in the row.

To reduce the likelihood of learning a poor model or drawing a particularly bad set of synthetic data, this process is often repeated multiple times; thus, multiple imputation also provides a framework for analyzing the bias and variance of the synthetic data and for combining the data from multiple trials.

3.2 Methods Studied

Because we work with both discrete and continuous probability distributions, we will use $\Pr[]$ to denote a probability and $p()$ to denote a probability density. When the distribution is not defined and the analysis can hold for both continuous and discrete probability distributions, we use $\Pr[]$.

Additionally, recall from Chapter 2 that we use X to denote a database, n to denote its size – i.e. the number of rows – and $\mathcal{X}$ to denote the universe of possible rows in X.

Given a database X, our goal is to enable analysts to learn more about the original model that generated X. More specifically, we assume that each row $x \in X$ is generated i.i.d. from some underlying distribution $D(\theta)$, where D is a known family of distributions parameterized by unknown value(s) θ. Then, we want to release information that allows analysts to learn about θ. However, note that the assumption of X originating from the distribution is used solely for the purpose of analysis on the data, but does not play into an evaluation of privacy: that is, differential privacy takes into account only the randomness of the algorithm, not the randomness in data collection.

Our beliefs about θ before having seen the data are given by a prior distribution $\Theta(\theta_0)$ parameterized by value(s) θ_0. This prior θ_0 is assumed to be publicly known, and will generally take the form of a conjugate prior for the distribution family.[1]

The posterior distribution is given by Bayes' rule:

$$\Pr[\theta \mid X] = \frac{\Pr[X \mid \theta] \cdot \Pr[\theta]}{\Pr[X]}.$$

Example 3.1 (Bernoulli distribution with uniform prior)**.** A Bernoulli random variable $B \sim \text{Bern}(\theta)$, for $\theta \in [0, 1]$, is one taking on the value 1 with probability θ and the value 0 with probability $1 - \theta$. If we have a uniform prior over $[0, 1]$ on θ, the posterior probability distribution is

$$p(\theta \mid X) \propto \theta^{n_1}(1 - \theta)^{n_0},$$

were n_1 is the number of 1's in X and n_0 the number of 0's in X. □

Because multiple imputation involves multiple steps – first generating a value of θ and then generating synthetic data – we want to understand how each of these steps contributes to the privacy provided. Consequently, we will examine methods both for releasing θ and for releasing synthetic data.

We begin with two ways of generating θ:

> **Method** $\textsc{Param}_{\text{MAP}}$: Release θ_{MAP}, the most likely value of θ in the posterior distribution $\theta \mid X$.[2]
>
> **Method** $\textsc{Param}_{\text{Post}}$: Release a value of θ sampled from the posterior distribution $\theta \mid X$.

Example 3.2 (Bernoulli, cont.)**.** Maintaining the notation from the previous example, $\textsc{Param}_{\text{MAP}}$ would release $\frac{n_1}{n}$, while $\textsc{Param}_{\text{Post}}$ would release a value of θ drawn from the posterior distribution noted above. □

$\textsc{Param}_{\text{MAP}}$, being deterministic, will not even be (ε, δ)-differentially private by Proposition 2.13 for any distribution of interest. On the other hand, as $\textsc{Param}_{\text{Post}}$ introduces randomness in the sample, we will show in Chapters 6 and 7 that privacy becomes possible in the specific distribution families considered.

For synthetic data generation, we use $Z \in \mathcal{X}^m$ to denote a synthetic database of size m and $z \in \mathcal{X}$ to denote a single synthetic data point. Following the idea behind multiple imputation, we use either $\textsc{Param}_{\text{MAP}}$ or $\textsc{Param}_{\text{Post}}$ to generate θ, and then generate data points using that value of θ.

> **Method** $\textsc{Synth}_{\text{MAP}}$: Release Z consisting of m independent samples from $D(\theta_{\text{MAP}})$.
>
> **Method** $\textsc{Synth}_{\text{Post (One)}}$: Release Z consisting of m independent samples from $D(\theta)$, where θ is generated *once* using $\textsc{Param}_{\text{Post}}$.
>
> **Method** $\textsc{Synth}_{\text{Post (Many)}}$: Release Z consisting of m independent samples from $D(\theta)$, where θ is *regenerated* using $\textsc{Param}_{\text{Post}}$ for each point of the synthetic data. This is equivalent to releasing m i.i.d. samples from the posterior distribution.

Because $\textsc{Synth}_{\text{Post (One)}}$ and $\textsc{Synth}_{\text{Post (Many)}}$ are identical when $m = 1$, we refer to them collectively as $\textsc{Synth}_{\text{Post}}$ in this case.

Example 3.3 (Bernoulli, cont.)**.** Again maintaining the notation used thus far, for a Bernoulli variable with uniform prior, $\textsc{Synth}_{\text{MAP}}$ would release m independent data points, each of which is equal to 1 with probability $\frac{n_1}{n}$.

[1] A conjugate prior for a finitely parameterized family of distributions is one such that both the prior and posterior distributions fall within the family. This characteristic is generally convenient for computation because the only change between the prior and posterior distributions is that of the parameter(s).

[2] MAP stands for *maximum a posteriori*.

Meanwhile, $\textsc{Synth}_{\text{Post (One)}}$ would choose a value of θ and generate synthetic data as mentioned above, while $\textsc{Synth}_{\text{Post (Many)}}$ would release m independent data points, each of which is equal to 1 with probability $\frac{n_1+1}{n+2}$. □

In practice, synthetic data generation using multiple imputation is a combination of $\textsc{Synth}_{\text{Post (One)}}$ and $\textsc{Synth}_{\text{Post (Many)}}$: some number of θ's are generated, each of which is used to generate a synthetic dataset of some size. The number of distinct θ's used tends to resemble a small constant, while either the total number of synthetic data points generated over all parameters or the number of synthetic data points for each parameter is linear in the original. We have split this into two extreme cases in order to separate the effects of these two choices.

3.3 Distributions Studied

Now, we set up the two distributions whose privacy we analyze later in detail. These, along with their generalizations, are among the simplest distributions. However, they are both widely recurrent and provide examples of important categories of distributions – for instance, a discrete distribution (Bernoulli) compared to a continuous distribution (Gaussian).

3.3.1 The Bernoulli Distribution

Our first distribution is the Bernoulli, which we already saw in Example 3.1. As mentioned there, a Bernoulli random variable is one that takes only on two values, i.e. $\mathcal{X} = \{0, 1\}$. The distribution $\text{Bern}(\theta)$ is parameterized by a single value $\theta \in [0, 1]$, which represents the probability that the random variable is equal to 1. As in the example, we use n_0, n_1 to denote the number of $0's$ and $1's$, respectively, in X.

A binomial distribution is the sum of a n i.i.d. Bernoulli random variables, and its conjugate prior is the continuous Beta distribution $\text{Beta}(\alpha, \beta)$, with density

$$p\left(\theta\right)_{\alpha,\beta} = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathrm{B}(\alpha, \beta)},$$

where $\mathrm{B}(\alpha, \beta)$ is the Beta function

$$\mathrm{B}(\alpha, \beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}.$$

Because the size n of the data is publicly known, we allow the parameters $\alpha(n), \beta(n)$ to depend on n but not on the actual database X.

This prior can be interpreted as the analyst having previously seen $\alpha - 1$ examples of a 1 and $\beta - 1$ examples of a 0 before getting access to the data. Thus, when $\alpha = \beta = 1$, this results in a uniform prior.

Consequently, $\theta \mid X \sim \text{Beta}(n_1 + \alpha, n_0 + \beta)$, with posterior probability density for the parameter given by

$$p\left(\theta \mid X\right) = \frac{\theta^{n_1+\alpha-1}(1-\theta)^{n_0+\beta-1}}{\mathrm{B}(n_1+\alpha, n_0+\beta)}$$

and posterior probability for the next data point z seen given by

$$\Pr\left[z = 1 \mid X\right] = \frac{n_1+\alpha}{n+\alpha+\beta}.$$

The interpretation of α, β also suggests two important characteristics of the Beta distribution, which are also reflected in Figure 3.1. Firstly, the density is concentrated close to $\frac{n_1+\alpha-1}{n+\alpha+\beta-2}$, the fraction of 1's we have

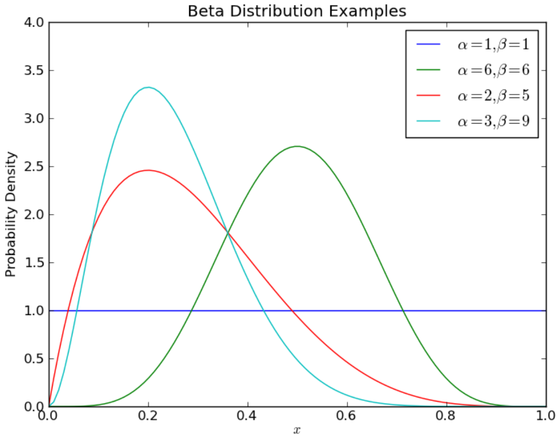

Figure 3.1: Beta distributions with various parameters. Note that $\alpha = 2, \beta = 5$ and $\alpha = 3, \beta = 9$ have the same ratio of 1's to 0's seen, but the latter has more total examples and hence concentrates closer to that ratio. Meanwhile, $\alpha = 3, \beta = 9$ and $\alpha = 6, \beta = 6$ have the same total number of examples seen but different ratios of 1's to 0's.

seen between the real and prior data. Secondly, the larger α and β are, the more probability is concentrated in this area – i.e. seeing more data points gives a stronger or more certain prior.

More concretely:

$$\begin{aligned} \mathrm{E}_{\alpha,\beta}\left[\theta \mid X\right] &= \frac{n_1 + \alpha}{n + \alpha + \beta} \\ \theta_{\mathrm{MAP}} = \mathrm{Mode}_{\alpha,\beta}\left[\theta \mid X\right] &= \frac{n_1 + \alpha - 1}{n + \alpha + \beta - 2} \\ \mathrm{Median}_{\alpha,\beta}\left[\theta \mid X\right] &\approx \frac{n_1 + \alpha - \frac{1}{3}}{n + \alpha + \beta - \frac{2}{3}}. \end{aligned}$$

3.3.2 The Gaussian Distribution with Known Variance

Our second distribution is the Gaussian (normal) distribution $\mathcal{N}(\mu, \sigma^2)$. For the Gaussian distribution, $\mathcal{X} = \mathbb{R}$, and the parameters consist of a mean μ (which is also the expectation and mode) and a standard deviation σ. The probability density function is

$$p_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

We focus only on the case where the standard deviation σ is pubicly known, but the mean μ is not. Consequently, unless specified otherwise, when we refer to the Gaussian distribution in the rest of this thesis, we will be referring to the one with known variance.

In this case, the conjugate prior on μ is another normal distribution $\mathcal{N}(\mu_0, \sigma_0^2)$. Because μ_0 just applies a horizontal shift to the density function, we assume that $\mu_0 = 0$ to simplify our calculations.

Then, if we let $X = (x_1, \ldots, x_n)$ denote the database and $\overline{x} = \frac{1}{n}\sum x_i$ the average value,[3] the posterior

[3] $\overline{x}$ is also known as the *sufficient statistic.*

distribution for μ is

$$\mu \mid X = \mu \mid \overline{x} \sim \mathcal{N}\left(\frac{\sigma_0^2 \overline{x}}{\frac{\sigma^2}{n} + \sigma_0^2}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\right)$$

and the posterior probability for the next data point z seen is

$$z \mid X = z \mid \overline{x} \sim \mathcal{N}(\mu \mid \overline{x}, \sigma^2) = \mathcal{N}\left(\frac{\sigma_0^2 \overline{x}}{\frac{\sigma^2}{n} + \sigma_0^2}, \sigma^2 + \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\right).$$

3.4 Other Techniques in Statistical Disclosure Limitation

Before moving on to further analysis of these methods, we take a step back to look at the utility of multiple imputation for statisticians, by comparing it to other techniques that have been developed for statistical disclosure limitation. Over time, these techniques have encompassed a number of categories, some of which assume the existence of a trustworthy curator to whom queries can be submitted while others release the data (or some version thereof) itself [AW89].

3.4.1 Query Restriction

The former category, where analysts are not allowed direct access to the data, includes *query restriction*: limitations on the number or types of queries allowed. For an example of query restriction in use, we can look to census data – often a common target of analysis:

Example 3.4. To protect privacy, the United States Census Bureau guarantees that categories contains enough individuals to mask the identity of any one. In the case of geographic categories, at least 100,000 individuals are needed to release data about that category; for other data, 10,000 are needed. The Census Bureau's query system also limits the queries that can be posed: users interested in a particular small-population area are limited to broader queries that those who are interested in, say, nation-wide values. Additionally, the system refuses to output results that do not apply to a sufficiently large group of individuals [Zay07]. □

So long as noise is not added to the query outputs, query restriction preserves the accuracy of the analysis, but makes it more difficult to gain access to data. For example, the Census Bureau initially restricts use of their query system to beta testers and those physically located at their own information centers. Other data holders may have their own processes for applying for access, all of which involve an additional cost for the analyst.

Query restriction also suffers from two other key constraints:

1. There may not be a guarantee that the data curator exists in the long term. While this may be less applicable to the Census Bureau, it is a much bigger problem in the case of individual data holders who would prefer not to take on the burden of handling queries for an unrestricted period of time.
2. Limitations on the amount of privacy loss or on the type of queries necessarily restrict the utility of the dataset, possibly with the result that analysis on the dataset must be stopped (or become sufficiently inaccurate as to be useless) after some period of time.

3.4.2 Data Suppression

To avoid these constraints, some statisticians began turning to releasing the data wholesale after making changes to preserve privacy [Sla13]. Broadly speaking, these changes can be grouped into two categories: data suppression and data perturbation [Dre11].

Data suppression techniques generally fall in line with the anonymization methods mentioned in the introduction: that is, they remove or generalize identifying fields. However, as we mentioned in Example 1.1 and Example 2.2, they have also been subject to a number of reindentification attacks in recent years. This suggests that the amount of data anonymized thus far has been insufficient: while the goal is to release as much data as possible without compromising privacy, it is unclear whether this is possible without removing many fields of significant analytical value. However, because data suppression does not alter not alter correlations in the original data, whatever (limited) analysis can be performed on the anonymized database tends to be fairly accurate.

3.4.3 Data Perturbation

On the other hand, *data perturbation* – whether by generating fully synthetic data or by simply altering individual columns – tends to distort the results of statistical analyses, and thus is often less preferred by analysts.

One method of perturbation is data swapping, which randomly exchanges the values of some identifying variables between rows. This might involve, for example, randomly choosing pairs of individuals in a database and swapping their incomes; as a result, an adversary who has identified an individual in the database based on other characteristics would not be able to guarantee that s/he has found the correct income as well. This, too, is a major technique used by the Census Bureau [Zay07]. Yet, it can often make the data unusable for analysts: swapping fields in as few as 1% of rows will still produce problematically low confidence intervals when performing inference [DR10].

Our focus – multiple imputation – is another form of data perturbation. Because the replaced fields are fully synthetic, this is believed to offer more privacy than other techniques such as data swapping. As such, some researchers have suggested that it may be the only reasonable privacy-preserving mechanism for releasing some datasets. Indeed, the Census Bureau has begun releasing some synthetically generated data – for instance, for traffic maps – which appear on manual examination to both preserve privacy for small groups and to reasonably resemble the original data [Zay07].

Additionally, multiple imputation is more accurate than swapping rows, at least when trying to infer the basic model that generated the data [Rub76]. This has also held up in some real trials (see, for example, [MKA$^+$08]). However, because the data is wholly synthetic, it is easy to imagine spurious inferences produced or real ones accidentally eliminated; while the hope is that synthetic data generation succeeds in preserving more relationships within the data than may be captured in the model used, it is unclear how much this actually succeeds in practice.

Chapter 4: Related Work on Differential Privacy and Bayesian Inference

4.1 Alternate Statistical Estimators

As we mentioned in the Introduction, there has been significant work on algorithms designed to provide differential privacy. These include algorithms that release statistical estimates similar to the information we would like to convey using the methods outlined in Chapter 3.2. In particular, while $\textsc{Param}_{\text{MAP}}$ is not differentially private by itself, there exist for many distributions differentially private estimators of $\textsc{Param}_{\text{MAP}}$ that are asymptotically accurate as the number of data points grows.[1] We mention a few examples of such estimators here.

First, Dwork and Lei gave a method – the Propose-Test-Release framework – for releasing estimations of distribution parameters with differential privacy [DL09]. In particular, Propose-Test-Release takes advantage of the idea that mechanisms with low sensitivity can be accurately released with the addition of Laplace noise (Theorem 2.20). Because in many cases – i.e. with high probability over the sampling of the data – the local sensitivity of the dataset at X will be low even when the global sensitivity cannot be universally bounded, the framework first tests with differentially privacy whether the sensitivity is low and, if so, outputs the mechanisms with added noise.

