Differential Privacy, Property Testing, and Perturbations

by

Audra McMillan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Mathematics)
in The University of Michigan
2018

Doctoral Committee:

Professor Anna Gilbert, Chair
Professor Selim Esedoglu
Professor John Schotland
Associate Professor Ambuj Tewari

Audra McMillan

amcm@umich.edu

ORCID iD: 0000-0003-4231-6110

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Anna Gilbert. Anna managed to strike the balance between support and freedom that is the goal of many advisors. I always felt that she believed in my ideas and I am a better, more confident researcher because of her.

I was fortunate to have a large number of pseudo-advisors during my graduate career. Martin Strauss, Adam Smith, and Jacob Abernethy deserve special mention. Martin guided me through the early stages of my graduate career and helped me make the transition into more applied research. Martin was the first to suggest that I work on privacy. He taught me that interesting problems and socially-conscious research are not mutually exclusive. Adam Smith hosted me during what became my favourite summer of graduate school. I published my first paper with Adam, which he tirelessly guided towards a publishable version. He has continued to be my guide in the differential privacy community, introducing me to the right people and the right problems. Jacob Abernethy taught me much of what I know about machine learning. He also taught me the importance of being flexible in my research and that research is as much as about exploration as it is about solving a pre-specified problem.

I would like to thank my collaborators, Ambuj Tewari and Chansoo Lee, and my committee members, John Schotland and Selim Esedoglu. I owe particular thanks to Chansoo for organising the Michigan privacy reading group. In its brief tenure it has become a bright spot in my week. The support of Karen Smith was crucial to my happiness as an early graduate student. Thank you to my fellow sixth years for being by my side throughout the process. I'm so proud of all of you and so excited to see what you do next.

Michael Newman has been my closest ally in everything for the last 6 years, so much so that it is impossible to know what to thank him for. I'll just say that Ann Arbor felt like home after I met him. Thank you to my sisters, Olivia and Laura, who have always wanted me to succeed and whom I have always wanted to make proud.

Finally, a big thank you to my greatest supporters, my parents. They raised me to value hard work and perseverance and instilled in me a sense of self worth that transcends my accomplishments and defeats. I am able to pursue what I love because they taught me not to fear failure. They have celebrated my every achievement and will be the first to have the champagne ready when my degree is confirmed.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Controlling the dissemination of information about ourselves has become a minefield in the modern age. We release data about ourselves every day and don't always fully understand what information is contained in this data. It is often the case that the combination of seemingly innocuous pieces of data can be combined to reveal more sensitive information about ourselves than we intended. Differential privacy has developed as a technique to prevent this type of privacy leakage. It borrows ideas from information theory to inject enough uncertainty into the data so that sensitive information is provably absent from the privatised data. Current research in differential privacy walks the fine line between *removing sensitive information* while allowing non-sensitive information to be released.

At its heart, this thesis is about the study of information. Many of the results can be formulated as asking a subset of the questions: does the data you have contain enough information to learn what you would like to learn? and how can I affect the data to ensure you can't discern *sensitive* information? We will often approach the former question from both directions: information theoretic lower bounds on recovery and algorithmic upper bounds.

We begin with an information theoretic lower bound for graphon estimation. This explores the fundamental limits of how much information about the underlying population is contained in a finite sample of data. We then move on to exploring the connection between information theoretic results and privacy in the context of linear inverse problems. We find that there is a discrepancy between how the inverse problems community and the privacy community view *good* recovery of information. Next, we explore black-box testing for privacy. We argue that the amount of information required to verify the privacy guarantee of an algorithm, without access to the internals of the algorithm, is lower bounded by the amount of information required to *break* the privacy guarantee. Finally, we explore a setting where imposing privacy is a help rather than a hindrance: online linear optimisation. We argue that private algorithms have the right kind of *stability* guarantee to ensure low regret for online linear optimisation.

Part I

Introduction

CHAPTER 1

Motivation and Background

Controlling the dissemination of information about ourselves has become a minefield in the modern age. We release data about ourselves every day and do not always fully understand what information is contained in this data. The privacy implications of some types of data release are obvious: medical records, browser histories and education transcripts obviously contain sensitive information. It is, however, often the case that seemingly innocuous data contains more sensitive information than we initially recognise. For example, Molina-Markham et al. showed that smart meter data from a home actually contains not only the amount of energy used, but also fine grain information about the activities of the residents. For a long time, we did not have a language to talk about this type of latent variable privacy release. As a result, sincere efforts to prevent privacy leakage failed, often resulting in blatant privacy violations. There is a long list of so-called *anonymised* datasets that were revealed to still contain personally identifiable information [Barbaro and Jr, 2006, Information Systems Audit and Control Association (ISACA), 2011, Shrivastva et al., 2014]. The field of privacy preserving data analysis has developed to ease privacy concerns while supporting meaningful data analysis.

Privacy preserving data analytics is predated by the field of information theory, which studies a similar problem from a different perspective. Information theory is the study of the fundamental limitations of extracting information from data in the presence of uncertainty. A key example is attempting to reconstruct information after it has passed through a noisy channel. *Information theoretic* limitations on reconstruction state not only that known reconstruction algorithms fail but that *any* reconstruction algorithm will fail [E. Shannon, 1948]. This idea epitomises the shift in thinking of modern privacy-preserving data analytics. We want to inject enough uncertainty into the data that sensitive information is provably absent from the privatised data.

A breakthrough in this field was the definition of *differential privacy*, which put the concept of privacy preservation in data analysis on a mathematically rigorous foundation. Differential privacy and its ilk allow us to argue rigorously about how much information about an individual can be learned from the output of a computation. A major contribution of the pioneering paper in this field was the stance that *sensitive information* is information that could not have been obtained if the individual was not in the dataset [Dwork et al., 2006]. As an example, suppose a study reveals that most people love Michigan. Even if I were not in the dataset, you have learnt something about me: I probably love Michigan. Differential privacy takes the stance that this is *not* a violation of my privacy because the impact on me is the same, whether or not my information was used in the study. This distinction between learning something about me and violating my privacy is what allows differentially private data analyses to obtain meaningful results. The field of differential privacy walks the fine line between *removing sensitive information* while allowing non-sensitive information to be released. It has become the gold standard for privacy-preserving data analysis and fostered a large body of work (See [Dwork, 2008, Dwork and Roth, 2014a] for surveys).

At its heart, this thesis is about the study of information. Many of the results can be formulated as asking a subset of the following questions:

- Does the data contain enough information to learn what you would like to learn?
- How can I affect the data to ensure you cannot discern *sensitive* information?

We will often approach the former question from both directions: information theoretic lower bounds on recovery and algorithmic upper bounds.

We begin in Chapter II with an information theoretic argument of the type that will be prevalent throughout this thesis. We establish fundamental limitations on graphon estimation. We then move on to exploring the connection between information theoretic results and privacy in the context of linear inverse problems in Chapter III. In Chapter IV, we explore the problem of determining, without access to the internals of an algorithm, how private it is. We argue that the amount of information required to verify the privacy guarantee of an algorithm is lower bounded by the amount of information required to *break* the privacy guarantee. Finally, Chapter V is dedicated to studying how tools from the differential privacy literature can be used in the design and analysis of online linear optimisation algorithms.

1.1 Overview of Results

1.1.1 Lower Bound on Graphon Estimation

Networks and graphs arise as natural modelling tools in many areas of science. In many settings, particularly in social networks, networks display some type of community structure. In these settings, one can model the structure of the network by something called a *graphon* or, if there are only a finite number of communities, a *stochastic block* *model*. In a stochastic block model, one considers each node of the graph as belonging to one of k communities and the probability that two nodes are connected depends on the communities they belong to. Given an observed network, graphon estimation is the problem of recovering the graphon model from which the graph was drawn.

In Part II, based on joint work with Adam Smith [McMillan and Smith, 2017], we explore the fundamental limits of graphon estimation for block graphons. That is, given an n-node network that was generated from a k-block graphon, how accurately can you recover the graphon? Our lower bound improves upon previously known lower bounds Klopp et al. [2015] for sparse graphs (graphs where the average degree ρ is sublinear in n) and rules out non-trivial estimation in the very sparse regime.

1.1.2 Local Differential Privacy for Physical Sensor Data

The development of wireless technology has allowed an increasing amount of lightweight (thermal, light, motion, etc.) sensors to be deployed. In many systems, the sensor measurements are a *linear function* of the data we would like to keep private. In Part III, based on joint work with Anna Gilbert [Gilbert and McMillan, 2018], we explore how we can exploit the ill-posedness of some linear inverse problems when designing locally differentially private algorithms for releasing sensor measurements.

This work had two main contributions. The first was noticing a connection between ill-conditionedness and privacy. A linear problem $y = Ax$ is ill-conditioned if only a small amount of noise in the measurement vector y is needed to prohibit accurate recovery of x. To the best of my knowledge this connection had not previously been made. We found that if a problem is well-conditioned then we necessarily need to add a significant amount of noise to maintain privacy. The converse, however, is not generally true: it is possible to have an ill-conditioned matrix and still need to add a considerable amount of noise to maintain privacy. The proof of this relies on analysing the spectral properties of the two conditions.

After instantiating our formulation with the heat kernel, our second contribution was an improved upper bound on the Earth Mover distance (EMD) error of basis pursuit denoising with the heat kernel on the one-dimensional unit interval. Our work indicates that it is possible to produce differentially private sensor measurements that both keep the exact locations of the heat sources private and permit recovery of the *general vicinity* of the sources. It was somewhat surprising that basis pursuit denoising, which is only known to work for well-conditioned matrices, is also effective for the heat kernel in the EMD.

Diffusion on graphs is used to model the path of a random walker in a graph, as well as the spread of rumours, viruses or information in a social network. In the final section of Part III, we instantiate our framework with the graph diffusion operator and discuss the relationship between connectivity and privacy. A particularly interesting example is community graphs, where we would like the measurement data to reveal the community the rumour started in, but not the exact person. We discuss promising experimental results.

1.1.3 Property Testing for Differential Privacy

Recently differential privacy has gained traction outside of theoretical research as several companies (Google, Apple, Microsoft, Census, etc.) have announced deployment of large-scale differentially private mechanisms [Erlingsson et al., 2014, Apple, 2017, Adowd and Schmutte, 2017, Ding et al., 2017]. This use of DP, while exciting, might be construed as a marketing tool used to encourage privacy-aware consumers to release more of their sensitive data to the company. In addition, the software behind the deployment of DP is typically proprietary since it ostensibly provides commercial advantage. This raises the question: with limited access to the software, can we verify the privacy guarantees of purportedly DP algorithms?

In this early stage of commercial DP algorithms, approaches to transparency have been varied. For some algorithms, like Google's RAPPOR, a full description of the algorithm has been released [Erlingsson et al., 2014]. On the other hand, while Apple has released a white paper [Differential Privacy Team, 2017] and a patent [Thakurta et al., 2017], there are still many questions about their exact implementations. In Part IV, based on joint work with Anna Gilbert, we explore verifying the privacy of blackbox algorithms in two extreme settings; when we are given *no information* about the black-box (except the domain and range), and the *full information* setting where we have an untrusted full description of the algorithm $\mathcal{A}$.

A central theme of this work is that verifying the privacy guarantees that corporations (or any entity entrusted with private data) claim requires compromise by either the verifier or algorithm owner. If the verifier is satisfied with only a weak privacy guarantee (random approximate DP with δ and γ small but not extremely small), then she or he can verify this with no side information from the algorithm owner. If the company is willing to compromise by providing information about the algorithm up-front, then much stronger privacy guarantees can be verified. Given this level of transparency, one might be tempted to suggest that the company provide source code instead. While verifying privacy given source code is an important and active area, there are many scenarios where the source code itself is proprietary. We have already seen instances where companies have been willing to provide detailed descriptions of their algorithms. In the full information setting we obtain a sublinear algorithm for verifying random approximate differential privacy on discrete distributions.

1.1.4 Online Linear Optimisation through the lens of Differential Privacy

Online learning is a common machine learning task where data becomes available in a sequential order. In online supervised learning, there is a function $f : X \to Y$ to be learned and at each time step, the learner outputs an updated estimate, f_t to f and receives a new data point $(x_t, f(x_t))$. The loss suffered $\ell(f_t(x_t), f(x_t))$ is a measure of how far $f_t(x_t)$ is from the true label, $f(x_t)$. The learner's goal is to minimise the loss they suffer after T rounds. This is usually measured against the loss they would have suffered if they'd played the best in hindsight answer. That is, the learner seeks to minimise

$$\text{Regret}_T = \sum_{t=1}^{T} \ell(f_t(x_t), f(x_t)) - \min_{f^* \in \mathcal{X}} \sum_{t=1}^{T} \ell(f^*(x_t), f(x_t)).$$

We typically assume that the data sequence is chosen adversarially.

Algorithms that perform well in the adversarial online setting tend to be *stable* under small changes in the data [Ross and Bagnell, 2011]. This notion is very similar to the requirement that differentially private algorithms should not be too dependent on the data. In Part V, based on joint work with Chansoo Lee, Jacob Abernethy and Ambuj Tewari [Abernethy et al., 2018], we explore the use of tools from differential privacy in the design and analysis of online learning algorithms. This lead to a minimax optimal algorithm for k-sparse online PCA and a connection between differential privacy and differential consistency, another smoothness notion.

Chansoo Lee did much of the heavy lifting for the work in this Part V. The ideas in Section 12.2.2 belong to him. Some of the prose in this Part was written by Chansoo, Jacob or Ambuj, although the presentation has been altered from the preprint [Abernethy et al., 2018] for the purposes of this thesis.

1.2 Preliminaries

In this section we introduce some of the notation, definitions and basic techniques that will be used throughout this thesis. Unless specified otherwise $\mathcal{A}$ is a randomised algorithm whose domain is the set of databases. We use P_D to denote the distribution on outputs for the input D. For any integer n, $[n] = \{1, \cdots, n\}$. The $\ell_\infty, \ell_1, \ell_2$ norms on $\mathcal{R}^n$ will be denoted by $\|\cdot\|_\infty$, $\|\cdot\|_1$ and $\|\cdot\|_2$. Typically Ω will denote our data universe and $\mathcal{D}$ is a distribution on Ω.

We will deal with many measures of closeness between distributions. We collect these definitions here for ease of reference.

Definition 1.1. Let P and Q be two distributions.

- Total Variance (TV) distance $\|P - Q\|_{\text{TV}} = \sup_E |P(E) - Q(E)|$

- Max divergence $D_\infty(P,Q) = \sup_E \ln \frac{P(E)}{Q(E)}$.
- δ-approximate max divergence $D_\infty^\delta(P,Q) = \sup_{E \text{ s.t. } P(E)\geq\delta} \ln \frac{P(E)-\delta}{Q(E)}$.
- Rényi divergence of order β $D_\beta(P\|Q) = \frac{1}{\beta-1} \ln \mathbb{E}_{x\sim Q} \left(\frac{P(x)}{Q(x)}\right)^\beta$.
- Kullbeck-Leibler (KL) divergence $D_{KL}(P\|Q) = \int_R P(x) \ln \frac{P(x)}{Q(x)} dx$.

where the $\sup_E$ is the supremum over all events E in the outcome space.

1.2.1 Concentration Inequalities

In this section we review some concentration inequalities for random variables. The results in this section can be thought of as non-asymptotic analogues of the central limit theorem: the sum of many random variables tends to concentrate around it's mean. In fact, they concentrate so quickly that we get exponentially decreasing bounds on the probability that the sum of random variables deviates from its mean. These results are the reason we feel suspicious if a series of coin flips has too many heads or if a dice lands on 1 too often. The interested reader is referred to Vershynin [2019] and Sridharan [2018] for a more in-depth introduction.

Let $X_1, \cdots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$$

be the average of the random variables. Our first inequality, Hoeffding's inequality, is a special case of the more general Azuma-Hoeffding inequality. We will frequently use Hoeffding's inequality to argue that if a trial has a p probability of success than after a small number of trials, we are very likely to have had approximately pn successes.

Lemma 1.2 (Hoeffding's Inequality)**.** *For any $t > 0$ we have*

$$\mathbb{P}(\bar{X} - \mathbb{E}[\bar{X}] \geq t) \leq e^{-\frac{2n^2t^2}{\sum_{i=1}^n (b_i-a_i)^2}}.$$

Hoeffding's inequality did not use any knowledge about the distribution of the random variables X_i. By taking the variance of the X_i's into account, Bernstein's inequality gets a tighter bound.

Lemma 1.3 (Bernstein's Inequality)**.** *Suppose $|X_i| \leq c$ with probability 1 and let $\sigma^2 = \frac{1}{n}\sum_{i=1}^n Var(X_i)$. Then for any $t > 0$ we have*

$$\mathbb{P}(\bar{X} - \mathbb{E}[\bar{X}] \geq t) \leq e^{-\frac{nt^2}{2\sigma^2+2ct/3}}$$

CHAPTER 2

An Introduction to Differential Privacy and its Variants

In this Chapter, we give an introduction to the privacy definitions and techniques that will play an important role in this thesis. For a comprehensive overview of the field, a textbook treatment can be found in [Dwork and Roth, 2014a], and surveys can be found in [Vadhan, 2016, Dwork, 2008, Ji et al., 2014].

2.1 The Definition

A randomised algorithm is differentially private (DP) if the output distributions do not change very much with small changes in the input. It is important to note that it is the *algorithm* that has the property of being DP, not the output. In particular, one can not look at data and decide it is differentially private, and hence safe to release; the guarantee lies with the method used to *produce* the privatised data. This is in contrast to privacy notions like k-anonymity.

Recall that for a database D and algorithm $\mathcal{A}$, we let P_D denote the output distribution on input D. We call two databases neighbours if they differ in a single datapoint. The idea behind DP is simple: on neighbouring databases, the distributions P_D and $P_{D'}$ should be close enough that given a sample from $P_{D''}$, an adversary cannot determine whether $D'' = D$ or $D'' = D'$. The strongest of these definitions is the original, *pure differential privacy*. There is now a broad literature on relaxations of pure DP which find various ways to sacrifice privacy, with a view towards allowing a strictly broader class of algorithms to be implemented.

Definition 2.1 (Data Distribution Independent Privacy Definitions)**.** A randomised algorithm $\mathcal{A}$ is

- ϵ-Pure Differentially Private (pDP) if $\sup_{D,D'} D_\infty(P_D, P_{D'}) \leq \epsilon$.
- (ϵ, δ)-Approximate Differentially Private (aDP) if $\sup_{D,D'} D_\infty^\delta(P_D, P_{D'}) \leq \epsilon$.

where the supremums are over all pairs of neigbouring databases D and D'.

The motivation behind the definition of neighbours is that D contains my true data and D' has my data replaced with a random datapoint. If an adversary cannot distinguish between D and D' then they cannot identify by true data. The neighbouring definition given here isn't applicable for all types of data, and can be replaced with more suitable definitions in different contexts. Differential privacy defined on the new neighbouring relation will have the same interpretation: from the output of the algorithm, it is impossible to distinguish between neighbouring datasets. We will use this idea in Part III.

Note that ϵ-pDP is exactly $(\epsilon, 0)$-aDP. The parameter δ can be thought of as our probability of failing to preserve privacy. To see this, suppose the distributions P_D output 0 with probability $1-\delta$, and a unique identifier for the database D with probability δ. Then this algorithm is $(0, \delta)$-DP. Thus, we typically want δ to be small enough that we can be almost guaranteed that we will not observe this difference in the distributions. In contrast, while it is desirable to have ϵ small, a larger ϵ still gives meaningful guarantees Dwork et al. [2011]. Typically one should think of δ as *extremely* small, $\delta \approx 10^{-8}$, and ϵ as *quite* small, $\epsilon \approx 0.1$.

Approximate DP allows for low probability outcomes that result in catastrophic failures in privacy. We can also consider allowing privacy protection failures on unlikely *inputs*. Philosophically, the downside of this approach is obvious: sometimes the outliers in society are those that need the most protection.

It is often the case in practical data mining tasks that one has a distribution on the data universe. That is, certain data points are more likely than others to be seen in a random sample. For example, suppose the data universe is how far people travelled from home in the summer of 1969. While a random dataset *might* contain a datapoint that is upwards of 238,900 miles, it is very *unlikely* to contain any datapoint larger than 12,500 miles. However, since DP is a *worst-case* guarantee it requires us to output approximately the same value whether or not Neil Armstrong, Buzz Aldrin, or Michael Collins are in our dataset.[1] Hence, DP can exhibit a poor utility/privacy tradeoff due to unlikely events. There is a growing body of literature on weaker versions of privacy that can provide improved utility in these types of circumstances Ebadi et al. [2015], Hall et al. [2012], Barber and Duchi [2014]. These definitions may protect the privacy of everyone except Neil, Buzz and Michael, whose data will be entirely leaked.

Let $\mathcal{D}$ be a distribution on the data universe Ω. For a database D and datapoint z, let $[D_{-1}, z]$ denote the neighbouring database where the first datapoint of D is replaced by z.

Definition 2.2 (Data Distribution Dependent Privacy Definitons)**.** An algorithm $\mathcal{A}$ is

- (ϵ, γ)-Random Pure DP (RpDP) if $\mathbb{P}\left(D_\infty(P_D, P_{[D_{-1},z]}) \leq \epsilon\right) \geq 1-\gamma.$
- $(\epsilon, \delta, \gamma)$-Random Approximate DP (RADP) if $\mathbb{P}\left(D_\infty^\delta(P_D, P_{[D_{-1},z]}) \leq \epsilon\right) \geq 1-\gamma.$

[1] Michael is the unsung hero of the Apollo 11 mission. He piloted the command module alone in lunar orbit while Neil and Buzz landed on the moon. He travelled at least 238,900 miles from home that summer.

where the probabilities are over $D \sim \mathcal{D}^n, z \sim \mathcal{D}$.

Similar to δ, γ represents the probability of catastrophic failure in privacy. Therefore, we require that γ is small enough that this event is extremely unlikely to occur.

2.2 Properties of Differential Privacy

A key component of modern privacy-preserving data analytics is that the sensitive information is provably absent from the privatised data. The following proposition asserts this claim. It says that no amount of ingenuity in post-processing the output of a private algorithm can reveal the sensitive data.

Proposition 2.3 (Post-processing Inequality)**.** *Let $\mathcal{A} : \mathbb{Z}^\Omega \to \mathcal{O}$ be an (ϵ, δ)-DP algorithm and let $f : \mathcal{O} \to \mathcal{O}'$ be an arbitrary randomised algorithm. Then $f \circ \mathcal{A} : \mathbb{Z}^\Omega \to \mathcal{O}'$ is (ϵ, δ)-DP.*

The DP guarantee is that any event in the outcome space is almost equally likely to occur whether the database is D or it's neighbour D'. The following lemma allows us to interpret the word *event* much more broadly. It allows us to say that any consequence of the release of $\mathcal{A}(D)$ was almost equally likely to occur if the input database was D'. For example, suppose the output $\mathcal{A}(D)$ is being used to decide health insurance premiums. The following lemma says that the expected increase in a person's health insurance premium is almost the same, whether or not their data is in D.

Lemma 2.4. *Suppose $\mathcal{A}$ is an (ϵ, δ)-DP algorithm and D, D' are neighbouring databases. Then for any non-negative function $f : \mathcal{B} \to [0, F]$, we have*

$$\mathbb{E}[f(\mathcal{A}(D))] \leq e^\epsilon \mathbb{E}[f(\mathcal{A}(D'))] + \delta F$$

where the expectation is over the randomness in $\mathcal{A}$.

Of course, Proposition 2.3 doesn't hold if the post-processing is allowed to involve re-accessing the data. Fortunately, the composition of two DP algorithms is also a DP algorithm with slightly weaker privacy guarantees. This property, combined with Proposition 2.3 allows us to use DP algorithms as building blocks in larger machine learning algorithms. We begin with the basic composition theorem for the composition of *independent* DP algorithms.

Lemma 2.5 (Basic Composition Theorem)**.** *Suppose for all $i \in [t]$ the algorithm $\mathcal{A}_i : \mathbb{Z}^\Omega \to \mathcal{O}_i$ is (ϵ_i, δ_i)-DP. Then their combination $\mathcal{A}_{[i]} : \mathbb{Z}^\Omega \to \prod_{i=1}^t \mathcal{O}_i$ defined by $\mathcal{A}_{[i]}(D) = (\mathcal{A}_1(D), \cdots, \mathcal{A}_t(D))$ is $(\sum_{i=1}^t \epsilon_i, \sum_{i=1}^t \delta_i)$-DP.*

Now, suppose a data analyst sequentially chooses $\mathcal{A}_i$, receives $\mathcal{A}_i(D_i)$, then chooses $\mathcal{A}_{i+1}$, and so on. It is often the case that the choice of algorithm $\mathcal{A}_{i+1}$ will depend on all the outputs the data analyst has received in the past, as well as all their past choices. That is, the $(i+1)$-th mechanism can be written as $\mathcal{A}_{i+1} : \mathbb{Z}^{\Omega} \times (\mathcal{A}_1 \times \cdots \times \mathcal{A}_i) \times (\mathcal{O}_1 \times \mathcal{O}_i) \to \mathcal{O}_{i+1}$. The composition of the algorithms $\mathcal{A}_1, \cdots, \mathcal{A}_t$ produced in this way is called the *t-fold adaptive composition.*

Notice that the databases D_i are also allowed to vary. Practically, this may occur because time has elapsed, more data collection has occurred, or a different type of data is being used. This is an important attribute because individuals typically have data about them spread across many different databases. Two vectors of databases $(D_1, \cdots, D_t)$ and $(D_1', \cdots, D_t')$ are called neighbours if they differ in exactly one person's data. This may mean that more than one datapoint changes in the union $\cup_{i=1}^t D_i$.

Theorem 2.6 (Advanced Composition Theorem)**.** *For all $\epsilon, \delta, \delta' > 0$, the t-fold adaptive composition of (ϵ, δ)-DP algorithms satisfies $(\epsilon', t\delta + \delta')$-DP for*

$$\epsilon' = \sqrt{2t \ln(1/\delta')\epsilon} + t\epsilon(e^{\epsilon} - 1)$$

Theorem 2.6 improves on Lemma 2.5 as it demonstrates a trade-off between ϵ and δ. At a small cost to the δ parameter, we can allow the ϵ parameter to increase at a rate of $\sqrt{t}$ rather than t.

2.3 Toolbox

In this section we discuss three key differentially private algorithms. This is far from an exhaustive list of DP tools but demonstrates the various important ways uncertainty in the input can be created. Dwork and Roth [2014a] is a excellent resource for a textbook treatment of the core DP algorithms.

2.3.1 Randomised Response

The first algorithm we are going to discuss, *randomised response*, actually predates differential privacy by several decades. It was originally proposed by Warner [1965] to encourage participation and truth-telling in surveys about sensitive personal information. The algorithm is appropriate for binary data, for example answers to YES/NO questions. It proceeds as follows: before answering a YES/NO question, the survey participant is asked to flip a coin with probability p of HEADS and $1-p$ of TAILS. The outcome of the coin flip is not disclosed to the data collector. If HEADS, then survey participant tells the truth, and if TAILS, the survey participant lies.

The premise is that the participant retains *plausible deniability* since the data analyst is uncertain of the outcome of the coin flip. None-the-less, the aggregate data is still useful. Suppose X is a random person's true data and $\hat{X}$ is the output after randomised response. The goal of the data analyst is to learn $\mathbb{P}(X = \text{YES})$. The analyst receives samples of $\hat{X}$, which give an empirical estimate to $\mathbb{P}(\hat{X} = \text{YES})$. This can be translated to an empirical estimate of $\mathbb{P}(X = \text{YES})$, with slightly worse error bounds, via the formula:

$$\mathbb{P}(\hat{X} = \text{YES}) = p\,\mathbb{P}(X = \text{YES}) + (1-p)(1 - \mathbb{P}(X = \text{YES})).$$

The closer p is to 1, the less plausible deniability the participant has, but the more accurate the estimate is. The correct choice of p required to attain ϵ-DP is contained in the following diagram where the LHS is the input to randomised response and the RHS is the output.

$$\begin{tikzcd}
\text{YES} \arrow[r, "\frac{e^\epsilon}{1+e^\epsilon}"] \arrow[dr, "\frac{1}{1+e^\epsilon}" near start, swap] & \text{YES} \\
\text{NO} \arrow[r, "\frac{e^\epsilon}{1+e^\epsilon}", swap] \arrow[ur, "\frac{1}{1+e^\epsilon}" near end, swap] & \text{NO}
\end{tikzcd}$$

Lemma 2.7. *Randomised response with* $p = \frac{e^\epsilon}{1+e^\epsilon}$ *is* ϵ*-DP.*

Randomised response is called a *local* differentially private algorithm because the data is made private before it is sent to the data analyst. In fact, in a statistical setting, randomised response is optimal in the small ϵ regime for binary data [Kairouz et al., 2016]. We will return to local differential privacy in Part III.

2.3.2 The Gaussian Mechanism

Numeric queries of the form $g : \mathbb{Z}^\Omega \to \mathbb{R}^n$ form a fundamental class of statistical queries. For example, learning a parametrised model of the data often lies in this category. Fittingly, the algorithm we will discuss in this section, the Gaussian mechanism, was one of the first differentially private mechanisms. The Gaussian mechanism involves adding Gaussian noise with carefully calibrated standard deviation. The amount of noise is related to two quantities: the amount of privacy desired, and how sensitive the function g is to a single individual's data. The latter quantity is captured by the ℓ_2 sensitivity of g:

$$\triangle_2 g = \max_{D,D' \text{neighbours}} \|g(D) - g(D')\|_2$$

Lemma 2.8 (The Gaussian Mechanism)**.** *[Dwork and Roth, 2014a] Let* $\epsilon > 0$*,* $\delta > 0$ *and* $\sigma = \frac{2\ln(1.25/\delta)\triangle_2 g}{\epsilon}$ *then*

$$\mathcal{A}(D) \sim g(D) + N(0, \sigma^2 I_n)$$

is an (ϵ, δ)*-differentially private algorithm.*

Unlike randomised response, the Gaussian mechanism achieves approximate differential privacy ($\delta > 0$) rather than pure differential privacy. The nonzero δ term comes from bounding the tail probability, since the ratio $\frac{P_D}{P_{D'}}$ is unbounded in the tail. The *Laplacian mechanism* where Laplacian noise, with carefully calibrated standard deviation, is added to the output $g(D)$ is a common alternative to the Gaussian mechanism. The Laplacian mechanism satisfies pure differential privacy. There has been some research on the optimal noise model for achieving local DP [Geng and Viswanath, 2016].

2.3.3 The Exponential Mechanism

The exponential mechanism is used when we would like to choose the optimal response to the data. Common machine learning tasks like least squares regression fall into this category, as well as non-numeric queries like "what is the most common ailment?"

Given an arbitrary range $\mathcal{O}$, a utility function $u : \mathbb{Z}^{\Omega} \times \mathcal{O} \to \mathbb{R}$ maps database/output pairs to utility scores. The goal of the exponential mechanism is to output a differentially private, near optimal, solution to $\arg\max_{o \in \mathcal{O}} u(D, o)$. As in the Gaussian mechanism, an important quantity will be how much the utility score depends on a single individuals data:

$$\triangle u = \max_{o \in \mathcal{O}} \max_{D, D' \text{neighbours}} |u(D, o) - u(D', o)|.$$

The exponential mechanism $\mathcal{M}(D, u)$ selects and outputs an element $o \in \mathcal{O}$ with probability proportional to $e^{\frac{\epsilon u(D,o)}{2\triangle u}}$.

Lemma 2.9. *The exponential mechanism is ϵ-DP.*

Since the probability of an element $o \in \mathcal{O}$ being chosen decays exponentially in $u(D, o)$, with high probability the exponential mechanism outputs an element of $\mathcal{O}$ whose utility is close to optimal.

