New Statistical Applications for Differential Privacy

Rob Hall

December 2012
CMU-ML-12-113

Carnegie Mellon

New Statistical Applications for Differential Privacy

Rob Hall

December 2012

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Department of Statistics
Dietrich College of Humanities and Social Sciences
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Stephen Fienberg, Chair, CMU
Larry Wasserman, CMU
Alessandro Rinaldo, CMU
Adam Smith, Pennsylvania State University, University Park

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Abstract

Differential privacy is a relatively recent development in the field of privacy-preserving data mining, which was formulated to give a mathematically rigorous definition of privacy. The concept has spawned a great deal of work regarding the development of algorithms which are privacy-preserving under this definition, and also work which seeks to understand the fundamental limitations of such algorithms. When the goal is statistical inference it is important to understand what set of analyses may be carried out in the privacy-preserving framework with reasonable accuracy, and which data summaries and results may be reported.

In this work we begin by examining fundamental limitations of differentially private procedures when the goal is to release a sparse histogram, or a contingency table. We also describe error bounds when the goal is model selection for a private contingency table. Through examples we will demonstrate the implications of these lower error bounds for statistical analysis with datasets of different sizes and dimensions.

Thus far, differentially private algorithms appear to be restricted to the release of finite dimensional vectors (e.g., regression coefficients, point estimates, SVM parameters). We next develop methods for releasing functional data while preserving differential privacy. Specifically, we show that when the function of interest lies inside a reproducing kernel Hilbert space, then it satisfies differential privacy to release the function plus a Gaussian process noise term. As examples we consider kernel density estimation, kernel support vector machines, and suitably smooth functions which lie in a particular Sobolev space.

Finally we propose a relaxed privacy definition called *random differential privacy* (RDP). Differential privacy requires that adding any new observation to a database will have small effect on the output of the data-release procedure. Random differential privacy requires that adding a *randomly drawn new observation* to a database will have small effect on the output. We show an analog of the composition property of differentially private procedures which applies to our new definition. We show how to release an RDP histogram and we show that RDP histograms are much more accurate than histograms obtained using ordinary differential privacy. We finally show an analog of the global sensitivity framework for the release of functions under our privacy definition. We demonstrate that an extension of this idea and describe how it relates to other looser privacy definitions that have begun appearing in the literature.

Acknowledgements

I am thankful to Steve Fienberg for being everything I needed in an adviser. He pushed me to succeeed, and also gave me the freedom to look at the problems I was interested in. I am also thankful to Larry Wasserman and Alessandro Rinaldo for their invaluable contributions to this thesis, specifically in helping me to understand Reproducing Kernel Hilbert Spaces. I am grateful to Adam Smith for being a part of my committee, and for demanding a level of precision in my writing of which I can be proud. Lastly to my friends and family, for their support throughout my time in academia.

Contents

1 Differential Privacy **1**

1.1 Introduction . . . 1

1.1.1 Outline . . . 3

1.1.2 Related Work . . . 3

1.2 Differential Privacy . . . 6

1.2.1 Interpretation . . . 6

1.2.2 Approximate Differential Privacy . . . 8

1.2.3 Composition . . . 9

1.2.4 Post-Processing . . . 10

1.3 Basic Methods . . . 10

1.3.1 Noise Addition . . . 11

1.3.2 Noise Addition for Approximate Differential Privacy . . . 13

1.3.3 The K-norm Mechanism . . . 14

1.3.4 The Exponential Mechanism . . . 16

1.3.5 Optimal Discrete Mechanisms via Linear Programming . . . 17

1.4 Relationship to Cryptographic Protocols . . . 18

1.5 Summary . . . 20

2 Differential Privacy and Minimaxity in Discrete Problems **21**

2.1 Risk and Minimaxity . . . 21

2.1.1 Risk Decomposition . . . 22

2.2 Information Theoretic Inequalities from Statistics . . . 23

2.3 Lower Bounds for Counting Queries . . . 25

2.3.1 A Minimax Lower Bound . . . 25

2.3.2 Uniform Lower Bounds for Counting Queries . . . 26

2.3.3 Computation of Minimax Methods . . . 27

2.4 Lower Bounds for Histograms . . . 30

2.4.1 Lower Bounds for Sparse Histograms . . . 32
2.4.2 Linear Functions of Histograms . . . 37
2.5 Summary . . . 39

3 Differential Privacy in Spaces of Functions 40
3.1 Finite Dimensional Techniques . . . 41
3.1.1 Expansion in an Orthonormal Basis . . . 41
3.2 Differential Privacy in a Reproducing Kernel Hilbert Space . . . 42
3.2.1 Reproducing Kernel Hilbert Space Basics . . . 42
3.2.2 Privacy via the Spectral Decomposition in an RKHS . . . 43
3.2.3 (α, β)-Differential Privacy in an RKHS . . . 46
3.2.4 Alternate View of Gaussian Process Noise . . . 47
3.2.5 Algorithms . . . 50
3.3 Examples . . . 51
3.3.1 Kernel Density Estimation . . . 51
3.3.2 A Sobolev Space of Functions . . . 51
3.3.3 Minimizers of Regularized Functionals in an RKHS . . . 54
3.4 Summary . . . 56

4 Kernel Density Estimation 58
4.1 Introduction . . . 58
4.2 Non Private Kernel Density Estimation . . . 59
4.3 Fourier Approximation of the Kernel Density Estimate . . . 59
4.3.1 The Truncated Fourier Approximation . . . 63
4.4 Differentially Private Versions of the Truncated Density Estimator . . . 65
4.4.1 Privacy via Laplace Noise Addition . . . 65
4.4.2 Privacy via "Caratheodory Noise" Addition . . . 66
4.4.3 Inertia of the Caratheodory Orbitope . . . 69
4.4.4 Sampling the Convex Body . . . 73
4.5 Summary . . . 74

5 Weaker Alternatives to Differential Privacy 76
5.1 Random Differential Privacy . . . 76
5.1.1 Remarks about Random Differential Privacy . . . 77
5.2 RDP For Discrete Random Variables . . . 78
5.2.1 Basic Technique . . . 79
5.2.2 Binomial Proportion Estimation . . . 80

5.2.3 Sparse Histograms 81
5.3 RDP via Sensitivity Analysis 84
5.4 Variants of Random Differential Privacy 88
5.5 Summary 88

6 Summary **90**

Chapter 1

Differential Privacy

1.1 Introduction

Over the last decade, a great deal of activity in machine learning and related fields focused on "privacy preserving data-mining" (see e.g., Vaidya *et al.* [2005]; Dwork [2008]) which spawned regular meetings and workshops (e.g., Privacy and Statistical Databases) and journals of its own (e.g., the Journal of Privacy and Confidentiality). As statistical analyses of large scale databases become more commonplace, and as results of such analyses are made available to interested parties, it has become increasingly important to respect the privacy of the individuals who's information comprises the data. The field of privacy preserving datamining seeks to build methods which approximate the answers that a statistical analysis would give—that is, to maintain some "utility" to the analysts, but in a way which may be deemed to protect the individuals in the data. Evidently a non-private technique has full utility in the sense that it gives the exact answer, whereas to maintain full privacy may preclude the data owner from releasing anything at all. Thus there is often a tradeoff between utility and privacy (see e.g., Fienberg *et al.* [2008, 2010]; Lebanon *et al.* [2006]; Yang *et al.* [2012]).

An orthogonal concern within the same field is the research in cryptographic protocols (see e.g., Goldreich [1998, 2004]; Lindell and Pinkas [2002, 2009]). There the goal is to allow multiple parties to perform a computation on the union of their data, but in a way which messages passed during the computation do not reveal information about the input. In this setting the output is not inherently protected, it is given to all parties upon termination of the protocol. Whether or not the output reveals information about the input depends on the nature of the function which was computed. Examples for regression and classification are given in e.g., Vaidya *et al.* [2005]; Fienberg *et al.* [2006, 2009].

During the 1980s and 1990s, the statistics community took interest in privacy preserving methods

(see e.g., Duncan and Lambert [1986, 1989]; Duncan and Pearson [1991]; Hwang [1986]). However as Dwork noted in Dwork and Smith [2010] the criteria used to judge whether a method was deemed "private" or not were ad-hoc or otherwise inadequate. This lead to the development of "Differential Privacy" beginning in around 2002. Just as cryptographers use the notion of "semantic security" (see e.g., Goldreich [1998, 2004]) to judge whether an encryption scheme is secure, the differential privacy gives a mathematically rigorous framework for privacy of data releases. Since its inception there has been a great deal of interest in differential privacy, with numerous surveys already written (see e.g., Dwork [2008]; Wasserman and Zhou [2010]; Dwork and Smith [2010]). What's more a great many statistical methods have been instantiated in this framework and demonstrated to have good theoretical properties (e.g., good utility as the size of the data increases). Examples include point estimation (Smith [2008]; Dwork and Lei [2009]), classification (Chaudhuri *et al.* [2011]; Rubinstein *et al.* [2010]), regression (Dwork and Lei [2009]), contingency table estimation (Rudelson *et al.* [2010]; Hardt *et al.* [2010]; Barak *et al.* [2007]; Fienberg *et al.* [2010]), histogram estimation (Wasserman and Zhou [2010]; Smith [2009]), density estimation by means of expansion into a truncated basis (Wasserman and Zhou [2010]), as well as modern machine learning methods such as recommendation systems (McSherry and Mironov [2009]) among others.

Nevertheless there remains something of a divide that between the theoretical literature on the differential privacy, which is located mostly in the computer science community, and the statistical analyses faced by researchers in e.g., the U.S. Census Bureau, who seek to apply these techniques to real problems. For example, to employ the techniques of differential privacy in the task of synthetic data generation it was necessary to relax the privacy criteria in order to maintain any utility (Machanavajjhala *et al.* [2008]). Likewise, Fienberg *et al.* [2010]; Yang *et al.* [2012]; Charest [2012] noted that for sparse contingency tables that the techniques suggested in e.g., Barak *et al.* [2007] tend to destroy utility of the data.

In essence it appears that the theory of differential privacy remains mired in the regime of the classical statistics of the early 20^{th} century. That is, the setting in which the goal is to release (or estimate) a vector of fixed dimension, where the sample size is allowed to grow to infinity. In these cases the additional noise required to achieve differential privacy is often on the same order as the sampling error Smith [2008]. It is theoretically satisfying that the asymptotics of the private estimator and of the non-private one have the same behavior, however with modern problems it is more common that the number of data samples is small relative to the dimension of the object being estimated, for example in the case of large sparse histograms and contingency tables . In these cases the noise required by the proposed techniques tends to swamp whatever utility was in the data. This occurs both in practise and also in the theory, when the dimension of the released vector is allowed to grow as the number of samples increases.

Our purpose is to bridge some of the apparent gap between the theoretical side of differential pri-

vacy research and the practical problems faced by agencies who wish to perform privacy preserving analyses.

1.1.1 Outline

We begin by recalling basic definitions and methods from the differential privacy literature, then we investigate the noise levels required for differential privacy in a number of statistical settings. Thereupon we may determine what questions may be answered appropriately while preserving privacy. To do so we use known and novel lower bounds for error levels. We start out with some discrete problems, such as the release of histograms and contingency tables. We give an analogue of the minimax risk for these types of releases. We also demonstrate that in the modern situation of a large sparse table or histogram, that the minimax rate may be lower. We also give methods which achieve these lower rates in practise.

We then describe how differential privacy may be applied when the goal is to release a function itself. The main types of functions we envision are kernel density estimates and the regression functions of support vector machines. We give techniques based on truncating an expansion of these functions in an orthonormal basis. We demonstrate that since these functions lay in a reproducing kernel Hilbert space, the basis of eigenfunctions are a natural choice for the expansion. In this case differential privacy is demonstrated even for the full (non-truncated) expansion, and we find that the resulting method corresponds to the addition of a certain Gaussian process to the private function.

Although these techniques work well they use a relaxation of differential privacy which might be unappealing in certain situations (for example, De [2011] calls this relaxation "much weaker" than the standard definition). Therefore we build up a second method for kernel density estimation which does not require this relaxation. We show that the basic technique of adding noise to the coefficients of the function in some orthonormal basis fails to maintain the convergence rate of the unadulterated kernel density estimator. We instead use an alternative technique based on the geometry of the underlying space of coefficients, which may be treated as a novel application of the method of Hardt and Talwar [2010]. Under this method we find that the convergence rate is preserved up to a logarithmic factor.

We then present some weakenings of differential privacy and explain how they can lead to better accuracy at the expense of a weaker privacy guarantee.

1.1.2 Related Work

Research into privacy preserving datamining began in the 1980s in both statistics and computer science. Most of the early work may be regarded as providing methods for database access which

preserved privacy either through randomization of the response, or by explicitly limiting the kinds of queries which may be asked of the database.

It was long argued that the superior method for privacy protection was to grant users only selective or partial access to the data (see e.g., Conway and Strip [1976]). Thereupon, at the discretion of some administrator, sensitive variables in the data were obscured, either by the addition of noise, through partitioning (e.g., replacing values which fall in some range with a unique identifier for that range), or by disassociating values from the individuals to whom they correspond. More recently a similar idea has been used in order to permit access to a medical database via a web portal curated by an administrator (see e.g,. O'Keefe and Good [2008]; Sparks *et al.* [2008]; Reiter [2003b]). There, a user is allowed to perform an analysis (e.g., a linear regression, survival analysis etc) through a web interface, whereupon the server either sanitizes the returned model (for example through the addition of noise), or returns nothing if the model is deemed too disclosive of individuals' data. The determination of which type of response is appropriate for a given analysis is made on the basis of some heuristic rules.

Another family of techniques may be regarded as replacing the data itself with something "similar" but different to the private database, and then allowing full access to that data. Reiss [1984]; Liew *et al.* [1985] replace the data with samples drawn from some distribution. More recently Reiter [2003a] applied a similar technique, which built a decision trees model of the sensitive variables based on the non-sensitive ones, and Reiter [2004] similarly built a mixture of Gaussians. In either case the private data was replaced by samples from the learned models. Here since the released data did not expressly correspond to individuals in the original data (since the released data were samples from some distribution) it was regarded as private. Other techniques perform data masking via the addition of e.g., normally distributed noise (e.g., Duncan and Pearson [1991]; Duncan and Stokes [2009]). In order to perform statistical inferences on the resulting data, measurement error models are typically used (see e.g., Fuller [1993]; Duncan *et al.* [2001]). The tradeoff between the disclosure risk and the accuracy of the inference was considered in Trottini *et al.* [2004].

Finally we note that another set of techniques are the so-called matrix masking of e.g., Ting *et al.* [2008]; Duncan and Pearson [1991]; Duncan *et al.* [2011]. When the database is regarded as a matrix of real values, then matrix masking refers to the family of techniques which multiply said matrix on either side by two different potentially random matrices (which may e.g., select particular rows and columns of the data) and finally adding a third matrix to the result. Ting *et al.* [2008] chose these random matrices specifically to preserve the sufficient statistics to a multivariate Gaussian model of the data, while masking the actual values themselves. Thus these techniques are well suited to statistical analysis of the private data.

In parallel to the development of methods for preserving privacy, was a line of work focused on determining appropriate measures of privacy protection. Agrawal and Srikant [2000] determines the

level of privacy by the sizes of the confidence intervals that may be determined for the individual data elements (for example when adding noise to the data itself, or to the result of some statistical analysis). Reiter [2005] considered the probability that a unique individual in the sanitized data may be linked to some external file (thus revealing his sensitive information).

Motivated by the prospect of such "linkage attacks" on sanitized data, Sweeney [2002] proposed the "k-anonymity" criteria. This was a standard of privacy with a parameter which controlled the probability that a correct linkage could be made. Eventually it was determined that there was still the prospect for sensitive information to leak from such methods, leading to the creation of further improvements (see e.g., Kifer and Gehrke [2006]; Li and Li [2007]). In hindsight, these criteria may be regarded as reactionary in that they each prevented only certain kinds of "attacks" on the sanitized data which had appeared in the literature.

Dinur and Nissim [2003] first attempted to quantify the privacy of data release. There, the type of analyses permitted on the database were restricted to counting queries (e.g., to determine how many individuals meet a certain criteria). Privacy was determined by whether an adversary with access to the answers of such queries can reconstruct a non-trivial fraction of the database. This eventually lead Blum *et al.* [2005] to conceptualize privacy as rendering the private database indistinguishable from hypothetical similar databases. If an attacker cannot distinguish between a large set of input databases, any of which may have produced the same results to his queries, then he cannot with confidence claim to have uncovered anything beyond the characteristics intrinsic to all the elements of that set. This lead Dwork [2006] to develop differential privacy .

Since the inception of differential privacy the overwhelming majority of computer science research in privacy preserving datamining has adopted that notion of privacy. Blum *et al.* [2005] demonstrated that enough linear queries of a database may be answered with sufficient accuracy to perform machine learning on the data, and Dwork *et al.* [2006a] extended this to secure multiparty protocols. These may be regarded as predominantly theoretical constructions which asserted the learnability of important classes of functions under the differential privacy criterion. The work which immediately followed these was concerned with the constructions of differential privacy preserving methods which are useful in practise. The first set of methods involved noise addition in which the magnitude of the additive noise depended on the "sensitivity" of the released function (Dwork [2006]). After this there were attempts to reduce the amount of noise either through more complicated measures of the sensitivity (Nissim *et al.* [2007]), or through more elaborate treatments of the functions themselves (Dwork *et al.* [2006b]). In order to permit statistical inferences of a type which may not admit the requisitely bounded sensitivity, Dwork and Lei [2009] proposed to employ the methods of robust statistics. The non-private versions of the robust estimators may be intuitively seen to be less disclosive, since the affect of an individual data point on the output is reduced.

1.2 Differential Privacy

Here we recall the definition of differential privacy and introduce some notation. Let $D = (d_1, \ldots, d_n) \in \mathcal{D}$ be an input database in which d_i represents a row or an individual, and where $\mathcal{D}$ is the space of all such databases of n elements. For two databases D, D', we say they are "adjacent" or "neighboring" and write $D \sim D'$ whenever they differ in one element. That is, D' is constructed from D by removing one element and inserting a different element. Thus $D \sim D'$ means that $|D| = |D'|$ and that the size of the symmetric difference is 2. In some other works databases are called "adjacent" whenever one database contains the other together with exactly one additional element.

We may characterize a non-randomized algorithm in terms of the function it outputs, e.g., $\theta : \mathcal{D} \to \mathbb{R}^d$. Thus we write $\theta_D = \theta(D)$ to mean the output when the input database is D. Thus, a computer program which outputs a vector may be characterized a family of vectors $\{\theta_D : D \in \mathcal{D}\}$. Likewise a computer program which is randomized may be characterized by the distributions $\{P_D : D \in \mathcal{D}\}$ it induces on the output space (for example some Euclidean space $\mathbb{R}^d$) when the input is D. Since the non-randomized algorithms are in essence degenerate forms of randomized algorithms (in which the distributions concentrate to single points in the output space), we only consider randomized algorithms, which are also referred to as "mechanisms." We define differential privacy using this characterization of randomized algorithms.

Definition 1.2.1 (Differential Privacy)**.** A mechanism $P = \{P_D : D \in \mathcal{D}\}$ on $(\Omega, \mathcal{A})$ is called (α, β)-differentially private whenever

$$\forall D \sim D', \forall A \in \mathcal{A} : P_D(A) \leq e^{\alpha} P_{D'}(A) + \beta, \tag{1.1}$$

where $\alpha, \beta \geq 0$ are parameters.

Typically the above definition is called "approximate differential privacy" whenever $\beta > 0$, and "$(\alpha, 0)$-differential privacy" is shortened to "α-differential privacy."

1.2.1 Interpretation

The definition is strong because it protects the data against an imagined adversary who has access to a great deal of side information. Specifically, an adversary who knows all but one element of D, and who observes the output of some private algorithm (which we treat as a sample from P_D) cannot with confidence determine the values of the other data element. We may formulate the guarantee given by differential privacy in terms of the hypothesis test that may be performed by said adversary.

Proposition 1.2.2. *For $D = \{x_1, \ldots, x_n\}$, $a \sim P_D$ where P_D is from some (α, β)-differentially*

private mechanism, when testing the hypothesis

$$H : D = D_H, \ vs \ V : D = D_V,$$

where

$$D_H = \{x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n\}, \quad D_V = \{x_1, \ldots, x_{i-1}, x', x_{i+1}, \ldots, x_n\},$$

any test function $\Psi(a)$ with size γ has power smaller than $e^\alpha \gamma + \beta$.

Proof. Note that the rejection region of the test $R(\Psi) = \{a : \Psi(a) = 1\}$ is a measurable set and so $R(\Psi) \in \mathcal{A}$. Also $D_H \sim D_V$ and so

$$\int_{R(\Psi)} dP_{D_V}(a) \leq e^\alpha \int_{R(\Psi)} dP_{D_H}(a) + \beta,$$

since P_{D_H} is the distribution corresponding to the null hypothesis and P_{D_V} corresponds to the alternative, these integrals give the power and size of the test respectively, leading to the stated result. □

The above proposition is a modified version of Lemma 2.4 of Wasserman and Zhou [2010] – the modification being to incorporate the parameter β and simplify the proof. Note that the extension to the case $V : x_i \neq x$ is straightforward, and the bound on the power holds uniformly throughout the alternative hypothesis (since the differential privacy condition holds uniformly for all neighboring data sets). Also note that in the proof x_j for $j \neq i$ could take on any values whatsoever. The interpretation is that the adversary when constructing the test Ψ may indeed know the exact values of these x_j. Whether he does or not, he is still unable to reliably decide the value of x_i. Therefore the adversary can be seen as having almost complete data, in the language of cryptography this is his "side information." When the x_i correspond to individuals the above is equivalent to the adversary knowing the identity of all but one individual in the database, but still being unable to decide the identity of the unknown one.

There is a second interpretation of differential privacy which is in the context of individuals deciding whether to deliberately misreport their values in some database. For example, deciding whether to give truthful responses in some survey, knowing that the responses will be made public via some differentially private method. As remarked in McSherry and Talwar [2007], differential privacy ensures that individuals have very little incentive to misreport values, where the incentive is controlled by the parameter α.

1.2.2 Approximate Differential Privacy

We note that the approximate differential privacy may be regarded as permitting the failure of the exact differential privacy criteria, with probability at most β, where the probability is due to P_D.

Proposition 1.2.3. *Suppose that, for all $D \sim D'$, there exists a set $A^\star_{D,D'} \in \mathcal{A}$ such that, for all $S \in \mathcal{A}$,*

$$S \subseteq A^\star_{D,D'} \Rightarrow P_D(S) \leq e^\alpha P_{D'}(S) \tag{1.2}$$

and

$$P_D(A^\star_{D,D'}) \geq 1 - \beta. \tag{1.3}$$

Then the mechanism $\{P_D\}$ achieves the (α, β)-DP.

Proof. Let $S \in \mathcal{A}$. Then,

$$\begin{aligned} P_D(S) &= P_D(S \cap A^\star) + P_D(S \cap A^{\star C}) \leq P_D(S \cap A^\star) + \beta \\ &\leq e^\alpha P_{D'}(S \cap A^\star) + \beta \leq e^\alpha P_{D'}(S) + \beta. \end{aligned}$$

The first inequality is due to (1.3), the second is due to (1.2) and the third is due to the subadditivity of measures. □

Thus the set $A^\star_{D,D'}$ may be regarded as a set on which α-differential privacy holds, for $\beta = 0$ then this is evidently a set of measure one and thus the complement of $A^\star_{D,D'}$ has measure zero under both P_D and $P_{D'}$ so the α-differential privacy condition holds everywhere, rather than almost everywhere. Note that the above is called "pointwise differential privacy" by Kasiviswanathan and Smith [2008].

The converse of the above is evidently not true. Consider a pair of distributions P, Q over $0, 1$ with

$$P(0) = Q(1) = \frac{1}{2}\left(1 + \frac{e^\alpha - 1}{e^\alpha + 1}\right), \quad P(1) = Q(0) = \frac{1}{2}\left(1 - \frac{e^\alpha - 1}{e^\alpha + 1}\right),$$

and so

$$\frac{P(0)}{Q(0)} = \frac{Q(1)}{P(1)} = e^\alpha,$$

then perturbing these distributions to e.g.,

$$P'(0) = Q'(1) = \frac{1}{2}\left(1 + \frac{e^\alpha - 1}{e^\alpha + 1}\right) + \frac{\beta}{e^\alpha + 1}, \quad P'(1) = Q'(0) = \frac{1}{2}\left(1 - \frac{e^\alpha - 1}{e^\alpha + 1}\right) - \frac{\beta}{e^\alpha + 1},$$

gives a pair of distributions which fulfil the (α, β)-differential privacy but for which there is no non-empty set of the output space in which the ratio of probabilities is bounded by e^α on each

subset. Therefore the interpretation of approximate differential privacy as having some "failure probability" given by β is not generally true. The above proposition is thus only helpful in that it gives a way to construct private methods.

As noted in De [2011], the approximate differential privacy is much weaker than the unadulterated differential privacy in some situations.

1.2.3 Composition

Note that the α-differential privacy has the property that it "composes" nicely.

Proposition 1.2.4. *Let $P = \{P_D : D \in \mathcal{D}\}$ on $(\Omega, \mathcal{A})$ be α-differentially private, and let $Q = \{Q_D : D \in \mathcal{D}\}$ on $(\Omega', \mathcal{A}')$ be α'-differentially private. Then the mechanism R_D on the product space $(\Omega \times \Omega', \mathcal{A} \otimes \mathcal{A}')$ given by*

$$R_D(A \times A') = P_D(A)Q_D(A'),\ \forall A \times A' \in \mathcal{A} \otimes \mathcal{A}',$$

satisfies $\alpha + \alpha'$-differential privacy. Here $\mathcal{A} \otimes \mathcal{A}'$ is the smallest σ-field which contains the sets of the form $A \times A'$ for all $A \in \mathcal{A}$ and $A' \in \mathcal{A}'$.

Proof.

$$R_D(A \times A') = P_D(A)Q_D(A') \leq e^{\alpha} P_{D'}(A) e^{\alpha'} Q_{D'}(A') = e^{\alpha+\alpha'} R_{D'}(A \times A').$$

□

Evidently a similar property holds for approximate differential privacy. This demonstrates first that the differential privacy guarantee given above degrades gracefully (rather than completely failing) when multiple programs are all producing output regarding the same dataset. Thus some burden is taken off the practitioner, since he does not have to determine what other privacy preserving analyses have taken place on his input data. Second it demonstrates that differentially private procedures may be constructed from differentially private sub-procedures. For example if the goal is to perform a regression, it may be necessary to choose the regularization parameter. However this choice itself depends on the data (and hence has the capability to leak information about the dataset). In light of the above proposition, one method is to perform the regression say k times, each time with a different value of the regularization parameter (e.g., the values taken from some grid which spans a few orders of magnitude), use some α/k-differentially private method to estimate the risks, and then choose the minimizer among those to be output. The procedure will then have the overall α-DP guarantee.

1.2.4 Post-Processing

Another appealing characteristic of the differential privacy is that arbitrary "post-processing" by the recipient of the output does not alter the privacy guarantee.

Proposition 1.2.5. *Let $\{P_D : D \in \mathcal{D}\}$ on $(\Omega, \mathcal{A})$ achieve the (α, β)-differential privacy, and let $f : \Omega \to S$ be some arbitrary measurable function, then the induced measures on $(S, \mathcal{B})$ given by*

$$Q_D(B) = P_D\left(f^{-1}(B)\right), B \in \mathcal{B}$$

also achieve the (α, β)-differential privacy.

Proof. Due to measurability of f, $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$ and so

$$Q_D(B) = P_D\left(f^{-1}(B)\right) \leq e^{\alpha} P_{D'}\left(f^{-1}(B)\right) + \beta = e^{\alpha} Q_{D'}(B) + \beta.$$

□

Note that the σ-field $\mathcal{B}$ may be taken as the one generated by f acting upon $\mathcal{A}$.