Example 4.1. Propose-Test-Release enables the (ε, δ)-differentially private release of medians and trimmed means[2] with additive noise $O\left(\frac{1}{n}\right)$. □

For both the Bernoulli and Gaussian distributions, in a large dataset releasing a trimmed mean is approximately the same as releasing $\textsc{Param}_{\text{MAP}}$. One alternate way to accomplish this, as we mention in later chapters, is to release $\textsc{Param}_{\text{MAP}} + \text{Lap}()$, for the right amount of Laplace noise. However, while this works cleanly for Bernoulli-distributed data, it only works for Gaussian-distributed data if we first limit the range of the data. On the other hand, releasing the median or trimmed mean for a Gaussian using Propose-Test-Release does not suffer from this constraint.

Then, in the Subsample and Aggregate framework, Nissim, Raskhodnikova, and Smith gave another way of estimating the median of a bounded dataset with error $\frac{1}{\text{poly}(n)}$ which also uses the idea of bounding the local sensitivity [NRS07].

Building on a variant of Subsample and Aggregate, Smith showed that for a statistical estimator T and a data distribution drawn i.i.d. from a distribution $\mathcal{D}$, there exists a differentially private version $\mathcal{M}_T$ of that T such that T and $\mathcal{M}_T$ converge in distribution so long as T (or a rescaled version thereof) converges to a normal distribution [Smi11, Smi08]. In the cases of the Bernoulli and Gaussian distributions, the condition of asymptotic normality holds for most of the methods specified in Chapter 3.2, including both the parameter release and synthetic data release ones. Meanwhile, compared to the variance of T, the variance of $\mathcal{M}_T$ is

[1] The accuracy is taken with respect to the randomness over both the algorithm and the data sample.

[2] An example of a trimmed mean would be the mean of the middle 50% of elements of a dataset.

within a $1 + o(1)$ factor, suggesting that little accuracy needs to be lost in the process of adding differential privacy [Smi11].

Consequently, we know that information releases similar to the methods we will examine are not incompatible with differential privacy; moreover, it is possible to obtain differential privacy cheaply, in the sense that not a great deal of accuracy needs to be sacrificed. We take this further by examining whether differential privacy can be obtained for free with the original methods, rather than requiring modified ones.

4.2 Bayesian Inference and Differential Privacy

There has also been some work specifically examining the privacy of Bayesian inference itself. Our work improves some of these previous bounds while also extending analyzing parameter or synthetic data release methods not looked at by others. More specifically:

Machanavajjhala et al. showed that synthetic categorical data – a generalization of Bernoulli data – can be ε-differentially private given a sufficiently strong prior. In the context of our framework, they showed that:

Theorem 4.2 ([MKA$^+$08]). *There exists a constant c such that setting $\alpha = \beta \geq \frac{cm}{\varepsilon}$ provides ε-differential privacy for* $\textsc{Synth}_{\text{Post (One)}}$*, where m is the number of synthetic data points released.*

In Chapter 6, we extend this bound in the Bernoulli (but not categorical) case to $\textsc{Synth}_{\text{MAP}}$ and $\textsc{Synth}_{\text{Post (Many)}}$, and show that it is tight in all three cases. However, when m is linear in n, the amount of prior data given by this bound is unacceptably large: $\frac{cm}{\epsilon}$ may be much larger than the size of the original database. Thus, we also give bounds for the amount of (ε, δ)-differential privacy; our bounds only require sublinear quantities[3] of prior data.

Additionally, Dimitrakakis et al. looked general conditions on a distribution that would imply differential privacy for Bayesian inference [DNMR13]. They showed that:

1. Posterior samplings from distributions that fulfill Lipschitz continuity[4] are ε-differentially private. Unfortunately, Lipschitz continuity does not hold for many distributions, including both the Bernoulli and Gaussian.
2. Posterior samplings from distributions that are Lipschitz continuous with high probability over the sampling of X obey a somewhat relaxed version of (ε, δ)-differentially private. This category of distributions is larger, and includes the exponential, Laplace, binomial (Bernoulli), and Gaussian (with known mean) distributions.[5]

Applied to the Bernoulli distribution, this gives the following:

Theorem 4.3 ([DNMR13]). *In the case of a Bernoulli distribution with Beta prior* $\text{Beta}(\alpha, \beta)$*, there exists a constant $c \approx 2.46$ such that* $\textsc{Param}_{\text{Post}}$ *is $(0, \delta)$-differentially private when $\alpha, \beta \geq \frac{c}{\delta^2}$.*

Because we want $\delta = o\left(\frac{1}{n}\right)$, this means that $\alpha, \beta = \omega(n)$, with the result that the actual data becomes negligible compared to the prior data as n grows large. Instead, we will show that it is possible to achieve

[3] Assuming that δ is not exponentially small in n.

[4] These are distributions where there exists come constant c such that for all neighboring databases X, X' and all choices of prior θ, $|\Pr[X \mid \theta] - \Pr[X' \mid \theta]| < c$ (or $|p(X \mid \theta) - p(X' \mid \theta)| < c$) .

[5] It does not, however, include the Gaussian distribution with known variance, which is the version of the Gaussian that we examine.

(ε, δ)-differential privacy for $\varepsilon > 0$ with only logarithmic dependence on $\frac{1}{\delta}$, so that for any fixed ε, $\alpha, \beta = o(n)$.[6]

Finally, Bassily et al. [BGKS13] sought to generalize the privacy framework to *distributional differential privacy* (or $(\varepsilon, \delta, \Delta)$-differentially private, which we mentioned in Definition 2.14), by:

1. Treating the dataset X as a random variable that contributes to the randomness providing privacy.
2. Restricting the universe of datasets X for which the privacy condition is required to hold.

They showed that for a finite family of distributions, so long as the ratio of probabilities for any output under different distributions is bounded, $\text{PARAM}_{\text{MAP}}$ satisfies ε-distributional differential privacy. However, the requirement that the family be finite means that, as we show in Chapter 6.3, it does not apply directly to the Bernoulli case.

4.3 Accuracy and Privacy

Although we do not look at accuracy in depth, it is worth a word about it before moving on. As we have mentioned, multiple imputation has often provided better accuracy than previous forms of statistical disclosure limitation, with good results when applied to real datasets [RRR03]. This is especially true for partially synthetic data, when only a handful of fields but not the entire database are synthetically generated [DBR08].

However, these results on accuracy has applied primarily to multiple imputation when the priors are not chosen with respect to differential privacy. Indeed, more recent work suggests that when the priors are chosen for privacy, more significant accuracy loss results.

First, Machanavajjhala et al., running their algorithms on a (categorical) dataset of commute distances, verified that the synthetic data tended to systematically overestimate the likelihood of rare categories [MKA+08]. When they restricted their data to more common categories, then the error decreased correspondingly. This issue tends to be most particularly pronounced when the data is sparse – e.g. if the number of categories is on the order of the number of data points – since then the number of prior data points stays significant relative to the number of real data points.

Secondly, also working with several real datasets and a categorical distribution, Charest found that even when the amount of privacy required is low – i.e. $\varepsilon = 1$, which is considered a fairly large value – large error rates still resulted. Even at this level of privacy, prior selection for differential privacy still led to error rates as large as 25% and models that fit poorly on the real data, although private synthetic data generation did work very well in a few of the cases [Cha12]. Meanwhile, in simulations, she found that because multiple imputation often assumes a noninformative prior, the estimates it produced on differentially private data were both significantly biased and led to large (and therefore unhelpful) confidence intervals on the estimates. Even modifying the multiple imputation framework to account for the different priors only partially reversed the problem, especially when a large amount of privacy was required, and still led to notably greater variance and bias than would have been the case with the real data [Cha11].

[6] Again, assuming that δ is not exponentially small in n.

Chapter 5: Parameter vs. Synthetic Data Release

One question of interest is not just how much privacy is afforded by multiple imputation, but also the relative contributions of the two steps – parameter generation and synthetic data generation – to privacy. In particular, if parameter release affords the same amount of privacy as synthetic data release, then we might as well release the parameters and leave the synthetic data generation to analysts themselves. Thus, before diving into specific distributions, we first give a comparison of synthetic data and parameter release methods.

We hypothesize that when the number of synthetic data points released by our methods is large, then for distributions of interest, an adversary gains roughly the same knowledge as s/he would from a parameter release. Table 5.1 recaps the synthetic data methods and their corresponding parameter release methods. In the table, releasing an infinite number of data points using the method in the first column is equivalent to releasing the parameter according to the method in the second column, in the sense that the amount of privacy loss is the same.

Synthetic data release method	Parameter release method
$\textsc{Synth}_{\text{MAP}}$	$\textsc{Param}_{\text{MAP}}$
$\textsc{Synth}_{\text{Post (One)}}$	$\textsc{Param}_{\text{Post}}$

Table 5.1: The correspondences between synthetic data release methods and the parameter release methods which were used to generate the model from which the synthetic data was drawn. We have not included $\textsc{Synth}_{\text{Post (Many)}}$ here because the parameters are redrawn for each data point in that case, so it is not quite analogous to the other correspondences.

The implications from parameter release to synthetic data release follow from post-processing, because generating a parameter allows us to generate an arbitrary number of synthetic data points using it.

Theorem 5.1. *For any family of distributions D and choice of prior, if $\textsc{Param}_{\text{MAP}}$ is (ε, δ)-differentially private, so is $\textsc{Synth}_{\text{MAP}}$. Similarly, if $\textsc{Param}_{\text{Post}}$ is (ε, δ)-differentially private, then so is $\textsc{Synth}_{\text{Post (One)}}$.*

For the other direction, the intuitive idea is that the more synthetic data points released, the better an approximation to the PDF $D(\theta)$ can be obtained using these data points; an adversary can use this approximate PDF to estimate θ itself. To formalize these, we also need to define what it means for the adversary to be able to use $D(\theta)$ to estimate θ.

Definition 5.2. Let $D(\theta)$ be a PDF, and let $D_m(\theta)$ denote the results of m i.i.d, trials of $D(\theta)$. We say that $D(\theta)$ *determines* θ if, for any θ, θ',

$$\lim_{m \to \infty} \left(\Pr\left[D_m(\theta) \in T_m\right] - \Pr\left[D_m(\theta') \in T_m\right] \right) = 1.$$

Example 5.3. The Bernoulli, Beta, and Gaussian distributions determine their parameters. □

In other words, for any two values of θ, an adversary could determine with arbitrarily high confidence which it is by examining the results of i.i.d. trials of $D(\theta)$.

Theorem 5.4. *Suppose that D is a family of distributions such that $D(\theta)$ determines θ. Then, for any fixed choice of n and prior,*[1] *and for any values of ε, δ, if* $\textsc{Param}_{\text{MAP}}$ *is non-constant there exists a value of m such that* $\textsc{Synth}_{\text{MAP}}$ *is not (ε, δ)-differentially private when releasing m synthetic data points.*

Proof. Fix two neighboring databases X, X' which give different results of $\textsc{Param}_{\text{MAP}}$. Let $T_1, T_2, \ldots$ be a sequence of outcomes of $\textsc{Synth}_{\text{MAP}}$ such that

$$\lim_{m \to \infty} \left(\Pr\left[\textsc{Synth}_{\text{MAP}m}(X) \in T_m\right] - \Pr\left[\textsc{Synth}_{\text{MAP}m}(X') \in T_m\right]\right) = 1.$$

Then, for sufficiently large m, $\Pr\left[\textsc{Synth}_{\text{MAP}m}(X) \in T_m\right] \geq e^{\varepsilon} \Pr\left[\textsc{Synth}_{\text{MAP}m}(X') \in T_m\right] + \delta$. □

We hypothesize that this is also true for the pair of $\textsc{Param}_{\text{Post}}$ and $\textsc{Param}_{\text{MAP}}$ given a stronger notion of "determines." In other words, we suspect that something similar to the following is true, but were not able to pinpoint a reasonable notion of determination that implies it.

Hypothesis 5.5. *Suppose that D is a family of distributions such that $D(\theta)$ determines θ. Then, for any fixed choice of n and prior, if* $\textsc{Synth}_{\text{Post (One)}}$ *is (ε, δ)-differentially private for all values of m, then* $\textsc{Param}_{\text{Post}}$ *is $(\varepsilon, c\delta)$-differentially private for any $c > 1$.*

Meanwhile, under the conditions that any two neighboring databases X, X' result in different posterior distributions for the parameter, it is possible to apply this idea to $\textsc{Synth}_{\text{Post (Many)}}$, because releasing many synthetic data points using $\textsc{Synth}_{\text{Post (Many)}}$ gives an approximation to the posterior distribution on the data:

Corollary 5.6. *Suppose that the posterior $\Theta(\theta_{post})$ is a family of distributions such that determines its parameter. Then, for any fixed choice of n, and for any values of ε, δ, if there exist two neighboring but nonidentical datasets X, X' will result in different values of θ_{post}, there exists a value of m such that* $\textsc{Synth}_{\text{Post (Many)}}$ *is not (ε, δ)-differentially private when releasing m synthetic data points.*

[1] In particular, we do not allow the choice of prior to depend on the number of synthetic data points m here.

Chapter 6: The Bernoulli-Binomial Distribution

Recall from Chapter 3.3 that:

- A Bernoulli random variable $\text{Bern}(\theta)$ is equal 1 with probability θ and 0 with probability $1 - \theta$.
- A binomial random variable $\text{Bin}(n, \theta)$ is the sum of n independent $\text{Bern}(\theta)$ variables.
- It has conjugate prior $\text{Beta}(\alpha, \beta)$, where $(\alpha - 1)$ and $(\beta - 1)$ can be thought of as the number of 1's and 0's, respectively, of prior data seen before the database X.

Additionally, we use n_1 and n_0 to denote the number of 1's and 0's, respectively, in X.

Perhaps unsurprisingly, increasing the amount of prior information by increasing α and β results in increased privacy: we can interpret this as weighting the algorithm away from the private database X and towards the public prior information.

Thus, in this chapter we provide a number of new upper and lower bounds on the quantity of prior data required to achieve differential privacy. Table 6.1 summarizes both our new bounds and those already shown by others.

Method	Privacy	Achievable?	Upper Bound	Lower Bound
$\textsc{Param}_{\text{MAP}}$	ε or (ε, δ)	No	n/a	n/a
$\textsc{Param}_{\text{Post}}$	ε	No	n/a	n/a
	$(0, \delta)$	Yes	$O\left(\frac{1}{\delta^2}\right)$ [DNMR13]	
	(ε, δ)	Yes	$O\left(\frac{1}{\varepsilon^2}\left(\ln n + \ln \frac{1}{\delta}\right)\right)$	$\Omega\left(\frac{1}{\varepsilon^2}\left(\ln \frac{\varepsilon}{\delta n}\right)\right)$
$\textsc{Synth}_{\text{MAP}}$	ε	Yes	$\Theta\left(\frac{m}{\varepsilon}\right)$	
	(ε, δ)	Yes	$O\left(\frac{1}{\varepsilon^2}\left(\ln \frac{1}{\delta}\right)\right)$ if $m \leq cn$	
$\textsc{Synth}_{\text{Post (One)}}$	ε	Yes	$\Theta\left(\frac{m}{\varepsilon}\right)$ [MKA+08]	
	(ε, δ)	Yes	$O\left(\frac{1}{\varepsilon^2}\left(\ln n + \ln \frac{1}{\delta}\right)\right)$	
$\textsc{Synth}_{\text{Post (Many)}}$	ε	Yes	$\Theta\left(\frac{m}{\varepsilon}\right)$	
	(ε, δ)	Yes	$O\left(\frac{1}{\varepsilon^2} \ln \frac{1}{\delta}\right)$ if $m \leq cn$	

Figure 6.1: The bounds for the Bernoulli distribution that we prove in this chapter along with those already found in related work. The bounds refer to those on T, where $T(n, \varepsilon, \delta)$ is the threshhold such that our methods are (ε, δ)-differentially private if and only if $\alpha, \beta \geq T$. Thus, upper bounds correlate with sufficient values of α and β required to achieve privacy, while lower bounds refer to necessary values of α and β.

Our Results

We show the parameter generation methods are not ε-differentially private...

Unsurprisingly, $\textsc{Param}_{\text{MAP}}$ is not even (ε, δ)-differentially private since it is deterministic. Unfortunately, we show that $\textsc{Param}_{\text{Post}}$ is also not ε-differentially private. Even if we relax our privacy definition to $(\varepsilon, \delta, \Delta)$-differential privacy in order to minimize the impact of unusual databases, however, these negative results still hold.

... but all the synthetic data generation methods can be made both ε-differentially private and (ε, δ)-differentially private with a sufficiently large prior.

We show that all of the synthetic data generation methods require $\Theta\left(\frac{m}{\varepsilon}\right)$ points of prior data, where m is the number of synthetic data points released, to achieve ε-differential privacy. Our results extend those of Machanavajjhala et al. [MKA+08], who showed this for $\textsc{Synth}_{\text{MAP}}$ only. Meanwhile, we also show that $\alpha, \beta = \Omega\left(\frac{1}{\varepsilon^2} \ln \frac{1}{\delta}\right)$ is sufficient for (ε, δ)-differential privacy.

Sometimes this prior is too large...