Lemma 2.10. *Fixing a database D, for any $t > 0$ we have:*

$$\mathbb{P}\left[u(\mathcal{M}(D, u)) \leq \max_{o \in \mathcal{O}} u(D, o) - \frac{2\triangle u}{\epsilon}(\ln |\mathcal{O}| + t)\right] \leq e^{-t}.$$

CHAPTER 3

Information Theory

In this section we review some of the tools from information theory that will be useful throughout this thesis. The majority of the results in this section quantify the fundamental limitations in information recovery in the presence of uncertainty. We begin with general results that hold true for any error metric. We then specialise to the binary case: *property testing*.

3.1 Estimation Lower Bounds

Let us begin by defining the estimation problem at hand. Let $\mathcal{P}$ denote a class of distributions on a sample space M and let $\theta : \mathcal{P} \to \Theta$ determine a function on $\mathcal{P}$. Given i.i.d. samples from $P \in \mathcal{P}$, our goal is estimate $\theta(P)$. For a distribution P, the quantity $\theta(P)$ can be a property of P like its mean or variance, or it can be a unique identifier for P like $\theta(P_D) = D$. Thus, this framework captures both estimating properties of distributions and recovering information under uncertainty.

An estimator is a function $\hat{\theta} : M^r \to \Theta$ that takes as input r independent samples from the distribution, $X \sim P^r$, and outputs at estimate $\hat{\theta}(X)$ to $\theta(P)$. Given a metric $d(\cdot, \cdot)$ on Θ we can define the *maximum risk* of an estimator as

$$\max_{P \in \mathcal{P}} \mathbb{E}_{P^r}[d(\hat{\theta}(X), \theta(P))].$$

The fundamental limitation of reconstruction can then be characterised as the least maximal risk that can be achieved by any estimator:

$$\mathcal{M}_r(\theta(P), d) := \inf_{\hat{\theta}} \max_{P \in \mathcal{P}} \mathbb{E}_{P^r}[d(\hat{\theta}(X), \theta(P))]$$

where the supremum is taken over all distributions $P \in \mathcal{P}$, and the infimum is over all estimator $\hat{\theta}$.

Our first lower bound result is Le Cam's inequality. The TV distance between two distributions is often referred to as the statistical distance because it measures how easy it is to distinguish between the distributions based on samples. Le Cam's inequality states that if there exists a pair of distributions that are close in TV distance but whose estimated parameters are far apart, then estimating the parameter must be difficult.

Lemma 3.1 (Le Cam's inequality)**.** *For any pair $P_0, P_1 \in \mathcal{P}$,*

$$\mathcal{M}_r(\theta(P), d) \geq \frac{d(\theta(P_0), \theta(P_1))}{8} \left(1 - \frac{1}{2}\int |P_0 - P_1|\right)^{2r}$$

Fano's inequality allows us to choose a family of distributions, rather than simply two distributions as in Le Cam's. It states that if there is a large family of distributions that are close in KL divergence (another statistical distance measure) but far in estimated parameters then it is difficult to estimate the parameter. If (Θ, d) is a metric space, $\alpha > 0$ and $T \subset \Theta$, then we define the α-packing number of T to be the largest number of disjoint balls of radius α that can fit in T, denoted by $\mathcal{M}(\alpha, T, d)$. For a collection S of probability distributions, the KL diameter is defined by

$$d_{KL}(S) = \sup_{p,q \in S} D_{\mathrm{KL}}(p \| q). \tag{3.1}$$

The following version of Fano's inequality is found in Yu [1997]. It will be the version utilised throughout the remainder of this thesis.

Lemma 3.2 (Fano's Inequality)**.** *Let (M, d) be a metric space and $\{\mathbb{P}_\theta \mid \theta \in M\}$ be a collection of probability measures. For any totally bounded $T \subset M$ and $\alpha > 0$,*

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{P}_\theta \left(d^2(\hat{\theta}(X), \theta(P)) \geq \frac{\epsilon^2}{4} \right) \geq 1 - \frac{D_{KL}(P_T) + 1}{\log \mathcal{M}(\alpha, T, d)} \tag{3.2}$$

where $P_T = \{\mathbb{P}_\theta \mid \theta \in T\}$.

Both Le Cam's inequality and Fano's inequality have many forms. The versions presented here will be the most useful to us.

3.2 Property Testing Lower Bounds

In property testing of distributions, the aim is to decide whether a distribution has a property or not based on samples from the distribution. For example, is the distribution log-concave? Lipschitz? Gaussian? In Part IV we will be considering whether a pair of distributions satisfies the requirements for DP.

A *property* can abstractly be defined as a subset $\tilde{\mathcal{P}} \subset \mathcal{P}$. Given $X \sim P^r$, the problem is to decide if $P \in \tilde{\mathcal{P}}$ or if P is far from $\tilde{\mathcal{P}}$. The meaning of *far* varies depending on the context but is typically defined via a metric on the space of distributions $\mathcal{P}$. The *error* of an algorithm solving the decision problem is the probability, over draws from the distributions, that the algorithm is wrong.

Definition 3.3 (Property testing)**.** A property testing algorithm with sample complexity r, proximity parameter α, and property $\tilde{\mathcal{P}} \subset \mathcal{P}$, takes r samples from P and

1. (Completeness) ACCEPTS with probability at least $2/3$ if $P \in \tilde{\mathcal{P}}$.
2. (Soundness) REJECTS with probability at least $2/3$ if P is α-far from $\tilde{\mathcal{P}}$.

Property testing is a special case of the general estimation framework where $\Theta = \{0, 1\}$, $\theta(P) = 1 - \chi_{\tilde{P}}$ and $d(x, y) = |x - y|$ is the 0-1 loss. Then, $\mathbb{E}_{P^r}[d(\hat{\theta}(X), \theta(P))]$ is the probability that the estimator $\hat{\theta}$ is wrong on the distribution P. Thus, we get the following corollary of Le Cam's inequality.

Corollary 3.4 (Le Cam's inequality for property testing)**.** *Let α be the proximity parameter and $\tilde{\mathcal{P}}$ be a property. Suppose $P \in \tilde{\mathcal{P}}$ and Q is α-far from $\tilde{\mathcal{P}}$. Then, any property testing algorithm with query complexity $2r$ must satisfy*

$$\|P^r - Q^r\|_{TV} \geq \frac{1}{3}$$

The previous results in this chapter were all based on the concept that if two distributions are close in statistical distance, then you need a lot of samples to distinguish between them. We are going to restrict now to discrete distributions on $[n]$. The final lower bound we will discuss is built on the idea of *low frequency blindness*. Low frequency elements, that is elements that have low probability of occuring, are unlikely to appear in a random sample. However, they may have a large effect on whether or not the property is satisfied so an estimator cannot simply assume these unseen elements have probability zero. Lemma 3.5 captures the idea that "no tester can extract useful information from low-frequency elements".

We need to consider a generalisation of binary properties. A property π of a distribution is a function $\pi : \triangle[n] \to \mathbb{R}$, where $\triangle[n]$ is the set of discrete distributions on $[n]$. It is called symmetric if for all permutations σ and distributions P we have $\pi(p) = \pi(p \circ \sigma)$. It is (α, β)-weakly-continuous if for all distributions P^+ and P^- satisfying $\|P^+ - P^-\|_{\text{TV}} \leq \beta$ we have $|\pi(P^+) - \pi(P^-)| \leq \alpha$.

Lemma 3.5 (Low frequency blindness)**.** *[Valiant, 2011] Given a symmetric property π on distributions on $[n]$ that is (α, β)-weakly-continuous, if there exists two distributions, P^+, P^- that are identical for any index occurring with probability at least $1/k$ in either distribution but $\pi(P^+) > b$ and $\pi(P^-) < a$, then no tester can distinguish between $\pi > b - \alpha$ and $\pi < a + \alpha$ in $k \cdot \frac{\beta}{1000 \cdot 2^{4\sqrt{\log n}}}$ samples.*

As with the other Lemmas in this chapter, the key to using Lemma 3.5 is to find a pair of distributions that are *close*, in that they agree for high frequency elements, but far in the property parameter.

Part II

When is Nontrivial Graphon Estimation Possible?

CHAPTER 4

Introduction

Networks and graphs arise as natural modelling tools in many areas of science. In many settings, particularly in social networks, networks display some type of community structure. In these settings one may consider the nodes of the graph as belonging to one of k communities, and two nodes are connected with a probability that depends on the communities they belong to. This type of structure is captured in the *k-block graphon* model, also known as the stochastic block model. The more communities we allow in the model, the richer the model becomes and the better we can hope to describe the real world. One can think of a general *graphon* model as an ∞-block graphon where each node is given a label in $[0, 1]$ rather than $[k]$.

Given an observed network, graphon estimation is the problem of recovering the graphon model from which the graph was drawn. In this part, we are concerned with the fundamental limits of graphon estimation for block graphons. That is, given a n-node network that was generated from a k-block graphon, how accurately can you recover the graphon? We consider the "nonparametric" setting, where k may depend on n. Our lower bounds apply even to estimation algorithms that know the true number of blocks k, though this quantity typically needs to be estimated [1].

In many real world networks, the average degree of the network is small compared to the number of nodes in the network. Graphons whose expected average degree is linear in n are called dense, while graphons whose expected average degree is sublinear in n are referred to as sparse. In this work, we prove a new lower bound for graphon estimation for sparse networks. In particular, our results rule out *nontrivial* estimation for very sparse networks (roughly, where $\rho = O(k^2/n^2)$). An estimator is nontrivial if its expected error is significantly better than an estimator which ignores the input and always outputs the same model. It follows from recent work [Mossel et al., 2014, 2015, Neeman and Netrapalli, 2014] that nontrivial estimation is impossible when $\rho = O(1/n)$. Ours is the first lower bound that rules out nontrivial graphon estimation for large k. Previous work by Klopp

[1] This Part is based on joint with Adam Smith [McMillan and Smith, 2017]

et al. [2015] provides other lower bounds on graphon estimation that are tight in several regimes. In work concurrent to ours [Klopp et al., 2016], the same authors provide a similar bound to the one presented here.

Block graphon models were introduced by Hoff et al. [2002] under the name latent position graphs. Graphons play an important role in the theory of graph limits (see [Lovász, 2012] for a survey) and the connection between the graph model and convergent graph sequences has been studied in both the dense and sparse settings [Borgs et al., 2006, 2008, 2014b,a,b]. Estimation for stochastic block models with a fixed number of blocks was introduced by Bickel and Chen [2009], while the first estimation of the general model was proposed by Bickel et al. [2011]. Many graphon estimation methods, with an array of assumptions on the graphon, have been proposed since [Lloyd et al., 2012, Tang et al., 2013, Latouche and Robin, 2016, Wolfe and Olhede, 2013, Chan and Airoldi, 2014, Airoldi et al., 2013, Yang et al., 2014, Gao et al., 2015, Abbe et al., 2016, Chatterjee, 2015, Abbe and Sandon, 2015]. Gao et al. [2015] provide the best known upper bounds in the dense setting while Wolfe and Olhede [2013], Borgs et al. [2015] and Klopp et al. [2015] give upper bounds for the sparse case.

4.1 Graphons

Definition 4.1 (Bounded Graphons and W-random graphs)**.** A (bounded) *graphon* W is a symmetric, measurable function $W : [0,1]^2 \to [0,1]$. Here, symmetric means that $W(x,y) = W(y,x)$ for all $(x,y) \in [0,1]^2$.

For any integer n, a graphon W defines a distribution on graphs on n vertices as follows: First, select n labels $\ell_1, \cdots, \ell_n$ uniformly and independently from $[0,1]$, and form an $n \times n$ matrix H where $H_{ij} = W(\ell_i, \ell_j)$. We obtain an unlabelled, undirected graph G by connecting the ith and jth nodes with probability H_{ij} independently for each (i,j). The resulting random variable is called a *W-random graph*, and denoted $G_n(W)$.

For $\rho \geq 0$, we say a graphon is ρ-bounded if W takes values in $[0,\rho]$ (that is, $\|W\|_\infty \leq \rho$).

We denote the set of graphs with n nodes by $\mathcal{G}_n$, the set of graphons by $\mathcal{W}$ and the set of ρ-bounded graphons by $\mathcal{W}_\rho$. If W is ρ-bounded, then the expected number of edges in $G_n(W)$ is at most $\rho\binom{n}{2} = O(\rho n^2)$. In the case that ρ depends on n and $\lim_{n\to\infty} \rho \to 0$, we obtain a sparse graphon.

We consider the estimation problem: given parameters n and ρ, as well as a graph $G \sim G_n(W)$ generated from an unknown ρ-bounded graphon W, how well can we estimate W?

A natural goal is to design estimators that produce a graphon $\hat{W}$ that is close to W in a metric such as L_2. This is not possible, since there are many graphons that are far apart in L_2, but that generate the same probability distribution on graphs. If there exists a measure preserving map $\phi : [0,1] \to [0,1]$ such that $W(\phi(x), \phi(y)) = W'(x,y)$ for all $x, y \in [0,1]$, then $G_n(W)$ and $G_n(W')$ are identically distributed. The converse is true if we instead only require $W(\phi(x), \phi(y)) = W'(x,y)$ almost everywhere. Thus, we wish to say that $\hat{W}$ approaches the *class* of graphons that generate $G_n(W)$. To this end, we use the following metric on the set of graphons,

$$\delta_2(W, W') = \inf_{\substack{\phi:[0,1]\to[0,1] \\ \text{measure-preserving}}} \|W_\phi - W'\|_2 \tag{4.1}$$

where $W_\phi(x,y) = W(\phi(x), \phi(y))$ and ϕ ranges over all measurable, measure-preserving maps. Two graphons W and W' generate the same probability distribution on the set of graphs if and only if $\delta_2(W, W') = 0$ (see [Lovász, 2012], for example).

Existing upper bounds for graphon estimation are based on algorithms that produce graphons of a particular form, namely *block graphons*, also called *stochastic block models* (even when it is not known that the true graphon is a block graphon).

Definition 4.2 (k-block graphon (stochastic block models))**.** For $k \in \mathbb{N}$, a graphon is a *k-block graphon* if there exists a partition of $[0,1]$ into k measurable sets $I_1, \cdots, I_k$ such that W is constant on $I_i \times I_j$ for all i and j.

We can associate a graphon of this form to every square matrix. Given a $k \times k$ symmetric matrix M, let $W[M]$ denote the k-block graphon with blocks $I_i = (\frac{i-1}{k}, \frac{i}{k}]$ that takes the value M_{ij} on $I_i \times I_j$.

4.2 Main result

We are concerned with the problem of estimating a graphon, W, given a graph sampled from $G_n(W)$. A graphon estimator is a function $\hat{W} : \mathcal{G}_n \to \mathcal{W}$ that takes as input a n node graph, that is generated according to W, and attempts to output a graphon that is close to W. The main contribution of this work is the development of the lower bound

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G \sim G_n(W)} [\delta_2(\hat{W}(G), W)] \geq \Omega\left(\min\left(\rho, \sqrt{\frac{\rho k^2}{n^2}}\right)\right). \tag{4.2}$$

Combined with previous work we can give the following lower bound on the error of graphon estimators.

Theorem 4.3. *For any positive integer* $2 \leq k \leq n$ *and* $0 < \rho \leq 1$,

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G \sim G_n(W)} \left[\delta_2(\hat{W}(G), W)\right] \geq \Omega\left(\min\left(\rho,\ \rho\sqrt[4]{\frac{k}{n}} + \sqrt{\frac{\rho k^2}{n^2}}\right)\right) \tag{4.3}$$

where $\inf_{\hat{W}}$ *is the infimum over all estimators* $\hat{W} : G_n \to \mathcal{G}$ *and* $\sup_W$ *is the supremum over all* k*-block,* ρ*-bounded graphons. If* ρn *is non-decreasing and there exists a constant* $c > 0$ *such that* $\rho n \geq c$ *then*

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G \sim G_n(W)} \left[\delta_2(\hat{W}(G), W) \right] \geq \Omega \left(\min \left(\rho,\ \rho \sqrt[4]{\frac{k}{n}} + \sqrt{\frac{\rho k^2}{n^2}} + \sqrt{\frac{\rho}{n}} \right) \right)$$

Note that k and ρ may depend on n. That is, the theorem holds if we consider sequences ρ_n and k_n. Our result improves on previously known results when $\rho = O\left(\left(\frac{k}{n}\right)^{3/2} \right)$—that is, when the graphs produced by the graphon are sparse. The upper bound

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G \sim G_n(W)} \left[\delta_2(\hat{W}(G), W) \right] \leq O \left(\min \left(\rho,\ \rho \sqrt[4]{\frac{k}{n}} + \sqrt{\frac{\rho k^2}{n^2}} + \sqrt{\frac{\rho \log k}{n}} \right) \right) \tag{4.4}$$

by Klopp et al. [2015] implies that our lower bound is almost tight. In particular, if k is constant and ρ is within the designated range then the lower bound in Theorem 4.3 is tight.

When $\rho = O\left(\frac{k^2}{n^2}\right)$, Theorem 4.3 implies that the error is $\Omega(\rho)$, which is the error achieved by the trivial estimator $\hat{W} = 0$. That is, in the sparse setting, the trivial estimator achieves the optimal error. To the authors' knowledge this is the first result that completely rules out nontrivial estimation in the case where k is large. Concurrent work [Klopp et al., 2016] provides similar bounds.

The bound

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G \sim G_n(W)} [\delta_2(\hat{W}, W)] \geq \Omega \left(\rho \sqrt[4]{\frac{k}{n}} \right) \tag{4.5}$$

is due to previous work of Klopp et al. [2015] and the bound

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G \sim G_n(W)} [\delta_2(\hat{W}, W)] \geq \Omega \left(\min \left(\rho, \sqrt{\frac{\rho}{n}} \right) \right) \tag{4.6}$$

for constant k follows from results of Mossel et al. [2014] and Banerjee [September 2016]. We give details on how to derive (4.6) from their results in the Appendix.

4.3 Techniques: Combinatorial Lower Bounds for δ_p

Our proof of the main theorem will involve Fano's lemma. As such, during the course of the proof we will need to lower bound the packing number, with respect to δ_2, of a large set of k-block graphons. Whilst easily upper bounded, little is known about lower bounds on δ_2. To the authors' knowledge, this work gives the first lower bound for the packing number of $\mathcal{W}_\rho$ with respect to δ_2. We will also give a combinatorial lower bound for the δ_2 metric that is easier to handle than the metric itself.

To understand our technical contributions, it helps to first understand a problem related to graphon estimation, namely that of estimating the matrix of probabilities H. Existing algorithms for graphon estimation are generally analysed in two phases: first, one shows that the estimator $\hat{W}$ is close to the matrix H (in an appropriate version of the δ_2 metric), and then uses (high probability) bounds on $\delta_2(W, W[H])$ to conclude that $\hat{W}$ is close to W. Klopp et al. [2015] show tight upper and lower bounds on estimation of H. One can think of our lower bound as showing that the lower bounds on estimation of H can be transferred to the problem of estimating W.

The main technical difficulty lies in showing that a given pair of matrices A, B lead to graphons that are far apart in the δ_2 metric. Even if A, B are far apart in, say, ℓ_2, they may lead to graphons that are close in δ_2. For consistency with the graphon formalism, we normalise the ℓ_2 metric on $k \times k$ matrices so that it agrees with the L_2 metric on the corresponding graphons. For a $k \times k$ matrix A,

$$\|A\|_2 := \Big(\frac{1}{k^2}\sum_{i,j\in[k]} A_{ij}^2\Big)^{1/2} = \|W[A]\|_2\,. \tag{4.7}$$

As an example of the discrepancy between the ℓ_2 and δ_2 metrics, consider the matrices

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The matrices A and B have positive distance in the ℓ_2 metric, $\|A - B\|_2 = \frac{2}{3}$, but

$$\delta_2(W[A], W[B]) \; = \; 0.$$

One can get an *upper bound* on $\delta_2(W[A], W[B])$ by restricting attention in the definition of δ_2 to functions ϕ that permute the blocks I_i. This leads to the following metric on $k \times k$ matrices which minimises over permutations of the rows and columns of one of the matrices:

$$\hat{\delta}_2(A, B) := \min_{\sigma \in \mathcal{S}_k} \|A_\sigma - B\|_2\ , \tag{4.8}$$

where A_σ is the matrix with entries $(A_\sigma)_{ij} = A_{\sigma(i),\sigma(j)}$. This metric arises in other work (e.g. [Lovász, 2012]), and it is well known that

$$\delta_2(W[A], W[B]) \le \hat{\delta}_2(A, B). \tag{4.9}$$

To prove lower bounds, we consider a new metric on matrices, in which we allow the rows and columns to be permuted separately. Specifically, let

$$\hat{\hat{\delta}}_2(A, B) := \min_{\sigma,\tau \in \mathcal{S}_k} \|A_{\sigma,\tau} - B\|_2\ , \tag{4.10}$$

where $A_{\sigma,\tau}$ is the $k \times k$ matrix with entries $(A_{\sigma,\tau})_{ij} = A_{\sigma(i),\tau(j)}$.

Lemma 4.4 (Lower bound for δ_2)**.** *For every two $k \times k$ matrices A, B,*

$$\hat{\hat{\delta}}_2(A, B) \leq \delta_2(W[A], W[B]). \tag{4.11}$$

Because $\hat{\hat{\delta}}_2$ is defined "combinatorially" (that is, it involves minimisation over a discrete set of size about $2^{2k \ln k}$, instead of over all measure-preserving injections), it is fairly easy to lower bound $\hat{\hat{\delta}}_2(A, B)$ for random matrices A, B using the union bound. In particular, it allows us to give bounds on the packing number of $\mathcal{W}_\rho$ with respect to the δ_2 metric. The following Proposition will be proved after the proof of Theorem 4.3.

Proposition 4.5. *There exists $C > 0$ such that the $C\rho$-packing number of $\mathcal{W}_\rho$, equipped with δ_2, is $2^{\Omega(k^2)}$, that is $\mathcal{M}(C\rho, \mathcal{W}_\rho, \delta_2) = 2^{\Omega(k^2)}$.*

Finally, we note that these techniques extend directly to the δ_p metric, for $p \in [1, \infty]$. That is, we may define $\delta_p, \hat{\delta}_p$ and $\hat{\hat{\delta}}_p$ analogously to the definitions above, and obtain the bounds

$$\hat{\hat{\delta}}_p(A, B) \leq \delta_p(W[A], W[B]) \leq \hat{\delta}_p(A, B), \tag{4.12}$$

along with similar lower bounds on the packing number.

4.4 Related work

Work on graphon estimation falls broadly into two categories; estimating the matrix H and estimating the graphon W. When estimating H, the aim is to produce a matrix that is close in the ℓ_2 metric to the true matrix of probabilities H that was used to generate the graph G. When estimating the graphon, our aim is the minimise the δ_2 distance between the estimate and the true underlying graphon W that was used to generate G.

Gao et al. studied the problem of estimating the matrix of probabilities H given an instance chosen from W when $\rho = 1$. They proved the following minimax rate for this problem when W is a k-block graphon:

$$\inf_{\hat{M}} \sup_{H} \mathop{\mathbb{E}}_{G \sim G_n(H)} \left[\frac{1}{n^2} \left\| \hat{M}(G) - H \right\|_2 \right] \asymp \sqrt{\frac{k^2}{n^2} + \frac{\log k}{n}} \tag{4.13}$$

where the infinimum is over all estimators $\hat{M}$ from G_n to the set of symmetric $n \times n$ matrices, the supremum is over all probability matrices H generated from k-block graphons. Klopp et al. extended this result to the sparse case, proving that for all $k \leq n$ and $0 < \rho \leq 1$,

$$\inf_{\hat{M}} \sup_{H} \mathop{\mathbb{E}}_{G \sim G_N(H)} \left[\frac{1}{n^2} \left\| \hat{M}(G) - H \right\|_2 \right] \geq \Omega \left(\min \left(\sqrt{\rho \left(\frac{k^2}{n^2} + \frac{\log k}{n} \right)}, \rho \right) \right) \tag{4.14}$$

where the supremum is over all probability matrices H generated from k-block, ρ-bounded graphons.

[Klopp et al., 2015, Corollary 3] also studied the problem of estimating the graphon W. They proved that Equation (4.4) holds for any k-block, ρ-bounded graphon, W, with $k \leq n$. They also exhibited the first lower bound (known to us) for graphon estimation using the δ_2 metric. They proved that Equation (4.5) holds for $\rho > 0$ and $k \leq n$.

The related problems of distinguishing a graphon with $k > 1$ from an Erdös-Rényi model with the same average degree (called the distinguishability problem) and reconstructing the communities of a given network (called the reconstruction problem) have also been widely studied. This problem is closely related to the problem of estimating H. Recent work by Mossel et al. [2014] and Neeman and Netrapalli [2014] establish conditions under which a k-block graphon is mutually contiguous to the Erdös-Rényi model with the same average degree. Contiguity essentially implies that no test could ever definitely determine which of the two graphons a given sample came from. There is a large body of work on algorithmic and statistical problems in this area and we have only cited work that is directly relevant here.

CHAPTER 5

Technical Contributions

5.1 Lower Bound for the δ_2 Metric

As mentioned earlier, the main technical contribution of this part is lower bounding the δ_2 metric by the more combinatorial $\hat{\hat{\delta}}_2$ metric. In this section we will prove the inequality given in Lemma 4.4.

Proposition 5.1. *Let W, W' be k-block graphons with blocks $I_i = [\frac{i-1}{k}, \frac{i}{k})$ and $\pi : [0,1] \to [0,1]$ be a measure-preserving map. Then there exists a probability distribution $\mathbb{P}$ on $\mathcal{S}_k$ such that*

$$\|W_\pi - W'\|_2^2 = \mathop{\mathbb{E}}_{\sigma,\tau\sim\mathbb{P}}[\|W_{\sigma,\tau} - W'\|_2^2]\,, \tag{5.1}$$

where the expectation is taken over σ, τ selected independently according to $\mathbb{P}$.

Proof. Let $a_i = \frac{i-1}{k}$ and $p_{ij} = \mu(I_i \cap \pi^{-1}(I_j))$. Now, consider a $k \times k$ matrix P with $P_{ij} = kp_{ij}$. Noting that $\sum_{j=1}^k p_{ij} = \mu(I_i) = 1/k$ and $\sum_{i=1}^k p_{ij} = \mu(\pi^{-1}(I_j)) = 1/k$, we can see that P is doubly stochastic, that is, the rows and columns of P sum to 1. Berkhoff's theorem states that any doubly stochastic matrix can be written as a convex combination of permutation matrices. Therefore, we have a probability distribution $\mathbb{P}$ on $\mathcal{S}_k$ such that $P = \sum_{\sigma\in\mathcal{S}_k} \mathbb{P}(\sigma)\sigma$ and $\sum_{\sigma\in\mathcal{S}_k} \mathbb{P}(\sigma) = 1$ and

$$\mathbb{P}(\sigma(i) = j) = \sum\{\mathbb{P}(\sigma) \mid \sigma(i) = j\} = P_{ij} = kp_{ij}. \tag{5.2}$$

Taking expectations over σ, τ selected independently from $\mathbb{P}$,

$$\begin{aligned}
\mathbb{E}[\|W_{\sigma,\tau} - W'\|_2^2] &= \sum_{\sigma,\tau} \mathbb{P}(\sigma)\mathbb{P}(\tau) \sum_{i,j} \frac{1}{k^2}(W(a_{\sigma(i)}, a_{\tau(j)}) - W'(a_i, a_j))^2 \\
&= \sum_{i,i',j,j'} \frac{1}{k^2}\mathbb{P}(\sigma(i) = i')\mathbb{P}(\tau(j) = j')(W(a_i, a_j) - W'(a_{i'}, a_{j'}))^2 \\
&= \sum_{i,i',j,j'} p_{ii'}p_{jj'}(W(a_i, a_j) - W'(a_{i'}, a_{j'}))^2 \\
&= \|W_\pi - W'\|_2^2\,.
\end{aligned}$$

□

Proof of Lemma 4.4. Proposition 5.1 implies that for all measure preserving maps $\pi : [0,1] \to [0,1]$ and matrices A and B we have

$$\|W[A]_\pi - W[B]\|_2 \geq \inf_{\sigma,\tau \in \mathcal{S}_k} \|W[A]_{\sigma,\tau} - W[B]\|_2 = \inf_{\sigma,\tau \in \mathcal{S}_k} \|A_{\sigma,\tau} - B\|_2 = \hat{\hat{\delta}}_2(A,B). \tag{5.3}$$

Since this is true for any π, we have $\delta_2(W[A], W[B]) \geq \hat{\hat{\delta}}_2(A,B)$. □

5.2 Proof of Theorem 4.3

To prove the main theorem we will use Fano's lemma to find a constant that lower bounds the probability that the estimation exceeds $\min\left(\rho, \sqrt{\frac{\rho k^2}{n^2}}\right)$, which then implies the appropriate lower bound on the expected δ_2 error. To that end, we aim to find a large set, T, of k-block graphons whose KL-diameter and ϵ-packing number with respect to δ_2 with $\epsilon = \min\left(\rho, \sqrt{\frac{\rho k^2}{n^2}}\right)$ can be bounded. Our proof is inspired by that of Gao et al.

The following lemma gives us a way to easily upper bound the KL divergence between the distributions induced by two different graphons.

Lemma 5.2. *For any graphons* $\frac{1}{4} \leq W, W' \leq \frac{3}{4}$, *we have*

$$D(G_n(W) \| G_n(W')) \leq 8n^2 \|W - W'\|_2^2 . \tag{5.4}$$

Proof. Let T be a variable denoting the choice of labels, so

$$\mathbb{P}_{G_n(W)}(G) = \int_{\ell \in [0,1]^n} \mathbb{P}_T(\ell) \mathbb{P}_{G_n(W)}(G | T = \ell) d\ell. \tag{5.5}$$

Now,

$$\begin{aligned} D(G_n(W) \| G_n(W')) &= \sum_{G \in G_n} \mathbb{P}_{G_n(W)}(G) \ln\left(\frac{\mathbb{P}_{G_n(W)}(G)}{\mathbb{P}_{G_n(W')}(G)}\right) \\ &\leq \sum_{G \in G_n} \int_{\ell \in [0,1]^n} \mathbb{P}_T(\ell) \mathbb{P}_{G_n(W)}(G|T = \ell) \ln\left(\frac{\mathbb{P}_{G_n(W)}(G|T=\ell)}{\mathbb{P}_{G_n(W')}(G|T=\ell)}\right) d\ell \\ &= \int_{\ell \in [0,1]^n} \mathbb{P}_T(\ell) D(\mathbb{P}_{G_n(W)}(\cdot | T = \ell) \| \mathbb{P}_{G_n(W')}(\cdot | T = \ell) d\ell, \end{aligned}$$

where the inequality follows from the log-integral inequality [Lapidoth, 2017, Theorem 30.12.4]. Now, the probability density function of T is the constant function 1 so it follows from [Gao et al., 2015, Proposition 4.2] that

$$\begin{aligned}
D(G_n(W)\|G_n(W')) &\le 8\int_{\ell\in[0,1]^n}\sum_{i,j=1}^{n}(W(\ell_i,\ell_j)-W'(\ell_i,\ell_j))^2 d\ell \\
&= 8\sum_{i,j=1}^{n}\int_{\ell\in[0,1]^n}(W(\ell_i,\ell_j)-W'(\ell_i,\ell_j))^2 d\ell \\
&\le 8n^2\int_{[0,1]^2}(W(x,y)-W'(x,y))^2 dxdy \\
&= 8n^2\left\|W-W'\right\|_2^2
\end{aligned}$$

□

Recall that we are aiming to define a large set of k-block matrices that are close in KL divergence, but that are ϵ-far apart with respect to δ_2 (with $\epsilon=\min(\rho,\sqrt{\frac{\rho k^2}{n^2}})$). The following lemma shows that there exists a large set of matrices who are pairwise far in Hamming distance, even after every possible permutation of the rows and columns. We will use this in the proof of Theorem 4.3 to define a large class of k-block graphons who are pairwise far in the $\hat{\hat{\delta}}_2$ metric and hence the δ_2 metric. This gives us a bound on packing number.