1.3 Basic Methods

Here we summarize some of the basic methods which have been used in the literature for the differentially private release of finite dimensional vectors. Many types of statistical analysis and machine learning methods are characterized by finite dimensional vectors, for example point estimation of finite dimensional models Smith [2008], histograms Dwork and Smith [2010], logistic regression coefficients Chaudhuri *et al.* [2011], parameters to a linear support vector machine Rubinstein *et al.* [2010]; Chaudhuri *et al.* [2011], contingency tables Barak *et al.* [2007] and others. For all these types of analysis we may interpret the non-private method as outputting a vector, depending on the input database. Thus we may characterize a particular analysis as a set of vectors

$$\{\theta_D : D \in \mathcal{D}\} \subseteq \mathbb{R}^p, \quad \theta_D = \theta(D),$$

in which $\theta(D)$ is the result of the non-private algorithm on input D, and we note that the dimension of each θ_D is the same. In the case that they are not, but the maximum dimension of θ_D over all D is bounded above, then this constraint may be worked around by embedding the vectors into some appropriate higher dimensional space. In moving to private versions of these vectors it will be important that they all have the same dimension, since otherwise the dimension itself would potentially reveal information about D.

When the goal is the differentially private output of some vector, the standard approach is to devise a randomized algorithm which outputs something "close" to the requisite vector with high probability. Thus, to construct a set of measures P_D over $\mathbb{R}^p$ which fulfil the differential privacy and for which the "risk"

$$\int_{\mathbb{R}^p} \left\| \theta_D - \tilde{\theta}_D \right\|_2^2 \, dP_D(\tilde{\theta}_D),$$

is small. We examine the risk in greater detail in the next chapter. Next we proceed with some comparisons of three methods for constructing the measures P_D which have appeared in the literature.

1.3.1 Noise Addition

The prototypical differentially private procedures are the addition of Laplace (also called Double-Exponential) noise when $\beta = 0$, and Gaussian noise, when $\beta > 0$. Sometimes these techniques are called "perturbation" (see e.g., Wasserman and Zhou [2010]). Thus each P_D is constructed by centering the appropriate noise distribution at θ_D. In each case the noise is calibrated to the "sensitivity" of the vector in question. If one element of D can potentially have a large impact on the vector, then a large amount of noise may be required.

Proposition 1.3.1. *The mechanism with output space $\mathbb{R}^d$ defined by*

$$dP_D(x) \propto exp\left\{ -\frac{\alpha}{\Delta_1} \|x - \theta_D\|_1 \right\}, \tag{1.4}$$

achieves the α-differential privacy whenever

$$\sup_{D \sim D'} \|\theta_D - \theta_{D'}\|_1 \leq \Delta_1. \tag{1.5}$$

Proof. The result follows directly from the subadditivity property of norms. This proof is from Dwork [2008]. First note that since all P_D range over $\mathbb{R}^d$, they all have the same normalizing constant and so

$$\begin{aligned} \frac{dP_D(x)}{dP_{D'}(x)} &= \exp\left\{ \frac{\alpha}{\Delta_1} \left(\|\theta_{D'} - x\|_1 - \|\theta_D - x\|_1 \right) \right\} \\ &\leq \exp\left\{ \frac{\alpha}{\Delta_1} \|\theta_D - \theta_{D'}\|_1 \right\} \\ &\leq \exp\left\{ \alpha \right\}. \end{aligned}$$

Finally since the ratio of densities is bounded we have

$$P_D(A) = \int_A dP_D \leq \int_A e^{\alpha} dP_{D'} = e^{\alpha} P_{D'}(A),$$

which is the desired result. □

The densities (1.4) correspond to the Laplace or double exponential pdf. This is a location family with the location given by θ_D. Thus the release of $\tilde{\theta} \sim P_D$ may be achieved by taking $\tilde{\theta} = \theta_D + L$ where L is drawn from the Laplace distribution having mean zero and scale parameter Δ_1/α.

The quantity Δ_1 above is referred to as the "global sensitivity" of the function θ. Note that it typically depends on n, the size of the input database. When for example θ is the mean of the data, and the data lay in a compact set, then the global sensitivity of θ is on the order $O(n^{-1})$. In light of the above remark, evidently the average magnitude of the added noise will be proportional to the sensitivity and inversely proportional to α.

Note that in the above proof the norm $\|\cdot\|_1$ could be replaced by any norm whatsoever (in both the density and the sensitivity), since the triangle inequality (i.e., the subadditivity property of the norm) was the entirety of the proof. This is the basis for the "K-norm" method due to Hardt and Talwar [2010] which is described below, but first note that we may replace the norm with a Euclidean norm and the resulting algorithm is still characterized by noise addition from a certain parametric family. For some fixed covariance matrix Σ denote the Euclidean norm

$$\|x\|_\Sigma^2 = \|\Sigma^{-1/2} x\|_2^2 = x^T \Sigma^{-1} x, \tag{1.6}$$

we have

Proposition 1.3.2. *The mechanism defined by*

$$dP_D(x) \propto exp\left\{-\frac{\alpha}{\Delta_\Sigma}\|x - \theta_D\|_\Sigma\right\}, \tag{1.7}$$

fulfils the α-differential privacy whenever

$$\sup_{D \sim D'} \|\theta_D - \theta_{D'}\|_\Sigma \leq \Delta_\Sigma. \tag{1.8}$$

Sampling such a density is equivalent to taking

$$\tilde{\theta} = \theta_D + X, \quad X \sim dP(x) \propto \exp\left\{-\frac{\alpha}{\Delta}\|x\|_\Sigma\right\}.$$

Noise Generation

We turn to the generation of such a random variable as suggested above. Denoting by $\|\cdot\|_2$ the usual Euclidean norm, we note that it suffices to generate

$$y \sim \frac{1}{Z}\exp\left\{-\frac{\alpha}{\Delta_\Sigma}\|y\|_2\right\},$$

where Z is the normalization constant, since $x = \Sigma^{1/2}y$ has distribution

$$x \sim \frac{|\Sigma^{-1/2}|}{Z}\exp\left\{-\frac{\alpha}{\Delta_\Sigma}\|\Sigma^{-1/2}x\|_2\right\} = \frac{1}{Z|\Sigma|^{1/2}}\exp\left\{-\frac{\alpha}{\Delta_\Sigma}\|x\|_\Sigma\right\}.$$

We decompose $y = r\theta$ in which θ is a vector having $\|\theta\|_2 = 1$ and being distributed uniformly over the unit sphere in $\mathbb{R}^d$, whereas $r \geq 0$ is a scalar having density

$$f(r) \propto S_d(r)\ \exp\left\{-\frac{\alpha}{\Delta_\Sigma}r\right\}, \quad S_d(r) = \frac{2\pi^{d/2}r^{d-1}}{\Gamma(d/2)},$$

the latter being the surface area of the sphere of radius r in d dimensions. We recognize this to be the Gamma distribution, having "shape" parameter d and "scale" parameter Δ_Σ/α. Therefore we have

$$r \sim \frac{\left(\frac{\alpha}{\Delta_\Sigma}\right)^d}{\Gamma(d)}r^{d-1}\exp\left\{-\frac{\alpha}{\Delta_\Sigma}r\right\}.$$

1.3.2 Noise Addition for Approximate Differential Privacy

Likewise approximate differential privacy may be achieved by the addition of an appropriately scaled Gaussian vector.

Proposition 1.3.3. *Under condition (1.8) the mechanism defined by the densities*

$$dP_D(x) \propto exp\left\{-\frac{1}{2}\left(\frac{\alpha}{c(\beta)\Delta_\Sigma}\right)^2 (x-\theta_D)^T\,\Sigma^{-1}\,(x-\theta_D)\right\}$$

achieves (α, β)-differential privacy whenever

$$c(\beta) \geq \sqrt{2\log\frac{2}{\beta}}. \tag{1.9}$$

Proof. Consider the ratio of the densities

$$\frac{dP_D(x)}{dP_{D'}(x)} = \exp\left\{\frac{\alpha^2}{2c(\beta)^2\Delta_\Sigma^2}\left[(x-\theta_{D'})\Sigma^{-1}(x-\theta_{D'}) - (x-\theta_D)^T\Sigma^{-1}(x-\theta_D)\right]\right\}.$$

This ratio exceeds e^α only when

$$2x^T\Sigma^{-1}(\theta_D - \theta_{D'}) + \theta_{D'}^T\Sigma^{-1}\theta_{D'} - \theta_D^T\Sigma^{-1}\theta_D \geq 2\frac{c(\beta)^2\Delta_\Sigma^2}{\alpha}.$$

We consider the probability of this set under P_D, in which case we have $x = \theta_D + \frac{c(\beta)\Delta_\Sigma}{\alpha}\Sigma^{1/2}z$, where z is an isotropic normal with unit variance. We have

$$\frac{c(\beta)\Delta}{\alpha}z^T\Sigma^{-1/2}(\theta_D - \theta_{D'}) \geq \frac{c(\beta)^2\Delta^2}{\alpha^2} - \frac{1}{2}(\theta_D - \theta_{D'})^T\Sigma^{-1}(\theta_D - \theta_{D'}).$$

Multiplying by $\frac{\alpha}{c(\beta)\Delta_\Sigma}$ and using (1.8) gives

$$z^T\Sigma^{-1/2}(\theta_D - \theta_{D'}) \geq \frac{c(\beta)\Delta_\Sigma}{\alpha} - \frac{\alpha\Delta_\Sigma}{2c(\beta)}.$$

Note that the left side is a normal random variable with mean zero and variance smaller than Δ_Σ^2. The probability of this set is increasing with the variance of said variable, and so we examine the probability when the variance equals Δ_Σ^2. We also restrict to $\alpha \leq 1$, and let $y \sim \mathcal{N}(0,1)$, yielding

$$\begin{aligned} P\left(z^TM^{-1/2}(\theta_D - \theta_{D'}) \geq \frac{c(\beta)\Delta_\Sigma}{\alpha} - \frac{\alpha\Delta_\Sigma}{2c(\beta)}\right) &\leq P\left(\Delta_\Sigma y \geq \frac{c(\beta)\Delta_\Sigma}{\alpha} - \frac{\alpha\Delta_\Sigma}{2c(\beta)}\right) \\ &\leq P\left(y \geq c(\beta) - \frac{1}{2c(\beta)}\right) \leq \beta \end{aligned}$$

where $c(\beta)$ is as defined in (1.9). Thus proposition 1.2.3 gives the differential privacy, the final inequality arises from the Gaussian tail inequality. □

In principle the above technique will typically add "less noise" than the former, in the sense that the noise variable will have a small magnitude with high probability, however this comes at the price of a weakened privacy guarantee.

1.3.3 The K-norm Mechanism

Here we give a modified version of the "K-norm" method due to Hardt and Talwar [2010]. It was originally proposed for the specific task of the release of some linear function of a database which was regarded as a vector of counts (i.e., a histogram). Here we modify the approach so that it

applies to the release of an arbitrary statistic.

Recall that associated with any convex body $K \subset \mathbb{R}^d$ is a seminorm

$$\|x\|_K = \inf \{r > 0 : x \in r \cdot K\},$$

where

$$r \cdot K = \{rv : v \in K\},$$

is the scaling of the set K by some real number r. Whenever K is centrosymmetric $\|\cdot\|_K$ is a norm since it also has the property that $\|v\|_K = 0$ holds only for $v = 0$. Evidently this seminorm has the requisite subadditivity needed for the usual construction of a differentially private mechanism (i.e., in the proof of Proposition 1.3.1), and therefore we have

Proposition 1.3.4. *The mechanism defined by the densities*

$$dP_D(x) \propto exp\left\{-\frac{\alpha}{\Delta_K}\|\theta_D - x\|_K\right\},$$

achieves the α-differential privacy whenever

$$\Delta_K \geq \sup_{D \sim D'} \|\theta_D - \theta_{D'}\|_K.$$

When we have for example

$$\{\theta_D : D \in \mathcal{D}\} \subseteq K \subset \mathbb{R}^d,$$

for some convex set K, then evidently $\Delta_K \leq 2$. Note that when K is a unit sphere in the Euclidean norm then the K-norm method is the same as (1.7) with $\Sigma = I$. Likewise when K is the unit ball in the ℓ_1 norm then the K-norm method becomes (1.4).

Finally we note the difference in the above presentation of the K-norm mechanism versus the presentation given in Hardt and Talwar [2010]. There the task at hand was specifically the release of a linear function of the database, where the latter was regarded as a vector of integers (the coordinates corresponding to the number of each "type" of individuals present in the database). Thus they considered the convex set given by the image of the appropriately dimensioned ℓ_1-ball under the requisite linear function. Evidently in this situation Δ_K may be taken as 1, which leads to their method.

What remains to be seen are a method for performing the sampling according to the above measures, and to determine the expected error due to this approach in the general case. Since the

$P_D(x)$ are clearly a location family having location θ_D, sampling these is equivalent to sampling

$$\tilde{\theta} = \theta_D + X, \quad X \sim dP(x) \propto \exp\left\{-\frac{\alpha}{\Delta_K}\|x\|_K\right\}.$$

In Hardt and Talwar [2010] it is demonstrated that the latter distribution may be treated as a composition of two parts just as in the sampling of (1.7), namely to take

$$X = rV, \quad r \sim \Gamma\left(d+1, \frac{\Delta_K}{\alpha}\right), \quad V \sim \text{Uniform}(K),$$

where the latter means the uniform distribution over K.

1.3.4 The Exponential Mechanism

Note that the above techniques all result in the output of real values. In some types of problems the output is a discrete structure rather than a real vector (for example, histograms, counts, contingency tables, and decision trees). In such cases the addition of real valued noise to the output might spoil some important properties of the output (for example, to cause the histogram to sum to some quantity other than n), or may be meaningless (for example to add noise to a decision tree). To address such problems the Exponential Mechanism was introduced by McSherry and Talwar [2007]. The basic idea may be regarded as introducing some metric to the discrete output space, and then applying a discrete analogue of the Laplace noise addition.

Proposition 1.3.5. *Let Ω be some set and $g : \Omega \times \Omega \to \mathbb{R}^{0+}$ the latter meaning the non-negative real numbers. Then for output $\omega_D \in \Omega$ the probability distributions*

$$P_D(\omega) \propto exp\left\{-\frac{\alpha}{2\Delta_g} g(\omega, \omega_D)\right\},$$

provide α-differential privacy whenever

$$\Delta_g \geq \sup_{\omega\in\Omega} \sup_{D\sim D'} |g(\omega, \omega_D) - g(\omega, \omega_{D'})|.$$

Proof. Define

$$P_D(\omega) = \frac{Q_D(\omega)}{Z_D}$$

with

$$Q_D(\omega) = \exp\left\{-\frac{\alpha}{2\Delta_g} g(\omega, \omega_D)\right\},$$

and the normalization constants

$$Z_D = \int_\Omega dQ_D(\omega).$$

Then for each $\omega \in \Omega$, when $D \sim D'$

$$e^{-\alpha/2} \leq \frac{Q_D(\omega)}{Q_{D'}(\omega)} \leq e^{\alpha/2},$$

and so

$$e^{-\alpha/2} \leq \frac{Z_D}{Z_{D'}} \leq e^{\alpha/2},$$

Thus finally

$$e^{-\alpha} \leq \frac{Q_D(\omega)/Z_D}{Q_{D'}(\omega)/Z_{D'}} \leq e^{\alpha}.$$

□

Note that if g is a metric then it suffices to take

$$\Delta_g \geq \sup_{D \sim D'} g(\omega_D, \omega_{D'}),$$

in the case $\Omega = \mathbb{R}$ and $g(a, b) = |a - b|$, then the similarity to the Laplace noise addition is clear. The difference is that in general the normalization constants may be different (for example when Ω is some interval in $\mathbb{R}$), which leads to the factor of two in the exponent of the probability density.

1.3.5 Optimal Discrete Mechanisms via Linear Programming

Suppose $\theta(D)$ takes on a finite number of values, for simplicity consider the counting queries $\theta(D) \in N = \{1, \ldots, n\}$ given by

$$\theta(D) = \sum_{i=1}^{n} \mathbf{1}\left\{x_i \in B\right\},$$

for some set B. The extension to various other discrete structures is straightforward. Under the above condition, note that a differentially private method for the release of $\theta(D)$ may be characterized by a set of n discrete probability distributions over N, since we lose nothing (see Gupte and Sundararajan [2010] appendix A) by regarding the set of probability distributions as being indexed by $\theta(D)$ rather than the D themselves, namely

$$P_D(i) = P_{\theta(D)}(i) = Q(i, \theta(D)).$$

Here Q may be regarded as a $n \times n$ matrix $Q = q_{i,j}$ of real numbers. The constraints that the P_D all obey differential privacy, and form valid probability distributions take the form

$$\sum_{i=1}^{m} q_{i,j} = 1, \ \forall j \in N, \tag{1.10}$$

$$q_{i,j} - e^{\alpha} q_{i,j+1} \leq 0, \quad q_{i,j+1} - e^{\alpha} q_{i,j} \leq 0, \quad \forall j = 1, \ldots, m-1, \forall i \in N, \tag{1.11}$$

$$q_{i,j} \geq 0, \forall i, j \in N.$$

Regarding Q as a vector in $\mathbb{R}^{m^2}$ (e.g., by stacking the columns together) the above constraints form a bounded closed convex polyhedron with a non-empty interior. It is bounded due to e.g. (1.10) which ensures that the interior is a subset of the n-fold cartesian product of n-dimensional simplices. It is clearly a convex polyhedron since all the constraints are linear inequalities, finally it certainly has a non-empty interior because differentially private methods exist for this problem (for example, take the Q given by the exponential mechanism).

As noted in Hardt and Talwar [2010] and also expanded in Gupte and Sundararajan [2010], these constraints may be used along with linear programming in order to give methods which are "optimal" in the sense of minimizing a certain Bayes' risk. We specify some loss function $\ell : N \times N \to \mathbb{R}$, which characterizes how bad it is when $\theta(D) = i$ and the private method outputs j. Also we specify some prior distribution Λ on N, then we may minimize the resulting Bayes' risk, which is given by a the expectation of the pointwise risks under the prior distribution

$$R(\Lambda) = \sum_{i=1}^{n} \Lambda(i) \sum_{j=1}^{n} \ell(i,j) P_i(j).$$

As this function is linear for fixed ℓ and Λ, and the above constraints are also linear, determination of the minimizer may be achieved via linear programming. Note that this technique yields an optimal method for the counting problem, in time polynomial in n. Therefore it may be applicable for problems of a small size.

1.4 Relationship to Cryptographic Protocols

Part of the literature on privacy preserving data mining concerns the construction of cryptographic protocols Vaidya *et al.* [2005]; Lindell and Pinkas [2009]; Goldreich [2004, 1998]. A cryptographic protocol is in essence a method to compute a particular function of the data (or in the language of cryptography a "functionality," which assigns a possibly different output to each user) in a way which prevents leakage of sensitive information during the computation. Cryptographic protocols

mainly apply to data mining in cases when the input data is split between two or more agencies (parties). An example is the analysis of the data held by the census bureau, combined with tax data held by the IRS. In this case since the parties involved are precluded from sharing the information with each other they may perform a cryptographic protocol which will yield the same response as though they had shared their data, but in a way which maintains the appropriate privacy.

A protocol for computing the functionality is just a sequence of steps, consisting of parties performing local computations, and sending intermediate messages to each other. There are various cryptographic models which are used to judge the level of security afforded by a protocol. One which is popular in the privacy preserving data mining literature is the so-called "semi-honest" (or "honest but curious") model. In this model it is assumed that parties will obey the protocol (and do not try to e.g., inject malformed data or otherwise subvert the protocol) but keep a transcript of all the messages they receive. Intuitively, a protocol is secure in this setting whenever the intermediate messages give no information about the secret inputs of other parties.

Formally, a protocol is secure so long as there exists a set of polynomial time algorithms (on for each party involved) which, when given only the input and output of party to that party outputs a random transcript of message which is *computationally indistinguishable* from the transcript generated by a real run of the protocol. See Goldreich [2004] for a definition and discussion of computational indistinguishability. In essence, if the distribution of the sequence of messages depends only on the private input and output of that party then we can simulate messages by drawing from this distribution (so long as the random number generator returns samples which are computationally indistinguishable from draws from the distribution). The existence of a simulator shows that intermediate messages do not depend on the private input of other parties, and so the protocol is secure in the sense that parties gain no more information about each other's private inputs than that revealed by the output of the protocol.

An example of a protocol which does not achieve this definition of security is one where all parties send their data to party 1, who computes the analysis locally on the combined data and then sends it back to all other parties. In this case the messages received by party 1 consist of the data of other parties, in general it is impossible to simulate these messages given only the input and output belonging to party 1. Examples of protocols for regression which achieve the security under the semi-honest model are given in Hall *et al.* [2011]; Fienberg *et al.* [2012].

On the surface, differential privacy may be regarded as "orthogonal" to cryptographic protocols, in the sense that it prevents the output from leaking sensitive information whereas the other protects the computation itself. Thus for a multiparty setting in which multiple agencies each have private data, it may be beneficial to make use of both privacy models. That is, one could build protocols which admit both differential privacy and cryptographic security, for example by adding appropriately scaled noise to the output in a secure way. Examples of methods which achieve this

are given in the original paper by Dwork and coauthors Dwork *et al.* [2006a].

On the other hand, differential privacy may yield an alternative to the cryptographic protocols. For example, if the messages to be sent during the protocol have appropriately bounded sensitivity, then noise could be added to those in order to ensure that they admit differential privacy. Then the entire protocol would achieve differential privacy due to the composition property. This may be an attractive alternative to cryptographic protocols, since it sidesteps the need for homomorphic encryption which is sometimes too computationally burdensome in practice (especially when dealing with large data). Beimel *et al.* [2011] discuss this explicitly.

1.5 Summary

In this section a few of the basic techniques for achieving differential privacy were reviewed. We summarize the applicability of the methods in table 1.1 below. We note that in principle these techniques all follow a very similar pattern. First the output space is metrized, then the sensitivity of the output vector is bounded, and finally an output is sampled from the output space, from a measure which depends on both the input and the choice of metric. The difference between the methods listed above is simply the metric used (except for the exponential mechanism which is the discrete analog to the other methods).

Name	Output Space	Metric	Privacy Guarantee
Laplace Mechanism	Euclidean space	ℓ_1 distance	α-DP.
"Gamma Mechanism"	Euclidean space	ℓ_2 distance	α-DP.
Gaussian Mechanism	Euclidean space	ℓ_2 distance	(α, β)-DP.
K-norm Mechanism	Euclidean space	Minkowski norm	α-DP.
Exponential Mechanism	Discrete space	Arbitrary	α-DP.

Table 1.1: Applicability of the basic techniques of Differential Privacy.

We note that table 1.1 leaves out some plausible methods which have so far not appeared in use. Namely the (α, β)-DP equivalents of the K-norm and Exponential Mechanisms. For example, it seems plausible that by squaring the metric where it appears in the resulting probability distribution, would lead to something similar to the Gaussian mechanism and which would admit the approximate differential privacy.

Chapter 2

Differential Privacy and Minimaxity in Discrete Problems

Evidently the techniques of the preceding section will output a random variable, having mean equal to the requisite quantity. We may characterize the "risk" of a differentially private procedure in terms of the expected deviation from the requisite output. This parallels the same concept in statistical estimation, where estimators with small risk are favored. There the goal is to output a quantity of interest about a particular probability distribution, and the noise is due to the random nature of the data. We note that we are not referencing the "disclosure risk" which arises in the so-called "risk-utility tradeoff" Fienberg *et al.* [2010]. We suppose that differential privacy has addressed the problem of disclosure risk, and simply ask what utility may be expected of the resulting procedures. In this sense, our risk is the opposite of the normal notion of "utility." Some concepts in this chapter have been considered before in the differential privacy literature, namely in De [2011] and Hardt and Talwar [2010]. We mention similarities to these works as they arise.

2.1 Risk and Minimaxity

We define the risk as the expectation of the error introduced by the private procedure, relative to the output of the corresponding non-private procedure.

Definition 2.1.1 (Risk)**.** For some function $\theta : \mathcal{D} \to \mathcal{Z}$ where $\mathcal{Z}$ is the output space equipped with some metric ℓ, and some set of distributions $P = \{P_D : D \in \mathcal{D}\}$ define the risk

$$R(\theta, P, D) = \int_{\mathcal{Z}} \ell(\theta_D, z)\ dP_D(z). \tag{2.1}$$

For some procedure P we consider the maximum of the risk as D ranges over $\mathcal{D}$. This is the

risk of the "hardest" input for that particular procedure

$$R^\star(\theta, P) = \max_{D \in \mathcal{D}} R(\theta, P, D).$$

In designing a procedure P, it is useful to understand the theoretical best possible performance we may anticipate in terms of $R^\star$. This quantity is exactly the analog of the "minimax" risk from statistics. Here the constraint on the family P is that imposed by differential privacy (1.1). Define by $DP(\alpha, \beta)$ the set of all P that satisfy the (α, β)-DP, then we define the minimax risk as

$$R^\star_{\alpha,\beta}(\theta) = \inf_{P \in DP(\alpha,\beta)} R^\star(\theta, P).$$

It is useful to understand the dependence of the quantity on:

- the number of samples n,
- the dimension of the output d,
- the specific metric (or "loss function") ℓ,
- the parameters α, β.

Note that the first two of the above are implicitly "baked" into the definitions of $\mathcal{D}$ and $\mathcal{Z}$ respectively. In the remainder of this chapter we undertake the determination of the above quantity when θ corresponds to some statistical analysis of the data. We recall and present some novel as well as known lower bounds in the minimax risks faced by differentially private procedures, and conclude with some experiments which demonstrate the impact of such noise on the analysis itself.

2.1.1 Risk Decomposition

From the perspective of statistical analysis, the input database D may be regarded as a sequence of draws from a probability distribution. Note that such a supposition does not change the privacy guarantee established by differential privacy, but will be used in the construction of weaker privacy models below. In statistics the data are regarded as a sample from some unknown distribution F, and the goal is typically to estimate some features of F. For example, the mean and variance of the distribution, the shape of its density function, or the relationship between different variables in a multivariate setting. When the computed function $\theta : \mathcal{D} \to \mathcal{Z}$ of Definition 2.1.1 corresponds to an estimation of a functional of the distribution F, then there is already a concept of risk associated with the measurement. Overloading notation to let $\theta(F) \in \mathcal{Z}$ correspond to the requisite property of the unknown distribution, the risk is defined by:

$$R(\theta, F) = \int_{\mathcal{D}} \ell(\theta(D), \theta(F)) dF(D)$$

Here the randomness is due to the unknown measure F, and is taken over the space of databases arising from F. Here ℓ is some metric on the output space $\mathcal{Z}$. Sometimes this quantity is described as the "sampling error" inherent to the procedure θ, as it represents the average amount of inaccuracy in determining $\theta(F)$ on the basis of a random sample. When the procedure is also required to fulfil differential privacy the risk may be written as

$$\begin{aligned} R(P, \theta, F) &= \int_{\mathcal{D}} \int_{\mathcal{Z}} \ell(z, \theta(F)) \; dP_D(z) \; dF(D) \\ &\leq \int_{\mathcal{D}} \int_{\mathcal{Z}} \ell(z, \theta(D)) + \ell(\theta(D), \theta(F)) \; dP_D(z) \; dF(D) \\ &\leq \sup_{D \in \mathcal{D}} \int_{\mathcal{Z}} \ell(z, \theta(D)) \; dP_D(z) + \int_{\mathcal{D}} \ell(\theta(D), \theta(F)) \; dF(D) \\ &= R^\star(\theta, P) + R(\theta, F). \end{aligned}$$

Where the first inequality is the triangle inequality. Typically minimax risks are known (sometimes only up to constant factors) for the latter term, therefore the effect of requiring differential privacy is considered benign whenever the former term grows no faster than the latter. This means that (at least asymptotically) the requirement of differential privacy does not drastically effect the performance of the estimator.

