Upper and Lower Bounds for Privacy and Adaptivity in Algorithmic Data Analysis

Upper and Lower Bounds for Privacy and Adaptivity in Algorithmic Data Analysis

A dissertation presented

by

Thomas Alexander Steinke

to

The School of Engineering and Applied Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Computer Science

Harvard University

Cambridge, Massachusetts

July 2016

Dissertation Advisor:
Professor Salil P. Vadhan

Author:
Thomas Alexander Steinke

Upper and Lower Bounds for Privacy and Adaptivity in Algorithmic Data Analysis

Abstract

The increasing collection and use of sensitive personal data raises important privacy concerns. Another concern arising from the use of data in empirical sciences is the danger of producing results that are not statistically valid due to failing to account for the influence of previous exploration of the data. This thesis studies formalisations of these issues and the relationship between them.

- We give an alternative definition of differential privacy, which is a formal privacy standard for protecting sensitive data. Our definition strikes a balance between the mathematical elegance of so-called pure differential privacy and the power of approximate differential privacy.

- We prove tighter upper and lower bounds for differential privacy. Namely, we bound the minimum size of a dataset that permits accurately answering simple "one-way marginal" queries subject to differential privacy. Our bounds are tight up to constant or $\log \log$ factors.

- We show fundamental limits for privacy by exhibiting a privacy attack that, given the aggregate statistics of a suitable dataset and the data of an individual, can determine whether or not the individual's data is part of the dataset. Our attack is particularly applicable to genetic data and is robust to errors in

the aggregate statistics. This attack is very similar to our lower bounds for differential privacy and demonstrates that differential privacy achieves the fundamental limit of privacy in this setting.

- We simplify, tighten, and extend the connection between differential privacy and generalisation established by Dwork et al. (STOC 2015). In particular, when data is analysed adaptively – that is, multiple analyses are performed and each may depend on the outcome of previous analyses – differentially private algorithms produce statistically valid results in the setting where the data is a sample from some larger population. Our results establish the tight connection between differential privacy and generalisation.

- We prove lower bounds in the adaptive data analysis setting that nearly match the upper bounds given by differential privacy. Namely, we show that, given n samples from an unknown distribution, we cannot answer more than $O(n^2)$ adaptive statistical queries about that distribution while guaranteeing statistical accuracy.

- We show that adaptivity also poses a problem in differential privacy. We show that, for certain classes of queries, it is much harder to answer queries in a differentially private manner if the queries are posed adaptively than if the queries are provided all at once.

All of our results rely on understanding the information-theoretic relationship between the input and output of a randomised algorithm. The criterion for protecting privacy or ensuring generalisation is that changing a single input point of the data analysis algorithm does not affect the output "too much" in a probabilistic sense.

Contents

List of Tables

List of Figures

Acknowledgments

My six years in graduate school have been a transformative experience both academically and personally. I could not imagine who I am now back when I decided to move a third of the way around the world to start my PhD in a new country. I am greatly indebted to the many people who helped me along the way.

Firstly, I would like to thank my wife Joy who has stuck with me since high school, even when we were living on different continents, and eventually left her life in New Zealand to join me in Boston. I want to thank my parents Günter and Marina and my siblings Martin and Nicole who have always given me a loving and supportive home that I have missed dearly.

My office – Maxwell Dworkin 138 – has always been a lively place and I thank my fellow office dwellers Jon, Justin, Varun, Colin, Zhengyu, Jiapeng, Scott, Mark, Tom, Anna, Zhenming, Jiayang, Victor, Jack, Jarek, Vasileios, Elad, Yi-hsiu, and Chin for their company. I also thank Rezza, Natalie, Chris, Anna, Bo, Daniel, Alice, and Richard for ensuring that I socialise outside of the office. I thank Carol Harlow for making administrative and organisational tasks painless for me.

My collaborators Jon Ullman, Adam Smith, Cynthia Dwork, Salil Vadhan, Uri Stemmer, Mark Bun, Kobbi Nissim, Omer Reingold, Andrew Wan, Varun Kanade, Justin Thaler, Michael Mitzenmacher, Raef Bassily, and Sitan Chen are all brilliant. This thesis would not exist without them and I thank them for everything they have taught me. In particular, I credit Jon with getting me working on the subject of this thesis. I also thank my dissertation committee: Salil Vadhan, Jelani Nelson, Boaz Barak, Madhu Sudan, and Leslie Valiant.

Of course, none of this would have been possible without my amazing advisor Salil. I continue to be amazed by his brilliance and energy, but also his generosity,

humility, and patience. I could not ask for better mentor and will forever think "what would Salil ask?" when I am stuck on a problem.

To Theodore

Chapter 1

Introduction

It is becoming increasingly easy to collect, store, and process large amounts of data, which can be extremely useful for science, government, and industry. However, using this data also presents challenges. In particular, much of this data is sensitive and cannot be made public without imperiling the privacy of the individuals it pertains to. Furthermore, a major concern in empirical sciences is the possibility of overfitting data and reaching conclusions that do not generalise to the larger population from which the data was drawn. This problem is exacerbated by adaptivity – that is, when an analysis of a dataset incorporates information from prior examination of the same data (such as through model selection). These two related issues are the focus of this thesis.

We study formalisations of these issues in the form of *differential privacy* and the *statistical query model* respectively. In both cases, the question boils down to understanding the relationship between the input and output of a randomised algorithm. Namely, we must understand how much "information" about the input is revealed by the output. Any useful algorithm must reveal some information about its input, but to preserve privacy and prevent overfitting the revealed information

must be controlled. In particular, we wish to protect "local" information about individuals or that is specific to the sample, while permitting the release of "global" properties of the input data or source population.

1.1 Differential Privacy

As more and more sensitive data is being collected about individuals, a big challenge is releasing useful information about this data in a way that does not compromise the privacy of the individuals concerned. Many ad hoc techniques have been used to "de-identify" data on a record-by-record basis before public release. However, these methods have suffered numerous failures; researchers have "re-identified" information about individuals [Swe97, BZ06, NS08, CKN+11] by linking the sensitive data to public records or other data that overlaps with the released information.

These failures led to the development of *differential privacy* [DMNS06, DKM+06]. Differential privacy applies ideas developed in the cryptography community to the problem of privacy-preserving data analysis and provides a *rigorous* and *quantitative* theory in which to study the problem. In the decade since its inception, differential privacy has developed a rich literature and this thesis makes several contributions thereto.

1.1.1 Background

We first discuss how privacy is formalised by differential privacy. When releasing information about a sensitive dataset, our ideal privacy criterion is that *anything that can be learnt about an individual from the released information, can be learnt without that individual's data being included*. This does not ensure that *nothing* about an individual can be learnt from the released information. For example, revealing that

smoking and lung cancer are strongly correlated reveals sensitive information about an individual if that individual is known to smoke; however, this correlation can be learnt without the use of the data of that individual, so we do not consider it to be a privacy violation.

If we require that *no* information about any individual is revealed, we cannot release any useful information. Thus, to permit the release of useful information, the privacy criterion must be quantitatively relaxed to allow the revelation of small amount of information about each individual. Differential privacy is a formal quantification of this relaxed criterion.

Differential privacy is a property of a randomised procedure or mechanism $\mathcal{M}$ that takes a sensitive dataset x as input and releases the output $\mathcal{M}(x)$.[1] We compare the output distribution $\mathcal{M}(x)$ with a hypothetical output distribution $\mathcal{M}(x')$ in which the input x is changed to x' by removing, adding, or modifying the data of a single individual. The requirement of differential privacy is that $\mathcal{M}(x)$ should be indistinguishable from $\mathcal{M}(x')$ for any inputs x and x' differing only on the data of a single individual:

Definition 1.1.1 (Differential Privacy [DMNS06, DKM+06])**.** *A mechanism $\mathcal{M}$ satisfies* (ε, δ)-differential privacy *if, for all datasets x and x' differing only on the data of a single individual and every potential set of outcomes S,*

$$\mathbb{P}\left[\mathcal{M}(x) \in S\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{M}(x') \in S\right] + \delta. \tag{1.1}$$

The special case where $\delta = 0$ is called pure *differential privacy, in which case we refer to ε-differential privacy, instead of $(\varepsilon, 0)$-differential privacy. The case where $\delta > 0$ is called* approximate *differential privacy.*

[1] We emphasise that differential privacy is a property of the *mechanism* releasing the information, not simply of the output.

Setting the parameters $\varepsilon = \delta = 0$ corresponds to revealing no information ($\mathcal{M}(x)$ and $\mathcal{M}(x')$ are identically distributed), whereas setting $\varepsilon > 0$ or $\delta > 0$ permits revealing some information about individuals.[2] The definition of differential privacy (1.1) is inherently probabilistic; as in cryptographic definitions, randomness is necessary to "obscure" or "hide" the individual information. Thus any non-trivial differentially private release of information requires randomisation.

Differential privacy is a very "robust" definition, as we would expect of a meaningful privacy guarantee. Namely it satisfies the following properties, which, for simplicity, we only state for pure differential privacy here.

- *Postprocessing:* If information is released in a differentially private manner, then additional analysis of that released information — including combining it with information from other sources — will not weaken the differential privacy guarantee.

- *Composition:* If one individual's data is used in multiple independent releases, then the combination of these releases satisfies differential privacy, as long as each release satisfies differential privacy on its own. However, the quantitative privacy guarantee degrades: if each release satisfies ε-differential privacy, then combining k such releases satisfies $k\varepsilon$-differential privacy.

- *Group privacy:* If information is shared by several individuals (such as relatives), differential privacy still protects this information. As for composition, the privacy guarantee degrades with the number of individuals we wish to protect simultaneously. That is, if x differs from x' on the data of at most k individuals

[2]The parameter ε (sometimes called the *privacy loss bound*) is usually thought of as a small constant no larger than 1, whereas δ is usually much smaller (no larger than 10^{-6}, as δ bounds the probability of a potential catastrophic failure).

and $\mathcal{M}$ satisfies ε-differential privacy, then $\mathbb{P}\left[\mathcal{M}(x) \in S\right] \leq e^{k\varepsilon} \cdot \mathbb{P}\left[\mathcal{M}(x') \in S\right]$ for all possible sets of outcomes S.

Arguably, composition is the signature property of differential privacy. Composition permits viewing differentially private analyses as part of a larger system. Indeed, analysis of private data does not occur in a vacuum — a single individual's data will be used multiple times over their lifetime. As such, it is vital to understand the risk to privacy posed by the accumulated information released by independent analyses. Furthermore, simple mechanisms can be composed to perform more complex analytical tasks.

Perhaps the most surprising feature of differential privacy is that, despite its strength as a privacy protection, it is amenable to meaningful data analysis. An extensive literature has been developed showing that a wide range of useful analyses can be carried out subject to differential privacy and its variants. (See the textbook on the subject [DR14].)

1.1.2 Contributions

Concentrated Differential Privacy (§2, [BS16])

The original definition of differential privacy (Definition 1.1.1) has withstood a decade of scrutiny; subsequent research has (generally) validated it as the appropriate formalisation of privacy protection. However, there remains room to tweak the definition slightly to obtain better results. In particular, the aforementioned composition property becomes more complex in the setting of approximate differential privacy.

A key result in differential privacy is the so-called advanced composition theorem [DRV10]: Suppose k differentially private analyses $\mathcal{M}_1, \cdots, \mathcal{M}_k$ are applied to the

same dataset, where each $\mathcal{M}_i$ is $(\varepsilon_i, \delta_i)$-differentially private. Then the combination of these analyses satisfies $(\varepsilon, \delta' + \sum_i \delta)$-differential privacy for any $\delta' > 0$ and

$$\varepsilon = \frac{1}{2}\sum_i \varepsilon_i^2 + \sqrt{2\log(1/\delta')\sum_i \varepsilon_i^2}.$$

This is a very powerful result, which is used widely in the literature to obtain near-optimal results. However, it is not a tight result and, in practice, obtaining a sharper understanding of how differential privacy composes is critical. Unfortunately, the tightest composition bound is unwieldy and is #P-hard to compute exactly [KOV15, MV16]. Approximate differential privacy can be mathematically inelegant. Pure differential privacy, on the other hand, has very elegant composition properties, but is a less practical and more restrictive definition.

Our work combines the mathematical elegance of pure differential privacy with the power of approximate differential privacy. For this, we consider an alternative definition of differential privacy, namely *concentrated differential privacy*, which was first defined by Dwork and Rothblum [DR16]. We present an alternative formulation of the concept of concentrated differential privacy in terms of the Rényi divergence between the distributions obtained by running an algorithm on neighboring inputs.

Using our reformulated definition of concentrated differential privacy, we prove sharper quantitative results, establish lower bounds, and raise a few new questions. In particular, we are able to show (with some caveats) that any mechanism satisfying concentrated differential privacy can be converted into a mechanism satisfying pure differential privacy with only a quadratic blowup in sample complexity. Our lower bounds demonstrate sharp limits on the power of concentrated differential privacy. Specifically, concentrated differential privacy is susceptible to so-called packing lower bounds [HT10], which are often "too pessimistic." To circumvent these lower bounds and also unify concentrated differential privacy with approximate

differential privacy, we give an appropriate definition of "approximate concentrated differential privacy."

Tighter Upper and Lower Bounds for Differential Privacy (§4, [SU15a])

A natural way to measure the tradeoff between privacy and utility is *sample complexity*—the minimum number of records n that is sufficient in order to publicly release an accurate approximation to a given set of statistics about the dataset, while satisfying differential privacy. Intuitively, it's easier to achieve these two goals when n is large, as each individual's data will have only a small influence on the aggregate statistics of interest. Conversely, the sample complexity n should increase as ε and δ decrease (which strengthens the privacy guarantee).

The sample complexity of achieving pure differential privacy is well-understood for many settings (e.g. [HT10]). The more general case of approximate differential privacy is less well understood. However, for the interesting special case of one-way marginals we are able to provide bounds that are tight up to constant (or, in some cases, $\log \log$) factors for almost all choices of parameters.

Specifically, we consider algorithms that compute an extremely simple and fundamental family of queries, namely the *one-way marginals* of the dataset. For a dataset $x \in \{\pm 1\}^{n \times d}$, the d one-way marginals are simply the mean of the bits in each of the d columns: $\overline{x} := \frac{1}{n} \sum_{i=1}^{n} x_i \in [\pm 1]^d$, where $x_i \in \{\pm 1\}^d$ is the i-th element or row of x. A mechanism M is accurate if, on input x, its output is close to $\overline{x}$. Accuracy may be measured in a *worst-case* sense—i.e. $\|M(x) - \overline{x}\|_\infty \leq \alpha$, meaning every one-way marginal is answered with accuracy α—or in an *average-case* sense—i.e. $\|M(x) - \overline{x}\|_1 \leq \alpha d$, meaning the marginals are answered with average accuracy α.

Some of the earliest results in differential privacy [DN03, DN04, BDMN05,

DMNS06] give a simple (ε, δ)-differentially private algorithm—the *Laplace mechanism*—that computes the one-way marginals of $x \in \{\pm 1\}^{n \times d}$ with average error α as long as

$$n \geq O\left(\min\left\{\frac{\sqrt{d\log(1/\delta)}}{\varepsilon\alpha}, \frac{d}{\varepsilon\alpha}\right\}\right). \tag{1.2}$$

The previous best lower bounds are $n \geq \Omega(d/\varepsilon\alpha)$ [HT10] for pure differential privacy and $n \geq \tilde{\Omega}(\sqrt{d}/\varepsilon\alpha)$ for approximate differential privacy with $\delta = o(1/n)$ [BUV14]. We prove an optimal lower bound that combines the previous lower bounds:

Theorem 1.1.2 (Main Theorem)**.** *For every $\varepsilon, \delta, \alpha \in (0, 0.1)$ and $n, d \in \mathbb{N}$ the following holds. Let $M : \{\pm 1\}^{n \times d} \to [\pm 1]^d$ be (ε, δ)-differentially private. Suppose $\mathop{\mathbb{E}}_{M}[\|M(x) - \overline{x}\|_1] \leq \alpha d$ for all $x \in \{\pm 1\}^{n \times d}$. If $e^{-\alpha\varepsilon n/5} \leq \delta \leq \varepsilon/(250n)^{1.1}$, then*

$$n \geq \Omega\left(\frac{\sqrt{d\log(1/\delta)}}{\varepsilon\alpha}\right).$$

Although there has been a long line of work developing methods to prove lower bounds in differential privacy (see [DN03, DMT07, DY08, KRSU10, HT10, NTZ13, BUV14] for a representative, but not exhaustive, sample), our result is the first to show that the sample complexity must grow by a multiplicative factor of $\sqrt{\log(1/\delta)}$.

The proof of our lower bound draws techniques from the literature on *fingerprinting codes* [BS98, Tar08] (which are a recurring theme in this thesis). The key to the proof is showing that the output of any accurate mechanism must have high "correlation" with its input for a suitably-chosen random input. On the other hand differential privacy implies that the correlation between the input and output must be small. Balancing these conflicting constraints yields the lower bound.

Our lower bound holds for mechanisms that bound the average error over the queries (we denote this as L_1 error). Thus, it also holds for algorithms that

bound the maximum error over the queries (we denote this as L_∞ error). The Laplace mechanism gives a matching upper bound for average error. In many cases bounds on the *maximum* error are preferable. For maximum error, the sample complexity of the best previous mechanisms degrades by an additional polylog(d) factor compared to (1.2): $n \geq O\left(\min\left\{\frac{\sqrt{d\log(1/\delta)\log d}}{\varepsilon\alpha}, \frac{d\log d}{\varepsilon\alpha}\right\}\right)$

Surprisingly, this degradation is not necessary. We present algorithms that answer every one-way marginal with α accuracy and improve on the sample complexity of the Laplace mechanism by a factor of $(\log d)^{\Omega(1)}$. Namely we obtain sample complexity

$$n = O\left(\min\left\{\frac{\sqrt{d\cdot\log(1/\delta)\cdot\log\log d}}{\varepsilon\alpha}, \frac{d}{\alpha\varepsilon}\right\}\right)$$

These algorithms demonstrate that the widely used technique of adding independent noise to each query is suboptimal when the goal is to achieve worst-case error guarantees.

Privacy Attacks (§5, [DSS+15])

An important aspect of motivating differential privacy research is demonstrating that the definition is not "too restrictive." That is, we must argue that we cannot achieve better utility using different techniques whilst providing "adequate" privacy protections. One concrete way to demonstrate this is to show that lower bounds for differential privacy are in fact lower bounds for any reasonable notion of privacy. In other words, we would like to demonstrate that lower bounds for differential privacy in fact correspond to practical *privacy attacks*.

A privacy attack takes seemingly innocuous released information and uses it to reveal the private details of individuals. Thus constructing a privacy attack demonstrates that releasing such information compromises privacy; a privacy attack

rules out any "reasonable" notion of privacy (not just differential privacy).

We show that the aforementioned differential privacy lower bounds for one-way marginals do correspond to a simple privacy attack. Namely we construct a simple *tracing attack*: Given a collection of (approximate) summary statistics about a dataset, the precise data of a single target individual, and a small amount of auxiliary information, the attack can determine whether or not the target is a member of the dataset.

Such a tracing attack is a concern in many natural situations where membership in the dataset is considered sensitive. For example, in a genome-wide association study, the dataset contains genomic information about a case group of individuals with a specific medical diagnosis and the released summary statistics are SNP allele frequencies (i.e. one-way marginals). In this scenario, tracing would compromise privacy by revealing the medical diagnosis of the target. Homer et al. [HSR+08] surprised the genomics research community by demonstrating that tracing is possible if exact summary statistics are released.

We show that one can perform a tracing attack even when the one-way marginals are considerably distorted before being released. The parameter regime where our attack becomes feasible is close to the setting where is becomes possible to use differential privacy to provably foil such an attack. Thus our privacy attack shows that differential privacy is not too restrictive in the one-way marginals setting – releasing more one-way marginals than is possible under differential privacy necessarily compromises privacy by permitting our tracing attack.

The analysis of our tracing attack can be viewed as extending the analysis of fingerprinting codes, which are used in lower bounds for differential privacy in Chapter 4. In particular, we show that fingerprinting codes "arise naturally" and do not need to be specially constructed. Cryptographic constructions of fingerprinting

codes may choose a specific data distribution, whereas our attack is intended to be applicable to real-world data where the data distribution may even be unknown. This necessitates a different and more general analysis.

The analysis of our attack requires assuming that the data is drawn from some *unknown* product distribution and an independent sample from that distribution is provided as auxiliary information. Furthermore, we assume that the mechanism releasing the distorted one-way marginals faces significant uncertainty about this unknown distribution. (Otherwise, knowledge of the distribution from which the data is drawn can be used to foil a tracing attack.) This uncertainty about the distribution is modelled by a "meta distribution." Our analysis works for a large, natural class of meta distributions, rather than a single distribution. These assumptions are considerably weaker than previous analyses of fingerprinting codes.

1.2 Adaptivity and Generalisation

Data is used to perform statistical analyses or train a machine learning algorithms. However, the object of interest when doing so is not the data itself, rather the goal of statistical inference or learning is to understand the *population* from which that data was collected. That is, we usually assume that there is some unknown probability distribution representing the "ground truth" and the data consists of independent samples from that distribution and the objective of data analysis is to draw conclusions about the unknown probability distribution from the known data. *Overfitting* occurs when data analysis produces a conclusion that accurately represents the data, but fails to reflect the true population, in which case we say that the conclusion does not *generalise*.

A major concern in empirical sciences and machine learning is preventing overfitting [GL14, Ioa05]. However, most techniques for ensuring generalisation (e.g. [Bon36, BH95] are predicated on the assumption that the data analysis occurs all at once. In particular, it is difficult to handle *adaptive data analysis* in which there are multiple rounds of data analysis where each round is informed by the conclusions of previous rounds. For example, if the same dataset is used to select a model and then fit that model, then standard hypothesis techniques do not accurately reflect how well that model fits the population. Likewise, the choice of hyperparameters (such as regularization weights) can cause overfitting if the same dataset is used at multiple stages in the model fitting process.

A recent line of work [DFH+15c, HU14, et sequalae] has provided a model in which to study the problem of ensuring generalisation in adaptive data analysis and revealed a deep connection to differential privacy. This thesis contributes to both sides of this connection; we provide tighter results showing how differential privacy prevents overfitting and strengthen impossibility results for adaptive data analysis building on techniques used for proving differential privacy lower bounds.

1.2.1 Background

Following Dwork et al. [DFH+15c] and Hardt and Ullman [HU14], we study adaptive data analysis in the statistical query model: There is a *population* $\mathcal{P}$ which is probability distribution on some data universe $\mathcal{X}$ and an *analyst* $\mathcal{A}$ who wishes to study the population. The analyst passes queries to a mechanism $\mathcal{M}$, which must provide answers to those queries. The analyst specifies queries one-by-one and each query may depend on the answers returned for previous queries. The mechanism is given n independent samples x from the population, but otherwise knows nothing about the population. The mechanism must ensure that all the answers it returns are

accurate with high probability. For this to be possible, how large does the sample size n need to be as a function of the type and number of queries and the desired accuracy guarantee?

For simplicity, consider the case where the analyst asks k statistical queries. That is, the analyst specifies predicates $q_1, \cdots, q_k : \mathcal{X} \to [0,1]$ and the answers $a_1, \cdots, a_k$ are considered accurate if $|a_j - \mathop{\mathbb{E}}_{z \sim \mathcal{P}} [q_j(z)]| \leq \alpha$ for all j. In the non-adaptive setting (where the queries are specified all at once and cannot depend on previous answers) we can simply use empirical answers — that is, $a_j = \frac{1}{n} \sum_{i=1}^{n} q_j(x_i)$ — and the sample complexity is $n = O(\log(k)/\alpha^2)$ by a Chernoff-Hoeffding bound combined with the union bound. However, this approach does not work if the queries may be adaptive: Using only $k = O(n)$ queries, the analyst can discern the sample x.[3] Once the analyst knows the sample, it can formulate a query q that evaluates to 1 on the sample points and 0 elsewhere. This demonstrates that the problem is nontrivial.

A simple and well-known approach is *sample splitting*: The sample x is split into k subsamples and each query is evaluated on a "fresh" subsample. Unfortunately, this requires high sample complexity, namely $n = \tilde{\Theta}(k/\alpha^2)$.

1.2.2 Contributions

Improved Upper Bounds (§3, [BNS+16a])

Dwork et al. [DFH+15c] gave the first improved upper bounds for adaptive data analysis. They showed that, if the mechanism $\mathcal{M}$ satisfies differential privacy and is empirically accurate (that is, the answers are accurate with respect to the sample x, but not necessarily with respect to the population $\mathcal{P}$), then the mechanism is

[3] Uniformly random queries suffice to approximately discern the sample. If answers are returned with arbitrary precision and $\mathcal{X} \subseteq \mathbb{N}$, then the entire sample can be revealed with a single query: $q(x) = n^{-x}$.

accurate with respect to the population. In other words, they showed that differential privacy ensures generalisation. Applying this connection to known differentially private algorithms (namely, the Laplace or Gaussian mechanism), they obtain sample complexity $n = \tilde{\Theta}(\sqrt{k}/\alpha^{2.5})$ for k statistical queries with accuracy α.

We give a simpler proof of the reduction of Dwork et al., which also achieves better parameters and applies to more general families of queries. In particular, for *k low-sensitivity queries* (which are a generalisation of statistical queries), we obtain sample complexity $n = \tilde{\Theta}(\sqrt{k}/\alpha^2)$. We also show that the definition of differential privacy can be relaxed while still providing meaningful generalisation guarantees.

Improved Lower Bounds (§6, [SU15b])

Hardt and Ullman [HU14] gave the first lower bounds for adaptive data analysis. They showed that, if the analyst and population are both chosen adversarially, then the mechanism needs $n = \tilde{\Omega}(\sqrt[3]{k}/\alpha)$ samples to answer k statistical queries to accuracy α. This holds under one of two assumptions: Either the the dimension $d = \log|\mathcal{X}|$ of the data satisfies $d \geq k$ or the mechanism is computationally bounded so that it cannot break cryptographic encryption with d-bit keys.

We extend the proof of Hardt and Ullman to obtain the lower bound $n = \Omega(\sqrt{k}/\alpha)$, which is tight up to constants in its dependence on k. This requires the construction of an object we call *interactive fingerprinting codes*, which are a generalisation of fingerprinting codes. Our construction of interactive fingerprinting codes uses many of the same techniques we have used in proving lower bounds for differential privacy (§4) and for analysing privacy attacks (§5). However, our construction of interactive fingerprinting codes extends the fingerprinting analysis in a different direction: we provide optimal "robustness" – that is, we are able to perform the fingerprinting attack even if, say, 99% of the answers are arbitrarily

corrupted, whereas previously it was only known how to withstand less than 2% corrupted answers [BUV14].

The Power of Adaptivity in Differential Privacy (§7, [BSU16])

Our results and those of Dwork et al. [DFH+15c] show that differentially private mechanisms which can answer adaptive queries are useful for generalisation. Not all mechanisms in the differential privacy literature are suited to adaptive queries (e.g. [BLR13]). However, most mechanisms can handle adaptive queries. Moreover, almost all lower bounds for differential privacy use a set of queries that is provided up front. Hence, in the cases where the adaptive upper bounds and non-adaptive lower bounds match, we see that there is no difference between adaptive and non adaptive queries. Thus we ask whether this is always the case — is the sample complexity needed to accurately answer adaptive queries subject to differential privacy always the same as when the queries are provided all at once?

We answer this question negatively by showing an exponential separation between the adaptive and non-adaptive cases. In fact we separate three cases – the *adaptive* setting where queries are provided one at a time and may depend on previous answers, the *online* setting where the queries are provided one at a time but may not depend on previous answers, and the *offline* setting where the queries are provided all at once. We construct a class of statistical queries for which the sample complexity is maximal in the online setting, but, in the offline setting, the sample complexity is exponentially smaller. We also construct a class of "search queries" that separate the online and adaptive settings.

Chapter 2

Concentrated Differential Privacy

2.1 Introduction

Differential privacy [DMNS06] is a formal mathematical standard for protecting individual-level privacy in statistical data analysis. In its simplest form, (pure) differential privacy is parameterised by a real number $\varepsilon > 0$, which controls how much "privacy loss"[1] an individual can suffer when a computation (i.e., a statistical data analysis task) is performed involving his or her data.

One particular hallmark of differential privacy is that it degrades smoothly and predictably under the *composition* of multiple computations. In particular, if one performs k computational tasks that are each ε-differentially private and combines the results of those tasks, then the computation as a whole is $k\varepsilon$-differentially private. This property makes differential privacy amenable to the type of modular reasoning used in the design and analysis of algorithms: When a sophisticated algorithm is

[1]The privacy loss is a random variable that quantifies how much information is revealed about an individual by a computation involving their data; it depends on the outcome of the computation, the way the computation was performed, and the information that the individual wants to hide. We discuss it informally in this introduction and define it precisely in Definition 2.1.2 on page 19.

comprised of a sequence of differentially private steps, one can establish that the algorithm as a whole remains differentially private.

A widely-used relaxation of pure differential privacy is *approximate* or (ε, δ)-differential privacy [DKM+06], which essentially guarantees that the probability that any individual suffers privacy loss exceeding ε is bounded by δ. For sufficiently small δ, approximate (ε, δ)-differential privacy provides a comparable standard of privacy protection as pure ε-differential privacy, while often permitting substantially more useful analyses to be performed.

Unfortunately, there are situations where, unlike pure differential privacy, approximate differential privacy is not a very elegant abstraction for mathematical analysis, particularly the analysis of composition. The "advanced composition theorem" of Dwork, Rothblum, and Vadhan [DRV10] (subsequently improved by [KOV15, MV16]) shows that the composition of k tasks which are each (ε, δ)-differentially private is $(\approx\sqrt{k}\varepsilon, \approx k\delta)$-differentially private. However, these bounds can be unwieldy; computing the tightest possible privacy guarantee for the composition of k arbitrary mechanisms with differing $(\varepsilon_i, \delta_i)$-differential privacy guarantees is #P-hard [MV16]! Furthermore, these bounds are not tight even for simple and natural privacy-preserving computations. For instance, consider the mechanism which approximately answers k statistical queries on a given database by adding independent Gaussian noise to each answer. Even for this basic computation, the advanced composition theorem does not yield a tight analysis.[2]

Dwork and Rothblum [DR16] recently put forth a different relaxation of differen-

[2]In particular, consider answering k statistical queries on a dataset of n individuals by adding noise drawn from $\mathcal{N}(0, (\sigma/n)^2)$ independently for each query. Each individual query satisfies $(O(\sqrt{\log(1/\delta)}/\sigma), \delta)$-differential privacy for any $\delta > 0$. Applying the advanced composition theorem shows that the composition of all k queries satisfies $(O(\sqrt{k}\log(1/\delta)/\sigma), (k+1)\delta)$-differential privacy for any $\delta > 0$. However, it is well-known that this bound can be improved to $(O(\sqrt{k\log(1/\delta)}/\sigma), \delta)$-differential privacy.

tial privacy called *concentrated differential privacy*. Roughly, a randomised mechanism satisfies concentrated differentially privacy if the privacy loss has small mean and is subgaussian. Concentrated differential privacy behaves in a qualitatively similar way as approximate (ε, δ)-differential privacy under composition. However, it permits sharper analyses of basic computational tasks, including a tight analysis of the aforementioned Gaussian mechanism.

Using the work of Dwork and Rothblum [DR16] as a starting point, we introduce an alternative formulation of the concept of concentrated differential privacy that we call "zero-concentrated differential privacy" (zCDP for short). To distinguish our definition from that of Dwork and Rothblum, we refer to their definition as "mean-concentrated differential privacy" (mCDP for short). Our definition uses the Rényi divergence between probability distributions as a different method of capturing the requirement that the privacy loss random variable is subgaussian.

2.1.1 Our Reformulation: Zero-Concentrated Differential Privacy

As is typical in the literature, we model a dataset as a multiset or tuple of n elements (or "rows") in $\mathcal{X}^n$, for some "data universe" $\mathcal{X}$, where each element represents one individual's information. A (privacy-preserving) computation is a randomised algorithm $M : \mathcal{X}^n \to \mathcal{Y}$, where $\mathcal{Y}$ represents the space of all possible outcomes of the computation.

Definition 2.1.1 (Zero-Concentrated Differential Privacy (zCDP))**.** *A randomised mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ is (ξ, ρ)-zero-concentrated differentially private (henceforth (ξ, ρ)-zCDP) if, for all $x, x' \in \mathcal{X}^n$ differing on a single entry and all $\alpha \in (1, \infty)$,*

$$\mathrm{D}_\alpha \left(M(x) \middle\| M(x') \right) \leq \xi + \rho\alpha, \tag{2.1}$$

where $\mathrm{D}_\alpha\left(M(x)\|M(x')\right)$ *is the Rényi divergence*[3] *of order* α *between the distribution of* $M(x)$ *and the distribution of* $M(x')$.

We define ρ*-zCDP to be* $(0,\rho)$*-zCDP.*[4]

Equivalently, we can replace (2.1) with

$$\mathbb{E}\left[e^{(\alpha-1)Z}\right] \leq e^{(\alpha-1)(\xi+\rho\alpha)}, \tag{2.2}$$

where $Z = \mathsf{PrivLoss}\left(M(x)\|M(x')\right)$ is the privacy loss random variable:

Definition 2.1.2 (Privacy Loss Random Variable)**.** *Let* Y *and* Y' *be random variables on* Ω*. We define the* privacy loss random variable between Y and Y' *– denoted* $Z = \mathsf{PrivLoss}\left(Y\|Y'\right)$ *– as follows. Define a function* $f : \Omega \to \mathbb{R}$ *by* $f(y) = \log(\mathbb{P}\left[Y = y\right]/\mathbb{P}\left[Y' = y\right])$.[5] *Then* Z *is distributed according to* $f(Y)$.

Intuitively, the value of the privacy loss $Z = \mathsf{PrivLoss}\left(M(x)\|M(x')\right)$ represents how well we can distinguish x from x' given only the output $M(x)$ or $M(x')$. If $Z > 0$, then the observed output of M is more likely to have occurred if the input was x than if x' was the input. Moreover, the larger Z is, the bigger this likelihood ratio is. Likewise, $Z < 0$ indicates that the output is more likely if x' is the input. If $Z = 0$, both x and x' "explain" the output of M equally well.

[3]Rényi divergence has a parameter $\alpha \in (1,\infty)$ which allows it to interpolate between KL-divergence ($\alpha \to 1$) and max-divergence ($\alpha \to \infty$). It should be thought of as a measure of dissimilarity between distributions. We define it formally in Section 2.2. Throughout, we assume that all logarithms are natural unless specified otherwise — that is, base $e \approx 2.718$. This includes logarithms in information theoretic quantities like entropy, divergence, and mutual information, whence these quantities are measured in *nats* rather than in *bits*.

[4]For clarity of exposition, we consider only ρ-zCDP in the introduction and give more general statements for (ξ,ρ)-zCDP later. We also believe that having a one-parameter definition is desirable.

[5]Throughout we abuse notation by letting $\mathbb{P}\left[Y = y\right]$ represent either the probability mass function or the probability density function of Y evaluated at y. Formally, $\mathbb{P}\left[Y = y\right]/\mathbb{P}\left[Y' = y\right]$ denotes the Radon-Nikodym derivative of the measure Y with respect to the measure Y' evaluated at y, where we require Y to be absolutely continuous with respect to Y', i.e. $Y \ll Y'$.

A mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ is ε-differentially private if and only if $\mathbb{P}[Z > \varepsilon] = 0$, where $Z = \mathsf{PrivLoss}(M(x)\|M(x'))$ is the privacy loss of M on arbitrary inputs $x, x' \in \mathcal{X}^n$ differing in one entry. On the other hand, M being (ε, δ)-differentially private is equivalent, up to a small loss in parameters, to the requirement that $\mathbb{P}[Z > \varepsilon] \leq \delta$.

In contrast, zCDP entails a bound on the *moment generating function* of the privacy loss Z — that is, $\mathbb{E}\left[e^{(\alpha-1)Z}\right]$ as a function of $\alpha - 1$. The bound (2.2) implies that Z is a *subgaussian* random variable[6] with small mean. Intuitively, this means that Z resembles a Gaussian distribution with mean $\xi + \rho$ and variance 2ρ. In particular, we obtain strong tail bounds on Z. Namely (2.2) implies that

$$\mathbb{P}[Z > \lambda + \xi + \rho] \leq e^{-\lambda^2/4\rho}$$

for all $\lambda > 0$.[7]

Thus zCDP requires that the privacy loss random variable is concentrated around zero (hence the name). That is, Z is "small" with high probability, with larger deviations from zero becoming increasingly unlikely. Hence we are unlikely to be able to distinguish x from x' given the output of $M(x)$ or $M(x')$. Note that the randomness of the privacy loss random variable is taken only over the randomness of the mechanism M.

[6] A random variable X being subgaussian is characterised by the following four equivalent conditions [Riv12]. (i) $\mathbb{P}\left[|X - \mathbb{E}[X]| > \lambda\right] \leq e^{-\Omega(\lambda^2)}$ for all $\lambda > 0$. (ii) $\mathbb{E}\left[e^{t(X-\mathbb{E}[X])}\right] \leq e^{O(t^2)}$ for all $t \in \mathbb{R}$. (iii) $\mathbb{E}\left[(X - \mathbb{E}[X])^{2k}\right] \leq O(k)^k$ for all $k \in \mathbb{N}$. (iv) $\mathbb{E}\left[e^{c(X-\mathbb{E}[X])^2}\right] \leq 2$ for some $c > 0$.

[7] We only discuss bounds on the upper tail of Z. We can obtain similar bounds on the lower tail of $Z = \mathsf{PrivLoss}(M(x)\|M(x'))$ by considering $Z' = \mathsf{PrivLoss}(M(x')\|M(x))$.

Comparison to the Definition of Dwork and Rothblum

For comparison, Dwork and Rothblum [DR16] define (μ, τ)-*concentrated differential privacy* for a randomised mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ as the requirement that, if $Z = \mathsf{PrivLoss}\left(M(x) \| M(x')\right)$ is the privacy loss for $x, x' \in \mathcal{X}^n$ differing on one entry, then

$$\mathbb{E}\left[Z\right] \leq \mu \qquad \text{and} \qquad \mathbb{E}\left[e^{(\alpha-1)(Z-\mathbb{E}[Z])}\right] \leq e^{(\alpha-1)^2 \frac{1}{2}\tau^2}$$

for all $\alpha \in \mathbb{R}$. That is, they require both a bound on the mean of the privacy loss and that the privacy loss is tightly concentrated around its mean. To distinguish our definitions, we refer to their definition as *mean-concentrated differential privacy* (or mCDP).

Our definition, zCDP, is a *relaxation* of mCDP. In particular, a (μ, τ)-mCDP mechanism is also $(\mu - \tau^2/2, \tau^2/2)$-zCDP (which is tight for the Gaussian mechanism example), whereas the converse is not true. (However, a partial converse holds; see Lemma 2.4.3.)

2.1.2 Results

Relationship between zCDP and Differential Privacy

Like Dwork and Rothblum's formulation of concentrated differential privacy, zCDP can be thought of as providing guarantees of (ε, δ)-differential privacy *for all* values of $\delta > 0$:

Proposition 2.1.3. *If M provides ρ-zCDP, then M is $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$-differentially private for any $\delta > 0$.*

We also prove a slight strengthening of this result (Lemma 2.3.7). Moreover, there is a partial converse, which shows that, up to a loss in parameters, zCDP is

equivalent to differential privacy with this $\forall \delta > 0$ quantification (see Lemma 2.3.9).

There is also a direct link from pure differential privacy to zCDP:

Proposition 2.1.4. *If M satisfies ε-differential privacy, then M satisfies $(\frac{1}{2}\varepsilon^2)$-zCDP.*

Dwork and Rothblum [DR16, Theorem 3.5] give a slightly weaker version of Proposition 2.1.4, which implies that ε-differential privacy yields $(\frac{1}{2}\varepsilon(e^\varepsilon - 1))$-zCDP; this improves on an earlier bound [DRV10] by the factor $\frac{1}{2}$.

We give proofs of these and other properties using properties of Rényi divergence in Sections 2.2 and 2.3.

Propositions 2.1.3 and 2.1.4 show that zCDP is an intermediate notion between pure differential privacy and approximate differential privacy. Indeed, many algorithms satisfying approximate differential privacy do in fact also satisfy zCDP.

Gaussian Mechanism

Just as with mCDP, the prototypical example of a mechanism satisfying zCDP is the *Gaussian mechanism*, which answers a real-valued query on a database by perturbing the true answer with Gaussian noise.

Definition 2.1.5 (Sensitivity)**.** *A function $q : \mathcal{X}^n \to \mathbb{R}$ has* sensitivity Δ *if for all $x, x' \in \mathcal{X}^n$ differing in a single entry, we have $|q(x) - q(x')| \leq \Delta$.*

Proposition 2.1.6 (Gaussian Mechanism)**.** *Let $q : \mathcal{X}^n \to \mathbb{R}$ be a sensitivity-Δ query. Consider the mechanism $M : \mathcal{X}^n \to \mathbb{R}$ that on input x, releases a sample from $\mathcal{N}(q(x), \sigma^2)$. Then M satisfies $(\Delta^2/2\sigma^2)$-zCDP.*

We remark that either inequality defining zCDP — (2.1) or (2.2) — is exactly tight for the Gaussian mechanism for all values of α. Thus the definition of zCDP seems tailored to the Gaussian mechanism.

Basic Properties of zCDP

Our definition of zCDP satisfies the key basic properties of differential privacy. Foremost, these properties include smooth degradation under composition, and invariance under postprocessing:

Lemma 2.1.7 (Composition). *Let $M : \mathcal{X}^n \to \mathcal{Y}$ and $M' : \mathcal{X}^n \to \mathcal{Z}$ be randomised algorithms. Suppose M satisfies ρ-zCDP and M' satisfies ρ'-zCDP. Define $M'' : \mathcal{X}^n \to \mathcal{Y} \times \mathcal{Z}$ by $M''(x) = (M(x), M'(x))$. Then M'' satisfies $(\rho + \rho')$-zCDP.*

Lemma 2.1.8 (Postprocessing). *Let $M : \mathcal{X}^n \to \mathcal{Y}$ and $f : \mathcal{Y} \to \mathcal{Z}$ be randomised algorithms. Suppose M satisfies ρ-zCDP. Define $M' : \mathcal{X}^n \to \mathcal{Z}$ by $M'(x) = f(M(x))$. Then M' satisfies ρ-zCDP.*

These properties follow immediately from corresponding properties of the Rényi divergence outlined in Lemma 2.2.2.

We remark that Dwork and Rothblum's definition of mCDP is not closed under postprocessing; we provide a counterexample in Section 2.4.1. (However, an arbitrary amount of postprocessing can worsen the guarantees of mCDP by at most constant factors.)

Group Privacy

A mechanism M guarantees *group privacy* if no small group of individuals has a significant effect on the outcome of a computation (whereas the definition of zCDP only refers to individuals, which are groups of size 1). That is, group privacy for groups of size k guarantees that, if x and x' are inputs differing on k entries (rather than a single entry), then the outputs $M(x)$ and $M(x')$ are close.

Dwork and Rothblum [DR16, Theorem 4.1] gave nearly tight bounds on the group privacy guarantees of concentrated differential privacy, showing that a $(\mu = \tau^2/2, \tau)$-

concentrated differentially private mechanism affords $(k^2\mu \cdot (1 + o(1)), k\tau \cdot (1 + o(1)))$-concentrated differential privacy for groups of size $k = o(1/\tau)$. We are able to show a group privacy guarantee for zCDP that is exactly tight and works for a wider range of parameters:

Proposition 2.1.9. *Let $M : \mathcal{X}^n \to \mathcal{Y}$ satisfy ρ-zCDP. Then M guarantees $(k^2\rho)$-zCDP for groups of size k — i.e. for every $x, x' \in \mathcal{X}^n$ differing in up to k entries and every $\alpha \in (1, \infty)$, we have*

$$\mathrm{D}_\alpha \left(M(x) \middle\| M(x') \right) \leq (k^2\rho) \cdot \alpha.$$

In particular, this bound is achieved (simultaneously for all values α) by the Gaussian mechanism. Our proof is also simpler than that of Dwork and Rothblum; see Section 2.5.

Lower Bounds

The strong group privacy guarantees of zCDP yield, as an unfortunate consequence, strong lower bounds as well. We show that, as with pure differential privacy, zCDP is susceptible to information-based lower bounds, as well as to so-called packing arguments [HT10, MMP+10, De12]:

Theorem 2.1.10. *Let $M : \mathcal{X}^n \to \mathcal{Y}$ satisfy ρ-zCDP. Let X be a random variable on $\mathcal{X}^n$. Then*

$$I\left(X; M(X)\right) \leq \rho \cdot n^2,$$

where $I(\cdot;\cdot)$ denotes the mutual information between the random variables (in nats, rather than bits). Furthermore, if the entries of X are independent, then $I(X; M(X)) \leq \rho \cdot n$.

Theorem 2.1.10 yields strong lower bounds for zCDP mechanisms, as we can construct distributions X such that, for any accurate mechanism M, $M(X)$ reveals a lot of information about X (i.e. $I(X; M(X))$ is large for any accurate M).

In particular, we obtain a strong separation between approximate differential privacy and zCDP. For example, we can show that releasing an accurate approximate histogram (or, equivalently, accurately answering all point queries) on a data domain of size k requires an input with at least $n = \Theta(\sqrt{\log k})$ entries to satisfy zCDP. In contrast, under approximate differential privacy, n can be *independent* of the domain size k [BNS13]! In particular, our lower bounds show that "stability-based" techniques (such as those in the propose-test-release framework [DL09]) are not compatible with zCDP.

Our lower bound exploits the strong group privacy guarantee afforded by zCDP. Group privacy has been used to prove tight lower bounds for pure differential privacy [HT10, De12] and approximate differential privacy [§4]. These results highlight the fact that group privacy is often the limiting factor for private data analysis. For (ε, δ)-differential privacy, group privacy becomes vacuous for groups of size $k = \Theta(\log(1/\delta)/\varepsilon)$. Indeed, stability-based techniques exploit precisely this breakdown in group privacy.

As a result of this strong lower bound, we show that any mechanism for answering statistical queries that satisfies zCDP can be converted into a mechanism satisfying pure differential privacy with only a quadratic blowup in its sample complexity. More precisely, the following theorem illustrates a more general result we prove in Section 2.7.

Theorem 2.1.11. *Let $n \in \mathbb{N}$ and $\alpha \geq 1/n$ be arbitrary. Set $\varepsilon = \alpha$ and $\rho = \alpha^2$. Let $q : \mathcal{X} \to [0,1]^k$ be an arbitrary family of statistical queries. Suppose $M : \mathcal{X}^n \to [0,1]^k$ satisfies ρ-zCDP and*

$$\mathop{\mathbb{E}}_{M}[\|M(x) - q(x)\|_\infty] \leq \alpha$$

for all $x \in \mathcal{X}^n$. Then there exists $M' : \mathcal{X}^{n'} \to [0,1]^k$ for $n' = 5n^2$ satisfying ε-differential

privacy and

$$\mathop{\mathbb{E}}_{M'}\left[\|M'(x) - q(x)\|_\infty\right] \leq 10\alpha$$

for all $x \in \mathcal{X}^{n'}$.

For some classes of queries, this reduction is essentially tight. For example, for k one-way marginals, the Gaussian mechanism achieves sample complexity $n = \Theta(\sqrt{k})$ subject to zCDP, whereas the Laplace mechanism achieves sample complexity $n = \Theta(k)$ subject to pure differential privacy, which is known to be optimal.

For more details, see Sections 2.6 and 2.7.

Approximate zCDP

To circumvent these strong lower bounds for zCDP, we consider a relaxation of zCDP in the spirit of approximate differential privacy that permits a small probability δ of (catastrophic) failure:

Definition 2.1.12 (Approximate Zero-Concentrated Differential Privacy (Approximate zCDP))**.** *A randomised mechanism* $M : \mathcal{X}^n \to \mathcal{Y}$ *is* δ*-approximately* (ξ, ρ)*-zCDP if, for all* $x, x' \in \mathcal{X}^n$ *differing on a single entry, there exist events* E *(depending on* $M(x)$*) and* E' *(depending on* $M(x')$*) such that* $\mathbb{P}[E] \geq 1 - \delta$, $\mathbb{P}[E'] \geq 1 - \delta$, *and for all* $\alpha \in (1, \infty)$,

$$\mathrm{D}_\alpha\left(M(x)|_E \middle\| M(x')|_{E'}\right) \leq \xi + \rho \cdot \alpha \quad \wedge \quad \mathrm{D}_\alpha\left(M(x')|_{E'} \middle\| M(x)|_E\right) \leq \xi + \rho \cdot \alpha,$$

where $M(x)|_E$ *denotes the distribution of* $M(x)$ *conditioned on the event* E*. We further define* δ*-approximate* ρ*-zCDP to be* δ*-approximate* $(0, \rho)$*-zCDP.*

In particular, setting $\delta = 0$ gives the original definition of zCDP. However, this definition unifies zCDP with approximate differential privacy:

Proposition 2.1.13. *If M satisfies (ε, δ)-differential privacy, then M satisfies δ-approximate $\frac{1}{2}\varepsilon^2$-zCDP.*

Approximate zCDP retains most of the desirable properties of zCDP, but allows us to incorporate stability-based techniques and bypass the above lower bounds. This also presents a unified tool to analyse a composition of zCDP with approximate differential privacy; see Section 2.8.

2.1.3 Related Work

Our work builds on the aforementioned prior work of Dwork and Rothblum [DR16].[8] We view our definition of concentrated differential privacy as being "morally equivalent" to their definition of concentrated differential privacy, in the sense that both definitions formalise the same concept.[9] (The formal relationship between the two definitions is discussed in Section 2.4.) However, the definition of zCDP generally seems to be easier to work with than that of mCDP. In particular, our formulation in terms of Rényi divergence simplifies many analyses.

Dwork and Rothblum prove several results about concentrated differential privacy that are similar to ours. Namely, they prove analogous properties of mCDP as we prove for zCDP (cf. Sections 2.1.2, 2.1.2, 2.1.2, and 2.1.2). However, as noted, some of their bounds are weaker than ours; also, they do not explore lower bounds. Note that our lower bounds apply equally to their definition of mCDP. The main technical contribution of our work is to prove stark lower bounds for the power of

[8] Although Dwork and Rothblum's work only appeared publicly in March 2016, they shared a preliminary draft of their paper with us before we commenced this work. As such, our ideas are heavily inspired by theirs.

[9] We refer to our definition as "zero-concentrated differential privacy" (zCDP) and their definition as "mean-concentrated differential privacy" (mCDP). We use "concentrated differential privacy" (CDP) to refer to the underlying *concept* formalised by both definitions.

CDP.

Several of the ideas underlying concentrated differential privacy are implicit in earlier works. In particular, the proof of the advanced composition theorem of Dwork, Rothblum, and Vadhan [DRV10] essentially uses the ideas of concentrated differential privacy. Their proof contains analogs of Propositions 2.1.7, 2.1.3, and 2.1.4.

We also remark that Tardos [Tar08] used Rényi divergence to prove lower bounds for cryptographic objects called *fingerprinting codes*. Fingerprinting codes turn out to be closely related to differential privacy [Ull13, BUV14, §4], and Tardos' lower bound can be (loosely) viewed as a kind of privacy-preserving algorithm.

2.1.4 Further Work

We believe that concentrated differential privacy is a useful tool for analysing private computations, as it provides both simpler and tighter bounds. We hope that CDP will be prove useful in both the theory and practice of differential privacy.

Furthermore, our lower bounds show that CDP can really be a much more stringent condition than approximate differential privacy. Thus CDP defines a "subclass" of all (ε, δ)-differentially private algorithms. This subclass includes most differentially private algorithms in the literature, but not all — the most notable exceptions being algorithms that use the propose-test-release approach [DL09] to exploit low local sensitivity.

This "CDP subclass" warrants further exploration. In particular, is there a "complete" mechanism for this class of algorithms, in the same sense that the exponential mechanism [MT07, BLR13] is complete for pure differential privacy? Namely, any purely differentially private algorithm can be viewed as an instantiation of the exponential mechanism. We would like to obtain a similarly general paradigm for

constructing CDP algorithms. Alternatively, can we obtain a simple (combinatorial or geometric) characterisation of the sample complexity needed to satisfy CDP? The ability to prove stronger and simpler lower bounds for CDP than for approximate DP may be useful for showing the limitations of certain algorithmic paradigms. For example, any differentially private algorithm that only uses the Laplace mechanism, the exponential mechanism, the Gaussian mechanism, and the "sparse vector" technique, along with composition and postprocessing will be subject to the lower bounds for CDP.

There is also room to examine how to interpret the zCDP privacy guarantee. In particular, we leave it as an open question to understand the extent to which ρ-zCDP provides a stronger privacy guarantee than the implied (ε, δ)-DP guarantees (cf. Proposition 2.1.3).

In general, much of the literature on differential privacy can be re-examined through the lens of CDP, which may yield new insights and results.

2.2 Rényi Divergence

Recall the definition of Rényi divergence:

Definition 2.2.1 (Rényi Divergence [Rén61, Equation (3.3)])**.** *Let P and Q be probability distributions on Ω. For $\alpha \in (1, \infty)$, we define the* Rényi divergence of order α between P and Q *as*

$$\begin{aligned}
\mathrm{D}_\alpha(P\|Q) &= \frac{1}{\alpha - 1} \log \left(\int_\Omega P(x)^\alpha Q(x)^{1-\alpha} \mathrm{d}x \right) \\
&= \frac{1}{\alpha - 1} \log \left(\mathop{\mathbb{E}}_{x \sim Q} \left[\left(\frac{P(x)}{Q(x)} \right)^\alpha \right] \right) \\
&= \frac{1}{\alpha - 1} \log \left(\mathop{\mathbb{E}}_{x \sim P} \left[\left(\frac{P(x)}{Q(x)} \right)^{\alpha - 1} \right] \right),
\end{aligned}$$

where $P(\cdot)$ and $Q(\cdot)$ are the probability mass/density functions of P and Q respectively or, more generally, $P(\cdot)/Q(\cdot)$ is the Radon-Nikodym derivative of P with respect to Q.[10] *We also define the KL-divergence*

$$\mathrm{D}_1\left(P\|Q\right) = \lim_{\alpha\to 1}\mathrm{D}_\alpha\left(P\|Q\right) = \int_\Omega P(x)\log\left(\frac{P(x)}{Q(x)}\right)\mathrm{d}x$$

and the max-divergence

$$\mathrm{D}_\infty\left(P\|Q\right) = \lim_{\alpha\to\infty}\mathrm{D}_\alpha\left(P\|Q\right) = \sup_{x\in\Omega}\log\left(\frac{P(x)}{Q(x)}\right).$$

Alternatively, Rényi divergence can be defined in terms of the privacy loss (Definition 2.1.2) between P and Q:

$$e^{(\alpha-1)\mathrm{D}_\alpha(P\|Q)} = \mathop{\mathbb{E}}_{Z\sim\mathsf{PrivLoss}(P\|Q)}\left[e^{(\alpha-1)Z}\right]$$

for all $\alpha\in(1,\infty)$. Moreover, $\mathrm{D}_1\left(P\|Q\right) = \mathop{\mathbb{E}}_{Z\sim\mathsf{PrivLoss}(P\|Q)}[Z]$.

We record several useful and well-known properties of Rényi divergence. We refer the reader to [vEH14] for further discussion of these (and many other) properties.

Lemma 2.2.2. *Let P and Q be probability distributions and $\alpha\in[1,\infty]$.*

- Non-negativity: $\mathrm{D}_\alpha\left(P\|Q\right)\geq 0$ *with equality if and only if $P=Q$.*
- Composition: *Suppose P and Q are distributions on $\Omega\times\Theta$. Let P' and Q' denote the marginal distributions on Ω induced by P and Q respectively. For $x\in\Omega$, let P'_x and Q'_x denote the conditional distributions on Θ induced by P and Q respectively, where x specifies the first coordinate. Then*

$$\mathrm{D}_\alpha\left(P'\middle\|Q'\right) + \min_{x\in\Omega}\mathrm{D}_\alpha\left(P'_x\middle\|Q'_x\right) \leq \mathrm{D}_\alpha\left(P\|Q\right) \leq \mathrm{D}_\alpha\left(P'\middle\|Q'\right) + \max_{x\in\Omega}\mathrm{D}_\alpha\left(P'_x\middle\|Q'_x\right).$$

[10] If P is not absolutely continuous with respect to Q (i.e. it is not the case that $P\ll Q$), we define $\mathrm{D}_\alpha\left(P\|Q\right)=\infty$ for all $\alpha\in[1,\infty]$.

In particular if P and Q are product distributions, then the Rényi divergence between P and Q is just the sum of the Rényi divergences of the marginals.

- Quasi-Convexity: *Let P_0, P_1 and Q_0, Q_1 be distributions on Ω, and let $P = tP_0 + (1-t)P_1$ and $Q = tQ_0 + (1-t)Q_1$ for $t \in [0,1]$. Then $\mathrm{D}_\alpha(P\|Q) \leq \max\{\mathrm{D}_\alpha(P_0\|Q_0), \mathrm{D}_\alpha(P_1\|Q_1)\}$. Moreover, KL divergence is convex:*

$$\mathrm{D}_1(P\|Q) \leq t\mathrm{D}_1(P_0\|Q_0) + (1-t)\mathrm{D}_1(P_1\|Q_1).$$

- Postprocessing: *Let P and Q be distributions on Ω and let $f : \Omega \to \Theta$ be a function. Let $f(P)$ and $f(Q)$ denote the distributions on Θ induced by applying f to P or Q respectively. Then $\mathrm{D}_\alpha(f(P)\|f(Q)) \leq \mathrm{D}_\alpha(P\|Q)$.*

  Note that quasi-convexity allows us to extend this guarantee to the case where f is a randomised mapping.

- Monotonicity: *For $1 \leq \alpha \leq \alpha' \leq \infty$, $\mathrm{D}_\alpha(P\|Q) \leq \mathrm{D}_{\alpha'}(P\|Q)$.*

Proof of Non-Negativity. Let $h(t) = t^\alpha$. Then $h''(t) = \alpha(\alpha-1)t^{\alpha-2} > 0$ for all $t > 0$ and $\alpha > 1$. Thus h is strictly convex. Hence $e^{(\alpha-1)\mathrm{D}_\alpha(P\|Q))} = \mathop{\mathbb{E}}_{x\sim Q}[h(P(x)/Q(x))] \geq h(\mathop{\mathbb{E}}_{x\sim Q}[P(x)/Q(x)]) = h(1) = 1$, as required. □

Proof of Composition.

$$\begin{aligned}
e^{(\alpha-1)\mathrm{D}_\alpha(P\|Q)} &= \int_{\Omega\times\Theta} P(x,y)^\alpha Q(x,y)^{1-\alpha}\mathrm{d}(x,y) \\
&= \int_\Omega P'(x)^\alpha Q'(x)^{1-\alpha} \int_\Theta P'_x(y)^\alpha Q'_x(y)^{1-\alpha}\mathrm{d}y\mathrm{d}x \\
&= \int_\Omega P'(x)^\alpha Q'(x)^{1-\alpha} e^{(\alpha-1)\mathrm{D}_\alpha(P'_x\|Q'_x)}\mathrm{d}x \\
&\leq \int_\Omega P'(x)^\alpha Q'(x)^{1-\alpha}\mathrm{d}x \cdot \max_x e^{(\alpha-1)\mathrm{D}_\alpha(P'_x\|Q'_x)} \\
&= e^{(\alpha-1)\mathrm{D}_\alpha(P'\|Q')} \cdot e^{(\alpha-1)\max_x \mathrm{D}_\alpha(P'_x\|Q'_x)}.
\end{aligned}$$

The other side of the inequality is symmetric. □

Proof of Quasi-Convexity. Unfortunately Rényi divergence is not convex for $\alpha > 1$. (Although KL-divergence is.) However, the following property implies that Rényi divergence is quasi-convex.

Lemma 2.2.3. *Let P_0, P_1, Q_0, Q_1 be distributions on Ω. For $t \in [0,1]$, define $P_t = tP_1 + (1-t)P_0$ and $Q_t = tQ_1 + (1-t)Q_0$ to be the convex combinations specified by t. Then*

$$e^{(\alpha-1)\mathrm{D}_\alpha(P_t\|Q_t)} \leq te^{(\alpha-1)\mathrm{D}_\alpha(P_1\|Q_1)} + (1-t)e^{(\alpha-1)\mathrm{D}_\alpha(P_0\|Q_0)}.$$

Moreover, the limit as $\alpha \to 1+$ gives

$$\mathrm{D}_1\left(P_t\|Q_t\right) \leq t\mathrm{D}_1\left(P_1\|Q_1\right) + (1-t)\mathrm{D}_1\left(P_0\|Q_0\right).$$

Proof of Lemma 2.2.3. Let $f(t) = e^{(\alpha-1)\mathrm{D}_\alpha(P_t\|Q_t)}$. Since the equality is clearly true for

$t = 0$ and $t = 1$, it suffices to show that $f''(t) \geq 0$ for all $t \in [0,1]$. We have

$$
\begin{aligned}
f(t) =& \int_\Omega P_t(x)^\alpha Q_t(x)^{1-\alpha} \mathrm{d}x, \\
f'(t) =& \int_\Omega \frac{\mathrm{d}}{\mathrm{d}t} P_t(x)^\alpha Q_t(x)^{1-\alpha} \mathrm{d}x \\
=& \int_\Omega \alpha P_t(x)^{\alpha-1} \left(\frac{\mathrm{d}}{\mathrm{d}t} P_t(x) \right) Q_t(x)^{1-\alpha} + (1-\alpha) P_t(x)^\alpha Q_t(x)^{-\alpha} \left(\frac{\mathrm{d}}{\mathrm{d}t} Q_t(x) \right) \mathrm{d}x \\
=& \int_\Omega \alpha P_t(x)^{\alpha-1} (P_1(x) - P_0(x)) Q_t(x)^{1-\alpha} + (1-\alpha) P_t(x)^\alpha Q_t(x)^{-\alpha} (Q_1(x) - Q_0(x)) \mathrm{d}x, \\
f''(t) =& \int_\Omega \alpha(\alpha-1) P_t(x)^{\alpha-2} \left(P_1(x) - P_0(x)\right)^2 Q_t(x)^{1-\alpha} \\
& + \alpha(1-\alpha) P_t(x)^{\alpha-1} \left(P_1(x) - P_0(x)\right) Q_t(x)^{-\alpha} \left(Q_1(x) - Q_0(x)\right) \\
& + (1-\alpha)\alpha P_t(x)^{\alpha-1} \left(P_1(x) - P_0(x)\right) Q_t(x)^{-\alpha} \left(Q_1(x) - Q_0(x)\right) \\
& + (1-\alpha)(-\alpha) P_t(x)^\alpha Q_t(x)^{-\alpha-1} \left(Q_1(x) - Q_0(x)\right)^2 \mathrm{d}x \\
=& \alpha(\alpha-1) \int_\Omega \begin{pmatrix} \sqrt{P_t(x)^{\alpha-2} Q_t(x)^{1-\alpha}} \left(P_1(x) - P_0(x)\right) \\ -\sqrt{P_t(x)^\alpha Q_t(x)^{-\alpha-1}} \left(Q_1(x) - Q_0(x)\right) \end{pmatrix}^2 \mathrm{d}x \\
\geq& 0.
\end{aligned}
$$

□

□

Proof of Postprocessing. Let $h(x) = x^\alpha$. Note that h is convex. Let $f^{-1}(y) = \{x \in \Omega : f(x) = y\}$. Let Q_y be the conditional distribution on $x \sim Q$ conditioned on

$f(x) = y$. By Jensen's inequality,

$$\begin{aligned}
e^{(\alpha-1)\mathrm{D}_\alpha(P\|Q)} &= \mathop{\mathbb{E}}_{x\sim Q}\left[\left(\frac{P(x)}{Q(x)}\right)^\alpha\right] \\
&= \mathop{\mathbb{E}}_{y\sim f(Q)}\left[\mathop{\mathbb{E}}_{x\sim Q_y}\left[h\left(\frac{P(x)}{Q(x)}\right)\right]\right] \\
&\geq \mathop{\mathbb{E}}_{y\sim f(Q)}\left[h\left(\mathop{\mathbb{E}}_{x\sim Q_y}\left[\frac{P(x)}{Q(x)}\right]\right)\right] \\
&= \mathop{\mathbb{E}}_{y\sim f(Q)}\left[h\left(\int_{f^{-1}(y)}\frac{Q(x)}{Q(f^{-1}(y))}\frac{P(x)}{Q(x)}\mathrm{d}x\right)\right] \\
&= \mathop{\mathbb{E}}_{y\sim f(Q)}\left[h\left(\frac{P(f^{-1}(y))}{Q(f^{-1}(y))}\right)\right] \\
&= e^{(\alpha-1)\mathrm{D}_\alpha(f(P)\|f(Q))}.
\end{aligned}$$

□

Proof of Monotonicity. Let $1 < \alpha \leq \alpha' < \infty$. Let $h(x) = x^{\frac{\alpha'-1}{\alpha-1}}$. Then

$$h''(x) = \frac{\alpha'-1}{\alpha-1}\left(\frac{\alpha'-1}{\alpha-1}-1\right)x^{\frac{\alpha'-1}{\alpha-1}-2} \geq 0,$$

so h is convex on $(0,\infty)$. Thus

$$\begin{aligned}
e^{(\alpha'-1)\mathrm{D}_\alpha(P\|Q)} &= h\left(e^{(\alpha-1)\mathrm{D}_\alpha(P\|Q)}\right) = h\left(\mathop{\mathbb{E}}_{x\sim P}\left[\left(\frac{P(x)}{Q(x)}\right)^{\alpha-1}\right]\right) \\
&\leq \mathop{\mathbb{E}}_{x\sim P}\left[h\left(\left(\frac{P(x)}{Q(x)}\right)^{\alpha-1}\right)\right] = e^{(\alpha'-1)\mathrm{D}_{\alpha'}(P\|Q)},
\end{aligned}$$

which gives the result. □

2.2.1 Composition and Postprocessing

The following lemma gives the postprocessing and (adaptive) composition bounds (extending Lemmas 2.1.7 and 2.1.8).

Lemma 2.2.4 (Composition & Postprocessing)**.** *Let $M : \mathcal{X}^n \to \mathcal{Y}$ and $M' : \mathcal{X}^n \times \mathcal{Y} \to$*

$\mathcal{Z}$. *Suppose M satisfies (ξ, ρ)-zCDP and M' satisfies (ξ', ρ')-zCDP (as a function of its first argument). Define $M'' : \mathcal{X}^n \to \mathcal{Z}$ by $M''(x) = M'(x, M(x))$. Then M'' satisfies $(\xi + \xi', \rho + \rho')$-zCDP.*

The proof is immediate from Lemma 2.2.2. Note that, while Lemma 2.2.4 is only stated for the composition of two mechanisms, it can be inductively applied to analyse the composition of arbitrarily many mechanisms.

2.2.2 Gaussian Mechanism

The following lemma gives the Rényi divergence between two Gaussian distributions with the same variance.

Lemma 2.2.5. *Let $\mu, \nu, \sigma \in \mathbb{R}$ and $\alpha \in [1, \infty)$. Then*

$$\mathrm{D}_\alpha \left(\mathcal{N}(\mu, \sigma^2) \middle\| \mathcal{N}(\nu, \sigma^2) \right) = \frac{\alpha(\mu - \nu)^2}{2\sigma^2}$$

Consequently, the Gaussian mechanism, which answers a sensitivity-Δ query by adding noise drawn from $\mathcal{N}(0, \sigma^2)$, satisfies $\left(\frac{\Delta^2}{2\sigma^2}\right)$-zCDP (Proposition 2.1.6).

Proof. We calculate

$$\begin{aligned}
&\exp\left((\alpha - 1)\mathrm{D}_\alpha \left(\mathcal{N}(\mu, \sigma^2) \middle\| \mathcal{N}(\nu, \sigma^2) \right)\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} \exp\left(-\alpha \frac{(x-\mu)^2}{2\sigma^2} - (1-\alpha)\frac{(x-\nu)^2}{2\sigma^2} \right) \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} \exp\left(-\frac{(x-(\alpha\mu+(1-\alpha)\nu))^2 - (\alpha\mu+(1-\alpha)\nu)^2 + \alpha\mu^2 + (1-\alpha)\nu^2}{2\sigma^2} \right) \mathrm{d}x \\
&= \mathop{\mathbb{E}}_{x \sim \mathcal{N}(\alpha\mu+(1-\alpha)\nu, \sigma^2)} \left[\exp\left(-\frac{-(\alpha\mu + (1-\alpha)\nu)^2 + \alpha\mu^2 + (1-\alpha)\nu^2}{2\sigma^2} \right) \right] \\
&= \exp\left(\frac{\alpha(\alpha-1)(\mu-\nu)^2}{2\sigma^2} \right).
\end{aligned}$$

□

For the multivariate Gaussian mechanism, Lemma 2.2.5 generalises to the following.

Lemma 2.2.6. *Let $\mu, \nu \in \mathbb{R}^d$, $\sigma \in \mathbb{R}$, and $\alpha \in [1, \infty)$. Then*

$$\mathrm{D}_\alpha \left(\mathcal{N}(\mu, \sigma^2 I_d) \middle\| \mathcal{N}(\nu, \sigma^2 I_d) \right) = \frac{\alpha \|\mu - \nu\|_2^2}{2\sigma^2}$$

Thus, if $M : \mathcal{X}^n \to \mathbb{R}^d$ is the mechanism that, on input x, releases a sample from $\mathcal{N}(q(x), \sigma^2 I_d)$ for some function $q : \mathcal{X}^n \to \mathbb{R}^d$, then M satisfies ρ-zCDP for

$$\rho = \frac{1}{2\sigma^2} \sup_{\substack{x, x' \in \mathcal{X}^n \\ \text{differing in one entry}}} \|q(x) - q(x')\|_2^2. \tag{2.3}$$

2.3 Relation to Differential Privacy

We now discuss the relationship between zCDP and the traditional definitions of pure and approximate differential privacy. There is a close relationship between the notions, but not an exact characterisation.

2.3.1 Pure DP versus zCDP

Pure differential privacy is exactly characterised by $(\xi, 0)$-zCDP:

Lemma 2.3.1. *A mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies ε-DP if and only if it satisfies $(\varepsilon, 0)$-zCDP.*

Proof. Let $x, x' \in \mathcal{X}^n$ be neighbouring. Suppose M satisfies ε-DP. Then $\mathrm{D}_\infty \left(M(x) \| M(x') \right) \leq \varepsilon$. By monotonicity,

$$\mathrm{D}_\alpha \left(M(x) \middle\| M(x') \right) \leq \mathrm{D}_\infty \left(M(x) \middle\| M(x') \right) \leq \varepsilon = \varepsilon + 0 \cdot \alpha$$

for all α. So M satisfies $(\varepsilon, 0)$-zCDP. Conversely, suppose M satisfies $(\varepsilon, 0)$-zCDP.

Then

$$\mathrm{D}_\infty\left(M(x)\middle\|M(x')\right) = \lim_{\alpha\to\infty}\mathrm{D}_\alpha\left(M(x)\middle\|M(x')\right) \le \lim_{\alpha\to\infty}\varepsilon + 0\cdot\alpha = \varepsilon.$$

Thus M satisfies ε-DP. □

We now show that ε-differential privacy implies $(\frac{1}{2}\varepsilon^2)$-zCDP (Proposition 2.1.4).

Proposition 2.3.2. *Let P and Q be probability distributions on Ω satisfying $\mathrm{D}_\infty(P\|Q) \le \varepsilon$ and $\mathrm{D}_\infty(Q\|P) \le \varepsilon$. Then $\mathrm{D}_\alpha(P\|Q) \le \frac{1}{2}\varepsilon^2\alpha$ for all $\alpha > 1$.*

Remark 2.3.3. *In particular, Proposition 2.3.2 shows that the KL-divergence $\mathrm{D}_1(P\|Q) \le \frac{1}{2}\varepsilon^2$. A bound on the KL-divergence between random variables in terms of their max-divergence is an important ingredient in the analysis of the advanced composition theorem [DRV10]. Our bound sharpens (up to lower order terms) and, in our opinion, simplifies the previous bound of $\mathrm{D}_1(P\|Q) \le \frac{1}{2}\varepsilon(e^\varepsilon - 1)$ proved by Dwork and Rothblum [DR16].*

In particular, taking $\alpha \to 1$ we have the following corollary.

Corollary 2.3.4. *If $\mathrm{D}_\infty(P\|Q) \le \varepsilon$ and $\mathrm{D}_\infty(Q\|P) \le \varepsilon$, then $\mathrm{D}_1(P\|Q) \le \frac{1}{2}\varepsilon^2$.*

Proof of Proposition 2.3.2. We may assume $\frac{1}{2}\varepsilon\alpha \le 1$, as otherwise $\frac{1}{2}\varepsilon^2\alpha > \varepsilon$, whence the result follows from monotonicity. We must show that

$$e^{(\alpha-1)\mathrm{D}_\alpha(P\|Q)} = \mathop{\mathbb{E}}_{x\sim Q}\left[\left(\frac{P(x)}{Q(x)}\right)^\alpha\right] \le e^{\frac{1}{2}\alpha(\alpha-1)\varepsilon^2}.$$

We know that $e^{-\varepsilon} \le \frac{P(x)}{Q(x)} \le e^\varepsilon$ for all x. Define a random function $A : \Omega \to \{e^{-\varepsilon}, e^\varepsilon\}$ by $\mathop{\mathbb{E}}_A[A(x)] = \frac{P(x)}{Q(x)}$ for all x. By Jensen's inequality,

$$\mathop{\mathbb{E}}_{x\sim Q}\left[\left(\frac{P(x)}{Q(x)}\right)^\alpha\right] = \mathop{\mathbb{E}}_{x\sim Q}\left[\left(\mathop{\mathbb{E}}_A[A(x)]\right)^\alpha\right] \le \mathop{\mathbb{E}}_{x\sim Q}\left[\mathop{\mathbb{E}}_A[A(x)^\alpha]\right] = \mathop{\mathbb{E}}_A[A^\alpha],$$

where A denotes $A(x)$ for a random $x \sim Q$. We also have $\mathop{\mathbb{E}}_A[A] = \mathop{\mathbb{E}}_{x\sim Q}\left[\frac{P(x)}{Q(x)}\right] = 1$.

From this equation, we can conclude that

$$\mathop{\mathbb{P}}_{A}\left[A = e^{-\varepsilon}\right] = \frac{e^{\varepsilon} - 1}{e^{\varepsilon} - e^{-\varepsilon}} \qquad \text{and} \qquad \mathop{\mathbb{P}}_{A}\left[A = e^{\varepsilon}\right] = \frac{1 - e^{-\varepsilon}}{e^{\varepsilon} - e^{-\varepsilon}}.$$

Thus

$$\begin{aligned} e^{(\alpha-1)\mathrm{D}_\alpha(P\|Q)} \leq & \mathop{\mathbb{E}}_{A}\left[A^\alpha\right] \\ = & \frac{e^{\varepsilon} - 1}{e^{\varepsilon} - e^{-\varepsilon}} \cdot e^{-\alpha\varepsilon} + \frac{1 - e^{-\varepsilon}}{e^{\varepsilon} - e^{-\varepsilon}} \cdot e^{\alpha\varepsilon} \\ = & \frac{\left(e^{\alpha\varepsilon} - e^{-\alpha\varepsilon}\right) - \left(e^{(\alpha-1)\varepsilon} - e^{-(\alpha-1)\varepsilon}\right)}{e^{\varepsilon} - e^{-\varepsilon}} \\ = & \frac{\sinh(\alpha\varepsilon) - \sinh((\alpha-1)\varepsilon)}{\sinh(\varepsilon)}. \end{aligned}$$

The result now follows from Lemma 2.3.5. □

Lemma 2.3.5.

$$0 \leq y < x \leq 2 \implies \frac{\sinh(x) - \sinh(y)}{\sinh(x-y)} \leq e^{\frac{1}{2}xy}.$$

This technical lemma may be "verified" numerically by inspecting a plot of $z = e^{\frac{1}{2}xy} \cdot \sinh(x-y) - (\sinh(x) - \sinh(y))$ for $(x, y) \in [0, 2]^2$.