Lemma 5.3. *There exists a set S of symmetric $k\times k$ binary matrices such that $|S|=2^{\Omega(k^2)}$ and, for every $B,B'\in S$ and $\sigma,\tau\in\mathcal{S}_k$, we have Ham$(B_{\sigma,\tau},B')=\Omega(k^2)$.*

Proof. Fix permutations σ and τ, and consider two randomly chosen symmetric binary matrices B,B'. For $i\le j$, let $X_{ij}=1$ if $B_{\sigma(i),\tau(j)}=B'_{i,j}$ and 0 otherwise so X_{ij} is a Bernoulli random variable with $\mathbb{E}[X_{ij}]=\frac{1}{2}$. Thus, by a Chernoff bound,

$$\mathbb{P}\left(\text{Ham}(B_{\sigma,\tau},B')<\frac{k^2}{6}\right)=\mathbb{P}\left(\sum_{i\le j}X_{ij}\le\frac{k^2}{6}\right)\le e^{\frac{-2\left(\frac{k^2}{6}-\frac{1}{2}\binom{k}{2}\right)^2}{\binom{k}{2}}}. \tag{5.6}$$

Therefore, for randomly chosen B,B',

$$\mathbb{P}\left(\exists\sigma,\tau\text{ s.t. Ham}(B_{\sigma,\tau},B')<\frac{k^2}{6}\right)\le e^{\frac{-2\left(\frac{k^2}{6}-\frac{1}{2}\binom{k}{2}\right)^2}{\binom{k}{2}}}(k!)^2=2^{-\Omega(k^2)}. \tag{5.7}$$

For a constant $c>0$, consider the process that selects 2^{ck^2} binary matrices $\{B_i\}_i$ uniformly at random. The probability that all pairs are at Hamming distance at least $k^2/6$ is at least $1-2^{2ck^2}2^{-\Omega(k^2)}$. Selecting c sufficiently small, we get that at least one such set S exists. □

We are not aware of an explicit construction of a large family of matrices that are far apart in $\hat{\hat{\delta}}_2$ metric; we leave such a construction as an open problem.

We now proceed to the proof of Theorem 4.3. We will use Lemma 5.3 to define a set T with packing number $2^{\Omega(k^2)}$. The elements of T are all close in $\|\cdot\|_\infty$ norm, so using Lemma 5.2 we get a bound on the KL diameter. We then directly apply these bounds via Fano's lemma.

Theorem 5.4. *For any positive integer $k \le n$ and $0 < \rho \le 1$,*

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G \sim G_n(W)} \left[\delta_2(\hat{W}(G), W)\right] \ge \Omega\left(\min\left(\rho, \sqrt{\frac{\rho k^2}{n^2}}\right)\right). \tag{5.8}$$

where $\inf_{\hat{W}}$ is the infimum over all estimators $\hat{W} : G_n \to \mathcal{G}$ and $\sup_W$ is the supremum over all k-block, ρ-bounded graphons.

Proof. Let S be a set satisfying the conditions of Lemma 5.3 and let $\eta = \min(1, \frac{k}{n\sqrt{\rho}})$. For $B \in S$, define

$$Q_B = \rho\left[\frac{1}{2}\mathbf{1} + c\eta(2B - \mathbf{1})\right], \tag{5.9}$$

where $\mathbf{1}$ is the all 1's matrix and c is some constant that we will choose later. That is, $(Q_B)_{ij} = \rho[\frac{1}{2} + c\eta]$ if $B_{ij} = 1$ and $(Q_B)_{ij} = \rho[\frac{1}{2} - c\eta]$ if $B_{ij} = 0$. Let $T = \{W[Q_B] \mid B \in S\}$. Using Lemma 5.2 we conclude that for all $W, W' \in T$, we have

$$D(G_n(W)\|G_n(W')) \le 8n^2(2c\rho\eta)^2 \le 32c^2k^2\rho \tag{5.10}$$

so $d_{KL}(T) = O(c^2k^2\rho)$.

Let $B, B' \in S$ and suppose $\sigma, \tau \in \mathcal{S}_k$. By construction,

$$\|(W[Q_B])_{\sigma,\tau} - W[Q_{B'}]\|_2^2 \ge \frac{1}{k^2}\text{Ham}(B_{\sigma,\tau}, B')(2\rho c\eta)^2 = \Omega(c^2\rho^2\eta^2). \tag{5.11}$$

Thus by Corollary 5.1,

$$\delta_2(W[Q_B], W[Q_{B'}]) \ge \hat{\hat{\delta}}_2(W[Q_B], W[Q_{B'}]) \ge \Omega(c\rho\eta). \tag{5.12}$$

Therefore, there exists $D > 0$ such that if $\epsilon = D\rho c\eta = D\min\left(c\rho, \frac{ck\sqrt{\rho}}{n}\right)$, we have $\log \mathcal{M}(\epsilon, T, \delta_2) = \Omega(k^2)$. Lemma 3.2 implies

$$\inf_{\hat{W}} \sup_{W} \Pr\left(\delta_2(\hat{W}, W) \ge \frac{\epsilon}{2}\right) \ge 1 - \frac{O(c^2k^2\rho) + 1}{\Omega(k^2)}. \tag{5.13}$$

We can choose c small enough that the right hand side is larger than a fixed constant for all k and n. By Markov's inequality we have

$$\inf_{\hat{W}} \sup_{W} \mathbb{E}\left[\delta_2(\hat{W}, W)\right] = \Omega(\epsilon) = \Omega\left(\min\left(\rho, \sqrt{\rho\frac{k^2}{n^2}}\right)\right). \tag{5.14}$$

□

Proof of Proposition 4.5. During the course of the proof of Theorem 4.3 we construct $2^{\Omega(k^2)}$ graphons in $\mathcal{W}_\rho$ that are pairwise at least $\Omega(\rho c\eta)$ apart in the δ_2 distance for any $c > 0$ such that $|c\eta| \leq \frac{1}{2}$. Therefore, for some $C > 0$, the $C\rho$-packing number of $\mathcal{W}_\rho$ is at least $2^{\Omega(k^2)}$. $\square$

5.3 Proof of Equation (1.6)

We show here how to derive the lower bound in (4.6) from the results of Mossel et al. [2014] and Banerjee [September 2016].

Let $(\Omega_n, \mathcal{F}_n)$ be a sequence of measurable spaces, each equipped with two probability measures, $\mathbb{P}_n$ and $\mathbb{Q}_n$. We say $\mathbb{P}_n$ and $\mathbb{Q}_n$ are mutually contiguous if for any sequence of events A_n, we have $\lim_{n\to\infty} \mathbb{P}_n(A_n) \to 0$ if and only if $\lim_{n\to\infty} \mathbb{Q}_n(A_n) \to 0$.

Lemma 5.5. *Let W_1 be a k-block graphon generated by a matrix with diagonal entries, p, and off-diagonal entries, q. Let W_2 be the Erdös-Rényi model with the same expected degree as W_1. If W_1 and W_2 are mutually contiguous then*

$$\inf_{\hat{W}} \sup_{W} \mathop{\mathbb{E}}_{G\sim G_n(W)} \left[\delta_2(\hat{W}(G), W)\right] \geq \Omega\left(\frac{|p-q|}{\sqrt{k}}\right). \tag{5.15}$$

where $\inf_{\hat{W}}$ is the infimum over all estimators $\hat{W} : G_n \to \mathcal{G}$ and $\sup_W$ is the supremum over all k-block, $\max(p,q)$-bounded graphons.

Proof. Let $\epsilon = |p-q|$, so $\delta_2(W_1, W_2) = \sqrt{\frac{(k-1)^2}{k^3}\epsilon^2 + \frac{(k-1)}{k^3}\epsilon^2} \geq C\frac{\epsilon}{\sqrt{k}}$, for some constant C. Suppose, for sake of contradiction, that there exists $\hat{W}$ such that

$$\sup_{W\in\mathcal{W}_\rho} \mathbb{E}_{G\sim G_n(W)} \left[\delta_2(\hat{W}(G), W)\right]$$

is not $\Omega\left(\frac{\epsilon}{\sqrt{k}}\right)$. Then there exists a subsequence $\{n_t\}_{t\in\mathbb{N}}$ such that

$$\sup_{W\in\mathcal{W}_\rho} \mathbb{E}_{G\sim G_{n_t}(W)} \left[\delta_2(\hat{W}(G), W)\right] \leq \frac{C}{2t}\frac{\epsilon}{\sqrt{k}}$$

for all $t \in \mathbb{N}$.

The above inequality, combined with Markov's inequality, implies

$$\lim_{t\to\infty} Pr_{G\sim G_{n_t}(W_1)}[\delta_2(\hat{W}(G), W_1) \geq \frac{C}{2}\frac{\epsilon}{\sqrt{k}}] \to 0 \tag{5.16}$$

and

$$\lim_{t\to\infty} Pr_{G\sim G_{n_t}(W_2)}[\delta_2(\hat{W}(G), W_2) \geq \frac{C}{2}\frac{\epsilon}{\sqrt{k}}] \to 0. \tag{5.17}$$

Table 5.1: Known results about contiguity of block graphons and the corresponding Erdös-Rényi model

Condition on p and q for contiguity to hold	Parameter regime	Lower bound on estimation error	Citation
$n(p-q)^2 \leq 2(p+q)$	$p = a/n, q = b/n$ for constants $a, b,\ k = 2$	$\Omega\left(\min\left(\rho, \sqrt{\frac{\rho}{n}}\right)\right)$	[Mossel et al., 2014]
$n(p-q)^2 \leq 2(p+q)$	$\rho n \to \infty$, $\rho n = o(n)$, $k = 2$	$\Omega\left(\min\left(\rho, \sqrt{\frac{\rho}{n}}\right)\right)$	[Banerjee, September 2016]
$\frac{n^2(p-q)^2(k-1)}{p+(k+1)q} \leq 2\log(k-1)$	$p = a/n, q = b/n$ for constants a, b	$\Omega\left(\min\left(\frac{\rho}{\sqrt{k}}, \sqrt{\frac{\rho \log k}{n}}\right)\right)$	[Neeman and Netrapalli, 2014], [Banks et al., 2016]

By Equation 5.17 and the contiguity of W_1 and W_2,

$$\lim_{t\to\infty} Pr_{G\sim G_{n_t}(W_1)}[\delta_2(\hat{W}(G), W_2) \geq \frac{C}{2}\frac{\epsilon}{\sqrt{k}}] \to 0. \tag{5.18}$$

Therefore, Equations 5.16 and 5.18 imply that for large enough n, there exists a graph G such that $\delta_2(\hat{W}(G), W_1) < \frac{C}{2}\frac{\epsilon}{\sqrt{k}}$ and $\delta_2(\hat{W}(G), W_2) < \frac{C}{2}\frac{\epsilon}{\sqrt{k}}$, which implies that $\delta_2(W_1, W_2) < C\frac{\epsilon}{\sqrt{k}}$, which is a contradiction. □

There are many results in the literature exploring when block graphons are contiguous with the corresponding Erdös-Rényi model. Table 5.1 summarises some of the known results in this area and translates them into lower bounds on the graphon estimation problem via Lemma 5.5. Let $\rho = \max(p, q)$.

Part III

Local Differential Privacy for Physical Sensor Measurements

CHAPTER 6

Introduction

Imagine dropping a few drops of ink into a glass of water. The ink drops spread out, forming complicated tendrils that coil back on each other, expanding quickly, until all of the ink has diffused and the liquid is a slightly darker shade than its original colour. There is no physical process by which you can make the diffusing ink coalesce *back* into its original droplets. This intuition is at the heart of what we call **computational cloaking**. Because it is physically impossible to reconstruct the ink droplet exactly, we should be able to hide or keep private in a precise sense its original location. When mathematicians and physicists refer to cloaking, they usually mean transformation optics [Greenleaf et al., 2009], the design of optical devices with special customised effects on wave propagation. In this part, we exploit the ill-conditionedness of inverse problems to design algorithms to release differentially private measurements of the physical system[1].

We are motivated by the explosion in the power and ubiquity of lightweight (thermal, light, motion, etc.) sensors. These data offer important benefits to society. For example, thermal sensor data now plays an important role in controlling HVAC systems and minimising energy consumption in smart buildings [Lin et al., 2002, Beltran et al., 2013]. However, these sensors also collect data inside intimate spaces, homes and workspaces, so the information contained in the data is sensitive. To continue with the example of thermal sensor data, one might consider sources of heat to be people, whose locations we aim to keep private.

Our work indicates that it is possible to produce locally differentially private sensor measurements that both keep the exact locations of the heat sources private and permit recovery of the *general vicinity* of the sources. That is, the locally private data can be used to recover an estimate, $\hat{f}$, that is close to the true source locations, f_0, in the Earth Mover Distance (EMD). This is the second aspect to our work: algorithms that reconstruct sparse signals with error guarantees with respect to EMD (rather than the more traditional ℓ_1 or ℓ_2 error in which accurate recovery is insurmountable).

[1] This Part is based on joint work with Anna Gilbert [Gilbert and McMillan, 2018]

6.1 Source Localization

Suppose that we have a vector f_0 of length n that represents the strengths and positions of our “sources.” The ith entry represents the strength of the source at position i. Further, suppose that we take m **linear** measurements of our source vector; we observe

$$y = Mf_0$$

where M represents some generic linear physical transformation of our original data. Let us also assume that the source vector f_0 consists of at most k sources (or k non-zero entries). The straightforward linear inverse problem is to determine f_0, given M and a noisy version of y. More precisely, given noisy measurements $\tilde{y} = Mf_0 + N(0, \sigma^2 I_m)$, can we produce an estimate $\hat{f}$ that is still *useful*?

For physical processes such as diffusion, intuitively, we can recover the approximate *geographic vicinity* of the source. This is exactly the concept of *closeness* captured by the Earth Mover Distance (EMD). Thus, in this part, we aim to recover $\hat{f}$ that is close to f_0 in the EMD. The EMD can be defined between any two probability distributions on a finite discrete metric space $(\Omega, d(\cdot,\cdot))$. It computes the amount of *work* required to transform one distribution into the other.

Definition 6.1. [Rubner et al., 2000] Let $P = \{(x_1, p_1), \cdots, (x_n, p_n)\}$ and $Q = \{(x_1, q_1), \cdots, (x_n, q_n)\}$ be two probability distributions on the discrete space $\{x_1, \cdots, x_n\}$. Now, let

$$\mathfrak{f}^* = \arg \min_{\mathfrak{f} \in [0,1]^{n \times n}} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathfrak{f}_{ij} d(x_i, x_j) \tag{6.1}$$

$$\text{s.t.} \quad \mathfrak{f}_{ij} \geq 0 \ \ \forall i, j \in [m], \ \ \sum_{j=1}^{n} \mathfrak{f}_{ij} \leq p_i \ \ \forall i \in [m],$$

$$\sum_{i=1}^{n} \mathfrak{f}_{ij} \leq q_i \ \ \forall i \in [n], \ \ \text{and} \ \ \sum_{i=1}^{n} \sum_{i=1}^{n} \mathfrak{f}_{ij} = 1.$$

then $\mathrm{EMD}(P, Q) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathfrak{f}^*_{ij} d(x_i, x_j)$.

6.2 Computational Cloaking Precisely

First, we clarify exactly what *information* we would like to keep private. We consider the coordinates of f_0 to be our data, that is the locations of the sources are what we would like to keep private. We assume that there exists a metric $d(\cdot,\cdot)$ on the set of possible source locations, which induces the EMD on the set of source vectors. For the remainder of this part, we will assume that the metric d is such that every pair of source locations is connected by a path that travels via neighbours.

When the matrix M represents a physical process, we usually cannot hope to keep the *existence* of a source private and also recover an estimation to f_0 that is close in the EMD. However, it may be possible to keep the exact location private while allowing recovery of the "general vicinity" of the source. In fact, we will show in Section 8.2 that this is possible for diffusion on the discrete 1-dimensional line and in Section 8.3 that we can generalise these results to diffusion on a general graph. We are going narrow our definition of "neighbouring" databases to capture this idea.

Definition 6.2. For $\alpha > 0$, two source vectors f_0 and f_0' are α-*neighbours* if

$$\mathrm{EMD}(f_0, f_0') \leq \alpha.$$

The larger α is, the less stringent the neighbouring condition is, so the more privacy we are providing. This definition has two important instances. We can move a source of weight 1 by α units, hiding the location of a large heat source (like a fire) within a small area. Also, we can move a source with weight α by 1 unit, hiding the location that small heat source (like a person) over a much larger area. We will usually drop the α when referring to neighbouring vectors.

A *locally differentially private* algorithm is a private algorithm in which the individual data points are made private before they are collated by the data analyst. In many of our motivating examples the measurements y_i are at distinct locations prior to being transmitted by a data analyst (for example, at the sensors). Thus, the "local" part of the title refers to the fact that we consider algorithms where each measurement, y_i, is made private individually. This is desirable since the data analyst (e.g. landlord, government) is often the entity the consumer would like to be protected against. Also, it is often the case that the data must be communicated via some untrusted channel [Walters et al., 2007, FTC, 2015]. Usually this step would involve encrypting the data, incurring significant computational and communication overhead. However, if the data is made private prior to being sent, then there is less need for encryption. We then wish to use this locally differentially private data to recover an estimate to the source vector that is close in the EMD. The structure of the problem is outlined in the following diagram:

$$\begin{tikzcd}
 & y_1 \arrow[r, "\mathcal{A}"] & \tilde{y_1} & \\
f_0 \arrow[r, "M"] & \left\{ \vdots \right. & \left. \vdots \right\} \arrow[r, "R"] & \hat{f} \\
 & y_m \arrow[r, "\mathcal{A}"] & \tilde{y_m} &
\end{tikzcd}$$

Design algorithms $\mathcal{A}$ and R such that:

1. (Privacy) For all neighbouring source vectors f_0 and f_0', indices i, and Borel measurable sets E we have

$$\mathbb{P}(\mathcal{A}((Mf_0)_i) \in E) \leq e^{\epsilon}\mathbb{P}(\mathcal{A}(Mf_0')_i \in E) + \delta.$$

2. (Utility) $\text{EMD}(f_0, \hat{f})$ is small.

6.3 Related Work

An in-depth survey on differential privacy and its links to machine learning and signal processing can be found in [Sarwate and Chaudhuri, 2013]. The body of literature on general and local differential privacy is vast and so we restrict our discussion to work that is directly related. There is a growing body of literature of differentially private sensor data [Liu et al., 2012, Li et al., 2015, Wang et al., 2016, Jelasity and Birman, 2014, Eibl and Engel, 2016]. Much of this work is concerned with differentially private release of aggregate statistics derived from sensor data and the difficulty in maintaining privacy over a period of time (called the continual monitoring problem).

Connections between privacy and signal recovery have been explored previously in the literature. Dwork et al. [2007] considered the recovery problem with noisy measurements where the matrix M has i.i.d. standard Gaussian entries. Let $x \in \mathbb{R}^n$, $y = Mx \in \mathbb{R}^m$ where $m = \Omega(n)$, $\rho < 0.239$. Suppose y' is a perturbed version of y such that a ρ fraction of the measurements are perturbed arbitrarily and the remaining measurements are correct to within an error of α. Then, Dwork et al. [2007] concludes that with high probability the constrained ℓ_1-minimization, $\min_{x,y} \|y - y'\|_1$ s.t. $Mx = y$, recovers an estimate, $\hat{x}$, s.t. $\|x - \hat{x}\|_1 \leq O(\alpha)$. This is a negative result for privacy. In particular, when $\alpha = 0$ it says that providing reasonably accurate answers to a 0.761 fraction of randomly generated weighted subset sum queries is blatantly non-private. Newer results of Bun et al. [2014] can be interpreted in a similar light where M is a binary matrix. Compressed sensing has also been used in the privacy literature as a way to reduce the amount of noise needed to maintain privacy [Li et al., 2011, Roozgard et al., 2016].

There are also several connections between sparse signal recovery and inverse problems [Farmer et al., 2013, Burger et al., 2010, Haber, 2008, Landa et al., 2011]. The heat source identification problem is severely ill-conditioned and, hence, it is known that noisy recovery is impossible in the common norms like ℓ_1 and ℓ_2. This has resulted in a lack of interest in developing theoretical bounds [Li et al., 2014], thus the mathematical analysis and numerical algorithms for inverse heat source problems are still very limited.

To the best of the author's knowledge, the three papers that are most closely related to this Part are Li et al. [2014], Beddiaf et al. [2015] and Bernstein and Fernandez-Granda [2017]. All these papers attempt to circumvent the condition number lower bounds by changing the error metric to capture “the recovered solution is geographically close to the true solution”, as in this paper. Our algorithm is the same as Li, et al., who also consider the Earth Mover Distance (EMD). Our upper bound is a generalisation of theirs to source vectors with more than one source. Beddiaf, et al. follows a line of work that attempts to find the sources using ℓ_2-minimisation and regularisation. They assume that the number of sources is known in advance, while our algorithm avoids this assumption by instead minimising the ℓ_1 norm to promote sparsity. The error metric they consider is different but related to the EMD, it measures the ℓ_2 distance between the true source locations and the recovered source locations. Experimental results in their paper suggest that their algorithm performs well with noisy data. Their paper however contains no theoretical performance bounds. In work that was concurrent to ours, Bernstein et al. also considered heat source location, framed as deconvolution of the Gaussian kernel. They proved that a slight variant of Basis Pursuit Denoising solves the problem exactly assuming enough sensors and sufficient separability between sources. They also arrive at a similar result to Theorem 8.6 for the noisy case [Bernstein and Fernandez-Granda, 2017, Theorem 2.7].

CHAPTER 7

Privacy of measurements and Ill-conditioned Matrices

7.1 The Private Algorithm

Because we assume that our sensors are lightweight computationally, the algorithm $\mathcal{A}$ is simply the Gaussian mechanism (Lemma 2.8): each sensor adds Gaussian noise locally to its own measurement before sending to the perturbed measurement to the central node. Recall that ill-conditioned linear inverse problems behave poorly under addition of noise. Intuitively, this should mean we need only add a small amount of noise to mask the original data. We show that this statement is partially true. However, there is a fundamental difference between the notion of a problem being ill-conditioned (as defined by the condition number) and being easily kept private. Let M_i be the ith column of M.

Proposition 7.1. *With $\alpha > 0$ and the definition of α-neighbours presented in Definition 6.2, we have*

$$\mathcal{A}(Mf_0) \sim Mf_0 + \frac{2\log(1.25/\delta)\triangle_2(M)}{\epsilon} N(0, I_m)$$

is a (ϵ, δ)-differentially private algorithm where

$$\triangle_2(M) = \alpha \max_{e_i, e_j \textit{ neighbours}} \|M_i - M_j\|_2$$

Proof. Suppose f_0 and f_0' are α-neighbours and let f_{kl} be the optimal flow from f_0 to f_0' (as defined in Definition 6.1) so $f_0 = \sum_{k,l} f_{kl} e_k$ and $f_0' = \sum_{k,l} f_{kl} e_l$, where e_k are the standard basis vectors. Then

$$\begin{aligned} \|Mf_0 - Mf_0'\|_2 &\leq \sum_{k,l} f_{kl} \|Me_k - Me_l\|_2 \\ &= \left(\sum_{k,l} f_{kl} d(e_k, e_l)\right) \max_{i,j\text{neighbours}} \|Me_j - Me_i\|_2 \\ &\leq \alpha \max_{i,j\text{neighbours}} \|Me_i - Me_j\|_2 \end{aligned}$$

Then the fact that the algorithm is (ϵ, δ)-differentially private follows from Lemma 2.8. □

Let $(s_1, \cdots, s_{\min\{n,m\}})$ be the spectrum of M, enumerated such that $s_i \leq s_{i+1}$. The *condition number*, $\kappa_2(M)$, is a measure of how ill-conditioned this inverse problem is. It is defined as

$$\kappa_2(M) := \max_{e,b \in \mathbb{R}^m \setminus \{0\}} \frac{\|M^+ b\|_2}{\|M^+ e\|_2} \frac{\|e\|_2}{\|b\|_2} = \frac{s_{\max\{m,n\}}(M)}{s_1(M)}$$

where M^+ is the pseudo inverse of M. The larger the condition number the more ill-conditioned the problem is Belsley et al. [1980].

The following matrix illustrates the difference between how ill-conditioned a matrix is and how much noise we need to add to maintain privacy. Suppose

$$M = \begin{pmatrix} 1 & 0 \\ 0 & \rho \end{pmatrix}$$

where $\rho < 1$ is small. While this problem is ill-conditioned, $\kappa_2(M) = 1/\rho$ is large, we still need to add considerable noise to the first coordinate of Mx_0 to maintain privacy.

A necessary condition for $\Delta_2(M)$ to be small is that the matrix M is *almost* rank 1, that is, the spectrum should be almost 1-sparse. In contrast the condition that $\kappa_2(M)$ is large is only a condition on the maximum and minimum singular values. The following lemma says that if the amount of noise we need to add, $\triangle_2(M)$, is small then the problem is necessarily ill-conditioned.

Lemma 7.2. *Let M be a matrix such that $\|M\|_2 = 1$ then*

$$\triangle_2(M) \geq \frac{\alpha}{\kappa_2(M)},$$

where α is the parameter in Definition 6.2.

Proof. Suppose e_i is a neighbouring source to e_1 then

$$\begin{aligned} \frac{1}{\kappa_2(M)} &= \min_{\text{rank} E < \min\{m,n\}} \|M - E\|_2 \\ &\leq \|M - [M_j \; M_2 \; M_3 \cdots M_n]\|_2 \\ &\leq \|M_1 - M_i\|_2. \end{aligned}$$

Since we could have replaced 1 and i with any pair of neighbours we have

$$\frac{1}{\kappa_2(M)} \leq \min_{i,j \text{neighbours}} \|M_i - M_j\|_2 = \frac{\Delta_2(M)}{\alpha}.$$

□

The following lemma gives a characterization of $\triangle_2(M)$ in terms of the spectrum of M. It verifies that the matrix M must be *almost* rank 1, in the sense that the spectrum should be dominated by the largest singular value.

Lemma 7.3. *If $\Delta_2(M) \leq \nu$, then $|\|M_i\|_2 - \|M_j\|_2| \leq \frac{\nu}{\alpha}$ for any pair of neighbouring locations e_i and e_j and $|\sum_{i\neq \min\{m,n\}} s_i| \leq \frac{(n+1)^{3/2}\rho\nu}{\alpha}$, where ρ is the diameter of the space of source locations.*

Conversely, if $|\sum_{i\neq \min\{m,n\}} s_i| \leq \frac{\nu}{\alpha}$ and $|\|M_i\|_2 - \|M_j\|_2| \leq \frac{\nu}{\alpha}$ then $\Delta_2(M) \leq 4\nu$.

Proof. Let e_i and e_j be neighbouring sources. Now, assume $\Delta_2(M) \leq \nu$ then

$$|\|M_i\|_2 - \|M_j\|_2| \leq \|M_i - M_j\|_2 \leq \frac{\nu}{\alpha}.$$

Suppose wlog that $\max_i \|M_i\|_2 = \|M_1\|_2$ and let $M' = [M_1 \cdots M_1]$ be the matrix whose columns are all duplicates of the first column of M. Recall that the trace norm of a matrix is the sum of its singular values and for any matrix, $\|M\|_{tr} \leq \sqrt{\min\{m,n\}}\|M\|_F$ and $\|M\|_2 \leq \|M\|_F$. Since M' is rank 1, $\|M'\|_{tr} = \|M'\|_2 = s_{\min\{m,n\}}$, thus,

$$\begin{aligned} |\sum_{i=1}^{\min\{m,n\}-1} s_i| &\leq |\|M\|_{tr} - \|M'\|_{tr}| + |\|M'\|_{tr} - s_{\min\{m,n\}}| \\ &\leq \|M' - M\|_{tr} + |\|M'\|_2 - \|M\|_2| \\ &\leq (\sqrt{\min\{n,m\}} + 1)\|M' - M\|_F \\ &\leq (\sqrt{\min\{n,m\}} + 1)\rho(n-1)\frac{\nu}{\alpha} \end{aligned}$$

Conversely, suppose $|\sum_{i\neq\min\{m,n\}} s_i| \leq \frac{\nu}{\alpha}$ and $|\|M_i\|_2 - \|M_j\|_2| \leq \frac{\nu}{\alpha}$. Using the SVD we know, $M = \sum s_i U_i \otimes V_i$ where U_i and V_i are the left and right singular values, respectively. Thus,

$$\begin{aligned} \|M_i - M_j\|_2 &= \|\sum_k s_k (U_k)_i V_k - \sum_k s_k (U_k)_j V_k\|_2 \\ &\leq s_{\min\{m,n\}}|(U_1)_i - (U_1)_j| + \frac{\nu}{\alpha}. \end{aligned}$$

Also, $\frac{\nu}{\alpha} \geq \|M_i\|_2 - \|M_j\|_2 \geq s_{\min\{m,n\}}(U_1)_i - \frac{\nu}{\alpha} - s_{\min\{m,n\}}(U_1)_j - \frac{\nu}{\alpha}$ so
$|(U_1)_i - (U_1)_j| \leq 3\frac{\nu}{\alpha s_{\min\{m,n\}}}$. □

7.2 Backdoor Access via Pseudo-randomness

It has been explored previously in the privacy literature that replacing a random noise generator with cryptographically secure pseudorandom noise generator in an efficient differentially private algorithm creates an algorithm that satisfies a weaker version of privacy, computational differential privacy [Mironov et al., 2009]. While differential privacy is secure against *any* adversary, computational differential privacy is secure against a *computationally bounded* adversary. In the following definition, κ is a *security* parameter that controls various quantities in our construction.

Definition 7.4 (Simulation-based Computational Differential Privacy (SIM-CDP) Mironov et al. [2009])**.** A family, $\{\mathcal{M}_\kappa\}_{\kappa\in\mathbb{N}}$, of probabilistic algorithms $\mathcal{M}_\kappa : \mathcal{D}^n \to \mathcal{R}_\kappa$ is ϵ_n-SIM-CDP if there exists a family of ϵ_κ-differentially private algorithms $\{\mathcal{A}_n\}_{n\in\mathbb{N}}$, such that for every probabilistic polynomial-time adversary $\mathcal{P}$, every polynomial $p(\cdot)$, every sufficiently large $\kappa \in \mathbb{N}$, every dataset $D \in \mathcal{D}^n$ with $n \leq p(\kappa)$, and every advice string z_κ of size at $p(\kappa)$, it holds that,

$$|\mathbb{P}[\mathcal{P}_\kappa(\mathcal{M}_\kappa(D) = 1) - \mathbb{P}[\mathcal{P}_\kappa(\mathcal{A}_\kappa(D) = 1)| \leq \text{negl}(\kappa).$$

That is, $\mathcal{M}_\kappa$ and $\mathcal{A}_\kappa$ are computationally indistinguishable.

The transition to pseudo-randomness, of course, has the obvious advantage that pseudo-random noise is easier to generate than truly random noise. In our case, it also has the additional benefit that, given access to the seed value, pseudo-random noise can be removed, allowing us to build a "backdoor" into the algorithm. Suppose we have a trusted data analyst who wants access to the most accurate measurement data, but does not have the capacity to protect sensitive data from being intercepted in transmission. Suppose also that this party stores the seed value of each sensor and the randomness in our locally private algorithm $\mathcal{A}$ is replaced with pseudo-randomness. Then, the consumers are protected against an eavesdropping computationally bounded adversary, and the trusted party has access to the noiseless [1] measurement data. This solution may be preferable to simply encrypting the data during transmission since there may be untrusted parties who we wish to give access to the private version of the data.