In order to understand which statistical tasks are amenable to privatization, and which are rendered impossible by the noise requirements of differential privacy, we turn to the construction of lower risk bounds for the term $R^\star(\theta, P)$. In this section we first recall some useful inequalities which are used in statistics for the purpose of lower bounding the risks inherent in certain estimation problems. We then demonstrate how these types of inequalities may be applied to give lower bounds on the added risk due to differential privacy when the problem is the release of a histogram. We note the similarities and differences compared to the method of Hardt and Talwar [2010] and De [2011] also.

2.2 Information Theoretic Inequalities from Statistics

We recall some inequalities from information theory and statistics, see Yu [1997] for a more detailed account. These will be used to give the lower risk bounds. First is an inequality due to Fano. Suppose we have n probability distributions on any space whatsoever, say P_i as $i = 1, \ldots, n$ on $\mathcal{Z}$ (the σ-field is not important). On the basis of one sample $z \sim P_j$ from one of these distributions we

try to identify j, thus we have some function $\delta : \mathcal{Z} \to \{1, \dots, n\}$ which is our estimator. Evidently if all the distributions were close together in some statistical distance this becomes difficult. Suppose we incur some loss $\ell(P_i, P_j) = \ell(i, j)$ whenever the observation arises from P_j and we have the estimate $\delta(z) = i$. Define the risk as

$$R(j, \delta) = \int_{\mathcal{Z}} \ell(j, \delta(z)) \, dP_j(z),$$

then Fanos inequality gives

$$\inf_{\delta} \max_{k \in \{1,\dots,n\}} R(k, \delta) \geq \frac{1}{2} \min_{i \neq j} \ell(i, j) \left(1 - \frac{\max_{i \neq j} K(P_i, P_j) + \log 2}{\log n} \right), \tag{2.2}$$

thus no matter the choice of δ, the maximum risk it attains over the set $\{1, \dots, n\}$ is bounded below. Typically, in using this inequality we do not take all possible distributions in whatever our parameter space, but rather try to construct some kind of packing. That is, we try to determine a subset of the distributions over which the term $\min_{i \neq j} \ell(i, j)$ is sufficiently large. In doing so it is important that the set also has a small diameter in the KL divergence, and a size which is exponentially large, in order to make the parenthetical term non-negative. Note that restricting to a subset of all available distributions makes the lower bound no stronger, and hence is still valid for the complete problem. The reason is in essence that the larger problem is at least as hard as the hardest sub-problem.

We now describe how this kind of inequality apply to the problem of risk lower bounds for differentially private methods. We restrict attention to the $(\alpha, 0)$-differential privacy, since it make the resulting techniques amenable to an information theoretic analysis. The reason is that with $\beta = 0$ the distributions P_D must all be absolutely continuous with respect to one another. What's more, whenever densities dP_D exist (i.e., whenever there is a dominating measure), quantities such as the KL divergence exist, and so for $D \sim D' \in \mathcal{D}$ we have

$$K(P_D, P_{D'}) = \int_{\mathcal{Z}} \log \frac{dP_D(z)}{dP_{D'}(z)} \, dP_D(z) \leq \alpha.$$

What's more, denoting by d_H the Hamming distance between databases of equal size, so that $d_H(D, D')$ gives the length of the shortest path between D and D' in which each step is to a neighboring database, then

$$K(P_D, P_{D'}) \leq \alpha d_H(D, D').$$

Thus the requisite KL term of (2.2) is dealt with. The use of the function δ may seem to have no analogue in the setting of privacy. After all, we try to determine the set of distributions P_i, in order to have small risk, whereas in statistics the distributions were given and we sought the best

estimator. Nevertheless the theory still applies, note that in our setting we may consider δ as some kind of post processing procedure which "cleans up" the output of some generic private procedure. For example when the P_D are the functions arising from noise addition then δ may be interpreted as the function which projects the resulting real vector onto some discrete lattice (in the case of count queries) for example. What's more, in Gupte and Sundararajan [2010] such forms of post processing were explicitly considered.

2.3 Lower Bounds for Counting Queries

First consider the simple case of counting queries, of the form

$$\theta(D) = \sum_{i=1}^{n} \mathbf{1}\left\{x_i \in B\right\},$$

where $D = \{x_1, \ldots, x_n\}$, $x_i \in \mathcal{X}$ and $B \subseteq \mathcal{X}$. These type of statistics could be used for e.g., the estimation of a binomial proportion parameter for some statistical model.

2.3.1 A Minimax Lower Bound

If we evaluate a private method for counting queries using the absolute value loss function, we have the following bound on the minimax risk

Proposition 2.3.1. *For $\alpha < 2$ any α-differentially private family $P_D = P_{\theta(D)}$ over $0, \ldots, n$ must obey*

$$\sup_D R(\theta, P, D) \geq \frac{e^{-2}}{\alpha} \approx \frac{0.135}{\alpha}.$$

Proof. Suppose it were possible to have risk everywhere upper bounded by e^{-2}/α, then we may take any value τ and through Markov's inequality find that for all values i that the counting query may return

$$P_i([i - \tau/2, i + \tau/2 - 1]) \geq 1 - \frac{e^{-2}}{\alpha\tau},$$

and so due to the α-differential privacy

$$P_i([i - \tau/2 + t\tau, i + \tau/2 + t\tau - 1]) \geq e^{-\alpha t\tau}\left(1 - \frac{e^{-2}}{\alpha\tau}\right).$$

Hence

$$P_i([i - \tau/2, i + \tau/2 - 1]^C) \geq P_i([i - 3\tau/2, i - \tau/2 - 1] \cup [i + \tau/2, i + 3\tau/2 - 1]) \geq 2e^{-\alpha\tau}\left(1 - \frac{e^{-2}}{\alpha\tau}\right),$$

taking $\tau = 2/\alpha$ we may lower bound the risk at $\theta(D) = i$ by

$$R \geq \frac{\tau}{2} P_i([i - \tau/2, i + \tau/2 - 1]^C) \geq \frac{2e^{-2}}{\alpha}\left(1 - 2e^{-2}\right) > \frac{e^{-2}}{\alpha},$$

which is a contradiction. □

2.3.2 Uniform Lower Bounds for Counting Queries

The above minimax result demonstrates that for every α-differentially private method for a count query, there is at least one input for which the risk is growing in the order shown. However the prospect exists that at many other inputs the risk is much lower. In other words, our use of the minimax risk may be giving too conservative a view of the difficulty of this problem. Here we demonstrate that this is in fact not the case. We show that for large enough n, any procedure which achieves the minimax risk up to a constant also has risk everywhere lower bounded away from zero, in fact bounded within a constant factor of the minimax risk itself.

It is important to note that for general (i.e., non-minimax) methods it is impossible to give a uniform lower bound to the risk, for example consider the method which always outputs 0. This is differentially private for any α, and the output in fact gives no information whatsoever about the input database, however in the case that the actual count was zero then this method achieves zero risk leading to only the trivial uniform lower bound. In fact this phenomenon is not specific to counting, the same kind of counterexample may be constructed irrespective of the output space.

The basic idea of the proof is to note that if the minimax risk is sufficiently small, then the released histogram will – with high probability – be contained in a ball around the true histogram. Since the probability that the output is in such a ball is not allowed to change much when the input changes, we see that for a particular input, there is some non-zero probability that the output is in the ball around a different true histogram, leading to non-zero risk.

Proposition 2.3.2. *Any α-differentially private method P_D which achieves*

$$\sup_D \sum_{i=0}^{n} \left|\theta_D - \tilde{\theta}\right| P_D(\tilde{\theta}) \leq \frac{c}{\alpha},$$

for some constant c and for all n, also obeys

$$\inf_D \sum_{i=0}^{n} \left|\theta_D - \tilde{\theta}\right| P_D(\tilde{\theta}) \geq \frac{c}{2\alpha} e^{-2c-\alpha}.$$

Proof. Note that for any D due to the uniform upper bound on the risk Markov's inequality gives

that for any integer τ

$$P_D\left([\theta_D - \frac{\tau}{2}, \theta_D + \frac{\tau}{2}]\right) \geq 1 - \frac{c}{\alpha\tau},$$

and so due to differential privacy

$$P_{D'}\left([\theta_D - \frac{\tau}{2}, \theta_D + \frac{\tau}{2}]\right) \geq \exp\left\{-\alpha|\theta_D - \theta_{D'}|\right\}\left(1 - \frac{c}{\alpha\tau}\right).$$

Thus for any D' whatsoever we may find a D having $|\theta_D - \theta_{D'}| = \tau$, and so

$$\begin{aligned}\sum_{i=0}^{n}\left|\theta_{D'} - \tilde{\theta}\right| P_{D'}(\tilde{\theta}) &\geq \frac{\tau}{2}P_{D'}\left([\theta_{D'} - \frac{\tau}{2}, \theta_{D'} + \frac{\tau}{2}]^C\right)\\ &\geq \frac{\tau}{2}P_{D'}\left([\theta_D - \frac{\tau}{2}, \theta_D + \frac{\tau}{2}]\right)\\ &\geq \frac{\tau}{2}e^{-\alpha\tau}\left(1 - \frac{c}{\alpha\tau}\right).\end{aligned}$$

Taking

$$\tau = \left\lceil \frac{2c}{\alpha} \right\rceil,$$

leads to

$$\sum_{i=0}^{n}\left|\theta_{D'} - \tilde{\theta}\right| P_{D'}(\tilde{\theta}) \geq \frac{c}{2\alpha}e^{-2c-\alpha}.$$

Since D' was arbitrary this bound holds uniformly. □

2.3.3 Computation of Minimax Methods

For counting queries a differentially private method which is minimax may be explicitly constructed for whatever the loss function.

First recall the technique of Section 1.3.5 which gave a linear programming technique for the construction of the exact minimizer of a Bayes' risk for this problem. Here it was seen that the differentially private methods for discrete problems correspond to a certain bounded convex polyhedron.

Note that also in these discrete problems the risk function may be regarded as a vector, for each differentially private method, since

$$R(D, P) = R(\theta(D), P) = \sum_{i=1}^{n}\ell(i, \theta D)P_{\theta(D)}(i),$$

we may take the vector $R(P) \in \mathbb{R}^{n+1}$ with

$$R_i(P) = R(i+1, P).$$

We may consider the "risk body" associated with the set of differentially private methods

$$B(n, \alpha, \beta) = \{R(P) : P \in DP(\alpha, \beta)\},$$

an example of these structures in the two dimensional case (i.e., when $n = 1$ and the output is either zero or one) where the loss function is the absolute value loss is shown in Figure 2.1. Evidently for whatever the dimension and whatever the loss function the risk body will be a closed convex polyhedron in $\mathbb{R}^{n+1}$, this is clear since it is simply a linear transformation of the closed convex polyhedron corresponding to the parameters of the valid differentially private methods.

Minimax Equalizer Rules

In the case that we use the absolute value loss function i.e.,

$$\ell(i, j) = |i - j|,$$

then one way to construct a minimax procedure for these problems is to restrict to the methods with $R_i(P) = R_j(P)$ for all i, j, namely the "equalizer rules" who's risk is a constant function. Then, the rule which minimizes the Bayes' risk under any prior, subject to the constraint of constant risk will be minimax. Equivalently we could search for a particular prior under which the minimizer of the Bayes risk is an equalizer rule, however it is more straightforward to amend the linear program of Section 1.3.5 to include the constraints

$$\sum_i \ell(i, j) q_{i,j} = \sum_i \ell(i, 1) q_{i,1}, j \in 2, \ldots, n.$$

First to see that a differentially private method exists which satisfies these constraints consider

$$P_i(j) = \begin{cases} \frac{1}{2} & j = 0 \text{ or } j = n \\ 0 & \text{o/w} \end{cases},$$

since this method is oblivious to the input (the distribution over outputs are the same) it clearly satisfies differential privacy for any choice of α, β. Also the risk is

$$R(i, P) = \frac{1}{2}|i - 0| + \frac{1}{2}|i - n| = \frac{n}{2},$$

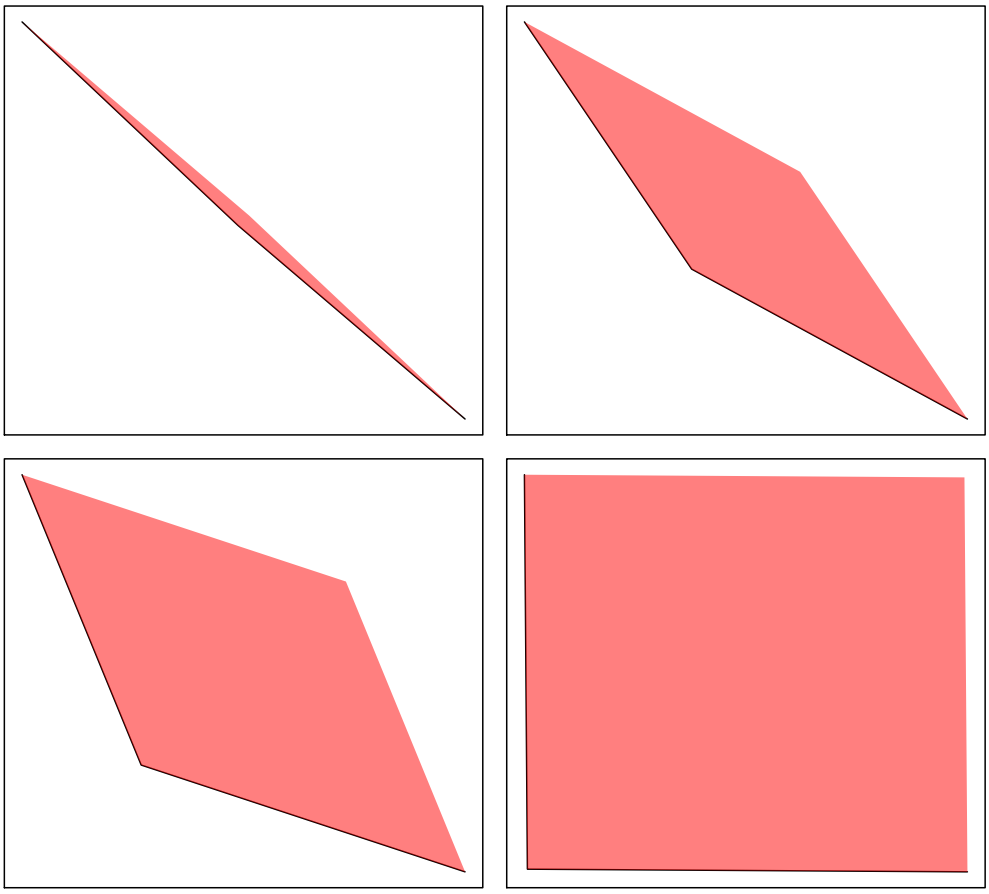


Figure 2.1: The risk body (shown as a shaded polygon) for the counting query for $n = 1$, the x-axis is the risk for $\theta = 0$ and the y-axis is the risk when $\theta = 1$. The four plots correspond to $\alpha \in \{0.05, 0.5, 1, 5\}$ respectively, the risk body becoming larger as α increases. Note that the lower left edge of each polygon correspond to the Bayes' rules computed in section 1.3.5.

which is constant. Hence the addition of the constraint above to the linear program leaves intact the non-emptiness of the feasible region, for the case of absolute value loss.

To see that the resulting methods will be minimax, note that if it were not then it would be uniformly dominated, i.e., there would be a different rule having everywhere smaller risk. Since the risk body is symmetric about the ray $R_1 = R_2 = \cdots = R_{n+1}$, for every two dimensional coordinate projection to R_{i+1}, R_{n-i} (due to the symmetry present in the loss function), we could thus construct a new equalizer rule having everywhere smaller risk than the above one, which is a contradiction.

In Figure 2.2 a plot is shown of the computed minimax risk for the count query with $n = 70$ over a range of α. We compare this to the curve $1/\alpha$ and see that while the minimax risk lies underneath the latter curve, as α decreases towards zero the two curves are brought closer together.

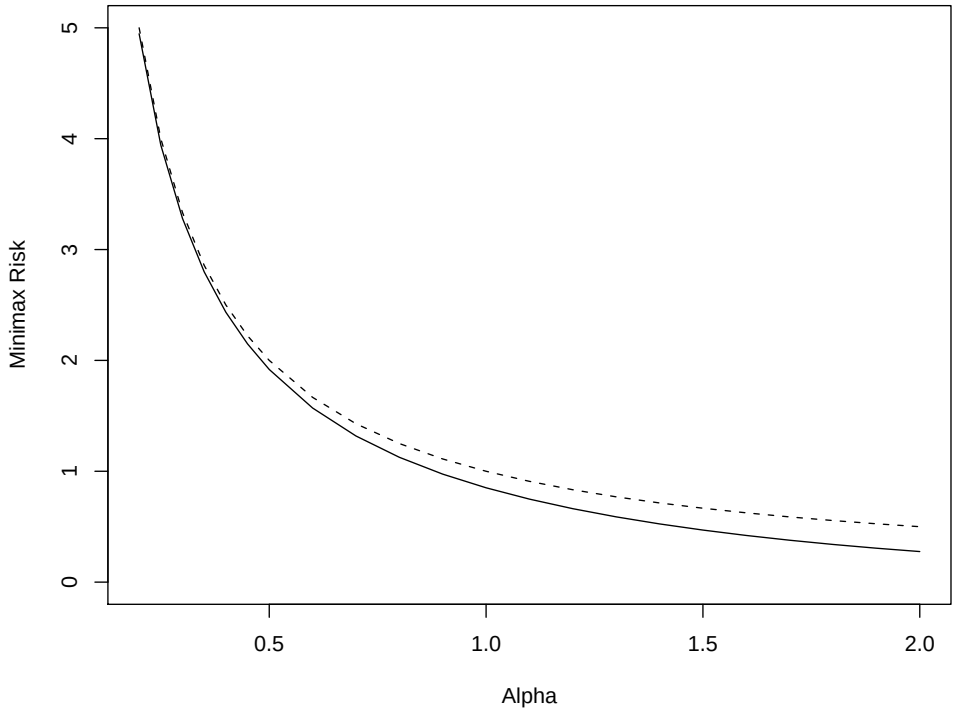


Figure 2.2: The minimax risk for the counting query with $n = 70$. The solid line is the explicitly computed minimax risk, whereas the dashed line corresponds to $1/\alpha$. The choice $n = 70$ is great enough that for all displayed α the risk is no longer increasing with n. For small enough α the risk appears to be very close to $1/\alpha$.

Therefore there is some effect for small α which the minimax lower bound of section 2.3.1 failed to acknowledge.

2.4 Lower Bounds for Histograms

A similar lower bound also applies to the release of a histogram, this consists of a function of the form

$$\theta : \mathcal{D} \to \mathbb{N}^p, \quad \theta_j(D) = \sum_{i=1}^{n} \mathbf{1}\left\{x_i \in B_j\right\},$$

where $x_i \in \mathcal{X}$, and the sets B_j form a partition of $\mathcal{X}$. This means that each point x_i contributes to exactly one coordinate θ_i. Hence the release of a histogram is "easier" than the release of p arbitrary counts (i.e., counts of non-disjoint sets).

First we restate a technical lemma which may be found in Yu [1997], which is extremely useful in simplifying the construction of a suitable packing for this problem.

Lemma 2.4.1. *In dimension $m \geq 6$, there exists a set of points $S \subseteq \{0,1\}^m$ having*

$$\min_{a \neq b \in S} \|a - b\|_1 \geq \frac{m}{3},$$

and

$$|S| \geq exp\left\{c_0 m\right\},$$

for some universal constant $c_0 > 0$.

Note that this set is a binary error correcting code with minimum distance $m/3$. We will now use this to demonstrate the following lower bound for the risk of differentially private histograms.

Proposition 2.4.2. *For the loss function $\ell(x,y) = \|x - y\|_1$, for $p > 6$ and when*

$$n \geq \frac{c_0(p-1)}{\alpha},$$

then the risk is bounded as

$$R^{\star}_{\alpha,0}(\theta) \geq c_1 \frac{p-1}{\alpha},$$

for some universal constant c_1.

Proof. Let $\tau > 0$ be some integer and consider the subset of histograms defined by

$$A = \left\{ \left(\sigma_1 \tau, \ldots, \sigma_{p-1}\tau, n - \sum_{i=1}^{p-1} \sigma_i \tau \right) : \sigma_i \in \{0,1\} \right\},$$

this set is evidently a hypercube of dimension $p-1$, and so by the above lemma whenever $p > 6$ we may construct $B \subset A$ having

$$\min_{a \neq b \in B} \|a - b\|_1 \geq \frac{(p-1)\tau}{3}, \quad |B| \geq \exp\left\{c_0(p-1)\right\}.$$

The differential privacy gives the requisite bound on the KL-divergence which is

$$\max_{D, D' : \theta(D), \theta(D') \in B} K(P_D, P_{D'}) \leq (p-1)\tau\alpha.$$

Finally using (2.2) we have

$$R^{\star}_{\alpha,0}(\theta) \geq \frac{(p-1)\tau}{6}\left(1 - \frac{\tau\alpha(p-1) + \log 2}{c_0(p-1)}\right),$$

Taking $\tau = c_0/\alpha$ leads to the desired result. $\square$

Note that the condition $p > 6$ may be removed by using Assouad's lemma rather than Fano's inequality (see Yu [1997]). The condition imposed on n in the statement ensures that the hypercube exists for our choice of τ. It may be regarded as a mild requirement since the statement is essentially that for each α, there is some n_0 above which the result holds. On the other hand if we treat $\alpha = \alpha(n)$ as a quantity which changes with n, where n is growing more slowly than $p/\alpha(n)$, then we cannot construct A as above for our choice of τ, instead we may take a hypercube having dimension $n\alpha/c_0$, which leads to a risk bound on the order of n. This is (up to a constant factor) the same risk as choosing a histogram uniformly at random from all the possible histograms having n observations in p bins (for any p), and implies that in such a regime only trivial risks are attainable.

Note that a straightforward method for differentially private histogram output is simply to add the Laplace noise to each bin (i.e., to use the technique of Proposition 1.3.1, with $\Delta_1 = 2$). In this case the risk is on the order of p/α, and hence equals the minimax risk up to a constant. This means that through the construction of more elaborate methods we may only improve the risk by some constant factor – it will never grow at a slower rate than the risk of the noise addition.

Relationship to Previous Work

We note some similarities and advantages as compared with the technique of Hardt and Talwar [2010]. The two are quite similar in that both required a packing of exponential size, and with distance appropriately bounded, however Hardt and Talwar [2010] were forced to make an important relaxation to the problem. Namely their lower bound holds only for the differentially private mechanisms which admit a kind of smooth extension. Consider a mechanism for histograms to be characterized by distributions indexed by the integer lattice points of $\mathbb{R}^p$, then their method only applies to mechanisms for which one may "fill in" the extra space, i.e., add further distributions at the non-integer lattice points, subject to the condition that

$$P_X(A) \leq e^{\alpha \|X-Y\|_1} P_Y(A),$$

as noted in De [2011] it is unclear how much of a restriction this represents. What's more in the appendix of Hardt and Talwar [2010] they give a relaxation which removes this requirement but which only holds as α is decreasing appropriately as n increases.

2.4.1 Lower Bounds for Sparse Histograms

The above mimimax arguments held over the entire space of histograms, in some cases we may be more concerned with performance in particular subsets of this space instead. For example we

may expect that the true histogram based on the database is sparse (that is, most of the cells are unoccupied), we ask if it is possible to improve the risk on these types of databases, possibly at the expense of minimaxity over the whole space.

The case of differentially private sparse contingency tables was considered in Cormode *et al.* [2011]. Contingency tables may be regarded as histograms with some additional structure, but for this section that structure is immaterial. They proposed various schemes in order to preserve the sparsity of the released table, since noise addition for example results in a table in which every cell has a non-zero value almost surely. The reason for their interest in this problem is that sparse tables may be much smaller data structures and therefore be more practical to e.g., share with other researchers, as well as allowing more efficient inference procedures. Here we take the simplest of their procedures and analyze it from the perspective of risk, and demonstrate that for sparse histograms the risk may be brought much lower than the linear function of p we had above.

We consider the differentially private release technique which is to add Laplace noise to each cell of the histogram and then to truncate to zero any cell with a resulting value that is below some predetermined threshold τ, namely we release

$$\tilde{\theta}_i = \begin{cases} 0 & \theta_i(D) + L_i \leq \tau \\ \theta_i(D) + L_i & \text{o/w}, \end{cases} \tag{2.3}$$

where L_i are iid Laplace variables with rate parameter $\alpha/2$. That such a method achieves the differential privacy is seen through Proposition 1.3.1, combined with Proposition 1.2.5, since the method may be regarded as post processing the result of Laplace noise addition. We now show that for a particular choice of τ which only depends on the dimension p and the parameter α, the error may be brought sublinear in p. Namely the error depends only logarithmically on p, and linearly on the number of occupied cells, which we denote by q.

Proposition 2.4.3. *The mechanism defined by (2.3) for a p dimensional histogram with*

$$\tau = \tau(p, \alpha) = \frac{2}{\alpha} \log p, \tag{2.4}$$

has risk smaller than

$$\frac{2q+1}{\alpha} \left(\log p + 1 \right),$$

whenever at most q cells of the input histogram are non-zero.

Proof. Consider the error in each cell, suppressing dependence on D in the interest of space

$$\mathbb{E}|\theta_i - \tilde{\theta}_i| = \begin{cases} \mathbb{E}[L_i \big| L_i > \tau] P(L_i > \tau) & \theta_i = 0 \\ \mathbb{E}[|L_i| \big| L_i > \tau - \theta_i] P(L_i > \tau - \theta_i) + \theta_i P(L_i \leq \tau - \theta_i) & \theta_i > 0 \end{cases}$$

In words, for $\theta_i = 0$ error is only made when L_i is sufficiently large. For $\theta_i > 0$ the value will be truncated to zero – resulting in error θ_i – unless L_i is large enough, in which case the error is given by the conditional expectation of L_i. We upper bound the error in the latter case by $\tau + 2/\alpha$. The reason is that in the event the thresholded value was originally negative, then the thresholding has certainly decreased the error, by moving the output value towards whatever the input positive value was. Likewise whenever the thresholded value was positive, the thresholding has moved it by no more than τ. Finally taking τ as in (2.4) leads to

$$P(L_i > \tau) = \frac{1}{2} e^{-\tau\alpha/2} = \frac{1}{2} e^{-\log p} = \frac{1}{2p},$$

and memorylessness of the exponential distribution leads to

$$\mathbb{E}[L_i \big| L_i > \tau] = \tau + \frac{2}{\alpha},$$

and so, since q gives the number of occupied cells we have

$$\begin{aligned} \mathbb{E}\|\theta - \tilde{\theta}\|_1 &\leq \frac{p-q}{2p}(\tau + \frac{2}{\alpha}) + q(\tau + \frac{2}{\alpha}) \\ &\leq \frac{2q+1}{\alpha} \left(\log p + 1\right). \end{aligned}$$

□

Thus we find that although this technique is not minimax (since the risk is on the order $p \log p$ for a dense histogram), it attains a far lower risk when the number of occupied cells is much smaller than p. We performed a two small experiments which are shown in Figure 2.3 and Figure 2.4. The first uses the "Edwards" dataset (see e.g., Charest [2012]) which has about one third of the cells populated, but mostly with low counts, thus the improvement due to using the sparse method is small. In the second experiment we used the National Longterm Care Survey data[1]. This is a very large histogram in which only 5% of the cells are occupied, and one cell in particular contains the majority of the counts. In this case the thresholded sparse histogram massively improves the error rate when compared to the naive noise addition method. It is interesting that this procedure yields a smaller error than Laplace noise addition for the first histogram we experimented with,

[1] available at `http://lib.stat.cmu.edu`

even though for that particular p and q the upper bound shown above is much larger than the $2p/\alpha$ error rate we get from the Laplace noise addition. In both examples we used datasets with large numbers of zero cells. We do not envision these methods to be successful for histograms which do not have this property.