For intuition, consider the third-order Taylor approximation to $\sinh(x)$ about 0:

$$\sinh(x) = x + \frac{1}{6}x^3 \pm O(x^5).$$

Then we can approximate

$$\begin{aligned} \frac{\sinh(x) - \sinh(y)}{\sinh(x-y)} \approx & \frac{x + \frac{1}{6}x^3 - (y + \frac{1}{6}y^3)}{(x-y) + \frac{1}{6}(x-y)^3} \\ = & 1 + \frac{3xy(x-y)}{6(x-y) + (x-y)^3} \\ \leq & 1 + \frac{1}{2}xy \\ \leq & e^{\frac{1}{2}xy}. \end{aligned}$$

Unfortunately, turning this intuition into an actual proof is quite involved. We

instead provide a proof from [hh]:

Proof. We need the following hyperbolic trigonometric identities.

$$\begin{aligned}
\cosh(w+z) &= \cosh(w)\cosh(z) + \sinh(w)\sinh(z) \\
&= \cosh(w)\cosh(z)\left(1+\tanh(w)\tanh(z)\right), \\
\sinh(x-y) &= \frac{1}{2}\left(e^{x-y}+1-1-e^{-(x-y)}\right) \\
&= \frac{1}{2}\left(e^{(x-y)/2}-e^{-(x-y)/2}\right)\left(e^{(x-y)/2}+e^{-(x-y)/2}\right) \\
&= 2\sinh\left(\frac{x-y}{2}\right)\cosh\left(\frac{x-y}{2}\right) \\
&= 2\sinh\left(\frac{x-y}{2}\right)\cosh\left(\frac{x}{2}\right)\cosh\left(\frac{-y}{2}\right)\left(1+\tanh\left(\frac{x}{2}\right)\tanh\left(\frac{-y}{2}\right)\right) \\
&= 2\sinh\left(\frac{x-y}{2}\right)\cosh\left(\frac{x}{2}\right)\cosh\left(\frac{y}{2}\right)\left(1-\tanh\left(\frac{x}{2}\right)\tanh\left(\frac{y}{2}\right)\right), \\
\sinh(x)-\sinh(y) &= \frac{1}{2}\left(e^{x}-e^{-x}-e^{y}+e^{-y}\right) \\
&= \frac{1}{2}\left(e^{(x-y)/2}-e^{-(x-y)/2}\right)\left(e^{(x+y)/2}+e^{-(x+y)/2}\right) \\
&= 2\sinh\left(\frac{x-y}{2}\right)\cosh\left(\frac{x+y}{2}\right) \\
&= 2\sinh\left(\frac{x-y}{2}\right)\cosh\left(\frac{x}{2}\right)\cosh\left(\frac{y}{2}\right)\left(1+\tanh\left(\frac{x}{2}\right)\tanh\left(\frac{y}{2}\right)\right).
\end{aligned}$$

We also use the fact that $0 \le \tanh(z) < \min\{z, 1\}$ for all $z > 0$. Thus

$$\frac{\sinh(x)-\sinh(y)}{\sinh(x-y)} = \frac{1+\tanh\left(\frac{x}{2}\right)\tanh\left(\frac{y}{2}\right)}{1-\tanh\left(\frac{x}{2}\right)\tanh\left(\frac{y}{2}\right)} = \frac{1+t}{1-t},$$

where $t = \tanh\left(\frac{x}{2}\right)\tanh\left(\frac{y}{2}\right) < 1$. Now

$$\frac{1+t}{1-t} \le e^{xy/2} \iff 1+t \le (1-t)e^{xy/2} \iff (e^{xy/2}+1)t \le e^{xy/2}-1$$

$$\iff t \le \frac{e^{xy/2}-1}{e^{xy/2}+1} = \frac{e^{xy/4}-e^{-xy/4}}{e^{xy/4}+e^{-xy/4}} = \tanh\left(\frac{xy}{4}\right).$$

So it only remains to show that $\tanh\left(\frac{x}{2}\right)\tanh\left(\frac{y}{2}\right) \leq \tanh\left(\frac{xy}{4}\right)$. This is clearly true when $y = 0$. Now

$$\frac{\partial}{\partial y}\tanh\left(\frac{x}{2}\right)\tanh\left(\frac{y}{2}\right) = \tanh\left(\frac{x}{2}\right)\cosh^{-2}\left(\frac{y}{2}\right)\frac{1}{2} \leq \cosh^{-2}\left(\frac{y}{2}\right)\frac{x}{4}$$

and

$$\frac{\partial}{\partial y}\tanh\left(\frac{xy}{4}\right) = \cosh^{-2}\left(\frac{xy}{4}\right)\frac{x}{4}.$$

Since $0 \leq y < x \leq 2$, we have $0 \leq xy/4 \leq y/2$ and, hence, $\cosh(y/2) \geq \cosh(xy/4)$. Thus

$$\frac{\partial}{\partial y}\tanh\left(\frac{x}{2}\right)\tanh\left(\frac{y}{2}\right) \leq \frac{\partial}{\partial y}\tanh\left(\frac{xy}{4}\right).$$

The inequality now follows by integration of the inequality on the derivatives. □

2.3.2 Approximate DP versus zCDP

The statements in this section show that, up to some loss in parameters, zCDP is equivalent to a family of (ε, δ)-DP guarantees for all $\delta > 0$.

Lemma 2.3.6. *Let $M : \mathcal{X}^n \to \mathcal{Y}$ satisfy (ξ, ρ)-zCDP. Then M satisfies (ε, δ)-DP for all $\delta > 0$ and*

$$\varepsilon = \xi + \rho + \sqrt{4\rho\log(1/\delta)}.$$

Thus to achieve a given (ε, δ)-DP guarantee it suffices to satisfy (ξ, ρ)-zCDP with

$$\rho = \left(\sqrt{\varepsilon - \xi + \log(1/\delta)} - \sqrt{\log(1/\delta)}\right)^2 \approx \frac{(\varepsilon - \xi)^2}{4\log(1/\delta)}.$$

Proof. Let $x, x' \in \mathcal{X}^n$ be neighbouring. Define

$$f(y) = \log\left(\frac{\mathbb{P}\left[M(x) = y\right]}{\mathbb{P}\left[M(x') = y\right]}\right).$$

Let $Y \sim M(x)$ and $Z = f(Y)$. That is, $Z = \mathsf{PrivLoss}\left(M(x) \| M(x')\right)$ is the privacy

loss random variable. Fix $\alpha \in (1, \infty)$ to be chosen later. Then

$$\mathbb{E}\left[e^{(\alpha-1)Z}\right] = \mathop{\mathbb{E}}_{Y \sim M(x)}\left[\left(\frac{\mathbb{P}\left[M(x) = Y\right]}{\mathbb{P}\left[M(x') = Y\right]}\right)^{\alpha-1}\right] = e^{(\alpha-1)\mathrm{D}_\alpha(M(x)\|M(x'))} \leq e^{(\alpha-1)(\xi+\rho\alpha)}.$$

By Markov's inequality

$$\mathbb{P}\left[Z > \varepsilon\right] = \mathbb{P}\left[e^{(\alpha-1)Z} > e^{(\alpha-1)\varepsilon}\right] \leq \frac{\mathbb{E}\left[e^{(\alpha-1)Z}\right]}{e^{(\alpha-1)\varepsilon}} \leq e^{(\alpha-1)(\xi+\rho\alpha-\varepsilon)}.$$

Choosing $\alpha = (\varepsilon - \xi + \rho)/2\rho > 1$ gives

$$\mathbb{P}\left[Z > \varepsilon\right] \leq e^{-(\varepsilon-\xi-\rho)^2/4\rho} \leq \delta.$$

Now, for any measurable $S \subset \mathcal{Y}$,

$$\begin{aligned}
\mathbb{P}\left[M(x) \in S\right] =& \mathbb{P}\left[Y \in S\right] \\
\leq& \mathbb{P}\left[Y \in S \wedge Z \leq \varepsilon\right] + \mathbb{P}\left[Z > \varepsilon\right] \\
\leq& \mathbb{P}\left[Y \in S \wedge Z \leq \varepsilon\right] + \delta \\
=& \int_{\mathcal{Y}} \mathbb{P}\left[M(x) = y\right] \cdot \mathbb{I}(y \in S) \cdot \mathbb{I}(f(y) \leq \varepsilon)\, \mathrm{d}y + \delta \\
\leq& \int_{\mathcal{Y}} e^{\varepsilon}\mathbb{P}\left[M(x') = y\right] \cdot \mathbb{I}(y \in S)\, \mathrm{d}y + \delta \\
=& e^{\varepsilon}\mathbb{P}\left[M(x') \in S\right] + \delta.
\end{aligned}$$

□

Lemma 2.3.6 is not tight. In particular, we have the following refinement of Lemma 2.3.6.

Lemma 2.3.7. *Let $M : \mathcal{X}^n \to \mathcal{Y}$ satisfy (ξ, ρ)-zCDP. Then M satisfies (ε, δ)-DP for all $\delta > 0$ and*

$$\varepsilon = \xi + \rho + \sqrt{4\rho \cdot \log(\sqrt{\pi \cdot \rho}/\delta)}.$$

Alternatively M satisfies (ε, δ)-DP for all $\varepsilon \geq \xi + \rho$ and

$$\delta = e^{-(\varepsilon-\xi-\rho)^2/4\rho} \cdot \min \begin{cases} \sqrt{\pi \cdot \rho} \\ \frac{1}{1+(\varepsilon-\xi-\rho)/2\rho} \\ \frac{2}{1+\frac{\varepsilon-\xi-\rho}{2\rho}+\sqrt{\left(1+\frac{\varepsilon-\xi-\rho}{2\rho}\right)^2+\frac{4}{\pi\rho}}} \end{cases}.$$

Note that the last of three options in the minimum dominates the first two options. We have included the first two options as they are simpler.

More generally, we have the following result.

Lemma 2.3.8. *Let P and Q be probability distributions on $\mathcal{Y}$ with $\mathrm{D}_\alpha(P\|Q) \leq \xi + \rho \cdot \alpha$ for all $\alpha \in (1, \infty)$. Then, for any $\varepsilon \geq \xi + \rho$ and*

$$\delta = e^{-(\varepsilon-\xi-\rho)^2/4\rho} \cdot \min \begin{cases} 1 \\ \sqrt{\pi \cdot \rho} \\ \frac{1}{1+(\varepsilon-\xi-\rho)/2\rho} \\ \frac{2}{1+\frac{\varepsilon-\xi-\rho}{2\rho}+\sqrt{\left(1+\frac{\varepsilon-\xi-\rho}{2\rho}\right)^2+\frac{4}{\pi\rho}}} \end{cases},$$

we have

$$P(S) \leq e^{\varepsilon} Q(S) + \delta$$

for all (measurable) S.

Proof. Define $f : \mathcal{Y} \to \mathbb{R}$ by $f(y) = \log(P(y)/Q(y))$. Let $Y \sim P$, $Y' \sim Q$ and let $Z = f(Y)$ be the privacy loss random variable. That is, $Z = \mathsf{PrivLoss}(P\|Q)$. For any

measurable $S \subset \mathcal{Y}$,

$$\begin{aligned}
P(S) =& \mathbb{P}\left[Y \in S\right] \\
=& \mathbb{P}\left[Y \in S \wedge f(Y) \leq \varepsilon\right] + \mathbb{P}\left[Y \in S \wedge f(Y) > \varepsilon\right] \\
=& \int_S P(y)\mathbb{I}[f(y) \leq \varepsilon]\mathrm{d}y + \mathbb{P}\left[Y \in S \wedge f(Y) > \varepsilon\right] \\
=& \int_S P(y)\mathbb{I}[P(y) \leq e^{\varepsilon}Q(y)]\mathrm{d}y + \mathbb{P}\left[Y \in S \wedge f(Y) > \varepsilon\right] \\
\leq& \int_S e^{\varepsilon}Q(y)\mathbb{I}[f(y) \leq \varepsilon]\mathrm{d}y + \mathbb{P}\left[Y \in S \wedge f(Y) > \varepsilon\right] \\
=& e^{\varepsilon}\mathbb{P}\left[Y' \in S \wedge f(Y') \leq \varepsilon\right] + \mathbb{P}\left[Y \in S \wedge f(Y) > \varepsilon\right] \\
=& e^{\varepsilon}\mathbb{P}\left[Y' \in S\right] - e^{\varepsilon}\mathbb{P}\left[Y' \in S \wedge f(Y') > \varepsilon\right] + \mathbb{P}\left[Y \in S \wedge f(Y) > \varepsilon\right] \\
\leq& e^{\varepsilon}\mathbb{P}\left[Y' \in S\right] + \left(\mathbb{P}\left[f(Y) > \varepsilon\right] - e^{\varepsilon}\mathbb{P}\left[f(Y') > \varepsilon\right]\right).
\end{aligned}$$

Thus we want to bound

$$\begin{aligned}
\delta =& \mathbb{P}\left[f(Y) > \varepsilon\right] - e^{\varepsilon}\mathbb{P}\left[f(Y') > \varepsilon\right] \\
=& \mathop{\mathbb{E}}_{Y}\left[\mathbb{I}[f(Y) > \varepsilon]\right] - \mathop{\mathbb{E}}_{Y'}\left[e^{\varepsilon}\mathbb{I}[f(Y') > \varepsilon]\right] \\
=& \mathop{\mathbb{E}}_{Y}\left[\mathbb{I}[f(Y) > \varepsilon]\right] - \int_{\mathcal{Y}} e^{\varepsilon}\mathbb{I}[f(y) > \varepsilon]Q(y)\mathrm{d}y \\
=& \mathop{\mathbb{E}}_{Y}\left[\mathbb{I}[f(Y) > \varepsilon]\right] - \int_{\mathcal{Y}} e^{\varepsilon}\mathbb{I}[f(y) > \varepsilon]\frac{Q(y)}{P(y)}P(y)\mathrm{d}y \\
=& \mathop{\mathbb{E}}_{Y}\left[\mathbb{I}[f(Y) > \varepsilon]\right] - \mathop{\mathbb{E}}_{Y}\left[e^{\varepsilon}\mathbb{I}[f(Y) > \varepsilon]e^{-f(Y)}\right] \\
=& \mathop{\mathbb{E}}_{Z}\left[\mathbb{I}[Z > \varepsilon]\left(1 - e^{\varepsilon - Z}\right)\right] \\
=& \mathop{\mathbb{E}}_{Z}\left[\max\left\{0, 1 - e^{\varepsilon - Z}\right\}\right] \\
=& \int_{\varepsilon}^{\infty}\left(1 - e^{\varepsilon - z}\right)\mathbb{P}\left[Z = z\right]\mathrm{d}z.
\end{aligned}$$

In particular,

$$\delta \leq \mathop{\mathbb{E}}_{Z}\left[\mathbb{I}[Z > \varepsilon]\right] = \mathbb{P}\left[Z > \varepsilon\right].$$

Alternatively, by integration by parts,

$$\delta = \int_{\varepsilon}^{\infty} e^{\varepsilon - z} \mathbb{P}\left[Z > z\right] \mathrm{d}z.$$

Now it remains to bound $\mathbb{P}\left[Z > z\right]$.

As in Lemma 2.3.6, by Markov's inequality, for all $\alpha > 1$ and $\lambda > \xi + \rho$,

$$\mathbb{P}\left[Z > \lambda\right] \leq \frac{\mathbb{E}\left[e^{(\alpha-1)Z}\right]}{e^{(\alpha-1)\lambda}} \leq \frac{e^{(\alpha-1)\mathrm{D}_\alpha(M(x)\|M(x'))}}{e^{(\alpha-1)\lambda}} \leq e^{(\alpha-1)(\xi+\rho\alpha-\lambda)}.$$

Choosing $\alpha = (\lambda - \xi + \rho)/2\rho > 1$ gives

$$\mathbb{P}\left[Z > \lambda\right] \leq e^{-(\lambda-\xi-\rho)^2/4\rho}.$$

Thus $\delta \leq \mathbb{P}\left[Z > \varepsilon\right] \leq e^{-(\varepsilon-\xi-\rho)^2/4\rho}$. Furthermore,

$$\begin{aligned}
\delta &= \int_{\varepsilon}^{\infty} e^{\varepsilon - z} \mathbb{P}\left[Z > z\right] \mathrm{d}z \\
&\leq \int_{\varepsilon}^{\infty} e^{\varepsilon - z} e^{-(z-\xi-\rho)^2/4\rho} \mathrm{d}z \\
&= \int_{\varepsilon}^{\infty} e^{-(z-\xi+\rho)^2/4\rho - \xi + \varepsilon} \mathrm{d}z \\
&= \frac{e^{\varepsilon-\xi} \cdot \sqrt{2\pi \cdot 2\rho}}{\sqrt{2\pi \cdot 2\rho}} \int_{\varepsilon}^{\infty} e^{-(z-\xi+\rho)^2/4\rho} \mathrm{d}z \\
&= e^{\varepsilon-\xi} \cdot \sqrt{2\pi \cdot 2\rho} \cdot \mathbb{P}\left[\mathcal{N}(\xi - \rho, 2\rho) > \varepsilon\right] \\
&= e^{\varepsilon-\xi} \cdot 2\sqrt{\pi \cdot \rho} \cdot \mathbb{P}\left[\mathcal{N}(0,1) > \frac{\varepsilon + \rho - \xi}{\sqrt{2\rho}}\right].
\end{aligned}$$

Define $h(x) = \mathbb{P}\left[\mathcal{N}(0,1) > x\right] \cdot \sqrt{2\pi} \cdot e^{x^2/2}$. Then

$$\delta \leq e^{\varepsilon-\xi} \cdot 2\sqrt{\pi \cdot \rho} \cdot \frac{h\left(\frac{\varepsilon+\rho-\xi}{\sqrt{2\rho}}\right)}{\sqrt{2\pi} \cdot e^{\left(\frac{\varepsilon+\rho-\xi}{\sqrt{2\rho}}\right)^2/2}} = \sqrt{2\rho} \cdot e^{-(\varepsilon-\xi-\rho)^2/4\rho} \cdot h\left(\sqrt{2\rho} + \frac{\varepsilon - \xi - \rho}{\sqrt{2\rho}}\right).$$

Now we can substitute upper bounds on h to obtain the desired results.

First we use $\mathbb{P}[\mathcal{N}(0,1) > x] \leq \frac{1}{2}e^{-x^2/2}$, which holds for all $x \geq 0$. That is, $h(x) \leq \sqrt{\pi/2}$, whence

$$\delta \leq \sqrt{2\rho} \cdot e^{-(\varepsilon-\xi-\rho)^2/4\rho} \cdot \sqrt{\frac{\pi}{2}}.$$

This rearranges to

$$\varepsilon \leq \xi + \rho + \sqrt{4\rho \cdot \log(\sqrt{\pi \cdot \rho}/\delta)}.$$

Alternatively, we can use $\mathbb{P}[\mathcal{N}(0,1) > x] \leq e^{-x^2/2}/\sqrt{2\pi}x$, which holds for all $x > 0$. That is, $h(x) \leq 1/x$ and

$$\delta \leq \sqrt{2\rho} \cdot e^{-(\varepsilon-\xi-\rho)^2/4\rho} \cdot h\left(\frac{\varepsilon-\xi+\rho}{\sqrt{2\rho}}\right) = \frac{2\rho}{\varepsilon+\rho-\xi} \cdot e^{-(\varepsilon-\xi-\rho)^2/4\rho} = \frac{e^{-(\varepsilon-\xi-\rho)^2/4\rho}}{1+(\varepsilon-\xi-\rho)/2\rho} \cdot .$$

Finally, we can use $\mathbb{P}[\mathcal{N}(0,1) > x] \leq e^{-x^2/2} \cdot 2/(\sqrt{x^2+8/\pi}+x)\sqrt{2\pi}$, which holds for all $x \geq 0$ [Due10, Equation (4)][Coo09, Equation (7)][AS64, Equation 7.1.13]. That is, $h(x) \leq 2/(\sqrt{x^2+8/\pi}+x)$ and

$$\delta \leq \sqrt{2\rho} \cdot e^{-(\varepsilon-\xi-\rho)^2/4\rho} \cdot h\left(\frac{\varepsilon-\xi+\rho}{\sqrt{2\rho}}\right) = \frac{2 \cdot e^{-(\varepsilon-\xi-\rho)^2/4\rho}}{1+\frac{\varepsilon-\xi-\rho}{2\rho}+\sqrt{\left(1+\frac{\varepsilon-\xi-\rho}{2\rho}\right)^2+\frac{4}{\pi\rho}}}.$$

□

Now we show a partial converse to Lemma 2.3.6.

Lemma 2.3.9. *Let $M : \mathcal{X}^n \to \mathcal{Y}$ satisfy (ε,δ)-DP for all $\delta > 0$ and*

$$\varepsilon = \hat{\xi} + \sqrt{\hat{\rho}\log(1/\delta)} \tag{2.4}$$

for some constants $\hat{\xi},\hat{\rho} \in [0,1]$. Then M is $\left(\hat{\xi} - \frac{1}{4}\hat{\rho} + 5\sqrt[4]{\hat{\rho}}, \frac{1}{4}\hat{\rho}\right)$-zCDP.

Thus zCDP and DP are equivalent up to a (potentially substantial) loss in parameters and the quantification over all δ.

Proof. Let $x, x' \in \mathcal{X}^n$ be neighbouring. Define

$$f(y) = \log\left(\frac{\mathbb{P}\left[M(x) = y\right]}{\mathbb{P}\left[M(x') = y\right]}\right).$$

Let $Y \sim M(x)$ and $Z = f(Y)$. That is, $Z = \mathsf{PrivLoss}\left(M(x) \| M(x')\right)$ is the privacy loss random variable. Let $Y' \sim M(x')$ and $Z' = f(Y')$. That is, $-Z'$ is the privacy loss random variable if we swap x and x'.

Let $\varepsilon, \delta > 0$ satisfy (1.1). By postprocessing, for all $t \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}\left[Z > t\right] =& \mathbb{P}\left[f(M(x)) > t\right] \\
\leq& e^{\varepsilon}\mathbb{P}\left[f(M(x')) > t\right] + \delta \\
=& e^{\varepsilon}\int_{\mathcal{Y}} \mathbb{P}\left[M(x') = y\right] \cdot \mathbb{I}(\mathbb{P}\left[M(x) = y\right] > e^{t}\mathbb{P}\left[M(x') = y\right])\mathrm{d}y + \delta \\
<& e^{\varepsilon}\int_{\mathcal{Y}} \mathbb{P}\left[M(x) = y\right] \cdot e^{-t} \cdot \mathbb{I}(\mathbb{P}\left[M(x) = y\right] > e^{t}\mathbb{P}\left[M(x') = y\right])\mathrm{d}y + \delta \\
=& e^{\varepsilon - t}\mathbb{P}\left[Z > t\right] + \delta,
\end{aligned}$$

whence $\mathbb{P}\left[Z > t\right] \leq \frac{\delta}{1 - e^{\varepsilon - t}}$. Then we can set $\varepsilon = \hat{\xi} + \lambda$ and $\delta = e^{-\lambda^2/\hat{\rho}}$ to obtain

$$\forall t > 0 \quad \mathbb{P}\left[Z > \hat{\xi} + t\right] \leq \inf_{0 < \lambda < t} \frac{e^{-\lambda^2/\hat{\rho}}}{1 - e^{\lambda - t}}.$$

In particular, for all $t \geq 0$,

$$\mathbb{P}\left[Z > \hat{\xi} + \sqrt[4]{\hat{\rho}} + t\right] \leq \frac{e^{-t^2/\hat{\rho}}}{1 - e^{-\sqrt[4]{\hat{\rho}}}} \leq \frac{2}{\sqrt[4]{\hat{\rho}}} e^{-t^2/\hat{\rho}}.$$

We use the inequality $1 + xe^{\xi} \leq e^{x+\xi}$ for all $x, \xi \geq 0$. We have

$$
\begin{aligned}
\mathbb{E}\left[e^{(\alpha-1)Z}\right] &= \int_0^\infty \mathbb{P}\left[e^{(\alpha-1)Z} > t\right] \mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\left[Z > \frac{\log t}{\alpha-1}\right] \mathrm{d}t \\
&= \int_{-\infty}^\infty \mathbb{P}\left[Z > z\right] \frac{\mathrm{d}t}{\mathrm{d}z} \mathrm{d}z \\
&= \int_{-\infty}^\infty (\alpha-1) e^{(\alpha-1)z} \cdot \mathbb{P}\left[Z > z\right] \mathrm{d}z \\
&\leq \int_{-\infty}^{\hat{\xi}+\sqrt[4]{\hat{\rho}}} (\alpha-1) e^{(\alpha-1)z} \cdot 1 \mathrm{d}z + \int_0^\infty (\alpha-1) e^{(\alpha-1)(t+\hat{\xi}+\sqrt[4]{\hat{\rho}})} \cdot \frac{2}{\sqrt[4]{\hat{\rho}}} e^{-t^2/\hat{\rho}} \mathrm{d}t \\
&= e^{(\alpha-1)(\hat{\xi}+\sqrt[4]{\hat{\rho}})} + (\alpha-1) e^{(\alpha-1)(\hat{\xi}+\sqrt[4]{\hat{\rho}})} \frac{2}{\sqrt[4]{\hat{\rho}}} \int_0^\infty e^{(\alpha-1)t} \cdot e^{-t^2/\hat{\rho}} \mathrm{d}t \\
&\leq e^{(\alpha-1)(\hat{\xi}+\sqrt[4]{\hat{\rho}})} + (\alpha-1) e^{(\alpha-1)(\hat{\xi}+\sqrt[4]{\hat{\rho}})} \frac{2}{\sqrt[4]{\hat{\rho}}} \int_{-\infty}^\infty e^{-(t-\frac{1}{2}(\alpha-1)\hat{\rho})^2/\hat{\rho} + \frac{1}{4}(\alpha-1)^2\hat{\rho}} \mathrm{d}t \\
&= e^{(\alpha-1)(\hat{\xi}+\sqrt[4]{\hat{\rho}})} + (\alpha-1) e^{(\alpha-1)(\hat{\xi}+\sqrt[4]{\hat{\rho}})} \frac{2}{\sqrt[4]{\hat{\rho}}} e^{\frac{1}{4}(\alpha-1)^2\hat{\rho}} \sqrt{\pi\hat{\rho}} \\
&= e^{(\alpha-1)(\hat{\xi}+\sqrt[4]{\hat{\rho}})} \left(1 + (\alpha-1) 2\sqrt{\pi} \sqrt[4]{\hat{\rho}} e^{\frac{1}{4}(\alpha-1)^2\hat{\rho}}\right) \\
&\leq e^{(\alpha-1)(\hat{\xi}+\sqrt[4]{\hat{\rho}})} e^{(\alpha-1)2\sqrt{\pi}\sqrt[4]{\hat{\rho}} + \frac{1}{4}(\alpha-1)^2\hat{\rho}} \\
&= e^{(\alpha-1)\left(\hat{\xi}+\sqrt[4]{\hat{\rho}}+2\sqrt{\pi}\sqrt[4]{\hat{\rho}}+\frac{1}{4}(\alpha-1)\hat{\rho}\right)} \\
&= e^{(\alpha-1)\left(\hat{\xi}+(1+2\sqrt{\pi})\sqrt[4]{\hat{\rho}}-\frac{\hat{\rho}}{4}+\frac{\hat{\rho}}{4}\alpha\right)}.
\end{aligned}
$$

Since $\mathbb{E}\left[e^{(\alpha-1)Z}\right] = e^{(\alpha-1)\mathrm{D}_\alpha(M(x)\|M(x'))}$, this completes the proof. □

2.4 Zero- versus Mean-Concentrated Differential Privacy

We begin by stating the definition of mean-concentrated differential privacy:

Definition 2.4.1 (Mean-Concentrated Differential Privacy (mCDP) [DR16])**.** *A randomised mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies (μ, τ)-mean-concentrated differential privacy if,*

for all $x, x' \in \mathcal{X}^n$ differing in one entry, and letting $Z = \mathsf{PrivLoss}\left(M(x) \| M(x')\right)$, we have

$$\mathbb{E}\left[Z\right] \leq \mu$$

and

$$\mathbb{E}\left[e^{\lambda\left(Z - \mathbb{E}[Z]\right)}\right] \leq e^{\lambda^2 \cdot \tau^2/2}$$

for all $\lambda \in \mathbb{R}$.

In contrast (ξ, ρ)-zCDP requires that, for all $\alpha \in (1, \infty)$, $\mathbb{E}\left[e^{(\alpha-1)Z}\right] \leq e^{(\alpha-1)(\xi+\rho\alpha)}$, where $Z \sim \mathsf{PrivLoss}\left(M(x) \| M(x')\right)$ is the privacy loss random variable. We now show that these definitions are equivalent up to a (potentially significant) loss in parameters.

Lemma 2.4.2. *If $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies (μ, τ)-mCDP, then M satisfies $(\mu - \tau^2/2, \tau^2/2)$-zCDP.*

Proof. For all $\alpha \in (1, \infty)$,

$$\mathbb{E}\left[e^{(\alpha-1)Z}\right] = \mathbb{E}\left[e^{(\alpha-1)(Z-\mathbb{E}[Z])}\right] \cdot e^{(\alpha-1)\mathbb{E}[Z]} \leq e^{(\alpha-1)^2\tau^2/2} \cdot e^{(\alpha-1)\mu} = e^{(\alpha-1)(\mu - \tau^2/2 + \tau^2/2\alpha)}.$$

□

Lemma 2.4.3. *If $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies (ξ, ρ)-zCDP,*
then M satisfies $(\xi + \rho, O(\sqrt{\xi + 2\rho}))$-mCDP.

Thus we can convert (μ, τ)-mCDP into $(\mu - \tau^2/2, \tau^2/2)$-zCDP and then back to $(\mu, O(\sqrt{\mu + \tau^2/2}))$-mCDP. This may result in a large loss in parameters, which is why, for example, pure DP can be characterised in terms of zCDP, but not in terms of mCDP.

We view zCDP as a relaxation of mCDP; mCDP requires the privacy loss to be "tightly" concentrated about its mean and that the mean is close to the origin.

The triangle inequality then implies that the privacy loss is "weakly" concentrated about the origin. (The difference between "tightly" and "weakly" accounts for the use of the triangle inequality.) On the other hand, zCDP directly requires that the privacy loss is weakly concentrated about the origin. That is to say, zCDP gives a subgaussian bound on the privacy loss that is centered at zero, whereas mCDP gives a subgaussian bound that is centered at the mean and separately bounds the mean.

There may be some advantage to the stronger requirement of mCDP, either in terms of what kind of privacy guarantee it affords, or how it can be used as an analytic tool. However, it seems that for most applications, we only need what zCDP provides.

To prove Lemma 2.4.2 make use of the following technical lemma.

Lemma 2.4.4. *Let X be a random variable. Then*

$$\mathbb{E}\left[e^{X-\mathbb{E}[X]}\right] \leq \frac{1}{2}\mathbb{E}\left[e^{2X}\right] + \frac{1}{2}\mathbb{E}\left[e^{-2X}\right].$$

Proof. Let X' be an independent copy of X and let $Y \in \{0,1\}$ be uniformly random and independent from X and X'. By Jensen's inequality,

$$\begin{aligned}
\mathbb{E}\left[e^{X-\mathbb{E}[X]}\right] &= \mathop{\mathbb{E}}_{X}\left[e^{\mathop{\mathbb{E}}_{X'}[X-X']}\right] \\
&\leq \mathop{\mathbb{E}}_{X,X'}\left[e^{X-X'}\right] \\
&= \mathop{\mathbb{E}}_{X,X'}\left[e^{2\mathop{\mathbb{E}}_{Y}[YX-(1-Y)X']}\right] \\
&\leq \mathop{\mathbb{E}}_{X,X',Y}\left[e^{2YX-2(1-Y)X'}\right] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{X,X'}\left[e^{2X-0}\right] + \frac{1}{2}\mathop{\mathbb{E}}_{X,X'}\left[e^{0-2X'}\right] \\
&= \frac{1}{2}\mathbb{E}\left[e^{2X}\right] + \frac{1}{2}\mathbb{E}\left[e^{-2X}\right].
\end{aligned}$$

□

Proof of Lemma 2.4.3. Let x and x' be neighbouring databases and $Z \sim \mathsf{PrivLoss}\left(M(x) \| M(x')\right)$ the privacy loss random variable We have $\mathbb{E}[Z] = \mathrm{D}_1\left(M(x) \| M(x')\right) \in [0, \xi + \rho]$ by non-negativity and zCDP. By zCDP, for all $\alpha \in (1, \infty)$, we have

$$\mathbb{E}\left[e^{(\alpha-1)Z}\right] = e^{(\alpha-1)\mathrm{D}_\alpha(M(x)\|M(x'))} \leq e^{(\alpha-1)(\xi+\rho\alpha)}$$

and

$$\begin{aligned}
\mathbb{E}\left[e^{-\alpha Z}\right] =& \mathop{\mathbb{E}}_{Y\sim M(x)}\left[\left(\frac{\mathbb{P}\left[M(x)=Y\right]}{\mathbb{P}\left[M(x')=Y\right]}\right)^{-\alpha}\right] \\
=& \mathop{\mathbb{E}}_{Y\sim M(x)}\left[\left(\frac{\mathbb{P}\left[M(x')=Y\right]}{\mathbb{P}\left[M(x)=Y\right]}\right)^{\alpha}\right] \\
=& e^{(\alpha-1)\mathrm{D}_\alpha(M(x')\|M(x))} \\
\leq& e^{(\alpha-1)(\xi+\rho\alpha)}.
\end{aligned}$$

By Lemma 2.4.4, for $\lambda \geq 1/2$,

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leq& \frac{1}{2}\mathbb{E}\left[e^{2\lambda Z}\right] + \frac{1}{2}\mathbb{E}\left[e^{-2\lambda Z}\right] \\
\leq& \frac{1}{2}e^{2\lambda(\xi+\rho(2\lambda+1))} + \frac{1}{2}e^{(2\lambda-1)(\xi+2\rho\lambda)} \\
\leq& e^{4(\xi+2\rho)\lambda^2}
\end{aligned}$$

and, for $\lambda \leq -1/2$,

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leq& \frac{1}{2}\mathbb{E}\left[e^{2\lambda Z}\right] + \frac{1}{2}\mathbb{E}\left[e^{-2\lambda Z}\right] \\
\leq& \frac{1}{2}e^{(-2\lambda-1)(\xi-2\rho\lambda)} + \frac{1}{2}e^{-2\lambda(\xi+\rho(1-2\lambda))} \\
\leq& e^{4(\xi+2\rho)\lambda^2}.
\end{aligned}$$

Now suppose $|\lambda| < 1/2$. Then

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \leq & \frac{1}{2}\mathbb{E}\left[e^{2\lambda Z}\right] + \frac{1}{2}\mathbb{E}\left[e^{-2\lambda Z}\right] \\
= & 1 + \sum_{k=1}^{\infty} \frac{(2\lambda)^{2k}}{(2k)!}\mathbb{E}\left[Z^{2k}\right] \\
\leq & 1 + (2\lambda)^2 \sum_{k=1}^{\infty} \frac{1}{(2k)!}\mathbb{E}\left[Z^{2k}\right] \\
= & 1 + 4\lambda^2\left(\frac{1}{2}\mathbb{E}\left[e^{Z}\right] + \frac{1}{2}\mathbb{E}\left[e^{-Z}\right] - 1\right) \\
\leq & 1 + 4\lambda^2\left(e^{\xi+2\rho} - 1\right) \\
\leq & e^{\lambda^2 O(\xi+2\rho)}.
\end{aligned}$$

□

2.4.1 Postprocessing and mCDP

We now give a family of counterexamples showing that mCDP is *not* closed under postprocessing (unlike zCDP).

Fix a parameter $\sigma > 0$, and consider the Gaussian mechanism for a single bit $M : \{-1, 1\} \to \mathbb{R}$, where $M(x)$ samples from $\mathcal{N}(x, \sigma^2)$. The mechanism M satisfies $(2/\sigma^2, 2/\sigma)$-mCDP (and also $(2/\sigma^2)$-zCDP).

Now consider the postprocessing function $T : \mathbb{R} \to \{-1, 0, 1\}$ defined as follows:

$$T(y) = \begin{cases} 1 & \text{if } y > t \\ -1 & \text{if } y < -t \\ 0 & \text{if } -t \leq y \leq t. \end{cases}$$

We examine the mCDP guarantees of the postprocessed mechanism $M' : \{-1, 1\} \to \{-1, 0, 1\}$ defined by $M'(x) = T(M(x))$:

Proposition 2.4.5. *Let $\sigma \geq 1$ and let $t \geq 6\sigma^3 + 1$. Then while the mechanism M is $(2/\sigma^2, 2/\sigma)$-mCDP, the postprocessed mechanism M' is not $(2/\sigma^2, 2/\sigma)$-mCDP.*

Proof. For each $x \in \{-1, 1\}$, let

$$p = \mathbb{P}\left[M'(x) = x\right] = \mathbb{P}\left[\mathcal{N}(0, \sigma^2) > t - 1\right],$$
$$q = \mathbb{P}\left[M'(x) = -x\right] = \mathbb{P}\left[\mathcal{N}(0, \sigma^2) > t + 1\right].$$

Note that $p > q$. Hence, for each $x \in \{-1, 1\}$,

$$\mathbb{P}\left[M'(x) = 0\right] = \mathbb{P}\left[\mathcal{N}(0, \sigma^2) \in [-t - 1, t - 1]\right] = 1 - p - q.$$

Let $f(y) = \log(\mathbb{P}\left[M(1) = y\right] / \mathbb{P}\left[M(-1) = y\right])$, and observe that

$$f(1) = \log \frac{p}{q}, \qquad f(-1) = \log \frac{q}{p}, \qquad f(0) = 0.$$

Now consider the privacy loss random variable $Z = \mathsf{PrivLoss}\left(M(1) \| M(-1)\right) = f(M(1))$. Then Z is distributed according to

$$Z = \begin{cases} \log \frac{p}{q} & \text{w.p. } p \\ \log \frac{q}{p} & \text{w.p. } q \\ 0 & \text{w.p. } 1 - p - q \end{cases}.$$

This gives $\mathbb{E}\left[Z\right] = (p - q)\log(p/q) \geq 0$. For $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] = \left(p\left(\frac{p}{q}\right)^{\lambda} + q\left(\frac{q}{p}\right)^{\lambda} + 1 - p - q\right) \cdot \left(\frac{p}{q}\right)^{-\lambda(p-q)}.$$

If M' were to satisfy $(2/\sigma^2, 2/\sigma)$-mCDP, we would have $\mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] \leq e^{2\lambda^2/\sigma^2}$ for all $\lambda > 0$. We will show that this does not hold for any setting of parameters

$\sigma \geq 1$ and $t \geq 6\sigma^3 + 1$, which shows that mCDP is not closed under postprocessing.[11]

Lemma 2.4.6. *The values p, q satisfy the following inequalities:*

1. $\sqrt{\frac{1}{2\pi}} \cdot \frac{\sigma}{t+\sigma-1} \cdot e^{-(t-1)^2/2\sigma^2} \leq p \leq \sqrt{\frac{1}{2\pi}} \cdot \frac{\sigma}{t-1} \cdot e^{-(t-1)^2/2\sigma^2}$
2. $\frac{p}{q} \geq e^{2t/\sigma^2}$

Proof. We have [Coo09, Equation (5)]

$$\frac{\sqrt{\frac{2}{\pi}} e^{-x^2/2} \cdot x}{x^2+1} \leq \mathbb{P}\left[\mathcal{N}(0,1) > x\right] \leq \frac{\sqrt{\frac{2}{\pi}} e^{-x^2/2}}{x}$$

for all $x \geq 0$. Thus

$$\frac{\sqrt{\frac{2}{\pi}} e^{-(t-1)^2/2\sigma^2} \cdot (t-1)}{\sigma \cdot ((t-1)^2/\sigma^2 + 1)} \leq p = \mathbb{P}\left[\mathcal{N}(0,1) > \frac{t-1}{\sigma}\right] \leq \frac{\sqrt{\frac{2}{\pi}} e^{-(t-1)^2/2\sigma^2} \cdot \sigma}{t-1}.$$

The first part now follows from the fact that $\sigma \leq t-1$.

To establish the second inequality, we write

$$\begin{aligned}
p &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{t-1}^{\infty} e^{-x^2/2\sigma^2}\, dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{t+1}^{\infty} e^{-(u-2)^2/2\sigma^2}\, du \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{t+1}^{\infty} e^{(4u-4)/2\sigma^2} \cdot e^{-u^2/2\sigma^2}\, du \\
&\geq e^{2t/\sigma^2} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{t+1}^{\infty} e^{-u^2/2\sigma^2}\, du \\
&= e^{2t/\sigma^2} \cdot q.
\end{aligned}$$

□

[11] For specific settings of parameters, this can be verified numerically. (For example, with the values $\sigma = 1$, $t = 3$, and $\lambda = 2$.)

By Lemma 2.4.6,

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] &= \left(p\left(\frac{p}{q}\right)^{\lambda} + q\left(\frac{q}{p}\right)^{\lambda} + 1 - p - q\right)\cdot\left(\frac{p}{q}\right)^{-\lambda(p-q)} \\ &\geq p\left(\frac{p}{q}\right)^{\lambda(1-(p-q))} \\ &\geq \sqrt{\frac{1}{2\pi}}\cdot\frac{\sigma}{t+\sigma-1}\cdot e^{-(t-1)^2/2\sigma^2}\cdot\left(e^{2t/\sigma^2}\right)^{\lambda(1-p)}.\end{aligned}$$

Now set $\lambda = \frac{t}{2}$. Then this quantity becomes

$$\sqrt{\frac{1}{2\pi}}\cdot\frac{\sigma}{t+\sigma-1}\cdot e^{-(t-1)^2/2\sigma^2}\cdot e^{t^2(1-p)/\sigma^2} = e^{t^2/2\sigma^2}\cdot\sqrt{\frac{1}{2\pi}}\cdot\frac{\sigma}{t+\sigma-1}\cdot\exp\left(\frac{t}{\sigma^2} - \frac{pt^2}{\sigma^2} - \frac{1}{2\sigma^2}\right). \quad (2.5)$$

We now examine the term

$$\begin{aligned}pt^2 &\leq \sqrt{\frac{1}{2\pi}}\frac{\sigma t^2 e^{-(t-1)^2/2\sigma^2}}{t-1} && \text{by Lemma 2.4.6} \\ &\leq \sqrt{\frac{1}{4\pi}}\frac{t^{5/2}e^{-(t-1)}}{t-1} && \text{for } t \geq 2\sigma^2+1 \\ &\leq \frac{1}{2} && \text{for } t \geq 3.\end{aligned}$$

We may thus bound (2.5) from below by

$$e^{t^2/2\sigma^2}\cdot\sqrt{\frac{1}{2\pi}}\cdot\frac{\sigma}{t+\sigma-1}\cdot\exp\left(\frac{t-1}{2\sigma^2}\right) \geq e^{t^2/2\sigma^2}\cdot\sqrt{\frac{1}{2\pi}}\cdot\frac{1}{2t}\cdot\exp\left(\frac{t-1}{2\sigma^2}\right)$$

The expression $\exp((t-1)/2\sigma^2)/2t$ is monotone increasing for $t \geq 2\sigma^2$. Thus, it is strictly larger than $\sqrt{2\pi}$ as long as $t \geq 6\sigma^3 + 1$. □

2.5 Group Privacy

In this section we show that zCDP provides privacy protections to small groups of individuals.

Definition 2.5.1 (zCDP for Groups)**.** *We say that a mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ provides (ξ, ρ)-zCDP for groups of size k if, for every $x, x' \in \mathcal{X}^n$ differing in at most k entries, we have*

$$\forall \alpha \in (1, \infty) \qquad \mathrm{D}_\alpha \left(M(x) \middle\| M(x') \right) \leq \xi + \rho \cdot \alpha.$$

The usual definition of zCDP only applies to groups of size 1. Here we show that it implies bounds for all group sizes. We begin with a technical lemma.

Lemma 2.5.2 (Triangle-like Inequality for Rényi Divergence)**.** *Let P, Q, and R be probability distributions. Then*

$$\mathrm{D}_\alpha \left(P \| Q \right) \leq \frac{k\alpha}{k\alpha - 1} \mathrm{D}_{\frac{k\alpha - 1}{k-1}} \left(P \| R \right) + \mathrm{D}_{k\alpha} \left(R \| Q \right) \tag{2.6}$$

for all $k, \alpha \in (1, \infty)$.

Proof. Let $p = \frac{k\alpha - 1}{\alpha(k-1)}$ and $q = \frac{k\alpha - 1}{\alpha - 1}$. Then $\frac{1}{p} + \frac{1}{q} = \frac{\alpha(k-1) + (\alpha - 1)}{k\alpha - 1} = 1$. By Hölder's inequality,

$$\begin{aligned}
e^{(\alpha-1)\mathrm{D}_\alpha(P\|Q)} &= \int_\Omega P(x)^\alpha Q(x)^{1-\alpha} \mathrm{d}x \\
&= \int_\Omega P(x)^\alpha R(x)^{-\alpha} \cdot R(x)^{\alpha-1} Q(x)^{1-\alpha} \cdot R(x) \mathrm{d}x \\
&= \mathop{\mathbb{E}}_{x \sim R} \left[\left(\frac{P(x)}{R(x)} \right)^\alpha \cdot \left(\frac{R(x)}{Q(x)} \right)^{\alpha - 1} \right] \\
&\leq \mathop{\mathbb{E}}_{x \sim R} \left[\left(\frac{P(x)}{R(x)} \right)^{p\alpha} \right]^{1/p} \cdot \mathop{\mathbb{E}}_{x \sim R} \left[\left(\frac{R(x)}{Q(x)} \right)^{q(\alpha-1)} \right]^{1/q} \\
&= e^{(p\alpha - 1)\mathrm{D}_{p\alpha}(P\|R)/p} \cdot e^{q(\alpha-1)\mathrm{D}_{q(\alpha-1)+1}(R\|Q)/q}.
\end{aligned}$$

Taking logarithms and rearranging gives

$$\mathrm{D}_\alpha \left(P \| Q \right) \leq \frac{p\alpha - 1}{p(\alpha - 1)} \mathrm{D}_{p\alpha} \left(P \| R \right) + \mathrm{D}_{q(\alpha-1)+1} \left(R \| Q \right).$$

Now $p\alpha = \frac{k\alpha-1}{k-1}$, $q(\alpha-1)+1 = k\alpha$, and

$$\frac{\frac{p\alpha-1}{p(\alpha-1)}}{\frac{k\alpha}{k\alpha-1}} = \frac{p\alpha-1}{p\alpha}\cdot\frac{k\alpha-1}{k(\alpha-1)} = \frac{\frac{k\alpha-1}{k-1}-1}{\frac{k\alpha-1}{k-1}}\cdot\frac{k\alpha-1}{k(\alpha-1)} = \frac{k\alpha-1-k+1}{k\alpha-1}\cdot\frac{k\alpha-1}{k(\alpha-1)} = 1,$$

as required. □

Proposition 2.5.3. *If $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies (ξ,ρ)-zCDP, then M gives $(\xi \cdot k\sum_{i=1}^{k}\frac{1}{i}, \rho\cdot k^2)$-zCDP for groups of size k.*

Note that

$$\sum_{i=1}^{k}\frac{1}{i} = 1 + \int_1^k \frac{1}{\lceil x\rceil}\mathrm{d}x \le 1 + \int_1^k \frac{1}{x}\mathrm{d}x = 1+\log k.$$

Thus (ξ,ρ)-zCDP implies $(\xi\cdot O(k\log k), \rho\cdot k^2)$-zCDP for groups of size k.

The Gaussian mechanism shows that $k^2\rho$ is the optimal dependence on ρ. However, $O(k\log k)\xi$ is not the optimal dependence on ξ: $(\xi,0)$-zCDP implies $(k\xi,0)$-zCDP for groups of size k.

Proof. We show this by induction on k. The statement is clearly true for groups of size 1. We now assume the statement holds for groups of size $k-1$ and will verify it for groups of size k.

Let $x, x' \in \mathcal{X}^n$ differ in k entries. Let $\hat{x}\in\mathcal{X}^n$ be such that x and $\hat{x}$ differ in $k-1$ entries and x' and $\hat{x}$ differ in one entry.

Then, by the induction hypothesis,

$$\mathrm{D}_\alpha\left(M(x)\|M(\hat{x})\right) \le \xi\cdot(k-1)\sum_{i=1}^{k-1}\frac{1}{i} + \rho\cdot(k-1)^2\cdot\alpha$$

and, by zCDP,

$$\mathrm{D}_\alpha\left(M(\hat{x})\middle\|M(x')\right) \le \xi+\rho\cdot\alpha$$

for all $\alpha\in(1,\infty)$.

By (2.6), for any $\alpha \in (1, \infty)$,

$$\begin{aligned}
\mathrm{D}_\alpha\left(M(x)\middle\|M(x')\right) \leq & \frac{k\alpha}{k\alpha-1}\mathrm{D}_{\frac{k\alpha-1}{k-1}}\left(M(x)\|M(\hat{x})\right)+\mathrm{D}_{k\alpha}\left(M(\hat{x})\middle\|M(x')\right) \\
\leq & \frac{k\alpha}{k\alpha-1}\left(\xi\cdot(k-1)\sum_{i=1}^{k-1}\frac{1}{i}+\rho\cdot(k-1)^2\cdot\frac{k\alpha-1}{k-1}\right)+\xi+\rho\cdot k\alpha \\
= & \xi\cdot\left(1+\frac{k\alpha}{k\alpha-1}(k-1)\sum_{i=1}^{k-1}\frac{1}{i}\right)+\rho\cdot\left(\frac{k\alpha}{k\alpha-1}(k-1)^2\frac{k\alpha-1}{k-1}+k\alpha\right) \\
= & \xi\cdot\left(1+\frac{k\alpha}{k\alpha-1}(k-1)\sum_{i=1}^{k-1}\frac{1}{i}\right)+\rho\cdot k^2\cdot\alpha \\
\leq & \xi\cdot\left(1+\frac{k}{k-1}(k-1)\sum_{i=1}^{k-1}\frac{1}{i}\right)+\rho\cdot k^2\cdot\alpha \\
= & \xi\cdot k\sum_{i=1}^{k}\frac{1}{i}+\rho\cdot k^2\cdot\alpha,
\end{aligned}$$

where the last inequality follows from the fact that $\frac{k\alpha}{k\alpha-1}$ is a decreasing function of α for $\alpha > 1$. □

2.6 Lower Bounds

In this section we develop tools to prove lower bounds for zCDP. We will use group privacy to bound the mutual information between the input and the output of a mechanism satisfying zCDP. Thus, if we are able to construct a distribution on inputs such that any accurate mechanism must reveal a high amount of information about its input, we obtain a lower bound showing that no accurate mechanism satisfying zCDP can be accurate for this data distribution.

We begin with the simplest form of our mutual information bound, which is an analogue of the bound of [MMP+10] for pure differential privacy:

Proposition 2.6.1. *Let $M : \mathcal{X}^n \to \mathcal{Y}$ satisfy (ξ, ρ)-zCDP. Let X be a random variable in*

$\mathcal{X}^n$. *Then*

$$I(X; M(X)) \leq \xi \cdot n(1 + \log n) + \rho \cdot n^2,$$

where I denotes mutual information (measured in nats, rather than bits).

Proof. By Proposition 2.5.3, M provides $(\xi \cdot n \sum_{i=1}^{n} \frac{1}{i}, \rho \cdot n^2)$-zCDP for groups of size n. Thus

$$\mathrm{D}_1\left(M(x) \middle\| M(x')\right) \leq \xi \cdot n \sum_{i=1}^{n} \frac{1}{i} + \rho \cdot n^2 \leq \xi \cdot n(1 + \log n) + \rho \cdot n^2$$

for all $x, x' \in \mathcal{X}^n$. Since KL-divergence is convex,

$$\begin{aligned} I(X; M(X)) &= \mathop{\mathbb{E}}_{x \leftarrow X}\left[\mathrm{D}_1\left(M(x) \| M(X)\right)\right] \\ &\leq \mathop{\mathbb{E}}_{x \leftarrow X}\left[\mathop{\mathbb{E}}_{x' \leftarrow X}\left[\mathrm{D}_1\left(M(x) \middle\| M(x')\right)\right]\right] \\ &\leq \mathop{\mathbb{E}}_{x \leftarrow X}\left[\mathop{\mathbb{E}}_{x' \leftarrow X}\left[\xi \cdot n(1 + \log n) + \rho \cdot n^2\right]\right] \\ &= \xi \cdot n(1 + \log n) + \rho \cdot n^2. \end{aligned}$$

□

The reason this lower bound works is the strong group privacy guarantee — even for groups of size n, we obtain nontrivial privacy guarantees. While this is good for privacy it is bad for usefulness, as it implies that even information that is "global" (rather than specific to a individual or a small group) is protected. These lower bounds reinforce the connection between group privacy and lower bounds [HT10, De12, §4].

In contrast, (ε, δ)-DP is not susceptible to such a lower bound because it gives a vacuous privacy guarantee for groups of size $k = O(\log(1/\delta)/\varepsilon)$. This helps explain the power of the propose-test-release paradigm.

Furthermore, we obtain even stronger mutual information bounds when the

entries of the distribution are independent:

Lemma 2.6.2. *Let $M : \mathcal{X}^m \to \mathcal{Y}$ satisfy (ξ, ρ)-zCDP. Let X be a random variable in $\mathcal{X}^m$ with independent entries. Then*

$$I\left(X; M(X)\right) \leq (\xi + \rho) \cdot m,$$

where I denotes mutual information (measured in nats, rather than bits).

Proof. First, by the chain rule for mutual information,

$$I(X; M(X)) = \sum_{i \in [m]} I(X_i; M(X) | X_{1 \cdots i-1}),$$

where

$$\begin{aligned} I(X_i; M(X) | X_{1 \cdots i-1}) &= \mathop{\mathbb{E}}_{x \leftarrow X_{1 \cdots i-1}} \left[I(X_i | X_{1 \cdots i-1} = x; M(X) | X_{1 \cdots i-1} = x) \right] \\ &= \mathop{\mathbb{E}}_{x \leftarrow X_{1 \cdots i-1}} \left[I(X_i; M(x, X_{i \cdots m})) \right], \end{aligned}$$

by independence of the X_is.

We can define mutual information in terms of KL-divergence:

$$\begin{aligned} I(X_i; M(x, X_{i \cdots m})) &= \mathop{\mathbb{E}}_{y \leftarrow X_i} \left[\mathrm{D}_1 \left(M(x, X_{i \cdots m}) | X_i = y \| M(x, X_{i \cdots m}) \right) \right] \\ &= \mathop{\mathbb{E}}_{y \leftarrow X_i} \left[\mathrm{D}_1 \left(M(x, y, X_{i+1 \cdots m}) \| M(x, X_{i \cdots m}) \right) \right]. \end{aligned}$$

By zCDP, we know that for all $x \in \mathcal{X}^{i-1}$, $y, y' \in \mathcal{X}$, and $z \in \mathcal{X}^{m-i}$, we have

$$\mathrm{D}_1 \left(M(x, y, z) \middle\| M(x, y', z) \right) \leq \xi + \rho.$$

Thus, by the convexity of KL-divergence,

$$\mathrm{D}_1 \left(M(x, y, X_{i+1 \cdots m}) \| M(x, X_{i \cdots m}) \right) \leq \xi + \rho$$

for all x and y. The result follows. □

More generally, we can combine dependent and independent entries as follows

Theorem 2.6.3. *Let $M : \mathcal{X}^n \to \mathcal{Y}$ satisfy (ξ, ρ)-zCDP. Take $n = m \cdot \ell$. Let $X^1, \cdots, X^m$ be independent random variables on $\mathcal{X}^\ell$. Denote $X = (X^1, \cdots, X^m) \in \mathcal{X}^n$. Then*

$$I\left(X; M(X)\right) \leq m \cdot \left(\xi \cdot \ell(1 + \log \ell) + \rho \cdot \ell^2\right),$$

where I denotes the mutual information (measured in nats, rather than bits).

Proof. By Proposition 2.5.3, M provides $(\xi \cdot \ell \sum_{i=1}^{\ell} \frac{1}{i}, \rho \cdot \ell^2)$-zCDP for groups of size ℓ. Thus

$$\mathrm{D}_1\left(M(x_1, \cdots, x_i, \cdots, x_m) \middle\| M(x_1, \cdots, x_i', \cdots, x_m)\right) \leq \xi \cdot \ell \sum_{i=1}^{\ell} \frac{1}{i} + \rho \cdot \ell^2 \leq \xi \cdot \ell(1 + \log \ell) + \rho \cdot \ell^2 \tag{2.7}$$

for all $x_1, \cdots, x_m, x_i' \in \mathcal{X}^\ell$.

By the chain rule for mutual information,

$$I(X; M(X)) = \sum_{i \in [m]} I(X_i; M(X) | X_{1 \cdots i-1}),$$

where

$$\begin{aligned} I(X_i; M(X) | X_{1 \cdots i-1}) &= \mathop{\mathbb{E}}_{x \leftarrow X_{1 \cdots i-1}} \left[I(X_i | X_{1 \cdots i-1} = x; M(X) | X_{1 \cdots i-1} = x)\right] \\ &= \mathop{\mathbb{E}}_{x \leftarrow X_{1 \cdots i-1}} \left[I(X_i; M(x, X_{i \cdots m}))\right], \end{aligned}$$

by independence of the X_is.

We can define mutual information in terms of KL-divergence:

$$I(X_i; M(x, X_{i\cdots m})) = \mathop{\mathbb{E}}_{y \leftarrow X_i} \left[\mathrm{D}_1 \left(M(x, X_{i\cdots m}) | X_i = y \| M(x, X_{i\cdots m}) \right) \right]$$
$$= \mathop{\mathbb{E}}_{y \leftarrow X_i} \left[\mathrm{D}_1 \left(M(x, y, X_{i+1\cdots m}) \| M(x, X_{i\cdots m}) \right) \right].$$

By (2.7) and the convexity of KL-divergence,

$$\mathrm{D}_1 \left(M(x, y, X_{i+1\cdots m}) \| M(x, X_{i\cdots m}) \right) \leq \xi \cdot \ell(1 + \log \ell) + \rho \cdot \ell^2$$

for all x and y. The result follows. □

2.6.1 Example Applications of the Lower Bound

We informally discuss a few applications of our information-based lower bounds to some simple and well-studied problems in differential privacy.

One-Way Marginals Consider $M : \mathcal{X}^n \to \mathcal{Y}$ where $\mathcal{X} = \{0,1\}^d$ and $\mathcal{Y} = [0,1]^d$. The goal of M is to estimate the attribute means, or one-way marginals, of its input database x:

$$M(x) \approx \overline{x} = \frac{1}{n} \sum_{i \in [n]} x_i.$$

It is known that this is possible subject to ε-DP if and only if $n = \Theta(d/\varepsilon)$ [HT10, §4]. This is possible subject to (ε, δ)-DP if and only if $n = \tilde{\Theta}(\sqrt{d \log(1/\delta)}/\varepsilon)$, assuming $\delta \ll 1/n$ [BUV14, §4].

We now analyse what can be accomplished with zCDP. Adding independent noise drawn from $\mathcal{N}(0, d/2n^2\rho)$ to each of the d coordinates of $\overline{x}$ satisfies ρ-zCDP. This gives accurate answers as long as $n \gg \sqrt{d/\rho}$.

For a lower bound, consider sampling $X_1 \in \{0,1\}^d$ uniformly at random. Set

$X_i = X_1$ for all $i \in [n]$. By Proposition 2.6.1,

$$I(X; M(X)) \leq n^2\rho$$

for any ρ-zCDP $M : (\{0,1\}^d)^n \to [0,1]^d$. However, if M is accurate, we can recover (most of) X_1 from $M(X)$, whence $I(X; M(X)) \geq \Omega(d)$. This yields a lower bound of $n \geq \Omega(\sqrt{d/\rho})$, which is tight up to constant factors.

Histograms (a.k.a. Point Queries) Consider $M : \mathcal{X}^n \to \mathcal{Y}$, where $\mathcal{X} = [T]$ and $\mathcal{Y} = \mathbb{R}^T$. The goal of M is to estimate the histogram of its input:

$$M(x)_t \approx h_t(x) = |\{i \in [n] : x_i = t\}|$$

For ε-DP it is possible to do this if and only if $n = \Theta(\log(T)/\varepsilon)$; the optimal algorithm is to independently sample

$$M(x)_t \sim h_t(x) + \mathsf{Laplace}(2/\varepsilon).$$

However, for (ε, δ)-DP, it is possible to attain sample complexity $n = O(\log(1/\delta)/\varepsilon)$ [BNS13, BNS16b, Theorem 3.13]. Interestingly, for zCDP we can show that $n = \Theta(\sqrt{\log(T)/\rho})$ is sufficient and necessary:

Sampling

$$M(x)_t \sim h_t(x) + \mathcal{N}(0, 1/\rho)$$

independently for $t \in [T]$ satisfies ρ-zCDP. Moreover,

$$\mathbb{P}\left[\max_{t\in[T]} |M(x)_t - h_t(x)| \geq \lambda\right] \leq T \cdot \mathbb{P}\left[|\mathcal{N}(0, 1/\rho)| > \lambda\right] \leq T \cdot e^{-\lambda^2\rho/2}.$$

In particular $\mathbb{P}\left[\max_{t\in[T]} |M(x)_t - h_t(x)| \geq \sqrt{\log(T/\beta)/\rho}\right] \leq \beta$ for all $\beta > 0$. Thus this algorithm is accurate if $n \gg \sqrt{\log(T)/\rho}$.

On the other hand, if we sample $X_1 \in [T]$ uniformly at random and set $X_i = X_1$

for all $i \in [n]$, then $I(X; M(X)) \geq \Omega(\log T)$ for any accurate M, as we can recover X_1 from $M(X)$ if M is accurate. Proposition 2.6.1 thus implies that $n \geq \Omega(\sqrt{\log(T)/\rho})$ is necessary to obtain accuracy.

This gives a strong separation between approximate DP and zCDP.

Approximate Maximisation Consider $M : \{\pm 1\}^n \to \{\pm 1\}^n$ where the goal is to maximise $\langle x, M(x)\rangle$ subject to ρ-zCDP.

One solution is randomised response [War65]: Each output bit i of M is chosen independently with

$$\mathbb{P}\left[M(x)_i = x_i\right] = \frac{e^\varepsilon}{e^\varepsilon + 1}.$$

This satisfies ε-DP and, hence, $\frac{1}{2}\varepsilon^2$-zCDP. And $\mathbb{E}\left[\langle x, M(x)\rangle\right] = n(e^\varepsilon - 1)/(e^\varepsilon + 1) = \Theta(n\varepsilon)$. Alternatively, we can independently choose the output bits i according to

$$M(x)_i = \operatorname{sign}\left(\mathcal{N}(x_i, \sqrt{2}/\rho)\right),$$

which satisfies ρ-zCDP.

It is known that these approaches are essentially optimal [DN03]. The mutual information bound for independent entries (Lemma 2.6.2) can be used to give a lower bound for zCDP:

Let $X \in \{\pm 1\}^n$ be uniformly random. By Lemma 2.6.2, since the bits of X are independent, we have $I(X; M(X)) \leq \rho \cdot n$ for any ρ-zCDP M. However, if M is accurate, we can recover part of X from $M(X)$ [DN03, §2], whence $I(X; M(X)) \geq \Omega(n)$.

Lower Bounds with Accuracy The above examples can be easily discussed in terms of a more formal and quantitative definition of accuracy. In particular, we consider the histogram example again:

Proposition 2.6.4. *If $M : [T]^n \to \mathbb{R}^T$ satisfies ρ-zCDP and*

$$\forall x \in [T]^n \qquad \mathop{\mathbb{E}}_{M}\left[\max_{t\in[T]} \left|M(x)_t - h_t(x)\right|\right] \leq \alpha n,$$

then $n \geq \Omega(\sqrt{\log(\alpha^2 T)/\rho\alpha^2})$.

Proof. Let $m = 1/10\alpha$ and $\ell = n/m$. For simplicity, assume that both m and n are integral.

Let $X_1, X_2, \cdots, X_m \in [T]^\ell$ be independent, where each X_i is ℓ copies of a uniformly random element of $[T]$. By Theorem 2.6.3,

$$I(X; M(X)) \leq \rho \cdot m \cdot \ell^2 = 10\rho\alpha n^2, \tag{2.8}$$

where $X = (X_1, \cdots, X_m) \in \mathcal{X}^n$. However,

$$\begin{aligned} I(X; M(X)) \geq & I(f(X); g(M(X))) \\ = & H(f(X)) - H(f(X)|g(M(X))) \\ = & H(X) - H(X|f(X)) - H(f(X)|g(M(X))) \end{aligned}$$

for any functions f and g, where H is the entropy (in nats). In particular, we let

$$f(x) = \{t \in T : \exists i \in [n] \;\; x_i = t\} \qquad \text{and} \qquad g(y) = \{t \in T : y_t \geq 5\alpha n\}.$$

Clearly $H(X) = m \log T$. Furthermore, $H(X|f(X)) \leq m \log m$, since X can be specified by naming m elements of $f(X)$, which is a set of at most m elements.

If

$$\max_{t\in[T]} \left|M(X)_t - h_t(X)\right| < 5\alpha, \tag{2.9}$$

then $g(M(X))$ contains exactly all the values in X — i.e. $f(X) = g(M(X))$. By Markov's inequality, (2.9) holds with probability at least $4/5$.

Now we can upper bound $H(f(X)|g(M(X)))$ by giving a scheme for specifying

$f(X)$ given $g(M(X))$. If (2.9) holds, we simply need one bit to say so. If (2.9) does not hold, we need one bit to say this and $m \log_2 T$ bits to describe $f(X)$. This gives

$$H(f(X)|g(M(X))) \leq \log 2 + \mathbb{P}\left[f(X) \neq g(M(X))\right] \cdot m \log T.$$

Combining these inequalities gives

$$I(X;M(X)) \geq m \log T - m \log m - \log 2 - \frac{1}{5} m \log T \geq \frac{4}{5} m \log(T m^{-5/4}) - 1 \geq \Omega(\log(\alpha^{1.25} T)/\alpha).$$

Combining this with (2.8) completes the proof. □

We remark that our lower bounds for zCDP can be converted to lower bounds for mCDP using Lemma 2.4.2.

2.7 Obtaining Pure DP Mechanisms from zCDP

We now establish limits on what more can be achieved with zCDP over pure differential privacy. In particular, we will prove that any mechanism satisfying zCDP can be converted into a mechanism satisfying pure DP with at most a quadratic blowup in sample complexity. Formally, we show the following theorem.

Theorem 2.7.1. *Fix $n \in \mathbb{N}$, $n' \in \mathbb{N}$, $k \in \mathbb{N}$ $\alpha > 0$, and $\varepsilon > 0$. Let $q : \mathcal{X} \to \mathbb{R}^k$ and let $\|\cdot\|$ be a norm on $\mathbb{R}^k$. Assume $\max_{x \in \mathcal{X}} \|q(x)\| \leq 1$.*

Suppose there exists a (ξ, ρ)-zCDP mechanism $M : \mathcal{X}^n \to \mathbb{R}^k$ such that for all $x \in \mathcal{X}^n$,

$$\mathop{\mathbb{E}}_{M}\left[\|M(x) - q(x)\|\right] \leq \alpha.$$

Assume $\xi \leq \alpha^2$, $\rho \leq \alpha^2$, and

$$n' \geq \frac{4}{\varepsilon\alpha}\left(\rho \cdot n^2 + \xi \cdot n \cdot (1 + \log n) + 1\right).$$

Then there exists a $(\varepsilon, 0)$-differentially private $M' : \mathcal{X}^{n'} \to \mathbb{R}^k$ satisfying

$$\mathop{\mathbb{E}}_{M'}\left[\|M'(x) - q(x)\|\right] \leq 10\alpha$$

and

$$\mathop{\mathbb{P}}_{M'}\left[\|M'(x) - q(x)\| > 10\alpha + \frac{4}{\varepsilon n'}\log\left(\frac{1}{\beta}\right)\right] \leq \beta$$

for all $x \in \mathcal{X}^{n'}$ and $\beta > 0$.

Before discussing the proof of Theorem 2.7.1, we make some remarks about its statement:

- Unfortunately, the theorem only works for families of statistical queries $q : \mathcal{X} \to \mathbb{R}^k$. However, it works equally well for $\|\cdot\|_\infty$ and $\|\cdot\|_1$ error bounds.
- If $\xi = 0$, we have $n' = O(n^2\rho/\varepsilon\alpha)$. So, if ρ, ε, and α are all constants, we have $n' = O(n^2)$. This justifies our informal statement that we can convert any mechanism satisfying zCDP into one satisfying pure DP with a quadratic blowup in sample complexity.
- Suppose $M : \mathcal{X}^n \to \mathbb{R}^k$ is the Gaussian mechanism scaled to satisfy ρ-zCDP and $\|\cdot\| = \|\cdot\|_1/k$. Then

  $$\alpha = \mathbb{E}\left[\|M(x) - q(x)\|\right] = \Theta\left(\sqrt{\frac{k}{\rho n^2}}\right).$$

  In particular, $n = \Theta(\sqrt{k/\rho\alpha^2})$. The theorem then gives us a ε-DP $M' : \mathcal{X}^{n'} \to \mathbb{R}^k$ with $\mathbb{E}\left[\|M'(x) - q(x)\|\right] \leq O(\alpha)$ for

  $$n' = \Theta\left(\frac{n^2\rho}{\varepsilon\alpha}\right) = \Theta\left(\frac{k}{\alpha^3\varepsilon}\right).$$

  However, the Laplace mechanism achieves ε-DP and $\mathbb{E}\left[\|M'(x) - q(x)\|\right] \leq \alpha$ with $n = \Theta(k/\alpha\varepsilon)$.

This example illustrates that the theorem is not tight in terms of α; it loses a $1/\alpha^2$ factor here. However, the other parameters are tight.

- The requirement that $\xi, \rho \leq \alpha^2$ is only used to show that

$$\max_{x \in \mathcal{X}^{n'}} \min_{\hat{x} \in \mathcal{X}^n} \|q(x) - q(\hat{x})\| \leq 2\alpha \tag{2.10}$$

using Lemma 2.7.5. However, in many situations (2.10) holds even when $\xi, \rho \gg \alpha^2$. For example, if $n \geq O(\log(k)/\alpha^2)$ or even $n \geq O(VC(q)/\alpha^2)$ then (2.10) is automatically satisfied.

The technical condition (2.10) is needed to relate the part of the proof with inputs of size n to the part with inputs of size n'.

Thus we can restate Theorem 2.7.1 with the condition $\xi, \rho \leq \alpha^2$ replaced by (2.10). This would be more general, but also more mysterious.

Alas, the proof of Theorem 2.7.1 is not constructive. Rather than directly constructing a mechanism satisfying pure DP from any mechanism satisfying zCDP, we show the contrapositive statement: any lower bound for pure DP can be converted into a lower bound for zCDP. Pure DP is characterised by so-called packing lower bounds and the exponential mechanism.

We begin by giving a technical lemma showing that for any output space and any desired accuracy we have a "packing" and a "net:"

Lemma 2.7.2. *Let $(\mathcal{Y}, d)$ be a metric space. Fix $\alpha > 0$. Then there exists a countable $T \subset \mathcal{Y}$ such that both of the following hold.*

- *(Net:) Either T is infinite or for all $y' \in \mathcal{Y}$ there exists $y \in T$ with $d(y, y') \leq \alpha$.*
- *(Packing:) For all $y, y' \in T$, if $y \neq y'$, then $d(y, y') > \alpha$.*

Proof. Consider the following procedure for producing T.

- Initialize $A \leftarrow \mathcal{Y}$ and $T \leftarrow \emptyset$.
- Repeat:
  - If $A = \emptyset$, terminate.
  - Pick some $y \in A$.
  - Update $T \leftarrow T \cup \{y\}$.
  - Update $A \leftarrow \{y' \in A : d(y', y) > \alpha\}$.

This procedure either terminates giving a finite T or runs forever enumerating a countably infinite T.

(Net:) If T is infinite, we immediately can dispense the first condition, so suppose the procedure terminates and T is finite. Fix $y' \in \mathcal{Y}$. Since the procedure terminates, $A = \emptyset$ at the end, which means y' was removed from A at some point. This means some $y \in T$ was added such that $d(y', y) \leq \alpha$, as required.

(Packing:) Fix $y \neq y' \in T$. We assume, without loss of generality, that y was added to T before y'. This means y' was not removed from A when y was added to T. In particular, this means $d(y', y) > \alpha$. □

It is well-known that a net yields a pure DP algorithm:

Lemma 2.7.3 (Exponential Mechanism [MT07, BLR13]). *Let $\ell : \mathcal{X}^n \times T \to \mathbb{R}$ satisfy $|\ell(x, y) - \ell(x', y)| \leq \Delta$ for all $x, x' \in \mathcal{X}^n$ differing in one entry and all $y \in T$. Then, for all $\varepsilon > 0$, there exists an ε-differentially private $M : \mathcal{X}^n \to T$ such that*

$$\mathop{\mathbb{P}}_{M}\left[\ell(x, M(x)) \leq \min_{y \in T} \ell(x, y) + \frac{2\Delta}{\varepsilon} \log\left(\frac{|T|}{\beta}\right)\right] \geq 1 - \beta$$

and

$$\mathop{\mathbb{E}}_{M}\left[\ell(x, M(x))\right] \leq \min_{y \in T} \ell(x, y) + \frac{2\Delta}{\varepsilon} \log |T|$$

for all $x \in \mathcal{X}^n$ and $\beta > 0$.

Proof. The mechanism is defined by

$$\mathop{\mathbb{P}}_{M}[M(x) = y] = \frac{e^{-\ell(x,y)\varepsilon/2\Delta}}{\sum_{y' \in T} e^{-\ell(x,y)\varepsilon/2\Delta}}.$$

The analysis can be found in [DR14, Theorems 3.10 and 3.11] and Lemma 3.7.1. □

We also show that a packing yields a lower bound for zCDP:

Lemma 2.7.4. *Let $(\mathcal{Y}, d)$ be a metric space and $q : \mathcal{X}^n \to \mathcal{Y}$ a function. Let $M : \mathcal{X}^n \to \mathcal{Y}$ be a (ξ, ρ)-zCDP mechanism satisfying*

$$\mathop{\mathbb{P}}_{M}[d(M(x), q(x)) > \alpha/2] \leq \beta$$

for all $x \in \mathcal{X}^n$. Let $T \subset \mathcal{Y}$ be such that $d(y, y') > \alpha$, for all $y, y' \in T$ with $y \neq y'$. Assume that for all $y \in T$ there exists $x \in \mathcal{X}^n$ with $q(x) = y$. Then

$$(1 - \beta) \log |T| - \log 2 \leq \xi \cdot n(1 + \log n) + \rho \cdot n^2.$$

In particular, if $\xi = 0$, we have

$$n \geq \sqrt{\frac{(1 - \beta) \log |T| - \log 2}{\rho}} = \Omega(\sqrt{\log |T|/\rho}).$$

Proof. Let $q^{-1} : T \to \mathcal{X}^n$ be a function such that $q(q^{-1}(y)) = y$ for all $y \in T$. Define $f : \mathcal{Y} \to T$ by

$$f(y) = \operatorname*{argmin}_{y' \in T} d(y, y')$$

(breaking ties arbitrarily). Then

$$\mathop{\mathbb{P}}_{M}\left[f(M(q^{-1}(y))) = y\right] \geq 1 - \beta$$

for all $y \in T$, as $\mathop{\mathbb{P}}_{M}\left[d(M(q^{-1}(y)), y) > \alpha/2\right] \leq \beta$ and $d(y', y) > \alpha$ for all $y' \in T \setminus \{y\}$.