Corollary 7.5 (Informal)**.** *Replacing the randomness in Proposition 8.5 with pseudo-randomness produces a local simulation-based computational differentially private algorithm for the same task. In addition, any trusted party with access to the seed of the random number generator can use the output of the private algorithm to generate the original data.*

[1] This data may still be corrupted by sensor noise that was not intentionally injected

CHAPTER 8

Recovery algorithm and Examples

We claimed that the private data is both useful and differentially private. In this Chapter we discuss recovering an estimate of f_0 from the noisy data $\tilde{y}$. Algorithms for recovering a sparse vector from noisy data have been explored extensively in the compressed sensing literature. However, theoretical results in this area typically assume that the measurement matrix M is sufficiently nice. Diffusion matrices are typically very far from satisfying the *niceness* conditions required for current theoretical results. Nonetheless, in this Chapter we discuss the use of a common sparse recovery algorithm, Basis Pursuit Denoising (BPD), for ill-conditioned matrices. The use of BPD to recover source vectors with the heat kernel was proposed by Li et al. [2014], who studied the case of a 1-sparse source vector.

We begin with a discussion of known results for BPD from the compressed sensing literature. While the theoretical results for BPD do not hold in any meaningful way for ill-conditioned diffusion matrices, we present them here to provide context for the use of this algorithm to recover a *sparse* vector. We then proceed to discussing the performance of BPD on private data in some examples: diffusion on the 1D unit interval and diffusion on general graphs.

8.1 Basis Pursuit Denoising

Basis Pursuit Denoising minimises the ℓ_1-norm subject to the constraint that the measurements of the proposed source vector $\hat{f}$ should be close in the ℓ_2-norm to the noisy sensor measurements. To simplify our discussion, let σ be the standard deviation of the noise added to the sensor measurements. The bound $\sigma\sqrt{m}$ in Algorithm 1 is chosen to ensure f_0 is a feasible point with high probability.

The bound $\sigma\sqrt{m}$ in Algorithm 1 is chosen to ensure f_0 is a feasible point with high probability.

```
Algorithm 1 $R$: Basis Pursuit Denoising
  Input: $M, \sigma > 0, \tilde{y}$
  $\hat{f} = \arg\min_{f\in[0,1]^n} \|f\|_1$   s.t. $\|Mf - \tilde{y}\|_2 \leq \sigma\sqrt{m}$
  Output: $\hat{f} \in [0,1]^n$
```

Lemma 8.1. *Hsu et al. [2012] Let $\nu \sim N(0, \sigma^2 I_m)$ then for all $t > 0$,*

$$\mathbb{P}[\|\nu\|_2^2 > \sigma^2(m + 2\sqrt{mt} + 2t)] \leq e^{-t}.$$

So for large m and small ρ, we have $\|\nu\|_2 \leq (1+\rho)\sigma\sqrt{m}$ with high probability.

8.1.1 Basis Pursuit Denoising for RIP matrices

In order to present the results in this section cleaner, rather than keeping track of $\sigma\sqrt{m}$ we introduce parameters $\alpha, \beta > 0$. Basis Pursuit Denoising

$$\arg\min_{f\in[0,1]^n} \|f\|_1 \quad \text{s.t. } \|Mf - \tilde{y}\|_2 \leq \alpha \tag{8.1}$$

is the convex relaxation of the problem we would like to solve, ℓ_0-minimisation:

$$\arg\min_{f\in[0,1]^n} \|f\|_0 \quad \text{s.t. } \|Mf - \tilde{y}\|_2 \leq \beta. \tag{8.2}$$

The minimum of the ℓ_0 norm is the *sparsest* solution. Unfortunately, this version of the problem is NP hard, so in order to find an efficient algorithm we relax to the ℓ_1 norm. The ℓ_1 norm is the "smallest" convex function that places a unit penalty on unit coefficients and zero penalty on zero coefficients. Since the relaxation is convex, we can use convex optimisation techniques to solve it. In the next section we'll discuss an appropriate optimisation algorithm. In this section, we focus on when the solution to the relaxed version (8.1) is similar to the solution for Equation (8.2).

We call the columns of M, denoted by M_i, atoms. We will assume for this section that $\|M_i\|_2 = 1$ for all i. Notice that the vector Mf_0 is the linear combination of the M_i with coefficients given by the entries of f_0 so we can think of recovering the vector f_0 as recovering the coefficients[1]. A key parameter of the matrix M is its *coherence*:

$$\mu = \max_{i,j} |\langle M_i, M_j\rangle|$$

Similar to $\triangle_2(M)$, the coherence is a measure how similar the atoms of M. The larger the coherence is, the more similar the atoms are, which makes them difficult to distinguish. For accurate sparse recovery, it is preferential for the coherence to be small. The following theorem relates the solutions to Equation (8.1) and (8.2).

[1] This is where BPD gets its name. We are pursuing the basis vectors that make up Mf_0.

Theorem 8.2. *[Tropp, 2004, Theorem C] Suppose $k \leq \frac{1}{3}\mu^{-1}$. Suppose f_{opt} is a k-sparse solution to Equation* (8.2) *with $\beta = \frac{\alpha}{\sqrt{1+6k}}$. Then the solution $\hat{f}$ produced from Algorithm 1 satisfies*

- *supp$(\hat{f})$ $\subset$supp(f_{opt})*
- *$\hat{f}$ is no sparser than a solution to Equation* (8.2) *with $\beta = \alpha$*
- $\|\hat{f} - f_{opt}\|_2 \leq \alpha\sqrt{3/2}$

Theorem 8.2 says that if the matrix M is coherent then the solution to the convex relaxation (Algorithm 1) is at least as sparse as a solution to (8.2) with error tolerance somewhat smaller than α. Also, $\hat{f}$ only recovers source locations that also appear in f_{opt}, although it may not recover all of the source locations that appear in f_{opt}. The final property bounds the weight assigned to any source identified in f_{opt} and not $\hat{f}$. If $\tilde{y} = Mf_0 + Z$ then the worst case discrepancy between f_0 and f_{opt} occurs when Z concentrates its weight on a single atom. In our case, the noise vector Z has i.i.d. Gaussian coordinates and hence is unlikely to concentrate its weight.

The key property for *exact* recovery of f_0, rather than f_{opt}, is that M is a near isometry on sparse vectors. A matrix M satisfies the *Restricted Isometry Property (RIP)* of order k with restricted isometry constant δ_k if δ_k is the smallest constant such that for all k-sparse vectors x,

$$(1 - \delta_k)\|x\|_2^2 \leq \|Mx\|_2^2 \leq (1 + \delta_k)\|x\|_2^2.$$

If f_0 is a feasible point and δ_k is small, then we can guarantee that f_0 and $\hat{f}$ are close in the ℓ_2 norm.

Theorem 8.3. *[Candès, 2008, Theorem 1.2] Assume $\delta_{2k} < \sqrt{2} - 1$ and $\|Z\|_2 \leq \alpha$ and f_0 is k-sparse. Then the solution $\hat{f}$ to* (8.1) *satisfies*

$$\|\hat{f} - f_0\|_2 \leq C\alpha$$

for some constant C.

The exact constant C is given explicitly in Candès [2008] and is rather small. For example, when $\delta_{2k} = 0.2$, we have $C \leq 8.5$.

Theorems 8.2 and 8.3 only provide meaningful results for matrices with small μ and δ_k. Unfortunately, the coherence and restricted isometry constants for ill-conditioned matrices, and in particular diffusion matrices, are both large. It is somewhat surprising then that BPD recovers well in the examples we will explore in the following sections.

8.1.2 Bregman Iteration

There has been considerable work on methods for solving convex constrained optimisation problems. Many of the proposed algorithms for Basis Pursuit Denoising are variations on *Bregman iteration* which we present in this section [Yin et al., 2008]. The more difficult part of the optimisation is converting from constrained optimisation to unconstrained optimisation, which we have more tools to solve.

Bregman iteration techniques are used to solve problems of the form

$$\arg\min_f J(f) \text{ s.t. } H(f) \leq \alpha$$

where J and H are both convex and $\min_f H(f) = 0$. The Bregman divergence is

$$D_J^p(f, f') = J(f) - J(f') - \langle x, f - f' \rangle, \;\; x \in \partial J(f')$$

where $\partial J(f')$ is the set of subdifferentials of J at f'. The Bregman divergence measures the gap between the function J and its tangent plane at f'.

Algorithm 2 Bregman Iteration

```
**Input:** $M, \lambda, \tilde{y}$
Initialise $u^0 = 0, p^0 = 0$.
**for** $t = 0, 1, \cdots$ **do**
    $f^{t+1} = \arg\min_f D_J^{p^t}(f, f^t) + \lambda H(f)$
    $p^{t+1} = p^k - \lambda \nabla H(u^{k+1})$
**end for**
```

A primary feature of Bregman divergences is that for any convex set C, the map $y \mapsto \arg\min_{f \in C} D_p(f, y)$ is a projection onto the set C. The update $f^{t+1} = \arg\min_f D_J^{p^t}(f, f^t) + \lambda H(f)$ is a regularised version of projection. It approximates the constrained problem $\min D^p J(f, f')$ s.t. $H(f) \leq \alpha$ by adding the objective function and the constraint together and attempting to minimise both simultaneously. The second update $p^{t+1} = p^k - \lambda \nabla H(u^{k+1})$ updates the subdifferential.

For Basis Pursuit Denoising, $J(f) = \|f\|_1$ and $H(f) = \|Mf - \tilde{f}\|_2$ and we can rewrite the updates as $f^0 = 0$ and $p^0 = \tilde{y}$,

$$f^{t+1} = \arg\min_f \|f\|_1 + \lambda \|Mf - p^t\|_2^2$$

$$p^{t+1} = p^t + (\tilde{y} - Mf^{t+1})$$

There are many methods for solving for f^{t+1}. We refer the interested reader to Li et al. [2014] for a discussion of a greedy method designed to find sparse solutions efficiently. Finally, the following theorem says Bregman iteration will converge to the correct answer eventually, although we have no guarantee on how quickly.

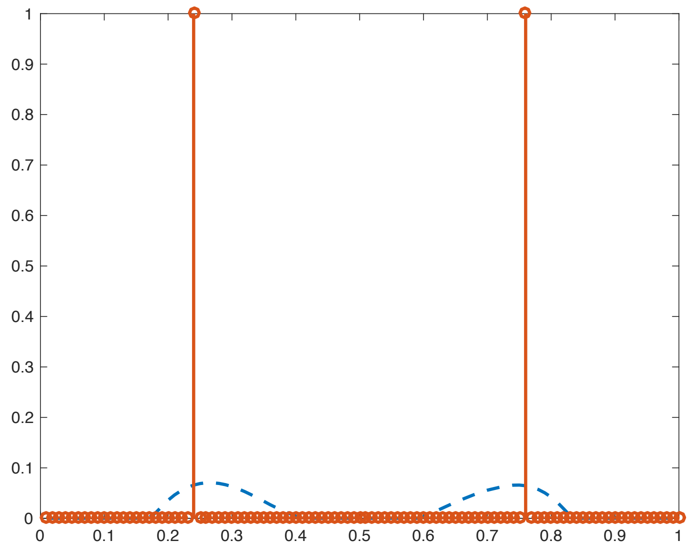


Figure 8.1: f_0 has unit peaks 0.76 and 0.24. The parameters are $n = 100$, $m = 50$, $T = 0.05$ and $\sigma = 0.1$. The red line is f_0, the blue (dashed) line is $\hat{f}$.

Theorem 8.4. *The constraint H monotonically decreases*

$$H(f^{t+1}) \leq H(f^t).$$

Also, if f_0 is a feasible point then as long as $\|\tilde{y} - Mf^{t+1}\|_2 > \alpha$ we have

$$D_J^{p^{t+1}}(f_0, f^{t+1}) < D_J^{p^t}(f_0, f^t)$$

For strictly convex J, the last statement implies that f^{t+1} is moving closer to f_0.

8.2 Diffusion on the Unit Interval

Let us define the linear physical transformation explicitly for heat source localization. To distinguish this special case from the general, we denote the measurement matrix by A (instead of M). For heat diffusion, we have a diffusion constant μ and a time t at which we take our measurements. Let $T = \mu t$ in what follows. Let $g(x, t) = \frac{1}{\sqrt{4\pi T}} e^{\frac{-|x|^2}{4T}}$. Let $n > 0$ and suppose the support of f is contained in the discrete set $\{\frac{1}{n}, \cdots, 1\}$. Let $m > 0$ and suppose we take m measurements at locations $\frac{1}{m}, \cdots, 1$ so y_i is the measurement of the sensor at location $\frac{i}{m}$ at time t and we have

$$y = Af_0 \quad \text{where} \quad A_{ij} = g\left(\frac{i}{n} - \frac{j}{m}, t\right).$$

The heat kernel, A, is severely ill-posed due to the fact that as heat dissipates, the measurement vectors for different source vectors become increasingly close Weber [1981]. Figure 8.1 shows the typical behaviour of Algorithm 1 with the matrix A. As can be seen in the

Table 8.1: Asymptotic upper bounds for private recovery assuming $\sqrt{T}(\frac{\sqrt{m}}{nT^{1.5}} + ke^{\frac{-\alpha^2}{4T}}) \leq c < 1$.

VARIABLE	EMD$\left(\frac{f_0}{\|f_0\|_1}, \frac{\hat{f}}{\|\hat{f}\|_1}\right)$
n	$O(1 + \frac{1}{\sqrt{n}})$
m	$O(1)$
T	$\min\{1,\ O(1 + \frac{1}{T} + T^{2.5}e^{-\alpha^2/4T})\}$

figure, this algorithm returns an estimate $\hat{f}$ that is indeed close to f_0 in the EMD but not close in more traditional norms like the ℓ_1 and ℓ_2 norms. This phenomenon was noticed by Li et al. [2014], who proved that if f_0 consists of a single source then EMD$(f_0, \hat{f})$ is small where $\hat{f} = R(\tilde{y})$.

Proposition 8.5. *With the definition of neighbours presented in Definition 6.2 and restricting to $f_0 \in [0,1]^n$ we have*

$$\triangle_2(A) = O\left(\frac{\alpha\sqrt{m}}{T^{1.5}}\right)$$

Proof. For all $i \in [n]$ we have

$$\begin{aligned}
\|A_i - A_{i+1}\|_2^2 &= \frac{1}{4\pi T}\sum_{j=1}^{m}\left(e^{\frac{-(\frac{i}{n}-\frac{j}{m})^2}{4T}} - e^{\frac{-(\frac{i+1}{n}-\frac{j}{m})^2}{4T}}\right)^2 \\
&= \frac{1}{4\pi T}\sum_{j=1}^{m} e^{\frac{-(\frac{i}{n}-\frac{j}{m})^2}{2T}}\left(1 - e^{\frac{(\frac{i}{n}-\frac{j}{m})^2-(\frac{i+1}{n}-\frac{j}{m})^2}{4T}}\right)^2 \\
&\leq \frac{1}{4\pi T}\max_{i\in[n]}\max_{j\in[m]}\left(1 - e^{\frac{(\frac{i}{n}-\frac{j}{m})^2-(\frac{i+1}{n}-\frac{j}{m})^2}{4T}}\right)^2 \sum_{j=1}^{m} e^{\frac{-(\frac{i}{n}-\frac{j}{m})^2}{2T}}
\end{aligned}$$

Now, $\sum_{j=1}^{m} e^{\frac{-(\frac{i}{n}-\frac{j}{m})^2}{2T}} \leq m$ and

$$\max_{i\in[n]}\max_{j\in[m]}\left(1 - e^{\frac{(\frac{i}{n}-\frac{j}{m})^2-(\frac{i+1}{n}-\frac{j}{m})^2}{4T}}\right)^2 \leq \max\{(1-e^{\frac{-3}{4nT}})^2, (1-e^{\frac{2}{4nT}})^2\} = O\left(\frac{1}{n^2T^2}\right).$$

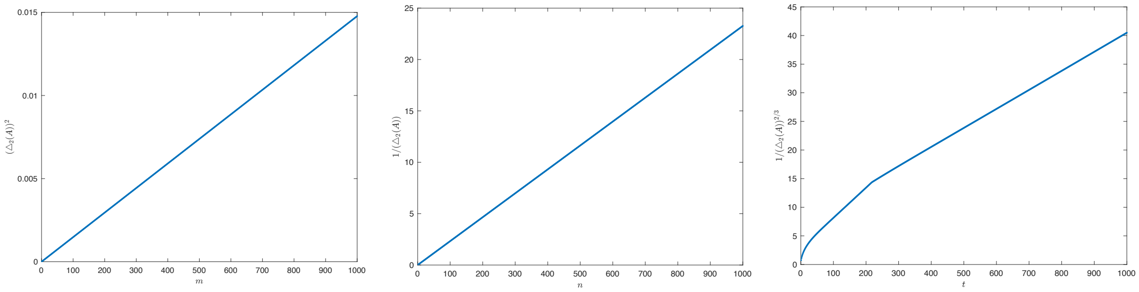
Therefore,

$$\|A_i - A_{i+1}\|_2 = O\left(\frac{\sqrt{m}}{nT^{1.5}}\right).$$

□

Figure 8.2 shows calculations of $\triangle_2(A)$ with varying parameters. The vertical axes are scaled to emphasise the asymptotics. These calculations suggest that the analysis in Proposition 8.5 is asymptotically tight in m, n and T.

Theorem 8.6. *Suppose that f_0 is a source vector, $\hat{y} = R(\tilde{y})$ and assume the following:*



(a) Dependence on m

(b) Dependence on n

(c) Dependence on t

Figure 8.2: Empirical results of computation of $\triangle_2(A)$. Unless specified otherwise, $m = 500$ and $t = 0.1$. In (8.2a), $n = 500$ and in (8.2c), $n = 1000$.

1. $m\sqrt{T/2} > 1$

2. $\sqrt{2T} < 1$

3. $|x_i - x_j| > \sqrt{2T} + 2A$ *for some* $A > 0$

then w.h.p.

$$\textit{EMD}\left(\frac{f_0}{\|f_0\|_1}, \frac{\hat{f}}{\|\hat{f}\|_1}\right) \leq \min\left\{1,\ O\left[\frac{1}{1-\min\{1,C\}}\left(\frac{1}{k}\sqrt{\frac{T^{1.5}C}{\sqrt{T}+1}} + k\min\{1,C\} + \frac{T^2C}{(T+1)k}\right)\right]\right\}$$

where $C = \min\left\{k,\ \sqrt{T}\left[\sigma + ke^{-A^2/4T}\right]\right\}$.

Assumptions 1 and 2 state that m needs to be large enough that for each possible source and we need to take the measurements before the heat diffuses too much. Assumption 3 says that the sources need to be sufficiently far apart. We can remove this assumption by noting that every source vector is close to a source vector whose sources are well separated and that for all f, f', $\|Af_0 - Af_0'\|_2 = O\left(\frac{\sqrt{m}}{T^{1.5}}\text{EMD}(f_0, f_0')\right)$.

The result is a generalisation to source vectors with more than one source of a result of Li et al. [2014] . Our proof is a generalisation of their proof and is contained in Section 8.4. In order to obtain a recovery bound for the private data, we set $\sigma = \left(\frac{\sqrt{m}}{nT^{1.5}}\right)$. The asymptotics of this bound are contained in Table 8.1. It is interesting to note that, unlike in the constant σ case, the error increases as $T \to 0$ (as well as when $T \to \infty$). This is because as $T \to 0$ the inverse problem becomes less ill-conditioned so we need to add more noise.

The following theorem gives a lower bound on the estimation error of the noisy recovery problem.

Theorem 8.7. *We have*

$$\inf_{\hat{f}} \sup_{f_0} \mathbb{E}[\textit{EMD}(f_0, \hat{f})] = \Omega\left(\min\left\{\frac{1}{2}, \frac{T^{1.5}\sigma}{\sqrt{m}}\right\}\right).$$

where $\inf_{\hat{f}}$ *is the infimum over all estimators* $\hat{f} : \mathbb{R}^m \to [0,1]^n$*,* $\sup_{f_0}$ *is the supremum over all source vectors in* $[0,1]^n$ *and* $\tilde{y}$ *is sampled from* $y + N(0, \sigma^2 I_m)$*.*

Note that this lower bound matches our upper bound asymptotically in σ and is slightly loose in T. It varies by a factor of $\sqrt{m}$ from our theoretical upper bound. Experimental results (contained in the extended version) suggest that the error decays like $O(1 + \frac{1}{\sqrt{m}})$. A consequence of Theorem 8.7 is that if two peaks are too close together, roughly at a distance of $O\left(\min\{\frac{1}{2}, \frac{T^{1.5}\sigma}{\sqrt{m}}\right)$, then it is impossible for an estimator to differentiate between the true source vector and the source vector that has a single peak located in the middle. Before we prove Theorem 8.7 we need following generalisation of the upper bound in Proposition 8.5.

Lemma 8.8. *Suppose* $\|f_0\|_1 = \|f_0'\|_1 = 1$ *then*

$$\|Af_0 - Af_0'\|_2 = O\left(\frac{\sqrt{m}}{T^{1.5}}\textit{EMD}(f_0, f_0')\right)$$

Proof. Firstly, consider the single peak vectors e_i and e_j. Then noting that $Ae_i = A_i$, we have from Proposition 8.5 that

$$\|Ae_i - Ae_j\|_2 \leq \sum_{l=0}^{j-i-1} \|Ae_{i+l} - Ae_{i+1+1}\|_2 \leq O\left(|i-j|\frac{\sqrt{m}}{nT^{1.5}}\right)$$

Now, let f_{ij} be the optimal flow from f_0 to f_0' as described in Definition 6.1 so $f_0 = \sum_{i,j} f_{ij}e_i$ and $f_0' = \sum_{ij} f_{ij}e_j$. Then

$$\begin{aligned}\|Af_0 - Af_0'\|_2 &\leq \sum_{ij} f_{ij}\|Ae_i - Ae_j\|_2 \\ &\leq O\left(\sum_{ij} f_{ij}\left|\frac{i}{n} - \frac{j}{n}\right|\frac{\sqrt{m}}{T^{1.5}}\right) \\ &= O\left(\frac{\sqrt{m}}{T^{1.5}}\text{EMD}(f_0, f_0')\right)\end{aligned}$$

□

Proof of Theorem 8.7. For any source vector f_0, let P_{f_0} be the probability distribution induced on $\mathbb{R}^m$ by the process $Af_0 + N(0, \sigma^2 I_m)$. Then the inverse problem becomes estimating which distribution P_{f_0} the perturbed measurement vector is sampled from. Let f_0

and f_0' be two source vectors. Then

$$\begin{aligned} D(P_{f_0}||P_{f_0'}) &= \sum_{i=1}^{m} \frac{((Af_0)_i - (Af_0')_i)^2}{2\sigma^2} \\ &= \frac{1}{2\sigma^2}\|Af_0 - Af_0'\|_2^2 \\ &\leq C\frac{m}{T^3\sigma^2}(\text{EMD}(f_0, f_0'))^2 \end{aligned}$$

for some constant C, where we use the fact that the KL-divergence is additive over independent random variables, along with Lemma 8.8. Now, let $a = \min\{\frac{1}{2}, \frac{T^{1.5}\sigma}{\sqrt{2C}\sqrt{m}}\}$. Let T be the set consisting of the following source vectors: $e_{1/2}$, $(1/2)e_{1/2-a/2} + (1/2)e_{1/2+a/2}$, $(1/4)e_{1/2-a} + (1/2)e_{1/2} + (1/4)e^{1/2+a}$, $(1/2)e_{1/2} + (1/2)e_{1/2+a}$, which are all at an EMD a from each other. Then $d_{KL}(T) + 1 \leq 3/2$ and $\log \mathcal{M}(a, T, \text{EMD}) = 2$. Thus, by Lemma 3.2,

$$\inf_{\hat{f}} \sup_{f_0} \mathbb{E}[\text{EMD}(f_0, \hat{f})] \geq \frac{3}{4}a = \Omega\left(\min\left\{\frac{1}{2}, \frac{T^{1.5}\sigma}{\sqrt{m}}\right\}\right).$$

□

8.3 Diffusion on Graphs

In this section we generalise to diffusion on an arbitrary graph. As usual, our aim is to protect the exact location of a source, while allowing the neighbourhood to be revealed. Diffusion on graphs models not only heat spread in a graph, but also the path of a random walker in a graph and the spread of rumours, viruses or information in a social network. A motivating example is *whisper* networks where participants share information that they would not like attributed to them. We would like people to be able to spread information without fear of retribution, but also be able to approximately locate the source of misinformation. The work in this section does not directly solve this problem since in our setting each node's data corresponds to their *probability* of knowing the rumour, rather than a binary *yes/no* variable. In future work, we would like to extend this work to designing whisper network systems with differential privacy guarantees. If a graph displays a community structure, then we would like to determine which community the source is in, without being able to isolate an individual person within that community.

Let G be a connected, undirected graph with n nodes. The $n \times n$ matrix W contains the edge weights so $W_{ij} = W_{ji}$ is the weight of the edge between node i and node j and the diagonal matrix D has D_{ii} equal to the sum of the i-th row of W. The graph Laplacian is $L = D - W$. As above, we also have a parameter controlling the rate of diffusion τ. Then if the initial distribution is given by f_0 then the distribution after diffusion is given by

Table 8.2: Examples of $\Delta_2(A_G)$ for some standard graphs. Each graph has n nodes.

G	$\Delta_2(A_G)^2$
COMPLETE GRAPH	$2e^{-\tau n}$
STAR GRAPH	$e^{-2\tau n} + \left(\frac{e^{-\tau n} - e^{-\tau}}{n-1}\right)^2 + \left(\frac{e^{-\tau n} - e^{-\tau}}{n-1} + e^{-\tau}\right)^2$

the linear equation $y = e^{-\tau L} f_0$ Thanou et al. [2017]. We will use A_G to denote the matrix $e^{-\tau L}$. Note that, unlike in the previous section, we have no heat leaving the domain (i.e., the boundary conditions are different).

The graph G has a metric on the nodes given by the shortest path between any two nodes. Recall that in Lemma 7.3 we can express the amount of noise needed for privacy, $\Delta_2(A_G)$, in terms of the spectrum of A_G. Let $s_1 \leq s_2 \leq \cdots \leq s_{\min\{n,m\}}$ be the eigenvalues of L then $e^{-\tau s_1} \geq \cdots \geq e^{-\tau s_{\min\{m,n\}}}$ are the eigenvalues of A_G. For any connected graph G, the Laplacian L is positive semidefinite and 0 is a eigenvalue with multiplicity 1 and eigenvector the all-ones vector.

Lemma 8.9. *For any graph G,*

$$\Delta_2(A_G) \;\leq\; \sum_{k=2}^{\min\{n,m\}} e^{-\tau s_i} |(U_i)_k - (U_j)_k|,$$

where U_i is the ith row of the matrix whose columns are the left singular vectors of L.

Proof. With set-up as in Lemma 7.3 we have

$$\begin{aligned} \|(A_G)_i - (A_G)_j\|_2 &= \| \sum_k e^{-\tau s_k} ((U_k)_i - (U_k)_j) V_k \|_2 \\ &\leq \sum_k e^{-\tau s_k} |(U_k)_i - (U_k)_j| \end{aligned}$$

Since the first eigenvector of L is the all ones vector, we $|(U_1)_i - (U_1)_j| = 0$. □

An immediate consequence of Lemma 8.9 is that $\Delta_2(A_G)$ is bounded above by

$$e^{-\tau s_2} \|U_i \;-\; U_j\|_1.$$

The second smallest eigenvalue of L, s_2, (called the *algebraic connectivity*) is related to the connectivity of the graph, in particular the graphs expanding properties, maximum cut, diameter and mean distance [Mohar, 1991]. As the graph becomes more connected, the rate of diffusion increases so the amount of noise needed for privacy decreases. The dependence on the rows of the matrix of left singular vectors is intriguing as these rows arise in several other areas of numerical analysis. Their ℓ_2 norms are called leverage scores Drineas et al. [2012] and they appear in graph clustering algorithms.

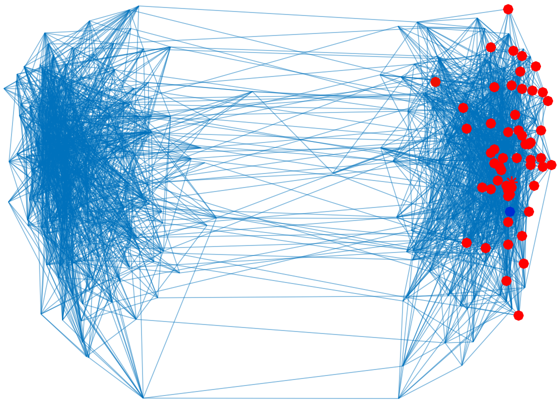


Figure 8.3: f_0 has a unit peak at one of the nodes on the left side. The graph G was drawn from a stochastic block model with intercommuntiy probability 5% and intracommunity probability 0.1%. The parameters are $\tau = 2, n = 500, \delta = 0.1, \epsilon = 4$.

Figure 8.3 shows the average behaviour of Algorithm 1 on a graph with community structure. Preliminary experiments suggest that provided τ is not too large or too small and ϵ is not too small, the correct community is recovered.

8.4 Appendix

We need some machinery before we can prove Theorem 8.6. The following lemma is from Li et al. [2014]. Since $T = \mu t$ is fixed we will let $g(x) = g(x, t)$.

Lemma 8.10. *[Li et al., 2014] Suppose $s_1 < x < s_2$ and $|s_1 - s_2| \leq \sqrt{2T}$ and consider the function $W(z) = -g'(s_2 - x)g(z - s_1) - g'(x - s_1)g(s_2 - z)$. Then $W(z)$ has a single maximum at x and*

$$W(x) - W(z) \quad \begin{cases} > W(x) - W(s_2 - \sqrt{2T}) & \textit{for } z \leq s_2 - \sqrt{2T} \\ \geq C_1 \|z - x\|_2^2 & \textit{for } z \in [s_2 - \sqrt{2T}, s_1 + \sqrt{2T}] \\ > W(x) - W(s_1 + \sqrt{2T}) & \textit{for } z \geq s_1 + \sqrt{2T} \end{cases}$$

where $C_1 = \inf_{z \in [s_2 - \sqrt{2T}, s_1 + \sqrt{2T}]} [-W''(z)/2] > 0$.

The following two lemmas are necessary for our proof of Theorem 8.6. For all $i \in [k]$ and $j \in [p]$, let $W_{ij}(z) = -g'(s_{i_{j+p}} - x_i)g(z - s_{i_j}) - g'(x_i - s_{i_j})g(s_{i_{j+p}} - z)$. Let $p = m\sqrt{T/2}$. We will often replace the distance between s_{i_j} and $s_{i_{j+p}}$ with $\sqrt{T/2}$ since it is asymptotically equal to the true distance $p/m = \lfloor m\sqrt{T/}\rfloor/m$ in m.