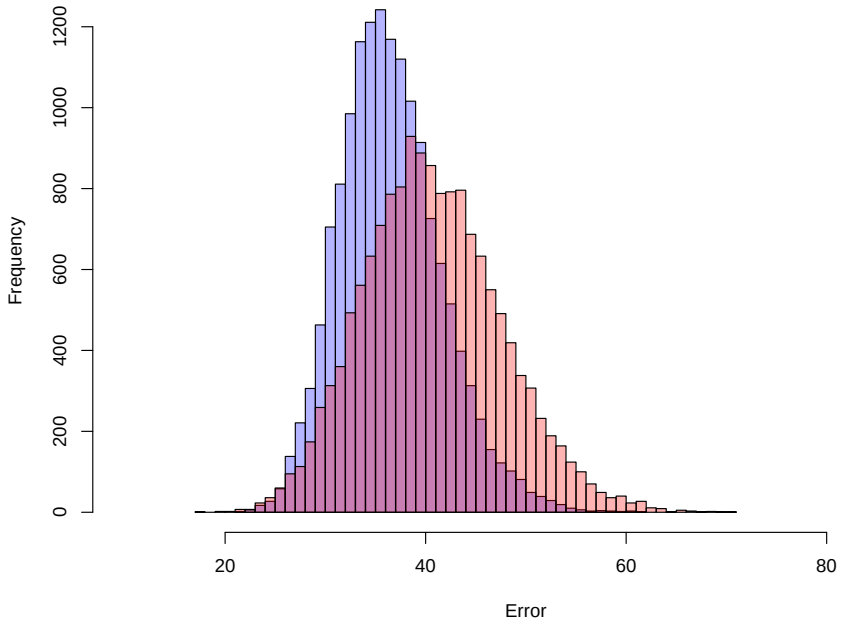


Figure 2.3: An experiment with the "Edwards" dataset Charest [2012]. This is a histogram having $p = 64$ cells of which $q = 22$ are occupied and $n = 70$. Displayed are the empirical errors over 15000 trials, for Laplace noise addition (in red) and the sparse histogram method (in blue). In both cases negative values resulting from noise addition were replaced with zeros.

We now construct a minimax lower bound for this problem. We thus restrict attention to the set of sparse histograms

$$\Theta_{p,q} = \left\{ \theta \in \mathbb{N}^p : \sum_{i=1}^{p} \theta_i = n, \sum_{i=1}^{p} \mathbf{1}\{\theta_i > 0\} \leq q \right\}.$$

Proposition 2.4.4. *For sufficiently large n and p, any α-differentially private method has*

$$\sup_{\theta \in \Theta_{p,q}} \mathbb{E}_{P_D} \left\| \theta(D) - \tilde{\theta} \right\|_1 \geq c \frac{q \log\left[(p-1)/q\right]}{\alpha},$$

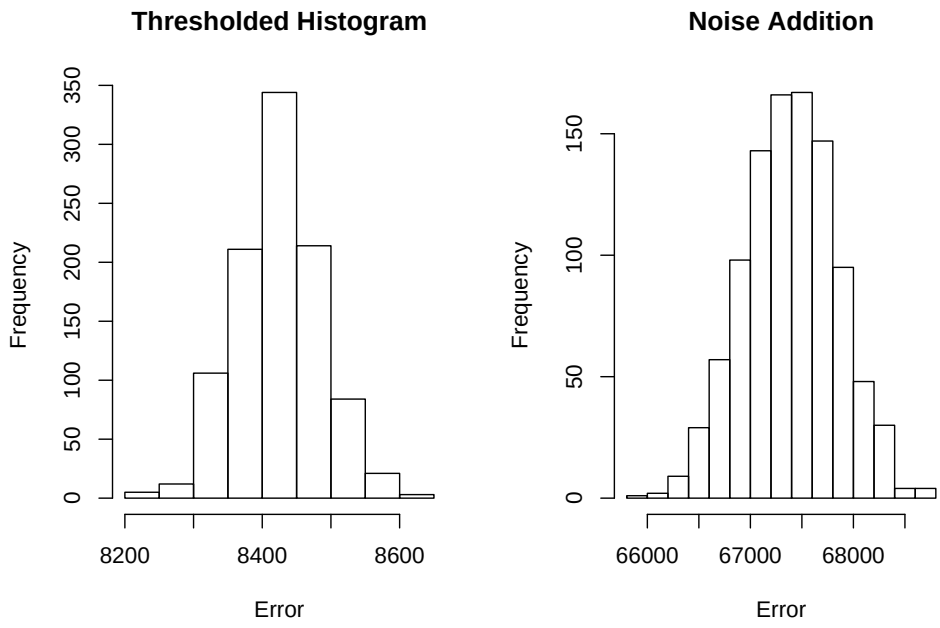


Figure 2.4: An experiment with the National Long Term Care Survey dataset . This is a histogram having $p = 2^{16}$ cells of which $q = 3152$ are occupied and $n = 21574$. Displayed are the empirical errors over 1000 trials, for the sparse method and for the non-sparse method. In both cases negative values resulting from noise addition were replaced with zeros.

for some universal constant $c > 0$.

Proof. As in the previous lower bound, consider a set of histograms

$$A = \left\{ (\eta\sigma_1, \ldots, \eta\sigma_{p-1}, n - \eta \sum_{i=1}^{p-1} \sigma_i), \sigma_i \in \{0,1\}, \|\sigma\|_1 = q \right\},$$

this set contains $(p-1)/q$ hypercubes of dimension q (by partitioning the coordinates into disjoint sets of size q and constructing the hypercube on each one). Then by using the hypercube packing of Lemma 2.4.1 and Fano's inequality we have

$$R \geq \frac{q\eta}{6} \left(1 - \frac{\alpha\eta q + \log 2}{c_0 q \log\left[(p-1)/q\right]} \right),$$

where c_0 is the constant from Lemma 2.4.1. Thus taking

$$\eta = \frac{c_0 \log\left[(p-1)/q\right]}{\alpha},$$

Leads to the desired result. □

Note the similarity between the above bound and the minimax risk bound for estimation of a sparse normal mean vector given in Johnstone [2011] (section 8.6).

Note that the bound given above is smaller than the risk of the thresholding method by an amount of the order $q \log q$. This may not be problematic since we suspect q to be small, but nevertheless illustrates an apparent gap between the lower bound and the error which was achievable using the basic technique. We note that in the case that q is somehow known ahead of time (which appears unlikely in practise) that the lower bound above may in fact be achieved. Namely if we take the thresholding estimator with

$$\tau = \frac{2}{\alpha} \log \frac{p}{q},$$

then the error is upper bounded by

$$\begin{aligned} \mathbb{E}\|\theta - \tilde{\theta}\|_1 &\leq \frac{q(p-q)}{2p}(\tau + \frac{2}{\alpha}) + q(\tau + \frac{2}{\alpha}) \\ &\leq \frac{3q}{\alpha}\left(\log \frac{p}{q} + 1\right). \end{aligned}$$

Note that while in principle this method achieves the correct error rate, in practise q is not known a-priori, and since τ is fixed prior to the data being observed, this method is not applicable except in rare occasions. Thus we prefer the original threshold since it does not depend on this unknown quantity and has an error rate within a small amount of the best available one.

2.4.2 Linear Functions of Histograms

A large body of the differential privacy literature concerns the release of a linear function of a histogram. Namely for some $m \times p$ matrix M we may be interested in releasing the quantity $M\theta(D)$, where θ produces some p-dimensional histogram as considered above. Evidently a simple method to apply to this problem is to construct a sanitized histogram $\tilde{\theta}$ which preserves the differential privacy, and then to releases $M\tilde{\theta}$ (or equivalently to release $\tilde{\theta}$ itself whereupon any linear function whatsoever may be computed by the recipients). As we may imagine however, this simple approach may introduce much more error into the requisite output vector than is strictly necessary. Here we consider a few kinds of matrices which are interesting to statisticians and determine the minimax risks available for the output of these linear functions.

Margins of Contingency Tables

First consider the case of one-way margins of a contingency table. A table corresponding to k binary attributes may be regarded as a histogram with 2^k cells. The one way margins are specified by a vector of length k, in which the i^{th} coordinate gives the number of elements in the table having value 1 for the i^{th} binary attribute. Thus in this case M is the $k \times 2^k$ matrix in which each column corresponds to a different binary vector.

Since $M\theta(D)$ is a k-dimensional vector of counts, an immediate way to perform a private release of this function is to use the minimax counting method from above for each coordinate. If we maintain α/k-differential privacy on each coordinate, then the composition property ensures that the overall release is α-differentially private. In this case in each element we have expected absolute value loss on the order of k/α, and since there are k coordinates and each is independent the overall expected ℓ_1 loss is on the order of k^2/α. We now show that this is the minimax rate, as above we first use the technical lemma about packing a hypercube, and then use Fano's inequality.

Proposition 2.4.5. *An α-differentially private method for the release of all the one way margins of a 2^k contingency table, has*

$$\sup_D \int \|M\theta_D - m\|_1 dP_D(m) \geq c_1 \frac{k^2}{\alpha}.$$

Proof. Let T be the packing set of lemma 2.4.1, the elements of T correspond to columns of M. For each element of T construct a histogram having τ elements in the cell corresponding to that column, and the remainder of the elements in the cell corresponding to all zeros. Then there are $e^{c_0 k}$ such histograms, and the differential privacy ensures that the KL divergence between distributions over the output space at any pair of these histograms is at most $\alpha\tau$. The ℓ_1 distance between the requisite linear functions of a pair of these histograms is evidently at least $\tau k/3$, thus Fano's inequality gives

$$R \geq \frac{k\tau}{8}\left(1 - \frac{\alpha\tau + \log 2}{c_0 k}\right),$$

and so taking

$$\tau = \frac{c_0 k}{2\alpha},$$

leads to the stated result. □

Thus for the one way margins it is only possible to improve the naive method by at most a constant factor. Note that using the packing we construct along with the method due to Hardt and Talwar [2010] yields the same result.

2.5 Summary

In this chapter we addressed the difficulty inherent to certain differentially private procedures, which we characterized by the minimax risk. Similar work was undertaken before in e.g., Hardt and Talwar [2010]; De [2011]. We restricted attention to the release of histograms and contingency tables, and were able to demonstrate a technique which achieves good accuracy in the case that the table is sparse. This is the typical situation when dealing with large scale tables. We also demonstrated lower bounds for errors for these problems and found our methods to be close to these lower bounds. Finally note that the minimax risk is not the only way to characterize the hardness of a private data release. We could also have considered the analog of the Bayes' risk from statistics. This would be similar to the minimax risk we use, but rather than concentrating on the worst case input, we would take a weighted average of the risks on different inputs, where the weights are specified by some distribution on the input space (the analog of the "prior" in Bayesian statistics).

Chapter 3

Differential Privacy in Spaces of Functions

When the goal is to release a function, we are presented with new challenges. For example, although the definition of differential privacy still makes sense, the techniques outlined in section 1.2 will not work since they operate over finite dimensional output spaces. What's more, since the function spaces are without a σ-finite dominating measure, even the techniques used to prove the differential privacy will not carry over. In this section we build a framework for differentially private release of functions, by considering limits of sequences of privacy preserving techniques on finite dimensional vectors.

We consider the family of functions over $T = \mathbb{R}^d$ (where appropriate we may restrict to a compact subset e.g., a unit cube of d-dimensions)

$$F = \{f_D : D \in \mathcal{D}\} \subset \mathbb{R}^T.$$

We will release some $\tilde{f}_D \sim P_D$ where P_D is a measure on $\mathbb{R}^T$ which depends on the input database. Following the techniques of Section 1.3, we may anticipate that a solution would be given by e.g., taking some distance function on the function space (for example, the uniform norm) and then sampling a distribution over $\mathbb{R}^T$ where the density at a point is given by the exponentiated negative distance to the input point. However there is a problem with this approach, which is that there is no analog of the Lebesgue measure in an infinite dimensional space (in fact there are no non-trivial sigma finite measures that are translation invariant). Thus we cannot construct mechanisms from densities as before, since the demonstration implicitly relied on the translation invariance of the underlying dominating measure. A workaround is to discretize the function space somehow and then define densities over the discretized version, but sampling sill still be problematic simply due

to the presumably fine level of discretization required for reasonable results.

In this chapter we propose two types of techniques for the release of functions, both may be regarded as the release of a finite dimensional quantity which somehow approximates the function. In the first case we expand the function in some basis, and add noise to the coefficients. Then the expansion may be truncated to a finite vector and released. In the second case we implicitly maintain a fully infinite dimensional representation of the function, add appropriate infinite dimensional noise, and then permit the user to repeatedly issue points at which to evaluate the function. This is a kind of on-line scenario but unlike those typically considered in the differential privacy literature, since in this case there will be no limit to the number of "queries" which can be answered, since the entire function is privacy preserving.

3.1 Finite Dimensional Techniques

We first consider a technique which represents the function f_D as a finite vector. This way, private release of the vector may be achieved via the techniques of Section 1.3, at which point a private version of the function may be reconstructed.

Thus we consider some mapping $\theta : \mathbb{R}^T \to \mathbb{R}^m$, and denote the image of F under this mapping as

$$\theta(F) = \{\theta(f_D) : f_D \in F\}.$$

The techniques of section 1.3 may evidently be used to release a vector $\tilde{\theta}$ with the requisite privacy guarantee, through the addition of a Gaussian vector in the case of approximate differential privacy, or the Gamma construction in the case of differential privacy.

Having released such a vector privately, we may consider a mapping back to the original function space

$$\xi : \mathbb{R}^m \to \mathbb{R}^T,$$

and reconstruct a function

$$\tilde{f}_D = \xi(\tilde{\theta}(f_D)).$$

3.1.1 Expansion in an Orthonormal Basis

A natural choice for $\theta(f)$ are the coordinates of f in some orthonormal basis of $\mathbb{R}^T$. Examples are the Fourier basis, the wavelet basis and the Hermite basis (see e.g., Wasserman [2006]). Denoting by $\{\psi_i\}_{i=1}^{\infty}$ a set of orthonormal basis functions with

$$\int_T \psi_i(x)\psi_j(x)\, d\lambda(x) = \mathbf{1}\{i = j\},$$

then we may take the coordinates

$$\theta_i(f) = \int_T f(x)\psi_i(x)\ d\lambda(x),$$

when $f \in \text{span}\{\psi_i\}$ we have

$$f(x) = \sum_{i=1}^{\infty} \theta_i(f)\psi_i(x),$$

and so we may define the function resulting from the truncation of θ to an m-dimensional vector as

$$\xi(\theta)(\cdot) = \sum_{i=1}^{m} \theta_i(f)\psi_i(x).$$

The remaining questions are which orthonormal basis to use, and how to choose m – the dimension of the finite vector θ. Clearly, the overall error in this approach decomposes into two parts, one due to the approximation error and one due to the noise addition

$$\mathbb{E}\int_T \left(f_D(x) - \tilde{f}_D(x)\right)^2\ d\lambda(x) = \int_T \left(f_D(x) - \xi[\theta(f_D)](x)\right)^2\ d\lambda(x) + \mathbb{E}\int_T \left(\xi[\theta(f_D)](x) - \tilde{f}_D(x)\right)^2\ d\lambda(x). \tag{3.1}$$

The choice of m allows a tradeoff between approximation error (which will tend to zero as m increases to infinity) and the privacy error (which will increase with m).

3.2 Differential Privacy in a Reproducing Kernel Hilbert Space

Whenever we have that $F \subseteq \mathcal{H}(K)$ where $\mathcal{H}(K)$ is an RKHS having reproducing kernel $K : T \times T \to \mathbb{R}$, we may make use of the additional structure in order to simplify both the correct determination of m as well as to bound the sensitivity of the finite dimensional vectors. First we recall some properties of RKHSs, then show how the particular kernel leads naturally to a particular choice of orthonormal basis into which to expand the functions (namely the eigenfunctions of the kernel itself).

3.2.1 Reproducing Kernel Hilbert Space Basics

We give some basic definitions for RKHSs below, and refer the reader to Bertinet and Agnan [2004] for a more detailed account. We first recall that the RKHS is generated from the closure of those functions which can be represented as finite linear combinations of the kernel:

$$\mathcal{H}_0(K) = \left\{ \sum_{i=1}^{n} \eta_i K_{x_i} \right\}$$

for some sequence $\eta_i \in \mathbb{R},\ x_i \in T$, and where $K_x = K(x, \cdot)$. For two functions

$$f = \sum_i \eta_i K_{x_i}, \quad g = \sum_j \mu_j K_{y_j},$$

the inner product is given by:

$$\langle f, g \rangle_{\mathcal{H}(K)} = \sum_{i \geq 1} \sum_{j \geq 1} \eta_i \mu_j K(x_i, y_j)$$

and the norm of f is

$$\|f\|^2_{\mathcal{H}(K)} = \langle f, f \rangle_{\mathcal{H}(K)}.$$

This gives rise to the "reproducing" nature of the Hilbert space, namely,

$$\langle K_x, K_y \rangle_{\mathcal{H}(K)} = K(x, y).$$

The functionals $\langle K_x, \cdot \rangle_{\mathcal{H}(K)}$ correspond to point evaluation:

$$\langle K_x, f \rangle_{\mathcal{H}(K)} = \sum_{i \geq 1} \theta_i K(x_i, x) = f(x).$$

The RKHS $\mathcal{H}(K)$ is then the union of $\mathcal{H}_0(K)$ with the limit points of the Cauchy sequences with respect to the norm induced by the inner product.

3.2.2 Privacy via the Spectral Decomposition in an RKHS

Recall that the reproducing kernel K may be written in terms of its eigenfunctions

$$K(x, y) = \sum_{i \geq 1} \lambda_i \psi_i(x) \psi_i(y),$$

in which the eigenfunctions and their respective eigenvalues are the solutions to the integral equation

$$\lambda_i \psi_i(a) = \int_T \psi_i(b) K(a, b)\ d\lambda(b),$$

and where we order the eigenvalues so that

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots$$

Likewise a function $f \in \mathcal{H}(K)$ may be uniquely written in terms of coordinates in the same basis

$$f = \sum_{i \geq 1} \theta_i(f) \psi_i(\cdot)$$

where $\theta_i(f) = \lambda_i \langle f, \psi_i \rangle_{\mathcal{H}(K)}$, and the inner product itself is represented as

$$\langle f, g \rangle_{\mathcal{H}(K)} = \sum_{i \geq 1} \frac{\theta_i(f)\theta_i(g)}{\lambda_i}.$$

Evidently the functions $\sqrt{\lambda_i}\psi_i(\cdot)$ form an orthonormal basis of $\mathcal{H}(K)$. What's more, by defining λ_i appropriately, the ψ_i form an orthonormal basis for $\mathcal{L}_2(T)$ (cf. Mercer's theorem).

We consider the release of the finite vector defined by the truncation to the first m terms of the spectral representation of the function. Thus we use the notation

$$\theta^m = (\theta_1, \ldots, \theta_m),$$

and construct the functions $\xi[\theta^m(f)]$ as above. Examples of these functions are given in Figure 3.1.

A similar approach was considered in Wasserman and Zhou [2010] in which the function to release was itself the truncation of the function defined by the empirical distribution of some sample data. It was also considered by Rubinstein *et al.* [2010] and Chaudhuri *et al.* [2011] in which a kernel SVM was truncated in the spectral expansion in order to obtain privacy. The difference here is that we assume explicitly that the functions are in the RKHS, which makes assessment of the sensitivity easier.

If we define the "RKHS sensitivity" of the family F of functions as

$$\Delta_K = \sup_{D \sim D'} \|f_D - f_{D'}\|_{\mathcal{H}(K)}, \tag{3.2}$$

then this leads to a valid upper bound for the "Mahalanobis sensitivity" of the finite vectors, where the covariance matrix is given by the diagonal matrix of eigenvalues λ_i corresponding to the first m eigenfunctions of the kernel.

Proposition 3.2.1. *Let $\Sigma \in \mathbb{R}^{m \times m}$ be diagonal with $\Sigma_{i,i} = \lambda_i$, then*

$$\sup_{D \sim D'} \|\theta^m(f_D) - \theta^m(f_{D'})\|_\Sigma \leq \Delta_K, \tag{3.3}$$

for any m.

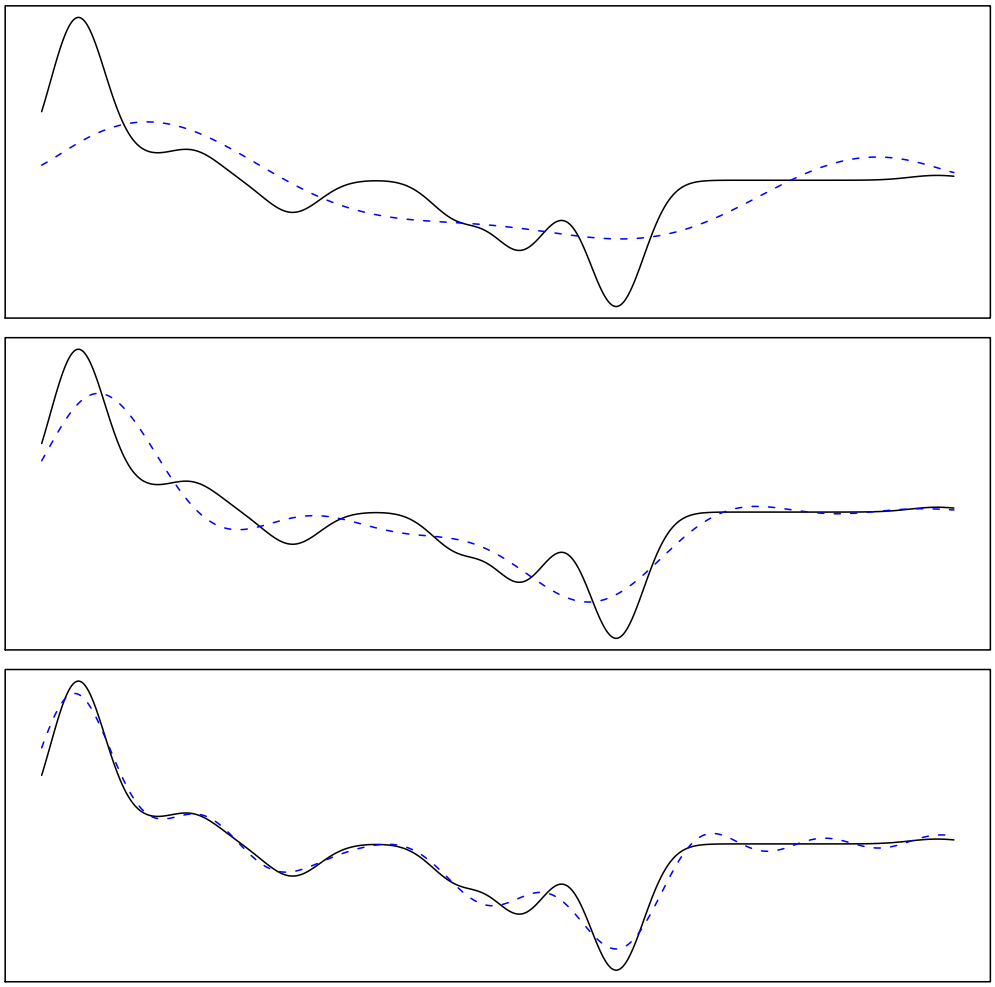


Figure 3.1: An example of the error due to finite truncation of a function in an RKHS. Here $\mathcal{H}(K)$ is the RKHS over $[0,1]$ with the Gaussian kernel. The black (solid line) function is being approximated by 5, 10, and 15 basis functions in the three plots, and the approximate function is shown as a dashed blue line. Evidently the approximation improves as the dimension increases.

Proof. For each $D \sim D'$ we have

$$\begin{aligned}
\|f_D - f_{D'}\|^2_{\mathcal{H}(K)} &= \sum_{i=1}^{\infty} \frac{(\theta_i(f_D) - \theta_i(f_{D'}))^2}{\lambda_i} \\
&\geq \sum_{i=1}^{m} \frac{(\theta_i(f_D) - \theta_i(f_{D'}))^2}{\lambda_i} \\
&= \|\theta^m(f_D) - \theta^m(f_{D'})\|^2_{\Sigma}.
\end{aligned}$$

□

3.2.3 (α, β)-Differential Privacy in an RKHS

In light of Propositions 1.3.3 and 3.2.1 we find that for the (α, β)-differentially private release of f_D it suffices to take

$$\tilde{f}_D^{m,(\alpha,\beta)}(\cdot) = \sum_{i=1}^{m} \left(\theta_i(f_D) + c(\alpha, \beta, \Delta_K)\sqrt{\lambda_i} Z_i\right) \psi_i(\cdot), \quad Z_i \sim \mathcal{N}(0,1), \tag{3.4}$$

where we define

$$c(\alpha, \beta, \Delta) = \sqrt{2 \log \frac{2}{\beta}} \frac{\Delta}{\alpha}. \tag{3.5}$$

We find

$$\mathbb{E} \int_T \left(f_D(x) - \tilde{f}_D^{m,(\alpha,\beta)}(x)\right)^2 \, d\lambda(x) = \sum_{i=m+1}^{\infty} \theta_i^2(f_D) + c^2(\alpha, \beta, \Delta_K) \sum_{i=1}^{m} \lambda_i,$$

note that now as m is increased, the first term on the right side vanishes, whereas the second term approaches

$$\sum_{i \geq 1} \lambda_i = \sum_{i \geq 1} \lambda_i \int_T \phi_i(x)\phi_i(x) \, d\lambda(x) = \int_T \sum_{i \geq 1} \lambda_i \psi_i(x)\psi_i(x) \, d\lambda(x) = \int_T K(x,x) \, d\lambda(x),$$

where the middle equality is due to the monotone convergence theorem. For kernels such as the Gaussian and exponential kernels, $K(x,x) = 1$ and so this quantity is simply $\lambda(T)$. Thus we may consider in a sense the limit of this technique as m tends to infinity, since the error (as measured in the integrated square error) will remain bounded. We note that the resulting function is

$$\tilde{f}_D^{\infty,(\alpha,\beta)}(\cdot) = f_D(\cdot) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} Z_i \psi_i(\cdot),$$

and recognize the second term of the sum as the Karhunen-Loève expansion of a Gaussian process having mean zero and covariance function given by the reproducing kernel K (see e.g., Bertinet and Agnan [2004]).

A Gaussian process indexed by T (see e.g., Adler [1990]; Adler and Taylor [2007]) is a collection of random variables $\{X_t : t \in T\}$, for which each finite subset is distributed as a multivariate Gaussian. A sample from a Gaussian process may be considered as a function $T \to \mathbb{R}$, by examining the so-called "sample path" $t \to X_t$. The Gaussian process is determined by the mean and covariance

functions, defined on T and T^2 respectively, as

$$m(t) = \mathbb{E}X_t, \quad K(s,t) = \mathrm{Cov}(X_s, X_t).$$

For any finite subset $S \subset T$, the random vector $\{X_t : t \in S\}$ has a normal distribution with the means, variances, and covariances given by the above functions.

Thus we may suspect that the addition of a Gaussian process to the function f_D will preserve privacy. Below we make this precise with a straightforward limiting argument.

Proposition 3.2.2. *For a family of functions $\{f_D : D \in \mathcal{D}\} \subseteq \mathcal{H}(K)$ which satisfies*

$$\sup_{X \sim X'} \|f_X - f_{X'}\|_{\mathcal{H}(K)} \leq \Delta_K, \tag{3.6}$$

then it satisfies (α, β)-DP to take each P_D to be the Gaussian process measure having mean function f_D and covariance function $c(\alpha, \beta, \Delta_K)$, with c defined as in (3.5).