Let Y be a uniformly random element of T and let $X = q^{-1}(Y)$. By the data

processing inequality and Proposition 2.6.1,

$$I(Y; f(M(q^{-1}(Y)))) = I(q(X); f(M(X))) \leq I(X; M(X)) \leq \xi \cdot n(1 + \log n) + \rho \cdot n^2.$$

However, $\mathbb{P}\left[f(M(q^{-1}(Y))) = Y\right] \geq 1 - \beta$. Denote $Z = f(M(q^{-1}(Y)))$ and let E be the indicator of the event that $Z = Y$. We have

$$I(Y; Z) = H(Y) - H(Y|Z) = H(Y) - H(Y, E|Z) = H(Y) - H(Y|E, Z) - H(E|Z).$$

Clearly $H(Y) = \log |T|$ and $H(E|Z) \leq H(E) \leq \log 2$. Moreover,

$$\begin{aligned} H(Y|E, Z) = &\mathop{\mathbb{E}}_{e \leftarrow E}\left[H(Y|Z, E = e)\right] \\ =&\mathbb{P}\left[Y = Z\right] \cdot 0 + \mathbb{P}\left[Y \neq Z\right] \cdot H(Y|Z, Y \neq Z) \\ \leq&\beta \cdot H(Y). \end{aligned}$$

Thus

$$I(Y; Z) \geq \log |T| - \log 2 - \beta \log |T|.$$

The result now follows by combining inequalities. □

We need one final technical lemma:

Lemma 2.7.5. *Let $q : \mathcal{X} \to \mathbb{R}^k$ satisfy $\max_{x \in \mathcal{X}} \|q(x)\| \leq 1$, where $\|\cdot\|$ is some norm. Let $M : \mathcal{X}^n \to \mathbb{R}^k$ satisfy (ξ, ρ)-zCDP and*

$$\mathop{\mathbb{E}}_{M}\left[\|M(x) - q(x)\|\right] \leq \alpha$$

for all $x \in \mathcal{X}^n$. For all n',

$$\max_{x \in \mathcal{X}^{n'}} \min_{\hat{x} \in \mathcal{X}^n} \|q(\hat{x}) - q(x)\| \leq 2\alpha + \sqrt{2(\xi + \rho)}.$$

Before proving Lemma 2.7.5, we complete the proof of Theorem 2.7.1 by combining Lemmas 2.7.2, 2.7.3, 2.7.4, and 2.7.5

Proof of Theorem 2.7.1. Apply Lemma 2.7.2 with $\mathcal{Y} = \left\{q(x) = \frac{1}{n}\sum_{i\in[n]} q(x_i) : x \in \mathcal{X}^n\right\} \subset \mathbb{R}^k$ and d being the metric induced by the norm to obtain $T \subset \mathcal{Y}$:

- (Net:) Either T is infinite or for all $y' \in \{q(x) : x \in \mathcal{X}^n\} \subset \mathbb{R}^k$ there exists $y \in T$ with $\|y - y'\| \leq 4\alpha$.
- (Packing:) For all $y, y' \in T$, if $y \neq y'$, then $\|y - y'\| > 4\alpha$.

By Markov's inequality

$$\mathop{\mathbb{P}}_{M}[\|M(x) - q(x)\| > 2\alpha] \leq \frac{1}{2}.$$

Thus, by Lemma 2.7.4,

$$\frac{1}{2}\log|T| - \log 2 \leq \xi \cdot n(1 + \log n) + \rho \cdot n^2.$$

This gives an upper bound on $|T|$. In particular, T must be finite.

Let $M' : \mathcal{X}^{n'} \to \mathbb{R}^k$ be the exponential mechanism (Lemma 2.7.3) instantiated with T and $\ell(x, y) = \|y - q(x)\|$. We have

$$\mathop{\mathbb{P}}_{M}\left[\|M(x) - q(x)\| \leq \min_{y\in T}\|y - q(x)\| + \frac{4}{\varepsilon n'}\log\left(\frac{|T|}{\beta}\right)\right] \geq 1 - \beta$$

and

$$\mathop{\mathbb{E}}_{M}[\|M(x) - q(x)\|] \leq \min_{y\in T}\|y - q(x)\| + \frac{4}{\varepsilon n'}\log|T|$$

for all $x \in \mathcal{X}^{n'}$. For $x \in \mathcal{X}^{n'}$, by the Net property and Lemma 2.7.5,

$$\begin{aligned}
\min_{y \in T} \|y - q(x)\| \leq & \min_{y \in T} \min_{y' \in \mathcal{Y}} \|y - y'\| + \|y' - q(x)\| \\
= & \min_{y' \in \mathcal{Y}} \left(\left(\min_{y \in T} \|y - y'\| \right) + \|y' - q(x)\| \right) \\
\leq & \min_{y' \in \mathcal{Y}} \left(4\alpha + \|y' - q(x)\| \right) \\
= & \min_{\hat{x} \in \mathcal{X}^n} \left(4\alpha + \|q(\hat{x}) - q(x)\| \right) \\
\leq & 4\alpha + 2\alpha + \sqrt{2(\xi + \rho)}.
\end{aligned}$$

Furthermore,

$$\frac{4}{\varepsilon n'} \log |T| \leq \frac{8}{\varepsilon n'} \left(\xi \cdot n(1 + \log n) + \rho \cdot n^2 + \log 2 \right) \leq 2\alpha.$$

The theorem now follows by combining inequalities. □

Finally we prove the technical Lemma 2.7.5. Essentially we show that if a private mechanism can accurately answer a set of queries with a given sample complexity, then those queries can be approximated on an unknown distribution with the same sample complexity. This is related to lower bounds on private sample complexity using Vapnik-Chervonenkis dimension e.g. [DR14, Theorem 4.8] and [BNSV15, Theorem 5.5].

First we state Pinsker's inequality [vEH14, Theorem 31].

Lemma 2.7.6 (Pinsker's Inequality)**.** *Let X and Y be random variables on $[-1, 1]$. Then*

$$\left| \mathbb{E}[X] - \mathbb{E}[Y] \right| \leq \sqrt{2 \mathrm{D}_1(X \| Y)}$$

We can also generalise Pinsker's inequality using Rényi divergence:

Lemma 2.7.7. *Let P and Q be distributions on Ω and $f : \Omega \to \mathbb{R}$. Then*

$$\left| \mathop{\mathbb{E}}_{x\sim P}[f(x)] - \mathop{\mathbb{E}}_{x\sim Q}[f(x)] \right| \leq \sqrt{\mathop{\mathbb{E}}_{x\sim Q}[f(x)^2]} \cdot \sqrt{e^{\mathrm{D}_2(P\|Q)} - 1}.$$

In particular, if $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies (ξ, ρ)-zCDP. Then, for any $f : \mathcal{Y} \to \mathbb{R}$ and all neighbouring $x, x' \in \mathcal{X}^n$,

$$\left| \mathbb{E}\left[f(M(x'))\right] - \mathbb{E}\left[f(M(x))\right] \right| \leq \sqrt{\mathbb{E}\left[f(M(x))^2\right]} \cdot \sqrt{e^{\xi+2\rho} - 1}.$$

Proof. By Cauchy-Schwartz,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim P}[f(x)] - \mathop{\mathbb{E}}_{x\sim Q}[f(x)] &= \mathop{\mathbb{E}}_{x\sim Q}\left[f(x)\left(\frac{P(x)}{Q(x)} - 1\right)\right] \\
&\leq \sqrt{\mathop{\mathbb{E}}_{x\sim Q}[f(x)^2]} \cdot \sqrt{\mathop{\mathbb{E}}_{x\sim Q}\left[\left(\frac{P(x)}{Q(x)} - 1\right)^2\right]}.
\end{aligned}$$

Now

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim Q}\left[\left(\frac{P(x)}{Q(x)} - 1\right)^2\right] &= \mathop{\mathbb{E}}_{x\sim Q}\left[\left(\frac{P(x)}{Q(x)}\right)^2 - 2\frac{P(x)}{Q(x)} + 1\right] \\
&= \mathop{\mathbb{E}}_{x\sim Q}\left[\left(\frac{P(x)}{Q(x)}\right)^2\right] - 2 + 1 \\
&= e^{\mathrm{D}_2(P\|Q)} - 1.
\end{aligned}$$

□

Proposition 2.7.8. *Let $q : \mathcal{X} \to \mathbb{R}^k$ satisfy $\max_{x\in\mathcal{X}} \|q(x)\| \leq 1$, where $\|\cdot\|$ is some norm. Let $M : \mathcal{X}^n \to \mathbb{R}^k$ satisfy (ξ, ρ)-zCDP and*

$$\mathop{\mathbb{E}}_{M}[\|M(x) - q(x)\|] \leq \alpha$$

for all $x \in \mathcal{X}^n$. Then, for any distribution $\mathcal{D}$ on $\mathcal{X}$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}^n, M}[\|M(x) - q(\mathcal{D})\|] \leq \alpha + \sqrt{2(\xi + \rho)}$$

and

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}^n}[\|q(x)-q(\mathcal{D})\|]\le 2\alpha+\sqrt{2(\xi+\rho)},$$

where $q(\mathcal{D})=\mathop{\mathbb{E}}_{z\sim\mathcal{D}}[q(z)]$.

Proof. First define the dual norm: For $x\in\mathbb{R}^k$,

$$\|x\|_*:=\max_{y\in\mathbb{R}^k:\|y\|=1}\langle x,y\rangle.$$

By definition, $\langle x,y\rangle=\langle y,x\rangle\le\|x\|_*\cdot\|y\|$ for all $x,y\in\mathbb{R}^k$. Moreover, $\|z\|=\max_{y\in\mathbb{R}^k:\|y\|_*=1}\langle z,y\rangle$ for all $z\in\mathbb{R}^k$.

Fix a distribution $\mathcal{D}$. Define $W:\mathcal{X}^n\to\mathbb{R}^k\times\mathbb{R}^k$ as follows. On input $x\in\mathcal{X}^n$, compute $a=M(x)$ and

$$s=\mathop{\mathrm{argmax}}_{v\in\mathbb{R}^k:\|v\|_*=1}\langle a-q(\mathcal{D}),v\rangle,$$

and output (a,s).

By postprocessing, W satisfies (ξ,ρ)-zCDP and

$$\mathop{\mathbb{E}}_{W}[\langle a-q(x),s\rangle|(a,s)=W(x)]\le\mathop{\mathbb{E}}_{W}[\|a-q(x)\|\cdot\|s\|_*|(a,s)=W(x)]=\mathop{\mathbb{E}}_{M}[\|M(x)-q(x)\|]\le\alpha \tag{2.11}$$

for all $x\in\mathcal{X}^n$.

The following is similar to Lemma 3.3.1. Let $x\sim\mathcal{D}^n$ and $y\sim\mathcal{D}$. Now

$$\begin{aligned}\mathop{\mathbb{E}}_{x,W}[\langle q(x),s\rangle\mid(a,s)=W(x)]&=\frac{1}{n}\sum_{i\in[n]}\mathop{\mathbb{E}}_{x,W}[\langle q(x_i),s\rangle\mid(a,s)=W(x)]\\&=\frac{1}{n}\sum_{i\in[n]}\mathop{\mathbb{E}}_{x,W}[f(x_i,W(x))]\end{aligned} \tag{2.12}$$

(letting $f : \mathcal{X} \times \mathbb{R}^k \times \mathbb{R}^k \to [-1, 1]$ be $f(z, a, s) = \langle q(z), s \rangle / \|s\|_*$)

$$\leq \frac{1}{n} \sum_{i \in [n]} \mathop{\mathbb{E}}_{x,y,W} [f(x_i, W(x_1, \cdots, x_{i-1}, y, x_{i+1}, \cdots, x_n))]$$
$$+ \sqrt{2\mathrm{D}_1 \left(f(x_i, W(x)) \| f(x_i, W(x_1, \cdots, x_{i-1}, y, x_{i+1}, \cdots, x_n))\right)}$$

(by Pinsker's inequality)

$$\leq \frac{1}{n} \sum_{i \in [n]} \mathop{\mathbb{E}}_{x,y,W} [f(y, W(x_1, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots, x_n))]$$
$$+ \sqrt{2\mathrm{D}_1 \left(W(x) \| W(x_1, \cdots, x_{i-1}, y, x_{i+1}, \cdots, x_n)\right)}$$

(by postprocessing and convexity and the fact that x_i and y are interchangeable)

$$\leq \frac{1}{n} \sum_{i \in [n]} \mathop{\mathbb{E}}_{x,y,W} [\langle q(y), s \rangle \mid (j, s, a) = W(x)]$$
$$+ \sqrt{2(\xi + \rho)}$$

(by zCDP)

$$= \mathop{\mathbb{E}}_{x,W} [\langle q(\mathcal{D}), s \rangle \mid (a, s) = W(x)] + \sqrt{2(\xi + \rho)}.$$

Combining (2.11) and (2.12) gives

$$\mathop{\mathbb{E}}_{x,M} [\|M(x) - q(\mathcal{D})\|] = \mathop{\mathbb{E}}_{x,W} [\langle a - q(\mathcal{D}), s \rangle \mid (a, s) = W(x)] \leq \alpha + \sqrt{2(\xi + \rho)}. \quad (2.13)$$

Finally, combining (2.13) and (2.11) gives

$$\mathop{\mathbb{E}}_{x}\left[\|q(x)-q(\mathcal{D})\|\right] \leq \mathop{\mathbb{E}}_{x,M}\left[\|M(x)-q(x)\|\right] + \mathop{\mathbb{E}}_{x,M}\left[\|M(x)-q(\mathcal{D})\|\right] \leq 2\alpha + \sqrt{2(\xi+\rho)}.$$

□

Proof of Lemma 2.7.5. Fix $x \in \mathcal{X}^{n'}$. Let $\mathcal{D}$ be the uniform distribution on elements of x so that $q(\mathcal{D}) = q(x)$. By Proposition 2.7.8,

$$\mathop{\mathbb{E}}_{\hat{x}\sim\mathcal{D}^n}\left[\|q(\hat{x})-q(\mathcal{D})\|\right] \leq 2\alpha + \sqrt{2(\xi+\rho)}.$$

In particular, there must exist $\hat{x} \sim \mathcal{D}^n$ such that $\|q(\hat{x})-q(\mathcal{D})\| \leq 2\alpha + \sqrt{2(\xi+\rho)}$, as required. □

2.8 Approximate zCDP

In the spirit of approximate DP, we propose a relaxation of zCDP:

Definition 2.8.1 (Approximate zCDP)**.** *A randomised mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ is δ-approximately (ξ,ρ)-zCDP if, for all $x, x' \in \mathcal{X}^n$ differing on a single entry, there exist events $E = E(M(x))$ and $E' = E'(M(x'))$ such that, for all $\alpha \in (1,\infty)$,*

$$\mathrm{D}_\alpha\left(M(x)|_E \middle\| M(x')|_{E'}\right) \leq \xi + \rho\cdot\alpha \quad \text{and} \quad \mathrm{D}_\alpha\left(M(x')|_{E'} \middle\| M(x)|_E\right) \leq \xi + \rho\cdot\alpha$$

and $\mathop{\mathbb{P}}_{M(x)}[E] \geq 1-\delta$ *and* $\mathop{\mathbb{P}}_{M(x')}[E'] \geq 1-\delta$.

Clearly 0-approximate zCDP is simply zCDP. Hence we have a generalisation of zCDP. As we will show later in this section, δ-approximate $(\varepsilon,0)$-zCDP is equivalent to (ε,δ)-DP. Thus we have also generalised approximate DP. Hence, this definition unifies both relaxations of pure DP.

Approximate zCDP is a three-parameter definition which allows us to capture many different aspects of differential privacy. However, three parameters is quite overwhelming. We believe that use of the one-parameter ρ-zCDP (or the two-parameter δ-approximate ρ-zCDP if necessary) is sufficient for most purposes.

It is easy to verify that the definition of approximate zCDP satisfies the following basic properties.

Lemma 2.8.2 (Composition & Postprocessing)**.** *Let $M : \mathcal{X}^n \to \mathcal{Y}$ and $M' : \mathcal{X}^n \times \mathcal{Y} \to \mathcal{Z}$ be randomised algorithms. Suppose M satisfies δ-approximate (ξ, ρ)-zCDP and, for all $y \in \mathcal{Y}$, $M'(\cdot, y) : \mathcal{X}^n \to \mathcal{Z}$ satisfies δ'-approximate (ξ', ρ')-zCDP. Define $M'' : \mathcal{X}^n \to \mathcal{Z}$ by $M''(x) = M'(x, M(x))$. Then M'' satisfies $(\delta + \delta' - \delta \cdot \delta')$-approximate $(\xi + \xi', \rho + \rho')$-zCDP.*

Lemma 2.8.3 (Tradeoff)**.** *Suppose $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies δ-approximate $(\xi, 0)$-zCDP. Then M satisfies δ-approximate ξ-zCDP and δ-approximate $\frac{1}{2}\xi^2$-zCDP.*

However, the strong group privacy guarantees of Section 2.5 no longer apply to approximate zCDP and, hence, the strong lower bounds of Section 2.6 also no longer hold. Circumventing these lower bounds is part of the motivation for considering approximate zCDP. However, approximate zCDP is not necessarily the only way to relax zCDP that circumvents our lower bounds:

Proving the group privacy bound requires "inflating" the parameter α: Suppose $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies ρ-zCDP and $x, x' \in \mathcal{X}^n$ differ on k entries. To prove $\mathrm{D}_\alpha\left(M(x) \| M(x')\right) \leq k^2\rho\alpha$, the proof of Proposition 2.5.3 requires a bound on $\mathrm{D}_{k\alpha}\left(M(x'') \| M(x''')\right)$ for $x'', x''' \in \mathcal{X}^n$ differing on a single entry.

Consider relaxing the definition of zCDP to only require the bound (2.1) or (2.2) to hold when $\alpha \leq m$:

Definition 2.8.4 (Bounded zCDP). *We say that $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies m-bounded (ξ, ρ)-zCDP if, for all $x, x' \in \mathcal{X}^n$ differing in only one entry and all $\alpha \in (1, m)$, $\mathrm{D}_\alpha\left(M(x) \| M(x')\right) \leq \xi + \rho \cdot \alpha$.*

This relaxed definition may also be able to circumvent the group privacy-based lower bounds, as our group privacy proof would no longer work for groups of size larger than m. We do not know what group privacy guarantees Definition 2.8.4 provides for groups of size $k \gg m$. This relaxed definition may be worth exploring, but is beyond the scope of our work.

2.8.1 Approximate DP Implies Approximate zCDP

We can convert approximate DP to approximate zCDP using the following lemma.

First we define a approximate DP version of the randomised response mechanism:

Definition 2.8.5. *For $\varepsilon \geq 0$ and $\delta \in [0, 1]$, define $\tilde{M}_{\varepsilon,\delta} : \{0,1\} \to \{0,1\} \times \{\perp, \top\}$ by*

$$\begin{aligned}
\mathbb{P}\left[\tilde{M}_{\varepsilon,\delta}(b) = (b, \top)\right] &= \delta, & \mathbb{P}\left[\tilde{M}_{\varepsilon,\delta}(b) = (1-b, \top)\right] &= 0, \\
\mathbb{P}\left[\tilde{M}_{\varepsilon,\delta}(b) = (b, \perp)\right] &= (1-\delta)\frac{e^\varepsilon}{1+e^\varepsilon}, & \mathbb{P}\left[\tilde{M}_{\varepsilon,\delta}(b) = (1-b, \perp)\right] &= (1-\delta)\frac{1}{1+e^\varepsilon}
\end{aligned}$$

for both $b \in \{0,1\}$.

The above mechanism is "complete" for approximate DP:

Lemma 2.8.6 ([KOV15], [MV16, Lemma 3.2]). *For every (ε, δ)-DP $M : \mathcal{X}^n \to \mathcal{Y}$ and all $x_0, x_1 \in \mathcal{X}^n$ differing in one entry, there exists a randomised $T : \{0,1\} \times \{\perp, \top\} \to \mathcal{Y}$ such that $T(\tilde{M}_{\varepsilon,\delta}(b))$ has the same distribution as $M(x_b)$ for both $b \in \{0,1\}$.*

Corollary 2.8.7. *If $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies (ε, δ)-DP, then M satisfies δ-approximate $(\varepsilon, 0)$-zCDP, which, in turn, implies δ-approximate $(0, \frac{1}{2}\varepsilon^2)$-zCDP.*

Proof. Fix neighbouring $x_0, x_1 \in \mathcal{X}^n$. Let $T : \{0,1\} \times \{\perp, \top\} \to \mathcal{Y}$ be as in Lemma 2.8.6.

Now we can write $M(x_b) = T(\tilde{M}_{\varepsilon,\delta}(b))$ for $b \in \{0,1\}$. Define events E_0 and E_1 by

$$E_b \equiv \left[\tilde{M}_{\varepsilon,\delta}(b) \in \{0,1\} \times \{\perp\}\right].$$

By definition, for both $b \in \{0,1\}$, $\underset{\tilde{M}_{\varepsilon,\delta}(b)}{\mathbb{P}}[E_b] = 1 - \delta$ and

$$M(x_b)|_{E_b} = T\left(\tilde{M}_{\varepsilon,\delta}(b)|_{\tilde{M}_{\varepsilon,\delta}(b) \in \{0,1\} \times \{\perp\}}\right) = T(\tilde{M}_{\varepsilon,0}(b)).$$

We have $\mathrm{D}_\infty\left(\tilde{M}_{\varepsilon,0}(b) \middle\| \tilde{M}_{\varepsilon,0}(1-b)\right) \leq \varepsilon$ for both $b \in \{0,1\}$. By postprocessing and monotonicity, this implies

$$\mathrm{D}_\alpha\left(M(x_b)|_{E_b} \middle\| M(x_{1-b})|_{E_{1-b}}\right) \leq \varepsilon$$

for both $b \in \{0,1\}$ and all $\alpha \in (1,\infty)$. Thus we have satisfied the definition of δ-approximate $(\varepsilon, 0)$-zCDP.

Applying Proposition 2.3.2 shows that this also implies δ-approximate $(0, \frac{1}{2}\varepsilon^2)$-zCDP. □

2.8.2 Approximate zCDP Implies Approximate DP

Lemma 2.8.8. *Suppose $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies δ-approximate (ξ, ρ)-zCDP. If $\rho = 0$, then M satisfies (ξ, δ)-DP. In general, M satisfies $(\varepsilon, \delta + (1-\delta)\delta')$-DP for all $\varepsilon \geq \xi + \rho$, where*

$$\delta' = e^{-(\varepsilon - \xi - \rho)^2/4\rho} \cdot \min\left\{\begin{array}{l} 1 \\ \sqrt{\pi \cdot \rho} \\ \frac{1}{1 + (\varepsilon - \xi - \rho)/2\rho} \\ \frac{2}{1 + \frac{\varepsilon - \xi - \rho}{2\rho} + \sqrt{\left(1 + \frac{\varepsilon - \xi - \rho}{2\rho}\right)^2 + \frac{4}{\pi\rho}}} \end{array}\right. .$$

Proof. Fix neighbouring $x, x' \in \mathcal{X}^n$ and let E and E' be the events promised by definition 2.8.1. We can assume, without loss of generality that $\mathbb{P}[E] = \mathbb{P}[E'] = 1 - \delta$.

Fix $S \subset \mathcal{Y}$. Then

$$\begin{aligned}
\mathbb{P}[M(x) \in S] =& \mathbb{P}[M(x) \in S \mid E] \cdot \mathbb{P}[E] + \mathbb{P}[M(x) \in S \mid \neg E] \cdot \mathbb{P}[\neg E] \\
\leq& \mathbb{P}[M(x) \in S \mid E] \cdot (1 - \delta) + \delta, \\
\mathbb{P}\left[M(x') \in S\right] =& \mathbb{P}\left[M(x') \in S \mid E'\right] \cdot \mathbb{P}\left[E'\right] + \mathbb{P}\left[M(x') \in S \mid \neg E'\right] \cdot \mathbb{P}\left[\neg E'\right] \\
\geq& \mathbb{P}\left[M(x') \in S \mid E'\right] \cdot (1 - \delta).
\end{aligned}$$

Firstly, if $\rho = 0$, then

$$\mathbb{P}[M(x) \in S \mid E] \leq e^{\xi} \mathbb{P}\left[M(x') \in S \mid E'\right]$$

and

$$\begin{aligned}
\mathbb{P}[M(x) \in S] \leq& \mathbb{P}[M(x) \in S \mid E] \cdot (1 - \delta) + \delta \\
\leq& e^{\xi} \mathbb{P}\left[M(x') \in S \mid E'\right] \cdot (1 - \delta) + \delta \leq e^{\xi} \mathbb{P}\left[M(x') \in S\right] + \delta,
\end{aligned}$$

which proves the first half of the lemma.

Secondly, by Lemma 2.3.8 (cf. Lemma 2.3.6), for all $\varepsilon \geq \xi + \rho$,

$$\mathbb{P}[M(x) \in S \mid E] \leq e^{\varepsilon} \mathbb{P}\left[M(x') \in S \mid E'\right] + e^{-(\varepsilon - \xi - \rho)^2/4\rho} \cdot \min \begin{cases} 1 \\ \sqrt{\pi \cdot \rho} \\ \frac{1}{1 + (\varepsilon - \xi - \rho)/2\rho} \\ \frac{2}{1 + \frac{\varepsilon - \xi - \rho}{2\rho} + \sqrt{\left(1 + \frac{\varepsilon - \xi - \rho}{2\rho}\right)^2 + \frac{4}{\pi\rho}}} \end{cases}.$$

Thus

$$\mathbb{P}\left[M(x)\in S\right] \leq e^{\varepsilon}\mathbb{P}\left[M(x')\in S\right] + \delta + (1-\delta)\cdot e^{-(\varepsilon-\xi-\rho)^2/4\rho}\cdot \min\begin{cases} 1 \\ \sqrt{\pi\cdot\rho} \\ \frac{1}{1+(\varepsilon-\xi-\rho)/2\rho} \\ \frac{2}{1+\frac{\varepsilon-\xi-\rho}{2\rho}+\sqrt{\left(1+\frac{\varepsilon-\xi-\rho}{2\rho}\right)^2+\frac{4}{\pi\rho}}} \end{cases}.$$

□

2.8.3 Application of Approximate zCDP

Approximate zCDP subsumes approximate DP. A result of this is that we can apply our tightened lemmas to give a tighter version of the so-called advanced composition theorem [DRV10].

Note that the following results are subsumed by the bounds of Kairouz, Oh, and Viswanath [KOV15] and Murtagh and Vadhan [MV16]. However, these bounds may be extended to analyse the composition of mechanisms satisfying CDP with mechanisms satisfying approximate DP. We believe that such a "unified" analysis of composition will be useful.

Applying Corollary 2.8.7, Lemma 2.8.2, and Lemma 2.8.8 yields the following result.

Corollary 2.8.9. *Let $M_1, \cdots, M_k : \mathcal{X}^n \to \mathcal{Y}$ and let $M : \mathcal{X}^n \to \mathcal{Y}^k$ be their composition. Suppose each M_i satisfies $(\varepsilon_i, \delta_i)$-DP. Set $\rho = \frac{1}{2}\sum_i^k \varepsilon_i^2$. Then M satisfies*

$$\left(\varepsilon, 1-(1-\delta')\prod_i^k(1-\delta_i)\right)\text{-DP}$$

for all $\varepsilon \geq \rho$ *and*

$$\delta' = e^{-(\varepsilon-\rho)^2/4\rho} \cdot \min \begin{cases} 1 \\ \sqrt{\pi \cdot \rho} \\ \frac{1}{1+(\varepsilon-\rho)/2\rho} \\ \frac{2}{1+\frac{\varepsilon-\rho}{2\rho}+\sqrt{\left(1+\frac{\varepsilon-\rho}{2\rho}\right)^2+\frac{4}{\pi\rho}}} \end{cases}.$$

A slight restatement is the following

Corollary 2.8.10. *Let* $M_1, \cdots, M_k : \mathcal{X}^n \to \mathcal{Y}$ *and let* $M : \mathcal{X}^n \to \mathcal{Y}^k$ *be their composition. Suppose each* M_i *satisfies* $(\varepsilon_i, \delta_i)$*-DP. Set* $\varepsilon^2 = \frac{1}{2}\sum_i^k \varepsilon_i^2$*. Then* M *satisfies*

$$\left(\varepsilon^2 + 2\lambda\varepsilon, 1 - (1-\delta')\prod_i^k (1-\delta_i)\right) \text{-DP}$$

for all $\lambda \geq 0$ *and*

$$\delta' = e^{-\lambda^2} \cdot \min \begin{cases} 1 \\ \sqrt{\pi} \cdot \varepsilon \\ \frac{1}{1+\lambda/\varepsilon} \\ \frac{2}{1+\frac{\lambda}{\varepsilon}+\sqrt{\left(1+\frac{\lambda}{\varepsilon}\right)^2+\frac{4}{\pi\varepsilon^2}}} \end{cases}.$$

Finally, by picking the second term in the minimum and using $1 - \prod_i (1-\delta_i) \leq \sum_i \delta_i$, we have the following simpler form of the lemma.

Corollary 2.8.11. *Let* $M_1, \cdots, M_k : \mathcal{X}^n \to \mathcal{Y}$ *and let* $M : \mathcal{X}^n \to \mathcal{Y}^k$ *be their composition. Suppose each* M_i *satisfies* $(\varepsilon_i, \delta_i)$*-DP. Then* M *satisfies*

$$\left(\frac{1}{2}\|\varepsilon\|_2^2 + \sqrt{2}\lambda\|\varepsilon\|_2, \sqrt{\frac{\pi}{2}} \cdot \|\varepsilon\|_2 \cdot e^{-\lambda^2} + \|\delta\|_1\right) \text{-DP}$$

for all $\lambda \geq 0$*. Alternatively* M *satisfies*

$$\left(\frac{1}{2}\|\varepsilon\|_2^2 + \sqrt{2\log(\sqrt{\pi/2} \cdot \|\varepsilon\|_2/\delta')} \cdot \|\varepsilon\|_2, \delta' + \|\delta\|_1\right) \text{-DP}$$

for all $\delta' \geq 0$.

In comparison to the composition theorem of [DRV10], we save modestly by a constant factor in the first term and, in most cases $\sqrt{\pi/2}\|\varepsilon\|_2 < 1$, whence the logarithmic term is an improvement over the usual advanced composition theorem.

Chapter 3

Adaptive Data Analysis

3.1 Introduction

Multiple hypothesis testing is a ubiquitous task in empirical research. A finite sample of data is drawn from some unknown population, and several analyses are performed on that sample. The outcome of an analysis is deemed significant if it is unlikely to have occurred by chance alone, and a "false discovery" occurs if the analyst incorrectly declares an outcome to be significant. False discovery has been identified as a substantial problem in the scientific community (see e.g. [Ioa05, GL14]). This problem persists despite decades of research by statisticians on methods for preventing false discovery, such as the widely used Bonferroni Correction [Bon36, Dun61] and the Benjamini-Hochberg Procedure [BH95].

False discovery is often attributed to misuse of statistics. An alternative explanation is that the prevalence of false discovery arises from the inherent *adaptivity* in the data analysis process—the fact that the choice of analyses to perform depends on previous interactions with the data (see e.g. [GL14]). Adaptivity is essentially unavoidable when a sequence of research groups publish research papers based on

overlapping data sets. Adaptivity also arises naturally in other settings, for example: in multistage inference algorithms where data are preprocessed (say, to select features or restrict to a principal subspace) before the main analysis is performed; in scoring data-based competitions [BH15]; and in the re-use of holdout or test data [DFH+15b, DFH+15a].

The general problem of adaptive data analysis was formally modeled and studied in recent papers by Dwork, Feldman, Hardt, Pitassi, Reingold, and Roth [DFH+15c] and by Hardt and Ullman [HU14]. The striking results of Dwork et al. [DFH+15c] gave the first nontrivial algorithms for provably ensuring statistical validity in adaptive data analysis, allowing for even an *exponential* number of tests against the same sample. In contrast, Chapter 6 and Hardt and Ullman [HU14] show inherent statistical and computational barriers to preventing false discovery in adaptive settings.

The key ingredient in Dwork et al. is a notion of "algorithmic stability" that is suitable for adaptive analysis. Informally, changing one input to a stable algorithm does not change it's output "too much." Traditionally, stability was measured via the change in the generalisation error of an algorithm's output, and algorithms stable according to such a criterion have long been known to ensure statistical validity in nonadaptive analysis [DW79a, DW79b, KR99, BE02, SSSS10]. Following a connection first suggested by McSherry,[1] Dwork et al. showed that differential privacy guarantees statistical validity in adaptive data analysis. This allowed them to repurpose known DP algorithms to prevent false discovery. A crucial difference from traditional notions of stability is that DP requires a change in one input lead to a small change in the *probability distribution* on the outputs. In this chapter, we refer

[1]See, e.g., [McS14], although the observation itself dates back at least to 2008 (personal communication).

to differential privacy as *DP stability*, to emphasise the relation to the literature on algorithmic stability and other notions of stability we study (KL- and TV-stability, in particular).

In this chapter, we extend the results of Dwork et al. along two axes. First, we give an *optimal* analysis of the statistical validity of DP-stable algorithms. As a consequence, we immediately obtain the best known bounds on the *sample complexity* (equivalently, the *convergence rate*) of adaptive data analysis. Second, we generalise the connection between DP stability and statistical validity to a much larger family of statistics. Our proofs are also significantly simpler than those of Dwork et al., and clarify the role of different stability notions in the adaptive setting.

3.1.1 Overview of Results

Adaptivity and Statistical Queries. Following the previous work on this subject [DFH+15c], we formalise the problem of adaptive data analysis as follows. There is a *distribution* $\mathcal{P}$ over some finite universe $\mathcal{X}$, and a *mechanism* $\mathcal{M}$ that does not know $\mathcal{P}$, but is given a set x consisting of n samples from $\mathcal{P}$. Using its sample, the mechanism must answer *queries* on $\mathcal{P}$. Here, a query q, coming from some family Q, maps a distribution $\mathcal{P}$ to a real-valued answer. The mechanism's answer a to a query q is *α-accurate* if $|a - q(\mathcal{P})| \leq \alpha$ with high probability. Importantly, the mechanism's goal is to provide answers that "generalise" to the underlying distribution, rather than answers that are specific to its sample.

We model adaptivity by allowing a *data analyst* to ask a sequence of queries $q_1, q_2, \ldots, q_k \in Q$ to the mechanism, which responds with answers $a_1, a_2, \ldots, a_k$. In the adaptive setting, the query q_j may depend on the previous queries and answers $q_1, a_1, \ldots, q_{j-1}, a_{j-1}$ arbitrarily. We say the mechanism is *α-accurate* given n samples for k adaptively chosen queries if, with high probability, when given a vector x of n

samples from an arbitrary distribution $\mathcal{P}$, the mechanism accurately responds to any adaptive analyst that makes at most k queries.

Dwork et al. [DFH+15c] considered the family of *statistical queries* [Kea93]. A statistical query q asks for the expected value of some function on random draws from the distribution. That is, the query is specified by a function $p : \mathcal{X} \to [0, 1]$ and its answer is $q(\mathcal{P}) = \mathbb{E}_{z \leftarrow_R \mathcal{P}}[p(z)]$.

The most natural way to answer a statistical query is to compute the *empirical answer* $\mathbb{E}_{z \leftarrow_R x}[p(z)]$, which is just the average value of the function on the given sample x.[2] It is simple to show that when k queries are specified *nonadaptively* (i.e. independent of previous answers), then the empirical answer is within $q(\mathcal{P}) \pm \alpha$ (henceforth, "α-accurate") with high probability so long as the sample has size $n \gtrsim \log(k)/\alpha^2$.[3] However, when the queries can be chosen adaptively, the empirical average performs much worse. In particular, there is an algorithm (based on [DN03]) that, after seeing the empirical answer to $k = O(\alpha^2 n)$ random queries, can find a query such that the empirical answer and the correct answer differ by α. Thus, the empirical average cannot be guaranteed to be accurate unless $n \gtrsim k/\alpha^2$, and so exponentially more samples are required to guarantee accuracy when the queries may be adaptive.

Answering Adaptive Statistical Queries. Surprisingly, Dwork et al. [DFH+15c], showed there are mechanisms that are much more effective than naïvely outputting the empirical answer. They show that empirically accurate differentially private

[2]For convenience, we will often use x as shorthand for the empirical distribution over x (viewed as a multiset). We use $z \leftarrow_R x$ to mean a random element chosen from the uniform distribution over the elements of x.

[3]This guarantee follows from bounding the error of each query using a Chernoff bound and then taking a union bound over all queries. The $\log k$ term corresponds to the Bonferroni correction in classical statistics.

mechanisms are accurate on the population and, by applying the Gausssian mechanism, they obtain a mechanisms that are accurate given only $n \gtrsim \sqrt{k}/\alpha^{2.5}$ samples, which is a significant improvement over the naïve mechanism when α is not too small. (See Table 3.1 for more detailed statements of their results, including results that achieve an *exponential* improvement in the sample complexity when $|\mathcal{X}|$ is bounded.)

Our first contribution is to give a simpler and quantitatively optimal analysis of the generalisation properties of stable algorithms, which immediately yields new accuracy bounds for adaptive statistical queries. In particular, we show that $n \gtrsim \sqrt{k}/\alpha^2$ samples suffice. Since $1/\alpha^2$ samples are required to answer a single nonadaptive query, our dependence on α is optimal. In Chapter 6, we will see that the $\sqrt{k}$ dependence is also optimal. (In particular, there is a lower bound of $n \gtrsim \sqrt{k}/\alpha + 1/\alpha^2$, but proving $n \gtrsim \sqrt{k}/\alpha^2$ remains open.)

Beyond Statistical Queries. Although statistical queries are surprisingly general [Kea93], we would like to be able to ask more general queries on the distribution $\mathcal{P}$ that capture a wider variety of machine learning and data mining tasks. To this end, we give the first bounds on the sample complexity required to answer large numbers of adaptively chosen *low-sensitivity queries* and *optimisation queries*, which we now describe.

Low-sensitivity queries are a generalisation of statistical queries. A query is specified by an arbitrary function $p : \mathcal{X}^n \to \mathbb{R}$ satisfying $|p(x) - p(x')| \leq 1/n$ for every $x, x' \in \mathcal{X}^n$ differing on exactly one element. The query applied to the population is defined to be $q(\mathcal{P}) = \mathbb{E}_{x \leftarrow_R \mathcal{P}^n}[p(x)]$. Examples include *distance queries* (e.g. "How far is the sample from being well-clustered?") and maxima of statistical queries (e.g. "What is the classification error of the best k-node decision tree?")

Optimisation queries are a broad generalisation of low-sensitivity queries to arbitrary output domains. The query is specified by a loss function $L : \mathcal{X}^n \times \Theta \to \mathbb{R}$ that is low-sensitivity in its first parameter, and the goal is to output $\theta \in \Theta$ that is "best" in the sense that it minimises the average loss. Specifically, $q(\mathcal{P}) = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}_{z \leftarrow_R \mathcal{P}^n}[L(z;\theta)]$. An important special case is when $\Theta \subseteq \mathbb{R}^d$ is convex and L is convex in θ, which captures many fundamental regression and classification problems.

Our sample complexity bounds are summarised in Table 3.1.

Query Type	Sample Complexity		Time per Query
	[DFH+15c]	This Work	
Statistical ($k \ll n^2$)	$\tilde{O}\left(\frac{\sqrt{k}}{\alpha^{2.5}}\right)$	$\tilde{O}\left(\frac{\sqrt{k}}{\alpha^2}\right)$	$\mathrm{poly}(n, \log \lvert\mathcal{X}\rvert)$
Statistical ($k \gg n^2$)	$\tilde{O}\left(\frac{\sqrt{\log \lvert\mathcal{X}\rvert} \cdot \log^{3/2} k}{\alpha^{3.5}}\right)$	$\tilde{O}\left(\frac{\sqrt{\log \lvert\mathcal{X}\rvert} \cdot \log k}{\alpha^3}\right)$	$\mathrm{poly}(n, \lvert\mathcal{X}\rvert)$
Low Sens. ($k \ll n^2$)	—	$\tilde{O}\left(\frac{\sqrt{k}}{\alpha^2}\right)$	$\mathrm{poly}(n, \log \lvert\mathcal{X}\rvert)$
Low Sens. ($k \gg n^2$)	—	$\tilde{O}\left(\frac{\log \lvert\mathcal{X}\rvert \cdot \log k}{\alpha^3}\right)$	$\mathrm{poly}(\lvert\mathcal{X}\rvert^n)$
Conv. Min. ($k \ll n^2$)	—	$\tilde{O}\left(\frac{\sqrt{dk}}{\alpha^2}\right)$	$\mathrm{poly}(n, d, \log \lvert\mathcal{X}\rvert)$
Conv. Min. ($k \gg n^2$)	—	$\tilde{O}\left(\frac{(\sqrt{d} + \log k) \cdot \sqrt{\log \lvert\mathcal{X}\rvert}}{\alpha^3}\right)$	$\mathrm{poly}(n, d, \lvert\mathcal{X}\rvert)$

Table 3.1: *Summary of Results. Here k = number of queries, n = number of samples, α = desired accuracy, $\mathcal{X}$ = universe of possible samples, d = dimension of parameter space Θ.*

3.1.2 Overview of Techniques

Our main result is a new proof, with optimal parameters, that a stable algorithm that provides answers to adaptive queries that are close to the empirical value on the sample gives answers that generalise to the underlying distribution. In particular,

we prove:

Theorem 3.1.1 (Main "Transfer Theorem")**.** *Let $\mathcal{M}$ be a mechanism that takes a sample $x \in \mathcal{X}^n$ and answers k adaptively-chosen low-sensitivity queries. Suppose that $\mathcal{M}$ satisfies the following:*

1. *For every sample x, $\mathcal{M}$'s answers are $(\alpha, \alpha\beta)$-accuratewith respect to the sample x — which we define to mean $\mathbb{P}\left[\max_{j \in k} \left|q_j(x) - a_j\right| \leq \alpha\right] \geq 1 - \alpha\beta$, where $q_1, \cdots, q_k : \mathcal{X}^n \to \mathbb{R}$ are the low-sensitivity queries that are asked and $a_1, \cdots, a_k \in \mathbb{R}$ are the answers given. The probability is taken only over $\mathcal{M}$'s random coins.*

2. *$\mathcal{M}$ satisfies $(\alpha, \alpha\beta)$-DP stability.*

Then, if x consists of n samples from an arbitrary distribution $\mathcal{P}$ over $\mathcal{X}$, $\mathcal{M}$'s answers are $(O(\alpha), O(\beta))$-accurate with respect to the population $\mathcal{P}$. That is,

$$\mathbb{P}\left[\max_{j \in k} \left|q_j(\mathcal{P}) - a_j\right| \leq O(\alpha)\right] \geq 1 - O(\beta),$$

where the probability is taken only over the choice of $x \leftarrow_R \mathcal{P}^n$ and $\mathcal{M}$'s random coins.

Our actual result is somewhat more general than Theorem 3.1.1. We show that the population-level error of a stable algorithm is close to its error on the sample, whether or not that error is low. Put glibly: stable algorithms cannot be wrong without realising it.

Compared to the results of [DFH+15c], Theorem 3.1.1 requires a quantitatively weaker stability guarantee—$(\alpha, \alpha\beta)$-stability, instead of $(\alpha, (\beta/k)^{1/\alpha})$-stability. It also applies to arbitrary low-sensitivity queries as opposed to the special case of statistical queries.

Our analysis differs from that of Dwork et al. in two key ways. First, we give a better bound on the probability with which a *single* low-sensitivity query output by

a DP-stable algorithm has good generalisation error. Second, we show a reduction from the case of *many* queries to the case of a single query that has no loss in parameters (in contrast, previous work took a union bound over queries, leading to a dependence on k, the number of queries).

Both steps rely on a thought experiment in which several "real" executions of a stable algorithm are simulated inside another algorithm, called a *monitor*, which outputs a function of the "real" transcripts. Because stability is closed under post-processing, the monitor is itself stable. Because it exists only as a thought experiment, the monitor can be given knowledge of the true distribution from which the data are drawn, and can use this knowledge to process the outputs of the simulated "real" runs. The monitor technique allows us to start from a basic guarantee, which states that a single query has good generalisation error with constant probability, and amplify the guarantee so that

- the generalisation error holds with very high probability, and
- the guarantee holds simultaneously over all queries in a sequence.

The proof of the basic guarantee follows the lines of existing proofs using algorithmic stability (e.g., [DW79a]), while the monitor technique and the resulting amplification statements are new.

The amplification of success probability is the more technically sophisticated of the two key steps. The idea is to run many (about $1/\beta$, using the notation of Theorem 3.1.1) copies of a stable mechanism on independently selected data sets. Each of these interactions results in a sequence of queries and answers. The monitor then selects the the query and answer pair from amongst all of the sequences that has the largest error. It then outputs this query as well as the index of the interaction that produced it. Our main technical lemma shows that the monitor will find a "bad"

query/dataset pair (one where the true and empirical values of the query differ) with at most constant probability. This implies that the each of the real executions outputs a bad query with probability $O(\beta)$. Relative to previous work, the resulting argument yields better bounds, applies to more general classes of queries, and even generalises to other notions of stability.

Optimality. In general, we cannot prove that our bounds are optimal. Even for nonadaptive statistical queries, $n \gtrsim \log(k)/\alpha^2$ samples are necessary, and in Chapter 6 we show that that $n \gtrsim \min\{\sqrt{k}, \sqrt{\log|\mathcal{X}|}\}/\alpha$ samples are necessary to answer adaptively chosen statistical queries. However, there remains a gap between the upper and lower bounds.

However, we can show that our connection between DP stability and generalisation is optimal (see Section 3.7 for details). Moreover, for every family of queries we consider, no DP-stable algorithm can achieve better sample complexity [BUV14, BST14]. Thus, any significant improvement to our bounds must come from using a weaker notion of stability or some entirely different approach.

Computational Complexity. Throughout, we will assume that the analyst only issues queries q such that the empirical answer $q(x)$ can be evaluated in time $\text{poly}(n, \log|\mathcal{X}|)$. When $k \ll n^2$ our algorithms have similar running time. However, when answering $k \gg n^2$ queries, our algorithms suffer running time at least $\text{poly}(n, |\mathcal{X}|)$. Since the mechanism's input is of size $n \cdot \log|\mathcal{X}|$, these algorithms cannot be considered computationally efficient. For example, if $\mathcal{X} = \{0,1\}^d$ for some dimension d, then in the non adaptive setting $\text{poly}(n, d)$ running time would suffice, whereas our algorithms require $\text{poly}(n, 2^d)$ running time. Unfortunately, this running time is known to be optimal, as Chapter 6 (building on an earlier

impossibility result [HU14] and hardness results in privacy [Ull13]) shows that, assuming exponentially hard one-way functions exist, there is no $\text{poly}(n, 2^{o(d)})$-time mechanism that accurately answers $k = \omega(n^2)$ statistical queries.

Stable / Differentially Private Mechanisms. Each of our results requires instantiating the mechanism with a suitable stable / differentially private algorithm. For statistical queries, the optimal mechanisms are the well known Gaussian and Laplace Mechanisms (slightly refined in Chapter 4) when k is small and the Private Multiplicative Weights Mechanism [HR10] when k is large. For arbitrary low-sensitivity queries, the Gaussian or Laplace Mechanism is again optimal when k is small, and for large k we can use the Median Mechanism [RR10].

When considering arbitrary search queries over an arbitrary finite range, the optimal algorithm is the Exponential Mechanism [MT07]. For the special case of convex minimisation queries over an infinite domain, we use the optimal algorithm of [BST14] when k is small, and when k is large, we use an algorithm of [Ull15] that accurately answers exponentially many such queries.

Other Notions of Stability Our techniques applies to notions of distributional stability other than differential privacy. In particular, defining stability in terms of total variation (TV) or KL divergence (KL) leads to bounds on the generalisation error that have polynomially, rather than exponentially, decreasing tails. See Section 3.4 for details.

3.2 Preliminaries

3.2.1 Queries

Given a distribution $\mathcal{P}$ over $\mathcal{X}$ or a sample $x = (x_1, \cdots, x_n) \in \mathcal{X}^n$, we would like to answer *queries* about $\mathcal{P}$ or x from some family Q. We will often want to bound the "sensitivity" of the queries with respect to changing one element of the sample. To this end, we use $x \sim x'$ to denote that $x, x' \in \mathcal{X}^n$ differ on at most one entry. We will consider several different families of queries:

- **Statistical Queries:** These queries are specified by a function $q : \mathcal{X} \to [0,1]$, and (abusing notation) are defined as

$$q(\mathcal{P}) = \mathop{\mathbb{E}}_{z \leftarrow_R \mathcal{P}}[q(z)] \qquad \text{and} \qquad q(x) = \frac{1}{n}\sum_{i \in [n]} q(x_i).$$

  The error of an answer a to a statistical query q with respect to $\mathcal{P}$ or x is defined to be

$$\mathrm{err}_x(q, a) = a - q(x) \qquad \text{and} \qquad \mathrm{err}^{\mathcal{P}}(q, a) = a - q(\mathcal{P}).$$

- **Δ-Sensitive Queries:** For $\Delta \in [0,1]$, $n \in \mathbb{N}$, these queries are specified by a function $q : \mathcal{X}^n \to \mathbb{R}$ satisfying $|q(x) - q(x')| \leq \Delta$ for every pair $x, x' \in \mathcal{X}^n$ differing in only one entry. Abusing notation, let

$$q(\mathcal{P}) = \mathop{\mathbb{E}}_{z \leftarrow_R \mathcal{P}^n}[q(z)].$$

  The error of an answer a to a Δ-sensitive query q with respect to $\mathcal{P}$ or x is defined to be

$$\mathrm{err}_x(q, a) = a - q(x) \qquad \text{and} \qquad \mathrm{err}^{\mathcal{P}}(q, a) = \mathop{\mathbb{E}}_{z \leftarrow_R \mathcal{P}^n}[\mathrm{err}_z(q, a)] = a - q(\mathcal{P}).$$

We denote the set of all Δ-sensitive queries by Q_Δ. If $\Delta = O(1/n)$ we say the query is *low sensitivity*. Note that $1/n$-sensitive queries are a strict generalisation of statistical queries.

- **Minimisation Queries:** These queries are specified by a loss function $L : \mathcal{X}^n \times \Theta \to \mathbb{R}$. We require that L has sensitivity Δ with respect to its first parameter, that is,

$$\sup_{\theta \in \Theta,\ x, x' \in \mathcal{X}^n,\ x \sim x'} |L(x;\theta) - L(x';\theta)| \leq \Delta\,.$$

Here Θ is an arbitrary set of items (sometimes called "parameter values") among which we aim to chose the item ("parameter") with minimal loss, either with respect to a particular input data set x, or with respect to expectation over a distribution $\mathcal{P}$.

The error of an answer $\theta \in \Theta$ to a minimisation query $L : \mathcal{X}^n \times \Theta \to \mathbb{R}$ with respect to x is defined to be

$$\mathrm{err}_x\left(L, \theta\right) = L(x, \theta) - \min_{\theta^* \in \Theta} L(x, \theta^*)$$

and, with respect to $\mathcal{P}$, is

$$\mathrm{err}^{\mathcal{P}}\left(L, \theta\right) = \mathop{\mathbb{E}}_{z \leftarrow_{\mathrm{R}} \mathcal{P}^n}\left[\mathrm{err}_z\left(L, \theta\right)\right] = \mathop{\mathbb{E}}_{z \leftarrow_{\mathrm{R}} \mathcal{P}^n}\left[L(z, \theta)\right] - \mathop{\mathbb{E}}_{z \leftarrow_{\mathrm{R}} \mathcal{P}^n}\left[\min_{\theta^* \in \Theta} L(z, \theta^*)\right].$$

Note that $\min_{\theta^* \in \Theta} \mathop{\mathbb{E}}_{z \leftarrow_{\mathrm{R}} \mathcal{P}^n}\left[L(z, \theta^*)\right] \geq \mathop{\mathbb{E}}_{z \leftarrow_{\mathrm{R}} \mathcal{P}^n}\left[\min_{\theta^* \in \Theta} L(z, \theta^*)\right]$, whence

$$\mathop{\mathbb{E}}_{z \leftarrow_{\mathrm{R}} \mathcal{P}^n}\left[L(z, \theta)\right] - \min_{\theta^* \in \Theta} \mathop{\mathbb{E}}_{z \leftarrow_{\mathrm{R}} \mathcal{P}^n}\left[L(z, \theta^*)\right] \leq \mathrm{err}^{\mathcal{P}}\left(L, \theta\right).$$

Note that minimisation queries (with $\Theta = \mathbb{R}$) generalise low-sensitivity queries: Given a Δ-sensitive $q : \mathcal{X}^n \to \mathbb{R}$, we can define $L(x;\theta) = |\theta - q(x)|$ to obtain a minimisation query with the same answer.

We denote the set of minimisation queries by Q_{min}. We highlight two special cases:

- *Minimisation for Finite Sets:* We denote by $Q_{min,D}$ the set of minimisation queries where Θ is finite with size at most D.
- *Convex Minimisation Queries:* If $\Theta \subset \mathbb{R}^d$ is closed and convex and $L(x;\cdot)$ is convex on Θ for every data set x, then the query can be answered nonprivately up to any desired error α, in time polynomial in d and α. We denote the set of all convex minimisation queries by Q_{CM}.

3.2.2 Mechanisms for Adaptive Queries

Our goal is to design a *mechanism* $\mathcal{M}$ that answers queries on $\mathcal{P}$ using only independent samples $x_1, \ldots, x_n \leftarrow_R \mathcal{P}$. Our focus is the case where the queries are chosen adaptively and adversarially.

Specifically, $\mathcal{M}$ is a stateful algorithm that holds a collection of samples $x_1, \ldots, x_n$ from the data universe $\mathcal{X}$, takes a query q from some family Q as input, and returns an answer a. We require that when $x_1, \ldots, x_n$ are independent samples from $\mathcal{P}$, the answer a is "close" to $q(\mathcal{P})$ in a sense that is appropriate for the family of queries. Moreover we require that this condition holds for every query in an adaptively chosen sequence $q_1, \ldots, q_k$. Formally, we define an accuracy game between a mechanism $\mathcal{M}$ and a stateful *data analyst* $\mathcal{A}$ in Figure 3.1.

Definition 3.2.1 (Accuracy)**.** *A mechanism $\mathcal{M}$ is (α, β)*-accurate with respect to the population for k adaptively chosen queries from Q given n samples in $\mathcal{X}$ *if for every adversary $\mathcal{A}$,*

$$\mathop{\mathbb{P}}_{\mathsf{Acc}_{n,k,Q}[\mathcal{M},\mathcal{A}]}\left[\max_{j\in[k]} \left|\mathrm{err}^{\mathcal{P}}\left(q_j, a_j\right)\right| \leq \alpha\right] \geq 1-\beta.$$

$\mathcal{A}$ chooses a distribution $\mathcal{P}$ over $\mathcal{X}$.
Sample $x_1, \ldots, x_n \leftarrow_R \mathcal{P}$, let $x = (x_1, \ldots, x_n)$. (Note that $\mathcal{A}$ does not know x.)
For $j = 1, \ldots, k$
  $\mathcal{A}$ outputs a query $q_j \in Q$.
  $\mathcal{M}(x, q_j)$ outputs a_j.
  (As $\mathcal{A}$ and $\mathcal{M}$ are stateful, q_j and a_j may depend on the history $q_1, a_1, \ldots, q_{j-1}, a_{j-1}$.)

Figure 3.1: *The Accuracy Game* $\mathsf{Acc}_{n,k,Q}[\mathcal{M}, \mathcal{A}]$

We will also use a definition of accuracy relative to the sample given to the mechanism, described in Figure 3.2.

$\mathcal{A}$ chooses $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$.
For $j = 1, \ldots, k$
  $\mathcal{A}$ outputs a query $q_j \in Q$.
  $\mathcal{M}(x, q_j)$ outputs a_j.
  (q_j and a_j may depend on the history $q_1, a_1, \ldots, q_{j-1}, a_{j-1}$ and on x.)

Figure 3.2: *The Sample Accuracy Game* $\mathsf{SampAcc}_{n,k,Q}[\mathcal{M}, \mathcal{A}]$

Definition 3.2.2 (Sample Accuracy)**.** *A mechanism $\mathcal{M}$ is (α, β)*-accurate with respect to samples of size n from $\mathcal{X}$ for k adaptively chosen queries from Q *if for every adversary $\mathcal{A}$,*

$$\mathop{\mathbb{P}}_{\mathsf{SampAcc}_{n,k,Q}[\mathcal{M},\mathcal{A}]}\left[\max_{j\in[k]} \left|\mathrm{err}_x\left(q_j, a_j\right)\right| \leq \alpha\right] \geq 1 - \beta.$$

3.2.3 DP Stability

In this context we choose the term "DP stability" rather than "differential privacy" to emphasise the conceptual relationship between this notion and other notions of algorithmic stability that have been studied in machine learning. We also emphasise that our work has a very different motivation to the motivation of differential privacy

— stable algorithms are desirable even when privacy is not a concern, such as when the data does not concern humans.

In our analysis, we will make crucial use of the fact that DP-stability (as well as the other notions of stability discussed in Section 3.4.1) is *closed under post-processing*.

Lemma 3.2.1 (Post-Processing). *Let $\mathcal{W} : \mathcal{X}^n \to \mathcal{R}$ and $f : \mathcal{R} \to \mathcal{R}'$ be a pair of randomised algorithms. If $\mathcal{W}$ is (ε, δ)-DP-stable then the algorithm $f(\mathcal{W}(x))$ is (ε, δ)-DP-stable.*

Stability for Interactive Mechanisms

The definition of differential privacy does not immediately apply to algorithms that interact with a data analyst to answer adaptively chosen queries. Such a mechanism does not simply take a sample x as input and produce an output. Instead, in the interactive setting, there is a mechanism $\mathcal{M}$ that holds a sample x and interacts with some algorithm $\mathcal{A}$. We can view this entire interaction between $\mathcal{M}$ and $\mathcal{A}$ as a single noninteractive meta algorithm that outputs the transcript of the interaction and define stability with respect to that meta algorithm. Specifically, we define the algorithm $\mathcal{W}[\mathcal{M}, \mathcal{A}](x)$ that simulates the interaction between $\mathcal{M}(x)$ and $\mathcal{A}$ and outputs the messages sent between them. The simulation is also parameterised by n, k, Q, although we will frequently omit these parameters when they are clear from context.

Input: A sample $x \in \mathcal{X}^n$
For $j = 1, \ldots, k$
Feed a_{j-1} to $\mathcal{A}$ and get a query $q_j \in Q$.
Feed q_j to $\mathcal{M}(x)$ and get an answer $a_j \in \mathcal{R}$.
Output $((q_1, a_1), \ldots, (q_k, a_k))$.

Figure 3.3: $\mathcal{W}_{n,k,Q}[\mathcal{M}, \mathcal{A}] : \mathcal{X}^n \to (Q \times \mathcal{R})^k$

Note that $\mathcal{W}[\mathcal{M}, \mathcal{A}]$ is a noninteractive mechanism, and its output is just the query-answer pairs of $\mathcal{M}$ and $\mathcal{A}$ in the sample accuracy game, subject to the mechanism being given the sample x. Now we can define the stability of an interactive mechanism $\mathcal{M}$ using $\mathcal{W}$.

Definition 3.2.3 (Stability of for Interactive Mechanisms)**.** *We say an interactive mechanism $\mathcal{M}$ is* (ε, δ)-DP-stable for k queries from Q *if for every adversary $\mathcal{A}$, the algorithm $\mathcal{W}_{n,k,Q}[\mathcal{M}, \mathcal{A}](x) : \mathcal{X}^n \to (Q \times \mathcal{R})^k$ is (ε, δ)-DP-stable.*

Composition of DP Stability

The definition above allows for *adaptive composition*. This follows directly from composition results of (ε, δ)-*differentially private* algorithms. A mechanism that is (ε, δ)-DP-stable for 1 query is $(\approx \varepsilon\sqrt{k}, \approx \delta k)$-stable for k adaptively chosen queries. See Chapter 2 for appropriate composition bounds.

3.3 From DP Stability to Accuracy for Low-Sensitivity Queries

In this section we prove our main result that any mechanism that is both accurate with respect to the sample and satisfies DP stability (with suitable parameters) is also accurate with respect to the population. The proof proceeds in two mains steps. First, we prove a lemma that says that there is no DP-stable mechanism that takes several independent sets of samples from the distribution and finds a query and a set of samples such that the answer to that query on that set of samples is very different from the answer to that query on the population. In Section 3.3.2 we prove this lemma for the simpler case of statistical queries and then in 3.3.3 we extend the

proof to the more general case of low-sensitivity queries.

The second step is to introduce a *monitoring algorithm*. This monitoring algorithm will simulate the interaction between the mechanism and the adversary on multiple independent sets of samples.

It will then output the least accurate query across all the different interactions. We show that if the mechanism is stable then the monitoring algorithm is also stable. By choosing the number of sets of samples appropriately, we ensure that if the mechanism has even a small probability of being inaccurate in a given interaction, then the monitor will have a constant probability of finding an inaccurate query in one of the interactions. By the lemma proven in the first step, no such monitoring algorithm can satisfy DP stability, therefore every stable mechanism must be accurate with high probability.

3.3.1 Warmup: A Single-Sample De-Correlated Expectation Lemma for Statistical Queries

As a warmup, in this section we give a simpler version of our main lemma for the case of statistical queries and a single sample. Although these results follow from the results of Section 3.3.3 on general low-sensitivity queries, we include the simpler version to introduce the main ideas in the cleanest possible setting.

Lemma 3.3.1. *Let $\mathcal{W} : \mathcal{X}^n \to Q$ be (ε, δ)-DP-stable where Q is the class statistical queries $q : \mathcal{X} \to [0,1]$. Let $\mathcal{P}$ be a distribution on $\mathcal{X}$ and let $x \leftarrow_R \mathcal{P}^n$. Then*[4]

$$\left| \mathop{\mathbb{E}}_{x,\mathcal{W}} [q(\mathcal{P}) \mid q = \mathcal{W}(x)] - \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} [q(x) \mid q = \mathcal{W}(x)] \right| \leq e^{\varepsilon} - 1 + \delta.$$

[4]The notation $\mathop{\mathbb{E}}_{x,\mathcal{W}} [q(\mathcal{P}) \mid q = \mathcal{W}(x)]$ should be read as "the expectation of $q(\mathcal{P})$, where q denotes the output of $\mathcal{W}(x)$." That is, the "event" being conditioned on is simply a definition of the random variable q.

Proof of Lemma 3.3.1. Before giving the proof, we set up some notation. Let $x = (x_1, \ldots, x_n)$. For a single element $x' \in \mathcal{X}$, and an index $i \in [n]$, we use $x_{i \to x'}$ to denote the new sample where the i-th element of x has been replaced by the element x'. Let $x' \leftarrow_{\mathrm{R}} \mathcal{P}$ be independent from x.

We can now calculate

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x,\mathcal{W}}[q(x) \mid q = \mathcal{W}(x)] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{x,\mathcal{W}}[q(x_i) \mid q = \mathcal{W}(x)] \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_0^1 \mathop{\mathbb{P}}_{x,\mathcal{W}}[q(x_i) > z \mid q = \mathcal{W}(x)]\,\mathrm{d}z
\end{aligned}$$

Now we can apply DP stability:

$$\begin{aligned}
&\leq \frac{1}{n}\sum_{i=1}^{n} \int_0^1 e^{\varepsilon} \mathop{\mathbb{P}}_{x,\mathcal{W}}[q(x_i) > z \mid q = \mathcal{W}(x_{i\to x'})] + \delta \mathrm{d}z \qquad \text{(by } (\varepsilon,\delta)\text{-DP stability)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{x',x,\mathcal{W}}[q(x_i) \mid q = \mathcal{W}(x_{i\to x'})] + \delta \right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{x',x,\mathcal{W}}\left[q(x') \mid q = \mathcal{W}(x)\right] + \delta \right) \\
&\qquad\qquad \text{(the pairs } (x_i, x_{i\to x'}) \text{ and } (x', x) \text{ are identically distributed)} \\
&= e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{x',x,\mathcal{W}}\left[q(x') \mid q = \mathcal{W}(x)\right] + \delta \\
&= e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{x,\mathcal{W}}[q(\mathcal{P}) \mid q = \mathcal{W}(x)] + \delta
\end{aligned}$$

An identical argument shows that

$$\mathop{\mathbb{E}}_{x,\mathcal{W}}[q(x) \mid q = \mathcal{W}(x)] \geq e^{-\varepsilon} \cdot \left(\mathop{\mathbb{E}}_{x,\mathcal{W}}[q(\mathcal{P}) \mid q = \mathcal{W}(x)] - \delta \right).$$

Therefore, using the fact that $|q(\mathcal{P})| \leq 1$ for any statistical query q and distribu-

tion $\mathcal{P}$, we have

$$\left| \mathop{\mathbb{E}}_{x,\mathcal{W}} [q(\mathcal{P}) \mid q = \mathcal{W}(x)] - \mathop{\mathbb{E}}_{x,\mathcal{W}} [q(x) \mid q = \mathcal{W}(x)] \right| \leq e^{\varepsilon} - 1 + \delta,$$

as desired. □

3.3.2 Warmup: A Multi-Sample De-Correlated Expectation Lemma for Statistical Queries

As a second warmup, in this section we give a simpler version of our main lemma for the case of statistical queries and *multiple* samples. That is, we consider a setting where there are many subsamples available to the algorithm. The multi-sample de-correlated expectation lemma says that a DP-stable algorithm cannot take a collection of samples $x_1, \ldots, x_T$ and output a pair (q, t) such that $q(\mathcal{P})$ and $q(x_t)$ differ significantly in expectation.

Lemma 3.3.2. *Let $\mathcal{W} : (\mathcal{X}^n)^T \to Q \times [T]$ be (ε, δ)-DP-stable where Q is the class statistical queries $q : \mathcal{X} \to [0, 1]$. Let $\mathcal{P}$ be a distribution on $\mathcal{X}$ and let $\mathbf{X} = (x_1, \ldots, x_T) \leftarrow_R (\mathcal{P}^n)^T$. Then*

$$\left| \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} [q(\mathcal{P}) \mid (q, t) = \mathcal{W}(\mathbf{X})] - \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} [q(x_t) \mid (q, t) = \mathcal{W}(\mathbf{X})] \right| \leq e^{\varepsilon} - 1 + T\delta.$$

Proof of Lemma 3.3.2. Before giving the proof, we set up some notation. Let $\mathbf{X} = (x_1, \ldots, x_T)$ be a set of T samples where each sample $x_t = (x_{t,1}, \ldots, x_{t,n})$. For a single element $x' \in \mathcal{X}$, and a pair of indices $(m, i) \in [T] \times [n]$, we use $\mathbf{X}_{(m,i)\to x'}$ to denote the new set of T samples where the i-th element of the m-th sample of $\mathbf{X}$ has been replaced by the element x'.

We can now calculate

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}[q(x_t) \mid (q,t) = \mathcal{W}(\mathbf{X})] \\
&= \sum_{m=1}^{T} \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}\left[\mathbf{1}_{\{t=m\}} \cdot q(x_m) \;\middle|\; (q,t) = \mathcal{W}(\mathbf{X})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{T} \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}\left[\mathbf{1}_{\{t=m\}} \cdot q(x_{m,i}) \;\middle|\; (q,t) = \mathcal{W}(\mathbf{X})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{T} \int_0^1 \mathop{\mathbb{P}}_{\mathbf{X},\mathcal{W}}\left[\mathbf{1}_{\{t=m\}} \cdot q(x_{m,i}) \;\geq z \middle|\; (q,t) = \mathcal{W}(\mathbf{X})\right] \mathrm{d}z
\end{aligned}
$$

Now we can apply (ε, δ)-DP stability.

$$
\begin{aligned}
&\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{T} \left(\int_0^1 e^{\varepsilon} \mathop{\mathbb{P}}_{\mathbf{X},\mathcal{W}}\left[\mathbf{1}_{\{t=m\}} \cdot q(x_{m,i}) \;\geq z \middle|\; (q,t) = \mathcal{W}(\mathbf{X}_{(m,i)\to x'})\right] + \delta\right) \mathrm{d}z \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by } (\varepsilon,\delta)\text{-DP stability)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{T} \left(e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{x',\mathbf{X},\mathcal{W}}\left[\mathbf{1}_{\{t=m\}} \cdot q(x_{m,i}) \;\middle|\; (q,t) = \mathcal{W}(\mathbf{X}_{(m,i)\to x'})\right] + \delta\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{T} \left(e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{x',\mathbf{X},\mathcal{W}}\left[\mathbf{1}_{\{t=m\}} \cdot q(x') \;\middle|\; (q,t) = \mathcal{W}(\mathbf{X})\right] + \delta\right) \\
&\qquad\qquad\qquad\qquad \text{(the pairs } (x_{m,i}, \mathbf{X}_{(m,i)\to x'}) \text{ and } (x', \mathbf{X}) \text{ are identically distributed)} \\
&= e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{x',\mathbf{X},\mathcal{W}}\left[q(x') \;\middle|\; (q,t) = \mathcal{W}(\mathbf{X})\right] + T\delta \\
&= e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}[q(\mathcal{P}) \mid (q,t) = \mathcal{W}(\mathbf{X})] + T\delta \\
&\leq \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}[q(\mathcal{P}) \mid (q,t) = \mathcal{W}(\mathbf{X})] + e^{\varepsilon} - 1 + T\delta \qquad \text{(since } q(\mathcal{P}) \in [0,1]\text{)}
\end{aligned}
$$

An identical argument shows that

$$
\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}[q(x_t) \mid (q,t) = \mathcal{W}(\mathbf{X})] \geq \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}[q(\mathcal{P}) \mid (q,t) = \mathcal{W}(\mathbf{X})] + (e^{-\varepsilon} - 1) - T\delta.
$$

□

3.3.3 A Multi-Sample De-Correlated Expectation Lemma

Here, we give the most general de-correlated expectation lemma that considers multiple samples and applies to the more general class of low-sensitivity queries.

Lemma 3.3.3 (Main Technical Lemma). *Let $\mathcal{W} : (\mathcal{X}^n)^T \to Q_\Delta \times [T]$ be (ε, δ)-DP-stable where Q_Δ is the class of Δ-sensitive queries $q : \mathcal{X}^n \to \mathbb{R}$. Let $\mathcal{P}$ be a distribution on $\mathcal{X}$ and let $\mathbf{X} = (x_1, \ldots, x_T) \leftarrow_R (\mathcal{P}^n)^T$. Then*

$$\left| \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} [q(\mathcal{P}) \mid (q,t) = \mathcal{W}(\mathbf{X})] - \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} [q(x_t) \mid (q,t) = \mathcal{W}(\mathbf{X})] \right| \leq 2(e^\varepsilon - 1 + T\delta)\Delta n.$$

We remark that if we use the weaker assumption that $\mathcal{W}$ is $(e^\varepsilon - 1 + \delta)$-TV stable, (defined in Section 3.4.1), then we would obtain the same conclusion but with the weaker bound of $2T(e^\varepsilon - 1 + \delta)\Delta n$. The advantage of using the stronger definition of DP stability is that we only have to decrease δ with T and not ε. This advantage is crucial because algorithms satisfying (ε, δ)-DP stability necessarily have a linear dependence on $1/\varepsilon$ but only a polylogarithmic dependence on $1/\delta$

Proof of Lemma 3.3.3. Let $\mathbf{X}' = (x'_1, \ldots, x'_T) \leftarrow_R (\mathcal{P}^n)^T$ be independent of $\mathbf{X}$. Recall that each element x_t of $\mathbf{X}$ is itself a vector $(x_{t,1}, \ldots, x_{t,n})$, and the same is true for each element x'_t of $\mathbf{X}'$. We will sometimes refer to the vectors $x_1, \ldots, x_T$ as the *subsamples of* $\mathbf{X}$.

We define a sequence of intermediate samples that allow us to interpolate between $\mathbf{X}$ and $\mathbf{X}'$ using a series of neighbouring samples. Formally, for $\ell \in \{0, 1, \ldots, n\}$ and $m \in \{0, 1, \ldots, T\}$, define $\mathbf{X}^{\ell,m} = (x_1^{\ell,m}, \ldots, x_T^{\ell,m}) \in (\mathcal{X}^n)^T$ by

$$x_{t,i}^{\ell,m} = \begin{cases} x_{t,i} & (t > m) \text{ or } (t = m \text{ and } i > \ell) \\ x'_{t,i} & (t < m) \text{ or } (t = m \text{ and } i \leq \ell) \end{cases}$$

By construction we have $\mathbf{X}^{0,1} = \mathbf{X}^{n,0} = \mathbf{X}$ and $\mathbf{X}^{n,T} = \mathbf{X}'$. Also $\mathbf{X}^{0,m} = \mathbf{X}^{n,m-1}$ for

$m \in [T]$. Moreover, pairs $(\mathbf{X}^{\ell,t}, \mathbf{X}^{\ell-1,t})$ are neighboring in the sense that there is a single subsample, x_t such that $x_t^{\ell,t}$ and $x_t^{\ell-1,T}$ are neighbors and for every $t' \neq t$, $x_{t'}^{\ell,t} = x_{t'}^{\ell-1,t}$.

For $\ell \in [n]$ and $m \in [T]$, define a randomised function $B^{\ell,m} : (\mathcal{X}^n)^T \times (\mathcal{X}^n)^T \to \mathbb{R}$ by

$$B^{\ell,m}(\mathbf{X}, \mathbf{Z}) = \begin{cases} q(\mathbf{z}_t) - q(\mathbf{z}_{t,-\ell}) + \Delta & t = m \\ 0 & t \neq m \end{cases} \quad \text{where } (q,t) = \mathcal{W}(\mathbf{X}),$$

where $\mathbf{z}_{t,-\ell}$ is the t-th subsample of $\mathbf{Z}$ with its ℓ-th element replaced by some arbitrary fixed element of $\mathcal{X}$.

We can now expand $\left| \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} [q(\mathcal{P}) - q(x_t) \mid (q,t) = \mathcal{W}(\mathbf{X})] \right|$ in terms of these intermediate samples and the functions $B^{\ell,m}$:

$$\begin{aligned}
&\left| \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} [q(\mathcal{P}) - q(x_t) \mid (q,t) = \mathcal{W}(\mathbf{X})] \right| \\
&= \left| \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}} \left[q(x_t') - q(x_t) \mid (q,t) = \mathcal{W}(\mathbf{X}) \right] \right| \\
&= \left| \sum_{\ell \in [n]} \sum_{m \in [T]} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}} \left[q(x_t^{\ell,m}) - q(x_t^{\ell-1,m}) \;\middle|\; (q,t) = \mathcal{W}(\mathbf{X}) \right] \right| \\
&\leq \sum_{\ell \in [n]} \left| \sum_{m \in [T]} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}} \left[q(x_t^{\ell,m}) - q(x_t^{\ell-1,m}) \;\middle|\; (q,t) = \mathcal{W}(\mathbf{X}) \right] \right| \\
&= \sum_{\ell \in [n]} \left| \sum_{m \in [T]} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}} \left[\begin{array}{l} \left(q(x_t^{\ell,m}) - q(x_{t,-\ell}^{\ell,m}) + \Delta \right) \\ -\left(q(x_t^{\ell-1,m}) - q(x_{t,-\ell}^{\ell-1,m}) + \Delta \right) \end{array} \;\middle|\; (q,t) = \mathcal{W}(\mathbf{X}) \right] \right| \\
&\qquad\qquad \text{(By construction, } x_{t,-\ell}^{\ell,m} = x_{t,-\ell}^{\ell-1,m}\text{)} \\
&= \sum_{\ell \in [n]} \left| \sum_{m \in [T]} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}} \left[B^{\ell,m}(\mathbf{X}, \mathbf{X}^{\ell,m}) - B^{\ell,m}(\mathbf{X}, \mathbf{X}^{\ell-1,m}) \right] \right| \qquad \text{(Definition of } B^{\ell,m}\text{.)}
\end{aligned}$$

Thus, it suffices to show that $\left| \sum_{m \in [T]} \mathbb{E} \left[B^{\ell,m}(\mathbf{X}, \mathbf{X}^{\ell,m}) - B^{\ell,m}(\mathbf{X}, \mathbf{X}^{\ell-1,m}) \right] \right| \leq 2(e^\varepsilon - 1 + T\delta)\Delta$ for all $\ell \in [n]$. To this end, we make a few observations.

1. Since q is Δ-sensitive, for every $\ell, m, \mathbf{X}, \mathbf{Z}$, we have $0 \leq B^{\ell,m}(\mathbf{X}, \mathbf{Z}) \leq 2\Delta$. Moreover, since $B^{\ell,m}(\mathbf{X}, \mathbf{Z}) = 0$ whenever $\mathcal{W}(\mathbf{X})$ outputs (q, t) with $t \neq m$, we have $\sum_{m \in [T]} \mathbb{E}\left[B^{\ell,m}(x, x^{\ell,m})\right] \leq 2\Delta$.