Lemma 8.11. *Using the assumptions of Theorem 8.6 we have*

$$\sum_{j=1}^{p} \inf_{z \in [s_{i_{j+p}} - \sqrt{2T}, s_{i_j} + \sqrt{2T}]} [-W''_{ij}(z)/2] \geq \Omega\left(\frac{m\sqrt{T/8} + 1}{T^{2.5}}\right)$$

Proof. Note first that $s_{i_{j+p}} - x_1 = \sqrt{T/2} - (x_i - s_{i_j})$ and $s_{i_{j+p}} - z = \sqrt{T/2} - (z - s_{i_j})$ for any $z \in [s_{i_{j+p}} - \sqrt{2T}, s_{i_j} + \sqrt{2T}]$. Let $z \in [s_{i_{j+p}} - \sqrt{2T}, s_{i_j} + \sqrt{2T}]$ then

$$\begin{aligned}
&- W_{ij}''(z) = g'(s_{i_{j+p}} - x_i)g''(z - s_{i_j}) + g'(x_i - s_{i_j})g''(s_{i_{j+p}} - z) \\
&= \frac{1}{16\pi T^3}\Bigg[(s_{i_{j+p}} - x_i)\left(1 - \frac{(z - s_{i_j})^2}{4T}\right)e^{\frac{-(s_{i_{j+p}}-x_i)^2-(z-s_{i_j})^2}{4T}} \\
&\qquad\qquad + (x_i - s_{i_j})\left(1 - \frac{(s_{i_{j+p}} - z)^2}{4T}\right)e^{\frac{-(x_i-s_{i_j})^2-(s_{i_{j+p}}-z)^2}{4T}}\Bigg] \\
&\geq \frac{1}{2T\sqrt{4\pi T}}\frac{1}{2T\sqrt{4\pi T}}e^{\frac{-5}{8}}\Bigg[(s_{i_{j+p}} - x_i)\left(1 - \frac{(z - s_{i_j})^2}{4T}\right) \\
&\qquad\qquad + (x_i - s_{i_j})\left(1 - \frac{(s_{i_{j+p}} - z)^2}{4T}\right)\Bigg] \\
&\geq \frac{e^{\frac{-5}{8}}}{16\pi T^3}\min\{(s_{i_{j+p}} - x_i), (x_i - s_{i_j})\}\left(2 - \frac{(z - s_{i_j})^2}{4T} - \frac{(s_{i_{j+p}} - z)^2}{4T}\right) \\
&\geq \frac{e^{\frac{-5}{8}}}{16\pi T^3}\min\{(s_{i_{j+p}} - x_i), (x_i - s_{i_j})\}\frac{3}{4}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{j=1}^{p} \inf_{z\in[s_{i_{j+p}}-\sqrt{2T}, s_{i_j}+\sqrt{2T}]}[-W_{ij}''(z)/2] &\geq \frac{3e^{\frac{-5}{8}}}{64\pi T^3}\sum_{j=1}^{p}\min\{(s_{i_{j+p}} - x_i), (x_i - s_{i_j})\} \\
&= \frac{3e^{\frac{-5}{8}}}{64\pi T^3}2\sum_{i=1}^{p/2}\frac{i}{m} \\
&= \frac{3e^{\frac{-5}{8}}}{64\pi T^3}2\frac{m\sqrt{T/8}(m\sqrt{T/8}+1)}{2m}
\end{aligned}$$

□

Lemma 8.12. *Using the assumptions of Theorem 8.6 we have*

$$\begin{aligned}
&\min_{i\in[k]}\min_{l:l/n\notin S_i}\sum_{j=1}^{p}(W_{ij}(x_i) - W_{ij}(l/n)) \\
&\qquad = \Omega\left(\frac{m\sqrt{T/2}(m\sqrt{T/2}+1)^2}{m^2}\frac{1}{T^{3.5}}\right).
\end{aligned}$$

Proof. From Lemma 8.10 we know for all l s.t. $l/n \notin S_i$ we have

$$W_{ij}(x_i) - W_{ij}(l/n) \geq \min\{W_{ij}(x_i) - W_{ij}(s_{i_{j+p}} - \sqrt{2T}), W_{ij}(x_i) - W_{ij}(s_{i_j} + \sqrt{2T})\}.$$

Let's start with

$$\begin{aligned}
&W_{ij}(x_i) - W_{ij}(s_{i_{j+p}} - \sqrt{2T}) \\
&\quad = -g'(s_{i_{j+p}} - x_i)\left(g(x_i - s_{i_j}) - g(s_{i_{j+p}} - \sqrt{2T} - s_{i_j})\right) \\
&\qquad\qquad - g'(x_i - s_{i_j})\left(g(s_{i_{j+p}} - x_i) - g(\sqrt{2T})\right).
\end{aligned}$$

Now, $g(z)$ is concave down for $z \in [-\sqrt{2T}, \sqrt{2T}]$ and λ-strongly concave on the interval $[-\sqrt{T/2}, \sqrt{T/2}]$ with $\lambda = \frac{-7}{16\sqrt{4\pi}T^{1.5}}e^{\frac{-1}{8}}$ so

$$g(x) - g(y) \begin{cases} \geq -g'(x)(y-x) & \text{for } x, z \in [-\sqrt{2T}, \sqrt{2T}] \\ \geq -g'(x)(y-x) + \frac{7}{32\sqrt{4\pi}T^{1.5}}e^{\frac{-1}{8}}(y-x)^2 & \text{for } x, z \in [-\sqrt{T/2}, \sqrt{T/2}] \end{cases}$$

Thus, since $|s_{i_{j+p}} - s_{i_j}| = p/m \sim \sqrt{T/2}$ and $|x_i - s_{i_j}| \leq \sqrt{T/2}$ and $|s_{i_{j+p}} - x_i| \leq \sqrt{T/2}$ we have

$$\begin{aligned}
&W_{ij}(x_i) - W_{ij}(s_{i_{j+p}} - \sqrt{2T}) \\
&\quad \geq -g'(s_{i_{j+p}} - x_i)\Bigg(- g'(x_i - s_{i_j})(s_{i_{j+p}} - x_i - \sqrt{2T}) \\
&\qquad\qquad + \frac{7}{32\sqrt{4\pi}T^{1.5}}e^{\frac{-1}{8}}(s_{i_{j+p}} - x_i - \sqrt{2T})^2\Bigg) \\
&\qquad\qquad - g'(x_i - s_{i_j})\left(-g'(s_{i_{j+p}} - x_i)(\sqrt{2T} - s_{i_{j+p}} - x_i)\right) \\
&\quad = g'(s_{i_{j+p}} - x_i)g'(x_i - s_{i_j})\frac{7}{32\sqrt{4\pi}T^{1.5}}e^{\frac{-1}{8}}(s_{i_{j+p}} - x_i - \sqrt{2T})^2 \\
&\quad \geq (s_{i_{j+p}} - x_i)(x_i - s_{i_j})e^{\frac{-(s_{i_{j+p}} - x_i)^2 - (x_i - s_{i_j})^2}{4T}}\frac{7}{512\sqrt{4\pi}T^{3.5}}e^{\frac{-1}{8}} \\
&\quad \geq (s_{i_{j+p}} - x_i)(x_i - s_{i_j})e^{\frac{-1}{8}}\frac{7}{512\sqrt{4\pi}T^{3.5}}e^{\frac{-1}{8}}
\end{aligned}$$

where the last inequality follow since $0 \leq x_i - s_{i_j} = \sqrt{T/2} - (s_{i_{j+p}} - x_i) \leq \sqrt{T/2}$. Now,

$$\begin{aligned}
\sum_{j=1}^{p}(s_{i_{j+p}} - x_i)(x_i - s_{i_j}) &\geq \sum_{j=1}^{p}\frac{j}{m}(\sqrt{T/2} - \frac{j}{m}) \\
&= \frac{p(p+1)(3m\sqrt{T/2} - 2p - 1)}{6m^2} \\
&\sim \frac{m\sqrt{T/2}(m\sqrt{T/2} + 1)^2}{6m^2}
\end{aligned}$$

□

We can now turn to our proof of the upper bound on the EMD error of Algorithm 1.

Proof of Theorem 8.6. Note that from Lemma 8.1, w.h.p. f_0 is a feasible point so $\|\hat{f}\|_1 \leq k$. Let $S_i = (x_i - \sqrt{T/2}, x_i + \sqrt{T/2}) \cap [0,1]$ for $i \in [k]$. Then we have that the S_i's are disjoint and each interval S_i contains $\sqrt{2T}m$ sensors. Let $p = \lfloor \sqrt{T/2}m \rfloor$ and let $s_{i_1} < \cdots < s_{i_p}$ be the locations of the sensors in S_i to the left of x_i and $s_{i_{p+1}} > \cdots > s_{i_{2p}}$ be the locations to the right. By Condition 3 we know that for any pair $l \in T_i$ and s_{c_j} where $i \neq c$ we have $|l/n - s_{c_j}| \geq A$.

For all $i \in [k]$ and $j \in [p]$, let

$$W_{ij}(z) = -g'(s_{i_{j+p}} - x_i)g(z - s_{i_j}) - g'(x_i - s_{i_j})g(s_{i_{j+p}} - z).$$

Let T_i be the set of all $l \in [n]$ such that $l/n \in (x_i - \frac{|x_i - x_{i-1}|}{2}) \cap [0,1]$ then

$$g(x_i - s_{i_j}) - \sum_{l \in T_i} \hat{f}_l g(l/n - s_{i_j}) \leq y_{i_j} - \hat{y}_{i_j} + \frac{\|\hat{x}\|_1}{\sqrt{4\pi T}} e^{\frac{-A^2}{4T}}.$$

Therefore,

$$\begin{aligned}
&\sum_{i=1}^{k}\sum_{j=1}^{p}\left| W_{ij}(x_i) - \sum_{l \in T_i} \hat{f}_l W_{ij}(l/n)\right| \\
&\quad = \sum_{i=1}^{k}\sum_{j=1}^{p}\Bigg[-g'(s_{j+p} - x_i)\left| g(x_i - s_{i_j}) - \sum_{l \in T_i} \hat{f}_l g(l/n - s_{i_j})\right| \\
&\qquad\qquad - g'(x_i - s_{i_j})\left| g(s_{i_{j+p}} - x_i) - \sum_{l \in T_i} \hat{f}_l g(s_{i_j} - l/n)\right|\Bigg] \\
&\quad \leq C_2\left[\|y - \hat{y}\|_1 + \frac{2p\|\hat{x}\|_1}{\sqrt{4\pi T}} e^{\frac{-A^2}{4T}}\right] \\
&\quad \leq C_2\left[\sqrt{m}\|y - \hat{y}\|_2 + \frac{2p\|\hat{x}\|_1}{\sqrt{4\pi T}} e^{\frac{-A^2}{4T}}\right] \\
&\quad \leq C_2\left[2\sigma m + \frac{mk}{\sqrt{2\pi}} e^{\frac{-A^2}{4T}}\right]
\end{aligned}$$

where $C_2 = \max_{i \in [k]} \max_{j \in [2p]} [-g'(|s_{i_j} - x_i|)]$ and the last inequality holds with high probability from Lemma 8.1. Also,

$$\begin{aligned}
W_{ij}(x_i) &\leq -g'(s_{i_{j+p}} - x_i)g(x_i - s_{i_j}) - g'(x_i - s_{i_j})g(s_{i_{j+p}} - x_i) \\
&\leq \frac{1}{16\pi T^2}|s_{i_{j+p}} - x_i| + \frac{1}{16\pi T^2}|x_i - s_{i_j}| \\
&\leq \frac{1}{8\pi T^{1.5}}
\end{aligned}$$

Therefore, since $\sum_{\ell\in T_i}\hat{f}_\ell \leq 1$ we have

$$\sum_{i=1}^{k}\sum_{j=1}^{p}\left|W_{ij}(x_i)-\sum_{l\in T_i}\hat{f}_l W_{ij}(l/n)\right| \leq \min\left\{\frac{km}{8\pi T}\ ,\ C_2\left[2\sigma m+\frac{mk}{\sqrt{2\pi}}e^{\frac{-A^2}{4T}}\right]\right\} = B.$$

Conversely, by Lemma 8.10, we have

$$\begin{aligned}
&\sum_{i=1}^{k}\sum_{j=1}^{p}\left|W_{ij}(x_i)-\sum_{l\in T_i}\hat{f}_l W_{ij}(l/n)\right| \\
&\geq \sum_{j=1}^{p}\left[\sum_{i=1}^{k}\left(1-\sum_{l\in T_i}\hat{f}_l\right)W_{ij}(x_i)+\sum_{i=1}^{k}\sum_{l\in T_i}\hat{f}_l(W_{ij}(x_i)-W_{ij}(l/n))\right] \\
&\geq \sum_{j=1}^{p}\sum_{i=1}^{k}\left(1-\sum_{l\in T_i}\hat{f}_l\right)W_{ij}(x_i)+\sum_{j=1}^{p}\sum_{i=1}^{k}\sum_{l:l/n\in S_i}\hat{f}_l C_1^j(x_i-l/n)^2+\sum_{i=1}^{k}\sum_{l:l/n\notin S_i}C_3\hat{f}_l \\
&\geq \sum_{j=1}^{p}\sum_{i=1}^{k}\left(1-\sum_{l\in T_i}\hat{f}_l\right)W_{ij}(x_i)+C_5\sum_{i=1}^{k}\sum_{l:l/n\in S_i}\hat{f}_l(x_i-l/n)^2+\sum_{i=1}^{k}\sum_{l:l/n\notin S_i}C_3\hat{f}_l
\end{aligned}$$

where $C_3 = \min_{i\in[k]}\min_{l:l/n\notin S_i}\sum_{j=1}^{p}(W_{ij}(x_i)-W_{ij}(l/n)) \geq 0$.

Now, by the uniformity of the sensor locations, $W_j = W_{i_j}(x_i) = W_{i'_j}(x_{i'})$ so

$$\sum_{j=1}^{p}\sum_{i=1}^{k}\left(1-\sum_{l\in T_i}\hat{f}_l\right)W_{ij}(x_i) = \sum_{j=1}^{p}\sum_{i=1}^{k}\left(1-\sum_{l\in T_i}\hat{f}_l\right)W_j \geq \sum_{j=1}^{p}W_j(k-\|\hat{f}\|_1) \geq 0.$$

Similarly the other two terms are both positive. Therefore, $\sum_{i=1}^{k}\sum_{l:l/n\notin S_i}\hat{f}_l \leq B/C_3$ or equivalently,

$$\sum_{i=1}^{k}\sum_{l:l/n\in S_i}\hat{f}_l \geq \|\hat{f}\|_1 - \min\{k, B/C_3\}.$$

This implies that most of the weight of the estimate $\hat{f}$ is contained in the intervals $S_1,\cdots,S_k$. Also,

$$\begin{aligned}
B &\geq \sum_{j=1}^{p}|W_{ij}(x_i)-\sum_{i\in T_i}\hat{f}_l W_{ij}(l/n)| \\
&\geq \sum_{j=1}^{p}W_{ij}(x_i)-\sum_{i\in T_i}\hat{f}_l W_{ij}(x_i) \\
&\geq \left(\sum_{j=1}^{p}W_{ij}(x_i)\right)\left(1-\sum_{l\in T_i}\hat{f}_l\right) = C_4\left(1-\sum_{l\in T_i}\hat{f}_l\right)
\end{aligned}$$

Therefore, $\sum_{l\in T_i} \hat{f}_l \geq 1 - \min\{1, B/C_4\}$ and

$$\begin{aligned}\sum_{l:l/n\in S_i} \hat{f}_l \leq \sum_{l\in T_i} \hat{f}_l = \|\hat{f}\|_1 - \sum_{a\neq i}\sum_{l\in T_a} \hat{f}_l \\ \leq \|\hat{f}\|_1 - (k-1)(1-\min\{1, B/C_4\}) \leq 1 + (k-1)\min\{1, B/C_4\}.\end{aligned}$$

This implies that the weight of estimate $\hat{f}$ contained in the interval S_i is not too much larger than the true weight of 1. Also,

$$\frac{\sum_{l:l/n\in S_i} \hat{f}_l}{\|\hat{f}\|_1} \leq \frac{1+(k-1)\min\{1, B/C_4\}}{k(1-\min\{1, B/C_4\})} = \frac{1}{k} + \frac{\min\{1, B/C_4\}}{1-\min\{1, B/C_4\}}$$

In order to upper bound the EMD$(\frac{f_0}{k}, \frac{\hat{f}}{\|\hat{f}\|_1})$ we need a flow, we are going to assign weight $\min\{\frac{\sum_{l:l/n\in S_i} \hat{f}_l}{\|\hat{f}\|_1}, \frac{1}{k}\}$ to travel to x_i from within S_i. The remaining unassigned weight is at most $k\frac{\min\{1,B/C_4\}}{1-\min\{1,B/C_4\}} + \frac{\min\{1,B/C_3\}}{k(1-\min\{1,B/C_4\})}$ and this weight can travel at most 1 unit in any flow. Therefore,

$$\begin{aligned}\text{EMD}\left(\frac{f_0}{k}, \frac{\hat{f}}{\|\hat{f}\|_1}\right) &\leq \sum_{i=1}^{k}\sum_{l:l/n\in S_i} \frac{\hat{f}_l}{\|\hat{f}\|_1}|x_i - l/n| + k\frac{\min\{1, B/C_4\}}{1-\min\{1, B/C_4\}} \\ &\quad + \frac{\min\{1, B/C_3\}}{k(1-\min\{1, B/C_4\})} \\ &\leq \frac{1}{k(1-\min\{1, B/C_4\})}\sqrt{\frac{B}{C_5}} + k\frac{\min\{1, B/C_4\}}{1-\min\{1, B/C_4\}} + \frac{\min\{1, B/C_3\}}{k(1-\min\{1, B/C_4\})} \qquad (8.3)\end{aligned}$$

Now, we need bounds on C_1, C_2, C_3 and C_4. Firstly, recall

$$C_1 = \inf_{z\in[s_2-\sqrt{2T}, s_1+\sqrt{2T}]} [-W''(z)/2] > 0.$$

The sensors s_{i_j} and $s_{i_{j+p}}$ are at a distance of p/m and recall that we only chose the sensors such that $|x_i - s_{i_j}| \geq 1/m$. Thus, any $z \in [s_2 - \sqrt{2T}, s_1 + \sqrt{2T}]$ we have either $|z - s_{i_j}| \leq \sqrt{T/8}$ or $|z - s_{i_{j+p}}| \leq \sqrt{T/8}$ so

$$\begin{aligned}-W''(z) &= g'(s_{i_{j+p}} - x_i)g''(z - s_{i_j}) + g'(x_i - s_{i_j})g''(s_{i_{j+p}} - z) \\ &\geq \frac{1}{16\pi T^3}\left(\frac{1}{m}\left(1 - \frac{1}{2T}\frac{T}{8}\right)e^{-\frac{1}{4T}\frac{T}{8}}\right)\end{aligned}$$

Therefore, $C_1 \geq \frac{17e^{\frac{-1}{32}}}{512}\frac{1}{m}\frac{1}{T^3}$. Next,

$$C_2 = \max_{i\in[k]}\max_{j\in[2p]} -g'(|x_i - s_j|) \leq \frac{1}{2\sqrt{4\pi T}}e^{\frac{-1}{4m^2T}}.$$

By Lemma 8.12 we have,

$$C_3 = \min_{i \in [k]} \min_{l : l/n \notin S_i} \sum_{j=1}^{p} (W_{ij}(x_i) - W_{ij}(l/n)) = \Omega\left(\frac{m\sqrt{T/2}(m\sqrt{T/2}+1)^2}{m^2}\frac{1}{T^{3.5}}\right).$$

Finally,

$$\begin{aligned}
C_4 &= \sum_{j=1}^{p} W_{ij}(x_i) \\
&= \frac{1}{8\pi T^2}\sum_{j=1}^{p}(s_{i_{j+p}} - x_i)e^{\frac{-(s_{i_{j+p}}-x_i)^2-(x_i-s_{i_j})^2}{4T}} + (x_i - s_{i_j})e^{\frac{-(s_{i_{j+p}}-x_i)^2-(x_i-s_{i_j})^2}{4T}} \\
&\geq \frac{1}{8\pi T^2}e^{\frac{-1}{8}}\sum_{j=1}^{p}(s_{i_{j+p}} - s_{i_j}) \\
&\geq \Omega\left(\frac{me^{\frac{-1}{8}}}{16\pi T}\right).
\end{aligned}$$

Lemma 8.11 gives $C_5 = \Omega\left(\frac{m\sqrt{T/8}+1}{T^{2.5}}\right)$. Putting all our bounds into (8.3) we gain the final result. □

Part IV

Property Testing for Differential Privacy

CHAPTER 9

Introduction

Recently differential privacy has gained traction outside of theoretical research as several companies (Google, Apple, Microsoft, Census, etc.) have announced deployment of large-scale differentially private mechanisms [Erlingsson et al., 2014, Apple, 2017, Adowd and Schmutte, 2017, Ding et al., 2017]. This use of DP, while exciting, might be construed as a marketing tool used to encourage privacy-aware consumers to release more of their sensitive data to the company. In addition, the software behind the deployment of DP is typically proprietary since it ostensibly provides commercial advantage. This raises the question: with limited access to the software, can we verify the privacy guarantees of purportedly DP algorithms?

Suppose there exists some randomised algorithm $\mathcal{A}$ that is claimed to be Ξ- differentially private and we are given query access to $\mathcal{A}$. That is, the domain of $\mathcal{A}$ is the set of databases and we have the power to choose a database D and obtain a (randomised) response $\mathcal{A}(D)$. How many queries are required to verify the privacy guarantee? We formulate this problem in the property testing framework for pure DP, approximate DP, random pure DP, and random approximate DP[1].

Definition 9.1 (Property testing with side information)**.** A property testing algorithm with query complexity q, proximity parameter α, privacy parameters Ξ and side information S, makes q queries to the black-box and:

1. (Completeness) ACCEPTS with probability at least $2/3$ if $\mathcal{A}$ is Ξ-private and S is accurate.

2. (Soundness) REJECTS with probability at least $2/3$ if $\mathcal{A}$ is α-far from being Ξ-private.

In this early stage of commercial DP algorithms, approaches to transparency have been varied. For some algorithms, like Google's RAPPOR, a full description of the algorithm has been released [Erlingsson et al., 2014]. On the other hand, while Apple has released

[1] This Part is based on joint work with Anna Gilbert.

a white paper [Differential Privacy Team, 2017] and a patent [Thakurta et al., 2017], there are still many questions about their exact implementations. We focus on the two extreme settings: when we are given *no information* about the black-box (except the domain and range), and the *full information* setting where we have an untrusted full description of the algorithm $\mathcal{A}$.

Both settings are subject to fundamental limitations. We first show that *verifying* privacy is at least as difficult as *breaking* privacy, even in the full information setting. That is, suppose r samples are sufficient to verify that an algorithm is Ξ-private. Then Theorem 10.2 implies that for every algorithm that is not Ξ-private, there exists some pair of neighbouring databases D and D' such that r samples from $\mathcal{A}(D)$ is enough to distinguish between D and D'. Differential privacy is designed so that this latter problem requires a large number of samples. This connection has the unfortunate implication that verifiability and privacy are directly at odds: *if a privacy guarantee is efficiently verifiable, then it mustn't be a strong privacy guarantee*.

For the remainder of this part we restrict to discrete distributions on $[n]$. Our upper and lower bounds in each setting are contained in Table 9.1. We rule out sublinear verification of privacy in every case except verifying approximate differential privacy in the full information setting. That is, for all other definitions of privacy, the query complexity for property testing for privacy is $\Omega(n)$.

Each privacy notion we consider is a relaxation of pure differential privacy. Generally, the privacy is relaxed in one of two ways: either privacy loss is allowed to occur on unlikely *outputs*, or privacy loss is allowed to occur on unlikely *inputs*. The results in Theorem 10.4 and the lower bounds in Table 9.1 imply that for efficient verification, we need to relax in *both* directions. That is, random approximate DP is the only efficiently verifiable privacy notion in the no information setting. Even then, we need about $1/\delta^2$ queries *per database* to verify (ϵ, δ)-approximate differential privacy. Theorem 10.8 shows that random approximate DP can be verified in (roughly) $O(\frac{4n(1+e^{2\epsilon})\log(1/\gamma)}{\gamma\delta^2})$ samples, where (roughly) δ and γ are the probabilities of choosing a disclosive output or input, respectively. This means verification is efficient if δ and γ are small but not too small. This may seem insufficient to those familiar with DP, where common wisdom decrees that δ and γ should be small enough that this query complexity is infeasibly large.

There have been several other relaxations of pure differential privacy proposed in the literature, chief among them Rényi DP [Mironov, 2017] and concentrated DP [Dwork and Rothblum, 2016]. These relaxations find various ways to sacrifice privacy, with a view towards allowing a strictly broader class of algorithms to be implemented. Similar to pure DP, Rényi and concentrated DP have the property that two distributions P and Q can be close in TV distance while the pair (P, Q) has infinite privacy parameters. Thus, many of the results for pure differential privacy in this part can be easily extended to Rényi and

Table 9.1: Per Database Query complexity bounds for property testing of privacy

	No Information	Full information
pDP	Unverifiable [Theorem 10.5]	$\Omega\left(\frac{1}{\beta\alpha^2}\right)$ [Theorem 10.12]
		$O\left(\frac{\ln n}{\alpha^2\beta^2}\right)$ [Theorem 10.11]
aDP	$\Omega(\max\{n^{1-o(1)}, \frac{1}{\alpha^2}\})$ [Theorem 10.7]	$O\left(\frac{\sqrt{n}}{\alpha^2}\right)$[Theorem 10.10]
	$O(\frac{n}{\alpha^2})$ [Theorem 10.8]	

concentrated DP. We leave these out of our discussion for brevity.

One might hope to obtain significantly lower query complexity if the property tester algorithm is given side information, even if the side information is untrusted. We find that this is true for both approximate DP and pure DP, if we allow the query complexity to depend on the side information. Recall, a randomised algorithm $\mathcal{A}$ can be abstracted as a set of distributions $\{P_D\}$ where $\mathcal{A}(D) \sim P_D$. We obtain a sublinear verifier for approximate DP. For pure DP, we find the quantity that controls the query complexity is

$$\beta = \inf_D \min_i P_D(i),$$

the minimum value of the collection of distributions. If β is large then efficient verification is possible: verifying that the pure differential privacy parameter is less than $\epsilon + \alpha$ requires $O(\frac{\ln n}{\alpha^2\beta^2})$ queries of each database (Theorem 10.11). Note that this is not sublinear since $\beta \leq 1/n$ and if $\beta = 0$ then we have no improvement on the no information setting. However, for reasonable β, this is a considerable improvement on the no information lower bounds and may be efficient for reasonable n.

A central theme of this work is that verifying the privacy guarantees that corporations (or any entity entrusted with private data) claim requires compromise by either the verifier or algorithm owner. If the verifier is satisfied with only a weak privacy guarantee (random approximate DP with δ and γ small but not extremely small), then they can achieve this with no side information from the algorithm owner. If the company is willing to compromise by providing information about the algorithm up-front, then much stronger privacy guarantees can be verified. Given this level of transparency, one might be tempted to suggest that the company provide source code instead. While verifying privacy given source code is an important and active area of research, there are many scenarios where the source code itself is proprietary. We have already seen instances where companies have been willing to provide detailed descriptions of their algorithms. In the full information case, we obtain our lowest sample complexity algorithms, including a sublinear algorithm for verifying approximate differential privacy.

This part proceeds as follows: we start by defining property testing for privacy in Section 9.1. We then proceed to the main contributions of this work:

- Verifying privacy is as hard as breaking privacy (Section 10.1).
- In the no information setting, verifying pure differential privacy is impossible while there is a finite query complexity property tester for approximate differential privacy (Section 10.3).
- If $\beta > 0$ then finite query complexity property testers exist for pure differential privacy in the full information setting (Section 10.4).
- A sublinear property tester exists for approximate differential privacy in the full information setting.

The main lower bounds and algorithmic upper bounds in this part are summarized in Table 9.1.

9.1 Background and Problem Formulation

We revert to the original definition of neighbouring: D, D' *neighbouring* if they differ on a single data point. We use the notation $\mathcal{A} = (P_0, P_1)$ to denote an algorithm that accepts on two databases 0 and 1 as input, and $\mathcal{A}(0) \sim P_0$ and $\mathcal{A}(1) \sim P_1$. We will only consider discrete distributions in this part, so P_D is a discrete distribution on $[n]$. For a distribution P, P^r represents r independent copies of P.

Our results vary heavily depending on whether $\delta = 0$ or $\delta > 0$. As such, we use approximate differential privacy or aDP to refer only to the case where $\delta > 0$ and pDP when $\delta = 0$.

9.1.1 Problem Formulation

Our goal is to answer the question *given these privacy parameters, is the algorithm $\mathcal{A}$ at least Ξ-private?* where Ξ is an appropriate privacy parameter. A property testing algorithm, which outputs ACCEPT or REJECT, answers this question if it ACCEPTS whenever $\mathcal{A}$ is Ξ-private, and *only* ACCEPTS if the algorithm $\mathcal{A}$ is close to being Ξ-private. A tester with side information may also REJECT simply because the side information is inaccurate.

We say that $\mathcal{A}$ is α-far from being Ξ-private if $\min_{\Xi'} \|\Xi' - \Xi\| > \alpha$, where the minimum is over all Ξ' such that $\mathcal{A}$ is Ξ'-private. The metrics used for each form of privacy are contained in Table 9.2. We introduce the scalar λ to penalise deviation in one parameter more than deviation in another parameter. For example, it is much worse to mistake a $(0, 0.1)$-RpDP algorithm for $(0, 0)$-RpDP than it is to mistake a $(0.1, 0)$-RpDP algorithm for $(0, 0)$-RpDP. We leave the question of *how much worse* as a parameter of the problem. However, we give the general guideline that if we want an α error to be tolerable in both ϵ and γ then $\lambda \approx \frac{\epsilon}{\gamma}$, which may be large, is an appropriate choice.

Table 9.2: Privacy notions, parameters and metrics.

Privacy Notion	Ξ	$\|\|\Xi - \Xi'\|\|$
pDP	ϵ	$\|\epsilon - \epsilon'\|$
aDP	(ϵ, δ)	$\|\delta_\epsilon - \delta'_\epsilon\|$
RpDP	(ϵ, γ)	$\min\{\|\epsilon - \epsilon'\|, \lambda\|\gamma - \gamma'\|\}$
RaDP	$(\epsilon, \delta, \gamma)$	$\min\{\|\delta_\epsilon - \delta'_\epsilon\|, \lambda\|\gamma - \gamma'\|\}$

The formal definition of a property tester with side information was given in Definition 9.1. A *no information* property tester is the special case when $S = \emptyset$. A *full information* property tester is the special case when $S = \{Q_D\}$ contains a distribution Q_D for each database D. We use Q_D to denote the distribution on outputs presented in the side information and P_D to denote the true distribution on outputs of the algorithm being tested. For $\alpha > 0$ and privacy parameter Ξ, a full information (FI) property tester for this problem satisfies:

1. (Completeness) Accepts with probability at least $2/3$ if the algorithm is Ξ-private and $P_D = Q_D$ for all D.
2. (Soundness) Rejects with probability at least $2/3$ if the algorithm is α-far from being Ξ-private.

We only force the property tester to ACCEPT if the side information is *exactly* accurate ($P_D = Q_D$). It is an interesting question to consider a property tester that is forced to ACCEPT if the side-information is *close* to accurate, for example in TV-distance. We do not consider this in this work as being close in TV-distance does not imply closeness of privacy parameters.

For a database D, we will refer to the process of obtaining a sample from P_D as *querying the black-box*. It will usually be necessary to input each database into the black-box multiple times. We will use m to denote the number of *unique* databases that are queries to the black-box and r to denote the number of times each database is input. We will only consider algorithms where the number of samples from P_D for each input database is r, so our query complexity is mr for each algorithm. Our aim is verify the privacy parameters using as few queries as possible.

9.1.2 Related Work

This work connects to two main bodies of literature. There are several works on verifying privacy with different access models that share the same motivation as this work. In terms of techniques, our work is most closely linked to recent work on property testing of distributions.

Several algorithms and tools have been proposed for formal verification of the DP guarantee of an algorithm [Barthe et al., 2014, Roy et al., 2010, Reed and Pierce, 2010, Gaboardi et al., 2013, Tschantz et al., 2011]. Much of this work focuses on verifying privacy given access to a description of the algorithm. There is a line of work [Barthe et al., 2014, Roy et al., 2010, Reed and Pierce, 2010, Gaboardi et al., 2013, Tschantz et al., 2011, Barthe et al., 2012, 2013, McSherry, 2009] using logical arguments (type systems, I/O automata, Hoare logic, etc.) to verify privacy. These tools are aimed at automatic (or simplified) verification of privacy of source code. There is another related line of work where the central problem is testing software for privacy leaks. This work focuses on *blatant* privacy leaks, such as a smart phone application surreptitiously leaking a user's email [Jung et al., 2008, Enck et al., 2010, Fan et al., 2012]. We are unaware of any work verifying DP assuming only black-box access to the private algorithm.