Proof of proposition 3.2.2. Denote an arbitrary measurable set of the infinite sequence of coefficients $\theta_1(f_D), \theta_2(f_D), \ldots$ by $A = \bigcap_{i \geq 1} A_i$ where A_i is a set of infinite sequences in $\mathbb{R}$ in which the i^{th} is restricted to lie in some measurable set. Since A is the limit of a decreasing sequence of sets $B_m = \bigcap_{i=1}^m A_i$, we have $P_X(A) = P_X(\lim_{m\to\infty} B_m) = \lim_{m\to\infty} P_X(B_m)$, and since differential privacy holds for each finite m, a simple limiting argument leads to the requisite privacy of the Gaussian process. □

We remark that under the restriction that T be compact, the error incurred due to privacy, when measured in mean $\mathcal{L}_2$ error, is given by the expectation of the square norm of the Gaussian process which is $\lambda(T)c(\alpha, \beta, \Delta) \sup_{x \in T} K(x, x)$, where we use λ to mean the Lebesgue measure (namely $\lambda(T) = 1$ for $T = [0, 1]^d$).

3.2.4 Alternate View of Gaussian Process Noise

The preceding section gave a demonstration of the differential privacy obtained via Gaussian process noise addition which was based on the spectral decomposition of the RKHS and correspondingly of the Gaussian process itself. In this section we remark that there is a second way to arrive at the same conclusion, by considering the release of finitely many function values (e.g., corresponding to evaluation on a grid of points) and allowing the size of the grid to grow to infinity.

Suppose the goal is to output function values at the points specified in some finite set $\{x_1, \ldots, x_m\}$, then we may consider the output of the vector

$$(f_D(x_1), \ldots, f_D(x_m)).$$

Evidently if the sensitivity of the resulting family of vectors could be bounded then the noise addition techniques of Section 1.3 could be used to release a private version of these vectors. It turns out that as above when the functions f_D lay in some RKHS then to upper bound the sensitivity is made simple. If we measure the sensitivity in a Euclidean norm in which we use the "Gram matrix" for the reproducing kernel and the set of points x_i then we have the following.

Proposition 3.2.3. *For $f \in \mathcal{H}(K)$, $m < \infty$ and $x_1, \ldots, x_m$ distinct points in T, we have*

$$\left\| \begin{pmatrix} K(x_1,x_1) & \cdots & K(x_1,x_m) \\ \vdots & \ddots & \vdots \\ K(x_m,x_1) & \cdots & K(x_m,x_m) \end{pmatrix}^{-1/2} \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_m) \end{pmatrix} \right\|_2^2 \leq \|f\|_{\mathcal{H}}^2.$$

To summarize the proof, the quantity in the statement of the theorem is just the square RKHS norm of the orthogonal projection of f to the subspace of $\mathcal{H}(K)$ spanned by the functions $K_{x_i} = K(x_i, \cdot)$.

Proof. Note that invertibility of the matrix is safely assumed due to Mercer's theorem. Denote by M the matrix having elements $M_{ij} = K(x_i, x_j)$. Denote by P the operator $\mathcal{H}(K) \to \mathcal{H}(K)$ defined by

$$P = \sum_{i=1}^{n} K_{x_i} \sum_{j=1}^{n} (M^{-1})_{i,j} \left\langle K_{x_j}, \cdot \right\rangle_{\mathcal{H}(K)}$$

We find this operator to be idempotent in the sense that $P = P^2$:

$$\begin{aligned} P^2 &= \sum_{i=1}^{n} K_{x_i} \sum_{j=1}^{n} (M^{-1})_{i,j} \left\langle K_{x_j}, \sum_{k=1}^{n} K_{x_k} \sum_{l=1}^{n} (M^{-1})_{k,l} \left\langle K_{x_l}, \cdot \right\rangle_{\mathcal{H}(K)} \right\rangle_{\mathcal{H}(K)} \\ &= \sum_{i=1}^{n} K_{x_i} \sum_{j=1}^{n} (M^{-1})_{i,j} \sum_{k=1}^{n} \left\langle K_{x_j}, K_{x_k} \right\rangle_{\mathcal{H}(K)} \sum_{l=1}^{n} (M^{-1})_{k,l} \left\langle K_{x_l}, \cdot \right\rangle_{\mathcal{H}(K)} \\ &= \sum_{i=1}^{n} K_{x_i} \sum_{j=1}^{n} (M^{-1})_{i,j} \sum_{k=1}^{n} M_{j,k} \sum_{l=1}^{n} (M^{-1})_{k,l} \left\langle K_{x_l}, \cdot \right\rangle_{\mathcal{H}(K)} \\ &= \sum_{i=1}^{n} K_{x_i} \sum_{l=1}^{n} (M^{-1})_{i,l} \left\langle K_{x_l}, \cdot \right\rangle_{\mathcal{H}(K)} \\ &= P. \end{aligned}$$

P is also self-adjoint due to the symmetry of M, i.e.

$$\begin{aligned}
\langle Pf, g\rangle_{\mathcal{H}(K)} &= \left\langle \sum_{i=1}^{n} K_{x_i} \sum_{j=1}^{n} (M^{-1})_{i,j} \left\langle K_{x_j}, f\right\rangle_{\mathcal{H}(K)}, g \right\rangle_{\mathcal{H}(K)} \\
&= \left\langle \sum_{i=1}^{n} \langle K_{x_i}, g\rangle_{\mathcal{H}(K)} \sum_{j=1}^{n} (M^{-1})_{i,j} K_{x_j}, f \right\rangle_{\mathcal{H}(K)} \\
&= \left\langle \sum_{j=1}^{n} K_{x_j} \sum_{i=1}^{n} (M^{-1})_{i,j} \langle K_{x_i}, g\rangle_{\mathcal{H}(K)}, f \right\rangle_{\mathcal{H}(K)} \\
&= \left\langle \sum_{j=1}^{n} K_{x_j} \sum_{i=1}^{n} (M^{-1})_{j,i} \langle K_{x_i}, g\rangle_{\mathcal{H}(K)}, f \right\rangle_{\mathcal{H}(K)} \\
&= \langle Pg, f\rangle_{\mathcal{H}(K)} .
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|f\|_{\mathcal{H}}^2 &= \langle f, f\rangle_{\mathcal{H}(K)} \\
&= \langle Pf + (f - Pf), Pf + (f - Pf)\rangle_{\mathcal{H}(K)} \\
&= \langle Pf, Pf\rangle_{\mathcal{H}(K)} + 2\langle Pf, f - Pf\rangle_{\mathcal{H}(K)} + \langle f - Pf, f - Pf\rangle_{\mathcal{H}(K)} \\
&= \langle Pf, Pf\rangle_{\mathcal{H}(K)} + 2\left\langle f, Pf - P^2 f\right\rangle_{\mathcal{H}(K)} + \langle f - Pf, f - Pf\rangle_{\mathcal{H}(K)} \\
&= \langle Pf, Pf\rangle_{\mathcal{H}(K)} + \langle f - Pf, f - Pf\rangle_{\mathcal{H}(K)} \\
&\geq \langle Pf, Pf\rangle_{\mathcal{H}(K)} \\
&= \langle f, Pf\rangle_{\mathcal{H}(K)} .
\end{aligned}$$

The latter term is nothing more than the left hand side in the statement. □

The proposition above demonstrates that the sensitivity of the vector of function values is less than the RKHS sensitivity of the family of functions. Therefore for any finite collection of points the approximate differential privacy may be achieved by the addition of an appropriately scaled Gaussian vector, where the covariance matrix is given by the Gram matrix of the reproducing kernel.

As above, this method exhibits a strong connection to Gaussian processes. Namely, the distribution of the noise added to the vector is exactly a finite dimensional distribution of the Gaussian process having mean zero and the covariance function given by the reproducing kernel.

3.2.5 Algorithms

There are two main modes in which functions f_D could be released by the holder of the data D to the outside parties. The first we describe as a "batch" setting in which the parties would designate some finite collection of points $x_1 \ldots, x_n \in T$. The database owner would compute $\tilde{f}_D(x_i)$ for each i and return the vector of results. At this point the entire transaction would end with only the collection of pairs $(x_i, \tilde{f}_D(x_i))$ being known to the outsiders. An alternative is the "online" setting in which outside users repeatedly specify points in $x_i \in T$, the database owner would reply with $\tilde{f}_D(x_i)$, but unlike the former setting he would remain available to respond to more requests for function evaluations. We name these settings "batch" and "online" for their resemblance of the batch and online settings typically considered in machine learning algorithms.

The batch method is nothing more than sampling a multivariate Gaussian, since the set $x_1, \ldots, x_n \in T$ specifies the finite dimensional distribution of the Gaussian process from which to sample. The released vector is simply

$$\begin{pmatrix} \tilde{f}_D(x_1) \\ \vdots \\ \tilde{f}_D(x_n) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} f_D(x_1) \\ \vdots \\ f_D(x_n) \end{pmatrix}, \frac{c(\beta)\Delta}{\alpha} \begin{pmatrix} K(x_1,x_1) & \cdots & K(x_1,x_n) \\ \vdots & \ddots & \vdots \\ K(x_n,x_1) & \cdots & K(x_n,x_n) \end{pmatrix} \right).$$

In the online setting, the data owner upon receiving a request for evaluation at x_i would sample the gaussian process conditioned on the samples already produced at $x_1, \ldots, x_{i-1}$. Let

$$C_i = \begin{pmatrix} K(x_1,x_1) & \cdots & K(x_1,x_{i-1}) \\ \vdots & \ddots & \vdots \\ K(x_{i-1},x_1) & \cdots & K(x_{i-1},x_{i-1}) \end{pmatrix}, \quad G_i = \begin{pmatrix} \tilde{f}_D(x_1) \\ \vdots \\ \tilde{f}_D(x_{i-1}) \end{pmatrix}, \quad V_i = \begin{pmatrix} K(x_1,x_i) \\ \vdots \\ K(x_{i-1},x_i) \end{pmatrix}.$$

Then,

$$\tilde{f}_D(x_i) \sim \mathcal{N}\left(V_i^T C_i^{-1} G_i,\ K(x_i,x_i) - V_i^T C_i^{-1} V_i\right).$$

The database owner may track the inverse matrix C_i^{-1} and after each request update it into C_{i+1}^{-1} by making use of Schur's complements combined with the matrix inversion lemma. Nevertheless we note that as i increases the computational complexity of answering the request will in general grow. In the very least, the construction of V_i takes time proportional to i. This may make this approach problematic to implement in practise.

3.3 Examples

We now give some examples in which the above technique may be used to construct private versions of functions in an RKHS.

3.3.1 Kernel Density Estimation

Let f_D be the kernel density estimator, where D is regarded as a sequence of points $x_i \in T$ as $i = 1, \ldots, n$ drawn from a distribution with density f. Let h denote the bandwidth. Assuming a Gaussian kernel, the estimator is

$$f_D(x) = \frac{1}{n(2\pi h^2)^{d/2}} \sum_{i=1}^{n} \exp\left\{\frac{-\|x - x_i\|_2^2}{2h^2}\right\}, \quad x \in T.$$

These functions reside in the RKHS with kernel

$$K(x, y) = \exp\left\{\frac{-\|x - y\|_2^2}{2h^2}\right\},$$

thus the RKHS sensitivity is

$$\begin{aligned}\|f_D - f_{D'}\|_{\mathcal{H}}^2 &= \left(\frac{1}{n(2\pi h^2)^{d/2}}\right)^2 \left(K(x_n, x_n) + K(x_n', x_n') - 2K(x_n, x_n')\right) \\ &\leq 2\left(\frac{1}{n(2\pi h^2)^{d/2}}\right)^2.\end{aligned}$$

We find the noise level to be on the same order as the noise due to the sampling of the points, therefore the privacy does not disrupt the rate of convergence of the estimator. In figure 3.2 we show the curve resulting from the application of the above method for the one-dimensional kernel density estimation.

3.3.2 A Sobolev Space of Functions

Consider the Sobolev space

$$H^1[0, 1] = \left\{f \in C[0, 1] : \int_0^1 (\partial f(x))^2 \, d\lambda(x) < \infty\right\}.$$

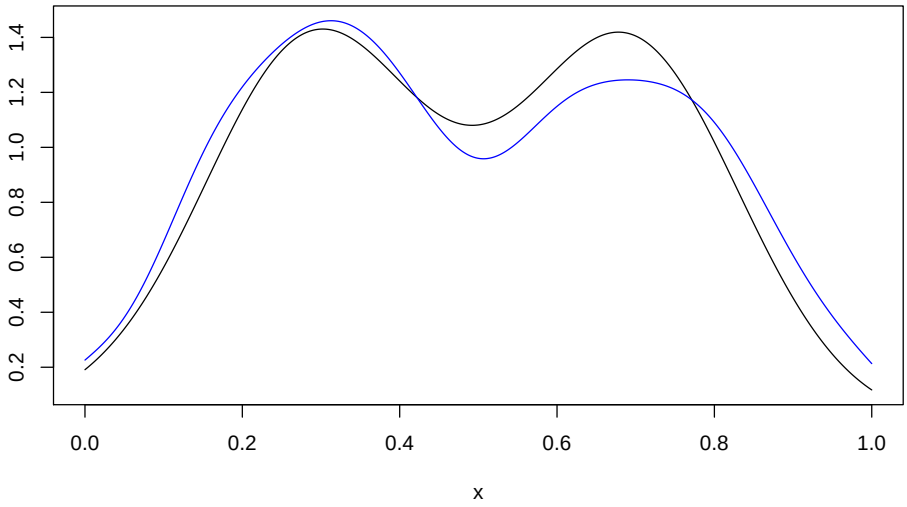


Figure 3.2: An example of a kernel density estimator (the black curve) and the released version (the blue curve). This uses the method developed in Section 3.3.1. Here we sampled $n = 100$ points from a mixture of two normals centered at 0.3 and 0.7 respectively. We use $h = 0.1$ and have $\alpha = 1$ and $\beta = 0.1$. The Gaussian Process is evaluated on an evenly spaced grid of 1000 points between 0 and 1. Note that gross features of the original kernel density estimator remain, namely the two peaks.

This is a RKHS with the kernel $K(x, y) = \exp\{-\gamma\ |x - y|\}$ for some parameter $\gamma > 0$. The norm in this space is given by

$$\|f\|_{\mathcal{H}}^2 = \frac{1}{2}\left(f(0)^2 + f(1))^2\right) + \frac{1}{2\gamma}\int_0^1 (\partial f(x))^2 + \gamma^2 f(t)^2\ d\lambda(t). \tag{3.7}$$

See e.g., Bertinet and Agnan [2004]; Parzen [1961]. Thus for a family of functions in one dimension which lay in the Sobolev space H^1, we may determine a noise level necessary to achieve the differential privacy by bounding the above quantity for the difference of two functions.

For functions over higher dimensional domains (as $[0, 1]^d$ for some $d > 1$) we may construct an RKHS by taking the d-fold tensor product of the above RKHS (see, in particular, Parzen [1963]; Aronszajn [1950] for details of the construction). The resulting space has the reproducing kernel

$$K(x, y) = \exp\{-\gamma\|x - y\|_1\},$$

and is the completion of the set of functions

$$\mathcal{G}_0 = \left\{ f : [0,1]^d \to \mathbb{R} : f(x_1, \ldots, x_d) = f_1(x_1) \cdots f_d(x_d), f_i \in H^1[0,1] \right\}.$$

The norm over this set of functions is given by:

$$\|f\|^2_{\mathcal{G}_0} = \prod_{j=1}^{d} \|f_i\|^2_{\mathcal{H}}. \tag{3.8}$$

The norm over the completed space agrees with the above on $\mathcal{G}_0$. The explicit form is obtained by substituting (3.7) into the right hand side of (3.8) and replacing all instances of $\prod_{j=1}^{d} f_j(x_j)$ with $f(x_1, \ldots, x_j)$.

Thus the norm in the completed space is defined for all f possessing all first partial derivatives which are all in $\mathcal{L}_2$. In figure 3.3 we repeat the example of figure 3.2, but using the above method instead. The kernel density estimator is an element of the Sobolev space, and therefore we once again may bound the sensitivity and add the appropriate Gaussian process. Note that the sample paths for the corresponding Gaussian process are no longer smooth.

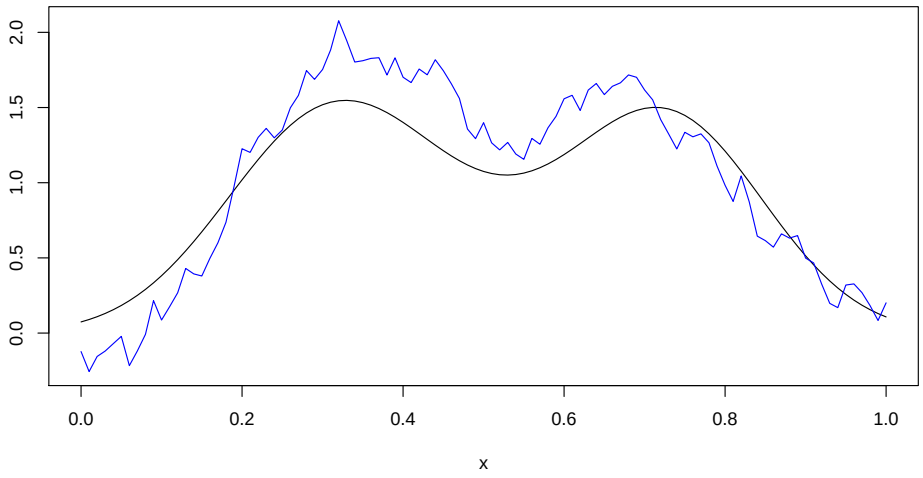


Figure 3.3: An example of a kernel density estimator (the black curve) and the released version (the blue curve). The setup is the same as in Figure 3.2, but the privacy mechanism developed in Section 3.3.2 was used instead. Note that the released function does not have the desirable smoothness of released function from Figure 3.2.

3.3.3 Minimizers of Regularized Functionals in an RKHS

The construction of the following section is due to Bousquet and Elisseeff [2002], who were interested in determining the sensitivity of certain kernel machines (among other algorithms) for the sake of bounding the generalization error of the output classifiers. Rubinstein (in e.g., Rubinstein *et al.* [2010]) noted that these bounds are useful for establishing the noise level required for differential privacy of support vector machines. They are also useful for our approach to privacy in a function space.

We consider classification and regression schemes in which the datasets $D = \{z_1, \ldots, z_n\}$ with $z_i = (x_i, y_i)$, where $x_i \in [0,1]^d$ are some covariates, and y_i is some kind of label, either taking values on $\{-1, +1\}$ in the case of classification or some taking values in some interval when the goal is regression. Thus the output functions are from $[0,1]^d$ to a subset of $\mathbb{R}$. The functions we are interested in take the form:

$$f_D = \arg\min_{g \in \mathcal{H}} \frac{1}{n} \sum_{z_i \in D} \ell(g, z_i) + \lambda \|g\|_{\mathcal{H}}^2 \tag{3.9}$$

where $\mathcal{H}$ is some RKHS to be determined, and ℓ is the so-called "loss function." We now repeat a definition from Bousquet and Elisseeff [2002] (using M in place of their σ to prevent confusion):

Definition 3.3.1 (M-admissible loss function Bousquet and Elisseeff [2002])**.** A loss function: $\ell(g, z) = c(g(x), y)$ is called M-admissible whenever c it is convex in its first argument and Lipschitz with constant M in its first argument.

We will now demonstrate that for (3.9), whenever the loss function is admissible, the minimizers on adjacent datasets may be bounded close together in RKHS norm. Denote the part of the optimization due to the loss function:

$$L_D(f) = \frac{1}{n} \sum_{z_i \in D} \ell(f, z_i).$$

Using the technique from the proof of lemma 20 of Bousquet and Elisseeff [2002] we find that since ℓ is convex in its first argument we have:

$$L_D(f_D + \eta(f_{D'} - f_D)) - L_D(f_D) \leq \eta(L_D(f_{D'}) - L_D(f_D)).$$

This also holds when f_D and $f_{D'}$ swap places. Summing the resulting inequality with the above and rearranging yields:

$$L_D(f_{D'} - \eta(f_{D'} - f_D)) - L_D(f_{D'}) \leq L_D(f_D) - L_D(f_D + \eta(f_{D'} - f_D)).$$

Due to the definition of $f_D, f_{D'}$ as the minimizers of their respective functionals we have:

$$L_D(f_D) + \lambda\|f_D\|_{\mathcal{H}}^2 \leq L_D(f_D + \eta(f_{D'} - f_D)) + \lambda\|f_D + \eta(f_{D'} - f_D)\|_{\mathcal{H}}^2$$
$$L_{D'}(f_{D'}) + \lambda\|f_{D'}\|_{\mathcal{H}}^2 \leq L_{D'}(f_{D'} - \eta(f_{D'} - f_D)) + \lambda\|f_{D'} - \eta(f_{D'} - f_D)\|_{\mathcal{H}}^2.$$

This leads to:

$$\begin{aligned}
0 &\geq \lambda\left(\|f_D\|_{\mathcal{H}}^2 - \|f_D + \eta(f_{D'} - f_D)\|_{\mathcal{H}}^2 + \|f_{D'}\|_{\mathcal{H}}^2 - \|f_{D'} - \eta(f_{D'} - f_D)\|_{\mathcal{H}}^2\right) \\
&\quad + L_D(f_D) - L_D(f_D + \eta(f_{D'} - f_D)) + L_{D'}(f_{D'}) - L_{D'}(f_{D'} - \eta(f_{D'} - f_D)) \\
&\geq 2\lambda\|\eta(f_{D'} - f_D)\|_{\mathcal{H}}^2 - L_D(f_{D'}) + L_D(f_{D'} - \eta(f_{D'} - f_D)) + L_{D'}(f_{D'}) - L_{D'}(f_{D'} - \eta(f_{D'} - f_D)) \\
&= 2\lambda\|\eta(f_{D'} - f_D)\|_{\mathcal{H}}^2 + \frac{1}{n}\left(\ell(z, f_{D'}) - \ell(z, f_{D'} - \eta(f_{D'} - f_D)) + \ell(z', f_{D'}) - \ell(z', f_{D'} - \eta(f_{D'} - f_D))\right)
\end{aligned}$$

Moving the loss function term to the other side and using the Lipschitz property leads to:

$$\|f_D - f_{D'}\|_{\mathcal{H}}^2 \leq \frac{M}{\lambda n}\|f_D - f_{D'}\|_\infty$$

What's more, the reproducing property together with Cauchy-Schwarz inequality yields:

$$|f_D(x) - f_{D'}(x)| = |\langle f_D - f_{D'}, K_x\rangle_{\mathcal{H}(K)}| \leq \|f_D - f_{D'}\|_{\mathcal{H}}\sqrt{K(x,x)}.$$

Combining with the previous result gives:

$$\|f_D - f_{D'}\|_{\mathcal{H}}^2 \leq \frac{M}{\lambda n}\|f_D - f_{D'}\|_{\mathcal{H}}\sqrt{\sup_x K(x,x)}$$

which gives

$$\|f_D - f_{D'}\|_{\mathcal{H}} \leq \frac{M}{\lambda n}\sqrt{\sup_x K(x,x)}.$$

For a soft-margin kernel SVM we have the loss function: $\ell(g, z) = (1 - yg(x))_+$, which means the positive part of the term in parentheses. Since the label y takes on either plus or minus one, we find this to be 1-admissible.

In figure 3.4 we give an example of our method to the release of a kernel support vector machine. The example is synthetic and operates on a two dimensional input space. We find that in areas of the input space with many data points (and hence where the SVM is fairly confident in its predictions) that the noise addition has not lead to catastrophically decreased accuracy. Where the input points are sparser, the SVM is not confident about the predictions (i.e., the regression function has a small absolute value) and so the noise addition leads to different predictions in these areas. This will presumably lead to the excess error due to privacy being well-behaved, but we leave

this for future work.

3.4 Summary

We have shown how to add random noise to a function in such a way that differential privacy is preserved. Our method differs from previous techniques, since it does not resort to the release of a finite dimensional projection of the function. It is interesting to note that our method only achieves the weaker notion of approximate differential privacy. Whether an analogous construction exists for the unadulterated differential privacy is a matter for future work, although we suspect it does not.

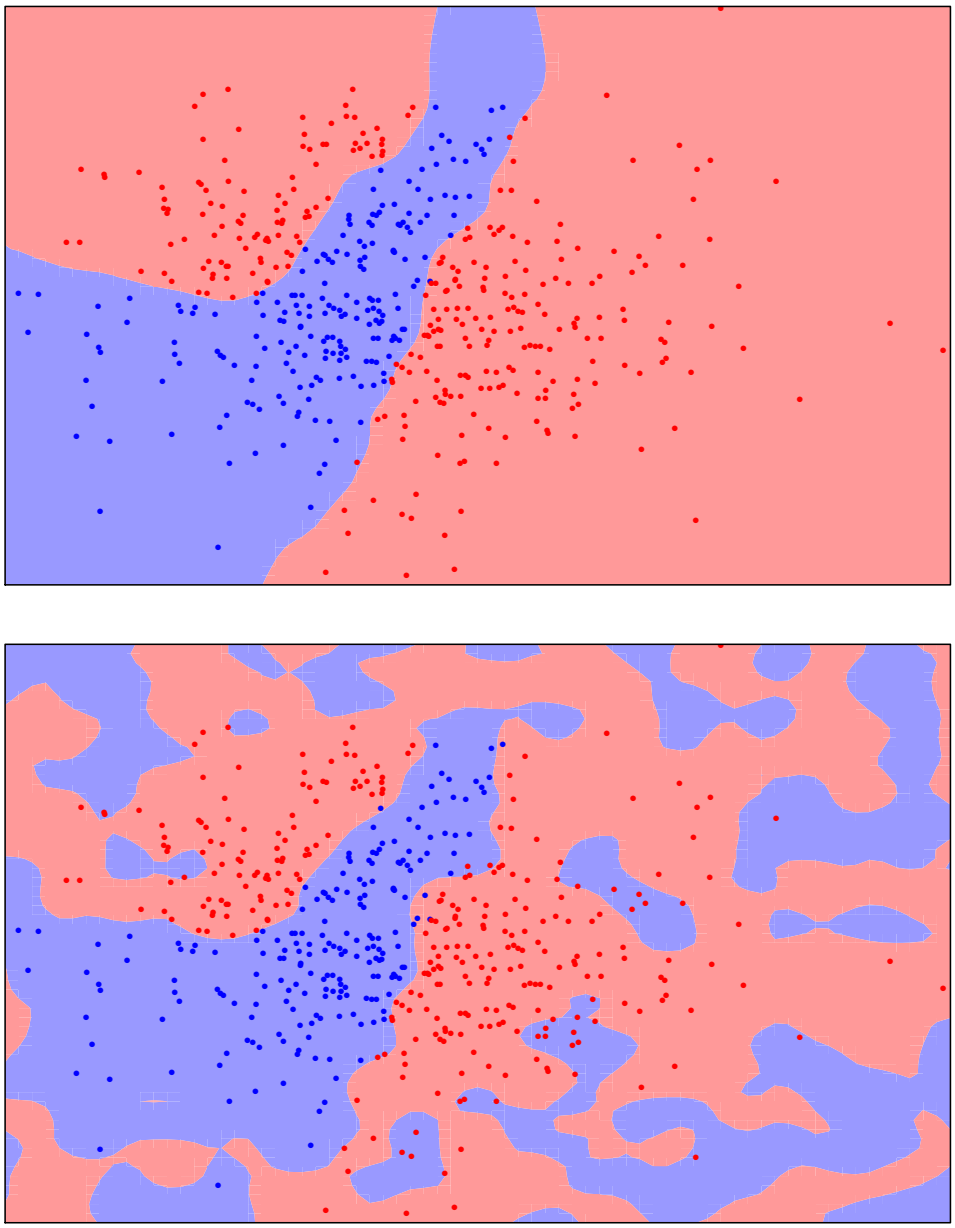


Figure 3.4: An example of a kernel support vector machine. In the top image are the data points, with the colors representing the two class labels. The background color corresponds to the class predicted by the learned kernel SVM. In the bottom image are the same data points, with the predictions of the private kernel SVM. This example uses the Gaussian kernel for classification.