In general, analysts prefer that the amount of prior data become negligibly small compared to the amount of actual data, so that the choice of prior is not particularly important. In our case this translates to $\alpha, \beta = o(n)$, i.e. that α and β are sublinear in the size of the actual data. Moreover, because synthetic data sets generated in practice often have size equal to that of the original, we would prefer that α and β remain sublinear in n even when this is true – i.e. $m = n$.

Consequently, our bound that synthetic data generation requires $\alpha, \beta = \Theta\left(\frac{m}{\varepsilon}\right)$ for ε-differential privacy is not satisfactory in practice, because it requires that our prior data is larger in size than our real database when $m = n$. However, we also show that in fact this is the smallest α, β for which we obtain ε-differential privacy, meaning that it is impossible to do any better.

...but our bounds provide smaller priors than previous results for both parameter release...

We show that $\alpha, \beta = \Omega\left(\frac{1}{\varepsilon^2} \ln \frac{1}{\delta}\right)$ is sufficient to obtain (ε, δ)-differential privacy for $\textsc{Param}_{\text{Post}}$. So long as $\delta = e^{-o(n)}$, this gives us a sub-linear dependence on n. As we mentioned in Chapter 4, this is an improvement over the $\alpha, \beta = \frac{1}{\delta^2} = \Omega(n^2)$ bound of Dimitrakakis et. al [DNMR13], which was not sublinear in n.

...and synthetic data release.

Similarly, we show that $\alpha, \beta = \Omega\left(\frac{1}{\varepsilon^2} \ln \frac{1}{\delta}\right)$ suffices to give (ε, δ)-differential privacy for up to $m \leq cn$ points of synthetic data and some constant c. In particular, this removes the dependence of the prior data size on m, a major improvement over the bounds for ε-differential privacy.

The proofs of these bounds are given in the subsequent sections. For the sake of brevity, in the proofs we often consider only the neighboring databases X' of X that have a single 1 replaced by a 0; the other case of neighboring databases follows symmetrically.

6.1 Releasing the Parameter

6.1.1 Param$_{\text{MAP}}$

As always, $\textsc{Param}_{\text{MAP}}$ is not private, although it can be modified to be so.

Corollary 6.1. $\textsc{Param}_{\text{MAP}}$ *is not (ε, δ)-differentially private for any choice of prior.*

Proof. This follows from Proposition 2.13 and the fact that θ_{MAP} is deterministic in X but not constant. □

However, because $\theta_{\text{MAP}} = \frac{n_1+\alpha}{n+\alpha+\beta}$, its sensitivity is $\frac{1}{n+\alpha+\beta}$, so we can obtain privacy by adding Laplace noise according to Theorem 2.20.

Corollary 6.2. *Releasing* $\textsc{Param}_{\text{MAP}} + \text{Lap}\left(\frac{1}{\varepsilon(n+\alpha+\beta)}\right)$ *is ε-differentially private.*

From the perspective of accuracy, adding Laplace noise may not be a bad deal. In particular, from the Laplace distribution we know that

$$\mathrm{E}_{\alpha,\beta}\left[\left|\text{Lap}\left(\frac{1}{\varepsilon(n+\alpha+\beta)}\right)\right|\right] = \frac{1}{\varepsilon(n+\alpha+\beta)} = O\left(\frac{1}{\varepsilon n}\right).$$

Meanwhile, if we assume that each row of X was randomly drawn according to $\text{Bern}(\theta_{\text{Real}})$ for θ_{Real} some constant, then

$$\text{Var}_{\alpha,\beta}\left[\theta_{MAP}\right] = \frac{n(\theta_{\text{Real}})(1-\theta_{\text{Real}})}{(n+\alpha+\beta)^2} < \frac{\theta_{\text{Real}}(1-\theta_{\text{Real}})}{n},$$

which means that the Laplace noise has an extra factor of $\frac{1}{\varepsilon\sqrt{n}}$ compared to the sampling standard deviation; if $n = \omega\left(\frac{1}{\varepsilon^2}\right)$, then the sampling variance dominates any noise introduced. Admittedly, this works best when θ_{Real} and $1-\theta_{\text{Real}}$ are not too small – a pattern that we will continue to see in later methods.

However, the distribution $\theta_{\text{MAP}} + \text{Lap}\left(\frac{1}{\varepsilon(n+\alpha+\beta)}\right)$ is also more awkward to work with than the posterior, and we lack a clean closed-form PDF for it.

6.1.2 Param$_{\text{Post}}$

As we show next, the situation improves somewhat when we use the randomized parameter release method $\textsc{Param}_{\text{Post}}$: although ε-differential privacy is still unachievable, (ε, δ)-differential privacy is achievable with the addition of a sublinear quantity of prior data.

First, the bad news:

Theorem 6.3. $\textsc{Param}_{\text{Post}}$ *is not ε-differentially private for any choice of prior.*

Proof. For neighboring databases X, X', we have

$$\frac{p_{\alpha,\beta}\left(\theta \mid X\right)}{p_{\alpha,\beta}\left(\theta \mid X'\right)} = \frac{\theta^{n_1+\alpha-1}(1-\theta)^{n_0+\beta-1}}{\theta^{n_1+\alpha-2}(1-\theta)^{n_0+\beta}} \cdot \frac{\frac{(n_1+\alpha-2)!(n_0+\beta)!}{(n+\alpha+\beta-1)!}}{\frac{(n_1+\alpha-1)!(n_0+\beta-1)!}{(n+\alpha+\beta-1)!}} = \frac{n_0+\beta}{n_1+\alpha-1} \cdot \frac{\theta}{1-\theta} \geq \frac{\beta}{n+\alpha-1} \cdot \frac{\theta}{1-\theta}.$$

Consequently, when we look at values of θ lying close to 1 (i.e. in an interval $[\theta', 1]$),

$$\frac{\Pr_{\alpha,\beta}\left[\theta \in [\theta', 1] \mid X\right]}{\Pr_{\alpha,\beta}\left[\theta \in [\theta', 1] \mid X'\right]} \geq \frac{\beta}{n+\alpha-1} \cdot \frac{\theta'}{1-\theta'}.$$

Regardless of the choice of prior, this goes to infinity as $\theta' \to 1$, so we cannot bound it from above by e^{ε}. □

This illustrates the primary difficulty for privately releasing θ: for values of θ lying close to either 0 or 1, switching a single data point in X changes our probability of outputting θ by a significant multiplicative factor.

This idea motivates our turning instead to (ε, δ)-differential privacy. Because we are most likely to encounter such an extreme value of θ when X is nearly all 0's or nearly all 1's, we can use α and β to ensure that the combined prior and real data contain a minimum number of both values. Unfortunately, because there is always a non-zero probability of seeing a fairly extreme value of θ regardless of the prior, this fails to gives us ε-differential privacy, as we just saw. However, by bounding the likelihood of seeing an extreme value of θ, we do obtain (ε, δ)-differential privacy.

Theorem 6.4. *There exists a constant c such that for sufficiently large n, $\textsc{Param}_{\text{Post}}$ is (ε, δ)-differentially private when*

$$\alpha, \beta \geq \frac{c}{\varepsilon^2}\left(\ln n + \ln\frac{1}{\delta}\right).$$

Because we assume that $\delta = o\left(\frac{1}{n}\right)$, the $\ln\frac{1}{\delta}$ is expected to be the dominant term above. Additionally, so long as $\delta = e^{-o(n)}$, this is sublinear in n.

To prove this, we use two lemmas:

Lemma 6.5. *Let θ denote the output of $\textsc{Param}_{\text{Post}}$. If*

$$\Pr_{\alpha,\beta}\left[\frac{e^{\varepsilon}(n_1+\alpha-1)}{(n_0+\beta)+e^{\varepsilon}(n_1+\alpha-1)} \geq \theta \geq \frac{(n_1+\alpha)}{(n_1+\alpha)+e^{\varepsilon}(n_0+\beta-1)}\right] \geq 1-\delta,$$

then $\textsc{Param}_{\text{Post}}$ is (ε, δ)-differentially private.

Lemma 6.6. *There exists a constant c such that for sufficiently large n, when $\alpha, \beta \geq \frac{c}{\varepsilon^2}\left(\ln n + \ln\frac{1}{\delta}\right)$, then with probability $1-\delta$, $\textsc{Param}_{\text{Post}}$ releases a value of θ within $\varepsilon \min\left(\frac{n_1+\alpha}{n+\alpha+\beta}, \frac{n_0+\beta}{n+\alpha+\beta}\right)$ of $\frac{n_1+\alpha}{n+\alpha+\beta}$.*

Of these two lemmas, the second is key: it shows that $\textsc{Param}_{\text{Post}}$ lies very close to its expectation with high probability. Meanwhile, the first simply pulls out some of the computations for finding the range of θ such that $\Pr_{\alpha,\beta}[\theta \mid X] \leq e^{\varepsilon}\Pr_{\alpha,\beta}[\theta \mid X']$; its proof can be found in the Appendix.

Proof of theorem. Note that[1]

$$\theta_{\text{MAP}} = \frac{n_1+\alpha-1}{n+\alpha+\beta-2} \approx \frac{n_1+\alpha-1}{n+\alpha+\beta-1} \approx \mathrm{E}_{\alpha,\beta}[\theta \mid X] = \frac{n_1+\alpha}{n+\alpha+\beta}.$$

First, we check that when θ lies outside the range given by Lemma 6.5, then it differs from its most likely value (and likewise from its expectation) by at least $\Omega(\varepsilon \min(\theta_{\text{MAP}}, 1-\theta_{\text{MAP}}))$.

To show this, we use the approximation above to get that

$$\frac{e^{\varepsilon}(n_1+\alpha-1)}{(n_0+\beta)+e^{\varepsilon}(n_1+\alpha-1)} \approx \frac{e^{\varepsilon}\theta_{\text{MAP}}}{1+(e^{\varepsilon}-1)\theta_{\text{MAP}}},$$

so

$$\begin{aligned}\theta > \frac{e^{\varepsilon}(n_1+\alpha-1)}{(n_0+\beta)+e^{\varepsilon}(n_1+\alpha-1)} \Rightarrow |\theta - \theta_{\text{MAP}}| &> \left|\frac{e^{\varepsilon}\theta_{\text{MAP}}}{1+(e^{\varepsilon}-1)\theta_{\text{MAP}}} - \theta_{\text{MAP}}\right| \\ &\geq \frac{\varepsilon\theta_{\text{MAP}}(1-\theta_{\text{MAP}})}{1+(e^{\varepsilon}-1)\theta_{\text{MAP}}} \\ &= \Omega(\varepsilon \min(\theta_{\text{MAP}}, 1-\theta_{\text{MAP}})),\end{aligned}$$

where the second line follows from $e^{\varepsilon} \geq 1+\varepsilon$ and the third from $(1-e^{\varepsilon})\theta_{\text{MAP}} \leq (e^1-1)\cdot 1$.

[1] In particular, the fudge factor above in the approximation sign is $O\left(\frac{1}{n+\alpha+\beta}\right)$, and hence will not affect the asymptotic end result. Thus, to focus on the main ideas of the proof we use these three fractions interchangeably in the proofs in this section.

Similarly,

$$\theta < \frac{(n_1 + \alpha - 1)}{(n_1 + \alpha - 1) + e^{\varepsilon}(n_0 + \beta)} \Rightarrow |\theta_{\text{MAP}} - \theta| > \left|\theta_{\text{MAP}} - \frac{\theta_{\text{MAP}}}{\theta_{\text{MAP}} + e^{\varepsilon}(1 - \theta_{\text{MAP}})}\right|$$
$$= \Omega(\varepsilon \min(\theta_{\text{MAP}}, 1 - \theta_{\text{MAP}})).$$

Since $\theta_{\text{MAP}} \approx \frac{n_1+\alpha}{n+\alpha+\beta}$, by Lemma 6.6 there exists a constant c such that when $\alpha, \beta \geq \frac{c}{\varepsilon^2}\left(\ln n + \ln \frac{1}{\delta}\right)$, then

$$\Pr\left[\theta > \frac{e^{\varepsilon}(n_1 + \alpha - 1)}{(n_0 + \beta) + e^{\varepsilon}(n_1 + \alpha - 1)} \text{ or } \theta < \frac{(n_1 + \alpha - 1)}{(n_1 + \alpha - 1) + e^{\varepsilon}(n_0 + \beta)}\right] \leq \delta.$$

Consequently, by Lemma 6.5, this mechanism is then (ε, δ)-differentially private. □

Proof of Lemma 6.6. Let $n' = n + \alpha + \beta - 2$, i.e. the number of total prior and new data points. Now, consider the following process for generating θ along with data points:

1. Sample θ uniformly from $[0, 1]$.
2. Let $X = (X_1, ..., X_{n'})$ be n' i.i.d. Bern(θ) random variables.

Conditioned on the last $(n'-n)$ data points containing $(\alpha-1)$ 1's and $(\beta-1)$ 0's, this means that $(X_1, \ldots, X_n)$ is distributed binomially with a Beta(α, β) prior on the probability – the same as our model for the real data.

Because the distribution of $\theta \mid X$ is determined by the sufficient statistic θ_{obs}, the distribution $\theta \mid X = X'$ is the same as that of $\theta \mid \sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n'$.

Let $X^{(\text{obs})} = (X_1^{(\text{obs})}, \ldots, X_{n'}^{(\text{obs})})$ be an observed value of X. Then, let $\theta_{\text{obs}} = \frac{\sum_{i=1}^{n'} X_i^{(\text{obs})}}{n'}$ and $\theta_{\text{exp}} = \mathrm{E}\left[\theta \mid X = X^{(\text{obs})}\right]$. By this notation, the lemma statement is equivalent to showing that

$$\Pr\left[|\theta_{\text{exp}} - \theta| > \varepsilon \min(\theta_{\text{exp}}, 1 - \theta_{\text{exp}}) \,\middle|\, \sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n'\right] \leq \delta$$

when α, β are sufficiently large. We do this by bounding $\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n' \text{ and } |\theta - \theta_{\text{exp}}| > \min(\theta_{\text{exp}}, 1 - \theta_{\text{exp}})\right]$ and $\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n'\right]$ separately, then applying Bayes' rule.

Without loss of generality assume that $\theta_{\text{obs}} \leq \frac{1}{2}$.

Bounding the numerator $\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n' \text{ and } |\theta - \theta_{\text{exp}}| > \min(\theta_{\text{exp}}, 1 - \theta_{\text{exp}})\right]$**:**

Note that $\theta_{\text{exp}} = \frac{\theta_{\text{obs}} n' + 1}{n' + 2}$ is very close to θ_{obs}; more specifically, conditioned on $\theta_{\text{obs}} \geq \frac{\alpha}{n'} = \Omega\left(\frac{1}{\varepsilon^2 n'}\right)$,

$$\theta_{\text{exp}} - \theta_{\text{obs}} = \frac{1 - 2\theta_{\text{obs}}}{n' + 2} < \varepsilon^2 \theta_{\text{obs}} \leq \frac{\varepsilon \theta_{\text{obs}}}{4}$$

for sufficiently small ε. Hence, our assumption that $\theta_{\text{obs}} \leq \frac{1}{2}$ also implies that $\min(\theta_{\text{exp}}, 1 - \theta_{\text{exp}}) \approx \theta_{\text{exp}}$.