Given sample access to two distributions P and Q and a distance measure $d(\cdot,\cdot)$, the question of determining between $d(P,Q) \leq a$ and $d(P,Q) \geq b$ is called *tolerant property testing*. This question is closely related to the question of whether $\mathcal{A} = (P,Q)$ is private. There is a large body of work exploring lower bounds and algorithmic upper bounds for tolerant testing using standard distances (TV, KL, χ^2, etc.) with both $a = 0$ and $a > 0$ [Daskalakis et al., 2018, Paninski, 2008, Batu et al., 2013, Acharya et al., 2015, Valiant and Valiant, 2014]. In our work, we draw most directly from the techniques of Valiant [2011].

A relevant paper in the intersection of privacy and property testing is Dixit et al. [2013]. The access model in this paper is different to ours but their goal is similar, Dixit et. al relate privacy testing to testing the Lipschitz property of functions and make progress on the latter.

CHAPTER 10

Technical Contributions

10.1 Lower Bounds via Distinguishability

We now turn to examining the fundamental limitations of property testing for privacy. We find that even in the full information setting, the query complexity to verifying privacy is lower bounded by the number of queries required to distinguish between two possible inputs. We expect the latter to increase with the strength of the privacy guarantee.

Definition 10.1. Databases D and D' are *r-distinguishable under $\mathcal{A}$* if there exists a testing algorithm such that given a description of $\mathcal{A}$ and $x \sim P_{D''}^r$ where $D'' \in \{D, D'\}$, it accepts with probability at least $2/3$ if $D'' = D$ and rejects with probability at least 2/3 if $D'' = D'$.

A major reason that DP has gained traction is that it is preserved even if the (randomised) algorithm is repeated (Lemma 2.6). That is, if $k > 0$ and $(P_D, P_{D'})$ is private, then $(P_D^k, P_{D'}^k)$ is private with slightly worse privacy parameters. The rate of decay of the privacy parameters (in k) varies with the notion of privacy. Typically we want the privacy parameters to start small enough that k has to be quite large before any pair of neighbouring databases can be distinguished between using the output.

The following theorem says that the per database query complexity of a privacy property testing algorithm is lower bounded by the minimal r such that two neighbouring databases are r-distinguishable under $\mathcal{A}$.

Theorem 10.2. *Consider any privacy definition, privacy parameter Ξ, and let $\alpha > 0$. Suppose there exists a Ξ-privacy property tester with proximity parameter α and (per database) query complexity r. Let $\mathcal{A}$ be an algorithm that is α-far from Ξ-private. If the privacy notion is*

- *pDP or aDP then there exists a pair of databases that are r-distinguishable under $\mathcal{A}$.*
- *RpDP or RaDP and $\Xi = (\epsilon, \delta, \gamma)$, then a randomly sampled pair of databases has probability at least $\gamma + \frac{\alpha}{\lambda}$ of being r-distinguishable.*

Proof. We start with pDP or aDP and suppose such a Ξ-privacy property testing algorithm exists. Let $\mathcal{A}$ be an algorithm that is α-far from Ξ-private. Since the privacy parameter is defined as a maximum over all neighbouring databases, there exists a pair of databases D and D' such that $(P_D, P_{D'})$ has the same privacy parameter as $\mathcal{A}$. We can design a tester algorithm that distinguishes between D and D' as follows: given input $x \sim P_{D''}^r$, first sample $y \sim P_D^r$. Then run the privacy property testing algorithm on $\mathcal{B} = (P_{D''}, P_D)$ with sample (x, y). If $D'' = D$ then $\mathcal{B}$ is 0-DP, so the property tester will accept with probability at least 2/3. If $D'' = D'$ then $\mathcal{B}$ is α-far from from Ξ-private so the property tester will reject with probability at least 2/3.

Finally, suppose such a (ϵ, γ)-RpDP property testing algorithm exists. Let $\mathcal{A}$ be an algorithm that is α-far from (ϵ, γ)-private so that, in particular, $\mathcal{A}$ is not $(\epsilon + \alpha, \gamma + \frac{\alpha}{\lambda})$-RpDP. Thus, if we randomly sample a pair of neighbouring databases D and D', with probability $\gamma + \frac{\alpha}{\lambda}$, $(P_D, P_{D'})$ is not $\epsilon + \alpha$-pDP. The remainder of the proof proceeds as above by noticing that the algorithm $(P_D, P_{D'})$ is α-far from (ϵ, γ)-RpDP and (P_D, P_D) is (ϵ, γ)-RpDP. The proof of almost identical for RaDP. □

10.2 Restriction to Two Distribution Setting

Differential privacy is inherently a local property. That is, verifying that $\mathcal{A}$ is Ξ-private means verifying that $(P_D, P_{D'})$ is Ξ-private, either always or with high probability, for pairs of neighbouring databases D and D'. We refer to the problem of determining whether a pair of distributions (P_0, P_1) satisfies Ξ-privacy as the *two database setting*. We argue in this section that the hard part of privacy property testing is the two database setting. For this reason, from Section 10.3 onwards, we only consider the two database setting.

An algorithm is non-adaptive if it chooses m pairs of distributions and queries the blackbox with each database r times. It does not choose its queries adaptively. The following is a non-adaptive algorithm for converting a tester in the two database setting to a random privacy setting.

Theorem 10.3 (Conversion to random privacy tester)**.** *If there exists a Ξ-privacy property tester for the two database setting with query complexity r per database and proximity parameter α, then there exists a privacy property tester for (Ξ, γ)-random privacy with proximity parameter 2α and query complexity*

$$O\left(r \log\left(\frac{2\lambda}{\alpha}\right) \frac{(\alpha/\lambda + \gamma)^2 + \alpha/\lambda}{(\alpha/\lambda)^2}\right).$$

Proof. The conversion is given in Algorithm 3. We first prove completeness. Suppose $\mathcal{A}$ is (γ, Ξ)-random private. Let

$$S = \{(D, D') \mid D, D' \text{ are neighbours and } (P_D, P_{D'}) \text{ is } \Xi\text{-private}\}$$

Algorithm 3 Random-privacy Property Tester

```
Input: A two distribution property tester $T$, $\alpha, \gamma > 0$, a data distribution $\mathcal{D}$
for $i = 1 : m$ do
    Sample $(D, D')$ neighbours from $\mathcal{D}$.
    for $j = 1 : \log(2\lambda/\alpha)$ do
        $x_{ij} = 1$ if $T(P_D, P_{D'})$ REJECTS
    end for
    $x_i = \lfloor \frac{1}{2} + \frac{1}{\log(2\lambda/\alpha)} \sum_{j=1}^{\log(2\lambda/\alpha)} x_{ij} \rfloor$
end for
$y = \frac{1}{m} \sum_{i=1}^{m} x_i$
if $y \leq \gamma + \frac{\alpha}{\lambda}$ then
    Output: ACCEPT
else
    Output: REJECT
end if
```

so $1 - \gamma \leq \mathbb{P}(S) \leq 1$. Our goal is to estimate $\mathbb{P}(S)$ using the empirical estimate given by $\frac{1}{m}\sum_{i=1}^{m} x_i$. We perform the property tester $\log(\lambda/\alpha)$ times on the pair $(P_D, P_{D'})$ to reduce the failure probability from $1/3$ to $\frac{\alpha}{2\lambda}$ so

$$\mathbb{E}[x_i] = \mathbb{P}(x_i = 1 \mid (D, D') \in S)\mathbb{P}(S) + \mathbb{P}(x_i = 1 \mid (D, D') \notin S)\mathbb{P}(S^c) \leq \frac{\alpha}{2\lambda} + \gamma.$$

Now,

$$\mathbb{P}\left(y \geq \gamma + \frac{\alpha}{\lambda}\right) \leq \mathbb{P}\left(y - \mathbb{E}[y] \geq \frac{\alpha}{2\lambda}\right) \leq e^{\frac{-m(\alpha/\lambda)^2}{2(\frac{\alpha}{2\lambda}+\gamma)^2+2(\alpha/\lambda)}} \leq \frac{1}{3}$$

where the first inequality follows from Bernstein's inequality [Sridharan, 2018]. Therefore, Algorithm 3 ACCEPTS with probability at least 2/3. To prove soundness suppose $\mathcal{A}$ is 2α-far from Ξ-private. Let

$$S = \{(D, D') \mid D, D' \text{ are neighbours and } (P_D, P_{D'}) \text{ is } 2\alpha\text{-far from } \Xi\text{-private}\}$$

so $1 \geq \mathbb{P}(S) \geq \gamma + 2\alpha/\lambda$. Then $\mathbb{E}[x_i] \geq \left(1 - \frac{\alpha}{2\lambda}\right)\left(\gamma + 2\frac{\alpha}{\lambda}\right) \geq \gamma + \frac{\alpha}{\lambda} + \frac{\alpha}{2\lambda}$. Therefore as above,

$$\mathbb{P}\left(y \leq \gamma + \frac{\alpha}{\lambda}\right) \leq \mathbb{P}\left(y - \mathbb{E}[y] \leq \frac{-\alpha}{2\lambda}\right) \leq \frac{1}{3}.$$

So, Algorithm 3 REJECTS with probability at least 2/3. □

Notice that if $\gamma \approx \frac{\alpha}{\lambda}$ then the query complexity is approximately $r\log(\frac{\lambda}{\alpha})\frac{\lambda}{\alpha} \approx \frac{r\log(\frac{1}{\gamma})}{\gamma}$. One shortcoming of the conversion algorithm in Theorem 10.3 is that we need to know the data distribution $\mathcal{D}$. We can relax to an approximation $\mathcal{D}'$ that is close in TV-distance, but it is not difficult to see that $\|\mathcal{D} - \mathcal{D}'\|_1 \leq \frac{\alpha}{\lambda}$ is necessary.

Theorem 10.4 (Lower bound)**.** *Let $\gamma, \alpha > 0$. Let r be a lower bound on the query complexity in the two distribution settting. If $\gamma + \frac{\alpha}{\lambda}$ is sufficiently small then any non-adaptive (Ξ, γ)-random privacy property tester with proximity parameter α has query complexity $\Omega(\max\{r, \frac{\lambda}{\alpha}\})$.*

We conjecture that the lower bound is actually $\Omega(r\frac{\lambda}{\alpha})$. If this is true then Theorem 10.3 gives an almost optimal conversion from the two information setting to the random setting.

Proof. A random privacy property tester naturally induces a property tester in the two distribution setting (P, Q) by setting $P_D = P$ for half the databases and $P_D = Q$ for the other half. Then $\{P_D\}$ is (Ξ, γ)-random private if (P, Q) is Ξ and α-far if (P, Q) is α-far. Therefore, the random privacy tester must use at least as many queries as a privacy tester in the two database setting.

Suppose (P, Q) is α-far from Ξ-private and the data universe is uniformly distributed. If $\gamma + \frac{\alpha}{\lambda}$ is small enough then there exists a pair of nested subsets $S' \subset S \subset \mathbb{Z}^\Omega$ such that

$$\mathbb{P}((S \times S^c) \cap \{(D, D') \mid D, D' \text{ neighbours }\}) = \gamma + \frac{\alpha}{\lambda}$$

and

$$\mathbb{P}((S' \times S'^c) \cap \{(D, D') \mid D, D' \text{ neighbours }\}) = \gamma.$$

Define $P_D = P$ if $D \in S$, $P_D = Q$ if $D \in S^c$, $Q_D = P$ if $D \in S'$ and $Q_D = Q$ if $D \in S'$. Then $\{P_D\}$ is (Ξ, γ)-random private and $\{Q_D\}$ is α-far from (Ξ, γ)-random private.

Recall that a non-adaptive property testing algorithm can query by randomly sampling a pair of neighbours D, D', and then sampling $P_D \times P_{D'}$. If N is the normalisation factor, the distributions $\frac{1}{N}\sum_{(D,D')\text{ neighbours}} P_D \times P_{D'}$ and $\frac{1}{N}\sum_{(D,D')\text{ neighbours}} Q_D \times Q_{D'}$ have total variation distance $2\|P - Q\|_{TV}\mathbb{P}((S\backslash S') \times S^c) \cap \{(D, D') \mid D, D' \text{ neighbours }\}) \leq \frac{\alpha}{\lambda}$. Therefore, it takes at least $\frac{\lambda}{\alpha}$ queries to distinguish between $\{P_D\}$ and $\{Q_D\}$. □

10.3 No Information Setting

We first show that no privacy property tester with finite query complexity exists for pDP. We then analyse a finite query complexity privacy property tester for aDP, as well query complexity lower bounds.

10.3.1 Unverifiability

The impossibility of testing pDP arises from the fact that very low probability events can cause the privacy parameters to increase arbitrarily. In each case we can design distributions P and Q that are close in TV-distance but for which the algorithm (P, Q) has arbitrarily large privacy parameters. This intuition allows us to use a corollary of Le Cam's inequality (Corollary 3.4) to prove our infinite lower bounds.

Theorem 10.5 (pDP lower bound)**.** *Let $\alpha > 0$ and $\epsilon > 0$. No ϵ-pDP property tester with proximity parameter α has finite query complexity.*

Proof. Let r be the query complexity of any pDP property tester. Let $A > 2\epsilon + \alpha$. Our goal is to prove that $r > \theta(e^A/A)$. If this is true for all A, the query complexity cannot be finite.

Consider algorithms, $\mathcal{A} = (P_0, P_1)$ and $\mathcal{B} = (Q_0, Q_1)$ where

$$P_0 = Q_0 = e^{-A-\epsilon}\chi_\psi \ + \ (1 \ - \ e^{-A-\epsilon})\chi_\omega,$$

$$P_1 = e^{-A}\chi_\psi + (1 - e^{-A})\chi_\omega \text{ and } Q_1 = e^{-2A}\chi_\psi + (1 - e^{-2A})\chi_\omega.$$

Then $\mathcal{A}$ is ϵ-pDP and $\mathcal{B}$ is α-far from ϵ-pDP. Now, by Pinsker's inequality,

$$\|(P_0^r, P_1^2), (Q_0^r, Q_1^r)\|_{\text{TV}} \ \leq \ \sqrt{\frac{r}{2}}\sqrt{\mathrm{D}_{\text{KL}}(P_0|Q_0)}.$$

Therefore, by Lemma 3.4,

$$\begin{aligned} r &\geq \frac{2}{9}\frac{1}{\mathrm{D}_{\text{KL}}(P_0|Q_0)} \\ &= \frac{2}{9}\frac{1}{e^{-A}\log(e^A) + (1 - e^{-A})\log\left(\frac{1-e^{-A}}{1-e^{-2A}}\right)} = \theta\left(\frac{e^A}{A}\right). \end{aligned}$$

□

We designed two distributions that are equal on a large probability set but for which the ratio $\frac{P_0(x)}{Q_0(x)}$ blows-up on a set with small probability. In Section 10.4 we will see that testing pure DP becomes possible if we make assumptions on the algorithm $\mathcal{A}$. The assumption we need will ensure that $\frac{P_0(x)}{Q_0(x)}$ is upper bounded, and similar results hold for RDP.

10.3.2 Property Testing for aDP in the No Information Setting

Fortunately, the situation is less dire for verifying aDP. Finite query complexity property testers do exist for aDP, although their query complexity can be very large. In the previous section, we relied on the fact that two distributions P and Q can be close in TV-distance while (P, Q) has unbounded privacy parameters. In this section, we first show this is not true for aDP, which sets it apart from the other privacy notions. We then prove that the query complexity is $\Omega\left(\max\left\{n^{1-o(1)}, \frac{1}{\alpha^2}\right\}\right)$, and there exists an algorithm that uses $O(\frac{4n(1+e^{2\epsilon})}{\alpha^2})$ queries per database. Define

$$\delta_\epsilon^{\mathcal{A}} \geq \max_{D, D' \text{neighbours}} \max_E P_D(E) - e^\epsilon P_{D'}(E). \tag{10.1}$$

An algorithm is (ϵ, δ)-aDP if and only if $\delta > \delta_\epsilon^*$. The following lemma shows the relationship between the aDP parameters and TV-distance.

Lemma 10.6. *Let* $\mathcal{A} = (P_0, P_1)$ *and suppose* $\mathcal{A}$ *is* (ϵ, δ)*-aDP and* $\alpha > 0$*. If* $\mathcal{B} = (Q_0, Q_1)$ *and*

1. $\|P_0 - Q_0\|_{TV} \leq \alpha$

2. $\|P_1 - Q_1\|_{TV} \leq \alpha$,

then $\mathcal{B}$ is $(\epsilon, \delta + (1+e^\epsilon)\alpha) - aDP$. Furthermore, if $\alpha \leq \frac{1-\delta}{1+e^\epsilon}$ then this bound is tight. That is, if $\delta_\epsilon^{\mathcal{A}} > 0$, then there exists an algorithm $\mathcal{B} = (Q_0, Q_1)$ such that conditions (1) and (2) hold but $\mathcal{B}$ is α-far from $(\epsilon, \delta_\epsilon^{\mathcal{A}})$.

Proof. For any event E,

$$Q_0(E) \leq P_0(E) + \alpha \leq e^\epsilon P_1(E) + \delta_{\mathcal{A}} + \alpha \leq e^\epsilon Q_1(E) + e^\epsilon\alpha + \alpha + \delta_{\mathcal{A}}.$$

Similarly, $Q_1(E) \leq e^\epsilon Q_0(E) + e^\epsilon\alpha + \alpha + \delta_{\mathcal{A}}$.

Conversely, let $\mathcal{A} = (P_0, P_1)$ and suppose $\delta_\epsilon^{\mathcal{A}} > 0$. There must exist an event E such that $P_0(E) = e^\epsilon P_1(E) + \delta_\epsilon^{\mathcal{A}}$. The condition on α can rewritten as $1 - \alpha \geq e^\epsilon\alpha + \delta$ so we must have that either $P_0(E) \leq 1 - \alpha$ or $P_1(E) \geq \alpha$.

First, suppose that $P_0(E) \leq 1 - \alpha$. Then there exists a distribution Q_0 such that $\|Q_0 - P_0\|_{\text{TV}} = \alpha$ and $Q_0(E) = P_0(E) + \alpha$. If we let $Q_1 = P_1$ then $Q_0(E) = e^\epsilon Q_1(E) + \alpha + \delta_\epsilon^{\mathcal{A}}$, which implies $\mathcal{B} = (Q_0, Q_1)$ is α-far from $(\epsilon, \delta_\epsilon^{\mathcal{A}})$-aDP.

Finally, suppose $P_1(E) \geq \alpha$. Then there exists a distribution Q_1 such that $\|Q_1 - P_1\|_{\text{TV}} = \alpha$ and $Q_1(E) = P_1(E) - \alpha$. Letting $Q_0 = P_0$, again $\mathcal{B} = (Q_0, Q_1)$ is α-far from $(\epsilon, \delta_\epsilon^{\mathcal{A}})$-aDP. □

Theorem 10.7 (Lower bound)**.** *Let $\alpha, \epsilon, \delta > 0$ and suppose $e^\epsilon/2 + \delta + \alpha < 1$. Any (ϵ, δ)-aDP property tester with proximity parameter α has query complexity*

$$r \geq \max\left\{n^{1-o(1)}, \frac{1}{\alpha^2}\right\}.$$

The proof of the $n^{1-o(1)}$ lower bound uses the low frequency blindness lemma (Lemma 3.5). For aDP, our property is $\pi((P, Q)) = \delta_\epsilon$, which is $((1 + e^\epsilon)\alpha, \alpha)$-weakly-continuous.

Proof. Let $P_0 = Q_0$ be the uniform distribution on $\{\psi, \omega\}$. Let

$$P_1 = \left(\frac{e^\epsilon}{2} + \delta\right)\chi_\psi + \left(1 - \frac{e^\epsilon}{2} - \delta\right)\chi_\omega$$

and

$$Q_1 = \left(\frac{e^\epsilon}{2} + \delta + \alpha\right)\chi_\psi + \left(1 - \frac{e^\epsilon}{2} - \delta - \alpha\right)\chi_\omega.$$

Then, (P_0, P_1) is (ϵ, δ)-aDP and (Q_0, Q_1) is α-far from (ϵ, δ)-aDP. Now,

$$D_{\text{KL}}(P_1|Q_1) = \left(\frac{e^\epsilon}{2} + \delta + \alpha\right)\ln\left(1 + \frac{\alpha}{\frac{e^\epsilon}{2} + \delta}\right) + \left(\frac{e^\epsilon}{2} - \delta - \alpha\right)\ln\left(1 - \frac{\alpha}{\frac{e^\epsilon}{2} - \delta}\right) \lesssim \alpha^2.$$

Algorithm 4 aDP Property Tester

```
Input: Universe size $n, \epsilon, \delta, \alpha > 0$
$\lambda = \max\{\frac{4n(1+e^{2\epsilon})}{\alpha^2}, \frac{12(1+e^{2\epsilon})}{\alpha^2}\}$
Sample $r$ ~Poi($\lambda$)
Sample $D_0 \sim P^r$, $D_1 \sim Q^r$
for $i \in [m]$ do
    $x_i$ = number of $i$'s in $D_0$
    $y_i$ = number of $i$'s in $D_1$
    $z_i = \frac{1}{r}(x_i - e^{\epsilon} y_i)$
end for
$z = \sum_{i=1}^{r} \max\{0, z_i\}$
if $z < \delta + \alpha$ then
    Output: ACCEPT
else
    Output: REJECT
end if
```

By the same argument as Theorem 10.5 we have $r = \Omega\left(\frac{1}{\alpha^2}\right)$.

Suppose $[n]$ is a disjoint union of the sets R_1, R_2 and R_3, all of which have cardinality $n/3$. Let $a = \frac{2\delta+\alpha}{3}, b = 2a, \eta = \frac{\delta-\alpha}{3}$ so $a + \eta = \delta$ and $b - \eta = \delta + \alpha$. Let

$$\begin{aligned} P_1 = P_0 = Q_0 = \frac{3a}{n}\chi_{R_1} & + \frac{3(1-a)}{n}\chi_{R_2} & \\ Q_1 = & \quad \frac{3(1-a)}{n}\chi_{R_2} & + \frac{3a}{n}\chi_{R_3}. \end{aligned}$$

Now, for (P_0, P_1), $\delta_\epsilon \leq a$ and for (Q_0, Q_1), $\delta_\epsilon \geq 2a = b$. Since the distributions agree on any index with probability greater than $\frac{3a}{n}$, Lemma 3.5 implies that no tester can distinguish between $\delta_\epsilon \geq b - \eta = \delta + \alpha$ and $\delta_\epsilon \leq a + \eta = \delta$ with less than $\frac{n}{3a}(1+e^{\epsilon})\eta = \frac{3n(\delta-\alpha)}{2\delta+\alpha} = \Omega(n)$ samples. □

At first glance, Theorem 10.7 doesn't look too bad. We should expect the sample complexity to scale like $1/\alpha^2$ since we need to have enough samples to detect the bad events. Our concern is the size of α. If we would like α to be the same order as δ, then our query complexity must scale as $\frac{1}{\delta^2}$. As we typically require δ to be extremely small (i.e. $\delta \approx 10^{-8}$), $\frac{1}{\delta}$ may be infeasibly large. If we are willing to accept somewhat larger δ, then $\frac{1}{\delta^2}$ may be reasonable.

We now turn our attention to Algorithm 4, a simple algorithm for testing aDP with query complexity $O(\frac{4n(1+e^{2\epsilon})}{\alpha^2})$. Its sample complexity matches the lower bound in Theorem 10.7 in α when n is held constant and in n when α is held constant. We are going to use a trick called *Poissonisation* to simplify the proof of soundness and completeness, as in Batu et al. [2013]. Suppose that, rather than taking r samples from P, the algorithm first samples r_1 from a Poisson distribution with parameter $\lambda = r$ and then takes r_1 samples from P. Let X_i be the random variable corresponding to the number of times the element $i \in [n]$ appears in the sample from P. Then X_i is distributed identically to the Poisson distribution with parameter $\lambda = p_i r$ and all the X_i's are mutually independent. Similarly, we sample r_2 from a Poisson distribution with parameter r and then take r_2 samples from Q. Let Y_i be the the number of times i appears in the sample from Q, so Y_i is Poisson with $\lambda = q_i r$ and the Y_i are independent.

Theorem 10.8 (Upper bound). *Let $\epsilon, \delta, \alpha > 0$. Algorithm 4 is a (ϵ, δ)-aDP property tester with proximity parameter 2α and sample complexity $O(\frac{4n(1+e^{2\epsilon})}{\alpha^2})$.*

Proof. Let $p_i = P(i)$ and $q_i = Q(i)$. Let $Z_i = \frac{1}{r}\left(X_i - e^{\epsilon} Y_i\right)$ so $\mathbb{E}[Z_i] = p_i - e^{\epsilon} q_i$ and $\text{Var}(Z_i) \leq \frac{p_i + e^{2\epsilon} q_i}{r}$. Note also that (P, Q) is (ϵ, δ)-DP if $\Delta := \sum_{i=1}^{n} \max\{0, \mathbb{E}[Z_i]\} \leq \delta$. First note that $\mathbb{E}[Z] \geq \Delta$. If $\mathbb{E}[Z_i] \leq 0$ then

$$\begin{aligned}
\mathbb{E}[\max\{0, Z_i\}] &= \int_0^{\infty} \mathbb{P}(\max\{0, Z_i\} > x) dx \\
&= \int_0^{\infty} \mathbb{P}(Z_i > x) dx \\
&\leq \int_0^{\infty} \mathbb{P}(Z_i - \mathbb{E}[Z_i] > x - \mathbb{E}[Z_i]) dx \\
&\leq \int_0^{\infty} \min\{1, \frac{VarZ_i}{(x - \mathbb{E}[Z_i])^2}\} dx \\
&= \int_0^{\sqrt{VarZ_i} + \mathbb{E}[Z_i]} 1 dx + \int_{\sqrt{VarZ_i} + \mathbb{E}[Z_i]}^{\infty} \frac{VarZ_i}{(x - \mathbb{E}[Z_i])^2} dx \\
&= \sqrt{VarZ_i} + \mathbb{E}[Z_i] + \sqrt{VarZ_i} \\
&\leq \sqrt{\frac{p_i + e^{2\epsilon} q_i}{r}}
\end{aligned}$$

If $\mathbb{E}[Z_i] > 0$ then $\mathbb{E}[\max\{0, X_i\}] \leq \mathbb{E}[Z_i] + \sqrt{\frac{p_i + e^{2\epsilon} q_i}{r}} = p_i - e^{\epsilon} q_i + \sqrt{\frac{p_i + e^{2\epsilon} q_i}{r}}$. Therefore,

$$\mathbb{E}[Z] \leq \Delta + \sum_{i=1}^{n} \sqrt{\frac{p_i + e^{2\epsilon} q_i}{r}} \leq \Delta + \sqrt{\frac{n}{r}(1 + e^{2\epsilon})}$$

Now, let Z_i' be an independent copy of Z_i then

$$\begin{aligned}
Var[\max\{0, Z_i\}] &= \frac{1}{2} \mathbb{E}[(\max\{0, Z_i\} - \max\{0, Z_i'\})^2] \\
&= \int_0^{\infty} \mathbb{P}((\max\{0, Z_i\} - \max\{0, Z_i'\})^2 \geq x) dx \\
&\leq \int_0^{\infty} \mathbb{P}((Z_i - Z_i')^2 \geq x) dx \\
&= VarZ_i = \frac{p_i + e^{2\epsilon} q_i}{r}.
\end{aligned}$$

So $VarZ \leq \frac{1+e^{2\epsilon}}{r}$. Therefore,

$$\begin{aligned}\mathbb{P}(|Z-\Delta|\geq\alpha) &\leq \mathbb{P}(|Z-\mathbb{E}[Z]|+|\mathbb{E}[Z]-\Delta|\geq\alpha)\\ &\leq \mathbb{P}(|Z-\mathbb{E}[Z]|\geq\alpha-\sqrt{\frac{n}{r}}(1+e^{2\epsilon}))\\ &\leq \frac{VarZ}{(\alpha-\sqrt{\frac{n}{r}}(1+e^{2\epsilon}))^2}\\ &\leq \frac{1+e^{2\epsilon}}{r(\alpha-\sqrt{\frac{n}{r}}(1+e^{2\epsilon}))^2},\end{aligned}$$

which is less than $1/3$ if $r \geq \max\{\frac{4n(1+e^{2\epsilon})^2}{\alpha^2}, \frac{12(1+e^{2\epsilon})}{\alpha^2}\}$. □

10.4 Full Information (FI) Setting

The situation is substantially rosier if we have side-information. Although there are some realistic scenarios where one may have *trusted* side-information, we will focus on *untrusted* side-information. In particular, we allow our property tester to REJECT simply because the provided side-information is inaccurate. We will see that the untrusted side-information can still be useful since verifying information is often easier than estimating it.

The usefulness of side-information in property testing is informally lower bounded by how easy it is to generate the same information, and how much the information tells us about the property. For example, the means of the distributions (P_0, P_1) do not tell us very much about whether or not the privacy guarantee is satisfied. If $\mathcal{A}$ is an unbiased estimate for a function f, then the following proposition states that *knowing the function the black-box is computing does not help in verifying pDP*.

Proposition 10.9. *Let $\alpha, \epsilon > 0$ and suppose the side information is (a, b), which are purported to be the means of P_0 and P_1. If there exists a Ξ-private algorithm (P_0, P_1) supported on $[n-1]$ with $\mathbb{E}(P_0) = a$ and $\mathbb{E}(P_0) = b$, then no ϵ-pDP property tester with side information (a, b) and proximity parameter α has finite query complexity.*

The requirement that a Ξ-private algorithm with the right side information exists is necessary. If no such algorithm exists, then the tester should always REJECT and requires no queries. Under the assumption that such an algorithm exists, it is reasonable to assume that there exists an algorithm with slightly smaller support.

Proof. Let $A > 0$ and $\mathcal{A} = (P_0, P_1)$ be the algorithm promised. That is, (P_0, P_1) is Ξ-private and supported on $[n-1]$. Let $Q_0 = (1-e^{-A})P_0 + e^{-A}\chi_n$ and $Q_1 = P_1$ so (Q_0, Q_1) is α-far from pDP. Now, $\|P_0 \times P_1 - Q_0 \times Q_1\|_{\text{TV}} \leq e^{-A}$ so it requires $e^A \to \infty$ samples to distinguish between the two distributions. □

Algorithm 5 aDP FI Property Tester

```
Input: Universe size $n$, $\epsilon, \delta, \alpha > 0$, $(Q_0, Q_1)$ and identify tester $T$ with sample complexity $r$
if $(Q_0, Q_1)$ is not $(\epsilon, \delta)$-aDP then
    Output: REJECT
else
    if $T(Q_0, x \sim P_0^r) =$ REJECT or $T(Q_1, x \sim P_1^r) =$ REJECT then
        Output: REJECT
    else
        Output: ACCEPT
    end if
end if
```

We will focus on what we call the *full information* setting: we are given sample access to $\mathcal{A}$ and a distribution Q_D for each database D. In contrast to the mean, this side information is very informative about the privacy of the algorithm. It is also difficult to generate based on samples. We can *estimate* it using $\Theta(\frac{n}{\alpha^2})$ [Chan et al., 2014a] queries of each database, where α is the accuracy in TV-distance. However, we already know that the only privacy notion for which an estimate is sufficient is aDP. It is also known that testing identity to a known distribution requires asymptotically less samples than estimating an unknown distribution.

Proposition 10.10. *There exists a identity tester T such that Algorithm 10.10 is a (ϵ, δ)-aDP FI property tester with query complexity $O\left(\frac{\sqrt{n}}{\alpha^2}\right)$ and proximity parameter α.*

This is our first, and only, sublinear query complexity property tester for privacy. Since closeness in TV-distance implies closeness in aDP, we only need to check that the true distributions are close to (Q_0, Q_1) and that (Q_0, Q_1) is (ϵ, δ)-aDP. The difficult part is testing closeness of the distributions, for which we borrow from Chan et al. [2014b].