2. By construction, $B^{\ell,m}(\mathbf{X}, \mathbf{Z})$ is (ε, δ)-DP-stable as a function of its first parameter $\mathbf{X}$. Stability follows by the post-processing lemma (Lemma 3.2.1) since $B^{\ell,m}$ is a post-processing of the output of $\mathcal{W}(\mathbf{X})$, which is assumed to be (ε, δ)-DP-stable.

3. Lastly, observe that the random variables $\mathbf{X}^{\ell,m}$ are identically distributed (although they are not independent). Namely, each one consists of nT independent samples from $\mathcal{P}$. Moreover, for every ℓ and m, the pair $(\mathbf{X}^{\ell,m}, \mathbf{X})$ has the same distribution as $(\mathbf{X}, \mathbf{X}^{\ell,m})$. Specifically, the first component is nT independent samples from $\mathcal{P}$ and the second component is equal to the first component with a subset of the entries replaced by fresh independent samples from $\mathcal{P}$.

Consider the random variables $B^{\ell,m}(\mathbf{X}, \mathbf{X}^{\ell,m})$ and $B^{\ell,m}(\mathbf{X}, \mathbf{X}^{\ell-1,m})$ for some $\ell \in [n]$ and $m \in [T]$. Using observations 2 and 3, we have

$$B^{\ell,m}(\mathbf{X}, \mathbf{X}^{\ell,m}) \sim B^{\ell,m}(\mathbf{X}^{\ell,m}, \mathbf{X}) \sim_{(\varepsilon,\delta)} B^{\ell,m}(\mathbf{X}^{\ell-1,m}, \mathbf{X}) \sim B^{\ell,m}(\mathbf{X}, \mathbf{X}^{\ell-1,m}),$$

where $\sim$ denotes having the same distribution and $\sim_{(\varepsilon,\delta)}$ denotes having (ε, δ)-DP close distributions.[5] Thus $B^{\ell,m}(\mathbf{X}, \mathbf{X}^{\ell-1,m})$ and $B^{\ell,m}(\mathbf{X}, \mathbf{X}^{\ell,m})$ are (ε, δ)-DP close.

[5]In the spirit of (ε, δ)-DP stability, we say that distributions A and B over $\mathcal{R}$ are (ε, δ)-DP close if for every $R \subseteq \mathcal{R}$, $\mathbb{P}[A \in R] \leq e^{\varepsilon} \cdot \mathbb{P}[B \in R] + \delta$ and vice versa.

Now we can calculate

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell-1,m})\right] &= \int_0^{2\Delta} \mathop{\mathbb{P}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell-1,m}) \geq z\right] \mathrm{d}z \\
&\leq \int_0^{2\Delta} \left(e^{\varepsilon} \cdot \mathop{\mathbb{P}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m}) \geq z\right] + \delta\right) \mathrm{d}z \\
&= e^{\varepsilon} \cdot \int_0^{2\Delta} \mathop{\mathbb{P}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m}) \geq z\right] \mathrm{d}z + 2\delta\Delta \\
&= e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m})\right] + 2\delta\Delta.
\end{aligned}$$

Thus we have

$$\begin{aligned}
\sum_{m\in[T]} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell-1,m})\right] &\leq e^{\varepsilon} \cdot \left(\sum_{m\in[T]} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X},\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m})\right]\right) + 2T\delta\Delta \\
&\leq \sum_{m\in[T]} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{X}',\mathcal{W}}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m})\right] + 2(e^{\varepsilon}-1)\Delta + 2\Delta T\delta.
\end{aligned}$$

Thus we have the desired upper bound on the expectation of

$$\sum_{m\in[T]} \mathbb{E}\left[B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell,m}) - B^{\ell,m}(\mathbf{X},\mathbf{X}^{\ell-1,m})\right].$$

The corresponding lower bound follows from an analogous argument. This completes the proof. □

3.3.4 From Multi-Sample De-Correlated Expectation to Accuracy

Now that we have Lemma 3.3.3, we can prove the following result that DP-stable mechanisms that are also accurate with respect to their sample are also accurate with respect to the population from which that sample was drawn.

Theorem 3.3.4 (Main Transfer Theorem)**.** *Let Q be a family of Δ-sensitive queries on $\mathcal{X}$. Assume that, for some $\alpha, \beta \in (0, .1)$, $\mathcal{M}$ is*

1. *$(\varepsilon = \alpha/64\Delta n, \delta = \alpha\beta/32\Delta n)$-DP-stable for k adaptively chosen queries from Q and*

2. $(\alpha' = \alpha/8, \beta' = \alpha\beta/16\Delta n)$*-accurate with respect to its sample for* n *samples from* $\mathcal{X}$ *for* k *adaptively chosen queries from* Q.

Then $\mathcal{M}$ *is* (α, β)*-accurate with respect to the population for* k *adaptively chosen queries from* Q *given* n *samples from* $\mathcal{X}$.

The key step in the proof is to define a monitoring algorithm that takes T separate samples $\mathbf{X} = (x_1, \ldots, x_T)$ and for each sample x_t, simulates an independent interaction between $\mathcal{M}(x_t)$ and $\mathcal{A}$. This monitoring algorithm then outputs the query with the largest error across all of the queries and interactions (kT queries in total). Since changing one input to $\mathbf{X}$ only affects one of the simulations, the monitoring algorithm will be stable so long as $\mathcal{M}$ is stable, without any loss in the stability parameter. On the other hand, if $\mathcal{M}$ has even a small chance β of answering a query with large error, then if we simulate $T \approx 1/\beta$ independent interactions, there is a constant probability that at least one of the simulations results in a query with large error. Thus, the monitor will be a stable algorithm that outputs a query with large error *in expectation*. By the multi-sample de-correlated expectation lemma, such a monitor is impossible, which implies that $\mathcal{M}$ has probability $\leq \beta$ of answering any query with large error.

Proof of Theorem 3.3.4. Let $\mathcal{M}$ be an interactive mechanism. Let $\mathcal{A}$ be an analyst and let $\mathcal{P}$ be the distribution chosen by $\mathcal{A}$. We define the following monitoring algorithm.

If $\mathcal{M}$ is stable then so is $\mathcal{W}$, and this fact follows easily from the post-processing lemma (Lemma 3.2.1):

Claim 3.3.5. *For every* $\varepsilon, \delta \geq 0$, *if the mechanism* $\mathcal{M}$ *is* (ε, δ)*-DP-stable for* k *adaptively chosen queries from* Q, *then for every* $\mathcal{P}$ *and* $\mathcal{A}$, *the monitor* $\mathcal{W}_{\mathcal{P},k,Q}[\mathcal{M}, \mathcal{A}]$ *is* (ε, δ)*-DP-stable.*

Input: $\mathbf{X} = (x_1, \ldots, x_T) \in (\mathcal{X}^n)^T$

For $t = 1, \ldots, T$:

Simulate $\mathcal{M}(x_t)$ and $\mathcal{A}$ interacting, let $q_{t,1}, \ldots, q_{t,k} \in Q$ be the queries of $\mathcal{A}$ and let

$a_{t,1}, \ldots, a_{t,k} \in \mathbb{R}$ be the corresponding answers of $\mathcal{M}$.

Let

$$(j^*, t^*) = \text{argmax}_{j \in [k],\, t \in [T]} \left| \text{err}^{\mathcal{P}} \left(q_{t,j}, a_{t,j} \right) \right| .$$

If $a_{t^*,j^*} - q_{t^*,j^*}(\mathcal{P}) \geq 0$, let $q^* = q_{t^*,j^*}$, otherwise let $q^* = -q_{t^*,j^*}$. (Q_Δ is closed under negation.)

Output: (q^*, t^*).

Figure 3.4: $\mathcal{W}(\mathbf{X}) = \mathcal{W}_{\mathcal{P}}[\mathcal{M}, \mathcal{A}](\mathbf{X})$:

Proof. If $\mathcal{M}$ is (ε, δ)-DP-stable for k adaptively chosen queries from Q then for every analyst $\mathcal{A}$ who asks k queries from Q, and every t the algorithm $\mathcal{W}'(x_t)$ that simulates the interaction between $\mathcal{M}(x_t)$ and $\mathcal{A}$ and outputs the resulting query-answer pairs is (ε, δ)-DP-stable. From this, it follows that the algorithm $\mathcal{W}'(\mathbf{X})$ that simulates the interactions between $\mathcal{M}(x_t)$ and $\mathcal{A}$ for every $t = 1, \ldots, T$ and outputs the resulting query-answer pairs is (ε, δ)-DP-stable. To see this, observe that if $\mathbf{X}, \mathbf{X}'$ differ only on one subsample x_t, then for every $t' \neq t$, $x_{t'} = x'_{t'}$ and thus the query-answer pairs corresponding to subsample t' are identically distributed regardless of whether we use $\mathbf{X}$ or $\mathbf{X}'$ as input to $\mathcal{W}$.

Observe that the algorithm $\mathcal{W}$ defined above is simply a post-processing of these kT query-answer pairs. That is, (q^*, t^*) depends only on $\{(q_{t,j}, a_{t,j}\}_{t \in [T], j \in [k]}$ and $\mathcal{P}$, and not on $\mathbf{X}$. Thus, by Lemma 3.2.1, $\mathcal{W}$ is (ε, δ)-DP-stable. □

We will use the $\mathcal{W}$ with $T = \lfloor 1/\beta \rfloor$. In light of Claim 3.3.5 and our assumption

that $\mathcal{M}$ is (ε, δ)-DP-stable, we can apply Lemma 3.3.3 to obtain

$$\left| \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} [q^*(\mathcal{P}) - q^*(x_{t^*}) \mid (q^*, t^*) = \mathcal{W}(\mathbf{X})] \right| \leq 2\left(e^{\alpha/64\Delta n} - 1 + T\left(\frac{\alpha\beta}{32\Delta n} \right) \right) \Delta n \leq \frac{\alpha}{8}. \tag{3.1}$$

To complete the proof, we show that if $\mathcal{M}$ is not (α, β)-accurate with respect to the population $\mathcal{P}$, then (3.1) cannot hold. To do so, we need the following natural claim about the output of the monitor.

Claim 3.3.6. $\mathop{\mathbb{P}}_{\mathbf{X},\mathcal{W}} \left[q^*(\mathcal{P}) - a_{q^*} > \alpha \right] > 1 - (1-\beta)^T$, *and* $q^*(\mathcal{P}) - a_{q^*} \geq 0$, *where* a_{q^*} *is the answer to* q^* *produced during the simulation.*

Proof. Since $\mathcal{M}$ fails to be (α, β)-accurate, for every $t \in [T]$,

$$\mathop{\mathbb{P}}_{x_t,\mathcal{M}} \left[\max_{j \in [k]} \left| q_{t,j}(\mathcal{P}) - a_{t,j} \right| > \alpha \right] > \beta. \tag{3.2}$$

We obtain the claim from (3.2) by using the fact that the T sets of query-answer pairs corresponding to different subsamples $x_1, \ldots, x_T$ are independent. That is, the random variables $\max_{j \in [k]} \left| q_{t,j}(\mathcal{P}) - a_{t,j} \right|$ indexed by $t \in [T]$ are independent. Since $q^*(\mathcal{P}) - a_{q^*}$ is simply the maximum of these independent random variables, the first part of the claim follows. Also, by construction, $\mathcal{W}$ ensures that

$$q^*(\mathcal{P}) - a_{q^*} \geq 0. \tag{3.3}$$

□

Claim 3.3.7. *If* $\mathcal{M}$ *is* (α', β')*-accurate for the sample but not* (α, β)*-accurate for the population, then*

$$\left| \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}} [q^*(\mathcal{P}) - q^*(x_{t^*}) \mid (q^*, t^*) = \mathcal{W}(\mathbf{X})] \right| \geq \alpha/4.$$

Proof. Now we can calculate

$$
\begin{aligned}
&\left|\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}\left[q^*(\mathcal{P}) - q^*(x_{t^*}) \mid (q^*,t^*) = \mathcal{W}(\mathbf{X})\right]\right| \\
&= \left|\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}\left[q^*(\mathcal{P}) - a_{q^*} \mid (q^*,t^*) = \mathcal{W}(\mathbf{X})\right] + \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}\left[a_{q^*} - q^*(x_{t^*}) \mid (q^*,t^*) = \mathcal{W}(\mathbf{X})\right]\right| \\
&\geq \left|\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}\left[q^*(\mathcal{P}) - a_{q^*} \mid (q^*,t^*) = \mathcal{W}(x)\right]\right| - \left|\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}\left[a_{q^*} - q^*(x_{t^*}) \mid q^* = \mathcal{W}(x)\right]\right| \\
&\geq \alpha(1-(1-\beta)^T) - \left|\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}\left[a_{q^*} - q^*(x_{t^*}) \mid (q^*,t^*) = \mathcal{W}(\mathbf{X})\right]\right| && \text{(Claim 3.3.6)} \\
&\geq \alpha(1-(1-\beta)^T) - \left(\alpha/8 + 2T\left(\frac{\alpha\beta}{16\Delta n}\right)\Delta n\right) && (3.4) \\
&\geq \alpha/2 - (\alpha/8 + \alpha/8) = \alpha/4 && (T = \lfloor 1/\beta \rfloor.)
\end{aligned}
$$

Line (3.4) follows from two observations. First, since $\mathcal{M}$ is assumed to be $(\alpha/8, \alpha\beta/16\Delta n)$-accurate for one sample, by a union bound, it is simultaneously $(\alpha/8, T(\alpha\beta/16\Delta n))$-accurate for all of the T samples. Thus, we have $a_{q^*} - q^*(x_{t^*}) \leq \alpha'$ except with probability at most $T(\alpha\beta/16\Delta n)$. Second, since q^* is a Δ-sensitive query, we always have $a_{q^*} - q^*(x_{t^*}) \leq 2\Delta n$.[6] □

Thus, if $\mathcal{M}$ is not (α,β)-accurate for the population, we will obtain a contradiction to (3.1). This completes the proof. □

3.4 Other Notions of Stability and Accuracy on Average

Definition 3.4.2 gives one notion of stability, namely DP stability. However, this is by no means the only way to formalise stability for our purposes. In this section we consider other notions of stability and the advantages they have.

[6]Without loss of generality, the answers of $\mathcal{M}$ can be truncated to an interval of width $2\Delta n$ that contains the correct answer $q^*(x_{t^*})$. Doing so will ensure $|a_{q^*} - q^*(x_{t^*})| \leq 2\Delta n$.

3.4.1 Other Notions of Algorithmic Stability

We will define here other notions of algorithmic stability, and in Section 3.4.2, we will show that such notions can provide expected guarantees for generalisation error which can be used to achieve accuracy on average.

Definition 3.4.1 (TV-Stability)**.** *Let $\mathcal{W} : \mathcal{X}^n \to \mathcal{R}$ be a randomised algorithm. We say that $\mathcal{W}$ is* ε-TV stable *if for every pair of samples that differ on exactly one element,*

$$\mathrm{d_{TV}}(\mathcal{W}(x), \mathcal{W}(x')) = \sup_{R \subseteq \mathcal{R}} \left| \mathbb{P}\left[\mathcal{W}(x) \in R\right] - \mathbb{P}\left[\mathcal{W}(x') \in R\right] \right| \leq \varepsilon.$$

Definition 3.4.2 (KL-Stability)**.** *Let $\mathcal{W} : \mathcal{X}^n \to \mathcal{R}$ be a randomised algorithm. We say that $\mathcal{W}$ is* ε-KL-stable *if for every pair of samples x, x' that differ on exactly one element,*

$$\mathop{\mathbb{E}}_{r \leftarrow_R \mathcal{W}(x)} \left[\log \left(\frac{\mathbb{P}\left[\mathcal{W}(x) = r\right]}{\mathbb{P}\left[\mathcal{W}(x') = r\right]} \right) \right] \leq 2\varepsilon^2$$

The post-processing property of DP stability (Lemma 3.2.1 in Section 3.2.3) also applies to the two stability notions above.

Lemma 3.4.1 (Stability Notions Preserved under Post-Processing)**.** *Let $\mathcal{W} : \mathcal{X}^n \to \mathcal{R}$ and $f : \mathcal{R} \to \mathcal{R}'$ be a pair of randomised algorithms. If $\mathcal{W}$ is {ε-TV, ε-KL, (ε, δ)-DP}-stable then the algorithm $f(\mathcal{W}(x))$ is {ε-TV, ε-KL, (ε, δ)-DP}-stable.*

Relationships Between Stability Notions ε-KL stability implies ε-TV stability by Pinsker's inequality. The relationship between DP stability defined in Section 3.2.3 and the above notions is more subtle. When $\varepsilon \leq 1$, $(\varepsilon, 0)$-DP stability implies ε-KL stability and thus also ε-TV stability. When $\varepsilon \leq 1$ and $\delta > 0$, (ε, δ)-DP stability implies $(2\varepsilon + \delta)$-TV stability. It also implies that $\mathcal{M}$ is "close" to satisfying 2ε-KL stability (cf. [DRV10] for more discussion of these notions).

As in Section 3.2.3, we define TV-stability and KL-stability of an interactive

mechanism $\mathcal{M}$ through a noninteractive mechanism that simulates the interaction between $\mathcal{M}$ and an adversary $\mathcal{A}$. The definition for these notions of stability is precisely analogous to Definition 3.2.3 for DP stability.

As with DP stability, both notions above allow for *adaptive composition*. In fact, ε-TV stability composes linearly—a mechanism that is ε-TV stable for 1 query is εk-stable for k queries. The advantage of the stronger notions of KL and DP stability is that they have a stronger composition. A mechanism that is ε-KL stable for 1 query is $(\varepsilon\sqrt{k})$-stable for k queries.

3.4.2 From TV Stability to Accuracy on Average

In this section we show that TV stable algorithms guarantee a weaker notion of accuracy on average for adaptively chosen queries.

3.4.3 Accuracy on Average

In Section 3.2.2 we defined accurate mechanisms to be those that answer accurately (either with respect to the population or the sample) with probability close to 1. In this section we define a relaxed notion of accuracy that only requires low error in expectation over the coins of $\mathcal{M}$ and $\mathcal{A}$.

Definition 3.4.3 (Average Accuracy)**.** *A mechanism $\mathcal{M}$ is α-*accurate on average with respect to the population for k adaptively chosen queries from Q given n samples in $\mathcal{X}$ *if for every adversary $\mathcal{A}$,*

$$\mathop{\mathbb{E}}_{\mathsf{Acc}_{n,k,Q}[\mathcal{M},\mathcal{A}]}\left[\max_{j\in[k]}\left|\mathrm{err}^{\mathcal{P}}\left(q_j,a_j\right)\right|\right]\leq\alpha.$$

We will also use a definition of accuracy relative to the sample given to the mechanism:

Definition 3.4.4 (Sample Accuracy on Average)**.** *A mechanism $\mathcal{M}$ is α-*accurate on average with respect to samples of size n from $\mathcal{X}$ for k adaptively chosen queries from Q *if for every adversary $\mathcal{A}$,*

$$\mathop{\mathbb{E}}_{\mathsf{SampAcc}_{n,k,Q}[\mathcal{M},\mathcal{A}]}\left[\max_{j\in[k]} \left|\mathrm{err}_x\left(q_j, a_j\right)\right|\right] \leq \alpha.$$

A De-Correlated Expectation Lemma

Towards our goal of proving that TV stability implies accuracy on average in the adaptive setting, we first prove a lemma saying that TV stable algorithms cannot output a low-sensitivity query such that the sample has large error for that query. In the next section we will show how this lemma implies accuracy on average in the adaptive setting.

Lemma 3.4.5. *Let $\mathcal{W} : \mathcal{X}^n \to Q_\Delta$ be an ε-TV stable randomised algorithm. Recall Q_Δ is the family of Δ-sensitive queries $q : \mathcal{X}^n \to \mathbb{R}$. Let $\mathcal{P}$ be a distribution on $\mathcal{X}$ and let $x \leftarrow_R \mathcal{P}^n$. Then*

$$\left|\mathop{\mathbb{E}}_{x,\mathcal{W}}\left[q(\mathcal{P}) \mid q = \mathcal{W}(x)\right] - \mathop{\mathbb{E}}_{x,\mathcal{W}}\left[q(x) \mid q = \mathcal{W}(x)\right]\right| \leq 2\varepsilon\Delta n.$$

Proof. The proof proceeds via a sequence of intermediate samples. Let $x' \leftarrow_R \mathcal{P}^n$ be independent of x. For $\ell \in \{0, 1, \ldots, n\}$, we define $x^\ell = (x_1^\ell, \ldots, x_n^\ell) \in \mathcal{X}^n$ by

$$x_i^\ell = \begin{cases} x_i & i > \ell \\ x_i' & i \leq \ell \end{cases}$$

By construction, $x^0 = x$ and $x^n = x'$, and intermediate samples x^ℓ interpolate between x and x'. Moreover, x^ℓ and $x^{\ell+1}$ differ in at most one entry, so that we can use the stability condition to relate $\mathcal{W}(x^\ell)$ and $\mathcal{W}(x^{\ell+1})$.

For every $\ell \in [n]$, we define $B^\ell : \mathcal{X}^n \times \mathcal{X}^n \to \mathbb{R}$ by

$$B^\ell(x, \mathbf{z}) = q(\mathbf{z}) - q(\mathbf{z}_{-\ell}) + \Delta, \text{ where } q = \mathcal{W}(x).$$

Here, $\mathbf{z}_{-\ell}$ is $\mathbf{z}$ with the ℓ-th element replaced by some arbitrary fixed element of $\mathcal{X}$.

Now we can write

$$\begin{aligned}
&\left| \mathop{\mathbb{E}}_{x,\mathcal{W}} [q(\mathcal{P}) - q(x) \mid q = \mathcal{W}(x)] \right| \\
&= \left| \mathop{\mathbb{E}}_{x,x',\mathcal{W}} \left[q(x') - q(x) \mid q = \mathcal{W}(x) \right] \right| \\
&= \left| \sum_{\ell=1}^{n} \mathop{\mathbb{E}}_{x,x',\mathcal{W}} \left[q(x^\ell) - q(x^{\ell-1}) \mid q = \mathcal{W}(x) \right] \right| \\
&\leq \sum_{\ell=1}^{n} \left| \mathop{\mathbb{E}}_{x,x',\mathcal{W}} \left[q(x^\ell) - q(x^{\ell-1}) \mid q = \mathcal{W}(x) \right] \right| \\
&= \sum_{\ell \in [n]} \left| \mathop{\mathbb{E}}_{x,x',\mathcal{W}} \left[\left(q(x^\ell) - q(x^\ell_{-\ell}) + \Delta \right) - \left(q(x^{\ell-1}) - q(x^{\ell-1}_{-\ell}) + \Delta \right) \;\middle|\; q = \mathcal{W}(x) \right] \right| \\
&\qquad\qquad \text{(Since } x^\ell_{-\ell} = x^{\ell-1}_{-\ell}\text{)} \\
&= \sum_{\ell \in [n]} \left| \mathop{\mathbb{E}}_{x,x',\mathcal{W}} \left[B^\ell(x, x^\ell) - B^\ell(x, x^{\ell-1}) \right] \right|. \qquad \text{(Definition of } B\text{)}
\end{aligned}$$

Thus, to prove the lemma, it suffices to show that for every $\ell \in [n]$,

$$\left| \mathop{\mathbb{E}}_{x,x',\mathcal{W}} \left[B^\ell(x, x^\ell) - B^\ell(x, x^{\ell-1}) \right] \right| \leq 2\Delta\varepsilon$$

To complete the proof, we will need a few observations. First, since q is Δ-sensitive, for every $\ell, x, \mathbf{z}$, we have $0 \leq B^\ell(x, \mathbf{z}) \leq 2\Delta$.

Second, observe that since $\mathcal{W}$ is assumed to be ε-TV stable, by the post-processing lemma (Lemma 3.2.1) $B^\ell(x, \mathbf{z})$ is ε-TV stable with respect to its first parameter x.

Finally, observe that the random variables $x^0, \ldots, x^n$ are identically distributed (although not independent). That is, every x^ℓ consists of n independent draws from $\mathcal{P}$. Moreover, for every ℓ, the pairs (x, x^ℓ) and (x^ℓ, x) are identically distributed.

Specifically, the first component is n independent samples from $\mathcal{P}$ and the second component is equal to the first component with a subset of the entries replaced by new independent samples from $\mathcal{P}$.

Combining, the second and third observation with the triangle inequality, we have

$$\begin{aligned}\mathrm{d_{TV}}\left(B^\ell(x,x^\ell),B^\ell(x,x^{\ell-1})\right) &\le \mathrm{d_{TV}}\left(B^\ell(x,x^\ell),B^\ell(x^\ell,x)\right)\\ &\quad+\mathrm{d_{TV}}\left(B^\ell(x^\ell,x),B^\ell(x^{\ell-1},x)\right)\\ &\quad+\mathrm{d_{TV}}\left(B^\ell(x^{\ell-1},x),B^\ell(x,x^{\ell-1})\right)\\ &\le 0+\varepsilon+0=\varepsilon.\end{aligned}$$

Using the observations above, for every $\ell\in[n]$ we have

$$\mathop{\mathbb{E}}_{x,x',\mathcal{W}}\left[B^\ell(x,x^\ell)-B^\ell(x,x^{\ell-1})\right]\le 2\Delta\cdot\mathrm{d_{TV}}\left(B^\ell(x,x^\ell),B^\ell(x,x^{\ell-1})\right)\le 2\Delta\varepsilon.$$

Thus we have the desired upper bound on the expectation of $B^\ell(x,x^\ell)-B^\ell(x,x^{\ell-1})$. The corresponding lower bound follows from an analogous argument. This completes the proof. □

From De-Correlated Expectation to Accuracy on Average

Theorem 3.4.6. *Let Q_Δ be the family of Δ-sensitive queries on $\mathcal{X}$. Assume that $\mathcal{M}$ is*

1. *$(\varepsilon=\alpha/4\Delta n)$-TV stable for k adaptively chosen queries from $Q=Q_\Delta$ and*
2. *$(\alpha'=\alpha/2)$-accurate on average with respect to its sample for n samples from $\mathcal{X}$ for k adaptively chosen queries from Q.*

Then $\mathcal{M}$ is α-accurate on average with respect to the population for k adaptively chosen queries from Q given n samples from $\mathcal{X}$.

The high level approach of the proof is to apply the Lemma 3.4.5 to a "monitoring algorithm" that watches the interaction between the mechanism $\mathcal{M}(x)$ and the analyst $\mathcal{A}$ and then outputs the *least accurate* query. Since $\mathcal{M}(x)$ is stable, the de-correlated expectation lemma says that the query output by the monitor will satisfy $q(\mathcal{P}) \approx q(x)$ in expectation, this implies that even for the least accurate query in the interaction between $\mathcal{M}(x)$ and $\mathcal{A}$, $q(\mathcal{P}) \approx q(x)$ in expectation Thus, if $\mathcal{M}$ is accurate with respect to the sample x, it is also accurate with respect to $\mathcal{P}$.

Proof of Theorem 3.4.6. Let $\mathcal{M}$ be an interactive mechanism and $\mathcal{A}$ be an analyst that chooses the distribution $\mathcal{P}$. We define the following monitoring algorithm. If $\mathcal{M}$

Input: $x \in \mathcal{X}^n$
Simulate $\mathcal{M}(x)$ and $\mathcal{A}$ interacting, let $q_1, \ldots, q_k \in Q$ be the queries of $\mathcal{A}$ and let
$a_1, \ldots, a_k \in \mathbb{R}$ be the corresponding answers of $\mathcal{M}$.
Let $j = \text{argmax}_{j=1,\ldots,k} \left|\text{err}^{\mathcal{P}}\left(q_j, a_j\right)\right|$.
If $a_j - q_j(\mathcal{P}) \geq 0$, let $q^* = q_j$, otherwise let $q^* = -q_j$. (Q_Δ is closed under negation.)
Output: q^*.

Figure 3.5: $\mathcal{W}(x) = \mathcal{W}_{\mathcal{P}}[\mathcal{M}, \mathcal{A}](x)$:

is stable then so is $\mathcal{W}$, and this fact follows easily from the post-processing lemma (Lemma 3.2.1).

Claim 3.4.7. *For every $\varepsilon \geq 0$, if the mechanism $\mathcal{M}$ is ε-TV stable for k adaptively chosen queries from Q, then for every $\mathcal{P}$ and $\mathcal{A}$, the monitor $\mathcal{W}_{\mathcal{P}}[\mathcal{M}, \mathcal{A}]$ is ε-TV stable.*

Proof of Claim 3.4.7. The assumption that $\mathcal{M}$ is ε-TV stable for k adaptively chosen queries from Q means that for every analyst $\mathcal{A}$ who asks k queries from Q, the algorithm $\mathcal{W}'(x)$ that simulates the interaction between $\mathcal{M}(x)$ and $\mathcal{A}$ and outputs the resulting query-answer pairs is ε-TV stable. Observe that the algorithm $\mathcal{W}$

defined above is simply a post-processing of these query-answer pairs. That is, q^* depends only on $q_1, a_1, \ldots, q_k, a_k$ and $\mathcal{P}$, and not on x. Thus, by Lemma 3.2.1, for every $\mathcal{P}$ and $\mathcal{A}$, the monitor $\mathcal{W}_{\mathcal{P}}[\mathcal{M}, \mathcal{A}]$ is ε-TV stable. □

In light of Claim 3.4.7 and our assumption that $\mathcal{M}$ is $(\alpha/4\Delta n)$-TV stable, we can apply Lemma 3.4.5 to obtain

$$\left| \mathop{\mathbb{E}}_{x,\mathcal{W}} [q^*(\mathcal{P}) - q^*(x) \mid q^* = \mathcal{W}(x)] \right| \leq 2 \left(\frac{\alpha}{4\Delta n} \right) \Delta n \leq \alpha/2. \tag{3.5}$$

To complete the proof, we show that if $\mathcal{M}$ is not α-accurate on average with respect to the population $\mathcal{P}$, then (3.5) cannot hold.

Claim 3.4.8. *If $\mathcal{M}$ is $(\alpha/2)$-accurate for the sample but not α-accurate for the population, then*

$$\left| \mathop{\mathbb{E}}_{x,\mathcal{W}} [q^*(\mathcal{P}) - q^*(x) \mid q^* = \mathcal{W}(x)] \right| \geq \alpha/2.$$

Proof of Claim 3.4.8. Using our assumptions, we can calculate as follows.

$$\begin{aligned}
&\left| \mathop{\mathbb{E}}_{x,\mathcal{W}} [q^*(\mathcal{P}) - q^*(x) \mid q^* = \mathcal{W}(x)] \right| \\
&= \left| \mathop{\mathbb{E}}_{x,\mathcal{W}} [q^*(\mathcal{P}) - a_{q^*} \mid q^* = \mathcal{W}(x)] + \mathop{\mathbb{E}}_{x,\mathcal{W}} [a_{q^*} - q^*(x) \mid q^* = \mathcal{W}(x)] \right| \\
&\geq \left| \mathop{\mathbb{E}}_{x,\mathcal{W}} [q^*(\mathcal{P}) - a_{q^*} \mid q^* = \mathcal{W}(x)] \right| - \left| \mathop{\mathbb{E}}_{x,\mathcal{W}} [a_{q^*} - q^*(x) \mid q^* = \mathcal{W}(x)] \right| \\
&> \alpha - \left| \mathop{\mathbb{E}}_{x,\mathcal{W}} [a_{q^*} - q^*(x) \mid q^* = \mathcal{W}(x)] \right| && (3.6) \\
&\geq \alpha - \alpha/2 && (3.7) \\
&= \alpha/2.
\end{aligned}$$

Line (3.6) follows from two observations. First, by construction of $\mathcal{W}$, we always have $q^*(\mathcal{P}) - a_{q^*} \leq 0$. Second, since $\mathcal{M}$ is assumed not to be α-accurate on average for the population, the expected value of $|q^*(\mathcal{P}) - a_{q^*}| > \alpha$. Since $\mathcal{W}$ ensures

that $a_{q^*} - q^*(\mathcal{P}) \geq 0$, we also have that the absolute value of the expectation of $q^*(\mathcal{P}) - a_{q^*}$ is greater than α. Line (3.7) follows from the assumption that $\mathcal{M}$ is $(\alpha/2)$-accurate on average for the sample. □

Thus, if $\mathcal{M}$ is not α-accurate on average for the population, we will obtain a contradiction to (3.5). This completes the proof. □

3.5 From Low-Sensitivity Queries to Optimisation Queries

In this section, we extend our results for low-sensitivity queries to the more general family of minimisation queries. To do so, we design a suitable monitoring algorithm for minimisation queries. As in our analysis of low-sensitivity queries, we will have the monitoring algorithm take as input many independent samples and simulate the interaction between $\mathcal{M}$ and $\mathcal{A}$ on each of those samples. Thus, if $\mathcal{M}$ has even a small probability of being inaccurate, then with constant probability the monitor will find a minimisation query that $\mathcal{M}$ has answered inaccurately. Previously, we had monitor simply output this query and applied Lemma 3.3.3 to arrive at a contradiction. However, since Lemma 3.3.3 only applies to algorithms that output a low-sensitivity query, we can't apply it to the monitor that outputs a minimisation query. We address this by having the monitor output the *error function* associated with the loss function and answer it selects, which is a low-sensitivity query. If we assume that the mechanism is accurate for its sample but not for the population, then the monitor will find a loss function and an answer with low error on the sample but large error on the population. Thus the error function will be a low-sensitivity query with very different answers on the sample and the population,

which is a contradiction. To summarise, we have the following theorem.

Theorem 3.5.1 (Transfer Theorem for Minimisation Queries). *Let $Q = Q_{min}$ be the family of Δ-sensitive minimisation queries on $\mathcal{X}$. Assume that, for some $\alpha, \beta \geq 0$, $\mathcal{M}$ is*

1. *$(\varepsilon = \alpha/128\Delta n, \delta = \alpha\beta/64\Delta n)$-DP-stable for k adaptively chosen queries from Q and*

2. *$(\alpha' = \alpha/8, \beta' = \alpha\beta/32\Delta n)$-accurate with respect to its sample for n samples from $\mathcal{X}$ for k adaptively chosen queries from Q.*

Then $\mathcal{M}$ is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q given n samples from $\mathcal{X}$.

The formal proof is nearly identical to that of Theorem 3.3.4, so we omit the full proof. Instead, we will simply describe the modified monitoring algorithm.

Input: $\mathbf{X} = (x_1, \ldots, x_T) \in (\mathcal{X}^n)^T$
For $t = 1, \ldots, T$:
Simulate $\mathcal{M}(x_t)$ and $\mathcal{A}$ interacting, let $L_{t,1}, \ldots, L_{t,k} \in Q$ be the queries of $\mathcal{A}$ and let
$\theta_{t,1}, \ldots, \theta_{t,k} \in \mathbb{R}$ be the corresponding answers of $\mathcal{M}$.
Let (t^*, j^*) be

$$(t^*, j^*) = \text{argmax}_{j \in [k],\, t \in [T]} \left| \text{err}^{\mathcal{P}} \left(L_{t,j}, \theta_{t,j} \right) \right|.$$

Let $q^*(x) = \text{err}_x(L_{t^*,j^*}, \theta_{t^*,j^*})$ (note, by construction, $q^* \in Q_{2\Delta}$, i.e. q^* is 2Δ-sensitive)
Output: (q^*, t^*).

Figure 3.6: $\mathcal{W}(\mathbf{X}) = \mathcal{W}_{\mathcal{P}}[\mathcal{M}, \mathcal{A}](\mathbf{X})$:

3.6 Applications

3.6.1 Low-Sensitivity and Statistical Queries

We now plug known stable mechanisms (designed in the context of differential privacy) in to Theorem 3.3.4 to obtain mechanisms that provide strong error guarantees with high probability for both low-sensitivity and statistical queries.

Corollary 3.6.1 (Theorem 3.3.4 and Theorem 4.4.1)**.** *There is a mechanism $\mathcal{M}$ that is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q_Δ where $\Delta = O(1/n)$ given n samples from $\mathcal{X}$ for*

$$n \geq O\left(\frac{\sqrt{k \cdot \log\log k} \cdot \log^{3/2}(1/\alpha\beta)}{\alpha^2}\right)$$

The mechanism runs in time $\mathrm{poly}(n, \log|\mathcal{X}|, \log(1/\beta))$ *per query.*

Corollary 3.6.2 (Theorem 3.3.4 and [RR10])**.** *There is a mechanism $\mathcal{M}$ that is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q_Δ where $\Delta = O(1/n)$ given n samples from $\mathcal{X}$ for*

$$n = O\left(\frac{\log|\mathcal{X}| \cdot \log k \cdot \log^{3/2}(1/\alpha\beta)}{\alpha^3}\right)$$

The mechanism runs in time $\mathrm{poly}(|\mathcal{X}|^n)$ *per query. The case where Δ is not $O(1/n)$ can be handled by rescaling the output of the query.*

Corollary 3.6.3 (Theorem 3.3.4 and [HR10])**.** *There is a mechanism $\mathcal{M}$ that is α-accurate on average with respect to the population for k adaptively chosen queries from Q_{SQ} given n samples from $\mathcal{X}$ for*

$$n = O\left(\frac{\sqrt{\log|\mathcal{X}|} \cdot \log k \cdot \log^{3/2}(1/\alpha\beta)}{\alpha^3}\right)$$

The mechanism runs in time $\mathrm{poly}(n, |\mathcal{X}|)$ *per query.*

3.6.2 Optimisation Queries

The results of the Section 3.5 can be combined with existing differentially private algorithms for minimising "empirical risk" (that is, loss with respect to the sample x) to obtain algorithms for answering adaptive sequences of minimisation queries. We provide a few specific instantiations here, based on known differentially private mechanisms.

Minimisation Over Arbitrary Finite Sets

Corollary 3.6.4 (Theorem 3.5.1 and [MT07])**.** *Let Θ be a finite set of size at most D. Let $Q \subset Q_{min}$ be the set of sensitivity-$1/n$ loss functions bounded between 0 and C. Then there is a mechanism $\mathcal{M}$ that is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q_{min} given*

$$n \geq O\left(\frac{\log(DC/\alpha) \cdot \sqrt{k} \cdot \log^{3/2}(1/\alpha\beta)}{\alpha^2}\right)$$

samples from $\mathcal{X}$. The running time of the mechanism is dominated by $O((k + \log(1/\beta)) \cdot D)$ evaluations of the loss function.

Convex Minimisation

We state bounds for convex minimisation queries for some of the most common parameter regimes in applications. In the first two corollaries, we consider 1-Lipschitz[7] loss functions over a bounded domain.

Corollary 3.6.5 (Theorem 3.5.1 and [BST14])**.** *Let Θ be a closed, convex subset of $\mathbb{R}^d$ set such that $\max_{\theta \in \Theta} \|\theta\|_2 \leq 1$. Let $Q \subset Q_{min}$ be the set of convex 1-Lipschitz loss functions*

[7] A loss function $L : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}$ is 1-Lipschitz if for every $\theta, \theta' \in \mathbb{R}^d$, $x \in \mathcal{X}$, $|L(\theta, x) - L(\theta', x)| \leq \|\theta - \theta'\|_2$.

that are $1/n$-sensitive. Then there is a mechanism $\mathcal{M}$ that is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q given

$$n = \tilde{O}\left(\frac{\sqrt{dk} \cdot \log^2(1/\alpha\beta)}{\alpha^2}\right)$$

samples from Q. The running time of the mechanism is dominated by $k \cdot n^2$ evaluations of the gradient ∇L.

Corollary 3.6.6 (Theorem 3.5.1 and [Ull15])**.** *Let Θ be a closed, convex subset of $\mathbb{R}^d$ set such that $\max_{\theta \in \Theta} \|\theta\|_2 \leq 1$. Let $Q \subset Q_{min}$ be the set of convex 1-Lipschitz loss functions that are $1/n$-sensitive. Then there is a mechanism $\mathcal{M}$ that is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q given*

$$n = \tilde{O}\left(\frac{\sqrt{\log|\mathcal{X}|} \cdot (\sqrt{d} + \log k) \cdot \log^{3/2}(1/\alpha\beta)}{\alpha^3}\right)$$

samples from $\mathcal{X}$. The running time of the mechanism is dominated by $\mathrm{poly}(n, |\mathcal{X}|)$ and $k \cdot n^2$ evaluations of the gradient ∇L.

In the next two corollaries, we consider 1-strongly convex[8], Lipschitz loss functions over a bounded domain.

Corollary 3.6.7 (Theorem 3.5.1 and [BST14])**.** *Let Θ be a closed, convex subset of $\mathbb{R}^d$ set such that $\max_{\theta \in \Theta} \|\theta\|_2 \leq 1$. Let $Q \subset Q_{min}$ be the set of 1-strongly convex, 1-Lipschitz loss functions that are $1/n$-sensitive. Then there is a mechanism $\mathcal{M}$ that is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q given*

$$n = \tilde{O}\left(\frac{\sqrt{dk} \cdot \log^{3/2}(1/\alpha\beta)}{\alpha^{3/2}}\right)$$

[8] A loss function $L : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}$ is 1-strongly convex if for every $\theta, \theta' \in \mathbb{R}^d$, $x \in \mathcal{X}$,

$$L(\theta', x) \geq L(\theta, x) + \langle \nabla L(\theta, x), \theta' - \theta \rangle + (1/2) \cdot \|\theta - \theta'\|_2^2,$$

where the (sub)gradient $\nabla L(\theta, x)$ is taken with respect to θ.

samples from $\mathcal{X}$. The running time of the mechanism is dominated by $k \cdot n^2$ evaluations of the gradient ∇L.

Corollary 3.6.8 (Theorem 3.5.1 and [Ull15])**.** *Let Θ be a closed, convex subset of $\mathbb{R}^d$ set such that $\max_{\theta \in \Theta} \|\theta\|_2 \leq 1$. Let $Q \subset Q_{min}$ be the set of 1-strongly convex 1-Lipschitz loss functions that are $1/n$-sensitive. Then there is a mechanism $\mathcal{M}$ that is (α, β)-accurate with respect to the population for k adaptively chosen queries from Q given*

$$n = \tilde{O}\left(\sqrt{\log|\mathcal{X}|} \cdot \left(\frac{\sqrt{d}}{\alpha^{5/2}} + \frac{\log k}{\alpha^3}\right) \cdot \log^{3/2}(1/\alpha\beta)\right)$$

samples from $\mathcal{X}$. The running time of the mechanism is dominated by $\mathrm{poly}(n, |\mathcal{X}|)$ and $k \cdot n^2$ evaluations of the gradient ∇L.

3.7 An Alternative Form of Generalisation and Tightness of Our Results

We now provide an alternative form of our generalisation bounds. The following Theorem is more general than Theorem 3.3.4 because it says that *no* DP-stable procedure that outputs a low-sensitivity can output a query that distinguishes the sample from the population (not just DP-stable procedures that are accurate for the sample).

First we prove the following technical lemma.

Lemma 3.7.1. *Let F be a finite set, $f : F \to \mathbb{R}$ a function, and $\eta > 0$. Define a random variable X on F by*

$$\mathbb{P}[X = x] = \frac{e^{\eta f(x)}}{C}, \qquad \text{where} \qquad C = \sum_{x \in F} e^{\eta f(x)}.$$

Then

$$\mathbb{E}\left[f(X)\right] \geq \max_{x \in F} f(x) - \frac{1}{\eta} \log |F|.$$

Proof. We have

$$f(x) = \frac{1}{\eta}\left(\log C + \log \mathbb{P}\left[X = x\right]\right).$$

Thus

$$\begin{aligned}
\mathbb{E}\left[f(X)\right] &= \sum_{x \in F} \mathbb{P}\left[X = x\right] f(x) \\
&= \sum_{x \in F} \mathbb{P}\left[X = x\right] \frac{1}{\eta}\left(\log C + \log \mathbb{P}\left[X = x\right]\right) \\
&= \frac{1}{\eta}\left(\log C - H(X)\right),
\end{aligned}$$

where $H(X)$ is the Shannon entropy of the distribution of X (measured in nats, rather than bits). In particular,

$$H(X) \leq \log |\text{support}(X)| = \log |F|,$$

as the uniform distribution maximises entropy. Moreover, $C \geq \max_{x \in F} e^{\eta f(x)}$, whence $\frac{1}{\eta} \log C \geq \max_{x \in F} f(x)$. The result now follows from these two inequalities. □

Theorem 3.7.2. *Let $\varepsilon \in (0, 1/3)$, $\delta \in (0, \varepsilon/4)$, and $n \geq \frac{1}{\varepsilon^2} \log(\frac{4\varepsilon}{\delta})$. Let $\mathcal{M} : \mathcal{X}^n \to Q_\Delta$ be (ε, δ)-DP-stable where Q_Δ is the class of Δ-sensitive queries $q : \mathcal{X}^n \to \mathbb{R}$. Let $\mathcal{P}$ be a distribution on $\mathcal{X}$, let $x \leftarrow_R \mathcal{P}^n$, and let $q \leftarrow_R \mathcal{M}(x)$. Then*

$$\mathbb{P}_{x, \mathcal{M}}\left[|q(\mathcal{P}) - q(x)| \geq 18\varepsilon\Delta n\right] < \frac{\delta}{\varepsilon}.$$

Intuitively, Theorem 3.7.2 says that "stability prevents overfitting." It says that no stable algorithm can output a low-sensitivity function that distinguishes its input from the population the input was drawn from (i.e. "overfits" its sample).

In particular, Theorem 3.7.2 implies that, if a mechanism $\mathcal{M}$ is stable and outputs q that "fits" its data, then q also "fits" the population. This gives a learning theory perspective on our results.

Proof. Consider the following monitor algorithm $\mathcal{W}$.

Input: $\mathbf{X} = (x_1, \ldots, x_T) \in (\mathcal{X}^n)^T$
Set $F = \emptyset$.
For $t = 1, \ldots, T$:
Let $q_t \leftarrow \mathcal{M}(x_t)$, and set $F = F \cup \{(q_t, t), (-q_t, t)\}$.
Sample (q^*, t^*) from F with probability proportional to $\exp\left(\frac{\varepsilon}{\Delta}\left(q^*(x_{t^*}) - q^*(\mathcal{P})\right)\right)$.
Output: (q^*, t^*).

Figure 3.7: $\mathcal{W}(\mathbf{X}) = \mathcal{W}_{\mathcal{P}}[\mathcal{M}](\mathbf{X})$:

We will use the monitor $\mathcal{W}$ with $T = \lfloor \epsilon/\delta \rfloor$. Observe that $\mathcal{W}$ only access its input through $\mathcal{M}$ (which is (ε, δ)-DP-stable) and the exponential mechanism (which is $(\varepsilon, 0)$-DP-stable). Thus, by composition and postprocessing, $\mathcal{W}$ is $(2\varepsilon, \delta)$-DP-stable. We can hence apply Lemma 3.3.3 to obtain

$$\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}\left[q^*(x_{t^*}) - q^*(\mathcal{P}) \mid (q^*, t^*) = \mathcal{W}(\mathbf{X})\right] \leq 2\left(e^{2\varepsilon} - 1 + T\delta\right)\Delta n < 8\varepsilon\Delta n. \tag{3.8}$$

Now we can apply Lemma 3.7.1 with $f(q,t) = q(x_t) - q(\mathcal{P})$ and $\eta = \frac{\varepsilon}{\Delta}$ to get

$$\mathop{\mathbb{E}}_{q^*,t^*}\left[f(q^*, t^*)\right] \geq \max_{(q,t)\in F} f(q,t) - \frac{\Delta}{\varepsilon}\log|F| = \max_{t\in[T]}|q_t(x_t) - q_t(\mathcal{P})| - \frac{\Delta}{\varepsilon}\log(2T). \tag{3.9}$$

Combining (3.8) and (3.9) gives

$$\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}\left[\max_{t\in[T]}|q_t(x_t) - q_t(\mathcal{P})|\right] - \frac{\Delta}{\varepsilon}\log(2T) \leq \mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}\left[q^*(x_{t^*}) - q^*(\mathcal{P}) \mid (q^*, t^*) = \mathcal{W}(\mathbf{X})\right] < 8\varepsilon\Delta n. \tag{3.10}$$

To complete the proof, we assume, for the sake of contradiction, that $\mathcal{M}$ has a high enough probability of outputting a query q such that $|q(\mathcal{P}) - q(x)|$ is large. To

obtain a contradiction from this assumption, we need the following natural claim (analogous to Claim 3.3.6) about the output of the monitor.

Claim 3.7.3. *If*

$$\mathop{\mathbb{P}}_{x,\mathcal{M}}\left[|q(\mathcal{P}) - q(x)| \geq 18\varepsilon\Delta n\right] \geq \frac{\delta}{\varepsilon},$$

then

$$\mathop{\mathbb{P}}_{\mathbf{X},\mathcal{W}}\left[\max_{t\in[T]} |q_t(x_t) - q_t(\mathcal{P})| \geq 18\varepsilon\Delta n\right] \geq 1 - \left(1 - \frac{\delta}{\varepsilon}\right)^T \geq \frac{1}{2}.$$

Thus

$$\mathop{\mathbb{E}}_{\mathbf{X},\mathcal{W}}\left[\max_{t\in[T]} |q_t(x_t) - q_t(\mathcal{P})|\right] \geq 9\varepsilon\Delta n. \tag{3.11}$$

Combining (3.10) and (3.11) gives

$$9\varepsilon\Delta n - \frac{\Delta}{\varepsilon}\log(2T) \leq 8\varepsilon\Delta n,$$

which simplifies to

$$\log(2\varepsilon/\delta) \geq \log(2T) \geq \varepsilon^2 n.$$

This contradicts the assumption that $n \geq \frac{1}{\varepsilon^2}\log(\frac{4\varepsilon}{\delta})$ and hence completes the proof.

□

3.7.1 Optimality

We now show that our connection between DP stability and generalisation (Theorem 3.7.2 and Theorem 3.3.4) is optimal.

Lemma 3.7.1. *Let $\alpha > \delta > 0$, let $n \geq \frac{1}{\alpha}$, and let $\Delta \in [0,1]$. Let $\mathcal{U}$ be the uniform distribution over $[0,1]$. There exists a $(0,\delta)$-DP-stable algorithm $\mathcal{A} : [0,1]^n \to Q_\Delta$ such that if $\mathbf{X} \leftarrow_R \mathcal{U}^n$ and if $q \leftarrow_R \mathcal{A}(\mathbf{X})$ then*

$$\Pr[q(\mathbf{X}) - q(\mathcal{U}) \geq \alpha\Delta n] \geq \frac{\delta}{2\alpha}.$$

Proof. Consider the following simple algorithm, denoted as $\mathcal{B}$. On input a database x, output x with probability δ, and otherwise output the empty database. Clearly, $\mathcal{B}$ is $(0,\delta)$-DP-stable. Now construct the following algorithm $\mathcal{A}$.

Input: A database $\mathbf{X} \in [0,1]^n$. We think of $\mathbf{X}$ as $\frac{1}{\alpha}$ databases of size αn each: $\mathbf{X} = (x_1, \ldots, x_{1/\alpha})$.
For $1 \leq i \leq 1/\alpha$ let $\hat{x}_i = \mathcal{B}(x_i)$.
Let $p : [0,1] \to \{0,1\}$ where $p(x) = 1$ iff $\exists i$ s.t. $x \in \hat{x}_i$.
Define $q_p : [0,1]^n \to \mathbb{R}$ where $q_p(x) = \Delta \sum_{x \in x} p(x)$ (note that q_p is a Δ-sensitive query, and that it is a statistical query if $\Delta = 1/n$).
Output: q_p.

Figure 3.8: $\mathcal{A} : [0,1]^n \to Q_\Delta$

As $\mathcal{B}$ is $(0,\delta)$-DP-stable, and as $\mathcal{A}$ only applies $\mathcal{B}$ on disjoint databases, we get that $\mathcal{A}$ is also $(0,\delta)$-DP-stable.

Suppose $\mathbf{X} = (x_1, \ldots, x_{1/\alpha})$ contains i.i.d. samples from $\mathcal{U}$, and consider the execution of $\mathcal{A}$ on $\mathbf{X}$. Observe that the predicate p evaluates to 1 only on a finite number of points from $[0,1]$, and hence, we have that $q_p(\mathcal{U}) = 0$. Next note that $q_p(\mathbf{X}) = \alpha\Delta n \cdot |\{i : \hat{x}_i = x_i\}|$. Therefore, if there exists an i s.t. $\hat{x}_i = x_i$ then $q(\mathbf{X}) - q(\mathcal{U}) \geq \alpha\Delta n$. The probability that this is not the case is at most

$$(1-\delta)^{1/\alpha} \leq e^{-\delta/\alpha} \leq 1 - \frac{\delta}{2\alpha},$$

ans thus, with probability at least $\frac{\delta}{2\alpha}$, algorithm $\mathcal{A}$ outputs a Δ-sensitive query q s.t. $q(\mathbf{X}) - q(\mathcal{U}) \geq \alpha\Delta n$. □

In particular, using Lemma 3.7.1 with $\alpha = \varepsilon$ shows that the parameters in Theorem 3.7.2 are tight.

Chapter 4

Bounds for Differential Privacy

4.1 Introduction

In this chapter we consider the tradeoff between privacy and utility when answering simple queries about a sensitive dataset. Specifically, we consider the *sample complexity* of differential private answers for *one-way marginals*—the minimum number of records n that is sufficient in order to publicly release a given set of statistics about the dataset, while achieving both differential privacy and accuracy.

The sample complexity of achieving pure differential privacy is well known for many settings (e.g. [HT10]). The more general setting of *approximate differential privacy* is less well understood. Recently, Bun, Ullman, and Vadhan [BUV14] showed how to prove strong lower bounds for approximate differential privacy that are optimal up to polylogarithmic factors for $\delta \approx 1/n$, which is essentially the weakest privacy guarantee that is still meaningful.[1] Since δ bounds the probability of a complete privacy breach, we would like δ to be very small. Thus we would like

[1] When $\delta \geq 1/n$ there are algorithms that are intuitively not private, yet satisfy $(0, \delta)$-differential privacy.

to quantify the cost (in terms of sample complexity) as $\delta \to 0$. In this chapter we give lower bounds for approximately differentially private algorithms that are nearly optimal for every choice of δ, and smoothly interpolate between pure and approximate differential privacy.

In particular, we consider algorithms that compute the *one-way marginals of the dataset*—an extremely simple and fundamental family of queries. For a dataset $x \in \{\pm 1\}^{n \times d}$, the d one-way marginals are simply the mean of the bits in each of the d columns. Formally, we define

$$\overline{x} := \frac{1}{n} \sum_{i=1}^{n} x_i \in [\pm 1]^d$$

where $x_i \in \{\pm 1\}^d$ is the i-th row of x. A mechanism M is said to be *accurate* if, on input x, its output is "close to" $\overline{x}$. Accuracy may be measured in a *worst-case* sense—i.e. $\|M(x) - \overline{x}\|_\infty \leq \alpha$, meaning every one-way marginal is answered with accuracy α—or in an *average-case* sense—i.e. $\|M(x) - \overline{x}\|_1 \leq \alpha d$, meaning the marginals are answered with average accuracy α.

Some of the earliest results in differential privacy [DN03, DN04, BDMN05, DMNS06] give a simple (ε, δ)-differentially private algorithm—the *Laplace mechanism*—that computes the one-way marginals of $x \in \{\pm 1\}^{n \times d}$ with average error α as long as

$$n \geq O\left(\min\left\{\frac{\sqrt{d \log(1/\delta)}}{\varepsilon\alpha}, \frac{d}{\varepsilon\alpha}\right\}\right). \tag{4.1}$$

The previous best lower bounds are $n \geq \Omega(d/\varepsilon\alpha)$ [HT10] for pure differential privacy and $n \geq \tilde{\Omega}(\sqrt{d}/\varepsilon\alpha)$ for approximate differential privacy with $\delta = o(1/n)$ [BUV14]. Our main result is an optimal lower bound that combines the previous lower bounds.

Theorem 4.1.1 (Main Theorem). *For every $\varepsilon, \delta, \alpha \in (0, 0.1)$ and $n, d \in \mathbb{N}$ the fol-*

lowing holds. Let $M : \{\pm 1\}^{n \times d} \to [\pm 1]^d$ *be* (ε, δ)*-differentially private. Suppose* $\mathop{\mathbb{E}}_{M}[\|M(x) - \bar{x}\|_1] \leq \alpha d$ *for all* $x \in \{\pm 1\}^{n \times d}$*. If* $e^{-\alpha\varepsilon n/5} \leq \delta \leq \varepsilon/(250n)^{1.1}$*„ then*

$$n \geq \Omega\left(\frac{\sqrt{d \log(1/\delta)}}{\varepsilon\alpha}\right).$$

Although there has been a long line of work developing methods to prove lower bounds in differential privacy (see [DN03, DMT07, DY08, KRSU10, HT10, NTZ13, BUV14] for a representative, but not exhaustive, sample), our result is the first to show that the sample complexity must grow by a multiplicative factor of $\sqrt{\log(1/\delta)}$.

We also remark that the assumption on the range of δ is necessary: The Laplace mechanism gives accuracy α and satisfies $(\varepsilon, 0)$-differential privacy when $n \geq O(d/\varepsilon\alpha)$. When $\delta = 2^{-\Theta(d)}$, our lower bound matches, up to constants, the upper bound of the Laplace mechanism for $\delta = 0$. So the lower bound cannot be strengthened for smaller values of δ. On the other hand, randomly sampling $O(1/\alpha^2)$ rows from the dataset and outputting the average of those rows gives accuracy α and satisfies $(0, O(1/n\alpha^2))$-differential privacy. So for $\delta \approx 1/n$ we cannot hope to prove lower bounds that grow with d or $1/\varepsilon$.

Lower bounds for answering one-way marginals have been shown to imply lower bounds for fundamental problems such as private convex empirical risk minimisation [BST14] and private principle component analysis [DTTZ14]. Our new lower bound for one-way marginals thus implies similar new lower bounds for these problems.

Finally, our techniques yield a simple alternative proof that $n \geq \Omega(d/\varepsilon\alpha)$ is necessary to achieve pure differential privacy while satisfying the accuracy condition in Theorem 4.1.1. We present this proof as a warmup in Section 4.3.1.

4.1.1 New Algorithms for Maximum Error

Our lower bound holds for mechanisms that bound the average error over the queries (we denote this as L_1 error). Thus, it also holds for algorithms that bound the maximum error over the queries (we denote this as L_∞ error). The Laplace mechanism gives a matching upper bound for average error. In many cases bounds on the maximum error are preferable. For maximum error, the sample complexity of the best previous mechanisms degrades by an additional $(\log d)^{\Omega(1)}$ factor compared to (4.1).

Surprisingly, this degradation is not necessary. We present algorithms that answer every one-way marginal with α accuracy and improve on the sample complexity of the Laplace mechanism by roughly a $\log d$ factor. These algorithms demonstrate that the widely used technique of adding independent noise to each query is suboptimal when the goal is to achieve worst-case error guarantees.

Our algorithm for pure differential privacy satisfies the following.

Theorem 4.1.2. *For every $\varepsilon, \alpha > 0$, $d \geq 1$, and $n \geq 4d/\varepsilon\alpha$, there exists an efficient mechanism $M : \{\pm 1\}^{n \times d} \to [\pm 1]^d$ that is $(\varepsilon, 0)$-differentially private and*

$$\forall x \in \{\pm 1\}^{n \times d} \quad \mathop{\mathbb{P}}_{M}\left[\|M(x) - \overline{x}\|_\infty \geq \alpha\right] \leq (2e)^{-d}.$$

In fact, the algorithm promised by Theorem 4.1.2 is oblivious, perturbing the answers with noise from a fixed distribution[2] and our algorithm is applicable to any set of d queries of sensitivity at most $2/n$, rather than just one-way marginals.

And our algorithm for approximate differential privacy is as follows.

---

[2]That is, $M(x)$ is simply $\overline{x} + Y$ (truncated to $[\pm 1]^d$), where Y is a single distribution and does not depend on x.

Theorem 4.1.3. *For every $\varepsilon, \delta, \alpha > 0$, $d \geq 1$, and*

$$n \geq O\left(\frac{\sqrt{d \cdot \log(1/\delta) \cdot \log\log d}}{\varepsilon\alpha}\right),$$

there exists an efficient mechanism $M : \{\pm 1\}^{n \times d} \to [\pm 1]^d$ that is (ε, δ)-differentially private and

$$\forall x \in \{\pm 1\}^{n \times d} \quad \mathop{\mathbb{P}}_{M}\left[\|M(x) - \overline{x}\|_\infty \geq \alpha\right] \leq \frac{1}{d^{\omega(1)}}.$$

The algorithm stipulated by Theorem 4.1.3 is also oblivious and is applicable to arbitrary low-sensitivity queries. The algorithm also satisfies concentrated differential privacy — in this setting there is no stark separation between the approximate and concentrated versions of differential privacy.

These algorithms improve over the sample complexity of the best known mechanisms for each privacy and accuracy guarantee by a factor of $(\log d)^{\Omega(1)}$. Namely, the Laplace mechanism requires $n \geq O(d \cdot \log d/\varepsilon\alpha)$ samples for pure differential privacy and the Gaussian mechanism requires $n \geq O(\sqrt{d \cdot \log(1/\delta) \cdot \log d}/\varepsilon\alpha)$ samples for approximate differential privacy. We conjecture that the Algorithm in Theorem 4.1.3 can be improved to match our lower bound — that is, we believe that the $\sqrt{\log\log d}$ factor is unnecessary.

4.1.2 Techniques

Lower Bounds: Our lower bound relies on techniques from the literature on *fingerprinting codes* [BS98]. Fingerprinting codes were originally developed in the cryptography literature for watermarking digital content, but several recent works have shown they are intimately connected to lower bounds for differential privacy and related learning problems [Ull13, BUV14, HU14, §6]. In particular, Bun et al. [BUV14] showed that fingerprinting codes can be used to construct an attack

Privacy	Accuracy	Type	Previous bound (n)	This work (n)
(ε, δ)	L_1 or L_∞	Lower	$\tilde{\Omega}\left(\frac{\sqrt{d}}{\alpha\varepsilon}\right)$ [BUV14]	$\Omega\left(\frac{\sqrt{d\log(1/\delta)}}{\alpha\varepsilon}\right)$
(ε, δ)	L_1	Upper	$O\left(\frac{\sqrt{d\cdot\log(1/\delta)}}{\alpha\varepsilon}\right)$ Gaussian	
(ε, δ)	L_∞	Upper	$O\left(\frac{\sqrt{d\cdot\log(1/\delta)\cdot\log d}}{\alpha\varepsilon}\right)$ Gaussian	$O\left(\frac{\sqrt{d\cdot\log(1/\delta)\cdot\log\log d}}{\varepsilon\alpha}\right)$
$(\varepsilon, 0)$	L_1 or L_∞	Lower	$\Omega\left(\frac{d}{\alpha\varepsilon}\right)$ [HT10]	
$(\varepsilon, 0)$	L_1	Upper	$O\left(\frac{d}{\alpha\varepsilon}\right)$ Laplace	
$(\varepsilon, 0)$	L_∞	Upper	$O\left(\frac{d\cdot\log d}{\alpha\varepsilon}\right)$ Laplace	$O\left(\frac{d}{\alpha\varepsilon}\right)$

Table 4.1: *Summary of sample complexity upper and lower bounds for privately answering d one-way marginals with L_1 error αd or L_∞ error α.*

demonstrating that any mechanism that accurately answers one-way marginals is not differentially private.

We do not use fingerprinting codes directly in our proof, but rather we extract some key ideas from their analysis. Fingerprinting codes are discussed in more detail in Chapter 6. The lower bounds in this Chapter can be viewed as a construction of a "weak" fingerprinting code.

The heart of our lower bound is a "correlation analysis," which we now describe.

For the special case of pure differential privacy, the proof proceeds as follows. First, we sample $x_1, x_1' \in \{\pm 1\}^d$ independently and uniformly at random and then set $x_1 = x_2 = \cdots = x_n$ and $x_1' = x_2' = \cdots = x_n'$. For a mechanism $M : \{\pm 1\}^{n \times d} \to [-1, 1]^d$, we consider the quantities $Z := \langle M(x), x_1 \rangle$ and $\hat{Z} := \langle M(x), x_1' \rangle$. If M is accurate, then $\mathbb{E}[Z] \geq \Omega(d)$. On the other hand, since $M(x)$ and x_1' are independent, we have that $\mathbb{E}\left[\hat{Z}\right] = 0$. By Hoeffding's inequality, we also have that $\mathbb{P}\left[\hat{Z} > \lambda\right] \leq 2^{-\Omega(\lambda^2/d)}$ for all $\lambda > 0$. If M is $(\varepsilon, 0)$-differentially private, then, by group privacy, $\mathbb{P}[Z > \lambda] \leq e^{n\varepsilon}\mathbb{P}\left[\hat{Z} > \lambda\right]$ – that is, since x and x'

differ in n elements, repeatedly applying the differential privacy guarantee yields the $e^{n\varepsilon}$ multiplicative bound (rather than e^{ε}, which would apply if x and x' differ in only one element). Appropriately combining these three facts yields the lower bound for pure differential privacy.

Now we describe the lower bound for approximate differential privacy, which uses ideas from the analysis of fingerprinting codes: First pick a random dataset $x_1, \cdots, x_n \in \{\pm 1\}^d$. Specifically, pick $p \in [-1,1]^d$ uniformly at random and then, conditioned on p, pick $x_1, \cdots, x_n \in \{\pm 1\}^d$ independently with $\mathbb{E}[x_i] = p$ for all $i \in [n]$. Then we argue that, if $M : \{\pm 1\}^{n \times d} \to [-1,1]^d$ is an accurate mechanism, there must be correlation between its input and output. Namely, defining $Z_i := \langle M(x), x_i - p \rangle$, we have

$$\sum_{i \in [n]} \mathbb{E}[Z_i] = \sum_{i \in [n]} \mathbb{E}[\langle M(x), x_i - p \rangle] \geq \Omega(d). \tag{4.2}$$

On the other hand, if the i^{th} element of x is removed or replaced — the result of which we denote by x^i — then $M(x^i)$ is independent of x_i (conditioned on p). Hence $\mathbb{E}\left[\langle M(x^i), x_i - p \rangle\right] = 0$. We name this quantity $\hat{Z}_i := \langle M(x^i), x_i - p \rangle$. In fact, by a Chernoff-Hoeffding bound, we can argue that, for each $i \in [n]$ and $\lambda > 0$,

$$\mathbb{P}\left[\hat{Z}_i > \lambda\right] = \mathbb{P}\left[\langle M(x^i), x_i - p \rangle > \lambda\right] \leq e^{-\Omega(\lambda^2/d)}. \tag{4.3}$$

Thus we have conflicting bounds: $\hat{Z}_i$ is small with high probability, while Z_i is large in expectation (for a random $i \in [n]$). Now the punchline: If M is (ε, δ)-differentially private, then $M(x)$ and $M(x^i)$ must be "close" and, hence, so must Z_i and $\hat{Z}_i$. Namely,

$$\forall \lambda \qquad \mathbb{P}[Z_i > \lambda] \leq e^{\varepsilon} \mathbb{P}\left[\hat{Z}_i > \lambda\right] + \delta. \tag{4.4}$$

Combining (4.2), (4.3), and (4.4) with some straightforward calculations yields the

lower bound $n \geq \Omega(\sqrt{d}/\varepsilon)$ for $\delta \approx 1/n$.

The final step is to interpolate between the lower bound for pure differential privacy ($\delta = 0$) and the "fingerprinting" lower bound ($\delta \approx 1/n$). To do so, we apply group privacy. For pure differential privacy we used group size n and for the fingerprinting lower bound we used group size 1. To interpolate, we use group size $k = \Theta(\log(1/n\delta))$. We will sample $x_1, \cdots, x_{n/k} \in \{\pm 1\}^d$ as in the fingerprinting lower bound bound. Then we repeat each item k times to obtain n items. Now we apply the same analysis as for the fingerprinting lower bound except now x^i will differ from x on k entries and we must apply group privacy in (4.4). When $k = n$ or $k = 1$ this reduces to the analyses described above and when $1 < k < n$ it provides intermediate lower bounds.

The key to the δ dependence in the lower bound is that smaller values of δ allow group privacy to be applied to larger group sizes while maintaining a strong enough privacy guarantee for the fingerprinting analysis to carry through.

Upper Bounds: Our algorithm for pure differential privacy and worst-case error is an instantiation of the exponential mechanism [MT07] using the L_∞ norm. That is, the mechanism samples $y \in \mathbb{R}^d$ with probability proportional to $\exp(-\eta \|y\|_\infty)$ and outputs $M(x) = \overline{x} + y$. In contrast, adding independent Laplace noise corresponds to using the exponential mechanism with the L_1 norm and adding independent Gaussian noise corresponds to using the exponential mechanism with the L_2 norm squared. Using this distribution turns out to give better tail bounds than adding independent noise.

For approximate differential privacy, we use a completely different algorithm. We start by adding independent Gaussian noise to each marginal. However, rather than using a union bound to show that each Gaussian error is small with high probability,

we argue that "most" errors are small. Namely, with the sample complexity that we allow M, we can ensure that all but a $1/\mathrm{polylog}(d)$ fraction of the errors are small with high probability. Now we "fix" the $d/\mathrm{polylog}(d)$ marginals that are bad. We repeatedly use the exponential mechanism [MT07] to find one bad error and the correct it by sampling fresh Gaussian noise. The key is that we only need to run this procedure $d/\mathrm{polylog}(d)$ times, which means we can afford the necessary sample complexity.

4.2 Preliminaries

A well known fact about differential privacy is that it generalises smoothly to datasets that differ on more than a single row. We say that two datasets $x, x' \in \{\pm 1\}^{n\times d}$ are *k-adjacent* if they differ by at most k rows, and we denote this by $x \sim_k x'$. The following statement is essentially folklore, and we refer the reader to [DR14] for a textbook proof.

Fact 4.2.1 (Group Differential Privacy)**.** *For every $k \geq 1$, if $M : \{\pm 1\}^{n\times d} \to \mathcal{R}$ is (ε, δ)-differentially private, then for every two k-adjacent datasets $x \sim_k x'$, and every subset $S \subseteq \mathcal{R}$,*

$$\mathbb{P}\left[M(x) \in S\right] \leq e^{k\varepsilon} \cdot \mathbb{P}\left[M(x') \in S\right] + \frac{e^{k\varepsilon} - 1}{e^{\varepsilon} - 1} \cdot \delta.$$

All of the upper and lower bounds for one-way marginals have a multiplicative $1/\alpha\varepsilon$ dependence on the accuracy α and the privacy loss ε. This is no coincidence, and follows from the following general statement, which is folklore.

Fact 4.2.2 (Dependence on α and ε)**.** *Let $\alpha, \varepsilon\delta \in (0, 1/10]$. Fix some norm $\|\cdot\|$.*

Suppose there exists a (ε, δ)-differentially private mechanism $M : \{\pm 1\}^{n\times d} \to [\pm 1]^d$

such that for every dataset $x \in \{\pm 1\}^{n \times d}$,

$$\mathop{\mathbb{E}}_{M}\left[\|M(x) - \overline{x}\|\right] \leq \alpha \left\|\vec{1}\right\|.$$

Then there exists a $(1, \delta/\varepsilon)$*-differentially private mechanism* $M' : \{\pm 1\}^{n' \times d} \to [\pm 1]^d$ *for* $n' = \lceil 20\alpha\varepsilon n \rceil$ *such that for every dataset* $x' \in \{\pm 1\}^{n' \times d}$,

$$\mathop{\mathbb{E}}_{M'}\left[\|M'(x') - \overline{x'}\|\right] \leq \frac{1}{10}\left\|\vec{1}\right\|.$$

This fact allows us to suppress the accuracy parameter α and the privacy loss ε when proving our lower bounds. Namely, if we prove a lower bound of $n' \geq n^*$ for all $(1, \delta)$-differentially private mechanisms $M' : \{\pm 1\}^{n' \times d} \to [\pm 1]^d$ with $\mathop{\mathbb{E}}_{M'}\left[\|M'(x') - \overline{x'}\|_p\right] \leq d^{1/p}/10$, then we obtain a lower bound of $n \geq \Omega(n^*/\alpha\varepsilon)$ for all $(\varepsilon, \varepsilon\delta)$-differentially private mechanisms $M : \{\pm 1\}^{n \times d} \to [\pm 1]^d$ with $\mathop{\mathbb{E}}_{M}\left[\|M(x) - \overline{x}\|_p\right] \leq \alpha d^{1/p}$. So we will simply fix the parameters $\alpha = 1/10$ and $\varepsilon = 1$ in our lower bounds.

Proof. Let $k = \lfloor \log(2)/\varepsilon \rfloor$. Given $x' \in \{\pm 1\}^{n' \times d}$, define $x \in \{\pm 1\}^{n \times d}$ to be k copies of x' followed by $n - kn'$ rows of all 1 entries. Then

$$n\overline{x} = n'\overline{x'} \cdot k + (n - kn')\vec{1} \quad \text{and} \quad \overline{x'} = \frac{n}{kn'}\overline{x} - \frac{n - kn'}{kn'}\vec{1}.$$

Define M' by

$$M'(x') = \frac{n}{kn'}M(x) - \frac{n - kn'}{kn'}\vec{1}.$$

Then

$$\mathop{\mathbb{E}}_{M'}\left[\left\|M'(x') - \overline{x'}\right\|\right] = \frac{n}{kn'}\mathop{\mathbb{E}}_{M}\left[\|M(x) - \overline{x}\|\right] \leq \frac{\alpha n}{kn'}\left\|\vec{1}\right\|.$$

Thus, if $n' \geq 20\alpha\varepsilon n$, we have $\mathop{\mathbb{E}}_{M'}\left[\left\|M'(x') - \overline{x'}\right\|\right] \leq \frac{1}{10}\left\|\vec{1}\right\|$, as $k \geq 1/2\varepsilon$. By Fact 4.2.1, M' is $\left(k\varepsilon, \frac{e^{k\varepsilon}-1}{e^{\varepsilon}-1} \cdot \delta\right)$-differentially private. By our choice of k, we have $k\varepsilon \leq \log 2 \leq 1$ and $\frac{e^{k\varepsilon}-1}{e^{\varepsilon}-1}\delta \leq \frac{e^{\log 2}-1}{\varepsilon}\delta = \delta/\varepsilon$, as required. □

4.3 Lower Bounds for Differential Privacy

4.3.1 Warmup: Lower Bound for Pure Differential Privacy

It is known [HT10] that any ε-differentially private mechanism that answers d one-way marginals requires $n \geq \Omega(d/\varepsilon)$ samples. We provide an alternative simple proof of this fact, which also serves as a warmup to our main proof.

Theorem 4.3.1. *Let $M : \{\pm 1\}^{n \times d} \to [\pm 1]^d$ be a ε-differentially private mechanism. Suppose*

$$\forall x \in \{\pm 1\}^{n \times d} \quad \mathop{\mathbb{E}}_{M} [\|M(x) - \overline{x}\|_1] \leq 0.9d$$

Then $n \geq \Omega(d/\varepsilon)$.

The proof uses a special case of Hoeffding's Inequality:

Lemma 4.3.2 (Hoeffding's Inequality)**.** *Let $X \in \{\pm 1\}^n$ be uniformly random and $a \in \mathbb{R}^n$ fixed. Then*

$$\mathop{\mathbb{P}}_{X} [\langle a, X \rangle > \lambda \|a\|_2] \leq e^{-\lambda^2/2}$$

for all $\lambda \geq 0$.

Proof of Theorem 4.3.1. Let $x_1, x_1' \in \{\pm 1\}^d$ be independent and uniform. Let $x \in \{\pm 1\}^{n \times d}$ be n copies of x_1 and, likewise, let $x' \in \{\pm 1\}^{n \times d}$ be n copies of x_1'. Let $Z = \langle M(x), x_1 \rangle$ and $\hat{Z} = \langle M(x'), x_1 \rangle$.

Now we give conflicting tail bounds for Z and $\hat{Z}$, which we can relate by privacy.