Chapter 4

Kernel Density Estimation

4.1 Introduction

In this chapter we seek to construct a method for the α-differentially private release of a kernel density estimator which maintains the minimax rate of convergence enjoyed by the non-private estimator (see e.g., Wasserman [2006]). We propose a technique which is similar in spirit to the approach of Wasserman and Zhou [2010] in which the function is made private by perturbation of its coordinates in some orthonormal basis. Note that the main difference between this section and the private density estimator given above is that we require the strict α-differential privacy, rather than the weaker (α, β)-DP which was achieved there.

There are myriad techniques which could plausibly be brought to bear to address this problem, but we seek to preserve the rate of convergence enjoyed by the non-private kernel density estimator. This is that rate at which the estimate converges to the true density (in mean integrated square error) as the number of samples is increased. This restriction rules out some conceptually simple methods such as e.g., moving the points into some grid and treating the problem as essentially a histogram.

The approach we use is to truncate the fourier expansion of the kernel density estimate. We determine the number of coefficients that are required in order for this truncation to preserve the correct convergence rate. We then demonstrate that simply adding the Laplace noise to each coefficient in the fourier basis – while maintaining privacy – fails to achieve the correct convergence rate. However we also note that this type of noise addition is essentially overkill for this problem, and we demonstrate a different method based on the K-norm mechanism. By tailoring the noise addition to the underlying geometry of the problem we are finally able to achieve the correct rate of convergence.

4.2 Non Private Kernel Density Estimation

We begin by reviewing some basic facts about the non-private estimator, and set up some notation.

First we denote the one-dimensional Gaussian kernel

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\},$$

and in the interest of compact equations we also denote

$$\phi_\sigma(x, y) = \sigma^{-1}\phi\left(\frac{x-y}{\sigma}\right).$$

In moving to d dimensions we make use of the isotropic Gaussian kernel which we denote

$$\phi^d(x) = \prod_{i=1}^{d} \phi(x_i), \quad \phi^d_\sigma(x, y) = \prod_{i=1}^{d} \phi_\sigma(x_i, y_i)$$

for $x, y \in \mathbb{R}^d$.

Recall the kernel density estimator based on some sample $X = (x^1, \ldots, x_n)$ (the superscript being used to denote the index since we reserve the subscript for the components of each x^i)

$$\hat{f}_X(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \phi^d\left(\frac{x - x^i}{h}\right) = n^{-1} \sum_{i=1}^{n} \phi^d_h(x^i, x). \tag{4.1}$$

When the data arrive iid from some underlying density f, and under suitable regularity conditions on f (see e.g., Wasserman [2006]) then the estimator (4.1) enjoys the minimax rate of convergence

$$\mathbb{E} \int_{[0,1]^d} \left(\hat{f}_X(x) - f(x)\right)^2 = c_1 h^4 + \frac{c_2}{nh^d} = O\left(n^{-4/(4+d)}\right), \tag{4.2}$$

where the latter rate is attained by choosing

$$h = h_n = O\left(n^{-1/(4+d)}\right). \tag{4.3}$$

4.3 Fourier Approximation of the Kernel Density Estimate

We focus on the unit interval, in which the Fourier basis functions

$$\psi_1(x) = 1, \quad \psi_{2j}(x) = \sqrt{2}\cos(2j\pi x), \quad \psi_{2j+1}(x) = \sqrt{2}\sin(2j\pi x), \tag{4.4}$$

form an orthonormal basis for the periodic functions. We thus consider the periodic approximation to the kernel

$$\tilde{\phi}_\sigma(x,y) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)\psi_i(y), \tag{4.5}$$

in which

$$\lambda_1 = 1, \quad \lambda_{2j} = \lambda_{2j+1} = \exp\left\{-2j^2\sigma^2\right\}.$$

In d dimensions we use the d-fold tensor product of the basis considered above, namely by taking

$$\tilde{\phi}^d_\sigma(x,y) = \prod_{j=1}^{d} \tilde{\phi}_\sigma(x_j,y_j) = \sum_{a_1=1}^{\infty} \cdots \sum_{a_d=1}^{\infty} \left(\prod_{j=1}^{d} \lambda_{a_j}\right) \left(\prod_{j=1}^{d} \psi_{a_j}(x_j)\right) \left(\prod_{j=1}^{d} \psi_{a_j}(y_j)\right).$$

We defined the λ_i so as to obtain the following result.

Lemma 4.3.1.

$$\int_{[0,1]} \tilde{\phi}_\sigma(x,y)\psi_i(x)\ dx = \int_{\mathbb{R}} \phi_\sigma(x,y)\psi_i(x)\ dx, \tag{4.6}$$

Proof. Since

$$\cos(x) = \frac{e^{ix}+e^{-ix}}{2}, \quad \sin(x) = \frac{e^{ix}-e^{-ix}}{2i},$$

where i is the imaginary unit, and the characteristic function of a Gaussian is

$$\int_{\mathbb{R}} \phi_\sigma(x,\mu)e^{itx}\ dx = \exp\left\{it\mu - \frac{\sigma^2t^2}{2}\right\},$$

we have

$$\int_{\mathbb{R}} \phi_\sigma(x,\mu)\cos(ax+b)\ dx = \exp\left\{-\frac{a^2\sigma^2}{2}\right\}\cos(a\mu+b),$$

and

$$\int_{\mathbb{R}} \phi_\sigma(x,\mu)\sin(ax+b)\ dx = \exp\left\{-\frac{a^2\sigma^2}{2}\right\}\sin(a\mu+b).$$

From which the result is immediate. □

Thus we anticipate that whenever the original functions $\phi_\sigma(\cdot,y)$ concentrate mostly in the unit interval (in the sense that their integral outside of this interval is small) that the above Fourier approximation will be good in the sense of small ℓ_2 error (integrated square error). We now quantify this error in the one dimensional case.

Proposition 4.3.2. *For*

$$\sqrt{\sigma} < \mu < 1 - \sqrt{\sigma}, \tag{4.7}$$

we have

$$\int_{[0,1]} \left(\phi_\sigma(x,\mu) - \tilde{\phi}_\sigma(x,\mu)\right)^2 \, dx = O\left(\sigma^p\right), \tag{4.8}$$

for any p.

Proof.

$$\begin{aligned}
\int_{[0,1]} \left(\phi_\sigma(x,\mu) - \tilde{\phi}_\sigma(x,\mu)\right)^2 \, dx &= \sum_{i=1}^{\infty} \left(\int_{[0,1]} \psi_i(x) \left[\phi_\sigma(x,\mu) - \tilde{\phi}_\sigma(x,\mu)\right] \, dx \right)^2 \\
&= \sum_{i=1}^{\infty} \left(\int_{[0,1]} \psi_i(x)\phi_\sigma(x,\mu) \, dx - \int_{\mathbb{R}} \psi_i(x)\phi_\sigma(x,\mu) \, dx \right)^2 \\
&= \sum_{i=1}^{\infty} \left(\sum_{j\neq 0} \int_{[j,j+1]} \psi_i(x)\phi_\sigma(x,\mu) \, dx \right)^2 \\
&= \sum_{i=1}^{\infty} \left(\sum_{j\neq 0} \int_{[0,1]} \psi_i(x)\phi_\sigma(x-j,\mu) \, dx \right)^2 \\
&= \int_{[0,1]} \left(\sum_{j\neq 0} \phi_\sigma(j, |\mu - x|) \right)^2 \, dx
\end{aligned}$$

the reasons for the equalities are (in order): Parseval's identity, the equality (4.6), breaking up an integral into pieces, periodicity of the bases, Parseval's identity again along with making use of symmetry. Under the condition (4.7) then $|x-\mu| < 1 - \sqrt{\sigma}$ for whatever $x \in [0,1]$ and so

$$\begin{aligned}
\sum_{j\neq 0} \phi_\sigma(j, |\mu - x|) &< \phi_\sigma(1, |y\mu - x|) + \int_{(-\infty,-1]} \phi_\sigma(j, |\mu - x|) \, dj + \int_{[2,\infty)} \phi_\sigma(j, |\mu - x|) \, dj \\
&\leq \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma}\right\} + 2\sigma \exp\left\{-\frac{1}{2\sigma^2}\right\}.
\end{aligned}$$

This quantity is clearly negligible in that it is eventually smaller than any polynomial in σ. □

In Figure 4.1 the resulting approximations are shown. When the means are sufficiently far into the interval, and when σ is small the kernel is approximately periodic already and so the error in the approximation is small as demonstrated above.

In d dimensions (treating the dimension as fixed) we again have approximation error being

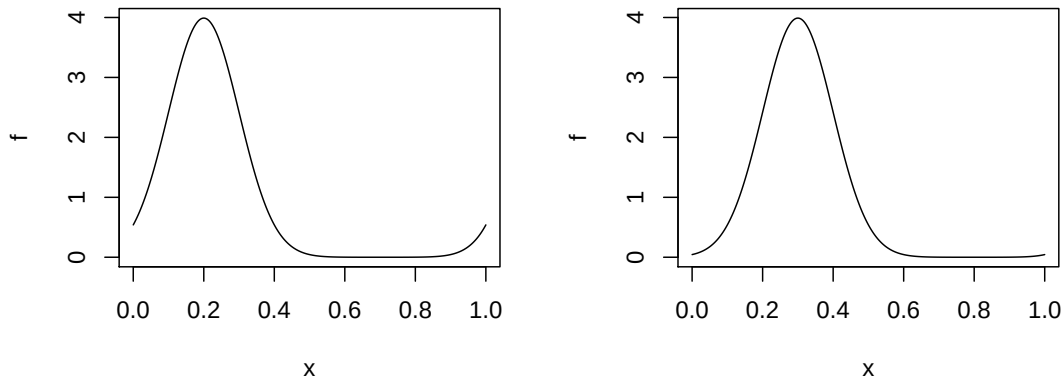


Figure 4.1: The periodic approximation to the Gaussian kernel with $\sigma = 0.1$ In the left $\mu = 0.2$ and so the periodic nature of the approximation is noticeable on the right hand side of the interval. In the right $\mu = 0.3$ and since this kernel puts more "mass" on the inside of the interval, this effect is reduced.

negligible as σ decreases to zero, since

$$\int_{[0,1]^d} \left(\tilde{\phi}_\sigma^d(x, \mu) - \phi_\sigma^d(x, \mu) \right)^2 \, dx = \prod_{j=1}^{d} \int_{[0,1]} \left(\tilde{\phi}_\sigma(x_j, \mu_j) - \phi_\sigma(x_j, \mu_j) \right)^2 \, dx_j.$$

We now consider conditions for the periodic approximation

$$\tilde{f}_X(x) = n^{-1} \sum_{i=1}^{n} \tilde{\phi}_h^d(x^i, x) \tag{4.9}$$

to achieve the correct convergence rate (4.2). To avoid complications arising from the condition (4.7) we could e.g., assume that f is supported in some subset of $[\eta, 1 - \eta]$ for some small constant $\eta > 0$, but this is not necessary. An alternative is to transform the data so that it lies in such an interval, do the density estimation there, then transform the resulting function back to the unit interval. For fixed η this corresponds to a density estimator for which the bandwidth h is off by a constant factor, and so does not affect the rate of convergence. Denote the transform as

$$\pi(x) = \eta + (1 - 2\eta)x,$$

then

$$\phi_{\sigma}^{d}\left(\pi(x),\pi(y)\right)=\phi_{\frac{\sigma}{1-2\eta}}^{d}\left(x,y\right),$$

and so the transformed version of (4.1) corresponds to the choice of $h'=h/(1-2\eta)$ which evidently achieves the correct rate since η is regarded as a constant. Then for sufficiently small h (sufficiently large n) we find

$$\int_{[0,1]^d}\left(n^{-1}\sum_{i=1}^{n}\phi_h^d\left(\pi(x^i),\pi(x)\right)-\tilde{\phi}_h^d\left(\pi(x^i),\pi(x)\right)\right)^2\,dx=O\left(h^p\right),$$

for any p, due to Proposition 4.3.2. For h on the order prescribed in (4.3) we evidently may take $p=4$ in order to see that the correct rate is maintained. For the remainder of this document we will suppose f to be supported within the interval $[\eta,1-\eta]$ rather than perform this transformation, in order to save space, but note this comes at no cost to generality of the approach. Evidently

$$\int_{[0,1]^d}\left(f_X(x)-\tilde{f}_X(x)\right)^2\,dx\leq\sup_{\mu\in[\eta,1-\eta]^d}\int_{[0,1]^d}\left(\phi_h^d(x,mu)-\tilde{\phi}_h^d(x,\mu)\right)^2\,dx=O(h^p),$$

for any p. Thus taking h as in (4.3) still leads to the correct rate under this approximation.

4.3.1 The Truncated Fourier Approximation

To reduce the kernel density estimator to a finite dimensional quantity we consider the truncation to the first $2m+1$ fourier coefficients (namely the constant basis function, and the sines and cosines up to frequency m). In one dimension this corresponds to

$$\tilde{f}_X^m(x)=n^{-1}\sum_{i=1}^{n}\left(1+\sum_{j=2}^{2m+1}\lambda_j\psi_j(x^i)\psi_j(x)\right)=1+\sum_{j=2}^{2m+1}\lambda_j\left(n^{-1}\sum_{i=1}^{n}\psi_j(x^i)\right)\psi_j(x).$$

In higher dimension we similarly truncate to the first $2m+1$ basis functions in each coordinate axis and consider the tensor product

$$\tilde{f}_X^m(x)=\sum_{a_1=1}^{2m+1}\cdots\sum_{a_d=1}^{2m+1}\left(\prod_{j=1}^{d}\lambda_{a_j}\right)\left(n^{-1}\sum_{i=1}^{n}\left(\prod_{j=1}^{d}\psi_{a_j}(x^i)\right)\right)\left(\prod_{j=1}^{d}\psi_{a_j}(x)\right).\tag{4.10}$$

We quantify the error between the truncation and the full fourier approximation in the following statement.

Proposition 4.3.3.

$$\sup_X \int_{[0,1]^d} \left(\tilde{f}_X(x) - \tilde{f}_X^m(x) \right)^2 = O\left(\frac{2^d}{mh^d} e^{-c_1 m^2 h^2} \right), \tag{4.11}$$

for some universal constant c_1*.*

Proof. We have

$$\int_{[0,1]^d} \left(\tilde{f}_X(x) - \tilde{f}_X^m(x) \right)^2 \leq \left(\sum_{i=1}^{\infty} \lambda_i^2 + \sum_{j=2m+2}^{\infty} \lambda_j^2 \right)^d - \left(\sum_{i=1}^{\infty} \lambda_i^2 \right)^d.$$

By bounding the sum with an integral and using the Gaussian tail inequality we obtain

$$\sum_{j=2m+2}^{\infty} \lambda_j^2 = 2 \sum_{j=m+1}^{\infty} e^{-4j^2h^2} \leq 2\sqrt{2\pi} \int_{[m,\infty)} \phi\left(\frac{j}{\frac{1}{2\sqrt{2}h}} \right) dj \leq \frac{c_1}{mh} e^{-c_2 m^2 h^2},$$

and likewise

$$\sum_{j=1}^{\infty} \lambda_j^2 \leq \frac{c_3}{h}.$$

Thus

$$\int_{[0,1]^d} \left(\tilde{f}_X(x) - \tilde{f}_X^m(x) \right)^2 \leq \frac{c_4 2^d}{mh^d} e^{-c_2 m^2 h^2}.$$

In the original statement we rename c_2 to c_1 since it is the only constant to appear. □

In light of the above, to ensure that the rate (4.2) is preserved it suffices to take h as in (4.3), and

$$m = O\left(\frac{\log n}{h} \right) = O\left(n^{1/(4+d)} \log n \right). \tag{4.12}$$

The reason is that the full expansion achieves the correct rate under this choice of h, and then the above choice of m ensures that the right hand side of (4.11) is on the same order as the requisite convergence rate. Note that we treat d as a constant.

So far then, we have a finite dimensional representation of the kernel density estimate, in which the parametrization does not depend on the data itself, and which does not spoil the convergence rate enjoyed by the unadulterated estimator.

4.4 Differentially Private Versions of the Truncated Density Estimator

We now turn to the construction of a private estimator based on the above.

4.4.1 Privacy via Laplace Noise Addition

To make a differentially private kernel density estimator it clearly suffices to add to each of the $O(m^d)$ coordinates of $\tilde{f}_X^m$ a Laplace random variate having scale parameter $O(n\alpha/m^d)$. Denoting the resulting random, private function

$$\tilde{f}_X^{m,\alpha}(x) = \sum_{a_1=1}^{2m+1} \cdots \sum_{a_d=1}^{2m+1} \left(\prod_{j=1}^{d} \lambda_{a_j} \right) \left(L_{a_1,\ldots,a_d} + n^{-1} \sum_{i=1}^{n} \left(\prod_{j=1}^{d} \psi_{a_j}(x^i) \right) \right) \left(\prod_{j=1}^{d} \psi_{a_j}(x) \right),$$

$$L_{a_1,\ldots,a_d} \overset{\text{iid}}{\sim} \text{Lap}\left(0, \frac{n\alpha}{(2m+1)^d}\right).$$

Regarding α as a constant and taking m in the order prescribed by (4.12) leads to an error on the order

$$\mathbb{E} \int_{[0,1]^d} \left(\tilde{f}_X^m - \tilde{f}_X^{m,\alpha} \right)^2 = \sum_{a_1=1}^{2m+1} \cdots \sum_{a_d=1}^{2m+1} \left(\prod_{j=1}^{d} \lambda_{a_j} \right) L_{a_1,\ldots,a_d} = O\left(\frac{m^{3d}}{n^2} \right) = O\left(n^{\frac{d-8}{4+d}} (\log n)^{3d} \right),$$

evidently this estimator fails to achieve the minimax rate in dimension greater than 3. We therefore look for an alternative noise model to use for perturbation. Nevertheless we note that this technique is reasonable to use for data up to 3 dimensions. We note that the technique described above is essentially the analog of the "orthogonal series expansion" estimator in Wasserman and Zhou [2010]. There the same problem occurs when the dimension grows, however they only considered the case of the one dimensional estimator which they found to enjoy the minimax rate.

That the possibility exists to construct a method with a faster convergence rate may be seen by noting that the noise added above is "overkill" for the problem. These independent Laplace draws would yield privacy even if the coordinate vectors of the functions for one dataset and all its neighbors formed a hypercube (with side length constant as m increases). To see this note that we treated each coefficient separately and made use of the composition property of differential privacy along with the appropriately adjusted α in order to achieve privacy. However, while it is true that neighboring datasets may differ in a constant fraction of the coordinates (e.g., by replacing a data point with one in the center of the interval, which has magnitude 1 in every cosine basis), the corresponding coefficient vectors do not fill up a hypercube as described above (e.g., they only

fill up a hyper rectangle in which the side length decreases as the dimension increases). Thus we anticipate that further study of these coefficient vectors will lead to an improved method. This is exactly what we pursue in the remainder of this chapter.

4.4.2 Privacy via "Caratheodory Noise" Addition

In this section we apply the K-norm method (Hardt and Talwar [2010] and also section 1.3 of this thesis) to the release of the truncated fourier approximation to the kernel density estimator. The actual construction of the method is straightforward, the difficult problems are the demonstration of the error rate (which involves the moments of inertia of a complicated convex body) and the development of a sampling procedure for the resulting noise. We show that the coordinate vectors lay in a convex body called the "Caratheodory Orbitope" Sanyal *et al.* [2009], and hence we name the resulting noise "Caratheodory Noise."

We first note that a Gaussian kernel in one dimension with mean μ, being approximated by the first $2m+1$ fourier coefficients corresponds to a point in $\mathbb{R}^{2m+1}$ with the coordinates

$$\Psi_m(\mu) = (1, \psi_2(\mu), \ldots, \psi_{2m+1}(\mu)),$$

with ψ_i as in (4.4). The release of (4.10) in the one dimensional case corresponds to the release of an average of these points

$$\bar{\Psi} = \frac{1}{n}\sum_{i=1}^{n} \Psi_m(x_i).$$

Note that $\Psi_m : [0,1] \to \mathbb{R}^{2m+1}$ defines a closed curve in space, which is inscribed upon the surface of a hypersphere in dimension $2m$ having radius $\sqrt{m}$. Evidently $\bar{\Psi}$ is a point in the convex hull of this curve

$$K_m = \operatorname{conv}\{\Psi_m(x), x \in [0,1]\}.$$

This body is known as the Universal Caratheodory orbitope, and has been studied in the algebraic geometry literature (see e.g., Sanyal *et al.* [2009]). The volume (as measured in the $2m$-dimensional subspace in which it lies) of a particular scaling of this body is given by Schoenberg [1954], from which we deduce the volume in question to be

$$\operatorname{Vol}_{2m}(K_m) = \frac{2^m \pi^m m!}{(2m)!} \approx \frac{\pi^m}{2^m m!},$$

where the latter is due to Sterling's approximation of the factorial function. Recall from Section 1.3 the "K-norm mechanism," for a one dimensional estimator we instantiate that here where the body

in question is K_m, namely by releasing the vector of fourier coefficients

$$\tilde{\Psi} \sim p_{\bar{\Psi}}(x) \propto \exp\left\{-\frac{\alpha}{\Delta}\|\bar{\Psi} - x\|_{K_m}\right\},$$

where

$$\Delta \geq \frac{1}{n} \sup_{a,b \in K_m} \|a - b\|_{K_m},$$

evidently it suffices to take $\Delta = 2/n$. As demonstrated in Hardt and Talwar [2010], sampling this distribution may be decomposed into two parts, and it suffices to release

$$\tilde{\Psi} = \bar{\Psi} + ru, \quad r \sim \Gamma(2m+1, \frac{2}{n\alpha}), \quad u \sim U(K_m),$$

in which we use $U(K)$ to mean the uniform distribution over K. For such a release the corresponding function is given by

$$\tilde{f}_X^{m,K} = 1 + \sum_{j=2}^{2m+1} \lambda_j \tilde{\Psi}_j \psi_j(x) = 1 + \sum_{j=2}^{2m+1} \lambda_j \left(ru_j + \frac{1}{n}\sum_{i=1}^{n} \psi_j(x^i) \right) \psi_j(x).$$

Since each $\lambda_j \leq 1$ and since the ψ_j form an orthonormal basis in $[0,1]$, the overall error rate for this approach is then

$$\mathbb{E}\|\tilde{f}_X - \tilde{f}_X^{m,K}\|^2 \leq \mathbb{E}\|\tilde{\Psi} - \bar{\Psi}\|^2 = \mathbb{E}r^2 \cdot \mathbb{E}\|u\|^2 = O\left(\frac{m^2}{n^2}\right)\mathbb{E}\|u\|^2.$$

In d dimensions, the released function corresponds to a point in the convex hull of the d-fold tensor product of the curve Ψ_m

$$K_m^d = \operatorname{conv}\left\{\operatorname{vec}\left(\Psi_m(x_1) \otimes \cdots \otimes \Psi_m(x_d)\right), x_i \in [0,1]\right\},$$

where we interpret $\otimes$ as the vector outer product, and the resulting body in $\mathbb{R}^{(2m+1)^d}$ as being formed from the points obtained by stacking the resulting arrays into large vectors. Once again the appropriate noise vector is obtained through the composition of a Gamma scalar and a uniform vector from this convex body. We therefore release

$$\tilde{\Psi} = \bar{\Psi} + ru, \quad r \sim \Gamma\left((2m+1)^d, \frac{2}{n\alpha}\right), \quad u \sim U(K_m^d),$$

The overall error rate is

$$\mathbb{E}\|\tilde{f}_X - \tilde{f}_X^{m,K}\|^2 = O\left(\frac{m^{2d}}{n^2}\right)\mathbb{E}\|u\|^2, \quad u \sim U(K_m^d).$$

The remaining questions of this chapter are therefore the rate at which this latter expectation increases, and the feasibility of performing the requisite sampling.

The above expectation is sometimes called the "moment of inertia" of K_m following the connection to the physics of rotating bodies. Here we determine the requisite quantity analytically. Note that since the first coordinate of Ψ_m is the constant 1, we consider the coordinate projection of the resulting body to the remaining $2m$ coordinates. Thus we define

$$K'_m = \operatorname{conv}\left\{\Psi'_m(x), x \in [0,1]\right\}.$$

with

$$\Psi'_m(\theta) = (\psi_2(\theta), \ldots, \psi_{2m+1}(\theta)),$$

and note that

$$\mathbb{E}\|u\|^2 = 1 + \mathbb{E}\|w\|^2, \quad u \sim U(K_m), \quad w \sim U(K'_m).$$

Define the inertia matrix $M(K'_m)$ with

$$M_{ij}(K'_m) = \frac{1}{\operatorname{Vol}(K'_m)} \int_{K'_m} x_i x_j,$$

Evidently the term $\mathbb{E}\|w\|^2$ which appeared in the above error rate for the one dimensional density estimator is just the trace of this matrix. We present the result first along with its salient consequences before dedicating some space to its demonstration.

Proposition 4.4.1.

$$M_{ij}(K'_m) = \begin{cases} \frac{1}{4m-3(i-1)} & i = j \\ 0 & i \neq j \end{cases}.$$

This immediately leads to

$$\mathbb{E}\|w\|^2 \leq 1, \quad w \sim U(K'_m),$$

where the inequality holds irrespective of m. Therefore in the one dimensional case, the above private density estimator has the convergence rate

$$\mathbb{E}\|f_X - \tilde{f}_X^{m,K}\|^2 = O\left(n^{-4/5}\right) + O\left(\frac{m^2}{n^2}\right) = O\left(n^{-4/5}\right),$$

taking m as prescribed above in (4.12). In d dimensions we note that

$$\mathbb{E}\|u\|^2 = \left(1 + \mathbb{E}\|w\|^2\right)^d \leq 2^d, \quad u \sim U(K_m^d), \quad w \sim U(K'_m). \tag{4.13}$$

In light of (4.13) for a multidimensional density estimator made private in the above fashion we

obtain the overall error rate of

$$\mathbb{E}\|\tilde{f}_X - \tilde{f}_X^{m,K}\|^2 = O\left(\frac{m^{2d}}{n^2}\right),$$

where we once again treat d as a constant. Finally taking m as in (4.12) with h as in (4.3) leads to

$$\mathbb{E}\|f_X - \tilde{f}_X^{m,K}\|^2 = O\left(\frac{(\log n)^{2d}}{h^{2d}n^2}\right) = O\left(\frac{(\log n)^{2d}}{n^{2d/(d+4)}n^2}\right) = O\left(n^{-4/(d+4)}\frac{(\log n)^{2d}}{n^{4/(d+4)}}\right) = O\left(n^{-4/(d+4)}\right),$$

where the last equality is since every positive power of n is eventually larger than $\log n$. Therefore we find that the method described above maintains the correct convergence rate. In fact as evident from the penultimate expression on the right, the rate at which the private estimator converges to the non-private one is substantially faster than the convergence of the latter to the true density.

4.4.3 Inertia of the Caratheodory Orbitope

In Schoenberg [1954] it was demonstrated that for curves of a given length in $\mathbb{R}^{2m}$, the one which maximizes the volume of the convex hull is given by an affine transform of Ψ_m. In this section we use some ideas of this paper in order to get at the inertia of this body rather than the volume. Following Schoenberg [1954] we proceed by approximating K_m with a sequence of polytopes having limit point K'_m. Thus we consider

$$K_{m,n} = \operatorname{conv}\left\{\Psi'_m\left(2\pi\frac{k}{n}\right) : k = 0, \ldots, n-1\right\},$$

evidently the above body is a polytope and so is the union of a set of disjoint simplices. One such representation Schoenberg [1954] is

$$K_m^n = \bigcup_{i_1 \le i_2 \le \ldots \le i_m} S_{i_1,\ldots,i_m},$$

where the simplices are

$$S_{i_1,\ldots,i_m} = \operatorname{conv}\left\{0, \Psi'_m\left(2\pi\frac{i_1}{n}\right), \Psi'_m\left(2\pi\frac{i_1+1}{n}\right), \ldots, \Psi'_m\left(2\pi\frac{i_m}{n}\right), \Psi'_m\left(2\pi\frac{i_m+1}{n}\right)\right\}.$$

Therefore we have

$$M(K_{m,n}) = \frac{1}{\operatorname{Vol}_{2m}(K_{m,n})} \sum_{i_1 \le i_2 \le \ldots \le i_m} \operatorname{Vol}_{2m}(S_{i_1,\ldots,i_m}) M(S_{i_1,\ldots,i_m}). \tag{4.14}$$

We first recall some basic geometric properties of simplices. Denote by Δ_d the "standard" simplex in d-dimensional Euclidean space, that is

$$\Delta_d = \left\{ x \in \mathbb{R}^d : \min_i x_i \geq 0,\ \sum_{i=1}^d x_i \leq 1 \right\},$$

evidently we have $\mathrm{Vol}_d(\Delta_d) = 1/d!$, since the unit hypercube in d dimensions is the union of $d!$ such disjoint simplices, each being rotated and translated appropriately.