This means that

$$\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n' \,\middle|\, |\theta - \theta_{\text{exp}}| > \varepsilon \theta_{\text{exp}}\right] \leq \Pr\left[\sum_{i=1}^{n'} X_i = p' n' \,\middle|\, |\theta - \theta_{\text{obs}}| > \frac{\varepsilon \theta_{\text{obs}}}{2}\right].$$

By a Chernoff bound,

$$\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n' \,\middle|\, |\theta - \theta_{\text{obs}}| > \frac{\varepsilon \theta_{\text{obs}}}{2}\right] \leq 2e^{-\varepsilon^2 \theta_{\text{obs}} n'/12},$$

implying that

$$\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n' \text{ and } |\theta - \theta_{\text{exp}}| > \varepsilon \theta_{\text{exp}}\right] \leq 2e^{-\varepsilon^2 \theta_{\text{obs}} n'/12}.$$

Bounding the denominator $\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n'\right]$**:**

By Stirling's approximation, $n! = \Theta\left(n^{n+\frac{1}{2}} e^{-n}\right)$, so

$$\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n' \,\middle|\, \theta = \theta_{\text{obs}}\right] = \Theta\left(\frac{1}{\sqrt{\theta_{\text{obs}}(1-\theta_{\text{obs}}) n'}}\right) = \Omega\left(\frac{1}{\sqrt{n'}}\right).$$

However, this is true not only for $\theta = \theta_{\text{obs}}$, but also for θ close to θ_{obs}. Namely, for all $\theta \in \left[\theta_{\text{obs}}, \theta_{\text{obs}} + \frac{1}{n'}\right]$,

$$\begin{aligned}
\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n' \,\middle|\, \theta\right] &= \binom{n'}{\theta_{\text{obs}} n'} (\theta)^{\theta_{\text{obs}} n'} (1-\theta)^{(1-\theta_{\text{obs}}) n'} \\
&= \Theta\left(\frac{1}{\sqrt{\theta_{\text{obs}}(1-\theta_{\text{obs}}) n'}} \left(\frac{\theta}{\theta_{\text{obs}}}\right)^{\theta_{\text{obs}} n'} \left(\frac{1-\theta}{1-\theta_{\text{obs}}}\right)^{(1-\theta_{\text{obs}}) n'}\right) \\
&= \Omega\left(\frac{1}{\sqrt{n'}}\right),
\end{aligned}$$

where the last step follows from $\frac{\theta}{\theta_{\text{obs}}} \geq 1$, $\frac{1}{\sqrt{\theta_{\text{obs}}(1-\theta_{\text{obs}})}} \geq \frac{1}{2}$, and

$$\left(\frac{1-\theta}{1-\theta_{\text{obs}}}\right)^{(1-\theta_{\text{obs}}) n'} \geq \left(1 - \frac{2}{n'}\right)^{(1-\theta_{\text{obs}}) n'} \geq \left(1 - \frac{2}{n'}\right)^{n'} = \Omega(1).$$

Consequently, because θ was chosen uniformly from $[0, 1]$, we get

$$\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n' \text{ and } \theta \in \left[\theta_{\text{obs}}, \theta_{\text{obs}} + \frac{1}{n'}\right]\right] = \Omega\left(\frac{1}{n'\sqrt{n'}}\right) \Rightarrow \Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n'\right] = \Omega\left(\frac{1}{n'\sqrt{n'}}\right).$$

Putting it all together:

Applying Bayes' Rule and plugging in the bounds obtained thus far, we have

$$\begin{aligned}
\Pr\left[|\theta - \theta_{\text{exp}}| > \varepsilon \theta_{\text{exp}} \,\middle|\, \sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n'\right] &= \frac{\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n' \text{ and } |\theta - \theta_{\text{exp}}| > \varepsilon \theta_{\text{exp}}\right]}{\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n'\right]} \\
&= O\left({n'}^{3/2} e^{-\varepsilon^2 \theta_{\text{obs}} n'/12}\right).
\end{aligned}$$

We know that $\theta_{\text{obs}} \geq \frac{\alpha}{n'}$. Hence, for some constant c, when $\alpha, \beta \geq \frac{c}{\varepsilon^2}\left(\ln n + \ln \frac{1}{\delta}\right)$, the probability above is less than δ. □

In fact, we can use a similar technique to show that this is also an asymptotic lower bound on α and β.

Theorem 6.7. *There exists a constant c such that for sufficiently large n, if* $\textsc{Param}_{\text{Post}}$ *is (ε, δ)-differentially private, then it must be true that*

$$\alpha, \beta \geq \frac{c}{\varepsilon^2}\left(\ln \frac{\varepsilon}{\delta n}\right).$$

Because we tend to think of ε is a constant and $\delta = \frac{1}{n^{\omega(1)}}$ this is effectively the same asymptotic bound as we saw in Theorem 6.4.

Lemma 6.8. *There exists a constant c such that for sufficiently large n, when $\alpha, \beta \leq \frac{c}{\varepsilon^2}\left(\ln \frac{1}{\delta n}\right)$, then when $n_1 = 0, n_0 = n$, with probability at least δ,* $\textsc{Param}_{\text{Post}}$ *releases a value of θ outside of the range $\left[e^{-\varepsilon} \cdot \frac{n_1+\alpha}{n+\alpha+\beta}, e^{\varepsilon} \cdot \frac{n_1+\alpha}{n+\alpha+\beta}\right]$.*

This lemma is similar to Lemma 6.6, but places a lower bound on the probability that $\textsc{Param}_{\text{Post}}$ releases a value of θ far from its expectation, instead of the probability that $\textsc{Param}_{\text{Post}}$ releases a value of θ close to its expectation. The proof of the lemma proceeds along similar lines as that of Lemma 6.6, and can be found in the Appendix.

Proof of theorem. As before, note that

$$\theta_{\text{MAP}} = \frac{n_1 + \alpha - 1}{n + \alpha + \beta - 2} \approx \frac{n_1 + \alpha}{n + \alpha + \beta}.$$

Because $\textsc{Param}_{\text{Post}}$ is (ε, δ)-differentially private, by Proposition 2.11,

$$\Pr_{\alpha,\beta}\left[\theta \geq e^{2\varepsilon} p\right] \leq \Pr_{\alpha,\beta}\left[\theta \geq \frac{e^{2\varepsilon}\theta_{\text{MAP}}}{1 + (e^{2\varepsilon} - 1)\theta_{\text{MAP}}}\right] \leq \delta \frac{1 - e^{-\varepsilon}}{.}$$

This is almost the of converse of Lemma 6.5. Then, by Lemma 6.8 (using 2ε and $\frac{\delta}{1-e^{-\varepsilon}} \approx \frac{\delta}{\varepsilon}$ in place of ε and δ), there exists a constant c so that $\alpha, \beta \geq \frac{c}{\varepsilon^2}\left(\ln \frac{\varepsilon}{\delta}\right)$. □

6.2 Releasing Synthetic Data

6.2.1 ε-differential privacy

We begin doing marginally better privacy-wise once we release data points drawn from the distribution parameterized by θ rather than releasing θ directly; the sampling of synthetic data points helps smooth out the cases where the value of θ drawn lands close to 0 or 1. More specifically, we are now able to obtain ε-differential privacy, which was not possible when releasing the parameter.

Theorem 6.9. *Releasing a single synthetic data point using* $\textsc{Synth}_{\text{MAP}}$ *is ε-differentially private if and only if $\alpha, \beta \geq 2 + \frac{1}{e^{\varepsilon}-1} \approx 2 + \frac{1}{\varepsilon}$.*

Proof. Call the synthetic data point z. Then, for all neighboring X, X',

$$\frac{\Pr_{\alpha,\beta}\left[z = 1 \mid X\right]}{\Pr_{\alpha,\beta}\left[z = 1 \mid X'\right]} = \frac{\frac{n_1+\alpha-1}{n+\alpha+\beta-2}}{\frac{n_1+\alpha-2}{n+\alpha+\beta-2}} = \frac{n_1 + \alpha - 1}{n_1 + \alpha - 2} \leq \frac{\alpha - 1}{\alpha - 2}.$$

When $\alpha - 1 = \frac{e^{\varepsilon}}{e^{\varepsilon}-1}$, this simplifies to e^{ε}, so we have ε-differential privacy.

From this proof, and the cases where $(n_0, n_1) = (n, 0)$ or $(0, n)$, we see that this is also the minimum α, β with which it is still possible to have ε-differentially privacy. □

Using basic composition (Theorem 2.17) for ε-differential privacy, we obtain the following corollary:

Corollary 6.10. $\textsc{Synth}_{\text{MAP}}$ *is ε-differentially private for m synthetic data points if and only if* $\alpha, \beta \geq 2 + \frac{1}{e^{\varepsilon/m}-1} \approx 2 + \frac{m}{\varepsilon}$.

Now, we turn to the case where the choice of θ is not deterministic, as it is was in $\textsc{Synth}_{\text{MAP}}$, but instead where θ is drawn from its posterior distribution. Intuitively, compared to $\textsc{Synth}_{\text{MAP}}$, we might expect that this method of choosing θ would increase the amount of privacy provided, because it inserts another layer of randomness; however, that turns out not actually to be the case.

Theorem 6.11. *Releasing a single synthetic data point using* $\textsc{Synth}_{\text{Post}}$ *is ε-differentially private if and only if* $\alpha, \beta \geq 1 + \frac{1}{e^{\varepsilon}-1} \approx 1 + \frac{1}{\varepsilon}$.

Proof. As before, call the synthetic data point z. Because $\Pr_{\alpha,\beta}[z = 1 \mid X] = \frac{n_1+\alpha}{n+\alpha+\beta}$,

$$\frac{\Pr_{\alpha,\beta}[z = 1 \mid X]}{\Pr_{\alpha,\beta}[z = 1 \mid X']} = \frac{n_1 + \alpha}{n_1 + \alpha - 1}.$$

In fact, this is more or less the same distribution for z and ratio $\frac{\Pr_{\alpha,\beta}[z=1 \mid X]}{\Pr_{\alpha,\beta}[z=1 \mid X']}$ as we saw with $z \leftarrow \textsc{Synth}_{\text{MAP}}$, except that $(\alpha-1)$ and $(\beta-1)$ have been exchanged for α and β. Thus, by the same reasoning, $\alpha, \beta = 1 + \frac{1}{e^{\varepsilon}-1}$ is the minimum prior needed for ε-differential privacy. □

Since $\textsc{Synth}_{\text{Post (Many)}}$ consists of m repeated independent trials of $\textsc{Synth}_{\text{Post}}$, we similarly get a bound on its ε-differential privacy by basic composition on Theorem 6.11:

Corollary 6.12. $\textsc{Synth}_{\text{Post (Many)}}$ *is ε-differentially private if and only if* $\alpha, \beta \geq 1 + \frac{1}{e^{\varepsilon/m}-1} \approx 1 + \frac{m}{\varepsilon}$.

We cannot, however, use composition on $\textsc{Synth}_{\text{Post (One)}}$ because θ is only generated once. In fact, we might expect the repeated draws of θ in $\textsc{Synth}_{\text{Post (Many)}}$ to decrease the privacy offered, because they decrease variance by hedging against the possibility of getting an extremal value of θ. However, this turns out not to be the case for ε-differential privacy: $\textsc{Synth}_{\text{Post (One)}}$ and $\textsc{Synth}_{\text{Post (Many)}}$ actually have the same degree of of ε-differential privacy.

Theorem 6.13 ([MKA$^+$08])**.** $\textsc{Synth}_{\text{Post (One)}}$*is ε-differentially private if and only if* $\alpha, \beta \geq 1 + \frac{m}{e^{\varepsilon}-1} \approx \frac{m}{e^{\varepsilon}}$.

In fact, Machanavajjhala et al. show that this is true even for the more general case of a multinomial distribution, where x is allowed to take on any of a finite number of values, with its conjugate prior of a Dirichlet distribution (see Chapter 6.4 for more on this generalization).

Proof. Let m_0, m_1 be the number of 0's and 1's in the synthetic data drawn. Then,

$$\Pr_{\alpha,\beta}[m_0, m_1 \mid X] = \int_{\theta=0}^{1} \frac{\theta^{n_1+\alpha-1}(1-\theta)^{n_0+\beta-1}}{\mathrm{B}(n_1+\alpha-1, n_0+\beta-1)} \cdot \theta^{m_1}(1-\theta)^{m_0} d\theta = \frac{\mathrm{B}(n_1+\alpha+m_1, n_0+\beta+m_2)}{\mathrm{B}(n_1+\alpha, n_0+\beta)}.$$

Then,

$$\frac{\Pr_{\alpha,\beta}[m_0, m_1 \mid X]}{\Pr_{\alpha,\beta}[m_0, m_1 \mid X']} = \frac{\frac{\mathrm{B}(n_1+\alpha+m_1, n_0+\beta+m_2)}{\mathrm{Beta}(n_1+\alpha, n_0+\beta)}}{\frac{\mathrm{B}(n_1+\alpha+m_1-1, n_0+\beta+m_2+1)}{\mathrm{Beta}(n_1+\alpha-1, n_0+\beta+1)}} = \frac{n_0+\beta}{n_1+\alpha-1} \cdot \frac{n_1+\alpha+m_1-1}{n_0+\beta+m_2}.$$

This is maximized when $m_1 = m$ and $m_2 = 0$, in which case it is equal to $\frac{n_1+\alpha+m-1}{n_1+\alpha-1}$. This fraction is maximized when $n_1 = 0$ and $n_0 = n$, in which case this just becomes $\frac{\alpha+m-1}{\alpha-1}$. Then,

$$\frac{\alpha+m-1}{\alpha-1} \leq e^{\varepsilon} \Leftrightarrow \alpha \geq 1 + \frac{m}{e^{\varepsilon}-1}.$$

Again, because we calculated the probability ratio exactly, this proof also shows that this is the minimum value of α, β required for ε-differential privacy. □

In this section, we have shown that all three of our synthetic data generation methods produce the same amount of ε-differential privacy. Using the composition theorems, this means that any combination of them – in particular, any number of values drawn for θ with $\text{S}\textsc{ynth}_{\text{Post}}$ and any number of data points generated from each – provides an amount of ε-differential privacy that depends only on the total number of synthetic data points.

Unfortunately, the amount of privacy provided is not necessarily satisfactory. For small values of m – i.e. $m = o(n)$ – the influence of the prior data will become negligible for large datasets. However, if we wanted to draw a synthetic dataset of size equal to the original – i.e. $m = n$ – then we would have to have more prior data than real data, causing a significant distortion.

6.2.2 (ε, δ)-differential privacy

Moving on to (ε, δ)-differential privacy, we have three ways to bound the privacy obtained:

1. By advanced composition (Theorem 2.18) on the bounds in the previous section.
2. By post-processing on the parameter release method $\text{P}\textsc{aram}_{\text{Post}}$.
3. By working directly with the probability distribution.

We walk through these in order.

First applying advanced composition to Theorems 6.9 and 6.11 gives:

Corollary 6.14. *There exists a constant c such that Methods $\text{S}\textsc{ynth}_{\text{MAP}}$ and $\text{S}\textsc{ynth}_{\text{Post (Many)}}$ are (ε, δ)-differentially private when $\alpha, \beta \geq \frac{c\sqrt{m \ln \frac{1}{\delta}}}{\varepsilon}$.*

While this does allow us to generate synthetic data using a sublinear number of data points in n, we would still prefer to have a smaller dependence on m, if possible.

We show that that, at least in the cases where $m \leq cn$ for an appropriate choice of c, it is possible to make the dependence logarithmic in both n and $\frac{1}{\delta}$ for all of the synthetic data generation methods.

For $\text{S}\textsc{ynth}_{\text{Post (One)}}$, the best such bound we obtained follows from post-processing on Theorem 6.4:

Corollary 6.15. *There exists a constant c such that for sufficiently large n, $\text{S}\textsc{ynth}_{\text{Post (One)}}$ is (ε, δ)-differentially private when*

$$\alpha, \beta \geq \frac{c}{\varepsilon^2}\left(\ln n + \ln \frac{1}{\delta}\right),$$

regardless of the number of synthetic data points generated.

While we cannot apply post-processing to $\text{S}\textsc{ynth}_{\text{MAP}}$ and $\text{S}\textsc{ynth}_{\text{Post (Many)}}$, we prove a similar bound in these cases, albeit with some limitation on the number of synthetic data points allowed.

Theorem 6.16. *There exist constants c_1, c_2 such that for sufficiently large n,* $\textsc{Synth}_{\text{MAP}}$ *and* $\textsc{Synth}_{\text{Post (Many)}}$ *are (ε, δ)-differentially private when $m \le c_1 n$ and $\alpha, \beta \ge \frac{c_2}{\varepsilon^2} \ln \frac{1}{\delta}$.*

Of the two bounds given by Corollary 6.14 and Theorem 6.16, the latter requires lower values of α, β so long as $m = \Omega\left(\frac{\sqrt{\ln \frac{1}{\delta}}}{\varepsilon}\right)$. Thus, for example, if we want m to be linear in n and $\delta = e^{-o(n)}$, the latter gives a stronger bound.

Proof. Let Z denote the synthetic dataset, m_1 the number of 1's in it, and m_0 the number of 0's. We would like for it to be true that

$$e^{-\varepsilon} \le \frac{\Pr_{\alpha,\beta}[Z \mid X]}{\Pr_{\alpha,\beta}[Z \mid X']} = \left(\frac{n_1+\alpha}{n_1+\alpha-1}\right)^{m_1} \left(\frac{n_0+\beta}{n_0+\beta+1}\right)^{m_0} \le e^{\varepsilon}.$$

Consequently, we want to bound by δ the probability that m_0 or m_1 land outside these boundaries.

Without loss of generality, suppose $n_1 + \alpha \le n_0 + \beta$, so that $\frac{n_1+\alpha}{n+\alpha+\beta} \le \frac{1}{2}$.

Let $\varepsilon' = \frac{\varepsilon}{c'}$ for a reasonably large constant c'. Then by a Chernoff bound,

$$\Pr_{\alpha,\beta}\left[\left|\frac{m_1}{m} - \frac{n_1+\alpha}{n+\alpha+\beta}\right| > \varepsilon' \frac{n_1+\alpha}{n+\alpha+\beta} \mid X\right] \le 2e^{-m\left(\frac{n_1+\alpha}{n+\alpha+\beta}\right)\varepsilon'^2/6},$$

so it is less than δ for $m \ge \frac{c'(n+\alpha+\beta)}{\varepsilon'^2 \alpha} \ln \frac{1}{\delta}$ and some constant c'.