By our hypothesis and Markov's inequality,

$$\begin{aligned}
\mathbb{P}\left[Z \leq d/20\right] =& \mathbb{P}\left[\langle M(x), x_1\rangle \leq 0.05d\right] \\
=& \mathbb{P}\left[\langle \overline{x}, x_1\rangle - \langle \overline{x} - M(x), x_1\rangle \leq 0.05d\right] \\
=& \mathbb{P}\left[\langle \overline{x} - M(x), x_1\rangle \geq 0.95d\right] \\
\leq& \mathbb{P}\left[\left\|\overline{x} - M(x)\right\|_1 \geq 0.95d\right] \\
\leq& \frac{\mathbb{E}\left[\left\|\overline{x} - M(x)\right\|_1\right]}{0.95d} \\
\leq& \frac{0.9}{0.95} < 0.95.
\end{aligned}$$

Since $M(x')$ is independent from x_1, we have

$$\forall \lambda \geq 0 \quad \mathbb{P}\left[\hat{Z} > \lambda\sqrt{d}\right] \leq \mathbb{P}\left[\langle M(x'), x_1\rangle > \lambda \left\|M(x')\right\|_2\right] \leq e^{-\lambda^2/2},$$

by Lemma 4.3.2. In particular, setting $\lambda = \sqrt{d}/20$ gives $\mathbb{P}\left[Z' > d/20\right] \leq e^{-d/800}$.

Now x and x' are datasets that differ in n rows, so group privacy implies that

$$\mathbb{P}\left[M(x) \in S\right] \leq e^{n\varepsilon}\mathbb{P}\left[M(x') \in S\right]$$

for all S. Thus

$$\frac{1}{20} < \mathbb{P}\left[Z > \frac{d}{20}\right] = \mathbb{P}\left[M(x) \in S_x\right] \leq e^{n\varepsilon}\mathbb{P}\left[M(x') \in S_x\right] = e^{n\varepsilon}\mathbb{P}\left[\hat{Z} > \frac{d}{20}\right] \leq e^{n\varepsilon}e^{-d/800},$$

where

$$S_x = \left\{y \in [\pm 1]^d : \langle y, x_1\rangle > \frac{d}{20}\right\}.$$

Rearranging $1/20 < e^{n\varepsilon}e^{-d/800}$, gives

$$n > \frac{d}{800\varepsilon} - \frac{\log(20)}{\varepsilon},$$

as required. □

4.3.2 Basic Lower Bound for Approximate Differential Privacy

We first show the basic version of the lower bound for approximate differential privacy:

Theorem 4.3.3. *Let $M : \{\pm 1\}^{n\times d} \to [\pm 1]^d$ be a $(\varepsilon = 1, \delta = 1/30n)$-differentially private mechanism. Suppose*

$$\forall x \in \{\pm 1\}^{n\times d} \quad \mathop{\mathbb{E}}_{M}[\|M(x) - \overline{x}\|_1] \leq \frac{d}{10}.$$

Then $n \geq \sqrt{d}/41$.

The key technical lemma is the fingerprinting lemma, which we prove in the next section.

Lemma 4.3.4 (Fingerprinting Lemma)**.** *Let $f : \{\pm 1\}^n \to [\pm 1]$. Then*

$$\mathop{\mathbb{E}}_{p\in[\pm 1], x_{1\cdots n}\sim p}\left[f(x)\cdot\sum_{i\in[n]}(x_i - p) + 2\,|f(x) - \overline{x}|\right] \geq \frac{1}{3}.$$

Proof of Theorem 4.3.3. First we define the input distribution: Sample $p \in [\pm 1]^d$ uniformly at random. Then sample $x \in \{\pm 1\}^{n\times d}$ from the product distribution with $\mathbb{E}\left[x_{i,j}\right] = p$ for all $i \in [n]$ and $j \in [d]$. For $i \in [n]$ define $x^i \in \{\pm 1\}^{n\times d}$ to be x with the i^{th} row resampled. Thus x and x^i are neighbouring and have identical marginal distributions. Define

$$Z_i = \langle M(x), x_i - p\rangle \qquad \text{and} \qquad Z = \frac{1}{n}\sum_{i\in[n]} Z_i = \langle M(x), \overline{x} - p\rangle,$$

where x_i is the i^{th} row of x. Also, define

$$\hat{Z}_i = \langle M(x^i), x_i - p\rangle.$$

First we show that privacy gives an upper bound on the expectation of Z:

Claim 4.3.5. *For all $i \in [n]$, $\mathbb{E}[|Z_i|] \le e^\varepsilon \sqrt{d} + \delta 2d$.*

Proof. Differential privacy implies that the distribution of Z_i is close to the distribution of $\hat{Z}_i$. In particular, since $|Z_i| \le 2d$,

$$\mathbb{E}[|Z_i|] = \int_0^{2d} \mathbb{P}[|Z_i| > z]\,\mathrm{d}z \le \int_0^{2d} e^\varepsilon \mathbb{P}\left[|\hat{Z}_i| > z\right] + \delta \mathrm{d}z = e^\varepsilon \mathbb{E}\left[|\hat{Z}_i|\right] + \delta 2d.$$

Thus we need only show that $\mathbb{E}\left[|\hat{Z}_i|\right] \le \sqrt{d}$. Consider a fixed value of p. Then $M(x^i)$ and $x_i - p$ are independent and $\mathbb{E}[x_i - p] = 0$, whence $\mathbb{E}\left[\hat{Z}_i\right] = 0$. For all $j \in [d]$, we have $\mathbb{E}\left[(x_{i,j} - p_j)^2\right] \le 1$. Since the entries of x_i are independent, we conclude that $\mathsf{Var}\left[\hat{Z}_i\right] \le d$. Thus

$$\mathbb{E}\left[|\hat{Z}_i|\right] \le \sqrt{\mathbb{E}\left[\hat{Z}_i^2\right]} = \sqrt{\mathsf{Var}\left[\hat{Z}_i\right] + \mathbb{E}\left[\hat{Z}_i\right]^2} \le \sqrt{d}.$$

□

Now accuracy will give us a lower bound on the expectation of Z:

Claim 4.3.6.

$$n\mathbb{E}[Z] + 2\mathbb{E}\left[\|M(x) - \overline{x}\|_1\right] \ge \frac{d}{3}.$$

Proof. By linearity of expectations, it suffices to show that, for all $j \in [d]$,

$$\mathbb{E}\left[nM(x)_j(\overline{x}_j - p_j) + 2|M(x)_j - \overline{x}_j|\right] \ge \frac{1}{3}.$$

We condition on the randomness of M and the values of $p_{j'}$ and $x_{i',j'}$ for all i and $j' \ne j$. Now the only randomness left is the choice of p_j and the j^{th} column of x. So (noting that M does not have access to p_j) we can write $M(x)_j = f(x_{1,j}, \cdots, x_{n,j})$ for some $f : \{\pm 1\}^n \to [-1, 1]$. Now we can apply the fingerprinting lemma to f to obtain the desired result. □

Since $n\mathbb{E}[Z] = \sum_{i\in[n]} Z_i$, we can combine the two claims to obtain

$$n\left(e^{\varepsilon}\sqrt{d} + \delta 2d\right) + 2\mathbb{E}\left[\|M(x) - \overline{x}\|_1\right] \geq \frac{d}{3}.$$

Substituting $\mathbb{E}\left[\|M(x) - \overline{x}\|_1\right] \leq d/10$, $\varepsilon = 1$, and $\delta = 1/30n$ gives $n \geq \sqrt{d}/15e$, as required. □

4.3.3 The Fingerprinting Lemma

In this section we prove the fingerprinting lemma. This section contains the key techniques that underlie fingerprinting codes [Tar08]. These ideas will appear again in Chapters 5 and 6.

The proof is broken into several lemmas. The first statement is somewhat mysterious but extremely powerful.

Lemma 4.3.7. *Let $f : \{\pm 1\}^n \to \mathbb{R}$. Define $g : [\pm 1] \to \mathbb{R}$ by*

$$g(p) = \mathop{\mathbb{E}}_{x_{1\cdots n}\sim p}\left[f(x)\right].$$

Then

$$\mathop{\mathbb{E}}_{x_{1\cdots n}\sim p}\left[f(x)\cdot\sum_{i\in[n]}(x_i - p)\right] = g'(p)\cdot(1-p^2).$$

How should we interpret this statement? Firstly $\mathop{\mathbb{E}}_{x_{1\cdots n}\sim p}\left[f(x)\cdot\sum_{i\in[n]}(x_i-p)\right]$ is the correlation between the input and output of f, which is the quantity we want to understand. The function g is a "symmetrisation" of f; it captures the interesting aspects (for our purposes) of f. In particular, $g(1) = f(1,1,\cdots,1)$ and $g(-1) = f(-1,-1,\cdots,-1)$. So as p varies from -1 to 1, $g(p)$ captures how f varies over inputs of varying Hamming weight. The derivative $g'(p)$ captures how rapidly f responds to a change in its input. Namely, $g'(p)$ represents how sensitive f is when a -1 is changed to a $+1$ on an input where the average of the bits is

about p. The Lemma statement is thus simply equating the correlation with this measure of the sensitivity of f.

Proof. Since $x^2 = 1$ for $x \in \{\pm 1\}$, we have the identity

$$\frac{\mathrm{d}}{\mathrm{d}p}\frac{1+xp}{2} = \frac{x}{2} = \frac{1+xp}{2}\frac{x-p}{1-p^2}$$

for all $x \in \{\pm 1\}$ and $p \in (-1, 1)$. By the product rule, we have

$$\frac{\mathrm{d}}{\mathrm{d}p}\prod_{i\in[n]}\frac{1+x_ip}{2} = \sum_{i\in[n]}\left(\frac{\mathrm{d}}{\mathrm{d}p}\frac{1+x_ip}{2}\right)\prod_{k\in[n]\setminus\{i\}}\frac{1+x_kp}{2} = \sum_{i\in[n]}\frac{x_i-p}{1-p^2}\prod_{k\in[n]}\frac{1+x_kp}{2}$$

for all $x \in \{\pm 1\}^n$ and $p \in (-1, 1)$. Sampling $x \sim p$ samples each $x \in \{\pm 1\}$ with probability $\frac{1+xp}{2}$. Thus sampling $x_{1\cdots n} \sim p$, samples each $x \in \{\pm 1\}^n$ with probability $\prod_{i\in[n]}\frac{1+x_ip}{2}$.

Now we can write

$$g(p) = \mathop{\mathbb{E}}_{x_{1\cdots n}\sim p}[f(x)] = \sum_{x\in\{\pm1\}^n} f(x)\prod_{i\in[n]}\frac{1+x_ip}{2}.$$

Using the above identities gives

$$\begin{aligned} g'(p) &= \sum_{x\in\{\pm1\}^n} f(x)\frac{\mathrm{d}}{\mathrm{d}p}\prod_{i\in[n]}\frac{1+x_ip}{2} \\ &= \sum_{x\in\{\pm1\}^n} f(x)\sum_{i\in[n]}\frac{x_i-p}{1-p^2}\prod_{k\in[n]}\frac{1+x_kp}{2} \\ &= \mathop{\mathbb{E}}_{x_{1\cdots n}\sim p}\left[f(x)\sum_{i\in[n]}\frac{x_i-p}{1-p^2}\right] \end{aligned}$$

□

The previous lemma considers a fixed value of p, whereas the next lemma takes an average over p. In particular, it gives an expression for the average of the

derivative $g'(p)$.

Lemma 4.3.8. *Let $g : [\pm 1] \to \mathbb{R}$ be a polynomial. Then*

$$\mathop{\mathbb{E}}_{p\in[\pm 1]}\left[g'(p)\cdot(1-p^2)\right] = 2\mathop{\mathbb{E}}_{p\in[\pm 1]}\left[g(p)\cdot p\right]. \tag{4.5}$$

If we discard the factor $(1-p^2)$, we would have

$$\mathop{\mathbb{E}}_{p\in[\pm 1]}\left[g'(p)\right] = \frac{1}{2}\int_{-1}^{+1} g'(p)\mathrm{d}p = \frac{g(1)-g(-1)}{2} \tag{4.6}$$

by the fundamental theorem of calculus. The factor $(1-p^2)$ simply yields a "smoothed" version of this simpler bound: the right hand side of (4.5) is the function $g(p)\cdot p$ averaged over the interval $[-1,1]$, whereas without the $(1-p^2)$ factor the right hand side of (4.6) is the average of the function $g(p)\cdot p$ over the endpoints $\{-1,1\}$.

Proof. Let $u(p) = 1-p^2$. By integration by parts and the fundamental theorem of calculus,

$$\begin{aligned}
\mathop{\mathbb{E}}_{p\in[\pm 1]}\left[g'(p)\cdot(1-p^2)\right] =& \frac{1}{2}\int_{-1}^{1} g'(p)(1-p^2)\mathrm{d}p \\
=& \frac{1}{2}\int_{-1}^{1} g'(p)u(p)\mathrm{d}p \\
=& \frac{1}{2}\int_{-1}^{1}\left(\frac{\mathrm{d}}{\mathrm{d}p}g(p)u(p)\right) - g(p)u'(p)\mathrm{d}p \\
=& \frac{1}{2}\left(g(1)u(1)-g(-1)u(-1)\right) - \frac{1}{2}\int_{-1}^{1} g(p)(-2p)\mathrm{d}p \\
=& 0 + \int_{-1}^{1} g(p)p\mathrm{d}p \\
=& 2\mathop{\mathbb{E}}_{p\in[\pm 1]}\left[g(p)\cdot p\right].
\end{aligned}$$

□

Combining Lemmas 4.3.7 and 4.3.8 yields the following version of the finger-

printing lemma.

Proposition 4.3.9. *Let $f : \{\pm 1\}^n \to \mathbb{R}$. Then*

$$\mathop{\mathbb{E}}_{p\in[\pm 1],x_{1\cdots n}\sim p}\left[f(x)\cdot\sum_{i\in[n]}(x_i-p)+(f(x)-p)^2\right]\geq\frac{1}{3}.$$

Proof. Define $g : [\pm 1] \to \mathbb{R}$ by

$$g(p)=\mathop{\mathbb{E}}_{x_{1\cdots n}\sim p}\left[f(x)\right].$$

By Lemmas 4.3.7 and 4.3.8,

$$\mathop{\mathbb{E}}_{p\in[\pm 1],x_{1\cdots n}\sim p}\left[f(x)\cdot\sum_{i\in[n]}(x_i-p)\right]=\mathop{\mathbb{E}}_{p\in[\pm 1]}\left[2g(p)p\right].$$

Moreover,

$$\mathop{\mathbb{E}}_{p\in[\pm 1],x_{1\cdots n}\sim p}\left[(f(x)-p)^2\right]=\mathop{\mathbb{E}}_{p\in[\pm 1],x_{1\cdots n}\sim p}\left[f(x)^2-2g(p)p+p^2\right]\geq 0-\mathop{\mathbb{E}}_{p\in[\pm 1]}\left[2g(p)p\right]+\frac{1}{3}.$$

□

An additional application of Jensen's inequality yields a slightly different statement:

Proposition 4.3.10. *Let $f : \{\pm 1\}^n \to \mathbb{R}$. Then*

$$\mathop{\mathbb{E}}_{p\in[\pm 1],x_{1\cdots n}\sim p}\left[f(x)\cdot\sum_{i\in[n]}(x_i-p)+(f(x)-\overline{x})^2\right]\geq\frac{1}{3}.$$

Proof. Define $g : [\pm 1] \to \mathbb{R}$ by

$$g(p)=\mathop{\mathbb{E}}_{x_{1\cdots n}\sim p}\left[f(x)\right].$$

By Lemmas 4.3.7 and 4.3.8,

$$\mathop{\mathbb{E}}_{p\in[\pm1],x_{1\cdots n}\sim p}\left[f(x)\cdot\sum_{i\in[n]}(x_i-p)\right]=\mathop{\mathbb{E}}_{p\in[\pm1]}\left[2g(p)p\right].$$

Moreover, by Jensen's inequality

$$\begin{aligned}\mathop{\mathbb{E}}_{p\in[\pm1],x_{1\cdots n}\sim p}\left[(f(x)-\overline{x})^2\right]&\geq\mathop{\mathbb{E}}_{p\in[\pm1]}\left[\left(\mathop{\mathbb{E}}_{x_{1\cdots n}\sim p}\left[f(x)-\overline{x}\right]\right)^2\right]\\&=\mathop{\mathbb{E}}_{p\in[\pm1]}\left[(g(p)-p)^2\right]\\&=\mathop{\mathbb{E}}_{p\in[\pm1]}\left[g(p)^2-2g(p)p+p^2\right]\\&=\mathop{\mathbb{E}}_{p\in[\pm1]}\left[g(p)^2\right]-\mathop{\mathbb{E}}_{p\in[\pm1],x_{1\cdots n}\sim p}\left[f(x)\cdot\sum_{i\in[n]}(x_i-p)\right]+\frac{1}{3}.\end{aligned}$$

Thus

$$\mathop{\mathbb{E}}_{p\in[\pm1],x_{1\cdots n}\sim p}\left[f(x)\cdot\sum_{i\in[n]}(x_i-p)+(f(x)-\overline{x})^2\right]\geq\mathop{\mathbb{E}}_{p\in[\pm1]}\left[g(p)^2\right]+\frac{1}{3}\geq\frac{1}{3}.$$

□

Finally it remains to replace ℓ_2 error with ℓ_1 error to obtain the desired statement:

Corollary 4.3.11 (Fingerprinting Lemma)**.** *Let $f:\{\pm1\}^n\to[\pm1]$. Then*

$$\mathop{\mathbb{E}}_{p\in[\pm1],x_{1\cdots n}\sim p}\left[f(x)\cdot\sum_{i\in[n]}(x_i-p)+2\left|f(x)-\overline{x}\right|\right]\geq\frac{1}{3}.$$

Proof. Since $|f(x)-\overline{x}|\leq 2$, we have $|f(x)-\overline{x}|^2\leq 2\,|f(x)-\overline{x}|$, which implies the result. □

4.3.4 The Full Lower Bound for Approximate Differential Privacy

First we prove a technical lemma similar to Claim 4.3.5.

Lemma 4.3.12. *Let $M : \left(\{\pm 1\}^d\right)^n \to [\pm 1]^d$ be (ε, δ)-differentially private. Let $p \in [\pm 1]^d$ and let $x_{1\cdots n} \sim p$. Then, for all $i \in [n]$,*

$$\mathop{\mathbb{E}}_{x}\left[\langle M(x), x_i - p\rangle\right] \leq \sqrt{2d\log(1/\delta)} + 2d \cdot (e^{\varepsilon} + 1)\delta.$$

Proof. Let x' be x with x_i resampled. Then $M(x')$ and x_i are independent. By Hoeffding's inequality,

$$\mathop{\mathbb{P}}_{x'}\left[\langle M(x'), x_i - p\rangle \geq \lambda\right] \leq e^{-\lambda^2/2d}$$

for all $\lambda > 0$. Since x and x' are neighbouring,

$$\mathop{\mathbb{P}}_{x}\left[\langle M(x), x_i - p\rangle \geq \lambda\right] \leq e^{\varepsilon - \lambda^2/2d} + \delta$$

for all $\lambda > 0$. Setting $\lambda = \sqrt{2d\log(1/\delta)}$ and using the fact that $\langle M(x), x_i - p\rangle \leq 2d$, we have

$$\mathop{\mathbb{E}}_{x}\left[\langle M(x), x_i - p\rangle\right] \leq \lambda + 2d \cdot \mathop{\mathbb{P}}_{x}\left[\langle M(x), x_i - p\rangle \geq \lambda\right] \leq \sqrt{2d\log(1/\delta)} + 2d \cdot (e^{\varepsilon} + 1)\delta.$$

□

Now we prove an intermediate version of our main lower bound.

Theorem 4.3.13. *Let $M : \left(\{\pm 1\}^d\right)^n \to [\pm 1]^d$ be (ε, δ)-differentially private. Suppose that, for all $x \in \left(\{\pm 1\}^d\right)^n$, we have*

$$\mathbb{E}\left[\|M(x) - \overline{x}\|_2^2\right] \leq \alpha^2 d.$$

If $\alpha^2 \leq 1/4$ and $(e^{\varepsilon} + 1)n\delta \leq 1/50$, then

$$n \geq \frac{\sqrt{d}}{25\sqrt{2\log(1/\delta)}}.$$

Proof. Let $p \in [\pm 1]^d$ be uniformly random and let $x_{1\cdots n} \sim p$. By Lemma 4.3.12 and

linearity of expectations,

$$\mathop{\mathbb{E}}_{x}\left[\left\langle M(x), \sum_{i\in[n]} x_i - p\right\rangle\right] \le n\cdot\left(\sqrt{2d\log(1/\delta)} + 2d\cdot(e^{\varepsilon}+1)\delta\right).$$

On the other hand, by Proposition 4.3.10 and linearity of expectations,

$$\mathop{\mathbb{E}}_{p,x}\left[\left\langle M(x), \sum_{i\in[n]} x_i - p\right\rangle + \|M(x)-\overline{x}\|_2^2\right] \ge \frac{d}{3}.$$

Thus

$$\frac{d}{3} - \alpha^2 d \le n\cdot\left(\sqrt{2d\log(1/\delta)} + 2d\cdot(e^{\varepsilon}+1)\delta\right).$$

rearranging and dividing by $\sqrt{d}$ gives

$$\sqrt{d}\cdot\left(\frac{1}{3} - \alpha^2 - 2(e^{\varepsilon}+1)n\delta\right) \le n\cdot\sqrt{2\log(1/\delta)}.$$

By our assumptions, $1/3 - \alpha^2 - 2(e^{\varepsilon}+1)n\delta \ge 1/25$. The result follows. □

Now we prove the main lower bound. Note that, by Fact 4.2.2, it suffices to consider a fixed value of α and ε.

Theorem 4.3.14. *Let $M : \left(\{\pm1\}^d\right)^n \to [\pm1]^d$ be $(1,\delta)$-differentially private. Suppose that, for all $x \in \left(\{\pm1\}^d\right)^n$, we have*

$$\mathbb{E}\left[\|M(x)-\overline{x}\|_1\right] \le \frac{d}{10}.$$

If $e^{-n/100} \le \delta \le (1/250n)^{1.1}$, then

$$n \ge \frac{\sqrt{d\log(1/\delta)}}{800}.$$

Proof. Let $M : \{\pm1\}^{n\times d} \to [\pm1]^d$ be a $(1,\delta)$-differentially private mechanism such that

$$\forall x \in \{\pm1\}^{n\times d} \quad \mathop{\mathbb{E}}_{M}\left[\|M(x)-\overline{x}\|_1\right] \le \frac{d}{10}. \tag{4.7}$$

Let k be an integer parameter to be chosen later. Let $n_k = \lfloor n/k \rfloor$. Let $M_k : \{\pm 1\}^{n_k \times d} \to [\pm 1]^d$ be the following mechanism. On input $x^* \in \{\pm 1\}^{n_k \times d}$, M_k creates $x \in \{\pm 1\}^{n \times d}$ by taking k copies of x^* and filling the remaining entries with 1s. Then M_k runs M on x and outputs $M(x)$.

By group privacy (Fact 4.2.1), M_k is a $\left(\varepsilon_k = k, \delta_k = \frac{e^k - 1}{e-1}\delta\right)$-differentially private mechanism. By the triangle inequality,

$$\left\| M_k(x^*) - \overline{x^*} \right\|_1 \le \left\| M(x) - \overline{x} \right\|_1 + \left\| \overline{x} - \overline{x^*} \right\|_1. \tag{4.8}$$

Now

$$\overline{x}_j = \frac{k \cdot n_k}{n} \overline{x_j^*} + \frac{n - k \cdot n_k}{n} 1.$$

Thus

$$\left| \overline{x}_j - \overline{x_j^*} \right| = \left| \left(\frac{k \cdot n_k}{n} - 1 \right) \overline{x_j^*} + \frac{n - k \cdot n_k}{n} \right| = \frac{n - k \cdot n_k}{n} \left| 1 - \overline{x_j^*} \right| \le 2 \frac{n - k \cdot n_k}{n}.$$

We have

$$\frac{n - k \cdot n_k}{n} = \frac{n - k\lfloor n/k \rfloor}{n} \le \frac{n - k(n/k - 1)}{n} = \frac{k}{n}.$$

Thus $\left\| \overline{x} - \overline{x^*} \right\|_1 \le 2dk/n$. Assume $k \le n/200$. Thus $\left\| \overline{x} - \overline{x^*} \right\|_1 \le d/100$. By (4.7), (4.8), and Hölder's inequality,

$$\mathop{\mathbb{E}}_{M_k} \left[\left\| M_k(x^*) - \overline{x^*} \right\|_2^2 \right] \le \mathop{\mathbb{E}}_{M_k} \left[\left\| M_k(x^*) - \overline{x^*} \right\|_1 \cdot \left\| M_k(x^*) - \overline{x^*} \right\|_\infty \right] \le \left(\frac{d}{10} + \frac{d}{100} \right) \cdot 2 \le \frac{d}{4}. \tag{4.9}$$

Now we can apply Theorem 4.3.13 to M_k: If

$$(e^{\varepsilon_k} + 1) n \delta_k = (e^k + 1) n \frac{e^k - 1}{e - 1} \delta \le \frac{1}{50}, \tag{4.10}$$

then

$$\frac{n}{k} \ge n_k \ge \frac{\sqrt{d}}{25\sqrt{2 \log(1/\delta)}}. \tag{4.11}$$

Now we pick the largest value of k satisfying (4.10) (and the assumption $k \le n/200$)

and obtain a lower bound on n from (4.11).

Rearranging (4.10) gives $e^{2k} - 1 \leq \frac{e-1}{50n\delta}$. We set

$$k = \left\lfloor \min\left\{ \frac{n}{200}, \frac{1}{2}\log\left(1 + \frac{e-1}{50n\delta}\right)\right\}\right\rfloor.$$

Assuming $\delta \geq e^{-n/100}$ implies that the second term in the minimum dominates and $k \geq \lfloor \log((e-1)/50n\delta)/2 \rfloor \geq \frac{1}{2}\log(1/\delta) - \frac{1}{2}\log(250n)$. Assuming $\delta \leq (1/250n)^{1.1}$ implies that $k \geq \frac{1}{22}\log(1/\delta)$. Thus (4.11) yields

$$n \geq \frac{k\sqrt{d}}{25\sqrt{2\log(1/\delta)}} \geq \frac{\sqrt{d\log(1/\delta)}}{800}.$$

□

Now we can prove the main theorem with the dependence on ε and α:

Proof Theorem 4.1.1. Fix a (ε, δ)-differentially private mechanism $M : \{\pm 1\}^{n\times d} \to [\pm 1]^d$ such that for every dataset $x \in \{\pm 1\}^{n\times d}$,

$$\mathop{\mathbb{E}}_{M}\left[\|M(x) - \overline{x}\|_1\right] \leq \alpha d.$$

By Fact 4.2.2, there exists a $(1, \delta/\varepsilon)$-differentially private mechanism $M' : \{\pm 1\}^{n'\times d} \to [\pm 1]^d$ for $n' = \lceil 20\alpha\varepsilon n\rceil$ such that for every dataset $x' \in \{\pm 1\}^{n'\times d}$,

$$\mathop{\mathbb{E}}_{M'}\left[\|M'(x') - \overline{x'}\|_1\right] \leq \frac{d}{10}.$$

Applying Theorem 4.3.14 to M' implies that, if $e^{-n'/100} \leq \delta/\varepsilon \leq (1/250n')^{1.1}$, then

$$n' = \lceil 20\alpha\varepsilon n\rceil \geq \frac{\sqrt{d\log(1/\delta)}}{800}.$$

Thus, if $e^{-\alpha\varepsilon n/5} \leq \delta \leq \varepsilon/(250n)^{1.1}$, then $n \geq \sqrt{d\log(1/\delta)}/16000\alpha\varepsilon - 1/20\alpha\varepsilon$. Note that the requirement $e^{-\alpha\varepsilon n/5} \leq \delta$ only becomes tight when $n\alpha\varepsilon = \Theta(d)$, so we can substitute $e^{-\Omega(d)} \leq \delta$ in place of this requirement. □

4.4 New Mechanisms for L_∞ Error

Adding independent noise seems very natural for one-way marginals, but it is suboptimal if one is interested in worst-case (i.e. L_∞) error bounds, rather than average-case (i.e. L_1) error bounds.

4.4.1 Pure Differential Privacy

Theorem 4.1.2 follows from Theorem 4.4.1. In particular, the mechanism $M : \{\pm 1\}^{n \times d} \to [\pm 1]^d$ in Theorem 4.1.2 is given by $M(x) = \overline{x} + Y$, where $Y \sim \mathcal{D}$ and $\mathcal{D}$ is the distribution from Theorem 4.4.1 with $\Delta = 2/n$.[3]

Theorem 4.4.1. *For all $\varepsilon > 0$, $d \geq 1$, and $\Delta > 0$, there exists a continuous distribution $\mathcal{D}$ on $\mathbb{R}^d$ with the following properties.*

- ***Privacy:*** *If $x, x' \in \mathbb{R}^d$ with $\|x - x'\|_\infty \leq \Delta$, then*

$$\mathop{\mathbb{P}}_{Y \sim \mathcal{D}}[x + Y \in S] \leq e^\varepsilon \mathop{\mathbb{P}}_{Y \sim \mathcal{D}}\left[x' + Y \in S\right]$$

  for all measurable $S \subseteq \mathbb{R}^d$.

- ***Accuracy:*** *For all $\alpha > 0$,*

$$\mathop{\mathbb{P}}_{Y \sim \mathcal{D}}[\|Y\|_\infty \geq \alpha] \leq \left(\frac{\Delta d}{\varepsilon \alpha}\right)^d e^{d - \alpha\varepsilon/\Delta}.$$

  In particular, if $d \leq \varepsilon\alpha/2\Delta$, then $\mathop{\mathbb{P}}_{Y \sim \mathcal{D}}[\|Y\|_\infty \geq \alpha] \leq (2e)^{-d}$.

- ***Efficiency:*** *$\mathcal{D}$ can be efficiently sampled.*

Proof. The distribution $\mathcal{D}$ is simply an instantiation of the exponential mechanism

[3]Note that we must truncate the output of M to ensure that $M(x)$ is always in $[\pm 1]^d$.

[MT07]. In particular, the probability density function is given by

$$\mathrm{pdf}_{\mathcal{D}}(y) \propto \exp\left(-\frac{\varepsilon}{\Delta}\|y\|_\infty\right).$$

Formally, for every measurable $S \subseteq \mathbb{R}^d$,

$$\mathop{\mathbb{P}}_{Y\sim\mathcal{D}}[Y \in S] = \frac{\int_S \exp\left(-\frac{\varepsilon}{\Delta}\|y\|_\infty\right) \mathrm{d}y}{\int_{\mathbb{R}^d} \exp\left(-\frac{\varepsilon}{\Delta}\|y\|_\infty\right) \mathrm{d}y}.$$

Firstly, this is clearly a well-defined distribution as long as $\varepsilon/\Delta > 0$.

Privacy is easy to verify: It suffices to bound the ratio of the probability densities for the shifted distributions. For $x, x' \in \mathbb{R}^d$ with $\|x' - x\|_\infty \leq \Delta$, by the triangle inequality,

$$\begin{aligned}\frac{\mathrm{pdf}_{\mathcal{D}}(x+y)}{\mathrm{pdf}_{\mathcal{D}}(x'+y)} &= \frac{\exp\left(-\frac{\varepsilon}{\Delta}\|x+y\|_\infty\right)}{\exp\left(-\frac{\varepsilon}{\Delta}\|x'+y\|_\infty\right)} = \exp\left(\frac{\varepsilon}{\Delta}\left(\|x'+y\|_\infty - \|x+y\|_\infty\right)\right) \\ &\leq \exp\left(\frac{\varepsilon}{\Delta}\|x'-x\|_\infty\right) \leq e^{\varepsilon}.\end{aligned}$$

Define a distribution $\mathcal{D}^*$ on $[0,\infty)$ to by $Z \sim \mathcal{D}^*$ meaning $Z = \|Y\|_\infty$ for $Y \sim \mathcal{D}$. To prove accuracy, we must give a tail bound on $\mathcal{D}^*$. The probability density function of $\mathcal{D}^*$ is given by

$$\mathrm{pdf}_{\mathcal{D}^*}(z) \propto z^{d-1} \cdot \exp\left(-\frac{\varepsilon}{\Delta}z\right),$$

which is obtained by integrating the probability density function of $\mathcal{D}$ over the infinity-ball of radius z, which has surface area $d2^d z^{d-1} \propto z^{d-1}$. Thus $\mathcal{D}^*$ is precisely the gamma distribution with shape d and mean $d\Delta/\varepsilon$. The moment generating function is therefore

$$\mathop{\mathbb{E}}_{Z\sim\mathcal{D}^*}\left[e^{tZ}\right] = \left(1 - \frac{\Delta}{\varepsilon}t\right)^{-d}$$

for all $t < \varepsilon/\Delta$. By Markov's inequality

$$\mathop{\mathbb{P}}_{Z\sim\mathcal{D}^*}[Z \geq \alpha] \leq \frac{\mathop{\mathbb{E}}_{Z\sim\mathcal{D}^*}\left[e^{tZ}\right]}{e^{t\alpha}} = \left(1 - \frac{\Delta}{\varepsilon}t\right)^{-d} e^{-t\alpha}.$$

Setting $t = \varepsilon/\Delta - d/\alpha$ gives the required bound.

It is easy to verify that $Y \sim \mathcal{D}$ can be sampled by first sampling a radius R from a gamma distribution with shape $d+1$ and mean $(d+1)\Delta/\varepsilon$ and then sampling $Y \in [\pm R]^d$ uniformly at random. To sample R we can set $R = \frac{\Delta}{\varepsilon}\sum_{i=0}^{d} \log U_i$, where each $U_i \in (0,1]$ is uniform and independent. This gives an algorithm (in the form of an explicit circuit) to sample $\mathcal{D}$ that uses only $O(d)$ real arithmetic operations, $d+1$ logarithms, and $2d+1$ independent uniform samples from $[0,1]$. □

We remark that the noise distribution in Theorem 4.4.1 is better than the Laplace mechanism *even for L_1 error* by a constant factor. In particular, the expected L_1 norm of the noise distribution in Theorem 4.4.1 is smaller than that of the Laplace mechanism with the same privacy level by a factor of $(2 - o_d(1))$. Moreover, the L_1 noise of the noise distribution in Theorem 4.4.1 stochastically dominates that of the Laplace mechanism:

Remark 4.4.2. *Fix $\varepsilon > 0$, $d \geq 1$, and $\Delta > 0$. Let $\mathcal{D}$ be the distribution on $\mathbb{R}^d$ from Theorem 4.4.1. Let $\mathcal{D}'$ be the distribution on $\mathbb{R}^d$ consisting of d independent samples from the Laplace distribution with scale $2d/\varepsilon n$. Both distributions provide ε-differential privacy when used to answer d one-way marginals. However*

$$\mathop{\mathbb{E}}_{Y\sim\mathcal{D}}\left[\|Y\|_1\right] = \frac{1+1/d}{2} \cdot \mathop{\mathbb{E}}_{Y'\sim\mathcal{D}'}\left[\left\|Y'\right\|_1\right]$$

and

$$\mathop{\mathbb{P}}_{Y\sim\mathcal{D}}\left[\|Y\|_1 > \alpha d\right] \leq \mathop{\mathbb{P}}_{Y'\sim\mathcal{D}'}\left[\left\|Y'\right\|_1 > \alpha d\right]$$

for all α.

4.4.2 Approximate Differential Privacy

We now describe our approximately differentially private mechanism in Figure 4.1.

Parameters: $\alpha \in (0,1)$.
Input: $x \in \{\pm 1\}^{n \times d}$.
Let

$$T = \left\lfloor \frac{2d}{\log^4 d} \right\rfloor, \qquad \sigma_0^2 = \frac{\alpha^2}{32 \log\log d}, \qquad \sigma_1^2 = \frac{\alpha^2}{8 \log^2 d}, \qquad \eta = \frac{2}{\alpha} \log^2 d.$$

For $j \in [d]$, sample a_j^0 from $\mathcal{N}(\overline{x}_j, \sigma_0^2)$.
For $t \in [T]$ do:
Sample $k_t \in [d]$ with

$$\mathbb{P}[k_t = k] = \frac{\exp(\eta |a_k^{t-1} - \overline{x}_k|)}{\sum_{j \in [d]} \exp(\eta |a_j^{t-1} - \overline{x}_j|)}.$$

Sample $a_{k_t}^t$ from $\mathcal{N}(\overline{x}_j, \sigma_1^2)$.
For $j \in [d] \setminus \{k_t\}$, $a_j^t = a_j^{t-1}$.
Output $a_1^T, \cdots, a_d^T$.

Figure 4.1: *Approximately DP Mechanism* $M : \{\pm 1\}^{n \times d} \to [\pm 1]^d$

Proof of Theorem 4.1.3. Firstly, we consider the privacy of M: The data is used in d applications of the Gaussian mechanism with variance σ_0^2 and sensitivity $2/n$, T applications of the Gaussian mechanism with variance σ_1^2, and T applications of the exponential mechanism each satisfying $(4\eta/n)$-differential privacy. Thus, by Lemmas 2.2.5 and 2.2.4 and Proposition 2.3.2 it satisfies ρ-zCDP for

$$\begin{aligned}
\rho =& d \cdot \frac{(2/n)^2}{2\sigma_0^2} + T \cdot \frac{(2/n)^2}{2\sigma_1^2} + T \cdot \frac{1}{2}(4\eta/n)^2 \\
\leq& \frac{64 d \log\log d}{\alpha^2 n^2} + \frac{2d}{\log^4 d} \cdot \frac{16 \log^2 d}{\alpha^2 n^2} + \frac{2d}{\log^4 d} \cdot \frac{32 \log^4 d}{\alpha^2 n^2} \\
=& \frac{d}{\alpha^2 n^2} \left(64 \log\log d + \frac{32}{\log^2 d} + 64 \right) \\
=& (64 + o(1)) \frac{d \log\log d}{\alpha^2 n^2}.
\end{aligned}$$

By Lemma 2.3.6 M satisfies (ε, δ)-differential privacy for all $\delta > 0$ and $\varepsilon = \rho + \sqrt{4\rho \log(1/\delta)}$.

Now we turn our attention to the accuracy of M: For $j \in [d]$, $t \in \{0, 1, \cdots, T\}$, and $\hat{\alpha} > 0$, let $X_j^t(\hat{\alpha}) \in \{0, 1\}$ be indicator of the event that $|a_j^t - \overline{x}_j| > \hat{\alpha}$ and let $X^t(\hat{\alpha}) = \sum_{j \in [d]} X_j^t(\hat{\alpha})$. That is, $X_j^t(\hat{\alpha})$ indicates whether the answer to the j^{th} marginal is $\hat{\alpha}$-accurate in the t^{th} round and $X^t(\hat{\alpha})$ represents the number of $\hat{\alpha}$-incaccurate answers in the t^{th} round.

The final answers are α-accurate if and only if $X^T(\alpha) = 0$. Thus we must show that $\mathbb{P}\left[X^T(\alpha) > 0\right] \leq \beta$, where

$$\beta = e^{-2d/\log^8 d} + \frac{d(d+1)}{d^{\log d}} = \frac{1}{d^{\omega(1)}}.$$

This follows from the following three claims:

(i) All but T of the initial answers are $\frac{\alpha}{2}$-accurate. i.e. $\mathbb{P}\left[X^0(\alpha/2) > T\right] \leq e^{-2d/\log^8 d}$.

(ii) In each of the T "fixing rounds," the exponential mechanism finds a $\alpha/2$-bad answer. i.e. $\mathbb{P}\left[X_{k_t}^{t-1}(\alpha/2) = 0 \mid X^{t-1}(\alpha) > 0\right] \leq \frac{1}{d^{\log d - 1}}$.

(iii) Each of the T resampled answers is $\alpha/2$-accurate. i.e. $\mathbb{P}\left[X_{k_t}^t(\alpha/2) = 1\right] \leq \frac{1}{d^{\log d}}$.

Claim (i) says that, with high probability, $X^0(\alpha/2) \leq T$. Claims (ii) and (iii) imply that, with high probability, $X^t(\alpha/2)$ strictly decreases in each round, as long as $X^t(\alpha) > 0$. Thus either $X^t(\alpha) = 0$ for some $t \in [T]$ or $X^T(\alpha/2) = 0$. Claim (iii) implies that if $X^t(\alpha) = 0$ at some point, then it remains 0 for the rest of the execution and $X^T(\alpha) = 0$ with high probability. So, as long as all the good events in claims (i-iii) happen, the final answers are α-accurate. A union bound shows that this happens with probability $1 - \beta$.

(i) Firstly, the random variables $X_1^0(\alpha/2), X_2^0(\alpha/2), \cdots, X_d^0(\alpha/2)$ are independent. For each $j \in [d]$,

$$\mathbb{E}\left[X_j^0(\alpha/2)\right] = \mathop{\mathbb{P}}_{G\sim\mathcal{N}(0,\sigma_0^2)}\left[|G| > \alpha/2\right] \le e^{-\alpha^2/8\sigma_0^2} \le \frac{1}{\log^4 d}.$$

Thus $\mathbb{E}\left[X^0(\alpha/2)\right] \le d/\log^4 d$. By Hoeffding's inequality,

$$\mathbb{P}\left[X_0(\alpha/2) \ge \mathbb{E}\left[X_0(\alpha/2)\right] + \lambda\right] \le e^{-2\lambda^2/d}$$

for all $\lambda > 0$. Setting $\lambda = d/\log^4 d$ verifies the first claim.

(ii) Now we must verify that, in each round, the exponential mechanism finds a bad query with high probability. We have

$$\begin{aligned}
\mathbb{P}\left[X_{k_t}^{t-1}(\alpha/2) = 0\right] =& \frac{\sum_{k\in[d]} \exp(\eta|a_k^{t-1} - \overline{x}_k|) \cdot \mathbb{I}(|a_k^{t-1} - \overline{x}_k| \le \alpha/2)}{\sum_{j\in[d]} \exp(\eta|a_j^{t-1} - \overline{x}_j|)} \\
\le& \frac{\exp(\eta\alpha/2)\cdot d}{\exp(\eta\alpha)\cdot X^{t-1}(\alpha)} \\
\le& d^{1-\log d},
\end{aligned}$$

assuming $X^{t-1}(\alpha) > 0$.

(iii) Finally, we have

$$\mathbb{P}\left[X_{k_t}^t(\alpha/2) = 1\right] = \mathop{\mathbb{P}}_{G\sim\mathcal{N}(0,\sigma_1^2)}\left[|G| > \alpha/2\right] \le e^{-\alpha^2/8\sigma_1^2} \le d^{-\log d}.$$

□

Chapter 5

Privacy Attacks

5.1 Introduction

Given a collection of (approximate) summary statistics about a dataset, and the precise data of a single target individual, under what conditions is it possible to determine whether or not the target is a member of the dataset? This *tracing* problem is the focus of this chapter.

Questions of this type arise in many natural situations in which membership in the dataset is considered sensitive; indeed, this is typically the reason for choosing to publish summary statistics, as opposed to releasing the raw data. In a scenario that is prominent in the literature, the dataset contains genomic information about a *case group* of individuals with a specific medical diagnosis, as in a genome-wide association study (GWAS), and the summary statistics are SNP allele frequencies, *i.e. one-way marginals*. Specifically, if each person's data consists of d binary attributes, we consider a mechanism that releases (an approximation to) the average value of the each of the d attributes. Homer et al. [HSR+08] demonstrated the privacy risks inherent in this scenario, presenting and analyzing a tracing algorithm for

membership in a GWAS case group, provided the attacker also has access to allele frequencies for a reference group of similar ancestral make-up as that of the case group.

It came as a surprise to the genomics research community that the trace amount of DNA contributed by an individual is enough to determine membership in the case group with high statistical confidence. The result had a major practical impact in the form of very restrictive policies governing access to allele frequency statistics in studies funded by the US National Institutes of Health and the Wellcome Trust. Follow-up analytical works provide alternative tests and asymptotic analyses of tradeoffs between the size of the test set, the size of a reference dataset, power, confidence, and number of measurements [SOJH09].

As in the follow-up works, the analysis in Homer et al. assumes that *exact* statistics are released, leaving open the possibility that the attack may be foiled if the statistics are distorted, for example, due to measurement error (which can be highly correlated across the statistics), or because noise is intentionally introduced in order to protect privacy.

In this chapter, we show that one can test if an individual is present in the case group even when the one-way marginals are considerably distorted before being released. We give a single tracing attack that applies to *all* mechanisms that produce sufficiently accurate estimates of the statistics in question, rather than to just the single mechanism that outputs exact statistics.

A line of work initiated by Dinur and Nissim [DN03] provides attacks of this flavor for certain kinds of statistics, showing that all mechanisms that release "too many" answers that are "too accurate" are subject to devastating "reconstruction attacks," which allow an adversary to determine the private data of almost all individuals in a dataset. These attacks, which immediately give lower bounds

on noise needed to avoid blatant non-privacy, have been extended in numerous works [DMT07, DY08, KRSU10, De12, KRS13, MN12, FMN13, NTZ13].

These reconstruction attacks do not generally apply in the setting of Homer et al., since they either require that the amount of noise introduced for privacy is very small (less than the sampling error), or require an exponential number of statistics, or do not apply to statistics that are as simple (namely, attribute frequencies), or require that the adversary have a significant amount of auxiliary information about the other individuals in the dataset.

Of course, complete reconstruction is an extreme privacy failure: the privacy of essentially every member of the dataset is lost! Conversely, protection from complete reconstruction is a very low barrier for a privacy mechanism. What if we are more demanding, and ask that an attacker not be able to determine whether an individual is present or absent from the dataset, that is, to *trace*? This in/out protection is the essence of differential privacy, and the question of how much noise is needed to ensure differential privacy, first studied in [HT10], has seen many recent developments [Ull13, BUV14, DNT14, HU14, §4,§6]. By shifting the goal from reconstructing to tracing, these works obtain lower bounds on noise for settings where reconstruction is impossible.

In particular, Chapter 4 and fingerprinting codes [BS98, Tar08, BUV14] can be viewed as providing tracing attacks that operate given attribute frequencies of the dataset.However, they require that the attribute frequencies of the underlying population are drawn from a particular (somewhat unnatural) distribution, and that the attacker has very accurate knowledge of these frequencies. We remark that such knowledge is the "moral equivalent," in this literature, to having a large reference population, in the genomics literature.

In this chapter, we generalise the attacks based on fingerprinting codes in several

ways to considerably broaden their applicability:

- The population's attribute frequencies can be drawn from any distribution on $[0,1]$ that is sufficiently smooth and spread out, including, for example, the uniform distribution on $[0,1]$ or a large subinterval. The tracing algorithm does not depend on the distribution.
- Instead of knowing the population attribute frequencies, it suffices for the attacker to have a *single* reference sample from the population.
- We show that similar attacks can be applied to Gaussian data (rather than binary data) for mechanisms that release too many attribute averages with nontrivial accuracy.

Our results provide a common generalisation of the fingerprinting results and the results of Homer et al., showing they are special cases of a much broader phenomenon.

Like the fingerprinting attacks of Chapter 4 and Bun et al. [BUV14], the lower bounds on noise implied by our attacks nearly match the upper bounds on noise sufficient to ensure the strong guarantees of differential privacy, for example, via the Gaussian or Laplace mechanisms [DN03, BDMN05, DMNS06, DKM+06, DRV10]). Thus, the cost in utility for avoiding our attacks is nearly the same as the cost for avoiding the much larger class of attacks that differential privacy prevents, where the dataset can be arbitrary and the attacker can know everything about it, except whether or not the target individual is present in the dataset.

5.1.1 Model and Assumptions

Distributional Assumption. The dataset consists of n independent samples from a *population*, which is given by a product distribution $\mathcal{P}_p$ on $\{\pm 1\}^d$. The vector

$p \in [-1,1]^d$ specifies the expectation of a sample from $\mathcal{P}_p$. That is, to sample $x \sim \mathcal{P}_p$, we set $x_j = 1$ with probability $(1+p_j)/2$ and set $x_j = -1$ with probability $(1-p_j)/2$, independently for each j.

The vector p represents unknown parameters of the population; p is unknown to both the mechanism and the privacy attacker.[1] The vector p is itself drawn from the product distribution $\mathcal{D}$ on $[-1,1]^d$ with the j^{th} marginal having probability density function $\rho_j : [-1,1] \to \mathbb{R}$. In the case of genomics, we can think of the distribution $\mathcal{D}$ as capturing, for example, differences between populations (although of course in reality this would not be a product distribution). Our attacks will succeed even if the mechanism knows $\mathcal{D}$ but the attacker does not, provided each ρ_j is sufficiently smooth and spread out *e.g.*, if ρ_j is uniform on a large enough subinterval of $[0,1]$).

Accuracy of the Mechanism. The (possibly randomized) *mechanism* $\mathcal{M}$ receives n independent samples $x_1, \cdots, x_n \in \{\pm 1\}^d$ drawn from $\mathcal{P}_p$ (after p is initially drawn from $\mathcal{D}$), and outputs a vector $q \in [-1,1]^d$ with $q \approx \bar{x} = \frac{1}{n}\sum_{i\in[n]} x_i \approx p$. That is, $\mathcal{M}$ provides approximate one-way marginals. We say $\mathcal{M}$ is *α-accurate* if for all $j \in [d]$ we have $\|q - p\|_\infty \leq \alpha$. For simplicity, we assume this holds with probability 1; if this is not the case, then this failure probability can be absorbed into the failure probability of our attack.

The Attacker. The *privacy attacker* $\mathcal{A}$ receives two samples in $\{\pm 1\}^d$, the target y and the reference z, where z is drawn independently from the population $\mathcal{P}_p$, together with the output q of $\mathcal{M}$ on a dataset $x_1, \ldots, x_n$, and produces an answer, either IN or OUT. The attacker's answer indicates whether or not it believes y is among the $x_1, \cdots, x_n$ given to $\mathcal{M}$. The attacker is guaranteed that the reference

[1] If the mechanism knows p then the problem becomes vacuous: it could simply ignore the data and publish p.

sample z is drawn from $\mathcal{P}_p$ independent from everything else. The attacker must satisfy two properties:

- *Soundness:* If y is drawn from $\mathcal{P}_p$ independent from the view of $\mathcal{M}$ (i.e. independent from q), then $\mathcal{A}$ should output IN with probability at most s.
- *Completeness:* Choose i uniformly from $[n]$ and set $y = x_i$. Then $\mathcal{A}$ should output IN with probability at least c. The probability is over all the random choices: i, x, z, and the coin flips of $\mathcal{A}$ and $\mathcal{M}$.

These conditions are interesting when $c \gg s$, as when $c \leq s$ they are trivially satisfied by having $\mathcal{A}$ always output IN with probability c. To interpret this, think of y as the data of a member of the population and $\mathcal{A}$ wants to determine whether or not y is in the dataset (case group) given to $\mathcal{M}$. For $\mathcal{A}$ to be considered successful we require that it can identify a random member of the dataset with reasonably high probability (given by the completeness parameter c), whilst, if y is not in the dataset, it is erroneously claimed otherwise with negligible probability (given by the (un)soundness parameter s). The reference sample z is some minimal auxiliary information about the population that $\mathcal{A}$ can use.

5.1.2 Our Results

Theorem 5.1.1 (Main – Informal)**.** *There is a universal constant $\alpha > 0$ such that for every $\delta > 0$, $n \in \mathbb{N}$, and $d \geq O(n^2 \log(1/\delta))$, there exists an attacker $\mathcal{A} : \{\pm 1\}^d \times [-1,1]^d \times \{\pm 1\}^d \to \{\text{IN}, \text{OUT}\}$ such that the following holds.*

Let $\mathcal{D}$ be a product distribution on $[-1,1]^d$ such that each marginal satisfies a technical smoothness condition (Definition 5.2.4). Let $\mathcal{M} : \{\pm 1\}^{n \times d} \to [-1,1]^d$ be α-accurate. Let

$p \sim \mathcal{D}$ *and* $x_1, \cdots, x_n, y, z \sim \mathcal{P}_p$. *Let* $q \sim \mathcal{M}(x_1, \cdots, x_n)$. *Then*

$$\mathbb{P}\left[\mathcal{A}(y, q, z) = \text{IN}\right] \leq \delta \quad \textit{and} \quad \mathbb{P}\left[\exists\, i \in [n] \;\; \mathcal{A}(x_i, q, z) = \text{IN}\right] \geq 1 - \delta.$$

Thus, if the first input (y) to $\mathcal{A}$ is a random independent element of the population, then $\mathcal{A}$ will accept with probability at most $s = \delta$ (the probability space includes the selection of y), but the first input is a random element of the dataset (x_i for a random i), $\mathcal{A}$ will accept with probability at least $c = (1 - \delta)/n$. Thus, the result is nontrivial when $\delta < (1 - \delta)/n$ (e.g. $\delta = o(1/n)$).

We discuss a number of features and extensions of the result.

Dimensionality Needed. The dimensionality d of the data needed for the attack is $d = \tilde{O}(n^2)$ for $\delta = 1/2n$, which is tight up to polylogarithmic factors for achieving constant accuracy α. Indeed, it is possible to answer $d = \tilde{\Omega}(n^2)$ one-way marginals with accuracy $\alpha = o(1)$, while satisfying the strong guarantee of $(o(1), 1/n^{\omega(1)})$-differential privacy [DN03, BDMN05, DKM+06, DMNS06, DRV10].[2] (Our tracing attack implies that no mechanism satisfying the above conditions can be $(0.1, 1/4n)$ differentially private.) For the one-way marginals we consider, the number of statistics released equals the dimensionality d of the data, but for richer families of statistics, the dimensionality is the more significant parameter. Indeed, many more than n^2 statistics can be released if the dimensionality d of the data is smaller than n^2—the algorithms of [BLR13, HR10, RR10, DRV10] can release a number of statistics that is nearly exponential in $n/\sqrt{d}$.

[2] An algorithm that operates on datasets is (ε, δ)-differentially private if for all datasets S, S' differing in the data of a single individual and every event E, the probability of E when the dataset is S is at most δ plus e^{ε} times the probability of E when the dataset is S'.

Beyond the $d = \Theta(n^2)$ Barrier. The price for our very weak assumptions – weakly accurate answers and only a single reference sample – is that we (provably) need $d = \Omega(n^2)$ and can only trace a single individual. With more accurate answers and a larger reference pool, a slightly modified version of our attacker can trace with smaller d, and can trace many individuals in the dataset: if the mechanism is α-accurate (for some $\alpha \geq n^{-1/2}$), and we are given roughly $1/\alpha^2$ independent reference samples from the distribution, then we trace when the dataset has dimension only $O(\alpha^2 n^2)$. Moreover, we can successfully trace $\Omega(1/\alpha^2)$ individuals in the dataset, yielding a completeness probability of $c = \Omega(1/\alpha^2 n)$ (Section 5.3).

Weaker Soundness Conditions. The soundness of our attack does not rely on any properties of the distribution $\mathcal{D}$, the accuracy of $\mathcal{M}$, the relation between d, n, and δ, or even the distribution of the rows $x_1, \ldots, x_n$. It only requires that, conditioned on q, y and z are sampled independently from the same product distribution. Thus, the attack can be carried out under only the latter assumption, and if it says IN, one can safely conclude $y \in \{x_1, \ldots, x_n\}$.

Higher-Power Attacks. Our completeness probability of $c = \Theta(1/\alpha^2 n)$ is essentially tight, as a mechanism $\mathcal{M}$ that outputs the averages on a subsample of size $O(1/\alpha^2)$ will be accurate but only allows tracing at most an $O(1/\alpha^2 n)$ fraction of individuals in the dataset

However, if we assume that $\mathcal{M}$ is *symmetric*, then we can get around this. That is, if we assume that $\mathcal{M}$ can be written as $\mathcal{M}(x_1, \cdots, x_n) = \mathcal{M}'(\overline{x})$ (where $\overline{x} = \frac{1}{n}\sum_{i \in [n]} x_i \in [-1, 1]^d$ is the average of the sample), then we can prove that

$$\forall i \in [n] \quad \mathbb{P}\left[\mathcal{A}(x_i, q, z) = \text{IN}\right] \geq 1 - \delta.$$

Note that with this high-power guarantee ($c \geq 1 - \delta$), it is meaningful to take δ to be a fixed constant (e.g. the standard significance level of .05).

The Distribution $\mathcal{D}$. As noted above, we impose a technical regularity condition on the distribution $\mathcal{D}$, requiring that its marginals ρ_j are sufficiently smooth and spread out. This includes distributions such as the uniform distribution on a large subinterval and the family of Beta distributions.

Some assumptions on $\mathcal{D}$ are necessary. For example, if each marginal ρ_j were supported on a subinterval of length at most $\alpha/2 \gg 1/\sqrt{n}$, then the mechanism could give accurate answers by just producing a vector $q \in [-1, 1]^d$ in the support of $\mathcal{D}$ and not using the dataset at all. This shows that the ρ_j need to be sufficiently "spread out". To see why "smoothness" is necessary, suppose that ρ_j were concentrated on two points p^* and p^{**} that are reasonably far apart (farther than 2α). Then the mechanism can simply test whether the average of the data elements exceeds $(p^* + p^{**})/2$ and, if so, output $\max\{p^*, p^{**}\}$; otherwise output $\min\{p^*, p^{**}\}$. While this mechanism is not differentially private (a guarantee against tracing in the worst case), with high probability over the choice of the dataset this mechanism is insensitive to small changes in the dataset, i.e., changing one row will not change the output. This makes tracing impossible.

5.1.3 Description of The Attack

Like the attacks in previous tracing work for the genomic setting [HSR+08, SOJH09, BRS+09, JYW+09, ZPL+11] and in the fingerprinting setting [Tar08, §4,§6], our attack uses a simple scoring function to make its decision. The scoring function works incrementally, with each marginal (SNP) making a separate contribution. The attack is described in full in Figure 5.1.

Input: $y, z \in \{\pm 1\}^d$ and $q \in [-1,1]^d$.
Compute $\langle y, q\rangle = \sum_{j\in[d]} y^j \cdot q^j$ and $\langle z, q\rangle = \sum_{j\in[d]} z^j \cdot q^j$.
If $\langle y, q\rangle - \langle z, q\rangle > \tau := \sqrt{8d\ln(1/\delta)}$, output IN; otherwise output OUT.

Figure 5.1: ***Our Privacy Attack*** $\mathcal{A}_{\delta,d}(y, q, z)$

The key features of the adversary are that it only sees the data of the user y being traced, plus a reference sample z (in addition, of course, to seeing the output q), and does not depend on the mechanism $\mathcal{M}$, the feature vector p, or the distribution $\mathcal{D}$ on p.

5.1.4 Comparison with Previous Work

As mentioned above, our model and results provide a common generalisation of lines of work from several fields.

1. Work in the genomics community [HSR+08, BRS+09, VH09, SOJH09, JYW+09] has so far focused on the case where *exact* statistics are available to the attacker ($\alpha = 0$ in our formalism). With a reference sample of $\Omega(n)$ individuals, they showed that $d = \Theta(n)$ attributes are necessary and sufficient, while with a constant-sized reference pool, $d = \Theta(n^2)$ is required [SOJH09]). Our first attack uses $\Theta(n^2 \cdot \log n)$ statistics with a reference pool of size 1, and makes only a minimal accuracy assumption (a constant bound α on the bias).

   Our second attack requires only $d = \tilde{O}(\alpha^2 n^2)$ statistics if the mechanism is α-accurate (for some $\alpha \geq n^{-1/2}$) and the reference pool is of size $O(\log(n)/\alpha^2)$, in which case it can also successfully trace $\Omega(1/\alpha^2)$ individuals in the dataset.

   Im et al. [IGNC12] use (exact) regression coefficients instead of marginals as the basis of an attack, with similar results to the case of marginals.

2. Work on fingerprinting attacks [BUV14, §4] corresponds to our setting of a constant α, but assumes that p is drawn from a specific distribution $\mathcal{D}$, and the attacker $\mathcal{A}$ knows p exactly (essentially, an infinite reference pool). The dimensions required in their attacks are similar to ours ($d = \Theta(n^2)$).

We note that previous work has focused on categorical data, but our results extend to the setting of normally-distributed real-valued data.

Other Work on Genetic Privacy. The literature contains attacks based on various types of published aggregate statistics, e.g., allele frequencies, genetic frequencies, and various quantitative phenotypes such as cholesterol levels [HSR+08, JYW+09, WLW+09, IGNC12]; see [EN14] for a survey. Particularly exciting (or troubling) is the work of Wang et al. [WLW+09] that exploits correlations among different SNPs. Not only do their attacks require relatively few SNPs, but they go beyond in/out privacy compromise, actually reconstructing SNPs of members of the case group. In our view, the message of these works and ours, taken as a whole, is that information combines in surprising ways, aggregation should not be assumed to provide privacy on its own, and rigorous approaches to controlling privacy risk are *necessary*.

5.2 Tracing with a Single Reference Sample

Now we analyse our attack (given in Figure 5.1) and thereby prove Theorem 5.1.1.

5.2.1 Soundness Analysis

Proposition 5.2.1 (Soundness). *Let $q, p \in [-1,1]^d$. Suppose $y, z \sim \mathcal{P}_p$ are independent from each other and from q. Then*

$$\mathbb{P}\left[\mathcal{A}_{\delta,d}(y,q,z) = \text{IN}\right] \leq \delta.$$

Proof. We can view p and q as fixed. Since y and z are identically distributed, $\mathbb{E}\left[\langle y,q\rangle - \langle z,q\rangle\right] = 0$. Since y and z are independent samples from a product distribution, we have that $\langle y,q\rangle - \langle z,q\rangle = \sum_{i\in[d]}(y^j - z^j)\cdot q^j$ is the sum of $2d$ independent random variables each of which is bounded by $\max\{\|y\|_\infty, \|z\|_\infty\}\cdot\|q\|_\infty \leq 1$. Thus, by a Hoeffding's inequality,

$$\mathbb{P}\left[\langle y,q\rangle - \langle z,q\rangle > \tau\right] \leq e^{-\tau^2/4d} = \delta,$$

as required. □

Remark 5.2.2. *Proposition 5.2.1 makes no assumptions about q. Thus soundness holds even if $\mathcal{M}$ is not accurate or if y, z are not sampled from the true population - they need only be sampled from the same product distribution.*

5.2.2 Correlation Analysis

To prove completeness we must show that $\langle x_i,q\rangle - \langle z,q\rangle > \tau$ with good probability for a random $i \in [n]$ when the mechanism's output is α-accurate. First we give a formal definition of accuracy:

Definition 5.2.3 (Accuracy). *We say $\mathcal{M} : \{\pm 1\}^{n\times d} \to [-1,1]^d$ is α-accurate if*

$$\|M(x) - \overline{x}\|_\infty \leq \alpha$$

for all $x \in \{\pm 1\}^{n\times d}$, where $\overline{x} \in [-1,1]^d$ is the average of the rows of x.

For simplicity we assume that the accuracy bound holds with probability 1. In many situations this may only hold with high probability, in which case we can absorb the failure probability into that of the attack.

We begin by showing that, under our regularity assumption on $\mathcal{D}$,

$$\mathbb{E}\left[\sum_{i\in[n]}\left(\langle x_i, q\rangle - \langle z, q\rangle\right)\right] \geq Cn\tau$$

for an appropriate constant $C > 1$.

Intuitively, $\sum_{i\in[n]}\langle x_i, q\rangle$ measures how much the output $q \in [-1,1]^d$ of $\mathcal{M}$ correlates with the input $x_1, \cdots, x_n \in \{\pm 1\}^d$ of $\mathcal{M}$, whereas $\langle z, q\rangle$ measures how much a random member of the population correlates with q. Thus we are proving that the output of $\mathcal{M}$ is more correlated with the input of $\mathcal{M}$ than with an independent sample from the population.

By linearity of expectations it suffices to show that $\mathbb{E}\left[\sum_{i\in[n]} x_i^j q^j - z^j q^j\right] \geq Cn\tau/d$ for each $j \in [d]$. We now focus on a fixed $j \in [d]$ and, for clarity, omit the superscript.

First some notation: Let $p \sim \rho$ denote that $p \in \mathbb{R}$ is drawn according to the probability distribution given by ρ (e.g. ρ is a probability density function $\rho : \mathbb{R} \to \mathbb{R}$). For $p \in [-1,1]$, let $x \sim p$ denote that $x \in \{\pm 1\}$ is drawn with $\mathbb{E}[x] = p$. Let $x_{1\cdots n} \sim \rho$ denote that $x_1, \cdots x_n \in \{\pm 1\}^n$ are drawn independently with $x_i \sim \rho$ for each $i \in [n]$.

The regularity condition we need is the following.

Definition 5.2.4 (Strong Distribution)**.** *Let ρ be a probability distribution on $[-1,1]$. Define $h_n^\rho : \{-n-1, -n+1, \cdots, n+1\} \to \mathbb{R}$ by*

$$h_n^\rho(t) = \frac{(n+1+t)(n+1-t)}{2(n+1)} \cdot \mathop{\mathbb{P}}_{p\sim\rho, x_{1\cdots n+1}\sim p}\left[\sum_{i\in[n+1]} x_i = t\right].$$

We say ρ is (ξ, n)-strong *if*

$$\sum_{t \in \{-n, -n+2, \cdots, n\}} \left| h_n^\rho(t-1) - h_n^\rho(t+1) \right| \leq \xi.$$

We say ρ is ξ-strong *if ρ is (ξ, n)-strong for all n.*

First let us unpack this definition: The definition bounds the *total variation* of the function h_n^ρ. So we require h_n^ρ to be smooth. The function h_n^ρ is the product of two terms. The first term is large (at most $(n+1)/2$) in the middle of the range and smoothly decreases towards zero at the ends of the range. The second term can be viewed as the probability mass function of a discretisation of the continuous distribution ρ: There are $n+2$ buckets $\{-n-1, -n+1, \cdots, n+1\}$. A sample $p \sim \rho$ is thrown into one of the $n+2$ buckets in a random fashion. The most likely bucket is the one closest to $(n+1) \cdot p$ and the probability of landing in a given bucket decays rapidly as we move away from the most likely bucket. Intuitively, being a strong distribution simply means that neighbouring buckets should contain a similar amount of probability mass.

We give some meaning to this definition in Section 5.2.4. Intuitively, it suffices for a distribution to have a "smooth" probability density function that is sufficiently "spread out." In particular, the uniform distribution on $[-1, 1]$ is 1-strong.

Now we relate the definition of a strong distribution to the correlation quantity of interest:

Lemma 5.2.5. *Let ρ be a (ξ, n)-strong probability distribution on $[-1, 1]$. Let $f : \{\pm 1\}^n \to [-1, 1]$. Then*

$$\left| \mathop{\mathbb{E}}_{p \sim \rho, x_{1 \cdots n} \sim p, z \sim p} \left[f(x) \sum_{i \in [n]} (x_i - z) \right] \right| \leq \xi.$$

Furthermore, the above inequality holds for all such f only if ρ is a (ξ, n)-strong distribution.

Proof. Define a random variable S_n^ρ on $\{-n-1, -n+1, \cdots, n-1, n+1\}$ as follows. First sample $p \sim \rho$. Then sample $x_{1\cdots n+1} \sim p$ and let $S_n^\rho = \sum_{i=1}^{n+1} x_i$.

Firstly, by symmetry the following are equivalent ways of sampling random variables $x_1, \cdots, x_n, z$.

- Sample $p \sim \rho$. Then sample $x_{1\cdots n} \sim p$ and $z \sim p$.
- Sample $s \sim S_n^\rho$. Then sample $x_1, \cdots, x_n, z \in \{\pm 1\}$ uniformly at random conditioned on $z + \sum_{i\in[n]} x_i = s$.
- Sample $s \sim S_n^\rho$. Then sample $z \in \{\pm 1\}$ with $\mathbb{E}[z] = \frac{s}{n+1}$. Then sample $x_1, \cdots, x_n \in \{\pm 1\}$ uniformly at random conditioned on $\sum_{i\in[n]} x_i = s - z$.

Thus we can rewrite the expectation using S_n^ρ:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{p\sim\rho, x_{1\cdots n}\sim p, z\sim p}\left[f(x)\sum_{i\in[n]}(x_i - z)\right] \\
&= \mathop{\mathbb{E}}_{s\sim S_n^\rho}\left[\mathop{\mathbb{E}}_{z\sim\frac{s}{n+1}}\left[\mathop{\mathbb{E}}_{x\in\{\pm1\}^n:\sum_{i\in[n]} x_i = s-z}\left[f(x)\sum_{i\in[n]}(x_i - z)\right]\right]\right] \\
&= \mathop{\mathbb{E}}_{s\sim S_n^\rho}\left[\mathop{\mathbb{E}}_{z\sim\frac{s}{n+1}}\left[\mathop{\mathbb{E}}_{x\in\{\pm1\}^n:\sum_{i\in[n]} x_i = s-z}[f(x)]\,(s - z - nz)\right]\right] \\
&= \mathop{\mathbb{E}}_{s\sim S_n^\rho}\left[\mathop{\mathbb{E}}_{z\sim\frac{s}{n+1}}[g(s-z)(s-z-nz)]\right],
\end{aligned}$$

where $g : \{-n, \cdots, n\} \to [-1, 1]$ given by

$$g(t) := \mathop{\mathbb{E}}_{x\in\{\pm1\}^n:\sum_{i\in[n]} x_i = t}[f(x)]$$

is the symmetrisation of f. Now we expand the expectations as sums:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{s\sim S_n^\rho}\left[\mathop{\mathbb{E}}_{z\sim\frac{s}{n+1}}\left[g(s-z)(s-z-nz)\right]\right]\\
&=\mathop{\mathbb{E}}_{s\sim S_n^\rho}\begin{bmatrix}\mathop{\mathbb{P}}_{z\sim\frac{s}{n+1}}[z=1] & \cdot & g(s-1)(s-1-n)\\ \mathop{\mathbb{P}}_{z\sim\frac{s}{n+1}}[z=-1] & \cdot & g(s+1)(s+1+n)\end{bmatrix}\\
&=\mathop{\mathbb{E}}_{s\sim S_n^\rho}\begin{bmatrix}\frac{n+1+s}{2(n+1)} & \cdot & g(s-1)(s-1-n)\\ \frac{n+1-s}{2(n+1)} & \cdot & g(s+1)(s+1+n)\end{bmatrix}\\
&=\mathop{\mathbb{E}}_{s\sim S_n^\rho}\left[\frac{(n+1+s)(n+1-s)}{2(n+1)}\cdot(g(s+1)-g(s-1))\right]\\
&=\frac{1}{2(n+1)}\sum_{s\in\{-n+1,-n+3,\cdots,n-1\}}\mathbb{P}\left[S_n^\rho=s\right]\cdot(n+1+s)(n+1-s)\cdot(g(s+1)-g(s-1))\\
&=\frac{1}{2(n+1)}\sum_{t\in\{-n+2,-n+4,\cdots,n\}}\mathbb{P}\left[S_n^\rho=t-1\right]\cdot(n+t)(n-t+2)\cdot g(t)\\
&-\frac{1}{2(n+1)}\sum_{t\in\{-n,-n+2\cdots,n-2\}}\mathbb{P}\left[S_n^\rho=t+1\right]\cdot(n+t+2)(n-t)\cdot g(t)\\
&=\frac{1}{2(n+1)}\sum_{t\in\{-n,-n+2,\cdots,n\}}g(t)\cdot\begin{pmatrix}\mathbb{P}\left[S_n^\rho=t-1\right]\cdot(n+t)(n-t+2)\\ -\mathbb{P}\left[S_n^\rho=t+1\right]\cdot(n+t+2)(n-t)\end{pmatrix}\\
&=\sum_{t\in\{-n,-n+2,\cdots,n\}}g(t)\cdot(h(t-1)-h(t+1)),
\end{aligned}
$$

where h is as in Definition 5.2.4. Now we can apply Hölder's inequality with the

definition of a strong distribution and g to conclude:

$$\begin{aligned}
&\left| \mathop{\mathbb{E}}_{p\sim\rho, x_{1\cdots n}\sim p, z\sim p} \left[f(x) \sum_{i\in[n]} (x_i - z) \right] \right| \\
&= \left| \mathop{\mathbb{E}}_{s\sim S_n^\rho} \left[\mathop{\mathbb{E}}_{z\sim \frac{s}{n+1}} \left[g(s-z)(s-z-nz) \right] \right] \right| \\
&= \left| \sum_{t\in\{-n,-n+2,\cdots,n\}} g(t)\cdot(h(t-1)-h(t+1)) \right| \\
&\leq \|g\|_\infty \sum_{t\in\{-n,-n+2,\cdots,n\}} |h(t-1)-h(t+1)| \\
&\leq \xi.
\end{aligned}$$

Note that there exists a g that makes this inequality tight, namely

$$g_{\text{tight}}(t) = \operatorname{sign}\left(h(t-1)-h(t+1)\right).$$

Setting $f_{\text{tight}}(x) = g_{\text{tight}}\left(\sum_{i\in[n]} x_i\right)$ shows that the lemma is tight. □

Now we translate Lemma 5.2.5 into the form we will use:

Corollary 5.2.6. *Let ρ be a (ξ, n)-strong probability distribution on $[-1,1]$. Let $M : \{\pm 1\}^n \to \mathbb{R}$ satisfy $|M(x) - \overline{x}| \leq \alpha$ for all $x \in \{\pm 1\}^n$. Then*

$$\mathop{\mathbb{E}}_{p\sim\rho, x_{1\cdots n}\sim p, z\sim p} \left[M(x) \sum_{i\in[n]} (x_i - z) \right] \geq \mathop{\mathbb{E}}_{p\sim\rho} \left[1 - p^2\right] - \alpha\xi.$$

This result should be compared to the fingerprinting lemma (Lemma 4.3.7) of the previous chapter. The key difference is that it is not specific to a single distribution ρ (Lemma 4.3.7 only holds for the uniform distribution, but is also "robust" in that it only requires a bound on the expected error, rather than the worst-case error.).

Proof. Write $M(x) = \overline{x} - \alpha \cdot f(x)$ for some $f : \{\pm 1\}^n \to [-1, 1]$. Now

$$\begin{aligned}
&\mathop{\mathbb{E}}_{p\sim\rho, x_{1\cdots n}\sim p, z\sim p}\left[M(x)\sum_{i\in[n]}(x_i - z)\right] \\
&= \mathop{\mathbb{E}}_{p\sim\rho, x_{1\cdots n}\sim p, z\sim p}\left[\overline{x}\sum_{i\in[n]}(x_i - z)\right] - \alpha \cdot \mathop{\mathbb{E}}_{p\sim\rho, x_{1\cdots n}\sim p, z\sim p}\left[f(x)\sum_{i\in[n]}(x_i - z)\right] \\
&\geq \mathop{\mathbb{E}}_{p\sim\rho, x_{1\cdots n}\sim p, z\sim p}\left[\overline{x}\sum_{i\in[n]}(x_i - z)\right] - \alpha \cdot \xi,
\end{aligned}$$

by Lemma 5.2.5. All that remains is the following calculation:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{p\sim\rho, x_{1\cdots n}\sim p, z\sim p}\left[\overline{x}\sum_{i\in[n]}(x_i - z)\right] \\
&= \mathop{\mathbb{E}}_{p\sim\rho, x_{1\cdots n}\sim p}\left[\overline{x}\cdot(\overline{x} - p)\cdot n\right] && (\text{since } \mathbb{E}[z] = p) \\
&= \mathop{\mathbb{E}}_{p\sim\rho, x_{1\cdots n}\sim p}\left[(\overline{x} - p)\cdot(\overline{x} - p)\cdot n\right] && (\text{since } \mathbb{E}[p\cdot(\overline{x} - p)] = 0) \\
&= \mathop{\mathbb{E}}_{p\sim\rho}\left[\mathop{\mathsf{Var}}_{x_{1\cdots n}\sim p}[\overline{x}]\cdot n\right] \\
&= \mathop{\mathbb{E}}_{p\sim\rho}\left[\mathop{\mathsf{Var}}_{x\sim p}[x]\right] \\
&= \mathop{\mathbb{E}}_{p\sim\rho}\left[1 - p^2\right].
\end{aligned}$$

□

We now make an observation that will allow the construction of a high-power attack for symmetric M. Suppose $f : \{\pm 1\}^n \to [-1, 1]$ can be written as $f(x) = f_*\left(\frac{1}{n}\sum_{i\in[n]} x_i\right)$ for some $f_* : [-1, 1] \to [-1, 1]$. Then, by symmetry, the conclusion of Corollary 5.2.6 can be altered to

$$\forall i \in [n] \quad \mathop{\mathbb{E}}_{p\sim\rho, x_1\cdots x_n\sim p, z\sim p}[f(x)\cdot(x_i - z)] \geq \frac{\mathop{\mathbb{E}}_{p\sim\rho}\left[1 - p^2\right] - \alpha\xi}{n}.$$

Formally, we have the following definition and Lemma.