Proposition 4.4.2. *The inertia matrix for the simplex is given by*

$$M_{ij}\left(\Delta_d\right) = \begin{cases} \frac{2}{(d+1)(d+2)} & i = j \\ \frac{1}{(d+1)(d+2)} & i \neq j \end{cases}. \tag{4.15}$$

The proof appears at the end of this section. Consider the convex hull formed by the origin and d more points in $\mathbb{R}^d$. We may denote such a body as

$$A\Delta_d = \left\{ Ax : x \in \Delta_d \right\},$$

in which A is the matrix having columns specified by the non-origin coordinates of the corners of the body. Evidently we have

$$M(A\Delta_d) = AM(\Delta_d)A^T,$$

since

$$\frac{1}{\mathrm{Vol}_d(A\Delta_d)} \int_{A\Delta_d} xx^T = \frac{1}{|A|\mathrm{Vol}_d(\Delta_d)} \int_{\Delta_d} (Ax)(Ax)^T |A| = A \left(\frac{1}{\mathrm{Vol}_d(\Delta_d)} \int_{\Delta_d} xx^T \right) A^T.$$

Returning to (4.14) and defining the matrix

$$A_{i_1,\ldots,i_m} = \left(\Psi'_m \left(2\pi \frac{i_1}{n} \right), \Psi'_m \left(2\pi \frac{i_1+1}{n} \right), \ldots, \Psi'_m \left(2\pi \frac{i_m}{n} \right), \Psi'_m \left(2\pi \frac{i_m+1}{n} \right) \right),$$

then we evidently have

$$\mathrm{Vol}_{2m}(S_{i_1,\ldots,i_m}) = \frac{1}{(2m)!} |A_{i_1,\ldots,i_m}|,$$

and

$$M(S_{i_1,\ldots,i_m}) = A_{i_1,\ldots,i_m} M(\Delta_{2m}) A^T_{i_1,\ldots,i_m}.$$

Passing into the limit as $n \to \infty$ we obtain

$$M_{ij}(K'_m) = \int_{\theta_1} \cdots \int_{\theta_m} \frac{D(\theta_1, \ldots, \theta_m)}{(2m+1)(2m+2)} \left(6 \sum_{k=1}^{m} \psi_i(\theta_k)\psi_j(\theta_k) + 4 \sum_{l \neq m}^{m} \psi_i(\theta_l)\psi_j(\theta_m) \right) \tag{4.16}$$

in which we defined

$$D(\theta_1, \ldots, \theta_m) = \frac{|B(\theta_1, \ldots, \theta_m)|}{\int_{\theta_1} \cdots \int_{\theta_m} |B(\theta_1, \ldots, \theta_m)|},$$

where B is the matrix given by

$$B(\theta_1, \ldots, \theta_m) = \left(\Psi'_m(\theta_1), \frac{\partial}{\partial \theta_1} \Psi'_m(\theta_1), \ldots, \Psi'_m(\theta_m), \frac{\partial}{\partial \theta_m} \Psi'_m(\theta_m) \right).$$

Once more, following Schoenberg [1954] we write the determinant as a sum of appropriately signed products of determinants of the 2×2 minors, in which each minor is made up of neighboring columns of B which both involve the same variable θ_i. This way, the multiple integral of each product decomposes into a product of integrals. What's more, the only such integrals which are non-zero are when each minor is made up of adjacent rows of the original matrix, involving functions having the same frequency. After integrating, we find the following

$$\int_{\theta_1} \cdots \int_{\theta_m} |B(\theta_1, \ldots, \theta_m)| = (2\pi)^m (m!)^2,$$

and

$$\int_{\theta_2} \cdots \int_{\theta_m} |B(\theta_1, \ldots, \theta_m)| = (2\pi)^{m-1} (m!)^2, \tag{4.17}$$

the last integral required for the inertia matrix is

$$\int_{\theta_3} \cdots \int_{\theta_m} |B(\theta_1, \ldots, \theta_m)| = \tag{4.18}$$
$$(2\pi)^{m-2}(m-2)! \sum_{j<i} \frac{m!}{ij} \left(2ij - \frac{(i+j)^2}{2} \cos[(i-j)(\theta_1 - \theta_2)] + \frac{(i-j)^2}{2} \cos[(i+j)(\theta_1 - \theta_2)] \right),$$

where we made use of the identity

$$\cos[k(\theta_1 - \theta_2)] = \sin(k\theta_1)\sin(k\theta_2) + \cos(k\theta_2)\cos(k\theta_2). \tag{4.19}$$

Evidently the integrals over the other sets of variables of the same sizes as the above will be equal due to the symmetry (or equivalently, by re-ordering columns in B which does not change the determinant).

Substituting (4.19) back into (4.18) and then inserting into (4.16) clearly leads to the demonstration that the requisite inertia matrix is diagonal. Thus we concentrate on the diagonal terms. We have

$$\int_{\theta_1} \cdots \int_{\theta_m} D(\theta_1, \ldots, \theta_m)\psi_i^2(\theta_1) = \frac{1}{2},$$

and

$$\int_{\theta_1} \cdots \int_{\theta_m} D(\theta_1, \ldots, \theta_m)\psi_i(\theta_1)\psi_i(\theta_2) = \frac{-8m^2 + (9\lceil i/2 \rceil - 3)m + 3}{m(m+1)(4m - 3\lceil i/2 \rceil + 3)},$$

which when substituted into (4.16) leads to the stated inertia.

Proof of Proposition 4.4.2. We determine the entry corresponding to x_1, the other diagonal terms being equal due to the symmetry of Δ_d.

$$\begin{aligned}
M_{11}(\Delta_d) &= d! \int_0^1 x_1^2 \int_0^{1-x_1} \cdots \int_0^{1-x_{d-1}} \, dx_d \cdots dx_2 \, dx_1 \\
&= d! \int_0^1 x_1^2 (1-x_1)^{d-1} \mathrm{Vol}_{d-1}(\Delta_{d-1}) \, dx_1 \\
&= d \int_0^1 x_1^2 (1-x_1)^{d-1} \, dx_1 \\
&= d \int_0^1 \sum_{i=2}^{d+1} (-1)^i \binom{d-1}{i-2} x_1^i \, dx_1 \\
&= d \sum_{i=0}^{d-1} (-1)^i \binom{d-1}{i} \frac{1}{i+3}.
\end{aligned}$$

Denote

$$f(d) = \sum_{i=0}^{d-1} (-1)^i \binom{d+2}{i+3} (i+1)(i+2),$$

then $f(d) = f(d-1)$ since

$$\begin{aligned}
f(d) &= \sum_{i=0}^{d-2} (-1)^i \left(\binom{d+1}{i+2} + \binom{d+1}{i+3} \right) (i+1)(i+2) + (-1)^{d-1} d(d+1) \\
&= f(d-1) + \sum_{i=0}^{d-2} (-1)^i \binom{d+1}{i+3} (i+1)(i+2) + (-1)^{d-1} d(d+1) \\
&= f(d-1) + d(d+1) \sum_{i=0}^{d-1} (-1)^i \binom{d-1}{i} \\
&= f(d-1)
\end{aligned}$$

and so since $f(1) = 2$,

$$M_{11}(\Delta_d) = \frac{f(d)}{(d+1)(d+2)} = \frac{2}{(d+1)(d+2)}.$$

The off diagonal terms are likewise equal due to the symmetry and so

$$\begin{aligned}
M_{12}(\Delta_d) &= d! \int_0^1 x_1 \int_0^{1-x_1} x_2(1-x_1-x_2)^{d-2} \frac{1}{(d-2)!} \\
&= d(d-1) \int_0^1 x_1 \int_0^{1-x_1} x_2(1-x_1-x_2)^{d-2} \\
&= d(d-1) \int_0^1 \left(\sum_{i=0}^{d-2} (-1)^i \binom{d-2}{i} (1-x_1)^d \frac{1}{i+2} \right) \\
&= d(d-1) \left(\sum_{i=0}^{d-2} (-1)^i \binom{d-2}{i} \frac{1}{i+2} \right) \int_0^1 x_1(1-x_1)^d \\
&= d(d-1) \left(\sum_{i=0}^{d-2} (-1)^i \binom{d-2}{i} \frac{1}{i+2} \right) \left(\sum_{i=0}^{d} (-1)^i \binom{d}{i} \frac{1}{i+2} \right).
\end{aligned}$$

Denote

$$g(d) = \sum_{i=0}^{d} (-1)^i \binom{d+2}{i+2} (i+1),$$

then following the same steps as above for f we have $g(d) = g(d-1)$, also $g(1) = 1$ and so

$$M_{12}(\Delta_d) = \frac{g(d-2)g(d)}{(d+1)(d+2)} = \frac{1}{(d+1)(d+2)}.$$

□

4.4.4 Sampling the Convex Body

The goal of this chapter is to demonstrate an approach for sampling uniformly from inside K_m. Evidently since K_m is contained in the hypersphere of radius $\sqrt{m}$, one possibility is to sample uniformly from the hypersphere, then to reject points which lay outside K_m. However due to the dramatic rate at which the volume of K_m decays relative to that of the containing hypersphere, this technique requires astronomical numbers of samples in order to be useful in high dimension. For example, to generate one sample requires on average $(2m)^{m-1}$ samples from the sphere.

We implement the "hit and run" approach of Vempala [2005]. This is a random walk inside K_m, where at each iteration a random direction is chosen, and a new point is chosen uniformly at random from the line segment given by the intersection of K_m with the ray passing through the current point and having the chosen direction.

All that is required to perform this sampling method efficiently is a technique to determine the extents of this line segment at each iteration. Note that per Sanyal *et al.* [2009] (theorem 5.2) the body K_m is isomorphic to a certain Hermitian Toeplitz spectrahedron. We have

$$(s_1, c_1, \ldots, s_m, c_m) \in K_m \Leftrightarrow \begin{pmatrix} 1 & c_1 + is_1 & \cdots & c_m + is_m \\ c_1 - is_1 & 1 & \cdots & c_{m-1} + is_{m-1} \\ \vdots & \ddots & \ddots & \vdots \\ c_m - is_m & c_{m-1} - is_{m-1} & \cdots & 1 \end{pmatrix} \succeq 0$$

where the latter means that the matrix is positive semidefinite, and where i is the imaginary unit. Note that this matrix is the sum of a real symmetric Toeplitz matrix and an imaginary skew-symmetric Toeplitz matrix. Finally note the connection between this representation and the so-called "trigonometric moment problem" described by Caratheodory. In this case it demonstrates that points in this convex hull correspond to equivalence classes of Borel probability measures on the interval $[0, 1]$ (where equivalence means equality of the first $2m$ fourier coefficients). Denote by $A(x)$ the matrix corresponding to a point x, and $B(x) = A(x) - I$. Let $u, v \in \mathbb{R}^{2m}$ and consider the ray $\{u + \eta v : \eta \in \mathbb{R}\}$. Assuming $u \in K_m$ then the intersection of this ray with K_m corresponds to

$$\{u + \eta v : A(u) + \eta B(v) \succeq 0\}.$$

Since during the algorithm u is an interior point of K_m, the matrix $A(u)$ is non-singular, so we may write the condition as

$$I + \eta A(u)^{-1} B(v) \succeq 0,$$

the endpoints of this line segment are therefore

$$\frac{-1}{\lambda_{\min}\left(A(u)^{-1}B(v)\right)}, \quad \frac{1}{\lambda_{\min}\left(-A(u)^{-1}B(v)\right)}$$

where $\lambda_{\min}(\cdot)$ gives the minimum eigenvalue of the matrix. Thus the "hit and run" MCMC approach may be readily implemented, only requiring subroutines to solve for maximum and minimum eigenvalues of matrices, the dimension of which are of the order $O(n^{1/(4+d)})$, which may be feasible in many cases.

4.5 Summary

In this chapter we presented a technique for the differentially private release of a kernel density estimator which did not spoil the convergence rate of the unadulterated non-private estimator. First we showed the difficulty of porting the technique of the previous chapter to satisfy the α-

differential privacy as opposed to the approximate differential privacy. We demonstrated that the Laplace mechanism fails to maintain the minimax rate for this problem, and then gave a new technique based on the K-norm mechanism which achieved the correct rate.

What remains to be seen is whether it is possible to sample the requisite uniform distribution over the convex body associated with this new method. For the case of a one-dimensional density estimator we gave an efficient MCMC scheme, however we did not give an efficient scheme for higher dimensional estimators. We suspect that the MCMC technique may be extended, by way of a kind of tensor product construction to represent the convex hull as another spectrahedron, however this remains to be seen in future work.

Chapter 5

Weaker Alternatives to Differential Privacy

The goal of this section will be to demonstrate some modifications to the adversary in differential privacy. We will then determine whether lower noise levels may be appropriate on certain statistical tasks.

5.1 Random Differential Privacy

We may replace the global quantifier $\forall D \sim D'$ in the definition of differential privacy with a milder condition and thus obtain a weaker notion of privacy. One conceptually appealing way is to view the dataset D as a random sample of an unknown distribution, and to require that the DP condition holds with high probability. This leads to "random differential privacy." First we regard the database $D = (x_1, \ldots, x_n)$ as arising from independent draws from some unknown underlying probability distribution

$$x_i \overset{\text{iid}}{\sim} P,$$

where P is unspecified. We then define random differential privacy as "differential privacy with high probability" with respect to this underlying probability measure.

Definition 5.1.1 ((α, β, γ)-Random Differential Privacy)**.** We say that a family of distributions $P = \{P_D : D \in \mathcal{D}\}$ is (α, β, γ)-Randomly Differentially Private when:

$$\mathbb{P}\left(\forall A \in \mathcal{A},\ P_D(A) \leq e^{\alpha} P_{D'}(A) + \beta\right) \geq 1 - \gamma$$

Where $\mathbb{P}$ is the $n+1$-fold product measure over $x_1, \ldots, x_n$ and also x'_n which is the element in D' not in D.

Although an α-DP procedure fulfils the requirement of $(\alpha, 0)$-RDP, the converse is not true. The reason is that the latter requires that the condition (that the ratio of densities be bounded) holds almost everywhere with respect to the unknown measure, whereas DP require that this condition holds uniformly everywhere in the space.

We next show an important property of the definition, namely, that RDP algorithms may be composed to give other RDP algorithms with different constants.

Proposition 5.1.2 (Composition)**.** *Suppose P, P' are families of distributions over $\mathcal{Z}, \mathcal{Z}'$ which are (α, β, γ)-RDP and $(\alpha', \beta', \gamma')$-RDP respectively. The release of outputs of both procedures achieves $(\alpha + \alpha', \gamma + \gamma')$-RDP.*

This result is simply an application of the union bound combined with the standard composition property of differential privacy. As an example, suppose it is required to release k different statistics of some data sample. If each one is released via a $(\alpha/k, \gamma/k)$-RDP procedure, then the overall release of all k statistics together achieves (α, γ)-RDP.

As we have relaxed the notion of differential privacy, our methods provide weaker guarantees with respect to individual privacy.

5.1.1 Remarks about Random Differential Privacy

As we have proposed a loosened definition which is alternative to Differential Privacy, some remarks are in order. First we explain how our definition may be thought of in words rather than symbols. We present the weaknesses and then the strengths of our privacy criteria as compared with differential privacy.

Recall that the condition of differential privacy is that the ratio of densities due to randomized algorithms on samples differing by one point, be bounded (or tends towards a bound super-polynomially fast). The effect of this condition is that an adversary who has $n - 1$ of the data points, cannot reliably test whether the n^{th} data point was a particular datum, since the power of his test would be bounded close to the level (Wasserman and Zhou [2010]). In essence if he wants to ensure probability 0.05 of not rejecting the correct data point, then he will only be able to reject about 5% of the candidates. In this regard he has failed to ascertain any information about the identity of the remaining data point.

In our definition we treat the data points as independent and identically distributed samples, which arise from an unknown distribution $\mathbb{P}$. We require that the original differential privacy condition holds with some probability (which we will take to be a high probability, for example $\gamma = 0.05$). This definition admits a number of interesting possibilities which are discussed below, but first we must be careful to describe the ramifications of our particular weakening of Differential Privacy. A reasonable way to consider our definition is that the usual DP criteria doesn't hold

uniformly over the space of data sets, but rather holds on "most" pairs of neighboring datasets, depending on γ. Although there are a number of ways to implement such a technique, a reasonable mental image to keep in mind is of data arising from a normal distribution. Then, a scheme would achieve RDP if it met the DP condition whenever none of the samples $x_1, \ldots, x_n, x'_n$ are sufficiently far into the tail of the distribution (i.e., if DP holds on a sphere of measure $1 - \gamma$). This example highlights what we envision to be the strongest criticism of the definition. Perhaps it is those extreme individuals (for example, the extremely wealthy or extremely sick) that require the most protection, and it is those individuals for which RDP expressly permits a lack of privacy. Therefore care must be taken when deciding whether it is appropriate to use this particular privacy definition. Finally we recall again that the size of the "tail" (and thus the number of exposed individuals in the above example) is controlled by γ, and therefore the above concern may not be an issue is γ is taken small enough.

Having presented what we foresee as the strongest criticism which could be leveled at our definition, we now turn to what we see are the most striking benefits of this relaxation over classical differential privacy. The first is that it may easily be applied in problems where the data lay in unbounded sets (for example real valued measurements). For example the sensitivity based approaches to DP typically require knowledge of the set in which the data lies, and when not available may lead to problems (although Dwork and Lei [2009] demonstrate some statistical applications in which unbounded sets may be handled by projecting the data to a suitable confidence set). In practise the absence of a bound on the data may lead the practitioner to e.g., use the most extreme values he observed to specify the radius of the set. However it is easy to see that this method will fail to achieve differential privacy.

The other advantage of RDP is that it gives some weak notion of privacy but while maintaining more utility in the released statistics. Evidently it will fail to hold up to a strong adversary such as the imagined adversary of section 1.2.1, however it may be suitable in certain conditions. What's more as shall be demonstrated below the noise magnitude required for RDP is often much smaller than that required for differential privacy.

5.2 RDP For Discrete Random Variables

Consider the special case that $x_i \in \{0, 1\}$. This situation also occurs when the function we release first applies some transform to the data so that the points become binary RVs, for example when we replace each x_i with the value of the indicator of some set, evaluated at that point. We consider the release of the mean of a set of such binary RVs. This statistic is given by $\hat{p}(x_1, \ldots, x_n) = n^{-1} \sum_i x_i$, and is an estimate of the quantity $p = \mathbb{P}(x_i = 1)$. Traditionally the Differentially Private approach has begun by noting that:

$$\sup_{x_1,\ldots,x_n,x_n'} |\hat{p}(x_1,\ldots,x_{n-1},x_n) - \hat{p}(x_1,\ldots,x_{n-1},x_n')| = n^{-1}$$

Which yields the approach of releasing $\tilde{p} = \hat{p} + \frac{1}{\alpha n}L$ where L is a standard double exponential random variable. When $\hat{p}$ is very close to 0 or 1, and when n is large, then we have evidence that p itself is close to 0 or 1. When this is the case, then it is unlikely that $\hat{p} \neq \hat{p}'$ (where we introduce this shorthand to mean the evaluations of the function $\hat{p}$ on the neighboring data samples. To see this, note that:

$$\mathbb{P}(\hat{p} \neq \hat{p}') = \mathbb{P}(x_n \neq x_n') = 2p(1-p)$$

The intent of this section is to demonstrate an approach that achieves RDP by releasing $\hat{p}$ directly whenever it is safe to do so (in the sense that the probability that a neighboring sample would result in a different estimate is below γ).

5.2.1 Basic Technique

Although there are several ways to obtain the requisite random differential privacy behavior, in this section we concentrate on one method which will be used twice. Therefore it is clearer to present the approach beforehand. We consider the $n+1$ samples $x_1,\ldots,x_{n+1}$ and the sufficient statistics $x_{(1)} < x_{(2)} < \cdots < x_{(n+1)}$, whatever ordering is used here is not important, it is only required to specify some unique ordering for all sets of $n+1$ points. These order statistics are the sufficient statistics since the distribution of the x_i is not assumed to be known, or to fall within any specific parametric family. Due to their sufficiency we may attain the correct $1-\gamma$ coverage level even without knowing the model which generates the data. We thus aim for γ-RDP conditioned on these sufficient statistics. First in the interest of compact equations, we define a new binary random variable which takes the value 1 whenever the DP criteria is met, and 0 otherwise. Denoting by Q_n the method which produces the private random output (the mechanism), we have

$$\eta(x_1,\ldots,x_{n+1},Q_n) = \mathbf{1}\left\{\forall Z \in \mathcal{Z} : Q_n(Z|x_1,\ldots,x_{n-1},x_n) \leq e^{\alpha}Q_n(Z|x_1,\ldots,x_{n-1},x_{n+1}) + \delta(n)\right\},$$

we suppress the dependence on $x_1,\ldots,x_{n+1}$, and Q_n. We aim for the following behavior

$$\mathbb{P}\left(\eta = 1 | x_{(1)},\ldots,x_{(n+1)}\right) \geq 1-\gamma,$$

then we have

$$\begin{aligned}\mathbb{P}(\eta = 1) &= \int_{\mathcal{X}^{n+1}} \mathbb{P}(\eta = 1 | x_{(1)}, \ldots, x_{(n+1)}) \; d\mathbb{P}(x_{(1)}, \ldots, x_{(n+1)}) \\ &= \int_{\mathcal{X}^{n+1}} 1 - \gamma \; d\mathbb{P}(x_{(1)}, \ldots, x_{(n+1)}) \\ &= 1 - \gamma,\end{aligned}$$

thus arriving at the correct overall $1 - \gamma$ coverage level. What is more, to attain RDP conditional on the sufficient statistics is conceptually simple and computationally tractable. First observe that the probability $P(\eta = 1 | x_{(1)}, \ldots, x_{(n+1)})$ is due to the rearrangement of the order statistics into the unordered sample $x_1, \ldots, x_{n+1}$. There are $(n+1)!$ ways to do the re-ordering and each one has equal probability. The reason for the latter is that the variables are assumed to be iid, and so the ordering does not impact the probability of a sample. Thus we take the $(n+1)!$ re-labelings, and ensure that Q_n achieves the differential privacy criteria (namely $\eta = 1$) on at least $\lceil (1-\gamma)(n+1)! \rceil$ of the re-labelings. This is all that is required to achieve the RDP conditional on the sufficient statistics. In the following sections, we apply this approach to both the estimation of a binomial proportion and to the estimation of a sparse histogram.

5.2.2 Binomial Proportion Estimation

First consider the $n+1$ bernoulli random variables: $x_1, \ldots, x_n, x_{n+1} \in \{0, 1\}^{n+1}$. For a sub-sample of size n we define the estimate of the binomial proportion as:

$$\hat{p}(x_1, \ldots, x_n) = n^{-1} \sum_{i=1}^{n} x_i$$

We consider the event that estimates based on the neighboring samples arrive at different values:

$$\begin{aligned}\xi(x_1, \ldots, x_{n+1}) &= \mathbf{1}\left\{\hat{p}(x_1, \ldots, x_{n-1}, x_n) \neq \hat{p}(x_1, \ldots, x_{n-1}, x_{n+1})\right\} \\ &= \mathbf{1}\left\{x_n \neq x_{n+1}\right\}\end{aligned}$$

The probability of this event conditioned on the sufficient statistics is simply the fraction of ways that the order statistics could be re-ordered which result in the last two samples taking different values. We define:

$$s(x_{(1)}, \ldots, x_{(n+1)}) = \sum_{i=1}^{n+1} x_{(i)}$$

Note that this function could not be evaluated in a real invocation of the procedure, since it depends on the neighboring sample which belongs to the supposed adversary. We have:

$$\mathbb{P}(\xi = 1|x_{(1)}, \ldots, x_{(n+1)}) = \frac{2s(n+1-s)(n-1)!}{(n+1)!} = \frac{2s}{n} - \frac{s}{n(n+1)} \leq \frac{2s}{n} \leq 2(\hat{p} + n^{-1})$$

This leads to the technique:

$$\tilde{p} = \begin{cases} \hat{p} & 2(\hat{p} + n^{-1}) \leq \gamma \\ \hat{p} + \frac{1}{n\alpha} L & \text{o/w} \end{cases} \tag{5.1}$$

The analysis begins by noting that the case in which the classical DP criteria fails to hold is either when both estimators are released exactly, and they differ, or when one estimator is released exactly whereas the other is released with the Laplace noise added to it. This is equivalent to the condition that the estimates differ and that at least one is released exactly (since if they are equal then it is impossible that only one is released exactly).

Proposition 5.2.1. *The technique of (5.1) achieves* $(\alpha, 0, \gamma)$*-RDP.*

Proof. Define:

$$m = n \min\{\hat{p}(x_1, \ldots, x_{n-1}, x_n), \hat{p}(x_1, \ldots, x_{n-1}, x_{n+1})\}$$

The condition that $\xi = 1$ leads to $s = m + 1$. The condition that at least one of the estimates be released exactly leads to $2(m+1) \leq n\gamma$. Thus we have that $\eta = 0$ only when $2s \leq n\gamma$.

$$\begin{aligned} \mathbb{P}\left(\eta = 0|x_{(1)}, \ldots, x_{(n+1)}\right) &= \mathbb{P}\left(\xi = 1, 2s \leq n\gamma|x_{(1)}, \ldots, x_{(n+1)}\right) \\ &= \mathbb{P}\left(\xi = 1|2s \leq n\gamma\right) \mathbb{P}\left(2s \leq n\gamma|x_{(1)}, \ldots, x_{(n+1)}\right) \\ &= \mathbb{P}\left(\xi = 1|2s \leq n\gamma\right) \mathbf{1}\left\{2s(x_{(1)}, \ldots, x_{(n+1)}) \leq n\gamma\right\} \\ &\leq \frac{n\gamma}{n} \\ &= \gamma \end{aligned}$$

□

5.2.3 Sparse Histograms

Computation of a histogram is via a function $\theta : \mathcal{D} \to \mathbb{N}^d$, with coordinates specified by

$$\theta_j(D) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{d_i \in B_j\},$$

where the B_j are mutually disjoint sets of the input space.

Proposition 5.2.2. *The release of $\tilde{\theta}_D = \theta_D + X$ achieves $(\alpha, 0, \gamma)$-RDP, where X is the random vector with coordinates given by*

$$x_j = \begin{cases} 0 & \theta_j(D) = 0 \text{ and } 2d \leq \gamma n \\ \frac{2}{n\alpha} L_j & o/w \end{cases}, \tag{5.2}$$

where the L_j are iid standard Laplace random variables.

In demonstrating RDP, we take the sample $x_1, \ldots, x_n, x_{n+1}$ and denote: $S = S(x_1, \ldots, x_n)$ and $S' = S(x_1, \ldots, x_{n-1}, x_{n+1})$. We consider the output distribution of our method when applied to each of the neighboring samples. The event that the ratio of densities fail to meet the requisite bound is a subset of the event where either $x_{n+1} \in S$ or $x_n \in S'$, and when $2k \leq \gamma n$. In the complement of this event then the partitions are the same, and the differing samples both fall within the block which receives the Laplace noise, so the DP condition is achieved. In demonstrating the RDP, we simply bound the probability of the aforementioned event, conditional on the order statistics.