Then, assuming that this holds,

$$\begin{aligned}\left(\frac{n_1+\alpha}{n_1+\alpha-1}\right)^{m_1} \left(\frac{n_0+\beta}{n_0+\beta+1}\right)^{m_0} &\le \left(\frac{n_1+\alpha}{n_1+\alpha-1}\right)^{m(1+\varepsilon')\frac{n_1+\alpha}{n+\alpha+\beta}} \left(\frac{n_0+\beta}{n_0+\beta+1}\right)^{m(1-\varepsilon')\frac{n_0+\beta}{n+\alpha+\beta}} \\ &= \left(\left(\frac{n_1+\alpha}{n_1+\alpha-1}\right)^{(1+\varepsilon')(n_1+\alpha)} \left(\frac{n_0+\beta}{n_0+\beta+1}\right)^{(1-\varepsilon')(n_1+\beta)}\right)^{\frac{m}{n+\alpha+\beta}}.\end{aligned}$$

Now, note that $\left(1-\frac{1}{k}\right)^k = \frac{1}{e}\left(1-O\left(\frac{1}{k}\right)\right)$ and $\left(1+\frac{1}{k}\right)^k = e\left(1-O\left(\frac{1}{k}\right)\right)$. Plugging this in to the previous equation gives

$$\left(\frac{n_1+\alpha}{n_1+\alpha-1}\right)^{(1+\varepsilon')(n_1+\alpha-1)} \left(\frac{n_0+\beta}{n_0+\beta+1}\right)^{(1-\varepsilon')(n_1+\beta+1)} = e^{O(\varepsilon')},$$

so

$$\left(\frac{n_1+\alpha}{n_1+\alpha-1}\right)^{m_1} \left(\frac{n_0+\beta}{n_0+\beta+1}\right)^{m_0} = \left(e^{O(\varepsilon')} \left(\frac{n_1+\alpha}{n_1+\alpha-1}\right)^{(1+\varepsilon')} \left(\frac{n_0+\beta}{n_0+\beta+1}\right)^{-(1-\varepsilon')}\right)^{\frac{m}{n+\alpha+\beta}}.$$

Since we are interested in $\frac{m}{n+\alpha+\beta} \le 1$, all the terms except the $e^{O(\varepsilon')}$ converge to 1 as $\alpha, \beta \to \infty$, so this entire value is just $e^{O(\varepsilon')}$.

For the other direction,

$$\begin{aligned}\left(\frac{n_1+\alpha}{n_1+\alpha-1}\right)^{m_1} \left(\frac{n_0+\beta}{n_0+\beta+1}\right)^{m_0} &\ge \left(\left(\frac{n_1+\alpha}{n_1+\alpha-1}\right)^{(1-\varepsilon')(n_1+\alpha)} \left(\frac{n_0+\beta}{n_0+\beta+1}\right)^{(1+\varepsilon')(n_1+\beta)}\right)^{\frac{m}{n+\alpha+\beta}} \\ &> \left(\left(\frac{n_1+\alpha}{n_1+\alpha-1}\right)^{(1-\varepsilon')(n_1+\alpha-1)} \left(\frac{n_0+\beta}{n_0+\beta+1}\right)^{(1+\varepsilon')(n_1+\beta+1)}\right)^{\frac{m}{n+\alpha+\beta}} \\ &\ge \left(e^{O(\varepsilon')}\left(1-O\left(\frac{1}{\alpha+\beta}\right)\right)\right)^{\frac{m}{n+\alpha+\beta}},\end{aligned}$$

which is at least $e^{-\varepsilon}$ when $\alpha, \beta \geq \frac{c'' m}{\varepsilon(n+\alpha+\beta)}$ for some constant c''.

If we set $\alpha, \beta = \frac{c_2}{\varepsilon^2} \ln \frac{1}{\delta}$ and $m = c_1 n \approx c_1(n + \alpha + \beta)$, then both of the required inequalities are satisfied so long as $\frac{c_1}{c_2} \leq c'$ and $\frac{c_1}{c_2} \leq \frac{1}{\varepsilon} \ln \frac{1}{\delta}$. □

Thus, we have shown that for each of the synthetic data generation methods, it is possible to release a linear number of synthetic data points in n with relatively small priors and (ε, δ)-differential privacy.

6.3 Average Case Differential Privacy

Now, we will show that in cases where we failed to obtain ε-differential privacy or (ε, δ)-differential privacy, the situation does not improve when we turn to average case differential privacy. In particular, we do not obtain meaningful $(\varepsilon, \delta, \Delta)$-differential privacy for $\textsc{Param}_{\text{MAP}}$ or $(\varepsilon, 0, \Delta)$-differential privacy for $\textsc{Param}_{\text{Post}}$.

Our model assumes that each of the rows of X is generated i.i.d, and that each row can only take on a the values $\{0, 1\}$. This lets us give a sufficient condition for $(\varepsilon, \delta, \Delta)$-differential privacy which is more similar to that from differential privacy, and with which it is a bit easier to work.

Proposition 6.17. *Let Δ be the set of distributions on X where each row is generated i.i.d from the same Bernoulli distribution. Let $\mathcal{A}$ be an algorithm depending only on $\sum_{i=1}^n X_i$. If for all n_1, n_1' with $|n_1 - n_1'| = 1$,*

$$\Pr_{\mathcal{D}}\left[\sum_{i=1}^n (X_i) = n_1\right] \leq e^{\varepsilon} \cdot \Pr_{\mathcal{D}}\left[\sum_{i=1}^n (X_i) = n_1'\right] + 2\delta,$$

then $\mathcal{A}$ is $(\varepsilon, \delta, \Delta)$-differentially private.

Proof. Let Sim be a simulator that randomly fills in the missing row X_i by choosing uniformly from $\{0, 1\}$, then runs $\mathcal{A}$ on the resulting database. Then, for all values $x \in \{0, 1\}$,

$$\begin{aligned}
\Pr_{\mathcal{D}}[\text{Sim}(X_{-i}) \in S \mid X_i = x] &= \frac{1}{2} \cdot \Pr_{\mathcal{D}}[\mathcal{A}(X) \in S \mid X_i = x] + \frac{1}{2} \cdot \Pr_{\mathcal{D}}[\mathcal{A}(X) \in S \mid X_i \neq x] \\
&\leq \frac{1}{2} \cdot \Pr_{\mathcal{D}}[\mathcal{A}(X) \in S \mid X_i = x] + \frac{1}{2} \cdot (e^{\varepsilon} \Pr_{\mathcal{D}}[\mathcal{A}(X) \in S \mid X_i = x] + 2\delta) \\
&= \frac{1 + e^{\varepsilon}}{2} \cdot \Pr_{\mathcal{D}}[\mathcal{A}(X) \in S \mid X_i = x] + \delta \\
&\leq e^{\varepsilon} \cdot \Pr_{\mathcal{D}}[\mathcal{A}(X) \in S \mid X_i = x] + \delta.
\end{aligned}$$

Similarly, for the other direction,

$$\begin{aligned}
e^{\varepsilon} \cdot \Pr_{\mathcal{D}}[\text{Sim}(X_{-i}) \in S \mid X_i = x] + \delta &= \frac{e^{\varepsilon}}{2} \cdot \Pr_{\mathcal{D}}[\mathcal{A}(X) \in S \mid X_i = x] + \frac{e^{\varepsilon}}{2} \cdot \Pr_{\mathcal{D}}[\mathcal{A}(X) \in S \mid X_i \neq x] + \delta \\
&\geq \frac{e^{\varepsilon}}{2} \cdot \Pr_{\mathcal{D}}[\mathcal{A}(X) \in S \mid X_i = x] + \frac{1}{2} \cdot \Pr_{\mathcal{D}}[\mathcal{A}(X) \in S \mid X_i = x] \\
&\geq \cdot \Pr_{\mathcal{D}}[\mathcal{A}(X) \in S \mid X_i = x].
\end{aligned}$$

□

For the remainder of this section, we let Δ be this set – i.e. all the distributions where each row of the database is chosen i.i.d. from a Bernoulli distribution. Then, the combination of Propositions 2.15 and 6.17 reduce establishing $(\varepsilon, \delta, \Delta)$-differential privacy to a similar condition to that of regular differential privacy; although we do not use Proposition 6.17 later one, we have included it above to show both sides of the implication.

Theorem 6.18. $\textsc{Param}_{\text{MAP}}$ *is not* $(\varepsilon, 0, \Delta)$*-differentially private.*

Proof. Because $\textsc{Param}_{\text{MAP}}$ is a deterministic function of $\sum_{i=1}^{n} X_i$, Proposition 2.15 implies that if $\textsc{Param}_{\text{MAP}}$ is $(\varepsilon, \delta, \Delta)$-differentially private, then for any $\mathcal{D} \in \Delta$,

$$\Pr_{\mathcal{D}}\left[\sum_{i=1}^{n} X_i = n_1\right] \leq e^{2\varepsilon} \cdot \Pr_{\mathcal{D}}\left[\sum_{i=1}^{n} X_i = n_1 + 1\right].$$

Let $\mathcal{D}$ be the distribution where each row is 1 with probability θ_{Real}. Then, we know that

$$\Pr_{\mathcal{D}}\left[\sum_{i=1}^{n} X_i = s\right] = \binom{n}{n_1} \theta_{\text{Real}}^{n_1} (1 - \theta_{\text{Real}})^{n - n_1},$$

so

$$\frac{\Pr_{\mathcal{D}}\left[\sum_{i=1}^{n} X_i = n_1\right]}{\Pr_{\mathcal{D}}\left[\sum_{i=1}^{n} X_i = n_1 + 1\right]} = \frac{n_1 + 1}{n - n_1} \cdot \frac{\theta_{\text{Real}}}{1 - \theta_{\text{Real}}},$$

which is an unbounded value as $\theta_{\text{Real}} \to 1$. □

In the previous proof, even if we try to restrict the possible distributions – say by restricting Δ so that $\theta_{\text{Real}}, 1 - \theta_{\text{Real}} \geq c$ for some constant c – the other term can still become arbitrarily small as n grows large. Consequently, even such a restriction would not give us asymptotic $(\varepsilon, 0, \Delta)$-differential privacy.

Corollary 6.19. *For sufficiently large n,* $\textsc{Param}_{\text{MAP}}$ *is not* $(\varepsilon, \delta, \Delta)$*-differentially private when*

$$\delta(1 + e^{\varepsilon}) = o\left(\frac{1 - e^{2\varepsilon}}{\sqrt{n}}(1 + e^{\varepsilon})\right).$$

Proof. As we noted in the proof of Lemma 6.6,

$$\Pr_{\mathcal{D}}\left[\sum_{i=1}^{n} \mathcal{X}_i = \theta_{\text{Real}} n\right] = \Omega\left(\frac{1}{\sqrt{n}}\right).$$

Consequently, letting X_{-1} be the last $n - 1$ rows of X,

$$\begin{aligned}
&\Pr_{\mathcal{D}}\left[\sum_{i=1}^{n} X_i = \theta_{\text{Real}}(n-1) - 1 \,\middle|\, X_1 = 0\right] - e^{2\varepsilon} \cdot \Pr_{\mathcal{D}}\left[\sum_{i=1}^{n} X_i = \theta_{\text{Real}}(n-1) - 1 \,\middle|\, X_1 = 1\right] \\
&= \Pr_{\mathcal{D}}\left[\sum_{i=2}^{n} X_{-1,i} = \theta_{\text{Real}}(n-1)\right] - e^{2\varepsilon} \cdot \Pr_{\mathcal{D}}\left[\sum_{i=2}^{n} X_{-1,i} = \theta_{\text{Real}}(n-1) - 1\right] \\
&= \Omega\left(\frac{1}{\sqrt{n-1}}\right)\left(1 - \frac{\binom{n-1}{\theta_{\text{Real}}(n-1)-1} \theta_{\text{Real}}^{\theta_{\text{Real}}(n-1)-1} (1 - \theta_{\text{Real}})^{(1-\theta_{\text{Real}})(n-1)+1}}{\binom{n-1}{\theta_{\text{Real}}(n-1)} \theta_{\text{Real}}^{\theta_{\text{Real}}(n-1)} (1 - \theta_{\text{Real}})^{(1-\theta_{\text{Real}})(n-1)}}\right), \\
&= \Omega\left(\frac{1}{\sqrt{n-1}}\right)\left(1 - e^{2\varepsilon} \cdot \frac{\theta_{\text{Real}}(n-1)}{(n-1)(1 - \theta_{\text{Real}}) + 1} \cdot \frac{1 - \theta_{\text{Real}}}{\theta_{\text{Real}}}\right) \\
&= \Omega\left(\frac{1 - e^{2\varepsilon}}{\sqrt{n-1}}\right).
\end{aligned}$$

Consequently, for sufficiently large n, this difference is greater than $\delta(1 + e^{\varepsilon})$, implying that $\textsc{Param}_{\text{MAP}}$ is not $(\varepsilon, \delta, \Delta)$-differentially private by Proposition 2.15. □

Because we expect at the least that $\delta = o\left(\frac{1}{n}\right)$, this means that $\textsc{Param}_{\text{MAP}}$ is not meaningfully $(\varepsilon, \delta, \Delta)$-differentially private. Similarly, we can prove that $(\varepsilon, 0, \Delta)$-differential privacy does not hold when it comes to $\textsc{Param}_{\text{Post}}$:

Theorem 6.20. $\textsc{Param}_{\text{Post}}$ *is not* $(\varepsilon, 0, \Delta)$*-differentially private.*

Proof. We have

$$\begin{aligned}
\frac{p_{\mathcal{D},\alpha,\beta}\left(\theta \mid X_1 = 1\right)}{p_{\mathcal{D},\alpha,\beta}\left(\theta \mid X_1 = 0\right)} &= \frac{\sum_{n_1=0}^{n-1}\binom{n-1}{n_1}\theta_{\text{Real}}^{n_1}(1-\theta_{\text{Real}})^{n-n_1-1}\frac{\theta^{\alpha+n_1}(1-\theta)^{\beta+n-n_1-2}(\alpha+\beta+n-1)!}{(\alpha+n_1)!(\beta+n-n_1-2)!}}{\sum_{n_1=0}^{n-1}\binom{n-1}{n_1}\theta_{\text{Real}}^{n_1}(1-\theta_{\text{Real}})^{n-n_1-2}\frac{\theta^{\alpha+n_1-1}(1-\theta)^{\beta+n-n_1-1}(\alpha+\beta+n-1)!}{(\alpha+n_1-1)!(\beta+n-n_1-1)!}} \\
&= \frac{\sum_{n_1=0}^{n-1}\binom{n-1}{n_1}\theta_{\text{Real}}^{n_1}(1-\theta_{\text{Real}})^{n-n_1-1}\binom{n+\alpha+\beta-2}{n_1+\alpha-1}\theta^{\alpha+n_1-1}(1-\theta)^{\beta+n-n_1-1}\cdot\frac{\theta}{(1-\theta)}\cdot\frac{\beta+n-n_1-1}{\alpha+n_1}}{\sum_{n_1=0}^{n-1}\binom{n-1}{n_1}\theta_{\text{Real}}^{n_1}(1-\theta_{\text{Real}})^{n-n_1-1}\binom{n+\alpha+\beta-2}{n_1+\alpha-1}\theta^{\alpha+n_1-1}(1-\theta)^{\beta+n-n_1-1}} \\
&\geq \frac{\theta}{1-\theta}\cdot\frac{\beta}{\alpha+n}\cdot\frac{\sum_{n_1=0}^{n-1}\binom{n-1}{n_1}\theta_{\text{Real}}^{n_1}(1-\theta_{\text{Real}})^{n-n_1-1}\binom{n+\alpha+\beta-2}{n_1+\alpha-1}\theta^{\alpha+n_1-1}(1-\theta)^{\beta+n-n_1-1}}{\sum_{n_1=0}^{n-1}\binom{n-1}{n_1}\theta_{\text{Real}}^{n_1}(1-\theta_{\text{Real}})^{n-n_1-1}\binom{n+\alpha+\beta-2}{n_1+\alpha-1}\theta^{\alpha+n_1-1}(1-\theta)^{\beta+n-n_1-1}} \\
&= \frac{\theta}{1-\theta}\cdot\frac{\beta}{\alpha+n},
\end{aligned}$$

which is unbounded, so we are done by Proposition 2.15. □

6.4 The Multinomial Distribution

Finally, we can interpret several of the synthetic data generation methods as instantiations of the exponential mechanism (Theorem 2.21). This enables us to generalize the ε-differential privacy upper-bounds from Bernoulli-distributed variables to categorically-distributed variables with less algebra on the distribution functions.

More specifically, a categorical random variable C is one that can take on a finite number of values $v_1, v_2, \ldots, v_k$ with $\Pr\left[C = v_i\right] = \theta_i$, such that $\sum_{i=1}^{k}\theta_i = 1$. Thus, a Bernoulli random variable is just a categorical random variable where $k = 2$.

The conjugate prior for the multinomial distribution – the combination of n i.i.d. categorical random variables – is the Dirichlet distribution $\text{Dir}(\alpha_1, \ldots, \alpha_k)$ over $[0, 1]^k$, which has probability density

$$p_{\alpha_1,\ldots,\alpha_k}\left(\theta_1, \ldots, \theta_k\right) \propto \prod_{i=1}^{k}\theta_i^{\alpha_i - 1}.$$

Thus, when $k = 2$, each coordinate of the Dirichlet distribution takes on a Beta distribution.