Definition 5.2.7. *A function $f : \mathcal{V}^n \to \mathcal{Y}$ (where $\mathcal{V}$ is a vector space) is* symmetric *if there exists a function $f_* : \mathcal{V} \to \mathcal{Y}$ such that $f(x) = f_*\left(\frac{1}{n}\sum_{i\in[n]} x_i\right)$ for all $x \in \{\pm 1\}^n$.*

Lemma 5.2.8. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be symmetric and let $X_1, \cdots, X_n \in \mathbb{R}$ be independent and identically distributed. Then*

$$\mathop{\mathbb{E}}_{X}\left[f(X)(X_k - \mathbb{E}\left[X_k\right])\right] = \frac{1}{n}\mathop{\mathbb{E}}_{X}\left[f(X)\sum_{i\in[n]}(X_i - \mathbb{E}\left[X_i\right])\right]$$

for all $k \in [n]$.

Proof. By Definition 5.2.7,

$$\mathop{\mathbb{E}}_{X}\left[f(X)\sum_{i\in[n]}(X_i - \mathbb{E}\left[X_i\right])\right] = \sum_{i\in[n]}\mathop{\mathbb{E}}_{X}\left[f_*\left(\frac{1}{n}\sum_{k\in[n]}X_k\right)(X_i - \mathbb{E}\left[X_i\right])\right] \tag{5.1}$$

Since $X_1, \cdots, X_n$ are independent and identically distributed, the pair $(\sum_{i\in[n]} X_i, X_k)$ is identically distributed for all k. Thus $f_*\left(\frac{1}{n}\sum_{i\in[n]} X_i\right)(X_k - \mathbb{E}\left[X_k\right])$, being a function of $(\sum_{i\in[n]} X_i, X_k)$, is identically distributed for each k. Consequently, all the terms in (5.1) are the same, which implies the lemma. □

We can summarise our expectation bounds as follows:

Proposition 5.2.9. *Suppose the distribution $\mathcal{D}$ is a product distribution in which each marginal ρ is (ξ, n)-strong and satisfies $\mathop{\mathbb{E}}_{p\sim\rho}\left[1 - p^2\right] \geq \gamma + \alpha\cdot\xi$. Suppose the mechanism $\mathcal{M} : \{\pm 1\}^{n\times d} \to [-1, 1]^d$ is α-accurate. Let $x_1, \cdots x_n, z \sim \mathcal{P}_p$ and $q \sim \mathcal{M}(x_1, \cdots, x_n)$.*

1. *Then we have*

$$\forall j \in [d] \quad \mathop{\mathbb{E}}_{p, x_1, \cdots, x_n, z}\left[\sum_{i\in[n]}\left(\langle x_i^j, q^j\rangle - \langle z^j, q^j\rangle\right)\right] \geq \gamma.$$

Moreover, this bound holds even when conditioned on all the randomness in columns other than j. That is, the bound holds when we condition on any values of p^{-j}, $\{x_i^{-j}\}_{i=1,\ldots,n}$, z^{-j}, and q^{-j} and the expectation is only over the remaining variables.

2. *If, in addition, $\mathcal{M}$ is symmetric, then*

$$\forall j \in [d]\ \forall i \in [n] \quad \mathop{\mathbb{E}}_{p,x_1,\cdots,x_n,z}\left[\langle x_i^j, q^j\rangle - \langle z^j, q^j\rangle\right] \geq \frac{\gamma}{n}$$

and hence

$$\forall i \in [n] \quad \mathop{\mathbb{E}}_{p,x_1,\cdots,x_n,z,\mathcal{M}}\left[\langle x_i, q\rangle - \langle z, q\rangle\right] \geq \frac{\gamma d}{n}.$$

Proof. We view z^{-j}, q^{-j}, x_i^{-j} as fixed and we average over the coins of $\mathcal{M}$. Now the only randomness is the choice of p^j and $z^j, x_1^j \cdots x_n^j \sim p^j$. Since $\mathcal{M}$ does not see p^j or z^j, we can write $q^j = f(x^j)$ for some $f : \{\pm 1\}^n \to [-1, 1]$. By the assumption that $\mathcal{M}$ is α-accurate, $|f(x) - \overline{x}| \leq \alpha$ for all $x \in \{\pm 1\}^n$. The result now follows from Corollary 5.2.6 and Lemma 5.2.8. □

5.2.3 Completeness Analysis

Now that we have shown that $\mathbb{E}\left[\sum_{i\in[n]}\left(\langle x_i, q\rangle - \langle z, q\rangle\right)\right]$ is large, we can turn this into a high probability statement.

Lemma 5.2.10. *Suppose the distribution $\mathcal{D}$ is a product distribution in which each marginal ρ is (ξ, n)-strong and satisfies $\mathop{\mathbb{E}}_{p\sim\rho}\left[1 - p^2\right] \geq \gamma + \alpha \cdot \xi$. Suppose the mechanism $\mathcal{M} : \{\pm 1\}^{n\times d} \to [-1, 1]^d$ is α-accurate. Assume $d > O(n^2 \log(1/\delta)/\gamma^2)$. Let $x_1, \cdots x_n, z \sim \mathcal{P}_p$ and $q \sim \mathcal{M}(x_1, \cdots, x_n)$. Then*

$$\mathop{\mathbb{P}}_{p,x_{1\cdots n},z,\mathcal{M}}\left[\sum_{i\in[n]}\left(\langle x_i, q\rangle - \langle z, q\rangle\right) < \frac{\gamma}{2d}\right] \leq \delta.$$

Moreover, if $\mathcal{M}$ is symmetric, then

$$\forall i \in [n] \quad \mathop{\mathbb{P}}_{p,x_{1\cdots n},z,\mathcal{M}}\left[\langle x_i, q\rangle - \langle z, q\rangle < \frac{\gamma d}{2n}\right] \leq \delta.$$

The formal proof of this Lemma is quite involved, but unenlightening. To

preserve flow, we defer it to a later section (page 199) and give a proof sketch here instead.

Proof Sketch. Write

$$\sum_{i\in[n]} (\langle x_i, q\rangle - \langle z, q\rangle) = \sum_{j\in[d]} q^j \cdot \sum_{i\in[n]} (x_i^j - z^j) =: \sum_{j\in[d]} A_j.$$

We have $\mathbb{E}\left[A_j\right] \geq \gamma$ for all $j \in [d]$. Suppose the A_j random variables were independent. Then we could apply a Chernoff bound. Using $|A_j| \leq 2n$, gives

$$\mathbb{P}\left[\left|\sum_{j\in[d]} A_j\right| > \frac{1}{2}\gamma d\right] \leq \exp\left(-\frac{(\gamma d/2)^2}{(4n)^2 d}\right) \leq \delta,$$

as required. The second half of the lemma is similar.

The A_j variables are not independent, but it turns out their sum concentrates nonetheless. The key observation is that $\mathbb{E}\left[A_j\right] \geq \gamma$ even if we condition on $A_1, \cdots, A_{j-1}, A_{j+1}, \cdots, A_d$. Namely

$$\mathbb{E}\left[A_j \mid A_1 = a_1, \cdots, A_{j-1} = a_{j-1}, A_{j+1} = a_{j+1}, \cdots, A_d = a_d\right] \geq \gamma$$

for all $j \in [d]$ and $a \in \mathbb{R}^d$. □

Now we can finally prove completeness.

Proposition 5.2.11 (Completeness)**.** *Suppose the distribution $\mathcal{D}$ is a product distribution in which each marginal ρ is (ξ, n)-strong and satisfies $\mathop{\mathbb{E}}_{p\sim\rho}\left[1 - p^2\right] \geq \gamma + \alpha \cdot \xi$. Assume $d > O(n^2 \log(1/\delta)/\gamma^2)$. Suppose the mechanism $\mathcal{M} : \{\pm 1\}^{n\times d} \to [-1, 1]^d$ is α-accurate. Let $x_1, \cdots, x_n, z \sim \mathcal{P}_p$ and $q = \mathcal{M}(x_1, \cdots, x_n)$. Then*

$$\mathop{\mathbb{P}}_{p, x_{1\cdots n}, z, \mathcal{M}}\left[\exists i \in [n] \quad \mathcal{A}_{\delta,d}(x_i, q, z) = \text{IN}\right] \geq 1 - \delta.$$

Proof. By Lemma 5.2.10, $\sum_{i\in[n]} (\langle x_i, q\rangle - \langle z, q\rangle) \geq \frac{\gamma}{2d} > n \cdot \tau = n \cdot 2\sqrt{d \log(1/\delta)}$

with high probability. Thus, with high probability, we have $\langle x_i, q\rangle - \langle z, q\rangle > \tau$ for at least one $i \in [n]$. □

We also state the high-power completeness we get from assuming that $\mathcal{M}$ is symmetric.

Proposition 5.2.12 (High-Power Completeness)**.** *Suppose the distribution $\mathcal{D}$ is a product distribution in which each marginal ρ is (ξ, n)-strong and satisfies $\mathop{\mathbb{E}}_{p\sim\rho}\left[1-p^2\right] \geq \gamma + \alpha\cdot\xi$. Assume $d > O(n^2\log(1/\delta)/\gamma^2)$. Suppose the mechanism $\mathcal{M} : \{\pm 1\}^{n\times d} \to [-1,1]^d$ is α-accurate and symmetric. Let $x_1, \cdots x_n, z \sim \mathcal{P}_p$ and $q \sim \mathcal{M}(x_1, \cdots, x_n)$. Then*

$$\forall i \in [n] \quad \mathop{\mathbb{P}}_{p, x_1, \cdots, x_n, z, \mathcal{M}}\left[\mathcal{A}_{\delta,d}(x_i, q, z) = \text{IN}\right] \geq 1-\delta.$$

Proof. By Lemma 5.2.10, for all $i \in [n]$ we have $\langle x_i, q\rangle - \langle z, q\rangle \geq \frac{\gamma d}{2n} > \tau = 2\sqrt{d\log(1/\delta)}$ with high probability. Thus, for all $i \in [n]$, we have $\langle x_i, q\rangle - \langle z, q\rangle > \tau$ with high probability. □

5.2.4 Interpreting Strong Distributions

The notion of strong distributions is critical in the completeness analysis of our attack—it ensures that the output of $\mathcal{M}$ correlates with its input. In this section we show that this condition is met by a large class of distributions and give some intuition for its meaning. First we restate Definition 5.2.4.

Definition 5.2.4. *Let ρ be a probability distribution on $[-1,1]$. Define $h_n^\rho : \{-n-1, -n+1, \cdots, n+1\} \to \mathbb{R}$ by*

$$h_n^\rho(t) = \frac{(n+1+t)(n+1-t)}{2(n+1)} \cdot \mathop{\mathbb{P}}_{p\sim\rho, x_{1\cdots n+1}\sim p}\left[\sum_{i\in[n+1]} x_i = t\right].$$

We say ρ is (ξ, n)-strong *if*

$$\sum_{t \in \{-n, -n+2, \cdots, n\}} \left| h_n^\rho(t-1) - h_n^\rho(t+1) \right| \leq \xi.$$

We say ρ is ξ-strong *if ρ is (ξ, n)-strong for all n.*

To gain some intuition for the meaning of the definition, we consider some example distributions that do *not* satisfy the strong distribution assumption.

(i) Suppose ρ is a point mass on p^*. Then S_n^ρ is a (shifted and scaled) binomial distribution and h_n^ρ has high total variation. In this situation a simple mechanism $\mathcal{M}$ can prevent tracing: simply outputting $q = p^*$ will be accurate with high probability, but this allows the output of $\mathcal{M}$ to be (almost) independent from its input. Tracing is thus impossible, as there is almost no difference between the IN and OUT cases.

(ii) Example (i) can be generalised: Any distribution supported on a short interval is not strong. s

(iii) Suppose ρ is supported on two points p^* and p^{**} that are far apart. Then S_n^ρ is a convex combination of shifted and scaled binomial distributions.

This corresponds to a mechanism $\mathcal{M}$ that knows p^* and p^{**} and returns one of the two if they are sufficiently accurate. Again, with high probability, the output of $\mathcal{M}$ is not sensitive to changes in the input. That means the output of $\mathcal{M}$ does not contain much information that is specific to its input. This makes tracing impossible.

(iv) Example (iii) can be generalised to any distribution supported on a small number of points. This can be generalised further to distributions supported on many short intervals.

The above examples demonstrate what a strong distribution avoids. Instead a strong distribution is "spread out" and "smooth."

The function h_n^ρ in Definintion 5.2.4 is somewhat unintuitive. We can give an alternative definition:

Lemma 5.2.13. *Let $U_1, \cdots, U_{n+1} \in [-1,1]$ be independent uniformly random variables and let $P \sim \rho$ be independent from $U_1, U_2, \cdots, U_{n+1}$. Let $U_{(1)} \geq U_{(2)} \geq \cdots \geq U_{(n)}$ denote the random variables in sorted order. Set $U_{(0)} = +1$ and $U_{(n+2)} = -1$. Then*

$$\mathop{\mathbb{P}}_{p\sim\rho, x_{1\cdots n+1}\sim p}\left[\sum_{i\in[n+1]} x_i = 2k - n - 1\right] = \mathop{\mathbb{P}}_{U_0,\cdots,U_n,P}\left[U_{(k)} \geq P > U_{(k+1)}\right]$$

for all $k \in \{0, 1, \cdots, n+1\}$.

Thus the function h_n^ρ from Definition 5.2.4 can be defined as

$$h_n^\rho(t) = \frac{(n+1+t)(n+1-t)}{2(n+1)} \cdot \mathop{\mathbb{P}}_{U_0,\cdots,U_n,P}\left[U_{\left(\frac{t+n+1}{2}\right)} \geq P > U_{\left(\frac{t+n+3}{2}\right)}\right].$$

Intuitively, $U_{(0)} \leq U_{(1)} \leq \cdots \leq U_{(n+2)}$ partition the interval $[-1,1]$ into $n+2$ subintervals. Now h_n^ρ captures the amount of probability mass from ρ falling into each of these subintervals. However, the partitioning is itself random, so the probability mass at a particular point does not fall into a single subintervale. However, $U_{(k)} \approx \frac{n+2-2k}{n+2}$, so this random partitioning approximately partitions the interval evenly.

Proof of Lemma 5.2.13. Let $U_1, U_2, \cdots, U_{n+1}$ and P be sampled as in the lemma statement. Now define random variables $x_1, \cdots, x_{n+1} \in \{\pm 1\}$ by

$$x_i = 1 \iff U_i \geq P.$$

If we view P as fixed, then $\mathbb{P}[x_i = 1] = (P+1)/2$ and $\mathbb{E}[x_i] = P$ for each i. Moreover, the distribution of $x_1, \cdots, x_{n+1}$ is $n+1$ independent conditioned on P.

We claim that, for any $k \in \{0, 1, \cdots, n+1\}$.,

$$\sum_{i \in [n+1]} x_i = 2k - n - 1 \iff U_{(k)} \leq P \leq U_{(k+1)}.$$

The lemma follows from this claim, as we have shown a coupling between the two probability spaces under which the two events coincide.

To see the claim, note that $\sum_{i \in [n+1]} x_i = 2k - n - 1$ if and only if k of the x_is are set to $+1$, which happens if and only if there are k choices of $i \in [n+1]$ with $U_i \geq P$. In turn this is equivalent to saying that the k^{th} largest U_i is greater than or equal to P, but the $(k+1)^{\text{th}}$ largest U_i is not — i.e. $U_{(k)} \geq P > U_{(k+1)}$. □

We can also characterise the limiting case (i.e. $n \to \infty$ rather than fixed n):

Proposition 5.2.14. *Let $\rho : [-1, 1] \to \mathbb{R}$ be a continuously differentiable probability density function. Then ρ is a ξ-strong distribution if and only if*

$$\int_{-1}^{+1} \left| \frac{\mathrm{d}}{\mathrm{d}p}(1 - p^2)\rho(p) \right| \mathrm{d}p \leq \xi. \tag{5.2}$$

Proposition 5.2.14 shows that ρ being a strong probability density function is equivalent to a bound on the total variation of $(1 - p^2)\rho(p)$. This function should be contrasted with h_n^ρ in Definition 5.2.4. Indeed, Proposition 5.2.14 is simply the result of taking $n \to \infty$ in Definition 5.2.4.

Proof of Proposition 5.2.14. Lemma 5.2.5 provides an exact characterisation of (ξ, n)-strong distributions. Namely, ρ is (ξ, n)-strong if and only if

$$\left| \mathop{\mathbb{E}}_{p \sim \rho, x_{1 \cdots n} \sim p, z \sim p} \left[f(x) \sum_{i \in [n]} (x_i - z) \right] \right| \leq \xi \tag{5.3}$$

for all $f : \{\pm 1\}^n \to [-1, 1]$. To show that ρ is ξ-strong we must show that (5.3) holds for all n and all f.

Fix n and $f : \{\pm 1\}^n \to [-1,1]$. Define $g : [-1,1] \to [-1,1]$ by

$$g(p) = \mathop{\mathbb{E}}_{x_{1\cdots n}\sim p}[f(x)]\,.$$

By Lemma 4.3.7, for any $p \in [-1,1]$,

$$\mathop{\mathbb{E}}_{x_{1\cdots n}\sim p}\left[f(x)\cdot\sum_{i\in[n]}(x_i-p)\right] = g'(p)\cdot(1-p^2).$$

Thus

$$\begin{aligned}\mathop{\mathbb{E}}_{p\sim\rho,x_{1\cdots n}\sim p,z\sim p}\left[f(x)\sum_{i\in[n]}(x_i-z)\right] &= \mathop{\mathbb{E}}_{p\sim\rho}\left[g'(p)(1-p^2)\right]\\ &= \int_{-1}^{+1} g'(p)(1-p^2)\rho(p)\mathrm{d}p.\end{aligned}$$

Now we apply integration by parts — that is, we integrate both sides of an application of the differentiation product rule:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}p}g(p)(1-p^2)\rho(p) =& g'(p)(1-p^2)\rho(p) + g(p)\frac{\mathrm{d}}{\mathrm{d}p}(1-p^2)\rho(p).\\ \int_{-1}^{+1}\frac{\mathrm{d}}{\mathrm{d}p}g(p)(1-p^2)\rho(p)\mathrm{d}p =& \int_{-1}^{+1} g'(p)(1-p^2)\rho(p)\mathrm{d}p + \int_{-1}^{+1} g(p)\frac{\mathrm{d}}{\mathrm{d}p}(1-p^2)\rho(p)\mathrm{d}p.\\ \int_{-1}^{+1} g'(p)(1-p^2)\rho(p)\mathrm{d}p =& \big(g(1)(1-1^2)\rho(1) - g(-1)(1-(-1)^2)\rho(-1)\big)\\ &- \int_{-1}^{+1} g(p)\frac{\mathrm{d}}{\mathrm{d}p}(1-p^2)\rho(p)\mathrm{d}p.\end{aligned}$$

Finally, we can apply Hölder's inequality:

$$\begin{aligned}\left|\mathop{\mathbb{E}}_{p\sim\rho,x_{1\cdots n}\sim p,z\sim p}\left[f(x)\sum_{i\in[n]}(x_i-z)\right]\right| &= \left|\int_{-1}^{+1} g'(p)(1-p^2)\rho(p)\mathrm{d}p\right| \\ &= \left|\int_{-1}^{+1} g(p)\frac{\mathrm{d}}{\mathrm{d}p}(1-p^2)\rho(p)\mathrm{d}p\right| \\ &\leq \|g\|_\infty \cdot \int_{-1}^{+1}\left|\frac{\mathrm{d}}{\mathrm{d}p}(1-p^2)\rho(p)\right|\mathrm{d}p \\ &\leq \int_{-1}^{+1}\left|\frac{\mathrm{d}}{\mathrm{d}p}(1-p^2)\rho(p)\right|\mathrm{d}p.\end{aligned}$$

This proves one side of the equivalence. The other side of the equivalence follows from the tightness of Hölder's inequality and the fact that, by choosing n large enough, we can make $g : [-1,1] \to [-1,1]$ arbitrarily close to the function that makes the inequality tight. □

Using the differentiation product rule and the triangle inequality, we can show that

$$\begin{aligned}\int_{-1}^{+1}\left|\frac{\mathrm{d}}{\mathrm{d}p}(1-p^2)\rho(p)\right|\mathrm{d}p &= \int_{-1}^{+1}\left|(1-p^2)\rho'(p)-2p\rho(p)\right|\mathrm{d}p \\ &\leq \int_{-1}^{+1}(1-p^2)\left|\rho'(p)\right|\mathrm{d}p + \int_{-1}^{+1}\left|2p\rho(p)\right|\mathrm{d}p \\ &\leq \int_{-1}^{+1}\left|\rho'(p)\right|\mathrm{d}p + 2.\end{aligned}$$

Thus, rather than bounding the total variation of $(1-p^2)\rho(p)$, it suffices to bound the total variation of ρ.

A bound on the total variation of the probability density function is a very natural "smoothness" condition. In particular, the uniform distribution, whose probability density function is the constant $\frac{1}{2}$, has zero total variation. Thus Proposition 5.2.14 justifies our assertion that strong distributions correspond to a smoothness condition.

Using Proposition 5.2.14 we can give some examples of strong distributions:

- The uniform distribution on $[-1, 1]$ is 1-strong.
- The uniform distribution on $[a, b]$ is ξ-strong for

$$\xi = \frac{2 - a^2 - b^2 + \int_a^b |2x| \mathrm{d}x}{b - a} \leq \frac{2}{b - a} + 2.$$

- The (scaled) Beta distribution, with $\rho(p) \propto (1 + p)^{u-1}(1 - p)^{v-1}$ (where $u > 0$ and $v > 0$ and the support is $[-1, 1]$), is $(4uv/(u + v))$-strong.

5.3 Tracing from Fewer Statistics

In the previous section we focused on tracing from very weak assumptions—weakly accurate answers and only a single reference sample. The price of these weak assumptions is that we (provably) need $d = \Omega(n^2)$ and can only trace a single individual. In this section we show that if the mechanism gives more accurate answers, then we can trace with smaller d, and can trace many individuals in the dataset. In exchange, we require a larger reference sample. More precisely, we show that if the mechanism is α-accurate (for some $\alpha \geq n^{-1/2}$), and we are given roughly $1/\alpha^2$ independent reference samples from the distribution, then we can trace when the dataset has dimension only $O(\alpha^2 n^2)$, and we can successfully trace $\Omega(1/\alpha^2)$ individuals in the dataset. We summarise our results in the following informal theorem, which effectively generalises Theorem 5.1.1 from the introduction.

Theorem 5.3.1 (Informal). *For every $\delta > 0$, $n \in \mathbb{N}$, $\alpha \geq 1/n^{1/2}$, $d \geq O(\alpha^2 n^2 \log(1/\delta))$, $m \geq O(\log(n)/\alpha^2)$, and $t \leq \Omega(1/\alpha^2)$, there exists an attacker $\mathcal{A}^* : \{\pm 1\}^d \times [\pm 1]^d \times (\{\pm 1\}^d)^{m+1} \to \{\mathrm{IN}, \mathrm{OUT}\}$ the following holds.*

Let $\mathcal{D}$ be a product distribution on $[-1, 1]^d$ such that each marginal satisfies a technical smoothness condition (Definition 5.2.4). Let $\mathcal{M} : \{\pm 1\}^{n \times d} \to [-1, 1]^d$ be α-accurate. Let

$p \sim \mathcal{D}$ *and* $x_1, \cdots, x_n, y, z_0, z_1, \ldots, z_m \sim \mathcal{P}_p$. *Let* $q \sim \mathcal{M}(x_1, \cdots, x_n)$. *Then*

$$\mathbb{P}\left[\mathcal{A}^*(y, q, (z_0, z_1, \ldots, z_m)) = \text{IN}\right] \leq \delta, \text{and}$$

$$\mathbb{P}\left[\left|\{i \in [n] \mid \mathcal{A}^*(x_i, q, (z_0, z_1, \ldots, z_m)) = \text{IN}\}\right| \geq t\right] \geq 1 - \delta.$$

The modified attack is described below. In the attack, y represents the targeted individual, q is a vector of the mechanism's answers, and $z_0, z_1, \ldots, z_m$ represent $m + 1$ independent reference samples from the distribution. The first reference sample z_0 is used exactly as before as an unbiased estimate of p. The remaining m samples $z_1, \ldots, z_m$ will be averaged to form an independent unbiased estimate of p with much lower variance. We will set $m \approx 1/\alpha^2$ so that this estimate is α-accurate.

Input: $y, z_0, z_1, \ldots, z_m \in \{\pm 1\}^d$, and $q \in [\pm 1]^d$.
Let $z = z_0$ and $w = (1/m) \sum_{i=1}^{m} z_i$.
Let $\eta := 2\alpha$ and let $\lfloor q - w \rceil_\eta \in [-\eta, \eta]^d$ be the entrywise truncation of $q - w$, to $[-\eta, \eta]$. (We believe that this truncation is unnecessary, but it is needed for our analysis.)
Compute

$$\langle y - z, \lfloor q - w \rceil_\eta \rangle = \sum_{j \in [d]} (y^j - z^j) \cdot \left\lfloor q^j - w^j \right\rceil_\eta .$$

If $\langle y - z, \lfloor q - w \rceil_\eta \rangle > \tau := 4\alpha\sqrt{d \log(1/\delta)}$, output IN; otherwise output OUT.

Figure 5.2: ***Attack with a Large Reference Sample*** $\mathcal{A}^*_{\delta,\alpha,d,m}(y, q, \vec{z})$

5.3.1 Soundness

Proposition 5.3.2 (Soundness)**.** *Fix any* $q, z_1, \ldots, z_m, p \in [-1, 1]^d$. *Suppose* $y, z_0 \sim \mathcal{P}_p$ *are independent from each other and from* $q, z_1, \ldots, z_m$. *Then*

$$\mathbb{P}\left[\mathcal{A}^*_{\delta,\alpha,d,m}(y, q, \vec{z}) = \text{IN}\right] \leq \delta.$$

Proof. Since y and z_0 are identically distributed, and $q, z_1, \ldots, z_m$ are fixed

$$\mathbb{E}\left[\langle y - z, \lfloor q - w \rceil_\eta \rangle\right] = 0$$

(recall $z = z_0$ and $w = (1/m)\sum_{i=1}^m z_i$). Since y and z_0 are independent samples from a product distribution, we have that $\langle y - z, \lfloor q - w \rceil_\eta \rangle = \sum_{i \in [d]} (y^j - z^j) \cdot \lfloor q - w \rceil_\eta^j$ is the sum of $2d$ independent random variables, each of which is bounded by $\eta = 2\alpha$. Thus, by Hoeffding's inequality,

$$\mathbb{P}\left[\langle y - z, \lfloor q - w \rceil_\eta \rangle > \tau\right] \leq e^{-\tau^2/16d\alpha^2} \leq \delta.$$

This completes the proof. □

5.3.2 Correlation Analysis

We have the following proposition, analogous to Proposition 5.2.9 in Section 5.2.2.

Lemma 5.3.3. *Let $\mathcal{M} : \{\pm 1\}^{n \times d} \to [-1, 1]^d$ be α-accurate, let $\eta = 2\alpha$, and let the distribution $\mathcal{D}$ be a product distribution where every marginal ρ is (ξ, n)-strong and satisfies $\mathop{\mathbb{E}}_{p \sim \rho}\left[1 - p^2\right] \geq \gamma + \alpha\xi$. Consider the following experiment. Let $p \sim \mathcal{D}$, let $x_1, \ldots, x_n, z_0, z_1, \ldots, z_m \sim \mathcal{P}_p$, and $q \sim \mathcal{M}(x_1, \ldots, x_n)$. Then for every $j \in [d]$,*

$$\mathbb{E}\left[\sum_{i \in [n]} (x_i^j - z^j) \lfloor q - w \rceil_\eta^j\right] \geq \gamma - 4n \cdot e^{-\alpha^2 m/2},$$

where $z = z_0$ and $w = (1/m)\sum_{i=1}^m w_i$.

Moreover, this statement holds even when we condition on everything pertaining to columns other than j. That is, the bound on the expectation holds when we condition on any value of p^{-j}, $\{x_i^{-j}\}_{i=1,\ldots,n}$, $\{z_i^{-j}\}_{j=0,1,\ldots,m}$, and q^{-j} and the randomness is taken only over the remaining variables.

Proof. Since $\mathcal{M}$ is α-accurate and the distribution is (ξ, n)-strong, by Proposition

5.2.9

$$\mathbb{E}\left[\sum_{i\in[n]}(x_i^j - z^j)\cdot(q^j - w^j)\right] \geq \gamma.$$

So it remains to show that

$$\mathbb{E}\left[\sum_{i\in[n]}(x_i^j - z^j)(q^j - w^j - \lfloor q - w\rceil_\eta^j)\right] \leq 4ne^{-\alpha^2 m/2}.$$

Since $\left|\sum_{i\in[n]}(x_i^j - z^j)\cdot(q^j - w^j - \lfloor q^j - w^j\rceil_\eta)\right| \leq 4n$ and $\sum_{i\in[n]}(x_i^j - z^j)(q^j - w^j - \lfloor q^j - w^j\rceil_\eta) = 0$ when $|q^j - w^j| \leq \eta$, it suffices to show that $\mathbb{P}\left[|q^j - w^j| > \eta\right] \leq e^{-\alpha^2 m/2}$. By accuracy, we have $|q^j - p^j| \leq \alpha$, and by a Chernoff bound, we have $\mathbb{P}\left[|p^j - w^j| > \alpha\right] \leq e^{-\alpha^2 m/2}$. This completes the proof. □

Proposition 5.3.4. *Suppose the distribution $\mathcal{D}$ is a product distribution in which each marginal ρ is (ξ, n)-strong and satisfies $\mathbb{E}_{p\sim\rho}\left[1 - p^2\right] \geq \gamma + \alpha\xi$. Suppose $\mathcal{M} : \{\pm 1\}^{n\times d} \to [-1,1]^d$ is α-accurate. Let $d > O(\alpha^2 n^2 \log(1/\delta)/\gamma^2)$ and $m \geq 2\log(24n/\gamma)/\alpha^2$. Let $x_1, \ldots, x_n, z_0, z_1, \ldots, z_m \sim \mathcal{P}_p$. Let $q \sim \mathcal{M}(x_1, \ldots, x_n)$ Then*

$$\mathbb{P}\left[\sum_{i\in[n]}\left(\langle x_i - z, \lfloor q - w\rceil_\eta\rangle\right) < \frac{\gamma d}{2}\right] \leq \delta$$

(recall $z = z_0$, $w = (1/m)\sum_{i=1}^m z_i$, and $\eta = 2\alpha$).

The proof of Proposition 5.3.4 is analogous to that of Lemma 5.2.10 and is presented in Section 5.4.2.

Proposition 5.3.4 establishes a lower bound on the sum of the expected scores. Next we will upper bound the 2-norm of the expected scores. Upper bounding the 2-norm will establish that the scores are "spread out," so there must be many (roughly $1/\alpha^2$) expected scores that are large (larger than the threshold τ).

Our analysis relies on the following technical lemma.

Lemma 5.3.5. *Let $X_1, \cdots, X_n \in \mathbb{R}$ be independent random variables such that $\mathbb{E}[X_i] = 0$ and $\mathbb{E}\left[X_i^2\right] \leq 1$ for every $i \in [n]$. Let $Y \in \mathbb{R}$ be another (not necessarily independent) random variable. Then*

$$\sum_{i \in [n]} \mathbb{E}[X_i Y]^2 \leq \mathbb{E}\left[Y^2\right].$$

Proof. For $i \in [n]$, let $c_i = \mathbb{E}[X_i Y]$. Define $h : \mathbb{R}^n \to \mathbb{R}$ by $h(x) = \sum_{i \in [n]} c_i x_i$. Then

$$\mathbb{E}\left[h(X)^2\right] = \sum_{i,j \in [n]} c_i c_j \mathbb{E}\left[X_i X_j\right] \leq \sum_{i \in [n]} c_i^2$$

and

$$\mathbb{E}[h(X)Y] = \sum_{i \in [n]} c_i \mathbb{E}[X_i Y] = \sum_{i \in [n]} c_i^2.$$

Thus

$$0 \leq \mathbb{E}\left[(h(X) - Y)^2\right] = \mathbb{E}\left[h(X)^2\right] - 2\mathbb{E}[h(X)Y] + \mathbb{E}\left[Y^2\right] \leq \sum_{i \in [n]} c_i^2 - 2\sum_{i \in [n]} c_i^2 + \mathbb{E}\left[Y^2\right].$$

Rearranging gives

$$\sum_{i \in [n]} c_i^2 \leq \mathbb{E}\left[Y^2\right],$$

as required. □

Lemma 5.3.6. *Fix $p \in [-1,1]^d$ and let $\mathcal{M} : \{\pm 1\}^{n \times d} \to [-1,1]^d$ be any mechanism. Fix any w and let $x_1, \cdots, x_n, z_0 \sim \mathcal{P}_p$ and $q \sim \mathcal{M}(x_1, \cdots, x_n)$. Then for every $j \in [d]$,*

$$\sqrt{\sum_{i \in [n]} \mathbb{E}\left[\langle x_i^j - z^j, \lfloor q^j - w^j \rceil_\eta \rangle\right]^2} \leq \eta\sqrt{2}$$

(recall $z = z_0$). Moreover, this statement holds even when we condition on everything pertaining to columns other than j. That is, the bound holds when we condition the expectations on any value of $\{x_i^{-j}\}_{i=1,\ldots,n}, z_0^{-j}$, and q^{-j} and the randomness is taken only over the remaining variables.

Proof. We apply Lemma 5.3.5 with $X_i = x_i^j - z^j$ and $Y = \lfloor q^j - w^j \rceil_\eta$. □

Once again, we would like to apply a concentration result to turn our bound on the sum of the squares of the expected scores into a high confidence bound on the sum of the squares of the scores themselves. Once again, this issue is complicated by a lack of independence. Nonetheless, we prove a suitable concentration bound for the sum of the squares of the scores in Proposition 5.4.11. Using this concentration bound we can prove the following.

Proposition 5.3.7. *Fix $p \in [-1,1]^d$ and let $\mathcal{M} : \{\pm 1\}^{n \times d} \to [-1,1]^d$ be any mechaniam. Assume $d \geq 64(n + \sqrt{\log(1/\delta)})$. Let $x_1, \cdots, x_n, z_0, z_1, \cdots, z_m \sim \mathcal{P}_p$, and let $q \sim \mathcal{M}(x_1, \cdots, x_n)$. Then*

$$\mathbb{P}\left[\sqrt{\sum_{i\in[n]} \langle x_i - z, \lfloor q - w \rceil_\eta \rangle^2} \leq 2\eta d\right] \geq 1 - \delta$$

(recall $z_0 = z$ and $w = (1/m)\sum_{i=1}^n z_i$).

Proof. By applying the triangle inequality to Lemma 5.3.6, we have

$$\sqrt{\sum_{i\in[n]} \mathbb{E}\left[\langle x_i - z, \lfloor q - w \rceil_\eta \rangle\right]^2} \leq d\eta\sqrt{2}.$$

By Theorem 5.4.11, for any $\lambda > 0$,

$$\mathbb{P}\left[\sqrt{\sum_{i\in[n]} \langle x_i - z, \lfloor q - w \rceil_\eta \rangle^2} > \lambda + d\eta\sqrt{2}\right] \leq \exp\left(\frac{nd}{2} - \frac{\lambda^2}{16\eta^2}\right).$$

The theorem follows by setting $\lambda = 4\eta\sqrt{\frac{nd}{2} + \log(1/\delta)} \leq \frac{\eta d}{2}$. □

Combining Proposition 5.3.4 with Proposition 5.3.7, we can show that, with high probability, the attack says IN for many target individuals x_i. To do so, we need the following elementary lemma.

Lemma 5.3.8. *Let $\sigma \in \mathbb{R}^n$ satisfy $\sum_{i\in[n]} \sigma_i \geq A$ and $\sum_{i\in[n]} \sigma_i^2 \leq B^2$. Then*

$$\left|\left\{i \in [n] : \sigma_i > \frac{A}{2n}\right\}\right| \geq \left(\frac{A}{2B}\right)^2.$$

Proof. Let $\tau = A/2n$ and $S = \{i \in [n] : \sigma_i > \tau\}$. Let $\sigma_S \in \mathbb{R}^{|S|}$ denote the restriction of σ onto the coordinates indexed by S. Then

$$\begin{aligned} A \leq \sum_{i\in[n]} \sigma_i &= \sum_{i\in[n]\setminus S} \sigma_i + \sum_{i\in S} \sigma_i \\ &\leq (n - |S|)\tau + \|\sigma_S\|_1 \\ &\leq n\tau + \sqrt{|S|} \cdot \|\sigma_S\|_2 \\ &\leq n\tau + \sqrt{|S|} \cdot \|\sigma\|_2 \\ &\leq n\tau + \sqrt{|S|} \cdot B. \end{aligned}$$

Rearranging gives

$$|S| \geq \left(\frac{A - n\tau}{B}\right)^2 = \left(\frac{A}{2B}\right)^2,$$

as required. □

Proposition 5.3.9 (Completeness with a Large Reference Sample)**.** *Suppose the distribution $\mathcal{D}$ is a product distribution in which each marginal ρ is (ξ, n)-strong and satisfies $\mathop{\mathbb{E}}_{p\sim\rho}\left[1 - p^2\right] \geq \gamma + \alpha\xi$. Suppose $\mathcal{M} : \{\pm 1\}^{n\times d} \to [-1, 1]^d$ is α-accurate. Let $d > O(\alpha^2 n^2 \log(1/\delta)/\gamma^2)$ and $m \geq 2\log(24n/\gamma)/\alpha^2$. Let $x_1, \ldots, x_n, z_0, z_1, \ldots, z_n \sim \mathcal{P}_p$. Let $q \sim \mathcal{M}(x_1, \ldots, x_n)$. Then*

$$\mathbb{P}\left[\left|\left\{i \in [n] : \mathcal{A}^*_{\delta,\alpha,d,m}(x_i, q, \vec{z}) = \text{IN}\right\}\right| \geq \frac{\gamma^2}{256\alpha^2}\right] \geq 1 - 2\delta.$$

Proof. By Proposition 5.3.4, with probability at least $1 - \delta$,

$$\sum_{i\in[n]} \left(\langle x_i - z, \lfloor q - w \rceil_\eta \rangle\right) \geq \frac{\gamma d}{2} =: A.$$

By Proposition 5.3.7, with probability at least $1 - \delta$,

$$\sqrt{\sum_{i \in [n]} \langle x_i - z, \lfloor q - w \rceil_\eta \rangle^2} \leq 2\eta d =: B.$$

By a union bound, both of these events occur with probability at least $1 - 2\delta$. Assuming they both occur, Lemma 5.3.8 implies

$$\left| \left\{ i \in [n] : \langle x_i - z, \lfloor q - w \rceil_\eta \rangle \geq \frac{A}{2n} \right\} \right| \geq \left(\frac{A}{2B} \right)^2 = \left(\frac{\gamma}{16\alpha} \right)^2.$$

We have $A/2n = \gamma d / 4n \geq \tau = 4\alpha\sqrt{d \log(1/\delta)}$, which implies the result. □

5.4 Concentration Bounds

The following concentration result implies the concentration results we use in the earlier sections.

Theorem 5.4.1. *Let $X \in \mathbb{R}^{n \times d}$ be a random matrix such that the columns are independent—that is, $X^1, X^2, \cdots, X^d \in \mathbb{R}^n$ are independent random variables. Let $Y \in \mathbb{R}^d$ be a random variable that possibly depends on X. Suppose that $\mathop{\mathbb{E}}_X \left[e^{tX_{i,j}} \right] \leq e^{ct^2}$ for all $t \in \mathbb{R}$, $i \in [n]$, and $j \in [d]$. Assume $\|Y\|_\infty \leq \alpha$ with certainty. Let $a \in \mathbb{R}^n$. Define*

$$Z_j = (a^T X)_j Y_j \in \mathbb{R} \qquad \text{and} \qquad Z = \sum_{j \in [d]} Z_j = a^T X Y \in \mathbb{R}.$$

Suppose $\mathbb{E}\left[Z_j \mid Z_{j+1} = z_{j+1}, \cdots, Z_d = z_d \right] \geq \gamma_j$ for all $j \in [d]$ and $z \in \mathbb{R}^d$. Let $\gamma = \sum_{j \in [d]} \gamma_j$. Then

$$\mathop{\mathbb{P}}_Z \left[Z < \gamma - \lambda \right] \leq \exp\left(\frac{-\lambda^2}{16cd\alpha^2 \|a\|_1^2} \right)$$

for all $\lambda > 0$.

In Lemma 5.2.10, $X_{i,j} = x_i^j - z^j$, $Y = q = M(x)$. The vector a specifies which subset of scores we are interested in (either $a = \vec{1}_n$ for the sum of all scores or $a = e_i$

for a single score).

Note that if X has bounded entries, it satisfies the condition of Theorem 5.4.1:

Lemma 5.4.2 (Hoeffding's Lemma)**.** *Let $X \in [a, b]$ be a random variable. Then*

$$\mathbb{E}\left[e^{t(X-\mathbb{E}[X])}\right] \leq e^{t^2(b-a)^2/8}$$

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

Likewise, if X has Gaussian entries, we can apply Theorem 5.4.1:

Lemma 5.4.3. *Let g be a standard Gaussian. Then, for all $t \in \mathbb{R}$, $\mathop{\mathbb{E}}_{g}\left[e^{tg}\right] = e^{t^2/2}$ and, if $0 \leq t < 1/2$, $\mathop{\mathbb{E}}_{g}\left[e^{tg^2}\right] = 1/\sqrt{1-2t}$.*

To prove Theorem 5.4.1 we need the following lemmas.

Lemma 5.4.4 (Hölder's Inequality)**.** *Let $X_1, \cdots, X_n \in \mathbb{R}$ be (possibly dependent) random variables. Let $\alpha_*, \alpha_1, \cdots, \alpha_n \in [1, \infty]$ with $1/\alpha_* = \sum_{i\in[n]} 1/\alpha_i$. Then*

$$\mathop{\mathbb{E}}_{X}\left[e^{\alpha_* \sum_{i\in[n]} X_i}\right]^{1/\alpha_*} \leq \prod_{i\in[n]} \mathop{\mathbb{E}}_{X_i}\left[e^{\alpha_i X_i}\right]^{1/\alpha_i}.$$

Lemma 5.4.5. *Let $X, Y \in \mathbb{R}$ be (possibly dependent) random variables. Suppose $|Y| \leq 1$. Then*

$$\mathop{\mathbb{E}}_{X,Y}\left[e^{XY-\mathop{\mathbb{E}}_{X,Y}[XY]}\right] \leq \mathop{\mathbb{E}}_{X,\xi}\left[e^{2\xi X}\right],$$

where $\xi \in \{\pm 1\}$ is uniform and independent of X and Y.

Proof. By Lemma 2.4.4,

$$\mathop{\mathbb{E}}_{X,Y}\left[e^{XY-\mathop{\mathbb{E}}_{X,Y}[XY]}\right] \leq \mathop{\mathbb{E}}_{X,Y,\xi}\left[e^{2\xi XY}\right] = \mathop{\mathbb{E}}_{X,Y,\xi}\left[e^{2\xi X|Y|}\right].$$

Define $\zeta \in \{0, 1\}$ to be a "randomised rounding" of $|Y|$, namely $\mathop{\mathbb{E}}_{\zeta}[\zeta \mid Y] = |Y|$. By Jensen's inequality,

$$\mathop{\mathbb{E}}_{X,Y}\left[e^{XY-\mathop{\mathbb{E}}_{X,Y}[XY]}\right] \leq \mathop{\mathbb{E}}_{X,Y,\xi}\left[e^{2\xi X|Y|}\right] = \mathop{\mathbb{E}}_{X,Y,\xi}\left[e^{\mathop{\mathbb{E}}_{\zeta}[2\xi X\zeta]}\right] \leq \mathop{\mathbb{E}}_{X,Y,\xi,\zeta}\left[e^{2\xi X\zeta}\right].$$

By Jensen's inequality $e^0 = e^{\mathop{\mathbb{E}}_{\xi}[2\xi X]} \leq \mathop{\mathbb{E}}_{\xi}\left[e^{2\xi X}\right]$. Thus

$$\begin{aligned}\mathop{\mathbb{E}}_{X,Y}\left[e^{XY - \mathop{\mathbb{E}}_{X,Y}[XY]}\right] \leq & \mathop{\mathbb{E}}_{X,Y,\xi,\zeta}\left[e^{2\xi X\zeta}\right] = \mathop{\mathbb{E}}_{X,Y}\left[\mathop{\mathbb{P}}_{\zeta}[\zeta = 0]\, e^0 + \mathop{\mathbb{P}}_{\zeta}[\zeta = 1] \mathop{\mathbb{E}}_{\xi}\left[e^{2\xi X}\right]\right] \\ \leq & \mathop{\mathbb{E}}_{X,Y}\left[\mathop{\mathbb{E}}_{\xi}\left[e^{2\xi X}\right]\right] = \mathop{\mathbb{E}}_{X,\xi}\left[e^{2\xi X}\right],\end{aligned}$$

as required. □

Lemma 5.4.6. *Let $X \in \mathbb{R}^n$ be a random variable. Suppose $\mathbb{E}\left[e^{tX_i}\right] \leq e^{ct^2}$ for all $t \in \mathbb{R}$ and $i \in [n]$. Let $Y \in [-\alpha, \alpha]$ be a random variable that possibly depends on X . Let $a \in \mathbb{R}^n$. Define*

$$Z = a^T XY \in \mathbb{R}.$$

Then

$$\mathbb{E}\left[e^{t(Z - \mathbb{E}[Z])}\right] \leq e^{4c\alpha^2 t^2 \|a\|_1^2}$$

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

Proof. We may assume, without loss of generality, that $\|a\|_1 = 1$ and $\alpha = 1$. By assumption, $\mathbb{E}\left[e^{ta_i X_i}\right] \leq e^{ct^2 a_i^2}$ for all $i \in [n]$ and $t \in \mathbb{R}$. Now we apply Lemma 5.4.4 with $\alpha_i = 1/|a_i| \in [1, \infty]$ and $\alpha_* = 1$:

$$\mathbb{E}\left[e^{ta^T X}\right] = \mathbb{E}\left[e^{\alpha_* t a^T X}\right]^{1/\alpha_*} \leq \prod_{i \in [n]} \mathbb{E}\left[e^{\alpha_i t a_i X_i}\right]^{1/\alpha_i} \leq \prod_{i \in [n]} e^{c\alpha_i t^2 a_i^2} = e^{ct^2 \|a\|_1} = e^{ct^2}$$

for all $t \in \mathbb{R}$. By Lemma 5.4.5,

$$\mathop{\mathbb{E}}_{Z}\left[e^{t(Z - \mathbb{E}[Z])}\right] = \mathop{\mathbb{E}}_{X,Y}\left[e^{ta^T XY - \mathop{\mathbb{E}}_{X,Y}\left[ta^T XY\right]}\right] \leq \mathop{\mathbb{E}}_{X,\xi}\left[e^{2\xi t a^T X}\right] \leq e^{4ct^2}$$

for all $t \in \mathbb{R}$, as required. □

Lemma 5.4.7. *Let $X \in \mathbb{R}^{n \times d}$ be a random variable such that the columns are independent. Suppose that $\mathbb{E}\left[e^{tX_{i,j}}\right] \leq e^{ct^2}$ for all $t \in \mathbb{R}$, $i \in [n]$, and $j \in [d]$. Let $Y \in [-\alpha, \alpha]^d$ be a*

random variable that possibly depends on X. Let $a \in \mathbb{R}^n$. For $j \in [d]$, define

$$Z_j = (a^T X)_j Y_j \in \mathbb{R} \qquad \textit{and} \qquad \mu_j(z) = \mathbb{E}\left[Z_j \mid Z_{j+1} = z_{j+1}, \cdots, Z_d = z_d\right].$$

Let $Z = \sum_{j\in[d]} Z_j = a^T XY$ and $\mu(z) = \sum_{j\in[d]} \mu_j(z)$. Then

$$\mathbb{E}\left[e^{t(Z-\mu(Z))}\right] \leq e^{4cd\alpha^2 t^2 \|a\|_1^2}$$

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

Proof. Firstly, by Lemma 5.4.6,

$$\begin{aligned}
&\mathbb{E}\left[e^{t(Z_j-\mu_j(z))} \mid Z_{j+1} = z_{j+1}, \cdots, Z_d = z_d\right] \\
&= \mathbb{E}\left[e^{t(Z_j-\mathbb{E}\left[Z_j \mid Z_{j+1}=z_{j+1},\cdots,Z_d=z_d\right])} \mid Z_{j+1} = z_{j+1}, \cdots, Z_d = z_d\right] \\
&\leq e^{4c\alpha^2 t^2 \|a\|_1^2}
\end{aligned}$$

for all $t \in \mathbb{R}$, $j \in [d]$, and $z \in \mathbb{R}^d$.

Now we prove by induction on $k \in [d]$ that

$$\mathbb{E}\left[e^{t\sum_{j\in[k]} Z_j-\mu_j(Z)} \mid Z_{k+1} = z_{k+1}, \cdots, Z_d = z_d\right] \leq e^{4ck\alpha^2 t^2 \|a\|_1^2}$$

for all $t \in \mathbb{R}$ and $z \in \mathbb{R}^d$, from which the lemma follows by setting $k = d$.

The base case $k = 1$ is immediate from Lemma 5.4.6. Finally, the induction step:

$$\begin{aligned}
&\mathbb{E}\left[e^{t\sum_{j\in[k]} Z_j - \mu_j(Z)} \mid Z_{k+1} = z_{k+1}, \cdots, Z_d = z_d\right] \\
&= \sum_{z_k^*} \mathbb{P}\left[Z_k = z_k^*\right] \mathbb{E}\left[e^{t\sum_{j\in[k-1]} Z_j - \mu_j(Z)} \cdot e^{t(Z_k - \mu_k(Z))} \mid Z_k = z_k^*, Z_{k+1} = z_{k+1}, \cdots, Z_d = z_d\right] \\
&= \sum_{z_k^*} \mathbb{P}\left[Z_k = z_k^*\right] \cdot e^{t(z_k^* - \mu_k(z))} \cdot \mathbb{E}\left[e^{t\sum_{j\in[k-1]} Z_j - \mu_j(Z)} \mid Z_k = z_k^*, Z_{k+1} = z_{k+1}, \cdots, Z_d = z_d\right] \\
&\leq \sum_{z_k^*} \mathbb{P}\left[Z_k = z_k^*\right] \cdot e^{t(z_k^* - \mu_k(z))} \cdot e^{4c(k-1)\alpha^2 t^2 \|a\|_1^2} \\
&= \mathbb{E}\left[e^{t(Z_k - \mu_k(z))} \mid Z_{k+1} = z_{k+1}, \cdots, Z_d = z_d\right] \cdot e^{4c(k-1)\alpha^2 t^2 \|a\|_1^2} \\
&\leq e^{4ct^2\|a\|_1^2} \cdot e^{4c(k-1)t^2\|a\|_1^2},
\end{aligned}$$

as required. □

Proof of Theorem 5.4.1. The assumption that $\mathbb{E}\left[Z_j \mid Z_{j+1} = z_{j+1}, \cdots, Z_d = z_d\right] \geq \gamma_j$ for all $j \in [d]$ and $z \in \mathbb{R}^d$. Implies $\mu_j(Z) \geq \gamma_j$ with certainty. Thus it remains to show that Z is close to $\mu(Z)$.

By Markov's inequality and Lemma 5.4.7,

$$\mathbb{P}\left[Z - \mu(Z) > \lambda\right] \leq \frac{\mathbb{E}\left[e^{t(Z-\mu(Z))}\right]}{e^{t\lambda}} \leq \frac{e^{4cd\alpha^2 t^2 \|a\|_1^2}}{e^{t\lambda}}.$$

Setting $t = \frac{\lambda}{8cd\alpha^2\|a\|_1^2}$ gives

$$\mathbb{P}\left[Z - \mu(Z) > \lambda\right] \leq e^{-t\lambda/2} = e^{\frac{-\lambda^2}{16cd\alpha^2\|a\|_1^2}},$$

as desired. □

5.4.1 Concentration of 2-Norm

Lemma 5.4.8. *Let $X \in \mathbb{R}^n$ be a product distribution. Suppose $\mathbb{E}\left[e^{tX_i}\right] \leq e^{ct^2}$ for all $t \in \mathbb{R}$ and $i \in [n]$. Let $Y \in [-\alpha, \alpha]$ be a random variable that possibly depends on X. Let $a \in \mathbb{R}^n$. Define*

$$Z = a^T XY \in \mathbb{R}.$$

Then

$$\mathbb{E}\left[e^{t(Z-\mathbb{E}[Z])}\right] \leq e^{4c\alpha^2 t^2 \|a\|_2^2}$$

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

Proof. We may assume, without loss of generality, that $\alpha = 1$. By assumption, $\mathbb{E}\left[e^{ta_i X_i}\right] \leq e^{ct^2 a_i^2}$ for all $i \in [n]$ and $t \in \mathbb{R}$. By independence,

$$\mathbb{E}\left[e^{ta^T X}\right] = \prod_{i\in[n]} \mathbb{E}\left[e^{ta_i X_i}\right] \leq \prod_{i\in[n]} e^{ct^2 a_i^2} = e^{ct^2\|a\|_2^2}$$

for all $t \in \mathbb{R}$. By Lemma 5.4.5,

$$\mathop{\mathbb{E}}_{Z}\left[e^{t(Z-\mathbb{E}[Z])}\right] = \mathop{\mathbb{E}}_{X}\left[e^{ta^T XY - \mathop{\mathbb{E}}_{X}[ta^T XY]}\right] \leq \mathop{\mathbb{E}}_{X,\xi}\left[e^{2\xi t a^T X}\right] \leq e^{4ct^2\|a\|_2^2}$$

for all $t \in \mathbb{R}$, as required. □

Lemma 5.4.9. *Let $X \in \mathbb{R}^n$ be a product distribution. Suppose $\mathbb{E}\left[e^{tX_i}\right] \leq e^{ct^2}$ for all $t \in \mathbb{R}$ and $i \in [n]$. Let $Y \in [-\alpha, \alpha]$ be a random variable that possibly depends on X. Define*

$$\vec{V} = XY \in \mathbb{R}^n.$$

Then

$$\mathbb{E}\left[e^{\frac{t^2}{2}\left\|\vec{V}-\mathbb{E}\left[\vec{V}\right]\right\|_2^2}\right] \leq e^{8nc\alpha^2 t^2}$$

for all $t \in [-1/4\sqrt{c}\alpha, 1/4\sqrt{c}\alpha]$.

Proof. Let $g \in \mathbb{R}^n$ be a standard multivariate Gaussian and

$$Z = g^T(V - \mathbb{E}[V]) \in \mathbb{R}.$$

By Lemma 5.4.3,

$$\mathop{\mathbb{E}}_{g}\left[e^{tZ}\right] = \prod_{i\in[n]} \mathop{\mathbb{E}}_{g_i}\left[e^{t(V_i - \mathbb{E}[V_i])g_i}\right] = \prod_{i\in[n]} e^{t^2(\vec{V}_i - \mathbb{E}[\vec{V}_i])^2/2} = e^{t^2\left\|\vec{V} - \mathbb{E}[\vec{V}]\right\|_2^2/2}.$$

By Lemmas 5.4.8 and 5.4.3,

$$\mathop{\mathbb{E}}_{g,V}\left[e^{tZ}\right] \le \mathop{\mathbb{E}}_{g}\left[e^{4c\alpha^2t^2\|g\|_2^2}\right] = \prod_{i\in[n]} \mathop{\mathbb{E}}_{g}\left[e^{4c\alpha^2t^2g_i^2}\right] = \left(\frac{1}{\sqrt{1 - 2\cdot 4c\alpha^2t^2}}\right)^n,$$

assuming $0 \le 4c\alpha^2t^2 < 1/2$. Thus

$$\mathop{\mathbb{E}}_{V}\left[e^{t^2\left\|V - \mathbb{E}[V]\right\|_2^2/2}\right] \le \left(\frac{1}{\sqrt{1 - 8c\alpha^2t^2}}\right)^n \le e^{8nc\alpha^2t^2},$$

as $1/\sqrt{1-x} \le e^x$ for $0 \le x \le 1/2$. □

Lemma 5.4.10. *Let $X \in \mathbb{R}^{n\times d}$ be a product distribution. Suppose $\mathbb{E}\left[e^{tX_{i,j}}\right] \le e^{ct^2}$ for all $t \in \mathbb{R}$, $i \in [n]$, and $j \in [d]$. Let $Y \in [-\alpha,\alpha]^d$ be a random variable that possibly depends on X. For $j \in [d]$, define*

$$\vec{V}^j = X^jY^j \in \mathbb{R}^n \qquad \text{and} \qquad \mu^j(\vec{v}^1,\cdots,\vec{v}^d) = \mathbb{E}\left[\vec{V}^j \mid \vec{V}^{j+1} = \vec{v}^{j+1},\cdots,\vec{V}^d = \vec{v}^d\right].$$

Let $\vec{V} = \sum_{j\in[d]} \vec{V}^j$ and $\mu(\vec{v}^1,\cdots,\vec{v}^d) = \sum_{j\in[d]} \mu^j(\vec{v}^1,\cdots,\vec{v}^d)$. Then

$$\mathbb{E}\left[e^{\frac{t^2}{2}\left\|\vec{V} - \mu(\vec{V}^1,\cdots,\vec{V}^d)\right\|_2^2}\right] \le e^{8ndc\alpha^2t^2}$$

for all $t \in [-1/4\sqrt{c}\alpha, 1/4\sqrt{c}\alpha]$.

The proof is analogous to that of Lemma 5.4.7.

Theorem 5.4.11. *Let $X \in \mathbb{R}^{n\times d}$ be a random matrix with independent entries. Suppose $\mathbb{E}\left[e^{tX_{i,j}}\right] \le e^{ct^2}$ for all $t \in \mathbb{R}$, $i \in [n]$, and $j \in [d]$. Let $Y \in [-\alpha,\alpha]^d$ be a random variable*

that possibly depends on X. For $j \in [d]$, define

$$\vec{V}^j = X^j Y^j \in \mathbb{R}^n.$$

Suppose that, for all $j \in [d]$ and $\vec{v}^1, \cdots, \vec{v}^d \in \mathbb{R}^n$,

$$\left\| \mathbb{E}\left[\vec{V}^j \mid \vec{V}^1 = \vec{v}^1, \cdots, \vec{V}^{j-1} \vec{v}^{j-1}, \vec{V}^{j+1} \vec{v}^{j+1}, \cdots, \vec{V}^d \vec{v}^d \right] \right\|_2 \leq \beta_j.$$

Let $\vec{V} = \sum_{j \in [d]} \vec{V}^j$. Then

$$\mathbb{P}\left[\left\| \vec{V} \right\|_2 > \lambda + \sum_{j \in [d]} \beta_j \right] \leq e^{\frac{nd}{2} - \frac{\lambda^2}{32c\alpha^2}}$$

for all $\lambda > 0$.

Proof. Let

$$\mu^j(\vec{v}^1, \cdots, \vec{v}^d) = \mathbb{E}\left[\vec{V}^j \mid \vec{V}^{j+1} = \vec{v}^{j+1}, \cdots, \vec{V}^d = \vec{v}^d \right]$$

for $j \in [d]$ and

$$\mu(\vec{v}^1, \cdots, \vec{v}^d) = \sum_{j \in [d]} \mu^j(\vec{v}^1, \cdots, \vec{v}^d).$$

By assumption $\left\| \mu^j(\vec{v}^1, \cdots, \vec{v}^d) \right\|_2 \leq \beta_j$ for all $j \in [d]$. Thus $\left\| \mu(\vec{v}^1, \cdots, \vec{v}^d) \right\|_2 \leq \sum_{j \in [d]} \beta_j$ by the triangle inequality. By Lemma 5.4.10, $\mathbb{E}\left[e^{\frac{t^2}{2} \left\| \vec{V} - \mu(\vec{V}^1, \cdots, \vec{V}^d) \right\|_2^2} \right] \leq e^{8ndc\alpha^2 t^2}$ for all $t \in [-1/4\sqrt{c}\alpha, 1/4\sqrt{c}\alpha]$. Thus, by Markov's inequality,

$$\mathbb{P}\left[\left\| \vec{V} - \mu(\vec{V}^1, \cdots, \vec{V}^d) \right\|_2 \geq \lambda \right] \leq \frac{\mathbb{E}\left[e^{\frac{t^2}{2} \left\| \vec{V} - \mu(\vec{V}^1, \cdots, \vec{V}^d) \right\|_2^2} \right]}{e^{\frac{t^2}{2}\lambda^2}} \leq e^{(8ndc\alpha^2 - \lambda^2/2)t^2}.$$

Setting $t = 1/4\sqrt{c}\alpha$ gives the result. □

5.4.2 Proofs of Concentration Lemmas

Now we prove the various concentration lemmas we need.

Lemma 5.4.12 (Restating Lemma 5.2.10)**.** *Suppose the distribution $\mathcal{D}$ is a product distribution in which each marginal ρ is (ξ, n)-strong and satisfies $\mathop{\mathbb{E}}_{p\sim\rho}\left[1-p^2\right] \geq \gamma + \alpha\cdot\xi$. Suppose the mechanism $\mathcal{M} : \{\pm 1\}^{n\times d} \to [-1,1]^d$ is α-accurate. Assume $d > O(n^2\log(1/\delta)/\gamma^2)$. Let $x_1, \cdots x_n, z \sim \mathcal{P}_p$ and $q \sim \mathcal{M}(x_1, \cdots, x_n)$. Then*

$$\mathop{\mathbb{P}}_{p, x_{1\cdots n}, z, \mathcal{M}}\left[\sum_{i\in[n]}\left(\langle x_i, q\rangle - \langle z, q\rangle\right) < \frac{\gamma}{2d}\right] \leq \delta.$$

Moreover, if $\mathcal{M}$ is symmetric, then

$$\forall i \in [n] \quad \mathop{\mathbb{P}}_{p, x_{1\cdots n}, z, \mathcal{M}}\left[\langle x_i, q\rangle - \langle z, q\rangle < \frac{\gamma d}{2n}\right] \leq \delta.$$

Proof of Lemma 5.2.10. Let $a = \vec{1} \in \mathbb{R}^n$, $X_{i,j} = x_i^j - z^j$, and $Y = q = \mathcal{M}(x)$. Now we have

$$\sum_{i\in[n]}\left(\langle x_i, q\rangle - \langle z, q\rangle\right) = Z = a^T X Y.$$

Lemma 5.4.2 implies $\mathbb{E}\left[e^{tX_{i,j}}\right] \leq e^{t^2/2}$ for all i, j, and t. Let $Z_j = a^T X^j Y^j = \sum_{i\in[n]}(x_i^j - z^j)q^j$. Proposition 5.2.9 shows that

$$\mathbb{E}\left[Z_j \mid Z_{j+1} = z_{j+1}, \cdots, Z_d = z_d\right] = \mathop{\mathbb{E}}_{p^j, x_1^j, \cdots x_n^j, z^j}\left[\sum_{i\in[n]}\left(\langle x_i^j, q^j\rangle - \langle z^j, q^j\rangle\right)\right] \geq \gamma$$

for all $j \in [d]$ and $z \in \mathbb{R}^d$. Thus Theorem 5.4.1 shows that

$$\mathop{\mathbb{P}}_{Z}\left[Z < \gamma d - \lambda\right] \leq \exp\left(\frac{-\lambda^2}{16cd\left\|a\right\|_1^2}\right)$$

for all $\lambda > 0$, where $c = 1/2$. In particular, setting $\lambda = \gamma d/2$ gives

$$\mathop{\mathbb{P}}_{Z}\left[Z < \gamma d/2\right] \leq \exp\left(\frac{-(\gamma d/2)^2}{8dn^2}\right) = \exp\left(\frac{-\gamma^2 d}{32n^2}\right) \leq \delta.$$

To prove the second part of the lemma, we set $a = \vec{e}_i$ instead. □

Proposition 5.4.13 (Restating Proposition 5.3.4)**.** *Suppose the distribution $\mathcal{D}$ is a product*

distribution in which each marginal ρ is (ξ, n)-strong and satisfies $\mathop{\mathbb{E}}_{p\sim\rho}\left[1-p^2\right] \geq \gamma + \alpha\xi$. Suppose $\mathcal{M} : \{\pm 1\}^{n\times d} \to [-1,1]^d$ is α-accurate. Let $d > O(\alpha^2 n^2 \log(1/\delta)/\gamma^2)$ and $m \geq 2\log(24n/\gamma)/\alpha^2$. Let $x_1, \ldots, x_n, z_0, z_1, \ldots, z_m \sim \mathcal{P}_p$. Let $q \sim \mathcal{M}(x_1, \ldots, x_n)$ Then

$$\mathbb{P}\left[\sum_{i\in[n]}\left(\langle x_i - z, \lfloor q - w\rceil_\eta\rangle\right) < \frac{\gamma d}{2}\right] \leq \delta$$

(recall $z = z_0$, $w = (1/m)\sum_{i=1}^m z_i$, and $\eta = 2\alpha$).

Proof of Proposition 5.3.4. Let $a = \vec{1}$, $X_{i,j} = (x_i^j - z^j)$, $Y_j = \lfloor q - w\rceil_\eta^j$, and $Z_j = (a^T X)_j Y_j = \sum_{i\in[n]}(x_i^j - z^j)\lfloor q - w\rceil_\eta^j$. By Lemma 5.3.3, the hypotheses of Theorem 5.4.1 are satisfied. Thus we have

$$\mathbb{P}\left[a^T XY < d\left(\gamma - 4ne^{-\alpha^2 m/2}\right) - \lambda\right] \leq e^{\frac{-\lambda^2}{8d\eta^2 n^2}}$$

for all $\lambda > 0$. Set $\lambda = \sqrt{8d\eta^2 n^2 \log(1/\delta)} \leq \gamma d/6$. Now $4ne^{-\alpha^2 m/2} \leq \gamma/6$, so the result follows. □

Chapter 6

Lower Bounds for Adaptive Data Analysis

6.1 Introduction

Empirical research commonly involves asking multiple "queries" on a finite sample drawn from some population (e.g., summary statistics, hypothesis tests, or learning algorithms). The outcome of a query is deemed significant if it is unlikely to have occurred by chance alone, and a "false discovery" occurs if the analyst incorrectly declares an observation significant. For decades statisticians have been devising methods for preventing false discovery, such as the "Bonferroni correction" [Bon36, Dun61] and the widely used and highly influential method of Benjamini and Hochberg [BH95] for controlling the "false discovery rate."

Nevertheless, false discovery persists across all empirical sciences, and both popular and scientific articles report on an increasing number of invalid research findings. Typically false discovery is attributed to misuse of statistics. However, another possible explanation is that methods for preventing false discovery do

not address the fact that data analysis is inherently *adaptive*—the choice of queries depends on previous interactions with the data.

As in Chapter 3 and [DFH+15c, HU14], we formalise the problem of adaptive data analysis in the statistical query model: there is an algorithm called the *mechanism* that is given n samples from an unknown distribution $\mathcal{P}$ over some finite universe $\mathcal{X} = \{0,1\}^d$, where the parameter d is the dimensionality of the distribution. The mechanism must answer *statistical queries* about $\mathcal{P}$. A statistical query q is specified by a predicate $p\colon \mathcal{X} \to \{0,1\}$ and the answer to a statistical query is

$$q(\mathcal{P}) = \mathop{\mathbb{E}}_{x\sim\mathcal{P}}[p(x)]\,.$$

The mechanism's answer a to a query q is *accurate* if $|a - q(\mathcal{P})| \leq \alpha$ with high probability (for suitably small α). Importantly, the goal of the mechanism is to provide answers that "generalise" to the underlying distribution, rather than answers that are specific to the sample. The latter is easy to achieve by outputting the empirical average of the query predicate on the sample.

The analyst makes a sequence of queries $q^1, q^2, \ldots, q^k$ to the mechanism, which responds with answers $a^1, a^2, \ldots, a^k$. In the adaptive setting, the query q^i may depend on the previous queries and answers $q^1, a^1, \ldots, q^{i-1}, a^{i-1}$ arbitrarily. We say the mechanism is *accurate* given n samples for k adaptively chosen queries if, when given n samples from an arbitrary distribution $\mathcal{P}$, the mechanism accurately responds to any adaptive analyst that makes at most k queries with high probability. A computationally efficient mechanism answers each query in time polynomial in n and d.[1]

When the queries are specified *non adaptively* (i.e. independent of previous

[1]We assume that the analyst only asks queries that can be evaluated on the sample in polynomial time.

answers), then the empirical average of each query on the sample is accurate with high probability as long as $k \leq 2^{o(n)}$. However, the situation turns out to be very different when the queries are asked adaptively. Using a connection to differential privacy, Chapter 3 (following Dwork et al. [DFH+15c]) shows that there is a computationally efficient mechanism that accurately answers $\tilde{\Omega}(n^2)$ many adaptively chosen queries. Furthermore, there is an exponential-time mechanism that can answer exponentially in n many queries. Unfortunately, [HU14], building on hardness results for differential privacy [Ull13, BUV14] showed that, assuming the existence of one-way functions, there is no computationally efficient algorithm that answers $\tilde{O}(n^3)$ queries. In this chapter we improve this result to $O(n^2)$:

Theorem 6.1.1 (Informal)**.** *Assuming the existence of one-way functions, there is no computationally efficient mechanism that, given n samples, is accurate on more than $O(n^2)$ adaptively chosen queries.*

Conceptually, our result further demonstrates that there is an inherent computational barrier to preventing false discovery in interactive data analysis. It also shows that in the worst case, an efficient mechanism for answering adaptively chosen statistical queries may as well be differentially private. That is, the mechanisms in Chapter 3 that answer $\tilde{\Omega}(n^2)$ queries are differentially private, and no efficient mechanism regardless of privacy can answer significantly more arbitrary, adaptively chosen queries.

As in [HU14], our hardness result applies whenever the dimensionality d of the data grows with the sample size such that 2^d is not polynomial in n.[2] This requirement is both mild and necessary. If $n \gg 2^d$ then the empirical distribution of the n samples will be close to the underlying distribution in statistical distance,

[2]This is under the stronger, but still standard, assumption that exponentially-hard one-way functions exist.

so every statistical query can be answered accurately given the sample. Thus, the dimensionality of the data has a major effect on the hardness of the problem. In fact, we can prove a nearly optimal *information theoretic* lower bound when the dimensionality of the data is much larger than n.

Theorem 6.1.2 (Informal)**.** *There is no mechanism (even a computationally unbounded one) that given n samples in dimension $d = O(n^2)$ is accurate on more than $O(n^2)$ adaptively chosen queries.*

Our result builds on the techniques of [HU14], who use *fingerprinting code* [BS98, Tar08] to prove their hardness result. In this work, we identify a variant called an *interactive fingerprinting code* [FT01], which abstracts the technique in [HU14] and gives a more direct way of proving hardness results for adaptive data analysis. A slightly weaker version of our results can be obtained using the nice recent construction of interactive fingerprinting codes due to Laarhoven et al. [LDR+13] as a black box.

Thus, we can summarize the contributions of this work as follows.

1. We identify *interactive fingerprinting codes* as the key combinatorial object underlying the hardness of preventing false discovery in adaptive environments, analogous to the way in which (non interactive) fingerprinting codes are the key combinatorial object underlying the hardness of differential privacy.

2. We use this connection to prove nearly optimal hardness results for preventing false discovery in interactive data analysis.

3. Our construction of interactive fingerprinting codes are *optimally robust*. In this context, optimal robustness means that all of our hardness results apply even when the mechanism answers only a $1/2 + \Omega(1)$ fraction of the queries

accurately. Robustness was identified by Bun et al. [BUV14] as an important property of fingerprinting codes which is necessary to prove their lower bounds. Bun et al. also constructed robust fingerprinting codes, but they were only able to tolerate corruption of a $1/75$ fraction of answers.

6.1.1 Techniques

The structure of our proof is rather simple, and closely follows the framework in [HU14]. We will design a challenge distribution $\mathcal{P}$ and a computationally efficient adaptive analyst $\mathcal{A}$ who knows $\mathcal{P}$. If any computationally efficient mechanism $\mathcal{M}$ is given n samples $S = \{x_1, \ldots, x_n\}$ drawn from $\mathcal{P}$, then our analyst $\mathcal{A}$ can use the answers of $\mathcal{M}$ to reconstruct the set S. Using this information, the adversary can construct a query on which S is not representative of $\mathcal{P}$.

Our adversary $\mathcal{A}$ and the distribution $\mathcal{P}$, like that of [HU14], is built from a combinatorial object with a computational "wrapper." The computational wrapper uses queries that cryptographically "hide" information from the mechanism $\mathcal{M}$. In our work he combinatorial object will be an *interactive fingerprinting code (IFPC)*. An IFPC is a generalisation of a *(standard) fingerprinting code*, which was originally introduced by Boneh and Shaw [BS98] as a way to watermark digital content.

An interactive fingerprinting code $\mathcal{F}$ is an efficient interactive algorithm that defeats any adversary $\mathcal{P}$ in the following game (with high probability). The adversary $\mathcal{P}$ picks $S \subset [N]$ unknown to $\mathcal{F}$. The goal of $\mathcal{F}$ is to identify S by making ℓ interactive queries to $\mathcal{P}$. $\mathcal{F}$ specifies each query by a vector $c \in \{\pm 1\}^N$. In response, the adversary $\mathcal{P}$ must simply output $a \in \{\pm 1\}$ such that $a = c_i$ for some $i \in [N]$. However, the adversary $\mathcal{P}$ is restricted to only see c_i for $i \in S$. At any time, $\mathcal{F}$ may *accuse* some $i \in [N]$. If $i \in S$ is accused, then i is removed from S (i.e. $S \leftarrow S \backslash \{i\}$), thereby further restricting $\mathcal{P}$. If $i \notin S$ is accused, then this is referred to

as a *false accusation*. To win, the interactive fingerprinting code $\mathcal{F}$ must accuse all of S, without making "too many" false accusations.

In contrast [HU14] use only standard fingerprinting codes. The difference between interactive and non interactive fingerprinting codes is that a non interactive fingerprinting code must give all ℓ queries to $\mathcal{P}$ at once, but is (necessarily) only required to identify one $i \in S$. The suboptimal parameters achieved by [HU14], as well as some of the additional technical work, are there result of having to boost non interactive fingerprinting codes to recover all of S. Using this new perspective of interactive fingerprinting codes, the technique of [HU14] can be seen as a construction of an interactive fingerprinting code with length $\ell = \tilde{O}(N^3)$ by concatenating N independent copies of Tardos' [Tar08] non interactive fingerprinting code of length $\ell = \tilde{O}(N^2)$.

However, one can construct more clever and shorter interactive fingerprinting codes. Specifically, Laarhoven et al. [LDR+13] (building on Tardos [Tar08]) give a construction that would be suitable for our application with $\ell = \tilde{O}(N^2)$. Extending their results, we give a new analysis of their interactive fingerprinting code as well as Tardos' non interactive fingerprinting code that allows us to achieve length $\ell = O(N^2)$ while still being sufficiently secure for our application.

Theorem 6.1.3 (Informal)**.** *For every N, there exists an interactive fingerprinting code with $\ell = O(N^2)$ that, except with negligible probability, makes at most $N/1000$ false accusations.*

This result suffices for the informal statements made above, but our construction is somewhat more general and has additional parameters and security properties, which we detail in Section 6.2.

6.1.2 Additional Related Work

Our work and [HU14] is part of a line of work connecting technology for secure watermarking to lower bounds for private and interactive data analysis tasks. This connection first appeared in the work of Dwork, Naor, Reingold, Rothblum, and Vadhan [DNR+09], who showed that the existence of *traitor-tracing schemes* implies hardness of differential privacy. Traitor-tracing schemes were introduced by Chor, Fiat, and Naor [CFN94], also for the problem of watermarking digital content. The connection between traitor-tracing and differential privacy was strengthened in [Ull13], which introduced the use of fingerprinting codes in the context of differential privacy, and used them to show optimal hardness results for certain settings. [BUV14] showed that fingerprinting codes can be used to prove nearly-optimal information-theoretic lower bounds for differential privacy, which established fingerprinting codes as the key information-theoretic object underlying lower bounds in differential privacy. Chapter 4 proves lower bounds for differential privacy using the ideas of fingerprinting, although fingerprinting codes are not explicitly used there.