The basic idea is to partition the space into two blocks and to release a noise-free histogram in one block, and use the classical differentially private histogram in the other. The partition will depend on the data itself. We consider $x \in \{0, \ldots, k\}$ as in a k-way histogram.
We consider the unadulterated histogram values:

$$\hat{p}_j = n^{-1} \sum_{i=1}^{n} 1\{x_i = j\}$$

For a sample $x_1, \ldots, x_n$, we denote:

$$U = U(x_1, \ldots, x_n) = \left\{ j : \hat{p}_j + \frac{1}{n} \leq \frac{\gamma}{2k} \right\}$$

Then we consider the release mechanism:

$$\tilde{p}_j = \begin{cases} \hat{p}_j & j \in U \\ \hat{p}_j + \frac{2}{n\alpha} L & \text{o/w} \end{cases} \tag{5.3}$$

In demonstrating RDP, we take the sample $x_1, \ldots, x_n, x_{n+1}$ and denote:

$$U = U(x_1, \ldots, x_n), U' = U(x_1, \ldots, x_{n-1}, x_{n+1})$$

The event that the DP criteria fails to hold is a subset of the event where either $x_{n+1} \in U$ or $x_n \in U'$. In the complement of this event then the partitions are the same, and the differing samples both fall within the block which receives the Laplace noise, so the DP condition is achieved. We consider the probability of the aforementioned event, in which DP fails:

$$\mathbb{P}\left(\eta = 0 | x_{(1)}, \ldots, x_{(n+1)}\right) \leq \mathbb{P}(x_n \in U' \text{ or } x_{n+1} \in U | x_{(1)}, \ldots, x_{(n+1)})$$

Proposition 5.2.3. *The technique of (5.3) satisfies the* (α, γ)*-RDP.*

Proof. Take:

$$c_j = \sum_{i=1}^{n+1} 1\{x_i = j\}$$

$$U^\star(x_1, \ldots, x_n, x_{n+1}) = \left\{ j : c_j \leq \frac{\gamma n}{2k} \right\}$$

We have that irrespective of the ordering of the points, $U, U' \subseteq U^\star$. We see that $j \in U^{\star C}$ implies $j \in U^C$ since:

$$\hat{p}_j + \frac{1}{n} \geq \frac{c_j}{n} > \frac{\gamma}{2k}$$

Thus $U^{\star C} \subseteq U^C$ and so we have $U \subseteq U^\star$. A symmetrical argument holds for $\hat{p}'_j$, which are the histogram values when x_{n+1} is used in place of x_n. We thus have:

$$\mathbb{P}(x_n \in U' \text{ or } x_{n+1} \in U | x_{(1)}, \ldots, x_{(n+1)}) \leq \mathbb{P}(x_n \in U^\star \text{ or } x_{n+1} \in U^\star | x_{(1)}, \ldots, x_{(n+1)})$$

The latter probability is just the fraction of ways in which the order statistics may be rearranged so that x_n, x_{n+1} have the requisite property. Due to the construction of $U^\star$, we have the property:

$$u(x_{(1)}, \ldots, x_{(n+1)}) = \sum_{i=1}^{n+1} \mathbf{1}\{x_{(i)} \in U^\star\} \leq \frac{\gamma n}{2}$$

Therefore the number of rearrangements having at least one of $x_n, x_{n+1} \in U^\star$ is

$$\begin{aligned}
\mathbb{P}(x_n \in U^\star \text{ or } x_{n+1} \in U^\star | x_{(1)}, \ldots, x_{(n+1)}) &= 1 - \frac{2}{(n+1)!}\binom{n+1-u}{2}(n-1)! \\
&= 1 - \frac{2}{(n+1)n}\binom{n+1-u}{2} \\
&= \frac{2un + u - u^2}{n(n+1)} \\
&\leq \frac{2u}{n+1} + \frac{u}{n(n+1)} \\
&\leq \frac{n\gamma}{n+1} + \frac{\gamma}{2(n+1)} \\
&< \frac{(n+1)\gamma}{n+1} \\
&= \gamma
\end{aligned}$$

Thus in summary we have:

$$\begin{aligned}
\mathbb{P}\left(\eta = 0 | x_{(1)}, \ldots, x_{(n+1)}\right) &\leq \mathbb{P}(x_n \in U' \text{ or } x_{n+1} \in U | x_{(1)}, \ldots, x_{(n+1)}) \\
&\leq \mathbb{P}(x_n \in U^\star \text{ or } x_{n+1} \in U^\star | x_{(1)}, \ldots, x_{(n+1)}) \\
&< \gamma,
\end{aligned}$$

which is the desired result. □

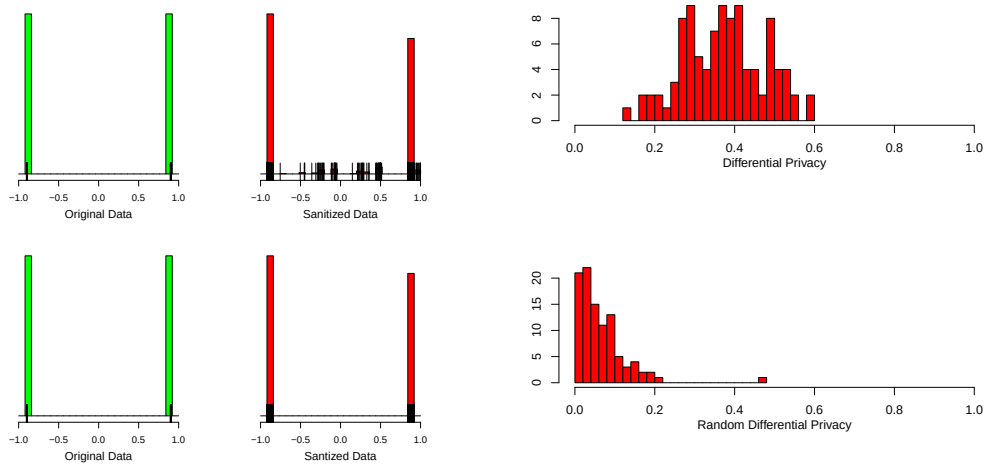
We also find the above technique to have good accuracy when d is large relative to n.

Proposition 5.2.4. *Suppose that* $2d \leq \gamma n$. *Let* $\theta_D = (\theta_1, \ldots, \theta_r, 0, \ldots, 0)$ *for some* $1 \leq r < k$. *Then* $||\theta_D - \tilde{\theta}_D||_1 = O_P(r/\alpha n)$.

We thus have a technique for which the risk is uniformly bounded above by $O(k/\alpha n)$ as with the DP technique, and which also enjoys the coordinate-wise upper bound on the risk. However in this regime, the risk is no longer uniformly lower bounded with a rate linear in k, since the upper bound is linear in r in the case of sparse vectors.

5.3 RDP via Sensitivity Analysis

We next demonstrate that RDP allows schemes for release of other kinds of statistics (besides histograms). We give the analog of the construction of proposition 1.3.1 for RDP. We consider the family of distributions with densities given by



(a) Original and synthetic data for DP (top) and RDP (bottom)

(b) Empirical error distribution for DP (top) and RDP (bottom)

Figure 5.1: An experiment demonstrating the improvement obtained when using RDP, when the task is the release of a very sparse histogram.

$$dP_D(z) \propto \exp\left\{\frac{-\alpha\,|z - g(D)|}{2s(D)}\right\}. \tag{5.4}$$

It is well known that when s is the constant function which gives an upper bound of the global sensitivity Dwork *et al.* [2006b] of g, this method enjoys the α-DP. As we allow s to depend on the data we may make use of the local sensitivity framework of Nissim *et al.* [2007]. Thus we require s to be "smooth" in the sense of Dwork *et al.* [2006b] with high probability, and also to be an upper bound on the sensitivity of g with high probability as g is given random inputs. Writing $s(D)$ as $s(x_1, \ldots, x_n)$ it follows from Nissim *et al.* [2007] that whenever

$$s(x_1, \ldots, x_n) \leq e^{\beta} s(x_1, \ldots, x'_n) \tag{5.5}$$

and

$$\sup_{x'_n} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x'_n)| \leq s(x_1, \ldots, x_n), \tag{5.6}$$

then the method (5.4) fulfils the (α, δ)-differential privacy (where $\delta = \delta(\beta)$ is given below). In order to get random differential privacy we allow both of these conditions to fail with some bounded probability.

We consider a special subset of functions for which the sensitivity depends only on the points that differ between the two samples. Namely we consider those functions g for which there exist

some suitable h which obeys

$$\sup_{D \sim D'} \left| g(D) - g(D') \right| = n^{-1} \sup_{x,x'} h(x, x').$$

Examples of functions satisfying this property are e.g., statistical point estimators Smith [2008] and regularized logistic regression estimates Chaudhuri *et al.* [2011]. In particular in these cases it is assumed that $\mathcal{D}$ is some compact subset of $\mathbb{R}^n$ and then e.g., $\sup_{x,x'} h(x, x') = \|x - x'\|_2$ gives the diameter of this set. We consider the empirical process based on h and the data sample D given by

$$F(D, t) = \frac{2}{n} \sum_{i=1}^{n/2} \mathbf{1} \left\{ h(x_i, x_{i+n/2}) \leq t \right\}$$

this is exactly an empirical CDF for the distribution of the random variable $h(x, x')$, based on $n/2$ independent samples. We may anticipate that sample quantiles of this empirical CDF will be close to the quantiles from the true CDF, which we denote by $H(t) = P(h \leq t)$. This is made precise by the DKW inequality (see e.g., Massart [1990]), which in this case yields

$$\mathbb{P}\left(\sup_t |H(t) - F(D, t)| \geq \epsilon \right) \leq 2e^{-n\epsilon^2}. \tag{5.7}$$

The above concentration result gives a means to establish that $s(D)$ may be a sensitivity bound which holds with high probability. For the bound to be smooth with high probability requires that sample quantiles of the above process will be close with high probability. Work due to Kiefer and others (see e.g, Kiefer [1967]; Arcones [1996]) demonstrates that this is the case.

Evidently a quantile of h, say h_θ leads to a quantile of the sensitivity of g since

$$\mathbb{P}\left(\left| g(D) - g(D') \right| > \tau \right) = \mathbb{P}\left(h(x, x') > n\tau \right),$$

thus using an appropriately scaled quantile of h as our function s will satisfy the probabilistic relaxation of (5.6). When using a sample quantile there are in essence two ways that this property will fail. The first is when the sample quantile is smaller than the true quantile, and the second is when the data leads to neighboring g with difference outside of the quantile. The former probability is obtained via DKW and the latter by selecting the appropriate quantile of the true distribution. Namely we may take $\hat{h}_\theta$ to be the θ sample quantile of h, and likewise h_θ to be the unknown true quantile, then

$$\mathbb{P}\left(h > \hat{h}_\theta \right) \leq \mathbb{P}\left(h_\theta > \hat{h}_\theta \right) + \mathbb{P}\left(h > h_\theta \right)$$

Now we turn attention to the condition (5.5), that the functions $s(D) = n\hat{h}_\theta$ based on samples

differing by one sample be close. We examine the Bahadur-Kiefer representation for quantiles of an empirical process

$$\hat{h}_\theta - h_\theta = \frac{F(D, h_\theta) - H(h_\theta)}{\partial H(h_\theta)} + O_p(n^{-1/2})$$

We desire the ratio

$$\frac{\hat{h}_\theta(D)}{\hat{h}_\theta(D')} \leq e^\beta,$$

with high probability. We also would like $\delta(\beta)$ to be negligible in n (a typical requirement when working with the approximate differential privacy). From Nissim *et al.* [2007] we have

$$\delta(\beta) = \exp\left\{-\frac{\alpha}{2\beta}\right\},$$

thus for example if we take $\beta = n^{-1/2}$ then we have that $\delta = e^{-\frac{\alpha}{2}\sqrt{n}}$ which is negligible (since we regard α as a constant which does not depend on n). We write

$$\frac{\hat{h}_\theta(D)}{\hat{h}_\theta(D')} \leq 1 + \frac{|\hat{h}_\theta(D) - \hat{h}_\theta(D')|}{\hat{h}_\theta(D')},$$

then

$$\frac{|\hat{h}_\theta(D) - \hat{h}_\theta(D')|}{\hat{h}_\theta(D')} = \frac{\frac{|F(D,h_\theta) - F(D',h_\theta)|}{\partial H(h_\theta)} + O_p(n^{-1/2})}{h_p + \frac{F(D',h_\theta) - H(h_\theta)}{\partial H(h_\theta)} + O_p(n^{-1/2})}$$

the DKW inequality and the triangle inequality lead to

$$\mathbb{P}(\sup_t |F(D,t) - F(D',t)| \geq \epsilon) \leq 4e^{-2\epsilon^2 n},$$

therefore

$$|F(D, h_\theta) - F(D', h_\theta)| = O_p(n^{-1/2}).$$

What's more, DKW directly gives

$$F(D', h_\theta) - H(h_\theta) = O_p(n^{-1/2}),$$

therefore since $\partial H(h_\theta)$ is constant with respect to n we have

$$\frac{|\hat{h}_\theta(D) - \hat{h}_\theta(D')|}{\hat{h}_\theta(D')} = \frac{O_p(n^{-1/2})}{h_\theta + O_p(n^{-1/2})}.$$

Thus we may achieve the approximate differential privacy with appropriately negligible δ in the limit as n grows to infinity.

5.4 Variants of Random Differential Privacy

The above variant of differential privacy arose from blindly wrapping the usual differential privacy criterion in a probabilistic statement. As such it may not be a particularly useful definition to the practitioners. Thus we may explore some closely related but different ideas in order to find weakenings under which meaningful statistical inferences are possible but which give a more appropriate privacy guarantee.

One obvious extension arises by considering the following scenario. The data owner has n data points which have arisen from some unknown distribution (e.g., due to some survey), the adversary has m points from the same distribution (e.g., from a similar survey done historically by some agency), and the overlap between the two sets is known to be at least k. This situation seems prima facie reasonable in the context of governmental surveys and census data. What's more due to its similarity to RDP it will admit similar analyses, and allow the construction of methods via a hypothesis testing construction (in which the size corresponds to the probability of a privacy breach). The above method also should be noted for its similarity to the method outlined in an upcoming paper due to Shlomo et. al.

What's more there are other ways to involve the probability of the underlying sample in the privacy statement. One possibility is "predictive differential privacy."

Definition 5.4.1 (Predictive Differential Privacy)**.** We say that P satisfies (α, γ)-*predictive differential privacy* if

$$P_D(Z \in A) \leq e^{\alpha} P_{D'} Q(Z \in A) \tag{5.8}$$

for all A and $D \sim D'$ where $D' \in \hat{S}^n(D)$ where the latter is a prediction interval for x_i based on dataset D, that satisfies $\mathbb{P}(x \in \hat{S}(D)) \geq 1 - \gamma$ (in which $\mathbb{P}$ is once again the $n+1$-fold product measure).

This definition is similar in spirit to random differential privacy, in that it admits a failure of the differential privacy condition whenever the "neighboring data" is unlikely under the joint distribution. In fact this is a strengthening of RDP in the sense that techniques which fulfill this definition will also be randomly differentially private.

5.5 Summary

In this chapter we gave a new privacy definition which is a strict relaxation of differential privacy. We showed how it can be applied to discrete problems like histogram estimation, and how it can be applied to continuous problems via a kind of "empirical" sensitivity analysis. We also demonstrated that accuracy is improved over a naive differentially private technique (the Laplace mechanism) for

histogram estimation. We also demonstrated that it can be useful in the case when the input space is not compact.

In our opinion the contributions of this section are obviated in light of the histogram release techniques shown in Chapter 2. The current chapter is more of a historic footnote in our search for improved accuracy out of differentially private techniques. After the development of these techniques we also found that it is possible to relax the requirement of compact support via the techniques in Dwork and Lei [2009], in some cases without sacrificing the convergence rate of the estimators in question. What remains to be seen what class of estimators this is true for, besides the ones shown in their paper.

Chapter 6

Summary

In this thesis we dealt with a few new methods for useful statistical inferences which provide the guarantee of differential privacy. A common thread was that we sought to preserve the statistical performance of the various techniques, while maintaining privacy. Some of our constructions left open avenues of work which may be pursued in the future.

First we demonstrated some techniques for discrete density estimators such as histograms and contingency tables. There we showed that in the case of high dimensional data, we may take advantage of sparsity in order to have some control over the breakdown in utility necessitated by the privacy guarantee. Although we demonstrated that the technique is close to optimal in its error rate, we hinted that the true optimal rate may be obtained by using a "False Discovery Rate" method rather than the truncation we used there. This will be interesting to explore in follow-on work.

Second we demonstrated how various function values estimators may be made private. We dealt with the entire class of functions which live in Reproducing Kernel Hilbert Spaces and showed a conceptually clean way to preserve the privacy of these. We gave essentially the natural extension of the finite dimensional differential privacy methods into an infinite dimensional space with certain structure. However the technique only gave a lesser privacy guarantee – that of the approximate differential privacy. It will be interesting to see whether there exists any similar method which is as easily applicable and which gives the proper privacy guarantee.

We then demonstrated for a specific task (kernel density estimation) how differential privacy may be obtained without sacrificing the statistical convergence rate of the estimator. Here the technique was more ad-hoc than the above, and it is not clear whether similar techniques will apply to different problems like this one. However for estimators which are the pointwise means of functions which lay sufficiently close to the span of the fourier basis, it seems likely that this technique can be made to apply.

Finally we gave a modification to the privacy criteria that allows for much greater utility but at the expense of a weaker guarantee. In light of recent work the actual methods we gave may not be so useful, but nevertheless we initiated an exploration into weaker but meaningful forms of privacy which we hope to see continued in the future.

Bibliography

Robert J. Adler and Jonathan E. Taylor. *Random Fields and Geometry (Springer Monographs in Mathematics)*. Springer, 1 edition, June 2007.

Robert J. Adler. *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*, volume 12 of *Lecture Notes–Monograph series*. Institute of Mathematical Statistics, 1990.

R. Agrawal and R. Srikant. Privacy-preserving data mining. *ACM SIGMOD Record*, 29(2):439–450, 2000.

Miguel A. Arcones. The Bahadur-Kiefer representation for u-quantiles. *The Annals of Statistics*, 24(3):1400–1422, 1996.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68, 1950.

Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '07, pages 273–282, New York, NY, USA, 2007. ACM.

Amos Beimel, Kobbi Nissim, and Eran Omri. Distributed private data analysis: On simultaneously solving how and what. *CoRR*, abs/1103.2626, 2011.

A. Bertinet and Thomas C. Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 128–138, 2005.

Olivier Bousquet and Andre Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Anne Sophie Charest. *Creation and Analysis of Differentially-Private Synthetic Datasets.* PhD thesis, Department of Statistics, Carnegie Mellon University, 2012.

K. Chaudhuri, C. Monteleoni, and A.D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

Richard Conway and David Strip. Selective partial access to a database. In *Proceedings of the 1976 annual conference*, ACM '76, pages 85–89, New York, NY, USA, 1976. ACM.

Graham Cormode, Cecilia Magdalena Procopiuc, Divesh Srivastava, and Thanh T. L. Tran. Differentially private publication of sparse data. *CoRR*, abs/1103.0825, 2011.

Anindya De. Lower bounds in differential privacy. *CoRR*, abs/1107.2183, 2011.

Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '03, pages 202–210, New York, NY, USA, 2003. ACM.

G.T. Duncan and D. Lambert. Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81:10–28, 1986.

G.T. Duncan and D. Lambert. The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, 7:207–217, 1989.

George Duncan and Robert Pearson. Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, 6(3):219–232, August 1991.

G.T. Duncan and L. Stokes. Data masking for disclosure limitation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1:83–92, 2009.

G.T. Duncan, S. Keller-McNulty, and L. Stokes. Disclosure risk vs. data utility: The R-U confidentiality map. Technical Report 121, National Institute of Statistical Sciences, December 2001.

George T. Duncan, Mark Elliot, and Juan-José Salazar-González. *Statistical confidentiality. Principles and practice.* Statistics for Social and Behavioral Sciences. Springer. New York, NY, 2011.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, STOC '09, pages 371–380, New York, NY, USA, 2009. ACM.

Cynthia Dwork and Adam Smith. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2):135–154, 2010.

C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. *EUROCRYPT*, pages 486–503, 2006.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. *Proceedings of the 3rd Theory of Cryptography Conference*, pages 265–284, 2006.

Cynthia Dwork. Differential privacy. *33rd International Colloquium on Automata, Languages and Programming*, pages 1–12, 2006.

Cynthia Dwork. Differential privacy: A survey of results. In *TAMC*, pages 1–19, 2008.

S.E. Fienberg, W.J. Fulp, A.B. Slavkovic, and T.A. Wrobel. "Secure" log-linear and logistic regression analysis of distributed databases. *Privacy in Statistical Databases: CENEX-SDC Project International Conference, PSD 2006*, pages 277–290, 2006.

S. Fienberg, Y. Nardi, A. Rinaldo, L. Wasserman, and S. Zhou. Quantization and the privacy-accuracy tradeoff. *Unpublished Manuscript*, 2008.

S.E. Fienberg, A.B. Slavkovic, and Y. Nardi. Valid statistical analysis for logistic regression with multiple sources. In P. Kantor and M. Lesk, editors, *Proc. Workshop on Interdisciplinary Studies in Information Privacy and Security—ISIPS 2008*. Springer-Verlag, New York, 2009.

Stephen E. Fienberg, Alessandro Rinaldo, and Xiaolin Yang. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. *Privacy in Statistical Databases*, pages 197–199, 2010.

Stephen Fienberg, Rob Hall, and Yuval Nardi. Achieving both valid and secure logistic regression analysis on aggregated data from different private sources. *Journal of Privacy and Confidentiality*, 4(1):189–220, 2012.

W.A. Fuller. Masking procedures for microdata. *Journal of Official Statistics*, 9:383–406, 1993.

O. Goldreich. *Modern Cryptography, Probabilistic Proofs, and Pseudorandomness*. Springer-Verlag, New York, 1998.

O. Goldreich. *Foundations of Cryptography: Volume 2 Basic Applications*. Cambridge University Press, 2004.

Mangesh Gupte and Mukund Sundararajan. Universally optimal privacy mechanisms for minimax agents. In *Proc. ACM SIGMOD*, pages 135–146, Indianapolis, Indiana, 2010.

Rob Hall, Stephen Fienberg, and Yuval Nardi. Secure multiparty linear regression based on homomorphic encryption. *Journal of Official Statistics*, 27(4):669–691, 2011.

Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. *STOC '10 Proceedings of the 42nd ACM symposium on Theory of computing*, pages 705–714, 2010.

Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. *Technical Report*, 2010.

Jiunn T. Hwang. Multiplicative errors-in-variables models with applications to recent data released by the u.s. department of energy. *Journal of the American Statistical Association*, 81(395):pp. 680–688, 1986.

Iain M. Johnstone. Gaussian estimation: Sequence and multiresolution models, 2011. pre-print, available online.

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

J. Kiefer. On Bahadur's representation of sample quantiles. *The Annals of Mathematical Statistics*, 38(5):1323–1342, 1967.

Daniel Kifer and Johannes Gehrke. l-diversity: Privacy beyond k-anonymity. In *In ICDE*, page 24, 2006.

Guy Lebanon, Monica Scannapieco, Mohamed R. Fouad, and Elisa Bertino. Beyond k-anonymity: A decision theoretic framework for assessing privacy risk. In *Privacy in Statistical Databases*, pages 217–232. Springer, 2006.

Ninghui Li and Tiancheng Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *In Proc. of IEEE 23rd Intl Conf. on Data Engineering (ICDE07*, 2007.

Chong K. Liew, Uinam J. Choi, Chung, and J. Liew. A data distortion by probability distribution. *ACM TRANSACTIONS ON DATABASE SYSTEMS*, 10:395–411, 1985.

Y. Lindell and B. Pinkas. Privacy preserving data mining. *Journal of Cryptology*, 15(3):177–206, 2002.

Yehuda Lindell and Benny Pinkas. Secure multiparty computation for privacy-preserving data mining. *Journal of Privacy and Confidentiality*, 1(1):59–98, 2009.

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

P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3), 1990.

Frank McSherry and Ilya Mironov. Differentially private recommender systems: Building privacy into the net. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 627–636, New York, NY, USA, 2009. ACM.

F. McSherry and K. Talwar. Mechanism Design via Differential Privacy. *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, pages 94–103, 2007.

K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. *Proceedings of the 39th annual ACM annual ACM symposium on Theory of computing*, pages 75–84, 2007.

Christine M. O'Keefe and Norm M. Good. A remote analysis server - What does regression output look like? In *Privacy in Statistical Databases*, pages 270–283, 2008.

E. Parzen. An Approach to Time Series Analysis. *The Annals of Mathematical Statistics*, 32(4):951–989, 1961.

E. Parzen. Probability density functionals and reproducing kernel hilbert spaces. *Proceedings of the Symposium on Time Series Analysis*, 196:155–169, 1963.

Steven P. Reiss. Practical data-swapping: the first steps. *ACM Trans. Database Syst.*, 9(1):20–37, March 1984.

J. P. Reiter. Using cart to generate partially synthetic, public use microdata. *Journal of Official Statistics*, pages 441–462, 2003.

J.P. Reiter. Model diagnostics for remote-access regression servers. *Statistics and Computing*, 13:371–380, 2003.

Jerome P. Reiter. Releasing multiply imputed, synthetic public-use microdata: An illustration and empirical. *Study, Journal of the Royal Statistical Society, A*, 168:185–205, 2004.

Jerome P. Reiter. Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100:1103–1112, December 2005.

Benjamin I. P. Rubinstein, Peter L. Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *Journal of Privacy and Confidentiality*, 2010.

Mark Rudelson, Adam Smith, and Jonathan Ullman. The price of privately releasing contingency tables and the spectra of random matrices with correlated rows. *STOC '10 Proceedings of the 42nd ACM symposium on Theory of computing*, pages 775–784, 2010.

Raman Sanyal, Frank Sottile, and Bernd Sturmfels. Orbitopes. *pre-print*, 2009. arXiv:0911.5436.

I. Schoenberg. An isoperimetric inequality for closed curves convex in even-dimensional euclidean spaces. *Acta Mathematica*, 91:143–164, 1954. 10.1007/BF02393429.

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

Adam Smith. Asymptotically optimal and private statistical estimation. In *Proceedings of the 8th International Conference on Cryptology and Network Security*, CANS '09, pages 53–57, 2009.

Ross Sparks, Chris Carter, John B. Donnelly, Christine M. O'Keefe, Jodie Duncan, Tim Keighley, and Damien McAullay. Remote access methods for exploratory data analysis and statistical modelling: Privacy-preserving analytics. *Computer Methods and Programs in Biomedicine*, 91(3):208–222, 2008.

Latanya Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002.

Daniel Ting, Stephen E. Fienberg, and Mario Trottini. Random orthogonal matrix masking methodology for microdata release. *Int. J. Inf. Comput. Secur.*, 2(1):86–105, January 2008.

M. Trottini, S.E. Fienberg, U.E. Makov, and M.M. Meyer. Additive noise and multiplicative bias as disclosure limitation techniques for continuous microdata: a simulation study. *Journal of Computational Methods in Sciences and Engineering*, 4:5–16, 2004.

J. Vaidya, Y. Zhu, and C. Clifton. *Privacy Preserving Data Mining (Advances in Information Security)*. Springer-Verlag, New York, 2005.

Santosh Vempala. Geometric random walks: a survey. *MSRI Volume on Combinatorial and Computational Geometry*, 52:577–616, 2005.

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

L. Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

Xiaolin Yang, Stephen E. Fienberg, and Alessandro Rinaldo. Differential privacy for protecting multi-dimensional contingency table data: Extensions and applications. *Journal of Privacy and Confidentiality*, 4(1):101–125, 2012.

Bin Yu. Assouad, Fano, and Le Cam. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

MACHINE LEARNING
DEPARTMENT

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA15213

Carnegie Mellon