Similarly to the Bernoulli case, the prior can be interpreted as us having already seen $(a_i - 1)$ values of v_i, for each i. The posterior distribution, if v_i occurs n_i times in X, is given by

$$\theta \mid X \propto \text{Dir}(\alpha_1 + n_1, \ldots, \alpha_1 + n_k).$$

Now, we can look at the actual synthetic data generation methods.

For $\textsc{Synth}_{\text{MAP}}$, recall that we proved in Theorem 6.9 that $\textsc{Synth}_{\text{MAP}}$ is ε-differentially private for a single data point z if and only if $\alpha, \beta \geq 2 + \frac{1}{e^{\varepsilon}-1}$. To view this in terms of the exponential mechanism, we use the utility function

$$-q(X, z) = \begin{cases} \ln\frac{n_1+\alpha-1}{n+\alpha+\beta-2} & z = 1 \\ \ln\frac{n_0+\beta-1}{n+\alpha+\beta-2} & z = 0 \end{cases},$$

i.e. so that $z = 1$ is output with probability $e^{-q(X,z)}$. Then, we can bound the sensitivity by $\ln \frac{\min(\alpha,\beta)}{\min(\alpha,\beta)-1}$. This means that $\textsc{Synth}_{\text{MAP}}$ is ε-differentially private so long as $\text{Sen}(q) \leq \frac{\varepsilon}{2}$, i.e. $\alpha, \beta \geq 1 + \frac{1}{e^{\varepsilon/2}-1}$. This is asymptotically the same as the exact bound, and is larger by approximately a constant factor of 2.

For multiple synthetic data points, it suffices to take the utility function q to be the sum of the function above over the individual data points; again, this gives us the same asymptotic bound, but with an extra constant factor of 2.

Extending the utility function to

$$-q(X,z) = \begin{cases} \ln \frac{n_i+\alpha_i-1}{n+\sum a_i - k} & z = i \end{cases},$$

we see the following:

Corollary 6.21. *For the categorical distribution with a Dirichlet prior,* $\textsc{Synth}_{\text{MAP}}$ *is ε-differentially private when $\alpha_1, \ldots, \alpha_k \geq 1 + \frac{1}{e^{\varepsilon/(2m)}-1}$.*

Because the posterior distributions for a single synthetic data point from $\textsc{Synth}_{\text{MAP}}$ and one from $\textsc{Synth}_{\text{Post (Many)}}$ are roughly the same, we similarly have:

Corollary 6.22. *For the categorical distribution with a Dirichlet prior,* $\textsc{Synth}_{\text{Post (Many)}}$ *is ε-differentially private when $\alpha_1, \ldots, \alpha_k \geq 1 + \frac{1}{e^{\varepsilon/(2m)}-1}$.*

Combined with Machanavajjhala et al.'s result that $\textsc{Synth}_{\text{Post (One)}}$ is ε-differentially private under these conditions on $\alpha_1, \ldots, \alpha_k$ [MKA$^+$08], we again have that all three of synthetic data generation methods are ε-differential privacy when $\alpha_1, \ldots, \alpha_k \geq \frac{2m}{\epsilon}$.

Chapter 7: The Gaussian Distribution with Known Variance

Now, we move on to the Gaussian distribution. Recall from Chapter 3.3 that:

- We focus on the Gaussian with $\mathcal{X} = \mathbb{R}$, unknown mean μ, and known variance σ^2.
- Its conjugate prior is another normal distribution with mean $\mu_0 = 0$ and variance σ_0^2.

However, we find that no differential privacy is achievable if $\mathcal{X} = \mathbb{R}$: changing a single data point to an extreme value allows us to move the probability distributions $\theta \mid X$ and $\theta \mid X'$ arbitrarily far apart.

Thus, to achieve any privacy we instead restrict our data points to lie in $[-R, R]$ – i.e. $\mathcal{X} = [-R, R]$. While this alone is also not quite sufficient for privacy, also conditioning our parameter choices for μ and synthetic data draws for z to lie in this range allows us to attain privacy in some cases.

Table 7.1 summarizes the bounds needed to achieve privacy when the appropriate conditions are met.

Method	Privacy	Achievable?	Lower Bound		Upper Bound
			m large	m small	
$\textsc{Param}_{\text{MAP}}$	ε or (ε, δ)	No	n/a		n/a
$\textsc{Param}_{\text{Post}}$	ε	Yes	$\Theta\left(\sqrt{\varepsilon}\right)$		
	(ε, δ)	Yes	$\Omega\left(\frac{\varepsilon\sqrt{n}}{\sqrt{\ln\frac{1}{\delta}}}\right)$		
$\textsc{Synth}_{\text{MAP}}$	ε	Yes	$\Theta\left(\sqrt{\frac{\varepsilon n}{m}}\right)$		
	(ε, δ)	Yes	$\Omega\left(\sqrt{\frac{\varepsilon n}{\sqrt{m\ln\frac{1}{\delta}}}}\right)$	$\Omega\left(\frac{\varepsilon n}{m\sqrt{\ln\frac{m}{\delta}}}\right)$	
$\textsc{Synth}_{\text{Post (One)}}$	ε	Yes	$\Omega\left(\sqrt{\varepsilon}\right)$		
	(ε, δ)	Yes	$\Omega\left(\frac{\varepsilon\sqrt{n}}{\sqrt{\ln\frac{1}{\delta}}}\right)$		
$\textsc{Synth}_{\text{Post (Many)}}$	ε	Yes	$\Theta\left(\sqrt{\frac{\varepsilon n}{m}}\right)$		
	(ε, δ)	Yes	$\Omega\left(\sqrt{\frac{\varepsilon n}{\sqrt{m\ln\frac{1}{\delta}}}}\right)$	$\Omega\left(\frac{\varepsilon n}{m\sqrt{\ln\frac{m}{\delta}}}\right)$	

Figure 7.1: The bounds for the Gaussian distribution that we prove in this chapter. The bounds refer to those on T, where $T(n, \varepsilon, \delta)$ is the threshhold such that our methods are (ε, δ)-differentially private if and only if $\frac{R}{\sigma} \leq T$. Lower bounds generally refer to a sufficient small value of $\frac{R}{\sigma}$ for privacy, while upper bounds refer to a necessary condition on $\frac{R}{\sigma}$. When privacy is achievable, we require the restriction that $\mathcal{X} = [-R, R]$ in all cases. For ε-differential privacy, we also require either that whichever is output of the parameter μ or the synthetic data points z also lie in $[-R, R]$.

As the table shows, the amount of privacy obtained can be thought of as a function of $\frac{R}{\sigma}$. If we think of the variance σ^2 as a constant, it makes intuitive sense that smaller values of R should lead to more privacy, since that removes all the low-probability outliers.

Unlike in the case of the Bernoulli, however, we cannot interpret the bounds in the table in terms of the strength of the prior on μ. On the whole, the prior is stronger when its variance σ_0^2 is small relative to σ^2; however, apart from having some effect on the convergence rate in the our asymptotic results, the actual impact of the choice of σ_0 appears to be relatively negligible.

Our Results

> **We show that except for Param$_{\text{MAP}}$, all of our methods can achieve both ε-differential privacy and (ε, δ)-differential privacy...**

To achieve (ε, δ)-differential privacy, it is sufficient for us to restrict the data points to $[-R, R]$, because the probability that the value(s) of μ or z output lies far outside that range grows exponentially small in the distance. However, to achieve ε-differential privacy, we also need the condition that whichever of μ and z is output also lies in $[-R, R]$.

> **... but while the conditions required for (ε, δ)-differential privacy are reasonable...**

We know the following general fact about the Gaussian distribution:

Proposition 7.1. *Let $x_1, \ldots, x_n$ be n i.i.d. random variables drawn from a Gaussian distribution with mean μ and variance σ^2. For any $\delta < 1$, there exists a constant c_δ such that with probability at least $1 - \delta$, all of $x_1, \ldots, x_n$ lie in the range $[\mu - c\sigma\sqrt{\ln n}, \mu + c\sigma\sqrt{\ln n}]$.*

Thus, to avoid cutting off data points with our restriction, we would prefer to be able to allow $\frac{R}{\sigma} = \Omega\left(\sqrt{n}\right)$. This is true for all of our (ε, δ)-differential privacy bounds:

- For Param$_{\text{Post}}$, and Synth$_{\text{Post (Many)}}$, we show it is sufficient to take $\frac{R}{\sigma} = O\left(\frac{\varepsilon\sqrt{n}}{\sqrt{\ln\frac{1}{\delta}}}\right)$, which gives a $\sqrt{n} = \omega(\sqrt{\ln n})$ dependence.
- For Synth$_{\text{MAP}}$ and Synth$_{\text{Post (One)}}$, we show that $\frac{R}{\sigma} = O\left(\sqrt{\frac{\varepsilon n}{\sqrt{m \ln\frac{1}{\delta}}}}\right)$ suffices. This too still fulfills the requirement that $\frac{R}{\sigma}$ is allowed to be $\Omega(\sqrt{n})$ when $m = n$. However, when m is small we show a different, stronger bound of $O\left(\frac{\varepsilon e}{m\sqrt{\ln\frac{m}{\delta}}}\right)$.

> **...those for ε-differential privacy are not.**

All of our bounds, for both parameter and synthetic data release, require $\frac{R}{\sigma} = O\left(\sqrt{\varepsilon}\right)$ if $m = n$. Consequently, the only way to meet this bound would be to either remove or alter a substantial number of data points, something that we would prefer not to do. In all cases except for Synth$_{\text{Post (One)}}$, we actually show that this is an asymptotically optimal bound. Thus, ε-differential privacy may not be meaningfully achievable in practice except perhaps when releasing a very small number of synthetic data points.

7.1 Releasing the Parameter

7.1.1 Param$_{\text{MAP}}$

In general, μ_{MAP} is deterministic but non-constant given the data, so again we have:

Theorem 7.2. $\textsc{Param}_{\text{MAP}}$ *is not* (ε,δ)*-differentially private for any choice of prior, even when* $\mathcal{X}$ *(and consequently also* θ_{MAP}*) is restricted to* $[-R,R]$.

As with θ_{MAP} in the Bernoulli distribution, however, we can release μ_{MAP} using Laplace noise. In this case, the restriction $\mathcal{X}=[-R,R]$ becomes necessary to obtain a finite sensitivity $\text{Sen}(\textsc{Param}_{\text{MAP}})=\frac{R}{n}$.

Corollary 7.3. *Releasing* $\textsc{Param}_{\text{MAP}}+\text{Lap}\left(\frac{R}{n\varepsilon}\right)$ *is* ε*-differentially private when* $\mathcal{X}$ *is restricted to* $[-R,R]$.

This has variance $\frac{2R^2}{n^2\varepsilon^2}$, compared to a variance for μ_{MAP} of $\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0}^2}$ for X if we think of it as a random database drawn from a Gaussian distribution. As the latter variance is approximately $\frac{\sigma^2}{n}$ for large n, the extra variance from the Laplace noise begins to be negligible when $\frac{R}{\sigma}=o\left(\varepsilon\sqrt{n}\right)$. Since this is $\omega(\sqrt{\ln n})$, like in the Bernoulli case, the Laplace mechanism appears to be a reasonable way of releasing the parameter.

7.1.2 Param$_{\text{Post}}$

However, unlike when we dealt with the restricted range of Bernoulli random variables, the fact that our data points can be anywhere in $\mathbb{R}$ prevents privacy even for $\textsc{Param}_{\text{Post}}$. In particular, allowing the data points to be unbounded makes it possible for the change in a single data point to overwhelm all the rest of the data – for example, consider a dataset that is clustered around 0 compared to one with a single large positive value.

Theorem 7.4. $\textsc{Param}_{\text{Post}}$ *is not* (ε,δ)*-differentially private for any choice of priors if* $\mathcal{X}=\mathbb{R}$.

Proof. If X, X' are neighboring datasets with average values $\overline{x},\overline{x}'$, respectively, then

$$\begin{aligned}
\ln\frac{p_{\sigma,\sigma_0}\left(\mu\,|X\right)}{p_{\sigma,\sigma_0}\left(\mu\,|X'\right)} &= \ln\frac{p_{\sigma,\sigma_0}\left(\mu\,|\overline{x}\right)}{p_{\sigma,\sigma_0}\left(\mu\,|\overline{x}'\right)}\\
&= \frac{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}{2}\left(\left(\mu-\frac{\sigma_0^2\overline{x}}{\frac{\sigma^2}{n}+\sigma_0^2}\right)^2-\left(\mu-\frac{\sigma_0^2\overline{x}'}{\frac{\sigma^2}{n}+\sigma_0^2}\right)^2\right)\\
&= \frac{\sigma_0^2\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}{2\left(\frac{\sigma^2}{n}+\sigma_0^2\right)}\left(\overline{x}-\overline{x}'\right)\left(2\mu-\frac{\sigma_0^2\left(\overline{x}+\overline{x}'\right)}{\frac{\sigma^2}{n}+\sigma_0^2}\right).
\end{aligned}$$

Because the data points are unbounded, if we consider any interval for μ (say, $[0,1]$) then so are both $(\overline{x}-\overline{x}')$ and $(\overline{x}+\overline{x}')$, so we cannot obtain ε-differential privacy.

Additionally, $(\overline{x}-\overline{x}')$ can be unbounded regardless of the value of μ that we are examining, so we cannot obtain (ε,δ)-differential privacy. More specifically, for any ε,δ (with $\delta<1$) and a fixed X, let U be such that[1]

$$\Pr_{\sigma,\sigma_0}\left[-U\le\mu\le U\,|X\right]>\frac{\delta}{1-e^{\varepsilon}}.$$

Then, we can change a single data point to make $\overline{x}'$ negative enough such that for all $\mu\le U$, $\ln\frac{p_{\sigma,\sigma_0}(\mu|\overline{x})}{p_{\sigma,\sigma_0}(\mu|\overline{x}')}>2\varepsilon$. By Proposition 2.11, the method is then not (ε,δ)-differentially private. □

[1] This is approximately $\frac{\delta}{\varepsilon}$, and can be greater than 1. If this is the case, then we instead set U such that $\Pr_{\sigma,\sigma_0}\left[-U\le\mu\le U\,|X\right]>\frac{\delta}{1-e^{-c\varepsilon}}\approx\frac{\delta}{c\varepsilon}$ for some c where the right-hand side is less than 1, and use $(c+1)\varepsilon$ in place of 2ε in the next line.

To simplify later calculations, let us denote the variance of $\mu \mid X$ by $\sigma_\mu^2 = \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}$ and the variance of $\overline{x}$ by $\sigma_{\overline{x}}^2 = \frac{\sigma^2}{n} + \sigma_0^2$. As $n \to \infty$, we have that $\sigma_\mu^2 \to \frac{\sigma^2}{n}$ and $\sigma_{\overline{x}}^2 \to \sigma_0^2$. The calculations from the previous proof can be rewritten as:

Lemma 7.5. *For any neighboring datasets X, X',*

$$\ln \frac{p_{\sigma,\sigma_0}\left(\mu \left| \overline{x} \right.\right)}{p_{\sigma,\sigma_0}\left(\mu \left| \overline{x}' \right.\right)} = \frac{n}{2\sigma^2}\left(\overline{x} - \overline{x}'\right)\left(2\mu - \frac{\sigma_0^2\left(\overline{x} + \overline{x}'\right)}{\sigma_{\overline{x}}^2}\right). \tag{*}$$

We have seen that the primary obstacle to bounding this value comes from the data points being unbounded, so we address this by restricting our dataset to lie in an interval $[-R, R]$ (for some real number R).

However, this also turns out to be insufficient by itself: as long as μ is unbounded, so is $\left(2\mu - \frac{\sigma_0^2\left(\overline{x}+\overline{x}'\right)}{\sigma_{\overline{x}}^2}\right)$.

Corollary 7.6. *Even restricting $\mathcal{X} = [-R, R]$, $\textsc{Param}_{\mathrm{Post}}$ is still not ε-differentially private.*

But, once we have restricted $\mathcal{X} = [-R, R]$, we can make one further restriction: if we already know that the data points lie in $[-R, R]$, then it is fairly meaningless to output a μ outside that range. Consequently, instead of drawing $\mu \mid X$ directly from the normal distribution, we can instead draw μ conditioned on it lying in $[-R, R]$ (but otherwise with the same relative probability densities as before). This makes it possible to obtain ε-differential privacy:

Theorem 7.7. *When both $\mathcal{X} = [-R, R]$ and μ is conditioned to lie in $[-R, R]$, $\textsc{Param}_{\mathrm{Post}}$ is ε-differentially private when $\frac{R}{\sigma} \leq \sqrt{\varepsilon}/2$. Meanwhile, $\textsc{Param}_{\mathrm{Post}}$ is not ε-differentially private if $\frac{R}{\sigma} \geq \sqrt{\varepsilon}$.*

Proof. We can individually bound the terms of $\ln \frac{p_{\sigma,\sigma_0}(\mu | X\,)}{p_{\sigma,\sigma_0}(\mu | X'\,)}$ as

$$\left|\overline{x} - \overline{x}'\right| \leq \frac{2R}{n}, \left|2\mu - \frac{\sigma_0^2\left(\overline{x}+\overline{x}'\right)}{\frac{\sigma^2}{n}+\sigma_0^2}\right| \leq 4R,$$

so combining and plugging into (*) we obtain

$$\left|\ln \frac{p_{\sigma,\sigma_0}\left(\mu \left| X \right.\right)}{p_{\sigma,\sigma_0}\left(\mu \left| X' \right.\right)}\right| \leq \frac{4R^2}{\sigma^2}.$$

The result follows from setting this to be less than ε. Up to constant factors, all the values in the analysis are achievable (by setting $\mu = -R, \overline{x} = R, \overline{x'} = R - \frac{2R}{n}$), so this is in fact also a lower bound. □

Because ε tends to be small, this means that we need all the data points to lie within less than a single standard deviation – an unrealistic expectation. As a result, although ε-differential privacy is theoretically possible for $\textsc{Param}_{\mathrm{Post}}$, it is unsatisfactory in practice.