Since there introduction by Boneh and Shaw [BS98] there has been extensive work on fingerprinting codes, most of which is beyond the scope of this discussion. For the standard, non-interactive definition of fingerprinting codes, [Tar08] gave an essentially optimal construction, which has been very influential in most of the subsequent work on the topic. The interactive model of fingerprinting codes was first studied by [FT01] under the name "dynamic traitor-tracing schemes." Formally their results are in a significantly different model and cannot be used to prove hardness of preventing false discovery. [Tas05] gave the first construction in the model we use, but achieved suboptimal code length. Recently Laarhoven, Doumen, Roelse, Škorić, and de Wegner [LDR+13], gave a construction with nearly optimal

length by generalising Tardos' code to the interactive setting. Their construction is quite similar to ours, but our analysis is substantively different and leads to sharper and more general guarantees (and we feel is more intuitive).

6.1.3 Organisation

In Section 6.2 we define and construct interactive fingerprinting codes, the main technical ingredient we use to establish our results. In Section 6.3 we show how interactive fingerprinting codes can be used to obtain hardness results for preventing false discovery. The definition of interactive fingerprinting codes is contained in Section 6.2.1 and is necessary for Section 6.3, but the remainder of Section 6.2 and Section 6.3 can be read in either order.

6.2 Interactive Fingerprinting Codes

In order to motivate the definition of interactive fingerprinting codes, it will be helpful to review the motivation for standard, non interactive fingerprinting codes.

Fingerprinting codes were introduced by Boneh and Shaw [BS98] for the problem of watermarking digital content (such as a movie or a piece of software). Consider a company that distributes some content to N users. Some of the users may illegally distribute copies of the content. To combat this, the company gives each user a unique version of the content by adding distinctive "watermarks" to it. Thus, if the company finds an illegal copy, it can be traced back to the user who originally purchased it. Unfortunately, users may be able to remove the watermarks. In particular, a coalition of users may combine their copies in a way that mixes or obfuscates the watermarks. A fingerprinting code ensures that, even if up to n users collude to combine their codewords, an illegal copy can be still be traced to at least

one of the users.

Formally, every user $i \in [N]$ is given a codeword $(c_i^1, c_i^2, \ldots, c_i^\ell) \in \{\pm 1\}^\ell$ by the fingerprinting code, which represents the combination of watermarks in that user's copy. A subset $S \subset [N]$ of at most n users can arbitrarily combine their codewords to create a "pirate codeword" $a = (a^1, a^2, \ldots, a^\ell) \in \{\pm 1\}^\ell$. The only constraint is so-called *consistency*—for every $j \in [\ell]$, if, for every colluding user $i \in S$, we have $c_i^j = b$, then $a^j = b$. That is to say, if each of the colluding users receives the same watermark, then their combined codeword must also have that watermark. Given a, the fingerprinting code must be able to trace at least one user $i \in S$. Tardos [Tar08] constructed optimal fingerprinting codes with $\ell = O(n^2 \log N)$.

A key drawback of fingerprinting codes is that we can only guarantee that a single user $i \in S$ is traced. This is inherent, as setting the pirate codeword a to be the codeword of a single user prevents any other user from being identified. We will see that this can be circumvented by moving to an interactive setting.

Suppose the company is instead distributing a *stream* of content (such as a TV series) to N users—that is, the content is not distributed all at once and the illegal copies are obtained whilst the content is being distributed (e.g. the episodes of the TV series appear on the internet before the next episode is shown). Again, the content is watermarked so that each user receives a unique stream and a subset $S \subset [N]$ of at most n users combine their streams and distribute an illegal stream. The company obtains the illegal stream and uses this to trace the colluding users S. As soon as the company can identify a colluding user $i \in S$, that user's stream is terminated (e.g. their subscription is cancelled). This process continues until every $i \in S$ has been traced and the distribution of illegal copies ceases.

Another twist on fingerprinting codes is robustness [BUV14]. Suppose that the consistency constraint only holds for $(1 - \beta)\ell$ choices of $j \in [\ell]$. That is to say, the

colluding users can somehow remove a β fraction of the watermarks. [BUV14] showed how to modify the Tardos fingerprinting code to be robust to a small constant fraction of inconsistencies. In this work, we show that robustness to any $\beta < 1/2$ fraction of inconsistencies can be achieved.

6.2.1 Definition and Existence

We are now ready to formally define interactive fingerprinting codes. To do so we make use of the following game between an adversary $\mathcal{P}$ and the fingerprinting code $\mathcal{F}$. Both $\mathcal{P}$ and $\mathcal{F}$ may be stateful. For a given execution of $\mathcal{F}$, we let

$\mathcal{P}$ selects a subset of users $S^1 \subseteq [N]$ of size n, unknown to $\mathcal{F}$.
For $j = 1, \ldots, \ell$:
  $\mathcal{F}$ outputs a column vector $c^j \in \{\pm 1\}^N$.
  Let $c^j_{S^j} \in \{\pm 1\}^{|S^j|}$ be the restriction of c^j to coordinates in S^j, which is given to $\mathcal{P}$.
  $\mathcal{P}$ outputs $a^j \in \{\pm 1\}$, which is given to $\mathcal{F}$.
  $\mathcal{F}$ accuses a (possibly empty) set of users $I^j \subseteq [N]$. Let $S^{j+1} = S^j \setminus I^j$.

Figure 6.1: $\mathsf{IFPC}_{N,n,\ell}[\mathcal{P}, \mathcal{F}]$

$C \in \{\pm 1\}^{N \times \ell}$ be the matrix with columns $c^1, \ldots, c^\ell$ and let $a \in \{\pm 1\}^\ell$ be the vector with entries $a^1, \ldots, a^\ell$. We want to construct the fingerprinting code so that, if a is consistent, then the tracer succeeds in recovering every user in S. For convenience, we will define the notation θ^j to be the number of rounds $1, \ldots, j$ in which a^j is not consistent with c^j. Formally, for a given execution of $\mathcal{F}$,

$$\theta^j = \left| \left\{ 1 \le k \le j \;\middle|\; \nexists\, i \in [N],\ a^k = c^k_i \right\} \right|.$$

Using this notation, a is β-consistent if $\theta^\ell \le \beta\ell$. We also define the notation ψ^j to be the number of users in $I^1, \ldots, I^j$ who are falsely accused (i.e. not in the coalition S^1).

Formally,

$$\psi^j = \left| \left(\bigcup_{1 \le k \le j} I^k \right) \setminus S^1 \right|.$$

Using this notation, we require $\psi^\ell \le \delta(N - |S^1|)$ - that is, the tracing algorithm does not make too many false accusations. "Too many" is defined as more than a δ-fraction of innocent users.

Definition 6.2.1 (Interactive Fingerprinting Codes)**.** *We say that an algorithm $\mathcal{F}$ is an n-collusion-resilient interactive fingerprinting code of length ℓ for N users robust to a β fraction of errors with failure probability ε and false accusation probability δ if for every adversary $\mathcal{P}$, it holds that*

$$\mathop{\mathbb{P}}_{\mathsf{IFPC}_{N,n,\ell}[\mathcal{P},\mathcal{F}]} \left[\left(\theta^\ell \le \beta\ell \right) \vee \left(\psi^\ell > \delta(N-n) \right) \right] \le \varepsilon$$

The length ℓ may depend on $N, n, \beta, \varepsilon, \delta$.

The constraint $\psi^\ell \le \delta N$ is called *soundness*—the interactive fingerprinting code should not make (many) false accusations. The constraint $\theta^\ell > \beta\ell$ is called *completeness*—the interactive fingerprinting code should force the adversary $\mathcal{P}$ to be inconsistent. Although it may seem strange that we make no reference to recovering the coalition S^1, notice that if $S^j \neq \emptyset$, then $\mathcal{P}$ can easily be consistent. Therefore, if the pirate cannot be consistent, it must be the case that $S^j = \emptyset$ for some j, meaning all of S^1 has been accused.

In the remainder of this section, we give a construction of interactive fingerprinting codes, and establish the following theorem.

Theorem 6.2.2 (Existence of Interactive Fingerprinting Codes)**.** *For every $1 \le n \le N$, $0 \le \beta < 1/2$, and $0 < \delta \le 1$, there is a n-collusion-resilient interactive fingerprinting code*

of length ℓ for N users robust to a β fraction of errors with failure probability

$$\varepsilon \leq \min\{\delta(N-n), 2^{-\Omega(\delta(N-n))}\} + \delta^{\Omega\left(\left(\frac{1}{2}-\beta\right)n\right)}$$

and false accusation probability δ for

$$\ell = O\left(\frac{n^2 \log(1/\delta)}{\left(\frac{1}{2}-\beta\right)^4}\right).$$

We remark on the parameters of our construction and how they relate to the literature.

Remark 6.2.3.

- *The expression for the failure probability ε is a bit mysterious. To interpret it, we fix $\beta = 1/2 - \Omega(1)$ and consider two parameter regimes: $\delta(N-n) \ll 1$ and $\delta(N-n) \gg 1$.*

  In the traditional parameter regime for fingerprinting codes $\delta(N-n) = \varepsilon' \ll 1$, and so no users are falsely accused. Then our fingerprinting code has length $O(n^2 \log((N-n)/\varepsilon'))$ and a failure probability of ε'. This matches the result of [LDR+13].

  However, if we are willing to tolerate falsely accusing a small constant fraction of users, then we can set, for example, $\delta(N-n) = .01N$, and our fingerprinting code will have length $O(n^2)$ and failure probability $2^{-\Omega(n)}$. To our knowledge, such large values of δ have not been considered before. It saves a logarithmic factor in our final result.

- *Our construction works for any robustness parameter $\beta < 1/2$. Previously [BUV14] gave a construction for $\beta = 1/75$ in the non-interactive setting. Previous constructions in the interactive setting do not achieve* any *robustness $\beta > 0$, even for the*

weaker model of robustness to erasures [BN08].

- *Our completeness condition differs subtly from previous work. We require that, with high probability,*

$$\theta^\ell = \left|\left\{1 \le k \le \ell \;\middle|\; \nexists\, i \in [N],\ a^k = c_i^k\right\}\right| > \beta\ell,$$

rather than the weaker condition

$$\left|\left\{1 \le k \le \ell \;\middle|\; \nexists\, i \in S^1,\ a^k = c_i^k\right\}\right| > \beta\ell.$$

While our version is less natural in the watermarking setting, it is important to our application to false discovery. Our interactive fingerprinting code ensures that the adversary cannot be consistent with respect to the population, rather than that it cannot be consistent with respect to the sample.

6.2.2 The Construction

Our construction and analysis is based on the optimal (non interactive) fingerprinting codes of Tardos [Tar08], and the robust variant by Bun et al. [BUV14]. The code is essentially the same, but columns are generated and shown to the adversary one at a time, and tracing is modified to identify users interactively.

We begin with some definitions and notation. For $0 \le a < b \le 1$, let $D_{a,b}$ be the distribution with support (a,b) and probability density function $\mu(p) = C_{a,b}/\sqrt{p(1-p)}$, where $C_{a,b}$ is a normalising constant.[3] For $\alpha, \zeta \in (0, 1/2)$, let $\overline{D}_{\alpha,\zeta}$ be the distribution on $[0,1]$ that returns a sample from $D_{\alpha,1-\alpha}$ with probability $1-2\zeta$ and 0 or 1 each with probability ζ.

[3]To sample from $D_{a,b}$, first sample $\varphi \in (\sin^{-1}(\sqrt{a}), \sin^{-1}(\sqrt{b}))$ uniformly, then output $\sin^2(\varphi)$ as the sample.

For $p \in [0,1]$, let $c \sim p$ denote that $c \in \{\pm 1\}$ is drawn from the distribution with $\mathbb{P}[c = 1] = p$ and $\mathbb{P}[c = -1] = 1 - p$. Let $c_{1\cdots n} \sim p$ denote that $c \in \{\pm 1\}^n$ is drawn from a product distribution in which $c_i \sim p$ independently for all $i \in [n]$.

Define $\phi^p : \{\pm 1\} \to \mathbb{R}$ by $\phi^0(c) = \phi^1(c) = 0$ and, for $p \in (0,1)$, $\phi^p(1) = \sqrt{(1-p)/p}$ and $\phi^p(-1) = -\sqrt{p/(1-p)}$. The function ϕ^p is chosen so that $\phi^p(c)$ has mean 0 and variance 1 when $c \sim p$.

Given parameters $1 \le n \le N$ and $0 < \delta, \beta < 1/2$
Set parameters:

$$\begin{aligned}
\alpha &= \frac{\left(\frac{1}{2} - \beta\right)}{4n} && \geq \Omega\left(\frac{\left(\frac{1}{2} - \beta\right)}{n}\right) \\
\zeta &= \frac{3}{8} + \frac{\beta}{4} && = \frac{1}{2} - \frac{1}{4}\left(\frac{1}{2} - \beta\right) \\
\sigma &= 64 \cdot \left\lceil \frac{6\pi n}{\left(\frac{1}{2} - \beta\right)^2} \right\rceil \cdot \left\lceil \log_e\left(\frac{32}{\delta}\right) \right\rceil && \leq O\left(\frac{n}{\left(\frac{1}{2} - \beta\right)^2} \log\left(\frac{1}{\delta}\right)\right) \\
\ell &= \left\lceil \frac{6\pi n}{\left(\frac{1}{2} - \beta\right)^2} \right\rceil \cdot \sigma && \leq O\left(\frac{n^2}{\left(\frac{1}{2} - \beta\right)^4} \log\left(\frac{1}{\delta}\right)\right)
\end{aligned}$$

Let $s_i^0 = 0$ for every $i \in [N]$.
For $j = 1, \ldots, \ell$:
  Draw $p^j \sim \overline{D_{\alpha,\zeta}}$ and $c_{1\cdots N}^j \sim p^j$.
  Issue $c^j \in \{\pm 1\}^N$ as a challenge and receive $a^j \in \{\pm 1\}$ as the response.
  For $i \in [N]$, let $s_i^j = s_i^{j-1} + a^j \cdot \phi^{p^j}(c_i^j)$.
  Accuse $I^j = \left\{ i \in [N] \mid s_i^j > \sigma \right\}$.

Figure 6.2: *The interactive fingerprinting code $\mathcal{F} = \mathcal{F}_{n,N,\delta,\beta}$*

The fingerprinting code $\mathcal{F}$ is defined in Figure 6.2. In addition to the precise setting of parameters, we have given asymptotic bounds to help follow the analysis. We now analyse $\mathcal{F}$ and establish Theorem 6.2.2. The proof of Theorem 6.2.2 is split

into Theorems 6.2.8 and 6.2.19. For convenience, define $I = \bigcup_{j \in [\ell]} I^j$.

6.2.3 Analysis Overview and Comparison

Comparison to Chapters 4 and 5

The key difference between the analysis we present here and that which appears in Chapters 4 and 5 is that here we focus on obtaining the strongest possible fingerprinting code. Namely, rather than assuming that all answers provided by $\mathcal{P}$ are α-accurate as before, we merely assume that all but a β fraction are consistent. This setting is the "orthodox" setting for fingerprinting codes, whereas the simplified analysis for α-accurate answers is novel to this thesis. This setting requires a substantially different analysis.

Scores

Intuitively, the quantity s_i^j, which we call the *score* of user i, measures the "correlation" between the answers $(a^1, \cdots, a^j)$ of $\mathcal{P}$ and the i-th codeword $(c_i^1, \cdots, c_i^j)$, using a particular measure of correlation that takes into account the choices $p^1, \ldots, p^j$. If s_i^j ever exceeds the threshold σ, meaning that the answers are significantly correlated with the i-th codeword, then we accuse user i. Thus, our goal is to show two things: *Soundness*, that the score of an *innocent* user (i.e. $i \notin S^1$) never exceeds the threshold, as the answers cannot be correlated with the unknown i-th codeword. And *completeness*, that the score of every *guilty* user (i.e. $i \in S^1$) will at some point exceed the threshold, meaning that the answers must correlate with the i-th codeword for every $i \in S^1$.

Soundness

The proof of soundness closely mirrors Tardos' analysis [Tar08] of the non-interactive case. If i is innocent, then, since $\mathcal{P}$ doesn't see the codeword $(c_i^1, \cdots, c_i^j)$ of the i^{th} user, there cannot be too much correlation. In this case, one can show that s_i^j is the sum of j independent random variables, each with mean 0 and variance 1, where we take the answers $a^1, \ldots, a^j$ as fixed and the randomness is over the choice of the unknown codeword. By analogy to Gaussian random variables, one would expect that $s_i^j \leq \sigma = \Theta(\sqrt{\ell \log(1/\delta)})$ with probability at least $1 - \delta$. Formally, the fact that the score in each round is not bounded prevents the use of a Chernoff bound. But nonetheless, in Section 6.2.4, we prove soundness using a Chernoff-like tail bound for s_i^j.

Completeness

To prove completeness, we must show that, for guilty users $i \in S^1$, we have $s_i^j > \sigma$ for some $j \in [\ell]$ with high probability. In Sections 6.2.5 and 6.2.5, we prove that if $\mathcal{P}$ gives consistent answers in a $1 - \beta$ fraction of rounds, then the sum of the scores for each of the guilty users is large. Specifically, in Theorem 6.2.17, we prove that with high probability

$$\sum_{i \in S^1} s_i^\ell \geq \Theta(\ell) \tag{6.1}$$

The constants hidden by the asymptotic notation are set to imply that, for at least one $i \in S^1$, the score s_i^ℓ is above the threshold $\sigma = \Theta(\ell/n)$. This step is not too different from the analysis of Tardos and Bun et al. [Tar08, BUV14] for the non-interactive case. To show that, for *every* $i \in S^1$, we will have $s_i^j > \sigma$ at some point, we must depart from the analysis of non-interactive fingerprinting codes. If $s_i^j > \sigma$, and user i is accused in round j, then the adversary will not see the suffix of

codeword $(c_i^{j+1}, \cdots, c_i^{\ell})$. By the same argument that was used to prove soundness, the answers will not be correlated with this suffix, so with high probability the score s_i^{ℓ} does not increase much beyond σ. Thus,

$$\sum_{i \in S^1} s_i^{\ell} \leq n \cdot O(\sigma) = \Theta(\ell). \tag{6.2}$$

The hidden constants are set to ensure that Equation (6.2) conflicts with Equation (6.1). Thus, we can conclude that $\mathcal{P}$ cannot give consistent answers for a $1 - \beta$ fraction of rounds. That is to say, $\mathcal{P}$ is forced to be inconsistent because all of S^1 is accused and eventually $\mathcal{P}$ sees none of the codewords and is reduced to guessing an answer a^j.

Establishing Correlation

Proving Equation (6.1) is key to the analysis. Our proof thereof combines and simplifies the analyses of [Tar08] and [BUV14]. For this high level overview, we ignore the issue of robustness and fix $\beta = 0$.

First we prove that the correlation bound holds in expectation and then we show that it holds with high probability using an Azuma-like concentration bound. (Again, as the random variables being summed are not bounded, we are forced to use a more tailored analysis to prove concentration.) We show that it holds in expectation for each round. In Proposition 6.2.14, we show that the concentration grows in expectation in each round. For every $j \in [\ell]$,

$$\mathbb{E}\left[\sum_{i \in S^1} s_i^j - s_i^{j-1}\right] = \mathbb{E}\left[\sum_{i \in S^1} a^j \cdot \phi^{p^j}(c_i^j)\right] \geq \Omega(1), \tag{6.3}$$

where the expectations are taken over the randomness of p^j, c^j, and a^j. Equation (6.3), combined with a concentration result, implies Equation (6.1).

The intuition behind Equation (6.3) and the choice of p^j is as follows. Consistency

guarantees that, if $c_i^j = b$ for all $i \in S^1$, then $a^j = b$. This is a weak correlation guarantee, but it suffices to ensure correlation between a^j and $\sum_{i \in S^1} c_i^j$. The affine scaling ϕ^{p^j} ensures that $\phi^{p^j}(c_i^j)$ has mean zero (i.e. is uncorrelated with a constant) and and unit variance (i.e. has unit correlation with itself). The expectation of $a^j \cdot \phi^{p^j}(c_i^j)$ can be interpreted as the i-th first-order Fourier coefficient of a^j as a function of c^j. To understand first-order Fourier coefficients, consider the "dictator" function: Suppose $a^j = c_{i^*}^j$ for some $i^* \in S^1$ - that is, $\mathcal{P}$ always outputs the i^*-th bit. Then

$$\mathop{\mathbb{E}}_{a^j, c^j, p^j} \left[a^j \sum_{i \in S^1} \phi^{p^j}(c_i^j) \right] = \mathop{\mathbb{E}}_{c^j, p^j} \left[c_{i^*}^j \cdot \phi^{p^j}(c_{i^*}^j) \right] = \mathop{\mathbb{E}}_{p^j} \left[2\sqrt{p^j(1-p^j)} \right] = \Theta(1).$$

This example can be generalised to a^j being an arbitrary function of $c_{S^1}^j$ using Fourier analysis. This calculation also indicates why we choose the probability density function of $p^j \sim D_{\alpha,1-\alpha}$ to be proportional to $1/\sqrt{p(1-p)}$.

To handle robustness ($\beta > 0$) we use the ideas of [BUV14]. With probability 2ζ each round is a "special" constant round—i.e. $c^j = (1)^N$ or $c^j = (-1)^N$. Otherwise it is a "normal" round where c^j is sampled as before. Intuitively, the adversary $\mathcal{P}$ cannot distinguish the special rounds from the normal rounds in which c happens to be constant. If the adversary gives inconsistent answers on normal rounds, then it must also give inconsistent answers on special rounds. Since there are many more special rounds than normal rounds, this means that a small number of inconsistencies in normal rounds implies a large number of inconsistencies on special rounds. Conversely, inconsistencies are absorbed by the special rounds, so we can assume there are very few inconsistencies in normal rounds. Thus $\mathcal{P}$ is forced to behave consistently on the normal rounds and the analysis on these rounds proceeds as before.

6.2.4 Proof of Soundness

We first show that no user is falsely accused except with probability $\delta/2$. This boils down to proving a concentration bound. Then another concentration bound shows that with high probability at most a δ fraction of users are falsely accused.

These concentrations bounds are essentially standard. However, we are showing concentration of sums of variables of the form $\phi^p(c)$, which may be quite large if $p \approx 0$ or $p \approx 1$. This technical problem prevents us from directly applying standard concentration bounds. Instead we open up the standard proofs and verify the desired concentration. We take the usual approach of bounding the moment generating function and using that to give a tail bound.

Lemma 6.2.4. *For $p \in [\alpha, 1-\alpha] \cup \{0,1\}$ and $t \in [-\sqrt{\alpha}/2, \sqrt{\alpha}/2]$, we have*

$$\mathop{\mathbb{E}}_{c\sim p}\left[e^{t\phi^p(c)}\right] \leq e^{t^2}.$$

Proof. If $p \in \{0,1\}$, $\phi^p = 0$ and the result is trivial. We have $\mathop{\mathbb{E}}_{c\sim p}[\phi^p(c)] = 0$, $\mathop{\mathbb{E}}_{c\sim p}\left[\phi^p(c)^2\right] = 1$, and, for $c \in \{\pm 1\}$, $|\phi^p(c)| \leq 1/\sqrt{\alpha}$. In particular, $|\phi^p(c)\cdot t| \leq 1/2$. For $u \in [-1/2, 1/2]$, we have $e^u \leq 1 + u + u^2$. Thus

$$\mathop{\mathbb{E}}_{c\sim p}\left[e^{t\phi^p(c)}\right] \leq 1 + t\mathop{\mathbb{E}}_{c\sim p}[\phi^p(c)] + t^2\mathop{\mathbb{E}}_{c\sim p}\left[\phi^p(c)^2\right] = 1 + t^2 \leq e^{t^2}.$$

□

Lemma 6.2.5. *Let $p_1 \cdots p_m \in [\alpha, 1-\alpha] \cup \{0,1\}$ and $c_1 \cdots c_m$ drawn independently with $c_i \sim p_i$. Let $a_1 \cdots a_m \in [-1,1]$ be fixed. For all $\lambda \geq 0$, we have*

$$\mathbb{P}\left[\sum_{i\in[m]} a_i\phi^{p_i}(c_i) \geq \lambda\right] \leq e^{-\lambda^2/4m} + e^{-\sqrt{\alpha}\lambda/4}.$$

Proof. By Lemma 6.2.4, for all $t \in [-\sqrt{\alpha}/2, \sqrt{\alpha}/2]$,

$$\mathop{\mathbb{E}}_{c}\left[e^{t\sum_{i\in[m]} a_i\phi^{p_i}(c_i)}\right] \leq \prod_{i\in[m]} \mathop{\mathbb{E}}_{c_i}\left[e^{ta_i\phi^{p_i}(c_i)}\right] \leq e^{t^2 m}.$$

By Markov's inequality,

$$\mathbb{P}\left[\sum_{i\in[m]} a_i\phi^{p_i}(c_i) \geq \lambda\right] \leq \frac{\mathbb{E}\left[e^{t\sum_{i\in[m]} a_i\phi^{p_i}(c_i)}\right]}{e^{t\lambda}} \leq e^{t^2 m - t\lambda}.$$

Set $t = \min\{\sqrt{\alpha}/2, \lambda/2m\}$. If $\lambda \in [0, m\sqrt{\alpha}]$, then

$$\mathbb{P}\left[\sum_{i\in[m]} a_i\phi^{p_i}(c_i) \geq \lambda\right] \leq e^{-\lambda^2/4m}.$$

On the other hand, if $\lambda \geq m\sqrt{\alpha}$, then

$$\mathbb{P}\left[\sum_{i\in[m]} a_i\phi^{p_i}(c_i) \geq \lambda\right] \leq e^{\alpha m/4 - \sqrt{\alpha}\lambda/2} \leq e^{-\sqrt{\alpha}\lambda/4}.$$

The result is obtained by adding these expressions. □

The following theorem shows how we can beat the union bound for tail bounds on partial sums.

Theorem 6.2.6 (Etemadi's Inequality [Ete85])**.** *Let $X_1 \cdots X_n \in \mathbb{R}$ be independent random variables. For $k \in [n]$, define $S_k = \sum_{i\in[k]} X_i$ to be the k^{th} partial sum. Then, for all $\lambda > 0$,*

$$\mathbb{P}\left[\max_{k\in[n]} |S_k| > 4\lambda\right] \leq 4 \cdot \max_{k\in[n]} \mathbb{P}\left[|S_k| > \lambda\right].$$

Proposition 6.2.7 (Individual Soundness)**.** *For all $i \in [N]$, we have*

$$\mathbb{P}\left[i \in I \setminus S^1\right] \leq 8(e^{-\sigma^2/64\ell} + e^{-\sigma\sqrt{\alpha}/16}) \leq \delta/2,$$

where the probability is taken over $\mathsf{IFPC}_{N,\leq N,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$ for an arbitrary $\mathcal{P}$.

Here $\mathsf{IFPC}_{N,\le n,\ell}$ denotes $\mathsf{IFPC}_{N,n,\ell}$ with the constraint $|S^1| = n$ replaced by the constraint $|S^1| \le n$.

Proof. Let $i \in [N] \setminus S^1$. Since the adversary does not see c_i^j for any $j \in [\ell]$, we may treat the answers of the adversary as fixed and analyse s_i^j as if c_i^j was drawn after the actions of the adversary are fixed. Thus, by Lemma 6.2.5, for every $j \in [\ell]$,

$$\mathbb{P}\left[s_i^j > \frac{\sigma}{4}\right] = \mathbb{P}\left[\sum_{k\in[j]} a^k \phi^{p^k}(c_i^k) > \frac{\sigma}{4}\right] \le e^{-\sigma^2/64\ell} + e^{-\sigma\sqrt{\alpha}/16}.$$

Likewise $\mathbb{P}\left[s_i^j < -\frac{\sigma}{4}\right] \le e^{-\sigma^2/64\ell} + e^{-\sigma\sqrt{\alpha}/16}$. Thus, by Theorem 6.2.6,

$$\mathbb{P}\left[i \in I\right] \le \mathbb{P}\left[\max_{j\in[\ell]} |s_i^j| > \sigma\right] \le 4 \max_{j\in[\ell]} \mathbb{P}\left[|s_i^j| > \frac{\sigma}{4}\right] \le 8(e^{-\sigma^2/64\ell} + e^{-\sigma\sqrt{\alpha}/16}) \le \frac{\delta}{2}.$$

□

Theorem 6.2.8 (Soundness)**.** *We have*

$$\mathbb{P}\left[|I \setminus S^1| > \delta(N - |S^1|)\right] \le \min\left\{\delta(N - |S^1|), e^{-\delta(N-|S^1|)/8}\right\},$$

where the probability is taken over $\mathsf{IFPC}_{N,\le N,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$ *for an arbitrary* $\mathcal{P}$*.*

Remark 6.2.9. *Interestingly, Theorem 6.2.8 does not require* $|S^1| \le n$ *– that is, it holds with respect to* $\mathsf{IFPC}_{N,\le N,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$*, rather than* $\mathsf{IFPC}_{N,n,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$*. It only requires that* $\mathcal{F}$ *does not see the codewords of users not in* S^1*.*

This is a useful if we are in a setting where $|S^1|$ *is unknown: if* $|S^1| > n$*, then the interactive fingerprinting code will still not make too many false accusations, even if it fails to identify all of* S^1*.*

Proof. Let $E_i \in \{0, 1\}$ be the indicator of the event $i \in I \setminus S^1$. The E_is for $i \in [N]$ are independent (conditioned on the choice of S^1 and p^j for $j \in [\ell]$). Moreover, by

Proposition 6.2.7, $\mathbb{E}[E_i] \le \delta/2$ for all $i \in [N]$. Thus, by a Chernoff bound,

$$\mathbb{P}\left[|I \backslash S^1| > \delta(N - |S^1|)\right] = \mathbb{P}\left[\sum_{i \in [N]\backslash S^1} E_i > \delta(N - |S^1|)\right] \le e^{-\delta(N-|S^1|)/8}.$$

If $\delta < 1/(N - |S^1|)$, then this is a very poor bound. Instead we use the fact that the E_is are discrete and Markov's inequality, which amounts to a union bound. For $\delta(N - |S^1|) < 1$, we have

$$\mathbb{P}\left[|I \backslash S^1| > \delta(N - |S^1|)\right] = \mathbb{P}\left[|I \backslash S^1| \ge 1\right] \le \mathbb{E}\left[\sum_{i \in [N]\backslash S^1} E_i\right] \le \frac{\delta(N-|S^1|)}{2} \le \delta(N-|S^1|).$$

□

The following lemma will be useful later.

Lemma 6.2.10. *For $i \in [N]$, let $j_i \in [\ell+1]$ be the first j such that $i \notin S^j$, where we define $S^{\ell+1} = \emptyset$. For any $S \subset [N]$,*

$$\mathbb{P}\left[\sum_{i \in S} s_i^\ell - s_i^{j_i - 1} > \lambda\right] \le e^{-\lambda^2/4|S|\ell} + e^{-\sqrt{\alpha}\lambda/4},$$

where the probability is taken over $\mathsf{IFPC}_{N,\le N,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$ for an arbitrary $\mathcal{P}$.

Proof. We have

$$\sum_{i \in S} s_i^\ell - s_i^{j_i - 1} = \sum_{i \in S} \sum_{j \in [\ell]} \mathbb{I}(j \ge j_i) a^j \phi^{p^j}(c_i^j).$$

Again, since the adversary doesn't see c_i^j for $j \ge j_i$, the random variables $\mathbb{I}(j \ge j_i)a^j$ and $\phi^{p^j}(c_i^j)$ are independent, so we can view $\mathbb{I}(j \ge j_i)a^j \in [-1, 1]$ as fixed. Now the result follows from Lemma 6.2.5. □

6.2.5 Proof of Completeness

To show that the fingerprinting code identifies guilty users we must lower bound the scores $\sum_{i \in S^1} s_i^\ell$. First we bound their expectation and then their tails.

Biased Fourier Analysis

For this section, assume that the adversary $\mathcal{P}$ is always consistent - that is, we have no robustness and $\beta = 0$. Robustness will be added in Section 6.2.5. Here we establish that the scores have good expectation, namely

$$\mathbb{E}\left[\sum_{i \in S^1} s_i^j - s_i^{j-1}\right] \geq \Omega(1)$$

for all $j \in [\ell]$. The score s_i^ℓ computes the 'correlation' between the bits given to user i and the output of the adversary. We must show that that the adversary's consistency constraint implies that there exists some correlation on average.

In this section we deviate from the proof in [Tar08]. We use biased Fourier analysis to give a more intuitive proof of the correlation bound.

We have the following lemma and proposition, which relate the correlation $a^j \cdot \sum_{i \in S^1} \phi^{p^j}(c_i^j)$ to the properties of a^j as a function of p^j. To interpret these imagine that f represents the adversary $\mathcal{P}$ with one round viewed in isolation – the fingerprinting code gives the adversary c^j and the adversary responds with $f(c^j_{S^j})$.

Firstly, the following lemma gives an interpretation of the correlation value for a fixed p^j; it is a rescaling of Lemma 4.3.7 with an alternative proof.

Lemma 6.2.11. *Let $f : \{\pm 1\}^n \to \mathbb{R}$. Define $g : [0,1] \to \mathbb{R}$ by $g(p) = \mathop{\mathbb{E}}_{c_{1\cdots n} \sim p}[f(c)]$. For any $p \in (0,1)$,*

$$\mathop{\mathbb{E}}_{c_{1\cdots n} \sim p}\left[f(c) \cdot \sum_{i \in [n]} \phi^p(c_i)\right] = g'(p)\sqrt{p(1-p)}.$$

Proof. For $p \in (0,1)$ and $s \subset [n]$, define $\phi_s^p : \{\pm 1\}^n \to \mathbb{R}$ by $\phi_s^p(c) = \prod_{i \in s} \phi^p(c_i)$. The functions ϕ_s^p form an orthonormal basis with respect to the product distribution with bias p – that is,

$$\forall s, t \subset [n] \quad \mathop{\mathbb{E}}_{c_{1\cdots n} \sim p} \left[\phi_s^p(c) \cdot \phi_t^p(c) \right] = \left\{ \begin{array}{ll} 1 & s = t \\ 0 & s \neq t \end{array} \right\}.$$

Thus, for any $p \in (0,1)$, we can write f in terms of these basis functions:

$$\forall c \in \{\pm 1\}^n \quad f(c) = \sum_{s \subset [n]} \tilde{f}^p(s) \phi_s^p(c),$$

where

$$\forall s \subset [n] \quad \tilde{f}^p(s) = \mathop{\mathbb{E}}_{c_{1\cdots n} \sim p} \left[f(c) \phi_s^p(c) \right].$$

This decomposition is a generalisation of Fourier analysis to biased distributions [O'D14, §8.4]. For $p, q \in (0,1)$, the expansion of f gives the following expressions

for $g(q)$, $g'(q)$ and $g'(p)$.

$$\begin{aligned}
g(q) &= \mathop{\mathbb{E}}_{c_{1\cdots n}\sim q}[f(c)] \\
&= \sum_{s\subset[n]} \tilde{f}^p(s) \mathop{\mathbb{E}}_{c_{1\cdots n}\sim q}\left[\phi_s^p(c)\right] \\
&= \sum_{s\subset[n]} \tilde{f}^p(s) \prod_{i\in s} \mathop{\mathbb{E}}_{c\sim q}[\phi^p(c)] \\
&= \sum_{s\subset[n]} \tilde{f}^p(s) \left(q\sqrt{\frac{1-p}{p}} - (1-q)\sqrt{\frac{p}{1-p}}\right)^{|s|}. \\
g'(q) &= \sum_{s\subset[n]:s\neq\emptyset} \tilde{f}^p(s)\cdot|s|\cdot\left(q\sqrt{\frac{1-p}{p}} - (1-q)\sqrt{\frac{p}{1-p}}\right)^{|s|-1}\cdot\left(\sqrt{\frac{1-p}{p}} + \sqrt{\frac{p}{1-p}}\right). \\
g'(p) &= \sum_{s\subset[n]:s\neq\emptyset} \tilde{f}^p(s)\cdot|s|\cdot 0^{|s|-1}\cdot\left(\sqrt{\frac{1-p}{p}} + \sqrt{\frac{p}{1-p}}\right) \\
&= \sum_{i\in[n]} \tilde{f}^p(\{i\})\cdot\left(\sqrt{\frac{1-p}{p}} + \sqrt{\frac{p}{1-p}}\right).
\end{aligned}$$

Note that $\tilde{f}^p(\{i\}) = \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c)\phi^p(c_i)]$ and, hence,

$$\mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}\left[f(c)\cdot\sum_{i\in[n]}\phi^p(c_i)\right] = \sum_{i\in[n]}\tilde{f}^p(\{i\}) = \frac{g'(p)}{\sqrt{\frac{1-p}{p}}+\sqrt{\frac{p}{1-p}}} = g'(p)\sqrt{p(1-p)}.$$

□

Now we can interpret the correlation for a random $p^j \sim D_{a,b}$.

Proposition 6.2.12. *Let $f : \{\pm1\}^n \to \mathbb{R}$. Define $g : [0,1] \to \mathbb{R}$ by $g(p) = \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c)]$. For any $0 \le a < b \le 1$,*

$$\mathop{\mathbb{E}}_{p\sim D_{a,b}}\left[\mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}\left[f(c)\cdot\sum_{i\in[n]}\phi^p(c_i)\right]\right] = \frac{g(b)-g(a)}{2\sin^{-1}(\sqrt{b})-2\sin^{-1}(\sqrt{a})} \ge \frac{g(b)-g(a)}{\pi}.$$

This effectively follows by integrating Lemma 6.2.11.

Proof. Let $\mu(p) = C_{a,b}/\sqrt{p(1-p)}$ be the probability density function for the distribution $D_{a,b}$ on the interval (a,b). By Lemma 6.2.11 and the fundamental theorem of calculus, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{p\sim D_{a,b}}\left[\mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}\left[f(c)\cdot\sum_{i\in[n]}\phi^p(c_i)\right]\right] &= \mathop{\mathbb{E}}_{p\sim D_{a,b}}\left[g'(p)\sqrt{p(1-p)}\right] \\
&= \int_a^b g'(p)\sqrt{p(1-p)}\mu(p)\mathrm{d}p \\
&= C_{a,b}\int_a^b g'(p)\mathrm{d}p \\
&= C_{a,b}\cdot(g(b)-g(a)).
\end{aligned}$$

It remains to show that $C_{a,b} = \left(2\sin^{-1}(\sqrt{b}) - 2\sin^{-1}(\sqrt{a})\right)^{-1} \geq 1/\pi$. This follows from observing that

$$C_{a,b}^{-1} = \int_a^b \frac{1}{\sqrt{p(1-p)}}\mathrm{d}p = \int_a^b \left(\frac{\mathrm{d}}{\mathrm{d}p}2\sin^{-1}(\sqrt{p})\right)\mathrm{d}p = 2\sin^{-1}(\sqrt{b}) - 2\sin^{-1}(\sqrt{a})$$

and

$$C_{a,b}^{-1} \leq C_{0,1}^{-1} = 2\sin^{-1}(1) - 2\sin^{-1}(0) = \pi.$$

□

Now we have a lemma to bring consistency into the picture. If f is consistent, $b \approx 1$, and $a \approx 0$, then

$$g(b) - g(a) \approx g(1) - g(0) = f((1)^n) - f((-1)^n) = 1 - (-1) = 2.$$

This gives a lower bound on the correlation for consistent f.

Lemma 6.2.13. *Let $f : \{\pm 1\}^n \to \{\pm 1\}$. Define $g : [0,1] \to [-1,1]$ by $g(p) = \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c)]$. Suppose $\alpha \in [0,1]$. Then $|g(1-\alpha) - g(1)| \leq 2n\alpha$ and $|g(\alpha) - g(0)| \leq 2n\alpha$.*

Proof. We have $\mathop{\mathbb{P}}_{c_{1\cdots n}\sim 1-\alpha}[X = (1)^n] = (1-\alpha)^n$ and

$$\begin{aligned} g(1-\alpha) - g(1) &= f((1)^n) \cdot \mathop{\mathbb{P}}_{c_{1\cdots n}\sim 1-\alpha}[c = (1)^n] + \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c)|c \neq (1)^n] \cdot \mathop{\mathbb{P}}_{c_{1\cdots n}\sim 1-\alpha}[c \neq (1)^n] - g(1) \\ &= g(1) \cdot (1-\alpha)^n + \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c)|c \neq (1)^n] \cdot (1 - (1-\alpha)^n) - g(1) \\ &= \left(g(1) - \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c)|c \neq (1)^n] \right) \cdot ((1-\alpha)^n - 1)\,. \end{aligned}$$

Now $\left| g(1) - \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c)|c \neq (1)^n] \right| \leq 2$ and $|(1-\alpha)^n - 1| \leq n\alpha$,
whence $|g(1-\alpha) - g(1)| \leq 2n\alpha$. The other half of the lemma is symmetric.

□

Robustness

We require the fingerprinting code to be robust to inconsistent answers. We show that the correlation is still good in the presence of inconsistencies.

For $f : \{\pm 1\}^n \to \{\pm 1\}$, define a random variable $\xi_{\alpha,\zeta}(f)$ by

$$\xi_{\alpha,\zeta}(f) = f(c) \cdot \sum_{i\in[n]} \phi^p(c_i) + \gamma \mathbb{I}\left(p \in \{0,1\} \wedge f(c) \neq 2p-1\right), \quad p \sim \overline{D_{\alpha,\zeta}}, \quad c_{1\cdots n} \sim p,$$

where $\mathbb{I}$ is the indicator function and $\gamma \in (0, 1/2)$ satisfies $\zeta\gamma/2 = (1-2\zeta)/\pi$ - that is,

$$\gamma := \frac{2}{\pi}\frac{1-2\zeta}{\zeta}.$$

The first term $f(c) \cdot \sum_{i\in[n]} \phi^p(c_i)$ measures the correlation as before. The second term
$\gamma \mathbb{I}\left(p \in \{0,1\} \wedge f(c) \neq 2p-1\right)$ measures inconsistencies. We will lower bound the expectation of $\xi_{\alpha,\zeta}(f)$, which amounts to saying "either there is good correlation *or* there is an inconsistency with good probability." Thus either the fingerprinting code is able to accuse users or the adversary is forced to be inconsistent.

The following bounds the expected increase in scores from one round of interaction.

Proposition 6.2.14. *Let $f : \{\pm 1\}^n \to \{\pm 1\}$ and $\alpha, \zeta \in (0, 1/2)$. Then*

$$\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] \geq \frac{2}{\pi}(1-2\zeta)(1-2n\alpha).$$

Proof. Define $g : [0,1] \to [-1,1]$ by $g(p) = \mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}[f(c)]$. Now

$$\begin{aligned}
\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] =& \mathop{\mathbb{P}}_{p\sim\overline{D}_{\alpha,\zeta}}[p=0]\cdot\gamma\mathbb{I}(f((-1)^n)=1) + \mathop{\mathbb{P}}_{p\sim\overline{D}_{\alpha,\zeta}}[p=1]\cdot\gamma\mathbb{I}(f((1)^n)=-1) \\
&+ \mathop{\mathbb{P}}_{p\sim\overline{D}_{\alpha,\zeta}}[p\in[\alpha,1-\alpha]]\cdot\mathop{\mathbb{E}}_{p\sim D_{\alpha,1-\alpha}}\left[\mathop{\mathbb{E}}_{c_{1\cdots n}\sim p}\left[f(c)\cdot\sum_{i\in[n]}\phi^p(c_i)\right]\right] \\
=&\zeta\cdot\gamma\left(\mathbb{I}(g(0)=1)+\mathbb{I}(g(1)=-1)\right) \\
&+(1-2\zeta)\cdot\frac{g(1-\alpha)-g(\alpha)}{2\sin^{-1}(\sqrt{1-\alpha})-2\sin^{-1}(\sqrt{\alpha})} \\
&\text{(by Proposition 6.2.12)} \\
\geq&\zeta\cdot\gamma\left(\frac{1+g(0)}{2}+\frac{1-g(1)}{2}\right)+(1-2\zeta)\cdot\frac{g(1-\alpha)-g(\alpha)}{\pi} \\
=&\frac{1-2\zeta}{\pi}\left(1+g(0)+1-g(1)+g(1-\alpha)-g(\alpha)\right) \\
\geq&\frac{1-2\zeta}{\pi}\left(2-|g(\alpha)-g(0)|-|g(1-\alpha)-g(1)|\right) \\
\geq&\frac{1-2\zeta}{\pi}(2-4n\alpha) \\
&\text{(by Lemma 6.2.13).}
\end{aligned}$$

□

Concentration

So far we have shown that the fingerprinting code achieves good correlation or the adversary is not consistent *in expectation*. However, we need this to hold with high

probability. Thus we now show that sums of $\xi_{\alpha,\zeta}(f)$ variables concentrate around their expectation.

Again, the proofs in this section are standard. However, the $\xi_{\alpha,\zeta}(f)$ variables can be quite unwieldy and we are thus unable to apply standard results directly. So instead we must open the proofs and verify that the concentration bounds hold. We proceed by bounding the moment generating function of $\xi_{\alpha,\zeta}(f)$ and then proving an Azuma-like concentration inequality. These calculations are not novel or insightful.

Proposition 6.2.15. *Let $f : \{\pm 1\}^n \to \{\pm 1\}$, $\alpha \in (0, 1/2)$, $\zeta \in [1/4, 1/2)$, and $t \in [-\sqrt{\alpha}/8, \sqrt{\alpha}/8]$. Then*

$$\mathbb{E}\left[e^{t(\xi_{\alpha,\zeta}(f) - \mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right])}\right] \leq e^{Ct^2},$$

where $C = \frac{64e^{n\alpha/4}}{\alpha}$.

Proof. We have

$$\xi_{\alpha,\zeta}(f) = f(c) \cdot \sum_{i \in [n]} \phi^p(c_i) + \gamma \mathbb{I}\left(p \in \{0,1\} \wedge f(c) \neq 2p - 1\right), \quad p \sim \overline{D_{\alpha,\zeta}}, \quad c_{1\cdots n} \sim p.$$

Let $Y = \sum_{i \in [n]} \phi^p(c_i)$. By Lemma 6.2.4 and independence,

$$\mathbb{E}\left[e^{tY}\right] = \mathop{\mathbb{E}}_{c_{1\cdots n} \sim p}\left[e^{t\sum_{i \in [n]} \phi^p(c_i)}\right] = \left(\mathop{\mathbb{E}}_{c \sim p}\left[e^{t\phi^p(c)}\right]\right)^n \leq e^{t^2 n}$$

for $t \in [-\sqrt{\alpha}/2, \sqrt{\alpha}/2]$. Pick $t \in \{\pm\sqrt{\alpha}/2\}$ such that

$$\sum_{k=0}^{\infty} \frac{t^{2k+1}}{(2k+1)!}\mathbb{E}\left[Y^{2k+1}\right] \geq 0.$$

Then by dropping positive terms, for all $j \geq 1$,

$$0 \leq \mathbb{E}\left[Y^{2j}\right] \leq \frac{(2j)!}{t^{2j}} \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbb{E}\left[Y^k\right] = \frac{(2j)!}{t^{2j}}\mathbb{E}\left[e^{tY}\right] \leq \frac{(2j)!}{t^{2j}} e^{nt^2} = \frac{4^j (2j)!}{\alpha^j} e^{n\alpha/4}.$$

Thus we have bounded the even moments of Y. By Cauchy-Schwartz, for $k =$

$2j+1 \geq 3$,

$$\begin{aligned}
\mathbb{E}\left[|Y|^k\right] \leq & \sqrt{\mathbb{E}\left[Y^{2j}\right] \cdot \mathbb{E}\left[Y^{2j+2}\right]} \\
\leq & \sqrt{\frac{4^j(2j)!}{\alpha^j} e^{n\alpha/4} \cdot \frac{4^{j+1}(2j+2)!}{\alpha^{j+1}} e^{n\alpha/4}} \\
= & \frac{2^k k!}{\alpha^{k/2}} e^{n\alpha/4} \sqrt{\frac{k+1}{k}}.
\end{aligned}$$

Since $|f(c)| \leq 1$, we have $\mathbb{E}\left[|f(c) \cdot Y|^k\right] \leq \mathbb{E}\left[|Y|^k\right] \leq 2^{k+1} k! e^{n\alpha/4}/\alpha^{k/2}$ for all $k \geq 2$. Since $\zeta \in [1/4, 1/2)$, we have $\gamma = (2/\pi)(1-2\zeta)/\zeta \in (0,1)$. Hence $\mathbb{E}\left[|\gamma \mathbb{I}\left(p \in \{0,1\} \wedge f(c) \neq 2p-1\right)|^k\right] \leq 1$ for all k. The map $u \mapsto |u|^k$ is convex for all $k \geq 2$, thus $|(x+y)/2|^k \leq (|x|^k + |y|^k)/2$ for all $k \geq 2$ and $x, y \in \mathbb{R}$. Combining these three facts, we have

$$\begin{aligned}
\mathbb{E}\left[|\xi_{\alpha,\zeta}(f)|^k\right] \leq & 2^{k-1} \mathbb{E}\left[|f(c) \cdot Y|^k + |\gamma \mathbb{I}(f(c) \neq f^*(c))|^k\right] \\
\leq & \frac{2^{2k} k! e^{n\alpha/4}}{\alpha^{k/2}} + 2^{k-1} \\
\leq & \frac{2^{2k+1} k! e^{n\alpha/4}}{\alpha^{k/2}}.
\end{aligned}$$

For $t \in [-\sqrt{\alpha}/8, \sqrt{\alpha}/8]$, we have

$$\begin{aligned}
\mathbb{E}\left[e^{t\xi_{\alpha,\zeta}(f)}\right] \leq & 1 + t\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] + \sum_{k=2}^{\infty} \frac{|t|^k}{k!} \mathbb{E}\left[|\xi_{\alpha,\zeta}(f)|^k\right] \\
\leq & 1 + t\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] + \sum_{k=2}^{\infty} \frac{|t|^k}{k!} \frac{2^{2k+1} k! e^{n\alpha/4}}{\alpha^{k/2}} \\
= & 1 + t\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] + 2e^{n\alpha/4} \sum_{k=2}^{\infty} \left(\frac{4|t|}{\sqrt{\alpha}}\right)^k \\
\leq & 1 + t\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] + 2e^{n\alpha/4} \sum_{k=2}^{\infty} \left(\frac{4|t|}{\sqrt{\alpha}}\right)^2 2^{-(k-2)} \\
= & 1 + t\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] + \frac{64 e^{n\alpha/4}}{\alpha} t^2 \\
\leq & e^{t\mathbb{E}\left[\xi_{\alpha,\zeta}(f)\right] + Ct^2}
\end{aligned}$$

□

Theorem 6.2.16 (Azuma-Doob-like Inequality). *Let $X_1 \cdots X_m \in \mathbb{R}$, $\mu_1 \cdots \mu_m \mathbb{R}$ and $\mathcal{U}_0 \cdots \mathcal{U}_m \in \Omega$ be random variables such that, for all $i \in [m]$,*

- *X_i is determined by $\mathcal{U}_i$,*
- *μ_i is determined by $\mathcal{U}_{i-1}$, and*
- *$\mathcal{U}_{i-1}$ is determined by $\mathcal{U}_i$.*

Suppose that, for all $i \in [m]$, $u \in \Omega$, and $t \in [-c, c]$,

$$\mathbb{E}\left[e^{t(X_i - \mu_i)} \mid \mathcal{U}_{i-1} = u\right] \leq e^{Ct^2}.$$

If $\lambda \in [0, 2Cmc]$, then

$$\mathbb{P}\left[\left|\sum_{i \in [m]} (X_i - \mu_i)\right| \geq \lambda\right] \leq 2e^{-\lambda^2/4Cm}.$$

If $\lambda \geq 2Cmc$, then

$$\mathbb{P}\left[\left|\sum_{i \in [m]} (X_i - \mu_i)\right| \geq \lambda\right] \leq 2e^{mCc^2 - c\lambda} \leq 2e^{-c\lambda/2}.$$

Proof. First we show by induction on $k \in [m]$ that, for all $u \in \Omega$ and $t \in [-c, c]$,

$$\mathbb{E}\left[e^{t\sum_{i=m-k+1}^{m}(X_i - \mu_i)} \mid \mathcal{U}_{m-k} = u\right] \leq e^{k \cdot Ct^2}.$$

This clearly holds for $k = 1$, as this is our supposition for $i = m$. Now suppose this

holds for some $k \in [m-1]$. For $u \in \Omega$ and $t \in [-c, c]$, we have

$$\begin{aligned}
&\mathbb{E}\left[e^{t\sum_{i=m-k}^{m}(X_i-\mu_i)} \mid \mathcal{U}_{m-(k+1)} = u\right] \\
&= \sum_{v\in\Omega} \mathbb{P}\left[\mathcal{U}_{m-k} = v \mid \mathcal{U}_{m-k-1} = u\right] \mathbb{E}\left[e^{t\sum_{i=m-k}^{m}(X_i-\mu_i)} \mid \mathcal{U}_{m-k} = v\right] \\
&= \sum_{v\in\Omega} \mathbb{P}\left[v \mid u\right] \mathbb{E}\left[e^{t(X_{m-k}-\mu_{m-k})} e^{t\sum_{i=m-k+1}^{m}(X_i-\mu_i)} \mid v\right] \\
&\text{(using shorthand } v \equiv \mathcal{U}_{m-k} = v \text{ and } u \equiv \mathcal{U}_{m-k-1} = u\text{)} \\
&= \sum_{v\in\Omega} \mathbb{P}\left[v \mid u\right] \mathbb{E}\left[e^{t(X_{m-k}-\mu_{m-k})} \mid v\right] \mathbb{E}\left[e^{t\sum_{i=m-k+1}^{m}(X_i-\mu_i)} \mid v\right] \\
&\text{(since } U_{m-k} = v \text{ determines } X_{m-k} \text{ and } \mu_{m-k}\text{)} \\
&\leq \sum_{v\in\Omega} \mathbb{P}\left[v \mid u\right] \mathbb{E}\left[e^{t(X_{m-k}-\mu_{m-k})} \mid v\right] e^{k\cdot Ct^2} \\
&\text{(by the induction hypothesis)} \\
&= \mathbb{E}\left[e^{t(X_{m-k}-\mu_{m-k})} \mid u\right] e^{k\cdot Ct^2} \\
&\leq e^{Ct^2} e^{k\cdot Ct^2} \\
&\text{(by our supposition for } i = m-k\text{)} \\
&= e^{(k+1)\cdot Ct^2}.
\end{aligned}$$

Thus, for all $t \in [-c, c]$, we have

$$\mathbb{E}\left[e^{t\sum_{i=1}^{m}(X_i-\mu_i)}\right] \leq e^{m\cdot Ct^2}.$$

By Markov's inequality we have

$$\mathbb{P}\left[\sum_{i\in[m]}(X_i-\mu_i) \geq \lambda\right] \leq \frac{\mathbb{E}\left[e^{t\sum_{i\in[m]}(X_i-\mu_i)}\right]}{e^{t\lambda}} \leq e^{mCt^2-t\lambda}$$

and

$$\mathbb{P}\left[\sum_{i\in[m]}(X_i-\mu_i) \leq -\lambda\right] \leq \frac{\mathbb{E}\left[e^{-t\sum_{i\in[m]}(X_i-\mu_i)}\right]}{e^{(-t)(-\lambda)}} \leq e^{mCt^2-t\lambda}$$

for all $t \in [0, c]$ and $\lambda > 0$. Set $t = \min\{c, \lambda/2mC\}$ to obtain the result. □

Bounding the Score

Now we can finally show that the scores are large with high probability.

Theorem 6.2.17 (Correlation Lower Bound)**.** *At the end of* $\mathsf{IFPC}_{N,n,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$ *for arbitrary* $\mathcal{P}$*, we have, for any* $\lambda \in [0, 17.5\ell/\sqrt{\alpha}]$*,*

$$\gamma \mathrm{err}^{\ell} + \sum_{i \in S^1} s_i^{\ell} \geq \frac{2}{\pi}(1 - 2\zeta)(1 - 2n\alpha)\ell - \lambda$$

with probability at least $1 - 2e^{-\frac{\lambda^2 \alpha}{280\ell}}$*.*

Proof. Since the adversary $\mathcal{P}$ is computationally unbounded and arbitrary, we may assume it is deterministic. We may also assume $n = |S^1|$ and that the adversary is able to see $c^j_{S^1}$ at each round. (This only gives the adversary more power.)

This means that for each $j \in [\ell]$ we can define a function $f^j : \{\pm 1\}^n \to \{\pm 1\}$ that only depends on the interaction up to round $j - 1$ (i.e. is a function of the state of $\mathcal{P}$ before it receives c^j) and satisfies $f^j(c^j_{S^j}) = a^j$. For $j \in [\ell]$, define

$$X_j := \gamma \cdot \mathbb{I}\left(p^j \in \{0, 1\} \wedge f^j(c^j_{S^1}) \neq 2p^j - 1\right) + f^j(c^j_{S^1}) \cdot \sum_{i \in S^1} \phi^{p^j}(c^j_i) \sim \xi_{\alpha,\zeta}(f^j),$$

where $\sim$ denotes having the same distribution. We have

$$\gamma \cdot (\mathrm{err}^j - \mathrm{err}^{j-1}) + \sum_{i \in S^1} (s_i^j - s_i^{j-1}) \leq X_j$$

and

$$\gamma \mathrm{err}^{\ell} + \sum_{i \in S^1} s_i^{\ell} \leq \sum_{j \in [\ell]} X_j \sim \sum_{j \in [\ell]} \xi_{\alpha,\zeta}(f^j).$$

Now we can apply the above lemmas to bound the expectation and tail of this random variable.

Firstly, Proposition 6.2.14 shows that

$$\mu_j := \mathbb{E}\left[X_j\right] = \mathbb{E}\left[\xi_{\alpha,\zeta}(f^j)\right] \geq \frac{2}{\pi}(1-2\zeta)(1-2n\alpha)$$

for all f^j. Moreover, by Proposition 6.2.15,

$$\mathbb{E}\left[e^{t(X^j-\mu_j)}\right] = \mathbb{E}\left[e^{t(\xi_{\alpha,\zeta}(f^j)-\mathbb{E}\left[\xi_{\alpha,\zeta}(f^j)\right])}\right] \leq e^{Ct^2}$$

for all $t \in [-\sqrt{\alpha}/8, \sqrt{\alpha}/8]$, where $C = 70/\alpha \geq 64e^{n\alpha/4}/\alpha$, as $\alpha \leq 1/4n$.

Define $\mathcal{U}_j = (f^1, p^1, c^1, \cdots, f^j, p^j, c^j, f^{j+1})$ for $j \in [\ell] \cup \{0\}$. Now $X_1 \cdots X_\ell$, $\mu_1 \cdots \mu_\ell$, and $\mathcal{U}_0, \cdots, \mathcal{U}_\ell$ satisfy the hypotheses of Theorem 6.2.16 with $C = 70/\alpha$, $c = \sqrt{\alpha}/8$, and $m = \ell$.

For $\lambda \in [0, 2Cmc] = [0, 17.5\ell/\sqrt{\alpha}]$, we have

$$\begin{aligned}\mathbb{P}\left[\sum_{j\in[\ell]} X_j \leq \frac{2}{\pi}(1-2\zeta)(1-2n\alpha)\ell - \lambda\right] \leq& \mathbb{P}\left[\left|\sum_{i\in[m]}(X_i-\mu_i)\right| \geq \lambda\right] \\ \leq& 2e^{-\lambda^2/4Cm} \leq 2e^{-\frac{\lambda^2\alpha}{280\ell}},\end{aligned}$$

as required. □

However, we can also prove that the scores are small with high probability. This follows from the fact that users with large scores are accused and therefore no user's score can be too large:

Lemma 6.2.18. *For all $\lambda > 0$,*

$$\mathbb{P}\left[\sum_{i\in S^1} s_i^\ell > \lambda + n\sigma + \frac{n}{\sqrt{\alpha}}\right] \leq e^{-\lambda^2/4n\ell} + e^{-\sqrt{\alpha}\lambda/4},$$

where the probability is taken over $\mathsf{IFPC}_{N,n,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$ for an arbitrary $\mathcal{P}$.

We will set $\lambda = \sigma$ and, since $1/\sqrt{\alpha} \leq \sigma$, we get that $\sum_{i\in S^1} s_i^\ell \leq 3\sigma n$ with high probability.

Proof. Let $j_i \in [\ell+1]$ be as in Lemma 6.2.10 – that is, $i \notin S^{j_i}$ and $i \in S^{j_i-1}$, where we define $S^{\ell+1} = \emptyset$ and $S^0 = [N]$. By the definition of j_i, s^j, and S^j, we have $s_i^{j_i-2} \le \sigma$ for all $i \in S^1$, as otherwise $i \in I^{j_i-2}$ and therefore $i \notin S^{j_i-1} = S^{j_i-2} \setminus I^{j_i-2}$. If $i \in S^1$, then $j_i = 1$ and $s_i^{j_i-1} = 0$. Thus

$$\sum_{i \in S^1} s_i^{j_i-1} = \sum_{i \in S^1} s_i^{j_i-2} + a^{j_i-1}\phi^{p^{j_i-1}}(c_i^{j_i-1}) \le \sum_{i \in S^1} \sigma + \frac{1}{\sqrt{\alpha}} \le n\sigma + \frac{n}{\sqrt{\alpha}}.$$

By Lemma 6.2.10,

$$\mathbb{P}\left[\sum_{i \in S^1} s_i^\ell - s_i^{j_i-1} > \lambda\right] \le e^{-\lambda^2/4n\ell} + e^{-\sqrt{\alpha}\lambda/4}.$$

The lemma follows. □

Now we show that the conflicting bounds of Theorem 6.2.17 and Lemma 6.2.18 imply completeness - that is, the adversary $\mathcal{P}$ cannot be consistent.

Theorem 6.2.19 (Completeness). *At the end of* $\mathsf{IFPC}_{N,n,\ell}[\mathcal{P}, \mathcal{F}_{N,n,\delta,\beta}]$ *for an arbitrary* $\mathcal{P}$*, we have* $\mathrm{err}^\ell > \beta\ell$ *with probability at least* $1 - \delta^{\frac{1}{2}(\frac{1}{2}-\beta)n}$*, assuming* $\left(\frac{1}{2} - \beta\right) n \ge 1$.

Proof. Suppose for the sake of contradiction that $\mathrm{err}^\ell \le \beta\ell$. By Lemma 6.2.18, $\sum_{i \in S^1} s_i^\ell \le \lambda + n\sigma + \frac{n}{\sqrt{\alpha}}$ with probability at least $1 - e^{-\lambda^2/4n\ell} - e^{-\sqrt{\alpha}\lambda/4}$. Set $\lambda = n\sigma \ge \frac{n}{\sqrt{\alpha}}$. Now we assume

$$\sum_{i \in S^1} s_i^\ell \le 3n\sigma,$$

which holds with probability at least $1 - e^{-n\sigma^2/4\ell} - e^{-\sqrt{\alpha}n\sigma/4}$. Then

$$\gamma \mathrm{err}^\ell + \sum_{i \in S^1} s_i^\ell \le \gamma\beta\ell + 3n\sigma. \tag{6.4}$$

By Theorem 6.2.17, with probability at least $1 - 2e^{-\frac{\lambda^2\alpha}{280\ell}}$,

$$\gamma \mathrm{err}^\ell + \sum_{i \in S^1} s_i^\ell \ge \frac{2}{\pi}(1-2\zeta)(1-2n\alpha)\ell - \lambda \tag{6.5}$$

for all $\lambda \in [0, 17.5\ell/\sqrt{\alpha}]$. Set $\lambda = \left(\frac{1}{2} - \beta\right)^2 \ell/2\pi$ and assme Equation (6.5) also holds.

Combining Equations (6.4) and (6.5) gives

$$\frac{2}{\pi}(1-2\zeta)(1-2n\alpha)\ell - \frac{\left(\frac{1}{2}-\beta\right)^2}{2\pi}\ell \leq \gamma\beta\ell + 3n\sigma. \tag{6.6}$$

We claim this is a contradiction, which then holds with high probability, thus proving the theorem.

Rearranging Equation (6.6) gives

$$\frac{2}{\pi}(1-2\zeta)(1-2n\alpha) \leq \frac{\left(\frac{1}{2}-\beta\right)^2}{2\pi} + \gamma\beta + \frac{3n\sigma}{\ell}. \tag{6.7}$$

Our setting of parameters gives

$$2n\alpha \leq \frac{\left(\frac{1}{2}-\beta\right)}{2} \quad \text{and} \quad \frac{3n\sigma}{\ell} \leq \frac{\left(\frac{1}{2}-\beta\right)^2}{2\pi}.$$

Substituting these into Equation (6.7) gives

$$\frac{2}{\pi}(1-2\zeta)\left(1 - \frac{1}{2}\left(\frac{1}{2}-\beta\right)\right) \leq \frac{\left(\frac{1}{2}-\beta\right)^2}{\pi} + \gamma\beta. \tag{6.8}$$

Now we use $1-2\zeta = \frac{1}{2}\left(\frac{1}{2}-\beta\right)$ and $\gamma = \frac{2}{\pi}\frac{1-2\zeta}{\zeta} = \frac{\left(\frac{1}{2}-\beta\right)}{\pi\zeta}$ to derive a contradiction from Equation (6.8):

$$\begin{aligned}
\frac{\left(\frac{1}{2}-\beta\right)}{\pi}\left(1 - \frac{1}{2}\left(\frac{1}{2}-\beta\right)\right) &\leq \frac{\left(\frac{1}{2}-\beta\right)^2}{\pi} + \frac{\left(\frac{1}{2}-\beta\right)}{\pi\zeta}\beta, \\
1 - \frac{1}{2}\left(\frac{1}{2}-\beta\right) &\leq \left(\frac{1}{2}-\beta\right) + \frac{\beta}{\zeta}, \\
\zeta\left(1 - \frac{3}{2}\left(\frac{1}{2}-\beta\right)\right) &\leq \beta.
\end{aligned}$$

Since $\zeta = \frac{1}{2} - \frac{1}{4}\left(\frac{1}{2} - \beta\right)$, we have

$$\zeta\left(1 - \frac{3}{2}\left(\frac{1}{2} - \beta\right)\right) = \frac{1}{2}\left(1 - \frac{1}{2}\left(\frac{1}{2} - \beta\right)\right)\left(1 - \frac{3}{2}\left(\frac{1}{2} - \beta\right)\right) > \frac{1}{2}\left(1 - 2\left(\frac{1}{2} - \beta\right)\right).$$

And

$$\beta = \frac{1}{2}\left(1 - 2\left(\frac{1}{2} - \beta\right)\right).$$

This gives a contradiction. The total failure probability is bounded by

$$e^{-n\sigma^2/4\ell} + e^{-\sqrt{\alpha}n\sigma/4} + 2e^{-\lambda^2\alpha/280\ell} \leq \left(\frac{\delta}{32}\right)^{16n} + \left(\frac{\delta}{32}\right)^{4n} + 2\left(\frac{\delta}{32}\right)^{\frac{1}{2}\left(\frac{1}{2}-\beta\right)n} \leq \delta^{\frac{1}{2}\left(\frac{1}{2}-\beta\right)n},$$

assuming $\left(\frac{1}{2} - \beta\right) n \geq 1$. □

6.2.6 Non-Interactive Fingerprinting Codes

Our construction and analysis also gives a construction of traditional non-interactive fingerprinting codes. First we give a formal definition of a fingerprinting code.

Definition 6.2.20 ((Non-Interactive) Fingerprinting Codes)**.** *A n-collusion resilient* (non-interactive) fingerprinting code *of length ℓ for N users robust to a β fraction of errors with failure probability ε and false accusation probability δ is a pair of random variables $C \in \{\pm 1\}^{N \times \ell}$ and $Trace : \{\pm 1\}^{\ell} \to 2^{[N]}$ such that the following holds. For all adversaries $\mathcal{P} : \{\pm 1\}^{n \times \ell} \to \{\pm 1\}^{\ell}$ and $S \subset [N]$ with $|S| = n$,*

$$\mathop{\mathbb{P}}_{C,Trace,\mathcal{P}}\left[\left(\left|\left\{1 \leq j \leq \ell : \nexists i \in [N] \;\; \mathcal{P}(C_S)^j = c_i^j\right\}\right| \leq \beta\ell\right) \wedge \left(Trace(\mathcal{P}(C_S)) = \emptyset\right)\right] \leq \varepsilon$$

and

$$\mathop{\mathbb{P}}_{C,Trace,\mathcal{P}}\left[\left|Trace(\mathcal{P}(C_S)) \cap ([N] \backslash S)\right| > \delta(N - n)\right] \leq \varepsilon,$$

where $C_S \in \{\pm 1\}^{n \times \ell}$ contains the rows of C given by S.

Our construction and analysis is readily adapted to the non-interactive setting.

We obtain the following theorem.

Theorem 6.2.21 (Existence of Non-Interactive Fingerprinting Codes)**.** *For every* $1 \leq n \leq N$, $0 \leq \beta < 1/2$, *and* $0 < \delta \leq 1$, *there is a n-collusion-resilient (non-interactive) fingerprinting code of length* ℓ *for* N *users robust to a* β *fraction of errors with failure probability*

$$\varepsilon \leq \min\{\delta(N-n), 2^{-\Omega(\delta(N-n))}\} + \delta^{\Omega\left(\left(\frac{1}{2}-\beta\right)n\right)}$$

and false accusation probability δ *for*

$$\ell = O\left(\frac{n^2 \log(1/\delta)}{\left(\frac{1}{2}-\beta\right)^4}\right).$$

6.3 Hardness of False Discovery

In this section we prove our main result - that answering $O(n^2)$ adaptive queries given n samples is hard. But first we must formally define the model in which we are working.

6.3.1 The Statistical Query Model

Given a distribution $\mathcal{P}$ over $\{0,1\}^d$, we would like to answer *statistical queries* about $\mathcal{P}$. A statistical query on $\{0,1\}^d$ is specified by a function $q : \{0,1\}^d \to [-1,1]$ and (abusing notation) is defined to be

$$q(\mathcal{P}) = \mathop{\mathbb{E}}_{x \leftarrow_R \mathcal{P}}[q(x)].$$

Our goal is to design a *mechanism* $\mathcal{M}$ that answers statistical queries on $\mathcal{P}$ using only iid samples $x_1, \ldots, x_n \leftarrow_R \mathcal{P}$. Our focus is the case where the queries are chosen adaptively and adversarially.

Specifically, $\mathcal{M}$ is a stateful algorithm that holds a collection of samples $x_1, \ldots, x_n$ in $\{0,1\}^d$, takes a statistical query q as input, and returns a real-valued answer $a \in [-1,1]$. We require that when $x_1, \ldots, x_n$ are iid samples from $\mathcal{P}$, the answer a is close to $q(\mathcal{P})$, and moreover that this condition holds for every query in an adaptively chosen sequence $q^1, \ldots, q^\ell$. Formally, we define the following game between an $\mathcal{M}$ and a stateful adversary $\mathcal{A}$.

$\mathcal{A}$ chooses a distribution $\mathcal{P}$ over $\{0,1\}^d$.
Sample $x_1, \ldots, x_n \leftarrow_{\mathrm{R}} \mathcal{P}$, let $x = (x_1, \ldots, x_n)$.
For $j = 1, \ldots, \ell$
  $\mathcal{A}$ outputs a query q^j.
  $\mathcal{M}(x, q^j)$ outputs a^j.
  (As $\mathcal{A}$ and $\mathcal{M}$ are stateful, q^j and a^j may depend on the history $q^1, a^1, \ldots, q^{j-1}, a^{j-1}$.)

Figure 6.3: $\mathsf{Acc}_{n,d,\ell}[\mathcal{M}, \mathcal{A}]$

Definition 6.3.1 (Accuracy)**.** *An mechanism $\mathcal{M}$ is (α, β, γ)-*accurate for ℓ adaptively chosen queries given n samples in $\{0,1\}^d$ *if for every adversary $\mathcal{A}$,*

$$\mathop{\mathbb{P}}_{\mathsf{Acc}_{n,d,\ell}[\mathcal{M},\mathcal{A}]} \left[\textit{For } (1-\beta)\ell \textit{ choices of } j \in [\ell],\ \left| \mathcal{M}(x, q^j) - q^j(\mathcal{P}) \right| \leq \alpha \right] \geq 1 - \gamma \,.$$

*As a shorthand, we will say that $\mathcal{M}$ is (α, β)-*accurate for ℓ queries *if for every $n, d \in \mathbb{N}$, $\mathcal{M}$ is $(\alpha, \beta, o_n(1))$-accurate for ℓ queries given n samples in $\{0,1\}^d$. Here, ℓ may depend on n and d and $o_n(1)$ is a function of n that tends to 0.*

We are interested in mechanisms that are both accurate and computationally efficient. We say that a mechanism $\mathcal{M}$ is *computationally efficient* if, when given samples $x_1, \ldots, x_n \in \{0,1\}^d$ and a query $q : \{0,1\}^d \to [-1,1]$, it runs in time $\mathrm{poly}(n, d, |q|)$. Here q will be represented as a circuit that evaluates $q(x)$ and $|q|$ denotes the size of this circuit.

6.3.2 Encryption Schemes

Our attack relies on the existence of a semantically secure private-key encryption scheme. An encryption scheme is a triple of efficient algorithms (Gen, Enc, Dec) with the following syntax:

- *Gen* is a randomized algorithm that takes as input a security parameter λ and outputs a λ-bit secret key. Formally, $sk \leftarrow_R Gen(1^\lambda)$.
- *Enc* is a randomized algorithm that takes as input a secret key and a message $m \in \{-1, 0, 1\}$ and outputs a ciphertext $ct \in \{0,1\}^{\text{poly}(\lambda)}$. Formally, $ct \leftarrow_R Enc(sk, m)$.
- *Dec* is a deterministic algorithm that takes as input a secret key and a ciphertext ct and outputs a decrypted message m'. If the ciphertext ct was an encryption of m under the key sk, then $m' = m$. Formally, if $ct \leftarrow_R Enc(sk, m)$, then $Dec(sk, ct) = m$ with probability 1.

Roughly, security of the encryption scheme asserts that no polynomial time adversary who does not know the secret key can distinguish encryptions of $m = 0$ from encryptions of $m = 1$, even if the adversary has access to a mechanism that returns the encryption of an arbitrary message under the unknown key. For convenience, we will require that this security property holds simultaneously for an arbitrary polynomial number of secret keys. The existence of an encryption scheme with this property follows immediately from the existence an ordinary semantically secure encryption scheme. We start with the stronger definition only to simplify our proofs. A secure encryption scheme exists under the minimal cryptographic assumption that one-way functions exist. The formal definition of security is not needed until Section 6.4.

6.3.3 The Attack

The adversary is specified in Figure 6.4. Observe that $\mathsf{Attack}_{n,d}$ is only well defined for pairs $n, d \in N$ for which $1 + \lceil \log_2(2000n) \rceil \leq d$, so that there exists a suitable choice of $\lambda \in \mathbb{N}$. Through this section we will assume that $n = n(d)$ is a polynomial in d and that d is a sufficiently large unspecified constant, which ensures that $\mathsf{Attack}_{n,d}$ is well defined.

The distribution $\mathcal{P}$:
Given parameters d, n, let $N = 2000n$, let $\lambda = d - \lceil \log_2(N) \rceil$.
Let (Gen, Enc, Dec) be an encryption scheme
For $i \in [N]$, let $sk_i \leftarrow_{\mathrm{R}} Gen(1^\lambda)$ and let $y_i = (i, sk_i) \in \{0,1\}^d$.
Let $\mathcal{P}$ be the uniform distribution over $\{y_1, \ldots, y_N\} \subseteq \{0,1\}^d$.

$\mathcal{M}$ samples $x_1, \ldots, x_n \leftarrow_{\mathrm{R}} \mathcal{P}$. Let $x = (x_1, \ldots, x_n)$.
Let $S \subseteq [N]$ be the set of unique indices i such that (i, sk_i) appears in x.