Proof. Chan et al. [2014b] proved that there exists a property tester T that takes as input a description of the discrete distribution P and $O(\frac{\sqrt{n}}{\alpha^2})$ samples from an distribution Q. If $P = Q$ then the tester ACCEPTS with probability at least 2/3 and if $\|P - Q\|_{\text{TV}} \geq \alpha$ then the tester REJECTS with probability at least 2/3. If we increase the sample complexity of the tester T by a constant factor then we can replace $2/3$ with $\sqrt{2/3}$.

To prove completeness, suppose $(P_0, P_1) = (Q_0, Q_1)$ and (Q_0, Q_1) is (ϵ, δ)-aDP. Since T ACCEPTS on both pairs (Q_0, P_0) and (Q_1, P_1) with probability at least $\sqrt{2/3}$, it ACCEPTS both with probability at least $2/3$. To prove soundness, suppose (P_0, P_1) is $(1 + e^\epsilon)\alpha$-far from (ϵ, δ)-aDP. Assume (Q_0, Q_1) is (ϵ, δ)-aDP because otherwise the tester REJECTS. It must be the case that either $\|Q_0 - P_0\|_{\text{TV}} \geq \alpha$ or $\|Q_1 - P_1\|_{\text{TV}} \geq \alpha$ because otherwise (P_0, P_1) would be $(\epsilon, (1 + e^\epsilon)\alpha)$-aDP by Lemma 10.6. Thus, with probability at least $2/3$ either $T(Q_0, x \sim P_0) =$ REJECT or $T(Q_1, x \sim P_1) =$ REJECT. □

Next, we show that for pDP, the side information allows us to obtain a finite query

Algorithm 6 pDP FI Property Tester

```
Input: Universe size $n$, $\epsilon, \alpha > 0$, $(Q_0, Q_1)$
$\lambda = \frac{\ln n}{\alpha^2\beta^2}$
Sample $r \sim$Poi$(\lambda)$
Sample $D_0 \sim P_0^r$, $D_1 \sim P_1^r$
for $i \in [m]$ do
   $x_i =$ number of $i$'s in $D_0$
   $y_i =$ number of $i$'s in $D_1$
end for
$\hat{\epsilon} = \sup_i \max\{\ln \frac{x_i}{y_i}, \ln \frac{y_i}{x_i}\}$
if $\hat{\epsilon} > \epsilon + 2\alpha$ then
   Output: REJECT
else
   if $\forall i \;\; e^{-\alpha} \leq \frac{x_i}{(Q_0)_i} \leq e^{\alpha}$ and $e^{-\alpha} \leq \frac{y_i}{(Q_1)_i} \leq e^{\alpha}$ then
      Output: ACCEPT
   else
      Output: REJECT
   end if
end if
```

complexity property tester. The side-information gives us an easy way to switch from a worst-case analysis to input specific upper bounds. We argue that

$$\beta = \min_E \min_D P_D(E),$$

where the first $\min$ is over events E and the second is over databases D, is the crucial quantity in understanding verifiability in the full information setting.

The lower bound proofs in the previous section all proceeded by finding two algorithms $\mathcal{A}$ and $\mathcal{B}$ that were close in TV-distance but had very different privacy parameters. The algorithms we chose all had one feature in common: the distributions P_D contained very low probability events. This property allowed us to drive the denominator of $\frac{Q_D(E)}{P_D(E)}$ to 0, and hence the privacy loss to ∞, while remaining close in TV-distance. This method works equally well in the full-information setting *if* low probability events exist in the distributions Q_D.

If the distribution Q_D does not have low probability events, then any distributions close to P_D must have bounded privacy parameters. To see this, suppose $(Q_0, Q_1) = (U, U)$ where U is the uniform distribution U on $\{\psi, \omega\}$. We can establish in approximately $\frac{1}{\alpha^2}$ samples whether or not P_0 and P_1 are both within TV-distance α of uniform. If not, then we REJECT. If so, then the worst case for privacy is $P_0 = (1/2 - \alpha)\chi_\psi + (1/2 + \alpha)\chi_\omega$ and $P_1 = (1/2 + \alpha)\chi_\psi + (1/2 - \alpha)\chi_\omega$. However, the increase in the pDP parameter from (Q_0, Q_1) to (P_0, P_1) is bounded by $\ln \frac{1/2+\alpha}{1/2-\alpha} \approx \alpha$.

Algorithm 6 is a full information property tester for pDP. Note that this algorithm is not sublinear in n since $\beta < \frac{1}{n}$.

Theorem 10.11 (pDP upper bound)**.** *Let $\epsilon > 0$ and $\alpha > 0$. Algorithm 6 is an ϵ-aDP FI property tester with proximity parameter 10α and query complexity $O\left(\frac{\ln n}{\alpha^2\beta^2}\right)$.*

Proof. We first show completeness. Suppose $(P_0, P_1) = (Q_0, Q_1)$ and $\mathcal{A}$ is ϵ-pDP. By the multiplicative Hoeffding's inequality,

$$\mathbb{P}\left(\frac{x_i}{(P_0)_i} \geq e^\alpha \text{ or } \frac{(P_0)_i}{x_i} \geq e^\alpha\right) \leq e^{-2r(e^\alpha-1)^2\beta^2} + e^{-2r(1-e^{-\alpha})^2\beta^2}.$$

Therefore,

$$\mathbb{P}\left(\exists i \text{ s.t. } \quad \frac{x_i}{(P_0)_i} \geq e^\alpha \text{ or } \frac{(P_0)_i}{x_i} \geq e^\alpha\right) \leq n(e^{-2r(e^\alpha-1)^2\beta^2} + e^{-2r(1-e^{-\alpha})^2\beta^2}) \leq \frac{1}{6}.$$

Thus with probability $2/3$ we have for all i, $e^{-\alpha} \leq \frac{x_i}{(P_0)_i} \leq e^\alpha$ and $e^{-\alpha} \leq \frac{y_i}{(P_1)_i} \leq e^\alpha$ and so

$$\frac{x_i}{y_i} = \frac{x_i}{(P_0)_i}\frac{(P_0)_i}{(P_1)_i}\frac{(P_1)_i}{y_i} \leq e^\alpha e^\epsilon e^\alpha.$$

This implies $\hat{\epsilon} \leq \epsilon + 2\alpha$ so we ACCEPT.

For soundness, we show that the ACCEPT conditions imply that (P_0, P_1) must be at least $(\epsilon + 10\alpha)$-pDP. The condition $e^{-\alpha} \leq \frac{x_i}{(Q_0)_i} \leq e^\alpha$ implies $|x_i - (Q_0)_i| \leq \max\{(e^\alpha - 1)(Q_0)_i, (1-e^{-\alpha})(Q_0)_i\}$. Also, by the additive Hoeffding's inequality

$$\begin{aligned}\mathbb{P}\Big(\exists i \text{ s.t. } |x_i - (P_0)_i| \geq (Q_0)_i \max\{(e^\alpha - 1), (1-e^{-\alpha})\}\Big) \\ \leq n\min\{e^{-r(e^\alpha-1)^2\beta^2}, e^{-r(1-e^{-\alpha})^2\beta^2}\} \\ \leq \frac{1}{6}.\end{aligned}$$

Therefore, with probability $2/3$,

$$|(P_0)_i - (Q_0)_i| \leq (Q_0)_i \max\{2(e^\alpha - 1), 2(1-e^{-\alpha})\} \leq 2\alpha(Q_0)_i$$

for sufficiently small α. This implies $\max\{\frac{(P_0)_i}{(Q_0)_i}, \frac{(Q_0)_i}{(P_0)_i}\} \leq e^{4\alpha}$. Similarly,

$$\max\{\frac{(P_1)_i}{(Q_1)_i}, \frac{(Q_1)_i}{(P_1)_i}\} \ \leq \ e^{4\alpha}.$$

Since $\hat{\epsilon} \leq \epsilon + 2\alpha$ we have

$$\frac{(P_0)_i}{(P_1)_i} = \frac{(P_0)_i}{(Q_0)_i}\frac{(Q_0)_i}{x_i}\frac{x_i}{y_i}\frac{y_i}{Q_1)_i}\frac{(Q_1)_i}{(P_1)_i} \leq e^{4\alpha+\alpha+\epsilon+2\alpha+\alpha+4\alpha}.$$

□

We now turn to lower bounding the query complexity of aDP testing in the FI setting. The sample complexity is tight in α but deviates by a factor of β.

Theorem 10.12 (pDP lower bound)**.** *Let $\alpha > 0$ and $\ln 2 > \epsilon > 0$. Given side information (Q_0, Q_1), any ϵ-pDP property tester with proximity parameter α has query complexity $\Omega\left(\frac{1}{\beta\alpha^2}\right)$.*

Proof. Let $\psi, \omega, \phi \in [n]$ and notice that $\beta < 1/2$ provided $n > 2$. To prove the lower bound let

$$Q_0 = \beta\chi_\psi + \beta\chi_\omega + (1-2\beta)\chi_\phi \text{ and } Q_1 = e^\epsilon\beta\chi_\psi + (2-e^\epsilon)\beta\chi_\omega + (1-2\beta)\chi_\phi$$

be the side-information. Then (Q_0, Q_1) is ϵ-pDP. Let

$$P_0 = e^{-\alpha}\beta\chi_\psi + (2-e^{-\alpha})\beta\chi_\omega + (1-2\beta)\chi_\phi \text{ and } P_1 = Q_1$$

so (P_0, P_1) is α-far from ϵ-pDP. Now,

$$\begin{aligned} D_{\text{KL}}(P_0|Q_0) &= \beta \ln \frac{\beta}{e^{-\alpha}\beta} + \beta \ln \frac{\beta}{(2-e^{-\alpha})\beta} = \beta\alpha + \beta \ln\left(1 - \frac{1-e^{-\alpha}}{2-e^{-\alpha}}\right) \\ &\leq \beta\alpha - \beta\alpha + \beta\alpha^2 = \beta\alpha^2. \end{aligned}$$

As in Theorem 10.5, we must have

$$r \geq \frac{2}{9}\frac{1}{D_{\text{KL}}(P_0|Q_0)} = \Omega\left(\frac{1}{\beta\alpha^2}\right).$$

□

Part V

Online Learning through the Lens of Differential Privacy

CHAPTER 11

Introduction

In the previous parts, data was released to us via some randomised process that obfuscated the information we were trying to learn. We viewed this randomisation as a hindrance to accurate recovery of the information. In Parts II and IV the uncertainty arose from the sampling process where we were uncertain that the sample accurately represented the underlying information. In Part III, we deliberately added noise to obscure the underlying information, and saw a direct negative relationship between the amount of noise added to the measurements and the accuracy of our recovered source vector. In the presence of both types of uncertainty, it is not always the case that the latter is counterproductive [Yu, 2013, Poggio et al., 2004]. In fact, perturbing the data can often lower the impact of the former type of uncertainty, sampling error, which is typically unavoidable. In this part of the thesis, we will explore how DP techniques can be used not only to maintain privacy, but to reduce regret in online learning algorithms. [1]

Online learning is the process of making predictions based on (possibly partial) knowledge of the outcome of previous predictions. Each round, the learner suffers a loss related to how wrong their prediction was. As an example, consider an advertising company deciding which adverts to show. Each day, they update their hypothesis on what each consumer will click on and lose revenue if their hypothesis is wrong. This framework has two main differences to the offline framework. Firstly, information is withheld from the learner and only released after they have made their prediction. Also, we consider the sequence of data to be adversarially chosen: the consumer is deliberately trying to mislead the advertiser. Unlike in the sampling setting, where we can bound the probability of obtaining a bad sample, an adversarially chosen sample is certain to be as unrepresentative of the underlying information as possible. Thus, the need to lower the impact of sampling error is even more prevalent in the online learning setting.

Stability of the output in the presence of small changes to input is a desirable feature of

[1] This Part is based on joint work with Chansoo Lee, Jacob Abernethy and Ambuj Tewari. Chansoo Lee did much of the heavy lifting and the ideas in Section 12.2.2 belong to him. Some of the prose in this Part was written by Chansoo, Jacob or Ambuj, although the presentation has been altered from the preprint [Abernethy et al., 2018] for the purposes of this thesis.

methods in statistics [Yu, 2013] and machine learning [Poggio et al., 2004]. Formal notions of stability and learning guarantees derived from them have been studied both for statistical [Bousquet and Elisseeff, 2002] and adversarial online learning [Ross and Bagnell, 2011]. Hardt et al. [2016] argues that training techniques used in deep learning promote stability as the key ingredient. In this Part, we use the DP lens to design and analyze algorithms for *online linear optimization*, an important family of problems in online learning. We use tools from DP to provide a clean analysis of the Follow-the-Perturbed-Leader algorithm in several settings.

We emphasize, at the outset, that our goal is *not* the design of low-regret algorithms that satisfy the privacy condition; indeed there is already substantial existing work along these lines [Jain et al., 2012, Thakurta and Smith, 2013, Tossou and Dimitrakakis, 2017, Agarwal and Singh, 2017]. Our goal is instead to show that, in and of itself, the DP methodology is quite well-suited to design randomized learning algorithms with excellent gaurantees. An excellent introduction to online learning can be found in Shalev-Shwartz [2012].

11.1 Online Linear Optimization

We define *Online Linear Optimization* (OLO) problem, which will be the main focus of this part of the thesis. Let $\mathcal{X} \subseteq \mathbb{R}^N$ be the learner's decision set, and $\mathcal{Y} \subseteq \mathbb{R}^N$ be the loss set. Suppose both sets are convex and bounded in dual norms and we refer to these norm bounds by $\|\mathcal{X}\|$ and $\|\mathcal{Y}\|_\star$ respectively.

We consider an oblivious adversary that chooses a sequence of loss vectors $\ell_t \in \mathcal{Y}$ ahead of time. At every round t, the learner chooses a vector $x_t \in \mathcal{X}$, and suffers loss $\langle x_t, \ell_t \rangle$. Note that the learner is allowed access to its private source of randomness in making its moves x_t. The learner's goal is to minimize the *expected regret* after T rounds:

$$\mathbb{E}\mathsf{Regret}_T = \mathbb{E} \textstyle\sum_{t=1}^T \langle x_t, \ell_t \rangle - \min_{x \in \mathcal{X}} \sum_{t=1}^T \langle x, \ell_t \rangle$$

where the expectations are over all of the learner's randomness. In the *loss-only setting*, losses are always positive: $\min_{x \in \mathcal{X}} \langle x, \ell_t \rangle \geq 0$ for all t. In the *loss/gain setting*, losses can be positive or negative. If $\mathcal{Y}$ and $\mathcal{X}$ are both bounded then the worst regret that can be suffered by any learner is linear in T.

We define a few shorthand notations: $L_t = \sum_{s=1}^t \ell_s$ is the cumulative loss, $x_t^* = \arg\min_{x \in \mathcal{X}} \langle x, L_t \rangle$ is the best action in hindsight at time t, and $L_t^* = \langle x_t^*, L_t \rangle$ is its total loss. A sequence $(a_1, \ldots, a_t)$ is abbreviated $a_{1:t}$.

11.2 Related work

The idea that answering a statistical query privately prevents overfitting has received significant attention in recent years [Dwork et al., 2015, Nissim and Stemmer, 2015, Bassily et al., 2016, Cummings et al., 2016]. In the online setting, the utility of the differential privacy as a stability notion for analysing *specific* online learning has been noted before. The connection between the exponential weight mechanism and the exponential mechanism (Lemma 2.9) was recognised in the early stages of DP [Dwork and Roth, 2014b]. Dwork et al. [2014] showed that the Gaussian mechanism results in a low-regret algorithm for online PCA.

Perturbation techniques for online learning have existed for some time. *Follow-the-Perturbed-Leader* (FTPL) relies on adding noise directly to the global loss objective [Kalai and Vempala, 2005]. Much of the work on analysing FTPL algorithms has been ad hoc and context specific [Kalai and Vempala, 2005, Theorem 2] [Devroye et al., 2013, van Erven et al., 2014, Syrgkanis et al., 2016]. Few recent works have proposed generic framework that provide some useful insights; Rakhlin et al. [2012] derived FTPL as a minimax strategy against a randomly simulated worst-case adversary. Abernethy et al. [2014] derived FTPL as a mirror descent with a stochastically smoothed dual function. However, none of the existing FTPL analyses make any explicit connections with DP as we do.

CHAPTER 12

Analysing Follow-the-Perturbed-Leader using Differential Privacy

In this chapter we describe a popular online learning algorithm that can be interpreted as a private mechanism, and use DP tools to prove generic regret bounds.

12.1 Follow-the-Leader

Let us begin by discussing the first algorithm that one might try: Follow-the-Leader (FTL). The FTL algorithm selects the exact optimal action based on the data seen so far:

$$x_{t+1}^{\text{FTL}} = \arg\min_{x \in \mathcal{X}} \langle x, L_{t-1} \rangle.$$

In the non-adversarial setting, FTL does just fine. In fact, a crucial fact used in statistical learning is that models (or classifiers) learnt on randomly sampled training data are likely to perform well on a random sample from the population. However, in the adversarial setting, FTL can suffer regret linear in T.

To see this consider the situation where $\mathcal{X} = \mathcal{Y} = [-1, 1]$. Suppose the adversary plays the sequence $\ell_1 = -0.5, \ell_2 = 1, \ell_3 = -1, \ell_4 = 1, \ell_5 = -1$, etc. A learner using FTL will respond by playing $x_1 = 0$, $x_2 = -1$, $x_3 = 1$, $x_4 = -1$, $x_5 = 1, \cdots$. Such a learner will incur regret T because they consistently incorrectly guess the sign of ℓ_t. A FTL learner is easily lead to overfitting the data: a small change in the adversaries data sequence can cause the learner to drastically alter their guess. The following result indicates that *stability* in the learner's predictions may be the key to better regret bounds.

Lemma 12.1. *Let $x_1, x_2, \cdots$ be the sequence of vectors produced by FTL. Then*

$$\mathsf{Regret}_T \leq \sum_{t=1}^{T} \langle x_t, \ell_t \rangle - \langle x_{t+1}, \ell_t \rangle$$

The proof of Lemma 12.1 can be found in [Shalev-Shwartz, 2012, Lemma 2.1].

12.2 Follow-the-Perturbed-Leader Analysis via One-Step Differential Privacy

We established in the previous section that it is in the learner's interest to play an action that is stable with respect to small changes in the input. Fortunately, differentially private algorithms are designed to achieve this guarantee! We will explore this connection through the analysis of a classical OLO algorithm: Follow-the-Perturbed Leader (FTPL). Given a noise distribution $\mathcal{D}$, FTPL selects the optimal action based on a perturbed version of the data seen so far:

$$x_{t+1}^{\text{FTPL}} = \arg\min_{x \in \mathcal{X}} \langle x, L_{t-1} + Z \rangle$$

where $Z \sim \mathcal{D}$. The name FTPL was coined by Kalai and Vempala [2005], although the main idea goes back to some of the earliest work in online learning [Hannan, 1957]. Note that, for oblivious adversaries, the expected regret does not depend on whether the same random Z is reused across all rounds or independent Z_t's are drawn at each round. We will assume the reused randomness case throughout this part.

The FTPL algorithm is closely related to a fictitious algorithm called Be-the-Perturbed-Leader (BTPL). BTPL also chooses the optimal action based on a perturbed version of the data, but it imagines that the learner knows ℓ_t when they are choosing x_t. That is,

$$x_{t+1}^{\text{BTPL}} = \arg\min_{x \in \mathcal{X}} \langle x, L_t + Z \rangle$$

where $Z \sim \mathcal{D}$. BTPL obviously performs better than FTPL since the learner *knows the future*. Differential privacy allows us to upper bound $\mathsf{Regret}(\text{FTPL})_T$ with a function of $\mathsf{Regret}(\text{BTPL})_T$, provided the map $L_t \mapsto L_t + Z$ is DP.

Definition 12.2 (One-step privacy)**.** An online learning algorithm is (ϵ, δ)*-one-step differentially private* if $D_\infty^\delta(x_t, x_{t+1}) \leq \epsilon$ for all $t = 1, \ldots, T$ given any loss sequence.

FTPL is (ϵ, δ)-one-step DP if $\mathcal{D}$ is chosen appropriately. One-step privacy is a powerful condition on the stability of FTPL (and online learning algorithms in general), from which we can derive generic regret bounds. The following theorem, relating privacy to regret, provides a powerful tool which we build on in the remainder of this part. In order to state the theorem more generally we introduce notation $\mathcal{A}^+$ for a fictitious algorithm that plays at time t what $\mathcal{A}$ would play at time $t+1$. If $\mathcal{A}$ is FTPL then $\mathcal{A}^+$ is BTPL.

Theorem 12.3. *If $\mathcal{A}$ is (ϵ, δ)-one-step DP for a loss-only OLO problem with $\epsilon \leq 1$, its expected regret is at most:*

$$2\epsilon L_T^* + 3\mathbb{E}[\mathsf{Regret}(\mathcal{A}^+)_T] + \delta \|\mathcal{X}\| \textstyle\sum_{t=1}^T \|\ell_t\|_\star$$

Proof. Using Lemma 2.4, we have for every t,

$$\mathbb{E}[\langle x_t, \ell_t \rangle)] \leq e^\epsilon \mathbb{E}[\langle x_{t+1}, \ell_t \rangle] + \delta \|\mathcal{X}\| \|\ell_t\|_\star.$$

By summing over t, we have

$$\begin{aligned}\mathbb{E}\left[\textstyle\sum_{t=1}^{T}\mathsf{Loss}(\mathcal{A})_t\right] &\le e^{\epsilon}\mathbb{E}[\sum_{t=1}^{T}\mathsf{Loss}(\mathcal{A}^{+})_t] + \delta\sum_{t=1}^{T}\|\mathcal{X}\|\|\ell_t\|_{\star}\\ &\le e^{\epsilon}(L_T^* + \mathbb{E}[\mathsf{Regret}(\mathcal{A}^{+})_T]) + \delta\textstyle\sum_{t=1}^{T}\|\mathcal{X}\|\|\ell_t\|_{\star}.\end{aligned}$$

Subtract L_T^* from each side and get:

$$(e^{\epsilon}-1)L_T^* + e^{\epsilon}\mathbb{E}[\mathsf{Regret}(\mathcal{A}^{+})_T] + \delta\|\mathcal{X}\|\textstyle\sum_{t=1}^{T}\|\ell_t\|_{\star}.$$

To complete the proof, we use the trivial upper bounds $e^{\epsilon} \le 1 + 2\epsilon \le 3$, which hold for $\epsilon \le 1$. □

Theorem 12.3 suggests a strategy for analysing FTPL algorithms: first show that the perturbation algorithm satisfies (ϵ, δ)-DP for some (ϵ, δ), then bound the regret of BTPL. Without perturbation BTPL has zero regret, since it will always choose the optimal action. In the presence of perturbation, BTPL suffers regret that does not grow in T but only in the magnitude of the noise and the size of $\mathcal{X}$ [Kalai and Vempala, 2005]:

$$\mathbb{E}[\mathsf{Regret}(\mathsf{BTPL})_T] \le \mathbb{E}_{Z\sim\mathcal{D}}[\sup_{x\in\mathcal{X}}\langle x, Z\rangle]. \tag{12.1}$$

12.2.1 General First-Order FTPL Bound

Abernethy et al. [2014] showed that FTPL with Gaussian noise is a universal OLO algorithm with regret $O(\|\mathcal{X}\|_2\|\mathcal{Y}\|_2\sqrt[4]{N}\sqrt{T})$. However, their analysis technique based on convex duality does not lead to first-order bounds in terms of L_T^* in the loss-only settings. In other words, one would prefer a bound that increases according to the loss of the best performing action, which grows linearly in T only in the most pessimistic scenarios.

With the analysis tools presented in this work, we establish that FTPL with Gaussian noise does enjoy, up to logarithmic factors, the first-order regret bound that scales in L_T^*. Put differently, FTPL with Gaussian noise is able to *adapt* to the input if there is a strong signal for the best action, a property that was not discovered in previous analysis.

Theorem 12.4. *Consider a* loss-only *OLO problem. Let* $R = \|\mathcal{X}\|_2\|\mathcal{Y}\|_2$. *FTPL with Gaussian noise achieves expected regret of order* $O\left(\sqrt[4]{N}\sqrt{RL_T^*\log T} + \sqrt{N}R\log T\right)$.

Proof. Let $\sigma = \epsilon^{-1}\|\mathcal{Y}\|_2 2\log(2/\delta)$, where ϵ, δ will be determined later. By Lemma 2.8 and Lemma 2.3, FTPL with $\mathcal{N}(0, \sigma I)$ is (ϵ, δ)-one-step DP with respect to $\mathcal{Y}$. Also note that the regret bound for the Gaussian BTPL is $\sigma\|\mathcal{X}\|_2\sqrt{2N}$.

We now apply Theorem 12.3 and get the regret bound:

$$2\epsilon L_T^* + 5\sigma\|\mathcal{X}\|_2\sqrt{N} + \delta\|\mathcal{X}\|_2\textstyle\sum_{t=1}^{T}\|\ell_t\|_* \le 2\epsilon L_T^* + 10\epsilon^{-1}R\sqrt{N}\log(2/\delta) + \delta TR.$$

Set $\delta = (2TR)^{-1}$, so that the last term becomes a constant. Then, choose

$$\epsilon = \min(\sqrt[4]{N}\sqrt{(R\log T)/L_T^*}, 1).$$

If $\epsilon = 1$, then we must have $L_T^* \leq \sqrt{N}R\log T$, then, which gives $O(\sqrt{N}R\log T)$ regret. Otherwise, we obtain $O(\sqrt[4]{N}\sqrt{RL_T^*\log T})$ regret. □

When $L_T^* \ll \|\mathcal{Y}\|_2 T$, our bound is a major improvement over $O(R\sqrt[4]{N}\sqrt{T})$ given by Abernethy et al. [2014]. For example, when $L_T^* = O(R\sqrt{T})$, our bound gives $O(R\sqrt[4]{NT})$. Note that we tuned both ϵ and δ as part of the analysis. This was allowed because low regret is our goal, rather than privacy.

12.2.2 Online Sparse PCA

In offline Principle Component Analysis (PCA), data in $\mathbb{R}^n$ are projected onto a k-dimensional subspace. The goal is, given data, to find the rank k projection matrix that minimises the compression loss $\sum_t \|Xx_t - x_t\|_2^2$. In the online version, at each round the learner chooses a k-dimensional subspace, or equivalently a rank k projection matrix X_t. The next datapoint is then revealed and the learner suffers the compression loss $\|X_t x_t - x_t\|_2^2$. Due to the structure of Regret_T, we can replace this loss function with $x_t^\top X_t x_t$. In this section we show that, for Online Sparse PCA, there is a simple FTPL algorithm that achieves the optimal regret.

Let $\mathrm{S^N}$ be the set of $N \times N$ symmetric real matrices, and $\lambda : \mathrm{S^N} \to \mathbb{R}^N$ be the function that outputs the eigenvalues of a matrix in decreasing order. The *spectral norm* of a matrix $X \in \mathrm{S^N}$ is the ℓ_∞-norm of $\lambda(X)$, denoted $\|\lambda(X)\|_\infty$. An *orthogonal invariant ensemble (OIE)* is a distribution over matrices such that for an arbitrary matrix A in its support, any orthogonal transformation of A is also in the support and has the same density as A.

Online *Sparse* PCA is an OLO problem where

$$\mathcal{X} = \{X : X \in \mathrm{S^N}, 0 \preceq X \preceq I \text{ and } \mathrm{tr}(X) = k\}$$

and $\mathcal{Y} = \{aa^\top : a \in \mathbb{R}^N, \|a\|_2 = 1\}$ is the set of rank-1 matrices with eigenvalue 1. The loss function is $a^\top Xa = \sum_{i,j}(aa^\top)_{ij}X_{ij}$, making it an N^2-dimensional problem. The optimal regret is $O(\sqrt{L^* k\log(N/k)} + k\log(N/k))$, while the best known FTPL algorithm by Dwork et al. [2014] achieves $O(N^{\frac{1}{4}}\sqrt{kL^*\log T})$ regret using Gaussian Orthogonal Ensemble.

The *Laplace-on-Diagonal Orthogonal Invariant Ensemble* (LOD) with scaling parameter $1/\epsilon$ has probability density function $p(Z) \propto \exp(-\epsilon\|\lambda(Z)\|_1)$. We propose the FTPL algorithm with Z distributed according to the LOD distribution for Online Sparse PCA. As per Theorem 12.3 we need to show both that BTPL has low regret and adding LOD noise satisfies DP.

Lemma 12.5. *LOD mechanism, defined as $\mathcal{M} : \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}$ with $\mathcal{M}(A) = A + Z$, where Z is a sample from $\mathrm{LOD}(u/\epsilon)$, is ϵ-differentially private with respect to the set $\{X \in \mathbb{R}^{N\times N} : \|\lambda(X)\|_1 \le u\}$.*

Proof. We will prove this by showing a generic reduction technique to the vector case. In particular, suppose that a distribution $\mathcal{D}$ over matrices has density function of the form $p(Z) = Cq(\|\lambda(Z)\|)$ for a normalizing constant C, arbitrary function of vectors q, and some norm $\|\cdot\|$. Then, we will show that the privacy guarantee of distribution $\mathcal{D}'$ over vectors whose density function is some constant times q extends to the matrices.

Let A, A', B be matrices. Then,

$$\tfrac{p(B-A)}{p(B-A')} = \tfrac{q(\|\lambda(B-A)\|)}{q(\|\lambda(B-A')\|)}.$$

By triangle inequality, $\|\lambda(B-A)\| - \|\lambda(B-A')\| \le \|\lambda(A-A')\|$. So,

$$\sup_{\substack{A,A',B\in\mathbb{R}^{N\times N}\\ \|\lambda(A-A')\|\le u}} \frac{p(B-A)}{p(B-A')} \le \sup_{\substack{a,a'\in\mathbb{R}^{N}\\ \|a-a'\|\le u}} \frac{q(a)}{q(a')}.$$

Hence, if adding a noise from $\mathcal{D}'$ achieves ϵ-DP with respect to a set of vectors bounded in $\|\cdot\|$, then adding a noise from $\mathcal{D}$ achieves ϵ-DP with respect to a set of matrices bounded in $\|\lambda(\cdot)\|$. □

Lemma 12.6 (k-sparse Online PCA)**.** *BTPL algorithm with Laplace-on-Diagonal ensemble for the k-sparse Online PCA problem has expected regret at most* $k(\log(N/k) + 1)/\epsilon$.

Proof. For Online k-Sparse PCA problem, $\sup_{x\in\mathcal{X}}\langle x, Z\rangle$ is the is the sum of k-largest eigenvalues of Z. When $\mathcal{D}$ is LOD ensemble, these eigenvalues follow the Laplace distribution. By [Neu, 2015, Lemma 10], the expected sum of k largest coordinates of an N-dimensional vector sampled from the Laplacian distribution is $k(\log(N/k)+1)/\epsilon$. □

Corollary 12.7. *FTPL with LOD achieves* $O(\sqrt{L^* k \log(N/k)} + k\log(N/k))$ *expected regret on Online k-Sparce PCA.*

Proof. Given Lemma 12.5 and Lemma 12.6 all we need to complete the proof is to apply Theorem 12.3 with $\epsilon = \min(\sqrt{k(1+\log(N/k))/L^*}, 1)$. □

CHAPTER 13

The Experts Setting, Hazard Rates and Privacy

We will now turn our attention to the classical online learning setting of *prediction with expert advice*, or often known as the *experts setting*. In short, one imagines a set of experts each making a prediction on every round, and a learner that must maintain a belief distribution over the experts, in order to form a merged prediction; for background please see, e.g., [Cesa-Bianchi and Lugosi, 2006, Littlestone and Warmuth, 1994, Freund and Schapire, 1997]. We show in this section first-order optimal regret bounds for experts setting that apply to FTPL with any finite hazard rate distribution.