However, we now know that the probability that μ lies far away from $\overline{x}, \overline{x}'$ is exponentially small in the number of standard deviations, which we can use to achieve (ε, δ)-differential privacy with a looser restriction on $\frac{R}{\sigma}$. In this case, it is no longer necessary to condition μ.

Theorem 7.8. *So long as $\left(\ln\frac{1}{\delta}\right)^2 \geq \frac{1}{2\varepsilon}$, when $\mathcal{X} = [-R, R]$, $\textsc{Param}_{\mathrm{Post}}$ is (ε, δ)-differentially private when*

$$\frac{R}{\sigma} \leq \frac{\varepsilon\sqrt{n}}{\sqrt{2\ln\frac{1}{\delta}}}.$$

Lemma 7.9 (Gaussian tail bound)**.** *If $k \geq \max\left(\sqrt{2\ln\frac{1}{\delta}}, \sqrt{\frac{\pi}{2}}\right)$, then*

$$\Pr\left[|\mu - \mathrm{E}\left[\mu \mid \overline{x}\right]| \geq k\sigma_\mu \mid \overline{x}\right] \leq \delta.$$

Proof of theorem. Let $k = \sqrt{2\ln\frac{1}{\delta}}$, so that there is no more than δ probability that

$$\left|\mu - \frac{\sigma_0^2\overline{x}}{\sigma_X^2}\right| = |\mu - \mathrm{E}\left[\mu \mid \overline{x}\right]| \geq k\sigma_\mu = \frac{k}{\sqrt{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}}.$$

Then, assuming that this holds,

$$\left|\ln\frac{p_{\sigma,\sigma_0}\left(\mu\,|\overline{x}\,\right)}{p_{\sigma,\sigma_0}\left(\mu\,|\overline{x}'\,\right)}\right| \leq \frac{n}{2\sigma^2}\cdot\frac{R}{n}\left(\frac{R}{n}+\frac{k}{\sqrt{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}}\right) \leq \frac{R}{2\sigma^2}\left(\frac{R}{n}+\frac{k\sigma}{\sqrt{n}}\right).$$

For this to be no more than ε, it suffices to take $n \geq \max\left(\frac{R^2}{\sigma^2\varepsilon}, \frac{R^2}{k^2\sigma^2\varepsilon^2}\right)$, or $\frac{R}{\sigma} \leq \frac{\varepsilon\sqrt{n}}{\sqrt{2\ln\frac{1}{\delta}}}$, so that each of the two terms in the sum is no more than $\varepsilon/2$. □

Proof of lemma. We can bound

$$\Pr\left[|\mu - \mathrm{E}\left[\mu \mid \overline{x}\right]| \geq k\sigma_\mu \mid \overline{x}\right] = 2\Pr\left[\mu \geq \mathrm{E}\left[\mu \mid \overline{x}\right] + k\sigma_\mu \mid \overline{x}\right] = \sqrt{\frac{2}{\pi}}\int_{a=k}^{\infty} e^{-\frac{k^2}{2}} \leq \sqrt{\frac{2}{\pi}}\frac{1}{k}e^{-k^2/2}.$$

For this to be less than δ, it suffices to take $k \geq \sqrt{2\log\frac{1}{\delta}}$ and $k \geq \sqrt{\frac{\pi}{2}}$. □

In this bound, so long as δ cannot decrease too quickly in n – i.e. $\delta = e^{-o(\sqrt{n})}$ – the value of $\frac{R}{\sigma}$ is now allowed to increase as n does, which is more satisfactory what what we saw in the bound for ε-differential privacy.

7.2 Releasing Synthetic Data

When we move on to generating synthetic data, unrestricted data points still prevent us from obtaining any privacy:

Theorem 7.10. Synth$_{\text{MAP}}$ *and* Synth$_{\text{Post}}$ *are not (ε, δ)-differentially private when $\mathcal{X} = \mathbb{R}$.*

Proof. For Synth$_{\text{Post}}$ and any $z \in \mathbb{R}$, we have that

$$\begin{aligned}\ln\frac{p_{\sigma,\sigma_0}\left(z\,|\overline{x}\,\right)}{p_{\sigma,\sigma_0}\left(z\,|\overline{x}'\,\right)} &= \frac{1}{2\left(\sigma^2+\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right)}\frac{\sigma_0^2}{\frac{\sigma^2}{n}+\sigma_0^2}\left(\overline{x}-\overline{x}'\right)\left(2z-\frac{\sigma_0^2(\overline{x}+\overline{x}')}{\frac{\sigma^2}{n}+\sigma_0^2}\right)\\ &= \frac{\sigma_0^2}{2\sigma_{\overline{x}}^2(\sigma^2+\sigma_\mu^2)}(\overline{x}-\overline{x}')\left(2z-\frac{\sigma_0^2\left(\overline{x}+\overline{x}'\right)}{\sigma_{\overline{x}}^2}\right).\end{aligned}$$

This is almost the same ratio as we saw for $\ln\frac{p_{\sigma,\sigma_0}(\mu|\overline{x}\,)}{p_{\sigma,\sigma_0}(\mu|\overline{x}'\,)}$, so the rest of the proof follows analogously to that of Theorem 7.4.

If instead we use $\textsc{Synth}_{\text{MAP}}$, then we obtain

$$\ln \frac{p_{\sigma,\sigma_0}\left(z\,|\overline{x}\right)}{p_{\sigma,\sigma_0}\left(z\,|\overline{x}'\right)} = \frac{1}{2\sigma^2}\left(\overline{x}-\overline{x}'\right)\left(2z-\left(\overline{x}+\overline{x}'\right)\right),$$

which is similarly not (ε, δ)-differentially private. □

Like we saw when generating the parameter, even restricting the data points to $[-R, R]$ does not give ε-differential privacy for generating synthetic data.

Corollary 7.11. *$\textsc{Synth}_{\text{MAP}}$ and $\textsc{Synth}_{\text{Post}}$ are not ε-differentially private when the data points are restricted to lie in $[-R, R]$.*

Again, however, adding another restriction gives us privacy:

Theorem 7.12. *When releasing a single synthetic data point z, $\textsc{Synth}_{\text{Post}}$ is ε-differentially private if $\mathcal{X} = [-R, R]$, z is conditioned to lie in $[-R, R]$, and $\frac{R}{\sigma} \leq \sqrt{\varepsilon n/2}$. This is asymptotically tight.*

Proof. We have

$$\frac{\sigma_0^2}{2\sigma_{\overline{x}}^2(\sigma^2+\sigma_\mu^2)} \leq \frac{1}{2\sigma^2}, |\overline{x}-\overline{x}'| \leq \frac{R}{n}, \left|2z - \frac{\sigma_0^2\left(\overline{x}+\overline{x}'\right)}{\sigma_{\overline{x}}^2}\right| \leq 4R,$$

which combined give

$$\left|\ln \frac{\Pr\left[z \mid \overline{x}\right]}{\Pr\left[z \mid \overline{y}\right]}\right| \leq \frac{2R^2}{n\sigma^2}.$$

Setting this to be less than ε gives the desired result. As in the proof of Theorem 7.7, this is almost an necessary condition – in particular, as $n \to \infty$, the first term approaches $\frac{1}{2\sigma^2}$ while the other two are reachabout up to a factor of $\left(1 - o\left(\frac{1}{n}\right)\right)$. □

We can compare this to the bound of $\sqrt{\epsilon}/2$ for obtaining ε-differential privacy with $\textsc{Param}_{\text{Post}}$. When releasing synthetic data, we were able to obtain an extra factor of $\sqrt{n}$ in the bound because the variance of z is always at least σ^2, whereas the variance of μ goes to 0 as the number of data points grows.

For $\textsc{Synth}_{\text{Post (Many)}}$, combining Theorem 7.12 with basic composition results in:

Corollary 7.13. *When releasing a single synthetic data point z, $\textsc{Synth}_{\text{Post (Many)}}$ is ε-differentially private if $\mathcal{X} = [-R, R]$, z is conditioned to lie in $[-R, R]$, and $\frac{R}{\sigma} \leq \sqrt{\varepsilon n/(2m)}$. This is asymptotically tight.*

We can use almost identical bounds to show the same result for $\textsc{Synth}_{\text{MAP}}$. In this case, it is possible to be slightly more precise about the bound:

Corollary 7.14. *When releasing a single synthetic data point z with $\mathcal{X} = [-R, R]$ and z is conditioned to lie in $[-R, R]$, $\textsc{Synth}_{\text{MAP}}$ is (ε, δ)-differentially private if and only if $\frac{R}{\sigma} \leq \sqrt{\frac{\varepsilon n}{2-\frac{1}{2n}}}$.*

Unlike our ε-differentially private bound on $\textsc{Param}_{\text{Post}}$, this allows the range of the data to grow with the number of data points.

Now, we move on to ε-differential privacy, where we again obtain a better bound.

Theorem 7.15. *For $n \geq \frac{\sigma^2}{\sigma_0^2}$, $\textsc{Synth}_{\text{MAP}}$ and $\textsc{Synth}_{\text{Post (Many)}}$, when $\mathcal{X} = [-R, R]$, are (ε, δ)-differentially private when*

$$\frac{R}{\sigma} \leq \frac{n\varepsilon}{2m\sqrt{2\ln\frac{m}{\delta}}}.$$

Proof. The proof follows by considering a single data point and then applying basic composition. By Lemma 7.9 (a tail bound on the Gaussian),

$$k \geq \sqrt{2 \log \frac{1}{\delta}} \Rightarrow \Pr\left[|z - E[z \mid \overline{x}]| \geq k\sqrt{\sigma_\mu^2 + \sigma^2} \mid \overline{x}\right] \leq \frac{1}{\delta}.$$

Assuming this is the case, then

$$\begin{aligned}\left|\ln \frac{p_{\sigma,\sigma_0}\left(z \,|\overline{x}\,\right)}{p_{\sigma,\sigma_0}\left(z \,|\overline{x}'\,\right)}\right| &= \frac{\sigma_0^2}{2\sigma_{\overline{x}}^2(\sigma^2+\sigma_\mu^2)}\left(\frac{R}{n}\right)\left(\frac{R}{n} + 2k\sqrt{\sigma^2+\sigma_\mu^2}\right). \\ &\leq \frac{1}{2\sigma^2}\left(\frac{R}{n}\right)\left(\frac{R}{n} + 2k\sqrt{\sigma^2+\sigma_\mu^2}\right).\end{aligned}$$

Setting each of these two terms to be no more than $\varepsilon/2$, and using $n \geq \frac{\sigma^2}{\sigma_0^2} \Rightarrow \sigma_\mu^2 \leq \sigma^2$ we see that this is ε-differentially private so long as $\frac{R}{\sigma} \leq \min\left(n\sqrt{\varepsilon}, \frac{n\varepsilon}{2\sqrt{2\ln\frac{1}{\delta}}}\right)$. □

In particular, this is still not very satisfactory in the case that $m = n$. In fact, when $m = n$, we would do better to use $\textsc{Synth}_{\text{Post (One)}}$ with the bound of $\frac{R}{\sigma} \leq \frac{\varepsilon\sqrt{n}}{\sqrt{2}\ln\frac{1}{\delta}}$ obtained in Theorem 7.8. To get a better bound for large values of m, we use advanced composition and Theorem 7.12:

Corollary 7.16. *There exists some constant c such that* $\textsc{Synth}_{\text{MAP}}$ *and* $\textsc{Synth}_{\text{Post (Many)}}$ *are (ε,δ)-differentially private when $\mathcal{X} = [-R, R]$ and*

$$\frac{R}{\sigma} \leq \sqrt{\frac{\varepsilon n}{4\sqrt{2m\ln\frac{1}{\delta}}}}.$$

Now, if we want to draw a linear number of synthetic data points in n, then this last result allows us to do so more realistically, since $\frac{R}{\sigma}$ is allowed to increase faster than $\sqrt{\ln n}$.

Chapter 8: Conclusion

In this thesis, we have examined five methods for performing Bayesian inference on a dataset to try to understand the extent to which multiple imputation already satisfies differential privacy.

Bernoulli vs. Gaussian Distributions

To do this, we examined in depth two distributions – the Bernoulli and the Gaussian – and proved new bounds on the conditions needed for differential privacy. In both cases, we have tried to provide some context on the conditions. For the Bernoulli, we interpreted them as a minimum quantity of prior data. Meanwhile, for the Gaussian, we interpreted the conditions as the maximum allowable range of the data.

Releasing Parameters vs. Releasing Synthetic Data

When releasing synthetic data, a parameter generation step is followed by a synthetic data generation step. Thus, one question we wanted to address was how much privacy was provided by the first compared to the second. While we might expect that, by adding an extra layer of randomness, synthetic data generation is more private than parameter drawing, it is not entirely clear whether this is the case.

Certainly, for the Bernoulli distribution, it was possible to obtain ε-differential privacy when releasing synthetic data, whereas it was not possible to do so when releasing parameters. Additionally, when the number of synthetic data points m is small (relative to the real database size n), we have clearly been able to obtain more privacy than when releasing synthetic data than when releasing of parameters.

Yet, as we noted in Chapter 5, releasing an arbitrarily large number of synthetic data points seems like it would yield a similar degree of privacy loss as releasing the parameter used for generating that data.

In the case of $m = n$, the (ε, δ)-differential privacy bounds we have proven for $\textsc{Synth}_{\text{Post (One)}}$ and $\textsc{Synth}_{\text{Post (Many)}}$ are fairly comparable to those we proved for $\textsc{Param}_{\text{Post}}$, which is a parameter release method. Consequently, if we want to release a synthetic data set of the same size as the original using our bounds, we have the same degree of inaccuracy as if we released a parameter using $\textsc{Param}_{\text{Post}}$. However, because we do not know whether our (ϵ, δ) bounds are tight, it is possible that synthetic data may actually provide more privacy than what we have been able to establish thus far.

What is perhaps more interesting here, though, is that we were not able to prove that $\textsc{Synth}_{\text{Post (One)}}$ and $\textsc{Synth}_{\text{Post (Many)}}$ provide more privacy than $\textsc{Synth}_{\text{MAP}}$, even though the latter's parameter is deterministic given the database: the amount of privacy provided in $\textsc{Synth}_{\text{MAP}}$ by the randomness of the synthetic data drawing alone still matched that of the parameter drawing in $\textsc{Param}_{\text{Post}}$.

Consequently, it may be that, for a reasonably (but not extraordinarily) large number of data points, both parameter drawing and synthetic data drawing can provide similar amounts of privacy, yet the combination of the two does not necessarily provide more.

ε-differential privacy vs. (ε, δ)-differential privacy

Perhaps unsurprisingly, we have seen very clear benefits when going from ε-differential privacy to (ε, δ)-differential privacy, in the sense that it is possible to achieve (ε, δ)-differential privacy with more reasonable priors or data restrictions. In both the distributions examined and for nearly all the methods,[1] ε-differential privacy was either unachievable wholesale or not practical when the number of synthetic data points desired was linear in n. As we proved that most of these ε-differential privacy bounds are tight, this means that ε-differential privacy for multiple imputation may not be particularly meaningful in practice. Meanwhile, in both distributions, (ε, δ)-differential privacy was more reasonably achievable for $\textsc{Param}_{\text{Post}}$ as well as for our three synthetic data generation methods.

Future Work

There are many potential directions for future work, including:

- **Establishing more lower bounds.** While we have given tight asymptotic bounds for the Bernoulli and Gaussian distributions for some pairs of privacy types and methods, for others we have only given a sufficient, not a necessary bound. Especially for (ε, δ)-differential privacy, being able to prove a tight bound would give more clarity for comparing the privacy afforded by private synthetic data generation to that afforded by parameter generation.
- **Establishing more $(\varepsilon, \delta, \Delta)$-differential privacy bounds.** We showed in Chapter 6.3 that there is also no corresponding average-case privacy for the Bernoulli distribution where either ε-differential privacy or (ε, δ)-differential privacy was not possible. However, it may be that the *quantity* of average-case privacy obtained is higher than that of regular privacy in cases where regular differential privacy already exists.
- **Examining more distributions.** In Chapter 6.4, we extended our ε-differential privacy bounds for the Bernoulli case to that of a categorical random variable, but have not yet done so for our (ε, δ)-differential privacy bounds. Alternatively, we suspect that many of the bounds in Chapter 7 for the Gaussian distribution could be extended to a multivariate normal distribution (perhaps even with the same proofs, replacing scalars by vectors or matrices as needed) – however, even if this is true we would still need to work out the dependence of the bounds on the degree of the multivariate normal. There are also many other important distributions that we have not touched, such as the exponential distribution or the Gaussian distribution with known mean and unknown variance.
- **Looking for more general criteria on distributions.** Ideally, going along the lines of the work by Dimitrakakis et al. [DNMR13], we would like to be able to formulate more general conditions on distributions that determine whether or not these methods are differentially private.
- **Evaluating the accuracy of these methods.** Given the work mentioned in Chapter 4.3, we do not know what the accuracy implications of these prior choices or data restrictions are. While we have tried to give some indication in both Chapters 6 and 7 of whether the priors or restrictions would be entirely unwieldy in theory, that is still far from saying that they would necessarily work well in practice at the sizes of real datasets.

  Additionally, in Chapter 2.3, we mentioned that subsampling can be used to gain ε-differential privacy at an accuracy cost of running algorithms on a database that is smaller in size by a factor of ε. This could be one way to run relatively unaltered versions of the Bayesian inference methods we have examined; however, the practical implications of the accuracy-privacy trade-off here are also unclear.

[1] With the exception of $\textsc{Param}_{\text{MAP}}$, which is neither ε-differentially private nor (ε, δ)-differentially private.

Chapter 9:   Bibliography

[AW89] Nabil R. Adam and John C. Worthmann. Security-control methods for statistical databases: A comparative study. *ACM Computer Survey*, 21(4):515–556, December 1989.

[BGKS13] R. Bassily, A. Groce, J. Katz, and A. Smith. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 439–448, Oct 2013.

[BLR11] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. *CoRR*, abs/1109.2229, 2011.

[Cha11] Anne-Sophie Charest. How can we analyze differentially-private synthetic datasets. *Journal of Privacy and Confidentiality*, 2(2), 2011.

[Cha12] Anne-Sophie Charest. Empirical evaluation of statistical inference from differentially-private contingency tables. In *Proceedings of the 2012 International Conference on Privacy in Statistical Databases*, PSD'12, pages 257–272, Berlin, Heidelberg, 2012. Springer-Verlag.

[CKN$^+$11] J.A. Calandrino, A. Kilzer, A. Narayanan, E.W. Felten, and V. Shmatikov. "you might also like:" privacy risks of collaborative filtering. In *Security and Privacy (SP), 2011 IEEE Symposium on*, pages 231–246, May 2011.

[DBR08] Jörg Drechsler, Stefan Bender, and Susanne Rässler. Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB establishment panel. *Trans. Data Privacy*, 1(3):105–130, December 2008.

[DL86] George T. Duncan and Diane Lambert. Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81(393):10–18, 1986.

[DL09] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing*, STOC '09, pages 371–380, New York, NY, USA, 2009. ACM.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.

[DNMR13] Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin Rubinstein. Robust, secure and private bayesian inference. (1306.1066), 2013.

[DR10] Jrg Drechsler and Jerome P. Reiter. Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105(492):1347–1357, 2010.

[DR13] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(34):211–407, 2013.

[Dre11] Jorg Drechsler. My understanding of the differences between the CS and the statistical approach to data confidentiality. In *The 4th IAB workshop on confidentiality and disclosure*. Institute for Employment Research, 2011.

[DRV10] Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 51–60, Washington, DC, USA, 2010. IEEE Computer Society.

[KLN$^+$08] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *CoRR*, abs/0803.0924, 2008.

[KRS12] Shiva Prasad Kasiviswanathan, Mark Rudelson, and Adam Smith. The power of linear reconstruction attacks. *CoRR*, abs/1210.2381, 2012.

[KS08] Shiva Prasad Kasiviswanathan and Adam Smith. A note on differential privacy: Defining resistance to arbitrary side information. *CoRR*, abs/0803.3946, 2008.

[MKA$^+$08] Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, ICDE '08, pages 277–286, Washington, DC, USA, 2008. IEEE Computer Society.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 94–103, Washington, DC, USA, 2007. IEEE Computer Society.

[NRS07] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing*, STOC '07, pages 75–84, New York, NY, USA, 2007. ACM.

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, SP '08, pages 111–125, Washington, DC, USA, 2008. IEEE Computer Society.

[oHS] US Department of Health and Human Services. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability act (HIPAA) Privacy Rule.

[Rei05] Jerome P Reiter. Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100(472):1103–1112, 2005.

[RRR03] Trivellore E Raghunathan, Jerome P Reiter, and Donald B Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics-Stockholm-*, 19(1):1–16, 2003.

[Rub76] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[Rub93] Donald B. Rubin. Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468, 1993.

[Sla13] Aleksandra Slavkovic. Overview of statistical disclosure limitation. 2013.

[Smi08] Adam Smith. Efficient, differentially private point estimators. *CoRR*, abs/0809.4794, 2008.

[Smi11] Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, pages 813–822, New York, NY, USA, 2011. ACM.

[Swe00] Latanya Sweeney. Simple demographics often identify people uniquely. 2000.

[Toc14] Anthony Tockar. Riding with the stars: Passenger privacy in the NYC taxicab dataset, 2014.

[Ull12] Jonathan Ullman. Answering n^{2+o(1)} counting queries with differential privacy is hard. *CoRR*, abs/1207.6945, 2012.

[Zay07] Laura Zayatz. Disclosure avoidance practices and research at the US Census Bureau: An update. *Journal of Official Statistics-Stockholm-*, 23(2):253, 2007.

Appendix A: Additional Proofs

From Chapter 2

Proposition 2.10. *Let $\mathcal{A}$ be an algorithm. Define*

$$S_\varepsilon(X, X') = \{x : \Pr[\mathcal{A}(X) = x] \leq e^\varepsilon \cdot \Pr[\mathcal{A}(X') = x]\}.$$

Then, if for all neighboring databases X, X',

$$\Pr[\mathcal{A}(X) \notin S_\varepsilon(X, X')] \leq \delta,$$

$\mathcal{A}$ is (ε, δ)-differentially private. This also holds if the output of $\mathcal{A}$ is a continuous random variable with probability density function p, i.e. if we define $S_\varepsilon(X, X') = \{x : p(\mathcal{A}(X) = x) \leq e^\varepsilon \cdot p(\mathcal{A}(X') = x)\}$.

Proof. For any set S and neighboring databases X, X', we have

$$\begin{aligned}\Pr[A(X) \in S] &= \Pr[A(X) \in S \cap S_\varepsilon(X, X')] + \Pr[A(X) \in S \cap S_\varepsilon(X, X')] \\ &\leq e^\varepsilon \cdot \Pr[A(X') \in S \cap S_\varepsilon(X, X')] + \delta \\ &\leq e^\varepsilon \cdot \Pr[A(X') \in S] + \delta.\end{aligned}$$

Similarly, $\Pr[A(X') \in S] \leq e^\varepsilon \cdot \Pr[A(X) \in S]$. □

Proposition 2.11. *Define $S_\varepsilon(X, X')$ as before. If there exist neighboring databases X, X' for which*

$$\Pr[\mathcal{A}(X) \notin S_\varepsilon(X, X')] > \delta,$$

then $\mathcal{A}$ is not $(\varepsilon/2, \delta(1 - e^{-\varepsilon/2}))$-differentially private.

Proof. Because $e^\varepsilon \cdot \Pr[\mathcal{A}(X') \notin S_\varepsilon(X, X')] \leq \Pr[\mathcal{A}(X) \notin S_\varepsilon(X, X')]$,

$$\begin{aligned}e^{\varepsilon/2} \cdot \Pr[\mathcal{A}(X') \notin S_\varepsilon(X, X')] + \delta(1 - e^{-\varepsilon/2}) &\leq e^{-\varepsilon/2} \cdot \Pr[\mathcal{A}(X) \notin S_\varepsilon(X, X')] + (1 - e^{-\varepsilon/2}) \cdot \Pr[\mathcal{A}(X) \notin S_\varepsilon(X, X')] \\ &= \Pr[\mathcal{A}(X) \notin S_\varepsilon(X, X')],\end{aligned}$$

contradicting $(\varepsilon/2, \delta(1 - e^{-\varepsilon/2}))$-differentially privacy. □

Proposition 2.13. *If an algorithm $\mathcal{A}$ is deterministic but not constant, then it is neither ε-differentially private nor (ε, δ)-differentially private.*

Proof. Consider databases X and Y such that $\mathcal{A}(X) \neq \mathcal{A}(Y)$. We can walk from X to Y by changing one row at a time – i.e. there exists a sequence of databases $X_1, \ldots, X_k$ such that X_i and X_{i+1} are neighbors for all $1 \leq i < k$ and with $X = X_1, Y = X_k$.

Because $\mathcal{A}(X_1) \neq \mathcal{A}(X_k)$, there must then exist some i such that $\mathcal{A}(X_i) \neq \mathcal{A}(X_{i+1})$. Letting $S = \{\mathcal{A}(X_i)\}$,

$$\Pr[\mathcal{A}(X_i) \in S] = 1 \not\leq \delta = e^\varepsilon \cdot \Pr[\mathcal{A}(X_{i+1}) \in S] + \delta,$$

so $\mathcal{A}$ is not (ε, δ)-differentially private. □

From Chapter 6

Lemma 6.5. *Let θ denote the output of* $\textsc{Param}_{\text{Post}}$*. If*

$$\Pr_{\alpha,\beta}\left[\frac{e^{\varepsilon}(n_1+\alpha-1)}{(n_0+\beta)+e^{\varepsilon}(n_1+\alpha-1)} \geq \theta \geq \frac{(n_1+\alpha)}{(n_1+\alpha)+e^{\varepsilon}(n_0+\beta-1)}\right] \geq 1-\delta,$$

then $\textsc{Param}_{\text{Post}}$ *is (ε,δ)-differentially private.*

Proof. First, note that if

$$\frac{e^{\varepsilon}(n_1+\alpha-1)}{(n_0+\beta)+e^{\varepsilon}(n_1+\alpha-1)} \geq \theta \geq \frac{(n_1+\alpha-1)}{(n_1+\alpha-1)+e^{\varepsilon}(n_0+\beta)},$$

then $e^{\varepsilon} \geq \frac{p_{\alpha,\beta}(\theta|X)}{p_{\alpha,\beta}(\theta|X')} = \frac{n_1+\alpha-1}{n_0+\beta}\cdot\frac{1-\theta}{\theta} \geq e^{-\varepsilon}$, where X' is obtained by replacing a 1 with a 0 in X. Similarly, if

$$\frac{e^{\varepsilon}(n_1+\alpha)}{(n_0+\beta-1)+e^{\varepsilon}(n_1+\alpha)} \geq \theta \geq \frac{(n_1+\alpha)}{(n_1+\alpha)+e^{\varepsilon}(n_0+\beta-1)},$$

then $e^{\varepsilon} \geq \frac{p_{\alpha,\beta}(\theta|X)}{p_{\alpha,\beta}(\theta|X')} \geq e^{-\varepsilon}$ for X' obtained by replacing a 0 in X.

Applying Proposition 2.10 gives us (ε,δ)-differential privacy when we look at the intersection of these two intervals for θ. □

Lemma 6.8. *There exists a constant c such that for sufficiently large n, when $\alpha,\beta \leq \frac{c}{\varepsilon^2}\left(\ln\frac{1}{\delta n}\right)$, then when $n_1 = 0, n_0 = n$, with probability at least δ,* $\textsc{Param}_{\text{Post}}$ *releases a value of θ outside of the range $\left[e^{-\varepsilon}\cdot\frac{n_1+\alpha}{n+\alpha+\beta}, e^{\varepsilon}\cdot\frac{n_1+\alpha}{n+\alpha+\beta}\right]$.*

Proof. We use the same notation and setup as in the proof of Lemma 6.6. By Bayes' rule,

$$\begin{aligned}
\Pr\left[\theta \leq e^{-\varepsilon}\theta_{\text{exp}} \;\middle|\; \sum_{i=1}^{n'} X_i = \theta_{\text{obs}}n'\right] &\geq \Pr\left[e^{-\varepsilon}\theta_{\text{exp}} \geq \theta \geq e^{-2\varepsilon}\theta_{\text{exp}} \;\middle|\; \sum_{i=1}^{n'} X_i = \theta_{\text{obs}}n'\right] \\
&= \frac{\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}}n' \;\middle|\; e^{-\varepsilon}\theta_{\text{exp}} \geq \theta \geq e^{-2\varepsilon}\theta_{\text{exp}}\right]\Pr\left[e^{-\varepsilon}\theta_{\text{exp}} \geq \theta \geq e^{-2\varepsilon}\theta_{\text{exp}}\right]}{\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}}n'\right]} \\
&\geq \frac{\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}}n' \;\middle|\; \theta = e^{-2\varepsilon}\theta_{\text{exp}}\right]\Pr\left[e^{-\varepsilon}\theta_{\text{exp}} \geq \theta \geq e^{-2\varepsilon}\theta_{\text{exp}}\right]}{\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}}n'\right]}.
\end{aligned}$$

Using Stirling's approximation,

$$\begin{aligned}
\Pr\left[\sum X_i = \theta_{\text{obs}} n' \;\middle|\; \theta = e^{-2\varepsilon}\theta_{\text{exp}}\right] &= \binom{n'}{\theta_{\text{obs}} n'} (e^{-2\varepsilon}\theta_{\text{exp}})^{\theta_{\text{obs}} n'} (1 - e^{-2\varepsilon}\theta_{\text{exp}})^{(1-\theta_{\text{obs}})n'} \\
&= \Theta\left(\frac{n'^{n'+\frac{1}{2}} e^{-n'} (e^{-2\varepsilon}\theta_{\text{exp}})^{\theta_{\text{obs}} n'} (1 - e^{-2\varepsilon}\theta_{\text{exp}})^{(1-\theta_{\text{obs}})n'}}{(\theta_{\text{obs}} n')^{\theta_{\text{obs}} n' + \frac{1}{2}} e^{-\theta_{\text{obs}} n'} ((1-\theta_{\text{obs}})n')^{(1-\theta_{\text{obs}})n' + \frac{1}{2}} e^{-(1-\theta_{\text{obs}})n'}} \right) \\
&= \Omega\left(\frac{(e^{-2\varepsilon})^{\theta_{\text{obs}} n'} (1 - e^{-2\varepsilon}\theta_{\text{exp}})^{(1-\theta_{\text{obs}})n'}}{(1-\theta_{\text{obs}})^{(1-\theta_{\text{obs}})n'} \sqrt{n'\theta_{\text{obs}}(1-\theta_{\text{obs}})}} \right) \\
&\approx \Omega\left(\frac{(e^{-2\varepsilon})^{\theta_{\text{obs}} n'} e^{-e^{-2\varepsilon}\theta_{\text{exp}}(1-\theta_{\text{obs}})n'} e^{\theta_{\text{obs}}(1-\theta_{\text{obs}})n'}}{\sqrt{n'\theta_{\text{obs}}(1-\theta_{\text{obs}})}} \right) \\
&\approx \Omega\left(\frac{e^{\theta_{\text{obs}} n'(-2\varepsilon + (1-\theta_{\text{obs}})(1-e^{-2\varepsilon}))}}{\sqrt{n'\theta_{\text{obs}}(1-\theta_{\text{obs}})}} \right).
\end{aligned}$$

Plugging this into the result of Bayes' rule, and using the facts that

$$\Pr\left[e^{-\varepsilon}\theta_{\text{exp}} \geq \theta \geq e^{-2\varepsilon}\theta_{\text{exp}}\right] = \theta_{\text{exp}}(e^{-\varepsilon} - e^{-2\varepsilon}) = \Theta\left(\theta_{\text{exp}}\varepsilon\right)$$

and

$$\Pr\left[\sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n'\right] = \Theta\left(\frac{1}{\sqrt{\theta_{\text{obs}}(1-\theta_{\text{obs}})n'}} \right),$$

we obtain

$$\Pr\left[\theta \leq e^{-\varepsilon}\theta_{\text{exp}} \;\middle|\; \sum_{i=1}^{n'} X_i = \theta_{\text{obs}} n'\right] = \Omega\left(\theta_{\text{exp}}\varepsilon \cdot e^{\theta_{\text{obs}} n'(-2\varepsilon + (1-\theta_{\text{obs}})(1-e^{-2\varepsilon}))}\right).$$

Note that $\theta_{\text{exp}} = \frac{\alpha}{n'} \approx \theta_{\text{obs}}$. Consequently, the above value is

$$\Omega\left(\frac{\varepsilon}{n'} \cdot e^{\alpha(-2\varepsilon + (1-\theta_{\text{obs}})(1-e^{-2\varepsilon}))}\right).$$

Using the Taylor expansion $e^{-2\varepsilon} \approx 1 - 2\varepsilon + 2\varepsilon^2$ to get that the negative of the exponent is $\Theta\left(\alpha\varepsilon^2\right)$, we see that this is greater than δ when the condition in the lemma holds. □