The experts setting is in fact a canonical online linear optimization problem: we let $\mathcal{X}$ be the probability simplex with N vertices, denoted Δ^N, and $\mathcal{Y} = [0,1]^N$. A central result in online learning is that the minimax regret in the experts setting is $O(\sqrt{L_T^* \log N} + \log N)$ [Abernethy et al., 2008]. Our main result in this section (Theorem 13.1) provides a generic sufficient condition for the distributions that FTPL can use to match the minimax regret.

Theorem 13.1. *For the loss-only experts setting, FTPL with Laplace, Gumbel, Frechet, Weibull, and Pareto noise (i.i.d. for each of N coordinates), with a proper choice of distribution parameters, all achieve $O(\sqrt{L_T^* \log N} + \log N)$ expected regret.*

Although we are not the first to find FTPL with the above regret bound, L_T^* bound for FTPL with any of the mentioned noise is not found in the literature. In fact, previous FTPL algorithms with L_T^* regret bound all relied on one-sided perturbation that *subtract* from the cumulative loss; Kalai and Vempala [2005] used the negative exponential noise and van Erven et al. [2014] the dropout noise that is effectively a negative Bernoulli noise.

Symmetric distributions, on the other hand, were previously shown to achieve only $O(\sqrt{T})$ regret: such as Gaussian noise [Abernethy et al., 2014], random-walk noise [Devroye et al., 2013], and a large family of symmetric noises [Rakhlin et al., 2012]. Our DP-based analysis shows that such discrepancy was merely due to the lack of proper analysis tools. We will use the analysis framework established in the previous Chapter: establishing privacy guarantees and bounding the regret of BTPL.

For this section, let $\mathcal{D}$ be an absolutely continuous distribution over $\mathbb{R}$ with probability density function $\mu_{\mathcal{D}}$ and cumulative density function $\Phi_{\mathcal{D}}$. Let $\tilde{f}_{\mathcal{D}}$ and $\mathcal{M}_{\mathcal{D}}$ be functions from $\mathbb{R}^N$ to $\mathbb{R}$ defined as $\tilde{f}_{\mathcal{D}}(x) = \mathbb{E}[\max_{i\in[N]}(x_i + Z_i)]$ and $\mathcal{M}_{\mathcal{D}}(x) = x + Z$, where Z in both definitions is a vector of N i.i.d. samples from $\mathcal{D}$.

In the experts setting, the output of FTPL is always a vertex of the simplex. In fact, the learner always plays $i^* = \arg\max_{i\in[N]}(x_i + Z_i)$. By Bertsekas [1972], we can swap the order of differentiation and expectation so $\nabla\tilde{f}_{\mathcal{D}}(x) = \mathbb{E}[e_{i^*}]$. That is, $\nabla_i\tilde{f}_{\mathcal{D}}(x)$ is the probability that FTPL algorithm would play $\mathbf{e}_i$ given a cumulative loss vector x.

13.1 Differential Consistency and Privacy

Abernethy et al. [2015], introduced a new kind of smoothness property called *differential consistency*. They showed that differential consistency is a key component of stable online learning algorithms. In this section, we show that privacy plays a similar role to differential consistency in the analysis of OLO algorithms.

Definition 13.2 (Differential Consistency)**.** We say that a function $f : \mathbb{R}^N \to \mathbb{R}$ is ϵ - differentially consistent if f is twice-differentiable and $\nabla^2_{ii} f \leq \epsilon\nabla_i f$ for all $i \in [N]$.

Since Differential consistency is inherently a *continuous* notion, we need to alter our privacy notion. *Lipschitz privacy* is a variant of DP where we have a continuous, rather than discrete, definition of neighbours.

Definition 13.3 (Lipschitz Privacy)**.** We say that a mechanism $\mathcal{M}$ is (ϵ, δ)-Lipschitz private with respect to a norm $\|\cdot\|$ if for all $a, a' \in \mathrm{dom}(\mathcal{M})$,

$$D^{\delta}_{\infty}(\mathcal{M}(a), \mathcal{M}(a')) \leq \epsilon\|a - a'\|.$$

The OLO analysis in [Abernethy et al., 2015] proceeds by showing that the differential consistency property can be used to bound the *divergence penalty* term of the expected regret of an FTPL algorithm, which measures how much the learner's guess depends on the most recently received loss function ℓ_{t-1}. The next proposition shows that we can by-pass this analysis by appealing to privacy.

Proposition 13.4. *If $\tilde{f}_{\mathcal{D}}$ is differentially consistent, then the mapping from $a \in \mathbb{R}^N$ to a random sample drawn from $\nabla\tilde{f}_{\mathcal{D}}(a)$ is ϵ-Lipschitz private with respect to $\|\cdot\|_1$.*

Conversely, if $\nabla\tilde{f}_{\mathcal{D}}(a)$ is ϵ-Lipschitz private with respect to $\|\cdot\|_1$ then $\tilde{f}_{\mathcal{D}}$ is differentially consistent.

Proof. First, note that the second derivative vector $\nabla^2_{i\cdot}\tilde{f}_{\mathcal{D}} = (\nabla^2_{i1}\tilde{f}_{\mathcal{D}}, \dots, \nabla^2_{iN}\tilde{f}_{\mathcal{D}})$ satisfies that the i-th coordinate is the only positive coordinate, and that its coordinates add up to 0 [Abernethy et al., 2014]. So, $\|\nabla^2_{i\cdot}\tilde{f}_{\mathcal{D}}\|_\infty = \nabla^2_{ii}\tilde{f}_{\mathcal{D}}$.

Define $q_i(u) = \nabla_i \tilde{f}_{\mathcal{D}}(a' + (a - a')u)$. Its derivative is

$$\begin{aligned} q_i'(u) &= \langle \nabla^2_{i\cdot} \tilde{f}_{\mathcal{D}}(a + (a - a')u), a' - a \rangle \\ &\leq \|\nabla^2_{i\cdot} \tilde{f}_{\mathcal{D}}(a + (a - a')u)\|_\infty \|a' - a\|_1 \\ &\leq \nabla^2_{ii} \tilde{f}_{\mathcal{D}}(a + (a' - a)u)\|a' - a\|_1 \\ &\leq \epsilon \nabla_i \tilde{f}_{\mathcal{D}}(a + (a' - a)u)\|a' - a\|_1 \\ &= \epsilon q_i(u)\|a' - a\|_1 \end{aligned}$$

The last inequality is from our differential consistency assumption. It follows that for any $u \in [0, 1]$, we have

$$\frac{q_i'(u)}{q_i(u)} = \frac{d}{du} \log(q_i(u)) \leq \epsilon \|a' - a\|_1$$

and therefore

$$\begin{aligned} \ln \tfrac{\nabla_i \tilde{f}_{\mathcal{D}}(a)}{\nabla_i \tilde{f}_{\mathcal{D}}(a')} &= \log q_i(1) - \log q_i(0) = \int_0^1 \frac{d}{du} \log(q_i(u))\, du \\ &\leq \epsilon \|a' - a\|_1. \end{aligned}$$

Now, suppose $\nabla \tilde{f}_{\mathcal{D}}(a)$ is ϵ-Lipschitz private so for all $i \in [N]$ we have

$$\ln \frac{\nabla_i \tilde{f}_{\mathcal{D}}(a)}{\nabla_i \tilde{f}_{\mathcal{D}}(b)} \leq \epsilon \|a - b\|_1.$$

Now, restricting to the case where a and b only differ in the ith coordinate this implies $\frac{\ln \nabla_i \tilde{f}_{\mathcal{D}}(a) - \nabla_i \tilde{f}_{\mathcal{D}}(b)}{|a_i - b_i|} \leq \epsilon$. Taking the limit as $b \to a$ we recover

$$\frac{\nabla_{ii} \tilde{f}_{\mathcal{D}}(a)}{\nabla_i \tilde{f}_{\mathcal{D}}(a)} = \frac{\partial}{\partial x_i} \ln \nabla_i \tilde{f}_{\mathcal{D}}(a) \leq \epsilon.$$

□

13.2 Proof of Theorem 13.1

The commonality between the noise models listed in Theorem 13.1 is that they have bounded Hazard rates. The Hazard rate of a distribution is a statistical tool for survival analysis that measures how fast a distribution's tail decays. The lower the Hazard rate, the slower the tail decays.

Definition 13.5 (Hazard Rate)**.** The hazard rate of $\mathcal{D}$ at a is $\mathrm{haz}_{\mathcal{D}}(a) = \frac{\mu_{\mathcal{D}}(a)}{1 - \Phi_{\mathcal{D}}(a)}$. The (maximum) hazard rate of $\mathcal{D}$ is

$$\mathrm{haz}_{\mathcal{D}} = \sup_{a \in \mathrm{support}(\mathcal{D})} \mathrm{haz}_{\mathcal{D}}(a).$$

Abernethy et al. [2015] showed that if $\mathrm{haz}_{\mathcal{D}} \leq \epsilon$, then $\tilde{f}_{\mathcal{D}}$ is ϵ-differentially consistent. The following corollary then follows directly from Proposition 13.4 since $\nabla \tilde{f}_{\mathcal{D}}(a)$ is the probability vector for FTPL.

Corollary 13.6. *If* $\mathrm{haz}_{\mathcal{D}} \leq \epsilon$, *then FTPL with* $\mathcal{D}^N$ *(sampling* N *i.i.d. samples from* $\mathcal{D}$ *to generate noise) is* ϵ*-one-step Lipschitz private.*

The final ingredient of our analysis is a bound on the BTPL regret.

Lemma 13.7. *for all distributions mentioned in Theorem 13.1,* $\mathsf{Regret}(\mathit{BTPL})_T$ *is of order* $(\log N)/\epsilon$.

Proof. By Equation (12.1),

$$\mathbb{E}[\mathsf{Regret}(\mathsf{BTPL})_T] \leq \mathbb{E}_{Z \sim \mathcal{D}}[\sup_{x \in \mathcal{X}} \langle x, Z \rangle] = \mathbb{E}_{Z \sim \mathcal{D}}[\max_i Z_i].$$

For each distribution, the expected maximum of N draws is asymptotically bounded by $(\log N)/\epsilon$ [Abernethy et al., 2015]. □

Proof of Theorem 13.1. All listed distributions have max hazard rate of ϵ (for the parameter choice, see Abernethy et al. [2015]). From Corollary 13.6 and post-processing immunity (Lemma 2.3), we conclude that FTPL with any of the listed distributions is ϵ-Lipschitz private with respect to $\|\cdot\|_1$. The loss set for experts setting, however, is bounded in the ∞-norm.

To address this gap, we will show that from the privacy perspective, the worst case is when ℓ_t has only one non-zero element and thus $\|\ell_t\|_1 = \|\ell_t\|_\infty$.

In the experts setting, the output of FTPL is always a vertex of the simplex. Consider an arbitrary noise vector Z. If $L_{t,i} + Z_i < L_{t,j} + Z_j$, then $L_{t,i} + z_i < L_{t,j} + Z_j + \alpha$ for any $\alpha > 0$. So, $\{Z \in \mathbb{R}^N : \mathbf{e}_i = \mathcal{O}(L_t + Z)\} \subseteq \{Z \in \mathbb{R}^N : \mathbf{e}_i = \mathcal{O}(L_t + Z + \ell^{(-i)})\}$ for any loss vector $\ell^{(-i)} \in \mathcal{Y}$ whose i-th coordinate is zero. In other words, adding any loss to coordinates other than i can only increase the probability of playing $\mathbf{e}_i$. So, for any fixed $\ell_{1,t-1} \in \mathcal{Y}^{t-1}$,

$$\begin{aligned}
\sup_{i \in [N], \ell_t \in \mathcal{Y}} \frac{\mathbb{P}[x_t^{\mathrm{FTPL}} = \mathbf{e}_i]}{\mathbb{P}[x_{t+1}^{\mathrm{FTPL}} = \mathbf{e}_i]} &= \sup_{i \in [N]} \frac{\mathbb{P}[x_t^{\mathrm{FTPL}} = \mathbf{e}_i]}{\inf_{\ell_t \in \mathcal{Y}} \mathbb{P}[x_{t+1}^{\mathrm{FTPL}} = \mathbf{e}_i]} \\
= \sup_{i \in [N]} \frac{\mathbb{P}[x_t^{\mathrm{FTPL}} = \mathbf{e}_i]}{\inf_{\ell_t : \|\ell_t\|_1 \leq 1} \mathbb{P}[x_{t+1}^{\mathrm{FTPL}} = \mathbf{e}_i]} \\
= \sup_{\ell_t : \|\ell_t\|_1 \leq 1} \sup_{i \in [N]} \frac{\mathbb{P}[x_t^{\mathrm{FTPL}} = \mathbf{e}_i]}{\mathbb{P}[x_{t+1}^{\mathrm{FTPL}} = \mathbf{e}_i]}.
\end{aligned}$$

Applying Theorem 12.3 with $\epsilon = \min(\sqrt{\mathsf{Regret}(\mathsf{BTPL})_T / L_T^*}, 1)$ completes the proof. □

Bibliography

Emmanuel Abbe and Colin Sandon. Recovering communities in the general stochastic block model without knowing the parameters. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 676–684. Curran Associates, Inc., 2015.

Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.

Jacob Abernethy, Manfred K Warmuth, and Joel Yellin. Optimal strategies from random walks. In *COLT*, 2008.

Jacob Abernethy, Chansoo Lee, Abhinav Sinha, and Ambuj Tewari. Online linear optimization via smoothing. In *COLT*, 2014.

Jacob Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *NIPS*, 2015.

Jacob Abernethy, Chansoo Lee, Ambuj Tewari, and Audra McMillan. Online linear optimization through the differential privacy lens. arXiv:1711.10019v2, 2018.

Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3591–3599. Curran Associates, Inc., 2015.

John M. Adowd and Ian M. Schmutte. Revisiting the economics of privacy: population statistics and confidentiality protection as public goods. https://www2.census.gov/ces/wp/2017/CES-WP-17-37.pdf, 2017.

Naman Agarwal and Karan Singh. The price of differential privacy for online learning. In *ICML*, 2017.

E. M. Airoldi, T. Costa, and S. Chan. A non-parametric perspective on network analysis: Theory and consistent estimation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26, pages 692–700, 2013.

Apple. What's new in ios: ios 10.0. *Apple Developer Guide*, 2017. URL `https://developer.apple.com/library/content/releasenotes/General/WhatsNewIniOS/Articles/iOS10.html#//apple_ref/doc/uid/TP40017084-SW1`.

Debapratim Banerjee. Contiguity results for planted partition models: the dense case. *arXiv:1609.02854v1*, September 2016.

Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *COLT*, 2016.

Michael Barbaro and Tom Zeller Jr. A face is exposed for AOL searcher no. 4417749. *The New York Times*, 2006.

Rina Foygel Barber and John C. Duchi. Privacy and statistical risk: Formalisms and minimax bounds. arXiv:1412.4451v1, 2014.

Gilles Barthe, Boris Köpf, Federico Olmedo, and Santiago Zanella Béguelin. Probabilistic relational reasoning for differential privacy. *SIGPLAN Not.*, 47(1):97–110, January 2012.

Gilles Barthe, George Danezis, Benjamin Gregoire, Cesar Kunz, and Santiago Zanella-Beguelin. Verified computational differential privacy with applications to smart metering. In *Proceedings of the 2013 IEEE 26th Computer Security Foundations Symposium*, CSF '13, pages 287–301, 2013.

Gilles Barthe, Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, César Kunz, and Pierre-Yves Strub. Proving differential privacy in hoare logic. In *Proceedings of the 2014 IEEE 27th Computer Security Foundations Symposium*, CSF '14, pages 411–424, 2014.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *STOC*, 2016.

Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4:1–4:25, February 2013.

Sara Beddiaf, Laurent Autrique, Laetitia Perez, and Jean-Claude Jolly. Heating source localization in a reduced time. 26, 02 2015.

David A. Belsley, Edwin Kuh, and Roy E. Welsch. *Regression Diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, New York, Chichester, 1980.

Alex Beltran, Varick L. Erickson, and Alberto E. Cerpa. Thermosense: Occupancy thermal based sensing for hvac control. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, BuildSys'13, pages 11:1–11:8, New York, NY, USA, 2013. ACM.

Brett Bernstein and Carlos Fernandez-Granda. Deconvolution of Point Sources: A Sampling Theorem and Robustness Guarantee. *arXiv:1707.00808*, July 2017.

D. P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals with an application in stochastic programming. In *Proceedings of the 1972 IEEE Conference on Decision and Control and 11th Symposium on Adaptive Processes*, pages 555–559, Dec 1972.

P. J. Bickel and A. Chen. A nonparametric view of network models and newman-girvan and other modularities. *Proc. Natl. Acad. Sci. USA*, 106:21068–21073, 2009.

P. J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *Ann. Statist.*, 39(5):2280–2301, 2011.

C. Borgs, J. T. Chayes, L. Lovász, V. Sós, and K. Vesztergombi. Counting graph homomorphisms. In *Topics in Discrete Mathematics (eds. M. Klazar, J. Kratochvil, M. Loebl, J. Matousek, R. Thomas, P.Valtr)*, pages 315–371. Springer, 2006.

C. Borgs, J. T. Chayes, L. Lovász, V. Sós, and K. Vesztergombi. Convergent graph sequences I: Subgraph frequencies, metric properties, and testing. *Adv. Math.*, 219:1801–1851, 2008.

C. Borgs, J. T. Chayes, H. Cohn, and Y. Zhao. An L^p theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions. *arXiv:1401.2906*, 2014a.

C. Borgs, J. T. Chayes, H. Cohn, and Y. Zhao. An L^p theory of sparse graph convergence II: LD convergence, quotients, and right convergence. *arXiv:1408.0744*, 2014b.

Christian Borgs, Jennifer T. Chayes, and Adam Smith. Private graphon estimation for sparse graphs. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 1369–1377, Cambridge, MA, USA, 2015. MIT Press.

Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2002.

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the Forty-sixth Annual ACM Symposium on Theory of Computing*, STOC '14, pages 1–10, New York, NY, USA, 2014. ACM.

Martin Burger, Yanina Landa, Nicolay M. Tanushev, and Richard Tsai. *Discovering a Point Source in Unknown Environments*, pages 663–678. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

Emmanuel J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589 – 592, 2008.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006. ISBN 978-0-521-84108-5.

S. H. Chan and E. M. Airoldi. A consistent histogram estimator for exchangeable graph models. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 32:208–216, 2014.

Siu-On Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pages 1844–1852, Cambridge, MA, USA, 2014a. MIT Press.

Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 1193–1203, Philadelphia, PA, USA, 2014b. Society for Industrial and Applied Mathematics. ISBN 978-1-611973-38-9.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *Ann. Statist.*, 43(1):177–214, 2015.

Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In *COLT*, 2016.

Constantinos Daskalakis, Gautam Kamath, and John Wright. *Which Distribution Distances are Sublinearly Testable?*, pages 2747–2764. 2018.

Luc Devroye, Gábor Lugosi, and Gergely Neu. Prediction by random-walk perturbation. In *COLT*, 2013.

Apple Differential Privacy Team. Learning with privacy at scale. https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html, 2017.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems 30*, December 2017.

Kashyap Dixit, Madhav Jha, Sofya Raskhodnikova, and Abhradeep Thakurta. Testing the lipschitz property over product distributions with applications to data privacy. In Amit Sahai, editor, *Theory of Cryptography*, pages 418–436, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.*, 13(1): 3475–3506, December 2012.

Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of ComputationTAMC*. Springer Verlag, April 2008.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9:211–407, August 2014a.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014b.

Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. arXiv:1603.01887v2, 2016.

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

Cynthia Dwork, Frank McSherry, and Kunal Talwar. The price of privacy and the limits of lp decoding. In *Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing*, STOC '07, pages 85–94, New York, NY, USA, 2007. ACM.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Differential privacy - a primer for the perplexed. In *Conf. of European Statisticians*, Joint UNECE/Eurostat work session on statistical data confidentiality, 2011.

Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *STOC*, 2014.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *STOC*, 2015.

Claude E. Shannon. A mathematical theory of communication. 27:379–423, 01 1948.

Hamid Ebadi, David Sands, and Gerardo Schneider. Differential privacy: Now it's getting personal. In *Proceedings of the 42Nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '15, pages 69–81, 2015.

Günther Eibl and Dominik Engel. Differential privacy for real smart metering data. *Computer Science - Research and Development*, pages 1–10, 2016.

William Enck, Peter Gilbert, Byung-Gon Chun, Landon P. Cox, Jaeyeon Jung, Patrick McDaniel, and Anmol N. Sheth. Taintdroid: An information-flow tracking system for realtime privacy monitoring on smartphones. In *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation*, OSDI'10, pages 393–407, 2010.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 1054–1067, New York, NY, USA, 2014. ACM.

Lejun Fan, Yuanzhuo Wang, Xueqi Cheng, and Shuyuan Jin. Quantitative analysis for privacy leak software with privacy petri net. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*, ISI-KDD '12, pages 7:1–7:9, 2012.

Brittan Farmer, Cassandra Hall, and Selim Esedolu. Source identification from line integral measurements and simple atmospheric models. *Inverse Problems and Imaging*, 7 (2):471–490, 2013.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1): 119 – 139, 1997.

FTC. Ftc staff report: Internet of Things: Privacy & Security in a Connected World. Technical report, Federal Trade Commission, January 2015.

Marco Gaboardi, Andreas Haeberlen, Justin Hsu, Arjun Narayan, and Benjamin C. Pierce. Linear dependent types for differential privacy. *SIGPLAN Not.*, 48(1):357–370, January 2013.

Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-optimal graphon estimation. *Ann. Statist.*, 43(6):2624–2652, 12 2015. doi: 10.1214/15-AOS1354.

Q. Geng and P. Viswanath. Optimal noise adding mechanisms for approximate differential privacy. *IEEE Transactions on Information Theory*, 62(2):952–969, Feb 2016.

Anna Gilbert and Audra McMillan. Local differential privacy for physical sensor data and sparse recovery. *Conference on Information Systems and Sciences*, 2018.

Allan Greenleaf, Yaroslav Kurylev, Matti Lassas, and Gunther Uhlmann. Cloaking devices, electromagnetic wormholes, and transformation optics. *SIAM Review*, 51(1):3–33, 2009.

E. Haber. Numerical methods for optimal experimental design of large-scale ill-posed problems. *Inverse Problems*, 24, 2008.

Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Random differential privacy. *Journal of privacy and confidentiality*, 4(2):43–59, 2012.

James Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 1957.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: stability of stochastic gradient descent. In *ICML*, 2016.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Information Systems Audit and Control Association (ISACA). Geolocation: Risk, issues and strategies. ISACA white paper, 2011.

Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta. Differentially private online learning. In *JMLR*, 2012.

Márk Jelasity and Kenneth P. Birman. Distributional differential privacy for large-scale smart metering. In *Proceedings of the 2Nd ACM Workshop on Information Hiding and Multimedia Security*, pages 141–146, New York, NY, USA, 2014. ACM.

Zhanglong Ji, Zachary C. Lipton, and Charles Elkan. Differential privacy and machine learning: a survey and review. arXiv:1412.7584, 2014.

Jaeyeon Jung, Anmol Sheth, Ben Greenstein, David Wetherall, Gabriel Maganis, and Tadayoshi Kohno. Privacy oracle: A system for finding application leaks with black box differential testing. In *Proceedings of the 15th ACM Conference on Computer and Communications Security*, CCS ’08, pages 279–288, 2008.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *J. Mach. Learn. Res.*, 17(1):492–542, January 2016. ISSN 1532-4435.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 2005.

Olga Klopp, Alexandre Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation (version 1). *arXiv:1507.04118v1*, July 2015.

Olga Klopp, Alexandre Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation (version 3). *arXiv:1507.04118v3*, March 2016.

Y. Landa, N. Tanushev, and R. Tsai. Discovery of point sources in the Helmholtz equation posed in unknown domains with obstacles. *Comm. in Math. Sci.*, 9:903–928, 2011.

A. Lapidoth. *A Foundation in Digital Communication*. Cambridge University Press, 2017.

Pierre Latouche and Stéphane Robin. Variational bayes model averaging for graphon functions and motif frequencies inference in w-graph models. *Statistics and Computing*, 26(6):1173–1185, 2016.

C. Li, P. Zhou, and T. Jiang. Differential privacy and distributed online learning for wireless big data. In *2015 International Conference on Wireless Communications Signal Processing (WCSP)*, pages 1–5, Oct 2015.

Yang D. Li, Zhenjie Zhang, Marianne Winslett, and Yin Yang. Compressive mechanism: Utilizing sparse representation in differential privacy. In *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society*, WPES '11, pages 177–182, New York, NY, USA, 2011. ACM.

Yingying Li, Stanley Osher, and Richard Tsai. Heat source identification based on L1 constrained minimization. *Inverse Problems and Imaging*, 1(1), 2014.

Craig Lin, Clifford C. Federspiel, and David M. Auslander. Multi-sensor single-actuator control of hvac systems. In *Proc. Int. Conf. Enhanced Building Operations*, Richardson, TX, October 2002.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

He Liu, Stefan Saroiu, Alec Wolman, and Himanshu Raj. Software abstractions for trusted sensors. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, MobiSys '12, pages 365–378, New York, NY, USA, 2012. ACM.

J. R. Lloyd, P. Orbanz, Z. Ghahramani, and D. M. Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems (NIPS)*, volume 25, pages 1007–1015, 2012.

László Lovász. Large networks and graph limits. *Amer. Math. Soc. Colloq. Publ.*, 60, 2012.

Audra McMillan and Adam Smith. When is non-trivial estimation possible for graphons and stochastic block models?? *Information and Inference: A Journal of the IMA*, page iax010, 2017.

Frank D. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, pages 19–30, 2009.

I. Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275, Aug 2017. doi: 10.1109/CSF.2017.11.

Ilya Mironov, Omkant Pandey, Omer Reingold, and Salil Vadhan. Computational differential privacy. In *Proceedings of the 29th Annual International Cryptology Conference on Advances in Cryptology*, CRYPTO '09, pages 126–142, Berlin, Heidelberg, 2009. Springer-Verlag.

Bojan Mohar. The laplacian spectrum of graphs. In *Graph Theory, Combinatorics, and Applications*, pages 871–898. Wiley, 1991.

Andrés Molina-Markham, Prashant Shenoy, Kevin Fu, Emmanuel Cecchet, and David Irwin. Private memoirs of a smart meter. In *Proceedings of the 2Nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, BuildSys '10, pages 61–66, New York, NY, USA, 2010. ACM.

Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probab. Theory Related Fields*, 162(3):431–461, 2014.

Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, STOC '15, pages 69–75, New York, NY, USA, 2015. ACM.

Joe Neeman and Praneeth Netrapalli. Non-reconstructability in the stochastic block model. *arXiv:1404.6304*, 2014.

Gergely Neu. First-order regret bounds for combinatorial semi-bandits. In *COLT*, 2015.

Kobbi Nissim and Uri Stemmer. On the generalization properties of differential privacy. *arXiv preprint arXiv:1504.05800*, 2015.

L. Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inf. Theor.*, 54(10):4750–4755, October 2008.

Tomaso Poggio, Ryan Rifkin, Sayan Mukherjee, and Partha Niyogi. General conditions for predictivity in learning theory. *Nature*, 428(6981):419, 2004.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Relax and randomize: From value to algorithms. In *NIPS*, 2012.

Jason Reed and Benjamin C. Pierce. Distance makes the types grow stronger: A calculus for differential privacy. In *Proceedings of the 15th ACM SIGPLAN International Conference on Functional Programming*, ICFP '10, pages 157–168, 2010.

Aminmohammad Roozgard, Nafise Barzigar, Pramode Verma, and Samuel Cheng. Genomic data privacy protection using compressed sensing. *Trans. Data Privacy*, 9(1): 1–13, April 2016.

Stéphane Ross and J Andrew Bagnell. Stability conditions for online learnability. *arXiv preprint arXiv:1108.3154*, 2011.

Indrajit Roy, Srinath T. V. Setty, Ann Kilzer, Vitaly Shmatikov, and Emmett Witchel. Airavat: Security and privacy for mapreduce. In *Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation*, NSDI'10, pages 20–20, 2010.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

A. D. Sarwate and K. Chaudhuri. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *IEEE Signal Processing Magazine*, 30(5):86–94, Sept 2013.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Found. Trends Mach. Learn.*, 4(2):107–194, February 2012.

Krishna Mohan Pd Shrivastva, M. A. Rizvi, and Shailendra Singh. Big data privacy based on differential privacy a hope for big data. In *Proceedings of the 2014 International Conference on Computational Intelligence and Communication Networks*, CICN '14, pages 776–781, Washington, DC, USA, 2014. IEEE Computer Society.

Karthik Sridharan. A gentle introduction to concentration inequalities. 02 2018.

Vasilis Syrgkanis, Haipeng Luo, Akshay Krishnamurthy, and Robert E Schapire. Improved regret bounds for oracle-based adversarial contextual bandits. In *NIPS*, 2016.

M. Tang, D. L. Sussman, and C. E. Priebe. *Ann. Statist.*, 41(3):1406–1430, 06 2013.

Abhradeep Thakurta, Andrew Vyrros, Umesh Vaishampayan, Gaurav Kapoor, Julien Freidiger, Vivek Sridhar, and Doug Davidson. Learning New Words. U.S. Patent 9,594,741 B1, March 14 2017.

Abhradeep Guha Thakurta and Adam Smith. (nearly) optimal algorithms for private online learning in full-information and bandit settings. In *NIPS*, 2013.

D. Thanou, X. Dong, D. Kressner, and P. Frossard. Learning heat diffusion graphs. *IEEE Transactions on Signal and Information Processing over Networks*, 3(3):484–499, Sept 2017.

Aristide Charles Yedia Tossou and Christos Dimitrakakis. Achieving privacy in the adversarial multi-armed bandit. In *AAAI*, 2017.

Joel Tropp. Just relax: Convex programming methods for subset selection and sparse approximation. 04-04, 01 2004.

Michael Carl Tschantz, Dilsun Kaynar, and Anupam Datta. Formal verification of differential privacy for interactive systems (extended abstract). *Electron. Notes Theor. Comput. Sci.*, 276:61–79, September 2011.

Salil Vadhan. The complexity of differential privacy. Harvard University Privacy Tools Project, 2016.

G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 51–60, Oct 2014.

Paul Valiant. Testing symmetric properties of distributions. *SIAM J. Comput.*, 40(6): 1927–1968, December 2011.

Tim van Erven, Wojciech Kotlowski, and Manfred K. Warmuth. Follow the leader with dropout perturbations. In *COLT*, 2014.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

J.P. Walters, Z. Liang, W. Shi, and V. Chaudhary. *Wireless Sensor Network Security: A Survey*, page 367. CRC Press: Boca Raton, FL, USA, 2007.

L. Wang, D. Zhang, D. Yang, B. Y. Lim, and X. Ma. Differential location privacy for sparse mobile crowdsensing. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1257–1262, Dec 2016.

Stanley Warner. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of American Statistical Association*, 60(309):63–69, 1965.

Charles F. Weber. Analysis and solution of the ill-posed inverse heat conduction problem. *International Journal of Heat and Mass Transfer*, 24(11):1783–1792, November 1981.

P.J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. *arXiv:1309.5936*, 2013.

Justin J. Yang, Qiuyi Han, and Edoardo M. Airoldi. Nonparametric estimation and testing of exchangeable graph models. In *Proceedings of 17th International Conference on Artificial Intelligence and Statistics*, pages 1060–1067, 2014.

Wotao Yin, Stanley Osher, Donald Goldfarb, and Jerome Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM J. Img. Sci.*, 1(1):143–168, March 2008.

Bin Yu. Assouad, Fano, and Le Cam. In David Pollard, Erik Torgersen, and GraceL. Yang, editors, *Festschrift for Lucien Le Cam*, pages 423–435. Springer New York, 1997.

Bin Yu. Stability. *Bernoulli*, 19(4):1484–1500, 2013.