Attack:
Initialise a n-collusion resilient interactive fingerprinting code $\mathcal{F}$ of length ℓ for N users robust to a β fraction of errors with failure probability $\varepsilon = \mathrm{negl}(n)$ and false accusation probability $\delta = 1/1000$.
Let $T^1 = \emptyset$.
For $j = 1, \ldots, \ell = \ell(N)$:
Let $c^j \in \{\pm 1\}^N$ be the column given by $\mathcal{F}$.
For $i = 1, \ldots, N$, let $ct_i^j = Enc(sk_i, c_i^j)$.
Define the query $q^j(i', sk')$ to be $Dec(sk', ct_{i'}^j)$ if $i' \notin T^j$ and 0 otherwise.
Let $a^j = \mathcal{M}(x; q^j)$ and round a^j to $\{\pm 1\}$ to obtain $\overline{a}^j$.
Give $\overline{a}^j$ to $\mathcal{F}$ and let $I^j \subseteq [N]$ be the set of accused users and $T^j = T^{j-1} \cup I^j$.

Figure 6.4: $\mathsf{Attack}_{n,d}[\mathcal{M}]$

6.3.4 Informal Analysis of the Attack

Before formally analysing the attack, we comment on the overall structure thereof. At a high level, the attack $\mathsf{Attack}_{n,d}[\mathcal{M}]$ runs the fingerprinting game $\mathsf{IFPC}_{N,n,\ell}[\mathcal{P}, \mathcal{F}]$,

where the mechanism $\mathcal{M}$ plays the rôle of the fingerprinting adversary $\mathcal{P}$. Each challenge c^j issued by $\mathcal{F}$ is passed to the mechanism in encrypted form as q^j. The mechanism must output an approximation a^j to the true answer

$$q^j(\mathcal{P}) = \frac{1}{N} \sum_{i \in [N] \setminus T^j} c_i^j.$$

In order to do this, the mechanism could decrypt q^j to obtain c^j for every j. However, the mechanism does not have all the necessary secret keys; it only has the secret keys corresponding to its sample S. Thus, by the security of the encryption scheme, any efficient mechanism effectively can only see $c^j_{S \setminus T^j}$. That is to say, if the mechanism is computationally efficient, then it has the same restriction as a fingerprinting adversary $\mathcal{P}$. Thus, any computationally efficient mechanism must lose the fingerprinting game, meaning it cannot answer every query (or even just a $\beta = 1/2 + \Omega(1)$ fraction of the queries) accurately.

One subtly arises since "accuracy" for the mechanism is defined with respect to the true answer $q^j(\mathcal{P}) = \frac{1}{N} \sum_{i \in [N] \setminus T^j} c_i^j$, whereas "accuracy" in the fingerprinting game is defined with respect to the average over all of c^j, that is $\frac{1}{N} \sum_{i \in [N]} c_i^j$. We deal with these subtleties by arguing that T^j, which is the number of users accused by the interactive fingerprinting code prior to the j-th query, is small. Here we use the fact that the fingerprinting code only allows a relatively small number of false accusations $N/1000$. Therefore $|T^j| \leq n + N/1000 \leq N/500$. As a result, the definition of accuracy guaranteed by the mechanism will be close enough to the definition of accuracy required for the interactive fingerprinting code to succeed in identifying the sample.

6.3.5 Analysis of the Attack

In this section we prove our main result:

Theorem 6.3.2 (Theorem 6.1.1)**.** *Assuming one-way functions exist, for all $\beta < 1/2$, there is a function $\ell(2000n, \beta) = O(n^2/\left(\frac{1}{2} - \beta\right)^4)$ such that there is no computationally efficient mechanism $\mathcal{M}$ that is $(0.99, \beta, 1/2)$-accurate for $\ell(2000n, \beta)$ adaptively chosen queries given n samples in $\{0,1\}^d$.*

We will start by establishing that the number of falsely accused users is small. That is, we have $|T^\ell \setminus S| \leq N/1000$ with high probability. This condition will follow from the security of the interactive fingerprinting code $\mathcal{F}$. However, security alone is not enough to guarantee that the number of falsely accused users is small, because security of $\mathcal{F}$ applies to adversaries that only have access to c_i^j for users $i \in S \setminus T^j$, whereas the queries to the mechanism depend on c_i^j for users $i \notin S \setminus T^j$. To remedy this problem we rely on the fact entries c_i^j for i outside of $S \setminus T^j$ are encrypted under keys sk_i that are not known to the mechanism. Thus, a computationally efficient mechanism "does not know" those rows. We can formalise this argument by comparing Attack to an IdealAttack (Figure 6.5) where these entries are replaced with zeros, and argue that the adversary cannot distinguish between these two attacks without breaking the security of the encryption scheme.

Claim 6.3.3. *For every mechanism $\mathcal{M}$, every polynomial $n = n(d)$, and every sufficiently large $d \in \mathbb{N}$,*

$$\mathbb{P}_{\mathsf{IdealAttack}_{n,d}[\mathcal{M}]}\left[|T^\ell \setminus S| > N/1000\right] \leq \mathrm{negl}(n)$$

Proof. This follows straightforwardly from a reduction to the security of the fingerprinting code. Notice that the query q^j does not depend on any entry c_i^j for $i \notin S \setminus T^{j-1}$. Thus, an adversary for the fingerprinting code who has access to

The distribution $\mathcal{P}$:
Given parameters d, n, let $N = 2000n$, and $\lambda = d - \lceil \log_2(N) \rceil$.
Let (Gen, Enc, Dec) be an encryption scheme
For $i \in [N]$, let $sk_i \leftarrow_R Gen(1^\lambda)$ and let $y_i = (i, sk_i) \in \{0,1\}^d$.
Let $\mathcal{P}$ be the uniform distribution over $\{y_1, \ldots, y_N\} \subseteq \{0,1\}^d$.

Choose samples $x_1, \ldots, x_n \leftarrow_R \mathcal{P}$, let $x = (x_1, \ldots, x_n)$.
Let $S \subseteq [N]$ be the set of unique indices i such that (i, sk_i) appears in x.

Recovery phase:
Initialise a n-collusion resilient interactive fingerprinting code $\mathcal{F}$ of length ℓ for N users robust to a β fraction of errors with failure probability $\varepsilon = \text{negl}(n)$ and false accusation probability $\delta = 1/1000$.
Let $T^1 = \emptyset$.
For $j = 1, \ldots, \ell = \ell(N)$:
Let $c^j \in \{\pm 1\}^N$ be the column given by $\mathcal{F}$.
For $i \in S$, let $ct_i^j = Enc(sk_i, c_i^j)$, for $i \in [N] \setminus S$, let $ct_i^j = Enc(sk_i, 0)$.
Define the query $q^j(i', sk')$ to be $Dec(sk', ct_{i'}^j)$ if $i' \notin T^j$ and 0 otherwise.
Let $a^j = \mathcal{M}(x; q^j)$ and round a^j to $\{\pm 1\}$ to obtain $\overline{a}^j$.
Give $\overline{a}^j$ to $\mathcal{F}$ and let $I^j \subseteq [N]$ be the set of accused users and $T^j = T^{j-1} \cup I^j$.

Figure 6.5: $\mathsf{IdealAttack}_{n,d}[\mathcal{M}]$

$c^j_{S \setminus T^{j-1}}$ can simulate the view of the mechanism. Since we have for any adversary $\mathcal{P}$

$$\mathop{\mathbb{P}}_{\mathsf{IFPC}_{N,n,\ell}[\mathcal{P},\mathcal{F}]} \left[\psi^\ell > (N-n)\delta \right] \leq \varepsilon,$$

we also have

$$\mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{M}]} \left[|T^\ell \setminus S| > N/1000 \right] \leq \text{negl}(n),$$

as desired. □

Now we can argue that an efficient mechanism cannot distinguish between the real attack and the ideal attack. Thus the conclusion that $|T^\ell \setminus S| \leq N/1000$ with high probability must also hold in the real game.

Claim 6.3.4. *Let Z_1 be the event $\left\{ |T^\ell \setminus S| > N/1000 \right\}$. Assume (Gen, Enc, Dec) is a*

computationally secure encryption scheme and let $n = n(d)$ be any polynomial. Then, if $\mathcal{M}$ is computationally efficient, for every sufficiently large $d \in \mathbb{N}$

$$\left| \mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{M}]} [Z_1] - \mathop{\mathbb{P}}_{\mathsf{Attack}_{n,d}[\mathcal{M}]} [Z_1] \right| \leq \mathrm{negl}(n)$$

The proof is straightforward from the definition of security, and is deferred to Section 6.4. Combining Claims 6.3.3 and 6.3.4 we easily obtain the following.

Claim 6.3.5. *For every computationally efficient mechanism $\mathcal{M}$, every polynomial $n = n(d)$, and every sufficiently large $d \in \mathbb{N}$,*

$$\mathop{\mathbb{P}}_{\mathsf{Attack}_{n,d}[\mathcal{M}]} \left[|T^\ell \setminus S| > N/1000 \right] \leq \mathrm{negl}(n)$$

Claim 6.3.5 will be useful because it will allow us to establish that an accurate mechanism must give answers that are consistent with the fingerprinting code. That is, using θ^ℓ to denote the number of inconsistent answers $\overline{a}^1, \ldots, \overline{a}^\ell$, we will have $\theta^\ell \ll \ell/2$ with high probability.

Claim 6.3.6. *If $\mathcal{M}$ is $(0.99, \beta, 1/2)$-accurate for $\ell = \ell(2000n)$ adaptively chosen queries then, for every polynomial $n = n(d)$ and every sufficiently large $d \in \mathbb{N}$,*

$$\mathop{\mathbb{P}}_{\mathsf{Attack}_{n,d}[\mathcal{M}]} \left[\theta^\ell \leq \beta\ell \right] \geq 1/2 - \mathrm{negl}(n)$$

Proof. In the attack, the mechanism's input consists of n samples from $\mathcal{P}$, and the total number of queries issued is ℓ. Therefore, by the assumption that $\mathcal{M}$ is $(0.99, \beta, 1/2)$-accurate for ℓ queries, we have

$$\mathbb{P}\left[\begin{array}{c} \text{For } (1-\beta)\ell \text{ choices of } j \in [\ell], \\ \left| \mathcal{M}(x, q^j) - \mathop{\mathbb{E}}_{(i,sk_i) \leftarrow_R \mathcal{P}} \left[q^j(i, sk_i) \right] \right| \leq 0.99 \end{array} \right] \geq 1/2. \tag{6.9}$$

Observe that, by construction, for every $j \in [\ell]$,

$$\begin{aligned}
&\left| \mathop{\mathbb{E}}_{(i,sk_i)\leftarrow_R \mathcal{P}} \left[q^j(i, sk_i) \right] - \mathop{\mathbb{E}}_{i\in[N]} \left[c_i^j \right] \right| \\
&= \left| \left(\frac{1}{N} \sum_{i\in[N]\setminus T^{j-1}} Dec(sk_i, ct_i^j) + \frac{1}{N} \sum_{i\in T^{j-1}} 0 \right) - \mathop{\mathbb{E}}_{i\in[N]} \left[c_i^j \right] \right| \\
&= \left| \left(\frac{1}{N} \sum_{i\in[N]\setminus T^{j-1}} c_i^j \right) - \frac{1}{N} \sum_{i\in[N]} c_i^j \right| \\
&= \left| -\frac{1}{N} \sum_{i\in T^{j-1}} c_i^j \right| \\
&\leq \frac{\left|T^{j-1}\right|}{N} \\
&\leq \frac{|T^{j-1} \setminus S| + |S|}{N}
\end{aligned} \tag{6.10}$$

where the second equality is because by construction $ct_i^j \leftarrow_R Enc(sk_i, c_i^j)$ and the inequality is because we have $c_i^j \in \{\pm 1\}$.

By Claim 6.3.5, and the fact that $T^{j-1} \subseteq T^\ell$, we have

$$\mathbb{P}\left[|T^{j-1} \setminus S| > N/1000 \right] \leq \mathrm{negl}(n).$$

Noting that $N/1000 + n < N/500$ and combining with (6.10), we have

$$\mathbb{P}\left[\forall\, j \in [\ell],\ \left| \mathop{\mathbb{E}}_{(i,sk_i)\leftarrow_R \mathcal{P}} \left[q^j(i, sk_i) \right] - \mathop{\mathbb{E}}_{i\in[n]} \left[c_i^j \right] \right| \leq 1/500 \right] \geq 1 - \mathrm{negl}(n) \tag{6.11}$$

Applying the triangle inequality to (6.9) and (6.11), we obtain

$$\mathbb{P}\left[\begin{array}{c} \text{For } (1-\beta)\ell \text{ choices of } j \in [\ell], \\ \left| \mathcal{M}(x, q^j) - \mathop{\mathbb{E}}_{i\in[N]} \left[c_i^j \right] \right| \leq 0.99 + 1/500 \end{array} \right] \geq 1/2 - \mathrm{negl}(n). \tag{6.12}$$

Fix a $j \in [\ell]$ such that a^j is 0.99-accurate for query q^j. If $c_i^j = 1$ for every $i \in [N]$, then $a^j = \mathcal{M}(x, q^j) \geq 1 - 0.99 - 1/500$, so the rounded answer $\overline{a}^j = 1$. Similarly

if $c_i^j = -1$ for every $i \in [N]$, $\overline{a}^j = -1$. Therefore there must exist $i \in [N]$ so that $\overline{a}^j = c_i^j$. Thus there are $(1-\beta)\ell$ choices of $j \in [\ell]$ for which this condition holds, so the number of errors θ^ℓ is at most $\beta\ell$. This completes the proof of the claim. □

As before, we can argue that the real attack and the ideal attack are computationally indistinguishable, and thus the mechanism must also give consistent answers in the ideal attack.

Claim 6.3.7. *Let Z_2 be the event $\left\{\theta^\ell \leq \beta\ell\right\}$. Assume (Gen, Enc, Dec) is a computationally secure encryption scheme and let $n = n(d)$ be any polynomial. Then if $\mathcal{M}$ is computationally efficient, for every $d \in \mathbb{N}$*

$$\left| \mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{M}]}[Z_2] - \mathop{\mathbb{P}}_{\mathsf{Attack}_{n,d}[\mathcal{M}]}[Z_2] \right| \leq \mathrm{negl}(n)$$

The proof is straightforward from the definition of security, and is deferred to Section 6.4. Combining Claims 6.3.6 and 6.3.7 we easily obtain the following.

Claim 6.3.8. *If $\mathcal{M}$ computationally efficient and $(0.99, \beta, 1/2)$-accurate for $\ell = \ell(2000n)$ adaptively chosen queries then for every polynomial $n = n(d)$ and every sufficiently large $d \in \mathbb{N}$,*

$$\mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{M}]}\left[\theta^\ell \leq \beta\ell\right] \geq 1/2 - \mathrm{negl}(n).$$

However, the conclusion of 6.3.8 can easily be seen to lead to a contradiction, because the security of the fingerprinting code assures that no attacker who only has access to $c^j_{S \setminus T^{j-1}}$ in each round $j = 1, \ldots, \ell$ can give answers that are consistent for $(1-\beta)\ell$ of the columns c^j. Thus, we have

Claim 6.3.9. *For every mechanism $\mathcal{M}$, every polynomial $n = n(d)$, and every sufficiently large $d \in \mathbb{N}$,*

$$\mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{M}]}\left[\theta^\ell \leq \beta\ell\right] \leq \mathrm{negl}(n)$$

Putting the above claims together, we obtain the main theorem:

Proof of Theorem 6.3.2. Assume for the sake of contradiction that there were such a mechanism. Theorem 6.2.2 implies that an interactive fingerprinting code of length $O(n^2/\left(\frac{1}{2}-\beta\right)^4)$ exists, so the attack can be carried out. By Claim 6.3.8 we would have

$$\mathbb{P}_{\mathsf{IdealAttack}_{n,d}[\mathcal{M}]}\left[\theta^\ell \leq \beta\ell\right] \geq 1/2 - \mathrm{negl}(n).$$

But, by Claim 6.3.9 we have

$$\mathbb{P}_{\mathsf{IdealAttack}_{n,d}[\mathcal{M}]}\left[\theta^\ell \leq \beta\ell\right] \leq \mathrm{negl}(n),$$

which is a contradiction. □

Note that the constants in the $(0.99, \beta, 1/2)$-accuracy assumption are arbitrary and have only been fixed for simplicity.

6.3.6 An Information-Theoretic Lower Bound

As in [HU14], we observe that the techniques underlying our computational hardness result can also be used to prove an information-theoretic lower bound when the dimension of the data is large. At a high level, the argument uses the fact that the encryption scheme we rely on only needs to satisfy relatively weak security properties, specifically security for at most $O(n^2)$ messages. This security property can actually be achieved against computationally unbounded adversaries provided that the length of the secret keys is $O(n^2)$. As a result, our lower bound can be made to hold against computationally unbounded mechanisms, but since the secret keys have length $O(n^2)$, we will require $d = O(n^2)$. We refer the reader to [HU14] for a slightly more detailed discussion, and simply state the following result.

Theorem 6.3.10 (Theorem 6.1.2). *For all $\beta < 1/2$, there is a function $\ell(2000n, \beta) = O(n^2/\left(\frac{1}{2} - \beta\right)^4)$ such that there is no mechanism $\mathcal{M}$ (even one that is computationally unbounded) that is $(0.99, \beta, 1/2)$-accurate for $\ell(2000n, \beta)$ adaptively chosen queries given n samples in $\{0,1\}^d$ when $d \geq \ell(2000n, \beta)$.*

6.4 Security Reductions from Sections 6.3

In Section 6.3 we made several claims comparing the probability of events in Attack to the probability of events in IdealAttack. Each of these claims follow from the assumed security of the encryption scheme. In this section we restate and prove these claims. Since the claims are all of a similar nature, the proof will be somewhat modular.

Before we begin recall the formal definition of security of an encryption scheme. Security is defined via a pair of mechanisms $\mathcal{E}_0$ and $\mathcal{E}_1$. $\mathcal{E}_1(sk_1, \ldots, sk_N, \cdot)$ takes as input the index of a key $i \in [N]$ and a message m and returns $Enc(sk_i, m)$, whereas $\mathcal{E}_0(sk_1, \ldots, sk_N, \cdot)$ takes the same input but returns $Enc(sk_i, 0)$. The security of the encryption scheme asserts that for randomly chosen secret keys, no computationally efficient adversary can tell whether or not it is interacting with $\mathcal{E}_0$ or $\mathcal{E}_1$.

Definition 6.4.1. *An encryption scheme (Gen, Enc, Dec) is* secure *if for every polynomial $N = N(\lambda)$, and every* poly(λ)*-time adversary $\mathcal{B}$, if $sk_1, \ldots, sk_N \leftarrow_R Gen(1^\lambda)$*

$$\left| \mathbb{P}\left[\mathcal{B}^{\mathcal{E}_0(sk_1,\ldots,sk_N,\cdot)} = 1\right] - \mathbb{P}\left[\mathcal{B}^{\mathcal{E}_1(sk_1,\ldots,sk_N,\cdot)} = 1\right] \right| = \text{negl}(\lambda)$$

We now restate the relevant claims from Section 6.3.

Claim 6.4.2 (Claim 6.3.4 Restated). *Let Z_1 be the event $\left\{\psi^\ell > N/8\right\}$.*

Assume (Gen, Enc, Dec) is a computationally secure encryption scheme and let $n = n(d)$

be any polynomial. Then if $\mathcal{M}$ is computationally efficient, for every $d \in \mathbb{N}$

$$\left| \mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{M}]}[Z_1] - \mathop{\mathbb{P}}_{\mathsf{Attack}_{n,d}[\mathcal{M}]}[Z_1] \right| \leq \mathrm{negl}(n)$$

Claim 6.4.3 (Claim 6.3.7 Restated)**.** *Let Z_2 be the event $\left\{\theta^\ell \leq \beta\ell\right\}$.*
Assume (Gen, Enc, Dec) is a computationally secure encryption scheme and let $n = n(d)$ be any polynomial. Then if $\mathcal{M}$ is computationally efficient, for every $d \in \mathbb{N}$

$$\left| \mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{M}]}[Z_2] - \mathop{\mathbb{P}}_{\mathsf{Attack}_{n,d}[\mathcal{M}]}[Z_2] \right| \leq \mathrm{negl}(n)$$

To prove both of these claims, for $c \in \{1, 2\}$ we construct an adversary $\mathcal{B}_c$ that will attempt to use $\mathcal{M}$ to break the security of the encryption. We construct $\mathcal{B}_c$ in such a way that its advantage in breaking the security of encryption is precisely the difference in the probability of the event Z_c between Attack and IdealAttack, which implies that the difference in probabilities is negligible. The simulator is given in Figure 6.6

Proof of Claims 6.4.2, 6.4.3. First, observe that for $c \in \{1, 2\}$, $\mathcal{B}_c$ is computationally efficient as long as $\mathcal{F}$ and $\mathcal{M}$ are both computationally efficient. It is not hard to see that our construction $\mathcal{F}$ is efficient and efficiency of $\mathcal{M}$ is an assumption of the claim. Also notice $\mathcal{B}$ can determine whether Z_c has occurred efficiently.

Now we observe that when the mechanism is $\mathcal{E}_1$ (the mechanism that takes as input i and m and returns $Enc(\overline{sk}_i, m)$), and $\overline{sk}_1, \ldots, \overline{sk}_N$ are chosen randomly from $Gen(1^\lambda)$, then the view of the mechanism is identical to $\mathsf{Attack}_{n,d}[\mathcal{M}]$. Specifically, the mechanism holds a random sample of pairs (i, sk_i) and is shown queries that are encryptions either under keys it knows or random unknown keys. Moreover, the messages being encrypted are chosen from the same distribution. On the other hand, when the mechanism is $\mathcal{E}_0$ (the mechanism that takes as input i and ct and

Simulate constructing and sampling from $\mathcal{P}$:
Given parameters d, n, let $N = 2000n$, let $\lambda = d - \lceil \log_2(2000n) \rceil$.
Sample users $u_1, \ldots, u_n \leftarrow_R [N]$, let S be the set of distinct users in the sample.
Choose new keys $sk_i \leftarrow_R Gen(1^\lambda)$ for $i \in S$.
For $i \in S$, let $x_i = (u_i, sk_{u_i})$, let $x = (x_1, \ldots, x_n)$.

Simulate the attack:
Let $T^1 = \emptyset$.
For $j = 1, \ldots, \ell = \ell(N)$:
Let c^j be the column given by $\mathcal{F}$.
For $i = 1, \ldots, N$:
If $i \in S$, let $ct_i^j = Enc(sk_i, c_i^j)$, otherwise as $\mathcal{E}$ for an encryption of c_i^j under
key $\overline{sk}_i$, that is $ct_i^j = \mathcal{E}_b(\overline{sk}_1, \ldots, \overline{sk}_N, i, c_i^j)$.
Define the query $q^j(i', sk')$ to be $Dec(sk', ct_{i'}^j)$ if $i' \notin T^j$ and 0 otherwise.
Let $a^j = \mathcal{M}(x; q^j)$ and round a^j to $\{\pm 1\}$ to obtain $\overline{a}^j$.
Give $\overline{a}^j$ to $\mathcal{F}$ and let $I^j \subseteq [N]$ be the set of accused users and $T^j = T^{j-1} \cup I^j$.

Output 1 if and only if the event Z_c occurs

Figure 6.6: $\mathcal{B}_{c,n,d}^{\mathcal{E}_b(\overline{sk}_1, \ldots, \overline{sk}_N, \cdot)}$

returns $Enc(\overline{sk}_i, 0)$), then the view of the mechanism is identical to $\mathsf{Attack}_{n,d}[\mathcal{M}]$. Thus we have that for $c \in \{1, 2\}$,

$$\left| \mathop{\mathbb{P}}_{\mathsf{IdealAttack}_{n,d}[\mathcal{M}]}[Z_c] - \mathop{\mathbb{P}}_{\mathsf{Attack}_{n,d}[\mathcal{M}]}[Z_c] \right|$$
$$= \left| \mathop{\mathbb{P}}_{\overline{sk}_1, \ldots, \overline{sk}_N \leftarrow_R Gen(1^\lambda)} \left[\mathcal{B}_{c,n,d}^{\mathcal{E}_0(\overline{sk}_1, \ldots, \overline{sk}_N, \cdot)} = 1 \right] - \mathop{\mathbb{P}}_{\overline{sk}_1, \ldots, \overline{sk}_N \leftarrow_R Gen(1^\lambda)} \left[\mathcal{B}_{c,n,d}^{\mathcal{E}_1(\overline{sk}_1, \ldots, \overline{sk}_N, \cdot)} = 1 \right] \right|$$
$$= \mathrm{negl}(\lambda) = \mathrm{negl}(d)$$

The last equality holds because we have chosen $N = 2000n(d) = \mathrm{poly}(d)$, and therefore we have $\lambda = d - \lceil \log N \rceil = d - O(\log d)$. This completes the proof of both claims. □

Chapter 7

The Power of Adaptivity in Differential Privacy

7.1 Introduction

A central question in the differential privacy is literature is: how much error do we need to answer a large set of queries $q_1, \ldots, q_k$? Before we can answer this question, we have to define a model of how the queries are asked and answered. The literature on differential privacy has considered three different interactive models[1] for specifying the queries:

- *The Offline Model:* The sequence of queries $q_1, \ldots, q_k$ are given to the algorithm together in a batch and the mechanism answers them together.
- *The Online Model:* The sequence of queries $q_1, \ldots, q_k$ is chosen in advance and then the mechanism must answer each query q_j before seeing q_{j+1}.

[1]Usually, the "interactive model" refers only to what we call the "adaptive model." We prefer to call all of these models interactive, since they each require an interaction with a data analyst who issues the queries. We use the term "interactive" to distinguish these models from one where the algorithm only answers a fixed set of queries.

- *The Adaptive Model:* The queries are not fixed in advance, each query q_{j+1} may depend on the answers to queries $q_1, \ldots, q_j$.

In all three cases, we assume that $q_1, \cdots, q_k$ are chosen from some family of allowable queries Q, but may be chosen adversarially from this family.

Differential privacy seems well-suited to the adaptive model. Arguably its signature property is that any adaptively-chosen sequence of differentially private algorithms remains collectively differentially private, with a graceful degradation of the privacy parameters [DMNS06, DRV10]. As a consequence, there is a simple differentially private algorithm that takes a dataset of n individuals and answers $\tilde{\Omega}(n)$ statistical queries in the adaptive model with error $o(1/\sqrt{n})$, simply by perturbing each answer independently with carefully calibrated noise. In contrast, the seminal lower bound of Dinur and Nissim and its later refinements [DN03, DY08] shows that there exists a fixed set of $O(n)$ queries that cannot be answered by any differentially private algorithm with such little error, even in the easiest offline model. For an even more surprising example, the private multiplicative weights algorithm of Hardt and Rothblum [HR10] can in many cases answer an exponential number of arbitrary, adaptively-chosen statistical queries with a strong accuracy guarantee, whereas [BUV14] show that the accuracy guarantee of private multiplicative weights is nearly optimal even for a simple, fixed family of queries.

These examples might give the impression that answering adaptively-chosen queries comes "for free" in differential privacy—that everything that can be achieved in the offline model can be matched in the adaptive model. Beyond just the lack of any separation between the models, many of the most powerful differentially private algorithms in all of these models use techniques from no-regret learning, which are explicitly designed for adaptive models.

Another motivation for studying the relationship between these models is the recent line of work connecting differential privacy to statistical validity for *adaptive data analysis* [HU14, DFH+15c, §3,§6], which shows that differentially private algorithms for adaptively-chosen queries in fact yield state-of-the-art algorithms for statistical problems unrelated to privacy. This connection further motivates studying the adaptive model and its relationship to the other models in differential privacy.

In this chapter, we show that these three models are actually distinct. In fact, we show exponential separations between each of the three models. These are the first separations between these models in differential privacy.

7.1.1 Our Results

Given a dataset x whose elements come from a data universe $\mathcal{X}$, a *statistical query on* $\mathcal{X}$ is defined by a predicate ϕ on $\mathcal{X}$ and asks "what fraction of elements in the dataset satisfy ϕ?" The answer to a statistical query lies in $[0,1]$ and our goal is to answer these queries up to some small additive error $\pm\alpha$, for a suitable choice of $0 < \alpha < 1$. If the mechanism is required to answer *arbitrary* statistical queries, then the offline, online, and adaptive models are essentially equivalent — the upper bounds in the adaptive model match the lower bounds in the offline model [DRV10, HR10, BUV14, §4]. However, we show that when the predicate ϕ is required to take a specific form, then it becomes strictly easier to answer a set of these queries in the offline model than it is to answer a sequence of queries presented online.

Theorem 7.1.1 (Informal)**.** *For every $n \in \mathbb{N}$, there exists a data universe $\mathcal{X}$ and a family of statistical queries Q on $\mathcal{X}$ such that,*

1. *there is a differentially private algorithm that takes a dataset $x \in \mathcal{X}^n$ and answers any set of $k = 2^{\Omega(\sqrt{n})}$ offline queries from Q up to error $\pm 1/100$ from Q, but*

2. *no differentially private algorithm can take a dataset $x \in \mathcal{X}^n$ and answer an arbitrary sequence of $k = O(n^2)$ online (but not adaptively-chosen) queries from Q up to error $\pm 1/100$.*

The constant $1/100$ is arbitrary — the negative results hold for larger values and the positive results hold for smaller values. The accuracy parameter does not play an important role in any of our results and should mostly be ignored.

This result establishes that the online model is strictly harder than the offline model. We also demonstrate that the adaptive model is strictly harder than the online model. Here, the family of queries we use in our separation is not a family of statistical queries, but is rather a family of *search queries* with a specific definition of accuracy that we will define later.

Theorem 7.1.2 (Informal)**.** *For every $n \in \mathbb{N}$, there is a family of "search" queries Q on datasets in $\mathcal{X}^n$ such that*

1. *there is a differentially private algorithm that takes a dataset $x \in \{\pm 1\}^n$ and accurately answers any online (but not adaptively-chosen) sequence of $k = 2^{\Omega(n)}$ queries from Q, but*

2. *no differentially private algorithm can take a dataset $x \in \{\pm 1\}^n$ and accurately answer an adaptively-chosen sequence of $k = O(1)$ queries from Q.*

We leave it as an interesting open question to separate the online and adaptive models for statistical queries, or to show that the models are equivalent for statistical queries.

Although Theorems 7.1.1 and 7.1.2 separate the three models, these results use somewhat contrived families of queries. Thus, we also investigate whether the models are distinct for *natural* families of queries that are of use in practical

applications. One very well studied class of queries is *threshold queries*. These are a family of statistical queries Q_{thresh} defined on the universe $[0,1]$ and each query is specified by a point $\tau \in [0,1]$ and asks "what fraction of the elements of the dataset are at most τ?" If we restrict our attention to so-called pure differential privacy (i.e. (ε, δ)-differential privacy with $\delta = 0$), then we obtain an exponential separation between the offline and online models for answering threshold queries.

Theorem 7.1.3 (Informal)**.** *For every $n \in \mathbb{N}$,*

1. *there is a pure differentially private algorithm that takes a dataset $x \in [0,1]^n$ and answers any set of $k = 2^{\Omega(n)}$ offline queries from Q_{thresh} up to error $\pm 1/100$, but*
2. *no pure differentially private algorithm takes a dataset $x \in [0,1]^n$ and answers an arbitrary sequence of $k = O(n)$ online (but not adaptively-chosen) queries from Q_{thresh} up to error $\pm 1/100$.*

We also ask whether or not such a separation exists for arbitrary differentially private algorithms (i.e. (ε, δ)-differential privacy with $\delta > 0$). Theorem 7.1.3 shows that, for pure differential privacy, online threshold queries have near-maximal sample complexity. That is, up to constants, the lower bound for online threshold queries matches what is achieved by the Laplace mechanism (cf. Theorem 4.3.1), which is applicable to arbitrary statistical queries. This may lead one to conjecture that adaptive threshold queries also require near-maximal sample complexity subject to approximate differential privacy. However, we show that this is not the case:

Theorem 7.1.4. *For every $n \in \mathbb{N}$, there is a differentially private algorithm that takes a dataset $x \in [0,1]^n$ and answers any set of $k = 2^{\Omega(n)}$ adaptively-chosen queries from Q_{thresh} up to error $\pm 1/100$.*

In contrast, for any offline set of k thresholds $\tau_1, \ldots, \tau_k$, we can round each element of the dataset up to an element in the finite universe $\mathcal{X} = \{\tau_1, \ldots, \tau_k, 1\}$

without changing the answers to any of the queries. Then we can use known algorithms for answering all threshold queries over any finite, totally ordered domain [BNS13, BNSV15] to answer the queries using a very small dataset of size $n = 2^{O(\log^*(k))}$. That is, we can answer a number of queries that is an exponential tower of height $\Omega(\log n)$, which is much more than exponential in n. We leave it as an interesting open question to settle the complexity of answering adaptively-chosen threshold queries in the adaptive model.

7.1.2 Techniques

Separating Offline and Online Queries

To prove Theorem 7.1.1, we construct a sequence of queries $q_1, \cdots, q_k$ such that, for all $j \in [k]$,

- q_j "reveals" the answers to $q_1, \cdots, q_{j-1}$, but
- $q_1, \cdots, q_{j-1}$ do not reveal the answer to q_j.

Thus, given the sequence $q_1, \cdots, q_k$ in the offline setting, the answers to $q_1, \cdots, q_{k-1}$ are revealed by q_k. So only q_k needs to be answered and the remaining query answers can be inferred. However, in the online setting, each query q_{j-1} must be answered before q_j is presented and this approach does not work. This is the intuition for our separation.

To prove the online lower bound, we build on the a lower bound for one-way marginal queries from Chapter 4 — unless $k \ll n^2$, there is no differentially private algorithm that answers k one-way marginal queries with constant accuracy. We are able to "embed" k marginal queries into the sequence of online queries $q_1, \cdots, q_k$. Thus a modification of the lower bound for one-way marginal queries applies in the online setting.

To prove the offline upper bound, we use the fact that every query reveals information about other queries. However, we must handle arbitrary sequences of queries, not just the specially-constructed sequences used for the lower bound. The key property of our family of queries is the following. Each element x of the data universe X requires k bits to specify. On the other hand, for any set of queries $q_1, \cdots, q_k$, we can specify $q_1(x), \cdots, q_k(x)$ using only $O(\log(nk))$ bits. Thus the effective size of the data universe given the queries is $\text{poly}(nk)$, rather than 2^k. Then we can apply a differentially private algorithm that gives good accuracy as long as the data universe has subexponential size [BLR13]. Reducing the size of the data universe is only possible once the queries have been specified; hence this approach only works in the offline setting.

Separating Online and Adaptive Queries

To prove Theorem 7.1.2, we start with the classical randomized response algorithm [War65]. Specifically, given a dataset $x \in \{\pm 1\}^n$, randomized response produces a new dataset $y \in \{\pm 1\}^n$ where each coordinate y_i is independently set to $+x_i$ with probability $(1+\alpha)/2$ and is set to $-x_i$ with probability $(1-\alpha)/2$. It is easy to prove that this algorithm is $(O(\alpha), 0)$-differentially private. What accuracy guarantee does this algorithm satisfy? By design, it outputs a vector y that has correlation approximately α with the dataset x — that is, $\langle y, x \rangle \approx \alpha n$. On the other hand, it is also easy to prove that there is no differentially private algorithm (for any reasonable privacy parameters) that can output a vector that has correlation at least $1/2$ with the sensitive dataset.

Our separation between the online and adaptive models is based on the observation that, if we can obtain $O(1/\alpha^2)$ "independent" vectors $y_1, \ldots, y_k$ that are each roughly α-correlated with x, then we can obtain a vector z that is $(1/2)$-correlated

with x, simply by letting z be the coordinate-wise majority of the y_js. Thus, no differentially private algorithm can output such a set of vectors. More precisely, we require that $\langle y_i, y_j \rangle \approx \alpha^2 n$ for $i \neq j$, which is achieved if each y_j is an independent sample from randomized response.

Based on this observation, we devise a class of queries such that, if we are allowed to choose k of these queries adaptively, then we obtain a set of vectors $y_1, \ldots, y_k$ satisfying the conditions above. This rules out differential privacy for $k = O(1/\alpha^2)$ adaptive queries. The key is that we can use adaptivity to ensure that each query asks for an "independent" y_j by adding the previous answers $y_1, \cdots, y_{j-1}$ as constraints in the search query.

On the other hand, randomized response can answer each such query with high probability. If a number of these queries is fixed in advance, then, by a union bound, the vector y output by randomized response is simultaneously an accurate answer to any collection of $2^{\Omega(n)}$ queries with high probability. Since randomized response is oblivious to the queries, we can also answer the queries in the online model, as long as they are not chosen adaptively.

At a high level, the queries that achieve this property are of the form "output a vector $y \in \{\pm 1\}^n$ that is approximately α-correlated with x and is approximately as uncorrelated as possible with the vectors $v_1, \ldots, v_m$." A standard concentration argument shows that randomized response gives an accurate answer to all the queries simultaneously with high probability. On the other hand, if we are allowed to choose the queries adaptively, then for each query q_i, we can ask for a vector y_i that is correlated with x but is as uncorrelated as possible with the previous answers $y_1, \ldots, y_{i-1}$.

Threshold Queries

For pure differential privacy, our separation between offline and online threshold queries uses a simple argument based on binary search. Our starting point is a lower bound showing that any purely differentially private algorithm that takes a dataset of n points $x_1, \ldots, x_n \in \{1, \ldots, T\}$ and outputs an *approximate median* of these points requires $n = \Omega(\log T)$. This lower bound follows from a standard application of the "packing" technique of Hardt and Talwar [HT10]. On the other hand, by using binary search, any algorithm that can answer $k = O(\log T)$ adaptively-chosen threshold queries can be used to find an approximate median. Thus, any purely differentially private algorithm for answering such queries requires a dataset of size $n = \Omega(k)$. Using the structure of the lower bound argument, we show that the same lower bound holds for online non-adaptive queries as well. In contrast, using the algorithms of [DNPR10, CSS11, DNRR15], we can answer k offline threshold queries on a dataset with only $n = O(\log k)$ elements, giving an exponential separation.

The basis of our improved algorithm for adaptive threshold queries under approximate differential privacy is a generalisation of the *sparse vector* technique [DNPR10, RR10, HR10] (see [DR14, §3.6] for a textbook treatment). Our algorithm makes crucial use of a *stability argument* similar to the propose-test-release techniques of Dwork and Lei [DL09]. To our knowledge, this is the first use of a stability argument for any online or adaptive problem in differential privacy and may be of independent interest. In particular, our algorithm is given an input $x \in X^n$, a threshold $t \in (0, 1)$, and an adaptive sequence of statistical (or low-sensitivity) queries $q_1, \cdots, q_k : X^n \to [0, 1]$ and, for each query q_j, it reports (i) $q_j(x) \geq t$, (ii) $q_j(x) \leq t$, or (iii) $t - \alpha \leq q_j(x) \leq t + \alpha$. The sample complexity of this algorithm is $n = O(\sqrt{c} \log(k/\varepsilon\delta)/\varepsilon\alpha)$, where k is the total number of queries, c is an upper bound on the number of times (iii) may be reported, and (ε, δ)-differential privacy

is provided. We call this the *Between Thresholds algorithm*.

Once we have this algorithm, we can use it to answer adaptively-chosen thresholds using an approach inspired by Bun et al. [BNSV15]. The high-level ideal is to sort the dataset $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ and then partition it into chunks of consecutive sorted elements. For any chunk, and a threshold τ, we can use the between thresholds algorithm to determine (approximately) whether τ lies below all elements in the chunk, above all elements in the chunk, or inside the chunk. Obtaining this information for every chunk is enough to accurately estimate the answer to the threshold query τ up to an error proportional to the size of the chunks. The sample complexity is dominated by the $O(\log k)$ sample complexity of our Between Thresholds algorithm multiplied by the number of chunks needed, namely $O(1/\alpha)$.

7.2 Preliminaries

7.2.1 Models of Interactive Queries

The goal of this work is to understand the implications of different ways to allow an adversary to query a sensitive dataset. In each of these models there is an algorithm $\mathcal{W}$ that holds a dataset $x \in \mathcal{X}^n$, and a fixed family of (statistical or search) queries Q on $\mathcal{X}^n$, and a bound k on the number of queries that $\mathcal{W}$ has to answer. There is also an adversary $\mathcal{A}$ that chooses the queries. The models differ in how the queries chosen by $\mathcal{A}$ are given to $\mathcal{W}$.

Offline

In the *offline* model, the queries $q_1, \ldots, q_k \in Q$ are specified by the adversary $\mathcal{A}$ in advance and the algorithm $\mathcal{W}$ is given all the queries at once and must provide answers. Formally, we define the following function $\mathsf{Offline}_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{W}} : \mathcal{X}^n \to Q^k \times \mathcal{Y}^k$ depending $\mathcal{A}$ and $\mathcal{W}$.

Input: $x \in X^n$.
$\mathcal{A}$ chooses $q_1, \cdots, q_k \in Q$.
$\mathcal{W}$ is given x and $q_1, \cdots, q_k$ and outputs $a_1, \cdots, a_k \in \mathcal{Y}$.
Output: $(q_1, \cdots, q_k, a_1, \cdots, a_k) \in Q^k \times \mathcal{Y}^k$.

Figure 7.1: $\mathsf{Offline}_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{W}} : \mathcal{X}^n \to Q^k \times \mathcal{Y}^k$

Online Non-Adaptive

In the *online non-adaptive* model, the queries $q_1, \ldots, q_k \in Q$ are again fixed in advance by the adversary, but are then given to the algorithm one at a time, and the algorithm must give an answer to query q_j before it is shown q_{j+1}. We define a function $\mathsf{Online}_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{W}} : \mathcal{X}^n \to Q^k \times \mathcal{Y}^k$ depending on the adversary $\mathcal{A}$ and the algorithm $\mathcal{W}$ as follows.

Input: $x \in X^n$.
$\mathcal{A}$ chooses $q_1, \cdots, q_k \in Q$.
$\mathcal{W}$ is given x.
For $j = 1, \ldots, k$:
  $\mathcal{W}$ is given q_j and outputs $a_j \in \mathcal{Y}$.[2]
Output: $(q_1, \cdots, q_k, a_1, \cdots, a_k) \in Q^k \times \mathcal{Y}^k$.

Figure 7.2: $\mathsf{Online}_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{W}} : \mathcal{X}^n \to Q^k \times \mathcal{Y}^k$

Online Adaptive

In the *online adaptive* model, the queries $q_1, \ldots, q_k \in Q$ are not fixed, and the adversary may choose each q_j based on the answers that the algorithm gave to the previous queries. We define a function $\mathsf{Adaptive}_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{W}} : \mathcal{X}^n \to Q^k \times \mathcal{Y}^k$ depending on the adversary $\mathcal{A}$ and the algorithm $\mathcal{W}$ as follows.

Input: $x \in X^n$.
$\mathcal{W}$ is given x.
For $j = 1, \ldots, k$:
  $\mathcal{A}$ chooses a query $q_j \in Q$.
  $\mathcal{W}$ is given q_j and outputs $a_j \in \mathcal{Y}$.
Output: $(q_1, \cdots, q_k, a_1, \cdots, a_k) \in Q^k \times \mathcal{Y}^k$.

Figure 7.3: $\mathsf{Adaptive}_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{W}} : \mathcal{X}^n \to Q^k \times \mathcal{Y}^k$

Definition 7.2.1 (Differential Privacy for Interactive Mechanisms)**.** *In each of the three cases — Offline, Online Non-Adaptive, or Online Adaptive — we say that $\mathcal{W}$ is (ε, δ)-differentially private if, for all adversaries $\mathcal{A}$, respectively* $\mathsf{Offline}_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{W}}$, $\mathsf{Online}_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{W}}$, *or* $\mathsf{Adaptive}_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{W}}$ *is (ε, δ)-differentially private.*

Definition 7.2.2 (Accuracy for Interactive Mechanisms)**.** *In each case — Offline, Online Non-Adaptive, or Online Adaptive queries — we say that $\mathcal{W}$ is (α, β)-accurate if, for all adversaries $\mathcal{A}$ and all inputs $x \in \mathcal{X}^n$,*

$$\mathop{\mathbb{P}}_{q_1, \cdots, q_k, a_1, \cdots, a_k} \left[\max_{j \in [k]} L_{q_j}(x, a_j) \leq \alpha \right] \geq 1 - \beta, \tag{7.1}$$

where $(q_1, \cdots, q_k, a_1, \cdots, a_k)$ is respectively drawn from one of $\mathsf{Offline}_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{W}}(x)$, $\mathsf{Online}_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{W}}(x)$, *or* $\mathsf{Adaptive}_{\mathcal{A} \overset{\rightarrow}{\leftarrow} \mathcal{W}}(x)$. *We also say that $\mathcal{W}$ is α-accurate if the above holds with (7.1) replaced by*

$$\mathop{\mathbb{E}}_{q_1, \cdots, q_k, a_1, \cdots, a_k} \left[\max_{j \in [k]} L_{q_j}(x, a_j) \right] \leq \alpha.$$

7.2.2 Search Queries

In this work we consider two general classes of queries on the dataset: statistical queries (which have been described earlier) and *search queries*.

Formally, a *search query* q on $\mathcal{X}^n$ is defined by a *loss function* $L_q : \mathcal{X}^n \times \mathcal{Y} \to [0, \infty)$, where $\mathcal{Y}$ is an arbitrary set representing the range of possible outputs. For a dataset $x \in \mathcal{X}^n$ and an output $y \in \mathcal{Y}$, we will say that y is *α-accurate for q on x* if $L_q(x, y) \leq \alpha$. In some cases the value of L_q will always be either 0 or 1. Thus we simply say that y is *accurate for q on x* if $L_q(x, y) = 0$. For example, if $\mathcal{X}^n = \{\pm 1\}^n$, we can define a search query by $\mathcal{Y} = \{\pm 1\}^n$, and $L_q(x, y) = 0$ if $\langle x, y \rangle \geq \alpha n$ and $L_q(x, y) = 1$ otherwise. In this case, the search query would ask for any vector y that has correlation α with the dataset.

Statistical queries are a special case of search queries: given a statistical query q on $\mathcal{X}^n$, we can define a search query L_q with $\mathcal{Y} = [0, 1]$ and $L_q(x, a) = |q(x) - a|$. Then both definitions of α-accurate align.

7.3 A Separation Between Offline and Online Queries

In this section we prove that online accuracy is strictly harder to achieve than offline accuracy, even for statistical queries. We prove our results by constructing a set of statistical queries that we call *prefix queries* for which it is possible to take a dataset of size n and accurately answer superpolynomially many offline prefix queries in a differentially private manner, but it is impossible to answer more than $O(n^2)$ online prefix queries while satisfying differential privacy.

We now define the family of prefix queries. These queries are defined on the

universe $X = \{\pm 1\}^* = \bigcup_{j=0}^{\infty} \{\pm 1\}^j$ consisting of all finite length binary strings.[3] For $x, y \in \{\pm 1\}^*$, we use $y \preceq x$ to denote that y is a *prefix* of x. Formally

$$y \preceq x \qquad \iff \qquad |y| \leq |x| \text{ and } \forall i = 1, \ldots, |y| \; x_i = y_i.$$

Definition 7.3.1. *For any finite set $S \subseteq \{\pm 1\}^*$ of finite-length binary strings, we define the* prefix query $q_S : \{\pm 1\}^* \to \{\pm 1\}$ *by*

$$q_S(x) = 1 \quad \iff \quad \exists y \in S \quad y \preceq x.$$

We also define

$$Q_{prefix} = \{q_S \mid S \subset \{\pm 1\}^*\}$$
$$Q^B_{prefix} = \{q_S \mid S \subset \{\pm 1\}^*, |S| \leq B\}$$

to be the set of all prefix queries and the set of prefix queries with sizes bounded by B, respectively.

7.3.1 Answering Offline Prefix Queries

We now prove that there is a differentially private algorithm that answers super-polynomially many prefix queries, provided that the queries are specified offline.

Theorem 7.3.2 (Answering Offline Prefix Queries)**.** *For every $\alpha, \varepsilon \in (0, 1/10)$, every $B \in \mathbb{N}$, and every $n \in \mathbb{N}$, there exists a*

$$k = \min\left\{2^{\Omega(\sqrt{\alpha^3 \varepsilon n})}, 2^{\Omega(\alpha^3 \varepsilon n / \log(B))}\right\}$$

and an $(\varepsilon, 0)$-differentially private algorithm $\mathcal{W}_{\mathsf{prefix}} : \mathcal{X}^n \times (Q^B_{prefix})^k \to \mathbb{R}^k$ that is

[3] All of the arguments in this section hold if we restrict to strings of length at most $k + \log n$. However, we allow strings of arbitrary length to reduce notational clutter.

$(\alpha, 1/100)$-*accurate for* k *offline queries from* Q^B_{prefix}.

We remark that it is possible to answer even more offline prefix queries by relaxing to (ε, δ)-differential privacy for some negligibly small $\delta > 0$. However, we chose to state the results for $(\varepsilon, 0)$-differential privacy to emphasize the contrast with the lower bound, which applies even when $\delta > 0$, and to simplify the statement.

Our algorithm for answering offline queries relies on the existence of a good differentially private algorithm for answering *arbitrary* offline statistical queries. For concreteness, the so-called "BLR mechanism" of Blum, Ligett, and Roth [BLR13] suffices, although different parameter tradeoffs can be obtained using different mechanisms. Differentially private algorithms with this type of guarantee exist only when the data universe is bounded, which is not the case for prefix queries. However, as we show, when the queries are specified offline, we can replace the infinite universe $\mathcal{X} = \{\pm 1\}^*$ with a finite, restricted universe $\mathcal{X}'$ and run the BLR mechanism. Looking ahead, the key to our separation will be the fact that this universe restriction is only possible in the offline setting. Before we proceed with the proof of Theorem 7.3.2, we will state the guarantees of the BLR mechanism.

Theorem 7.3.3 ([BLR13])**.** *For every* $0 < \alpha, \varepsilon \leq 1/10$ *and every finite data universe* $\mathcal{X}$*, if* $\mathcal{Q}_{\mathsf{SQ}}$ *is the set of all statistical queries on* $\mathcal{X}$*, then for every* $n \in \mathbb{N}$*, there is a*

$$k = 2^{\Omega(\alpha^3 \varepsilon n / \log |\mathcal{X}|)}$$

and an $(\varepsilon, 0)$*-differentially private algorithm* $\mathcal{W}_{\mathsf{BLR}} : \mathcal{X}^n \times \mathcal{Q}^k_{\mathsf{SQ}} \to \mathbb{R}^k$ *that is* $(\alpha, 1/100)$*-accurate for* k *offline queries from* $\mathcal{Q}_{\mathsf{SQ}}$.

We are now ready to prove Theorem 7.3.2.

Proof of Theorem 7.3.2. Suppose we are given a set of queries $q_{S_1}, \ldots, q_{S_k} \in Q^B_{\text{prefix}}$ and a dataset $x \in \mathcal{X}^n$ where $\mathcal{X} = \{\pm 1\}^*$. Let $S = \bigcup_{j=1}^k S_j$. We define the universe

$\mathcal{X}_S = S \cup \{\emptyset\}$ where $\emptyset$ denotes the empty string of length 0. Note that this universe depends on the choice of queries, and that $|\mathcal{X}_S| \leq kB + 1$. Since $\mathcal{X}_S \subset \mathcal{X}$, it will be well defined to restrict the domain of each query q_{S_j} to elements of $\mathcal{X}_S$.

Next, given a dataset $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, and a collection of sets $S_1, \ldots, S_k \subset \mathcal{X}$, we give a procedure for mapping each element of x to an element of $\mathcal{X}_S$ to obtain a new dataset $x^S = (x_1^S, \ldots, x_n^S) \in \mathcal{X}_S^n$ that is equivalent to x with respect to the queries $q_{S_1}, \ldots, q_{S_k}$. Specifically, define $r_S : \mathcal{X} \to \mathcal{X}_S$ by

$$r_S(x) = \text{argmax}_{y \in \mathcal{X}_S, y \preceq x} |y|.$$

That is, $r_S(x)$ is the longest string in $\mathcal{X}_S$ that is a prefix of x. We summarize the key property of r_S in the following claim

Claim 7.3.4. *For every $x \in \mathcal{X}$, and $j = 1, \ldots, k$, $q_{S_j}(r_S(x)) = q_{S_j}(x)$.*

Proof of Claim 7.3.4. First, we state a simple but important fact about prefixes: If y, y' are both prefixes of a string x with $|y| \leq |y'|$, then y is a prefix of y'. Formally,

$$\forall x, y, y' \in \{0,1\}^* \qquad (y \preceq x \ \wedge \ y' \preceq x \ \wedge \ |y| \leq |y'|) \implies y \preceq y'. \tag{7.2}$$

Now, fix any $x \in \mathcal{X}$ and any query q_{S_j} and suppose that $q_{S_j}(x) = 1$. Then there exists a string $y \in S_j$ such that $y \preceq x$. By construction, we have that $r_S(x) \preceq x$ and that $|r_S(x)| \geq |y|$. Thus, by (7.2), we have that $y \preceq r_S(x)$. Thus, there exists $y \in S_j$ such that $y \preceq r_S(x)$, which means $q_{S_j}(r_S(x)) = 1$, as required.

Next, suppose that $q_{S_j}(r_S(x)) = 1$. Then, there exists $y \in S_j$ such that $y \preceq r_S(x)$. By construction, $r_S(x) \preceq x$, so by transitivity we have that $y \preceq x$. Therefore, $q_{S_j}(x) = 1$, as required. □

Given this lemma, we can replace every row x_i of x with $x_i^S = r_S(x_i)$ to obtain a

new dataset x^S such that for every $j = 1, \ldots, k$,

$$q_{S_j}(x^S) = \frac{1}{n}\sum_{i=1}^{n} q_{S_j}(x_i^S) = \frac{1}{n}\sum_{i=1}^{n} q_{S_j}(x_i) = q_{S_j}(x).$$

Thus, we can answer $q_{S_1}, \cdots, q_{S_k}$ on $x^S \in \mathcal{X}_S^n$, rather than on $x \in \mathcal{X}^n$. Note that each row of x^S depends only on the corresponding row of x. Hence, for every set of queries $q_{S_1}, \ldots, q_{S_k}$, if $x \sim x'$ are adjacent datasets, then $x^S \sim x'^S$ are also adjacent datasets. Consequently, applying a (ε, δ)-differentially private algorithm to x^S yields a (ε, δ)-differentially private algorithm as a function of x.

In particular, we can give α-accurate answers to these queries using the algorithm $\mathcal{W}_{\mathsf{BLR}}$ as long as

$$k \leq 2^{\Omega(\alpha^3 \varepsilon n / \log |\mathcal{X}_S|)} = 2^{\Omega(\alpha^3 \varepsilon n / \log(kB+1))}.$$

Rearranging terms gives the bound in Theorem 7.3.2. We specify the complete algorithm $\mathcal{W}_{\mathsf{prefix}}$ in Figure 7.4.

$\mathcal{W}_{\mathsf{prefix}}(x; q_{S_1}, \ldots, q_{S_k})$:
Write $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, $S = \bigcup_{j=1}^{k} S_j$, $\mathcal{X}_S = S \cup \{\emptyset\}$.
For $i = 1, \ldots, n$, let $x_i^S = r_S(x_i)$ and let $x^S = (x_1^S, \ldots, x_n^S) \in \mathcal{X}_S^n$.
Let $(a_1, \ldots, a_k) = \mathcal{W}_{\mathsf{BLR}}(x^S; q_{S_1}, \ldots, q_{S_k})$.
Output $(a_1, \ldots, a_k)$.

Figure 7.4: $\mathcal{W}_{\mathsf{prefix}}$

□

7.3.2 A Lower Bound for Online Prefix Queries

Next, we prove a lower bound for online queries. Our lower bound shows that the simple approach of perturbing the answer to each query with independent noise is essentially optimal for prefix queries. Since this approach is only able to answer

$k = O(n^2)$ queries, we obtain an exponential separation between online and offline statistical queries for a broad range of parameters.

Theorem 7.3.5 (Lower Bound for Online Prefix Queries)**.** *There exists a function $k = O(n^2)$ such that for every sufficiently large $n \in \mathbb{N}$, there is no $(1, 1/30n)$-differentially private algorithm $\mathcal{W}$ that takes a dataset $x \in \mathcal{X}^n$ and is $(1/100, 1/100)$-accurate for k online queries from Q^n_{prefix}.*

In this parameter regime, our algorithm from Section 7.3.1 answers $k = e^{\tilde{\Omega}(\sqrt{n})}$ offline prefix queries, so we obtain an exponential separation.

Our lower bound is essentially the same as the lower bound for one-way marginals in Chapter 4. However, we cannot apply those bounds in a black-box manner, as we must show that the lower bound works adaptively. Recall the following key lemma from Section 4.3.3.

Lemma 7.3.6 (Fingerprinting Lemma)**.** *Let $f : \{\pm 1\}^n \to [-1, 1]$ be any function. Suppose p is sampled from the uniform distribution over $[-1, 1]$ and $c \in \{\pm 1\}^n$ is a vector of n independent bits, where each bit has expectation p. Letting $\bar{c}$ denote the coordinate-wise mean of c, we have*

$$\mathop{\mathbb{E}}_{p,c}\left[f(c) \cdot \sum_{i \in [n]} (c_i - p) + 2\left|f(c) - \bar{c}\,\right|\right] \geq \frac{1}{3}.$$

Proof of Theorem 7.3.5. First we define the distribution on the input dataset $x = (x_1, \ldots, x_n)$ and the queries $q_{S_1}, \cdots, q_{S_k}$.

Input dataset x:

- Sample $p^1, \cdots, p^k \in [-1, 1]$ independently and uniformly at random.
- Sample $c^1, \cdots, c^k \in \{\pm 1\}^n$ independently, where each c^j is a vector of n independent bits, each with expectation p^j.

- For $i \in [n]$, define

$$x_i = (\text{binary}(i), c_i^1, \cdots, c_i^k) \in \{\pm 1\}^{\lceil \log_2 n \rceil + k},$$

where $\text{binary}(i) \in \{\pm 1\}^{\lceil \log_2 n \rceil}$ is the binary representation of i where 1 is mapped to $+1$ and 0 is mapped to -1.[4] Let

$$x = (x_1, \ldots, x_n) \in \left(\{\pm 1\}^{\lceil \log_2 n \rceil + k}\right)^n.$$

Queries $q_{S_1}, \cdots, q_{S_k}$:

- For $i \in [n]$ and $j \in [k]$, define

$$z_{i,j} = (\text{binary}(i), c_i^1, \cdots, c_i^{j-1}, 1) \in \{\pm 1\}^{\lceil \log_2 n \rceil + j}.$$

- For $j \in [k]$, define $q_{S_j} \in Q^n_{\text{prefix}}$ by $S_j = \{z_{i,j} \mid i \in [n]\}$.

These queries are designed so that the correct answer to each query $j \in [k]$ is given by $q_{S_j}(x) = \overline{c}^j$:

Claim 7.3.7. *For every $j \in [k]$, if the dataset x and the queries $q_{S_1}, \ldots, q_{S_k}$ are constructed as above, then with probability 1,*

$$q_{S_j}(x) = \frac{1}{n} \sum_{i=1}^{n} q_{S_j}(x_i) = \frac{1}{n} \sum_{i=1}^{n} c_i^j = \overline{c}^j$$

Proof of Claim 7.3.7. We have

$$q_{S_j}(x_i) = 1 \iff \exists w \in S_j \ (w \preceq x_i) \iff \exists \ell \in [n] \ (z_{\ell,j} \preceq x_i).$$

By construction, we have $z_{\ell,j} \preceq x_i$ if and only if $\ell = i$ and $x_i^j = c_i^j = 1$, as required.

[4] This choice is arbitrary, and is immaterial to our lower bound. The only property we need is that $\text{binary}(i)$ uniquely identifies i and, for notational consistency, we require $\text{binary}(i)$ to be a string over the alphabet $\{\pm 1\}$.

Here, we have used the fact that the strings $\text{binary}(i)$ are unique to ensure that $z_{\ell,j} \preceq x_i$ if and only if $\ell = i$. □

We now show no differentially private algorithm $\mathcal{W}$ is capable of giving accurate answers to these queries. Let $\mathcal{W}$ be an algorithm that answers k online queries from Q^n_{prefix}. Suppose we generate an input dataset x and queries $q_{S_1}, \ldots, q_{S_k}$ as above, and run $\mathcal{W}(x)$ on this sequence of queries. Let $a^1, \ldots, a^k \in [-1, 1]$ denote the answers given by $\mathcal{W}$.

First, we claim that, if $\mathcal{W}(x)$ is accurate for the given queries, then each answer a^j is close to the corresponding value $\bar{c}^j = \frac{1}{n}\sum_{i=1}^n c_i^j$.

Claim 7.3.8. *If $\mathcal{W}$ is $(1/100, 1/100)$-accurate for k online queries from Q^n_{prefix}, then with probability 1 over the choice of x and $q_{S_1}, \ldots, q_{S_k}$ above,*

$$\mathop{\mathbb{E}}_{\mathcal{W}}\left[\sum_{j\in[k]} \left|a^j - \bar{c}^j\right|\right] \leq \frac{k}{10}.$$

Proof of Claim 7.3.8. By Claim 7.3.7, for every $j \in [k]$, $q_{S_j}(x) = \bar{c}^j$. Since, by assumption, $\mathcal{W}$ is $(1/100, 1/100)$-accurate for k online queries from Q^n_{prefix}, we have that with probability at least $99/100$,

$$\forall j \in [k] \quad \left|a^j - q_{S_j}(x)\right| \leq \frac{1}{100} \quad \Longrightarrow \quad \forall j \in [k] \quad \left|a^j - \bar{c}^j\right| \leq \frac{1}{100}$$

By linearity of expectation, this case contributes at most $k/100$ to the expectation. On the other hand, $|a^j - q_{S_j}(x)| \leq 2$, so by linearity of expectation the case where $\mathcal{W}$ is inaccurate contributes at most $2k/100$ to the expectation. This suffices to prove the claim. □

The next claim shows how the fingerprinting lemma (Lemma 7.3.6) can be applied to $\mathcal{W}$.

Claim 7.3.9.

$$\mathop{\mathbb{E}}_{p,x,q,\mathcal{W}}\left[\sum_{j\in[k]}\left(a^j\sum_{i\in[n]}(c_i^j-p^j)+2\left|a^j-\bar{c}^j\right|\right)\right]\geq\frac{k}{3}.$$

Proof. By linearity of expectation, it suffices to show that, for every $j\in[k]$,

$$\mathop{\mathbb{E}}_{p,x,q,\mathcal{W}}\left[a^j\sum_{i\in[n]}(c_i^j-p^j)+2\left|a^j-\bar{c}^j\right|\right]\geq\frac{1}{3}.$$

Since each column c^j is generated independently from the columns $c^1,\ldots,c^{j-1}$, c^j and p^j are independent from $q_{S_1},\cdots,q_{S_j}$. Thus, at the time $\mathcal{W}$ produces the output a^j, it does not have any information about c^j or p^j apart from its private input. (Although $\mathcal{W}$ later learns c^j when it is asked $q_{S_{j+1}}$.) For any fixed values of $c^1,\ldots,c^{j-1}$ and the internal randomness of $\mathcal{W}$, the answer a^j is a deterministic function of c^j. Thus we can apply Lemma 7.3.6 to this function to establish the claim. □

Combining Claims 7.3.8 and 7.3.9 gives

$$\mathop{\mathbb{E}}_{p,x,q,\mathcal{W}}\left[\sum_{j\in[k]}a^j\sum_{i\in[n]}(c_i^j-p^j)\right]\geq\frac{2k}{15}.$$

In particular, there exists some $i^*\in[n]$ such that

$$\mathop{\mathbb{E}}_{p,x,q,\mathcal{W}}\left[\sum_{j\in[k]}a^j(c_{i^*}^j-p^j)\right]\geq\frac{2k}{15n}. \tag{7.3}$$

To complete the proof, we show that (7.3) violates the differential privacy guarantee unless $n\geq\Omega(\sqrt{k})$.

To this end, fix any $p^1,\ldots,p^k\in[-1,1]$, whence $c_{i^*}^1,\cdots,c_{i^*}^k\in\{\pm1\}$ are independent bits with $\mathbb{E}\left[c^j\right]=p^j$. Let $\tilde{c}^1,\cdots,\tilde{c}^k\in\{\pm1\}$ be independent bits with $\mathbb{E}\left[\tilde{c}^j\right]=p^j$. The random variables $c_{i^*}^1,\cdots,c_{i^*}^k$ have the same marginal distribution as $\tilde{c}^1,\cdots,\tilde{c}^k$. However, $\tilde{c}^1,\cdots,\tilde{c}^k$ are independent from $a^1,\cdots,a^k$, whereas

$a^1, \cdots, a^k$ depend on $c^1_{i^*}, \cdots, c^k_{i^*}$. Consider the quantities

$$Z = \sum_{j \in [k]} a^j (c^j_{i^*} - p^j) \qquad \text{and} \qquad \tilde{Z} = \sum_{j \in [k]} a^j (\tilde{c}^j - p^j).$$

Differential privacy implies that Z and $\tilde{Z}$ have similar distributions. Specifically, if $\mathcal{W}$ is $(1, 1/30n)$-differentially private, then

$$\mathbb{E}\left[|Z|\right] = \int_0^{2k} \mathbb{P}\left[|Z| > z\right] \mathrm{d}z \leq \int_0^{2k} \left(e\mathbb{P}\left[|\tilde{Z}| > z\right] + \frac{1}{30n} \right) \mathrm{d}z = e\mathbb{E}\left[|\tilde{Z}|\right] + \frac{k}{15n},$$

as $|Z|, |\tilde{Z}| \leq 2k$ with probability 1.

Now $\mathbb{E}\left[|Z|\right] \geq \mathbb{E}\left[Z\right] \geq 2k/15n$, by (7.3). On the other hand, a^j is independent from $\tilde{c}^j$ and $\mathbb{E}\left[\tilde{c}^j - p^j\right] = 0$, so $\mathbb{E}\left[\tilde{Z}\right] = 0$. We now observe that

$$\mathbb{E}\left[|\tilde{Z}|\right]^2 \leq \mathbb{E}\left[\tilde{Z}^2\right] = \mathsf{Var}\left[\tilde{Z}\right] = \sum_{j \in [k]} \mathsf{Var}\left[a^j(\tilde{c}^j - p^j)\right] \leq \sum_{j \in [k]} \mathbb{E}\left[(\tilde{c}^j - p^j)^2\right] \leq k.$$

Thus, we have

$$\frac{2k}{15n} \leq \mathbb{E}\left[|Z|\right] \leq e\mathbb{E}\left[|\tilde{Z}|\right] + \frac{k}{15n} \leq e\sqrt{k} + \frac{k}{15n}.$$

The condition $2k/15n \leq e\sqrt{k} + k/15n$ is a contradiction unless $k \leq 225e^2n^2$. Thus, we can conclude that there exists a $k = O(n^2)$ such that no $(1, 1/30n)$-differentially private algorithm is accurate for more than k online queries from Q^n_{prefix}, as desired. This completes the proof. □

7.4 A Separation Between Adaptive and Non-Adaptive Online Queries

In this section we prove that even among online queries, answering adaptively-chosen queries can be strictly harder than answering non-adaptively-chosen queries.

Our separation applies to a family of search queries that we call *correlated vector queries.* We show that for a certain regime of parameters, it is possible to take a dataset of size n and privately answer an exponential number of fixed correlated vector queries, even if the queries are presented online, but it is impossible to answer more than a constant number of adaptively-chosen correlated vector queries under differential privacy.

The queries are defined on datasets $x \in \{\pm 1\}^n$ (hence the data universe is $\mathcal{X} = \{\pm 1\}$). For every query, the range $\mathcal{Y} = \{\pm 1\}^n$ is the set of n-bit vectors. We fix some parameters $0 < \alpha < 1$ and $m \in \mathbb{N}$. A query q is specified by a set V where $V = \{v^1, \ldots, v^m\} \subseteq \{\pm 1\}^n$ is a set of n-bit vectors. Roughly, an accurate answer to a given search query is any vector $y \in \{\pm 1\}^n$ that is approximately α-correlated with the input dataset $x \in \{\pm 1\}^n$ and has nearly as little correlation as possible with every v^j. By "as little correlation as possible with v^j" we mean that v^j may itself be correlated with x, in which case y should be correlated with v^j only insofar as this correlation comes through the correlation between y and x. Formally, for a query q_V, we define the loss function $L_{q_V} : \mathcal{X}^n \times \mathcal{X}^n \to \{0, 1\}$ by

$$L_{q_V}(x, y) = 0 \quad \iff \quad |\langle y - \alpha x, x \rangle| \leq \frac{\alpha^2 n}{100} \ \ \wedge \ \ \forall v^j \in V \ \left| \langle y - \alpha x, v^j \rangle \right| \leq \frac{\alpha^2 n}{100}.$$

We remark that the choice of $\alpha^2 n/100$ is somewhat arbitrary, and we can replace this choice with C for any $\sqrt{n} \ll C \ll n$ and obtain quantitatively different results. We chose to fix this particular choice in order to reduce notational clutter.

We let

$$Q_{\text{corr}}^{n,\alpha,m} = \{q_V \mid V \subseteq \{\pm 1\}^n, |V| \leq m\}$$

be the set of all correlated vector queries on $\{\pm 1\}^n$ for parameters α, m.

7.4.1 Answering Online Correlated Vector Queries

Provided that all the queries are fixed in advance, we can privately answer correlated vector queries using the randomized response algorithm. This algorithm simply takes the input vector $x \in \{\pm 1\}^n$ and outputs a new vector $y \in \{\pm 1\}^n$ where each bit y_i is independent and is set to x_i with probability $1/2 + \rho$ for a suitable choice of $\rho > 0$. The algorithm will then answer every correlated vector query with this same vector y. The following theorem captures the parameters that this mechanism achieves.

Theorem 7.4.1 (Answering Online Correlated Vector Queries)**.** *For every $0 < \alpha < 1/2$, there exists $k = 2^{\Omega(\alpha^4 n)}$ such that, for every sufficiently large $n \in \mathbb{N}$, there is a $(3\alpha, 0)$-differentially private algorithm $\mathcal{W}_{\mathsf{corr}}$ that takes a dataset $x \in \{\pm 1\}^n$ and is $(1/k)$-accurate for k online queries from $Q_{corr}^{n,\alpha,k}$.*

Proof Theorem 7.4.1. Our algorithm based on randomized response is presented in Figure 7.5 below.

$\mathcal{W}_{\mathsf{corr}}$:
Input: a dataset $x \in \{\pm 1\}^n$.
Parameters: $0 < \alpha < 1/2$.
For $i = 1, \ldots, n$, independently set

$$y_i = \begin{cases} +x_i & \text{with probability } \frac{1+\alpha}{2} \\ -x_i & \text{with probability } \frac{1-\alpha}{2} \end{cases}.$$

Let $y = (y_1, \ldots, y_n) \in \{\pm 1\}^n$, and answer each query with y.

Figure 7.5: $\mathcal{W}_{\mathsf{corr}}$

To establish privacy, observe that by construction each output bit y_i depends only on x_i and is independent of all x_j, y_j for $j \neq i$. Therefore, it suffices to observe

that if $0 < \alpha < 1/2$,

$$1 \leq \frac{\mathbb{P}[y_i = +1 \mid x_i = +1]}{\mathbb{P}[y_i = +1 \mid x_i = -1]} = \frac{1+\alpha}{1-\alpha} \leq e^{3\alpha}$$

and similarly

$$1 \geq \frac{\mathbb{P}[y_i = -1 \mid x_i = +1]}{\mathbb{P}[y_i = -1 \mid x_i = -1]} = \frac{1-\alpha}{1+\alpha} \geq e^{-3\alpha}.$$

To prove accuracy, observe that since the output y does not depend on the sequence of queries, we can analyse the mechanism as if the queries $q_{V_1}, \ldots, q_{V_k} \in Q_{\text{corr}}^{n,\alpha,k}$ were fixed and given all at once. Let $V = \bigcup_{j=1}^{k} V_j$, and note that $|V| \leq k^2$. First, observe that $\mathbb{E}[y] = \alpha x$. Thus we have

$$\mathop{\mathbb{E}}_{y}[\langle y - \alpha x, x\rangle] = 0 \quad \text{and} \quad \forall v \in V \ \mathop{\mathbb{E}}_{y}[\langle y - \alpha x, v\rangle] = 0$$

Since x and every vector in V is fixed independently of y, and the coordinates of y are independent by construction, the quantities $\langle y, x\rangle$ and $\langle y, v\rangle$ are each the sum of n independent $\{\pm 1\}$-valued random variables. Thus, we can apply Hoeffding's inequality[5] and a union bound to conclude

$$\mathop{\mathbb{P}}_{y}\left[|\langle y - \alpha x, x\rangle| > \frac{\alpha^2 n}{100}\right] \leq 2\exp\left(\frac{-\alpha^4 n}{20000}\right)$$
$$\mathop{\mathbb{P}}_{y}\left[\exists v \in V \text{ s.t. } |\langle y - \alpha x, v\rangle| > \frac{\alpha^2 n}{100}\right] \leq 2k^2\exp\left(\frac{-\alpha^4 n}{20000}\right)$$

The theorem now follows by setting an appropriate choice of $k = 2^{\Omega(\alpha^4 n)}$ such that $2(k^2+1)\cdot\exp\left(\frac{-\alpha^4 n}{20000}\right) \leq 1/k$. □

[5]We use the following statement of Hoeffding's Inequality: if $Z_1, \ldots, Z_n$ are independent $\{\pm 1\}$-valued random variables, and $Z = \sum_{i=1}^{n} Z_i$, then

$$\mathbb{P}\left[\Big|\ Z - \mathbb{E}[Z]\ \Big| > C\sqrt{n}\right] \leq 2e^{-C^2/2}$$

7.4.2 A Lower Bound for Adaptive Correlated Vector Queries

We now prove a contrasting lower bound showing that if the queries may be chosen adaptively, then no differentially private algorithm can answer more than a constant number of correlated vector queries. The key to our lower bound is that fact that adaptively-chosen correlated vector queries allow an adversary to obtain many vectors $y^1, \ldots, y^k$ that are correlated with x but pairwise nearly orthogonal with each other. As we prove, if k is sufficiently large, this information is enough to recover a vector $\tilde{x}$ that has much larger correlation with x than any of the vectors $y^1, \ldots, y^k$ have with x. By setting the parameters appropriately, we will obtain a contradiction to differential privacy.

Theorem 7.4.2 (Lower Bound for Correlated Vector Queries)**.** *For every $0 < \alpha < 1/2$, there is a $k = O(1/\alpha^2)$ such that for every sufficiently large $n \in \mathbb{N}$, there is no $(1, 1/20)$-differentially private algorithm that takes a dataset $x \in \{\pm 1\}^n$ and is $1/100$-accurate for k adaptive queries from $Q_{corr}^{n,\alpha,k}$*

We remark that the value of k in our lower bound is optimal up to constants, as there is a $(1, 1/20)$-differentially private algorithm that can answer $k = \Omega(1/\alpha^2)$ adaptively-chosen queries of this sort. The algorithm simply answers each query with an independent invocation of randomized response. Randomized response is $O(\alpha)$-differentially private for each query, and we can invoke the adaptive composition theorem [DMNS06, DRV10] to argue differential privacy for $k = \Omega(1/\alpha^2)$-queries.

Before proving Theorem 7.4.2, we state and prove the combinatorial lemma that forms the foundation of our lower bound.

Lemma 7.4.3 (Reconstruction Lemma)**.** *Fix parameters $0 \leq a, b \leq 1$. Let $x \in \{\pm 1\}^n$*

and $y^1, \cdots, y^k \in \{\pm 1\}^n$ *be vectors such that*

$$\forall 1 \le j \le k \quad \langle y^j, x \rangle \ge an$$
$$\forall 1 \le j < j' \le k \quad |\langle y^j, y^{j'} \rangle| \le bn.$$

Then, if we let $\tilde{x} = \mathrm{sign}(\sum_{j=1}^k y^j) \in \{\pm 1\}^n$ *be the coordinate-wise majority of* $y^1, \ldots, y^k$*, we have*

$$\langle \tilde{x}, x \rangle \ge \left(1 - \frac{2}{a^2 k} - \frac{2(b-a^2)}{a^2}\right) n.$$

Proof of Lemma 7.4.3. Let

$$\overline{y} = \frac{1}{k} \sum_{j=1}^k y^j \in [-1,1]^n.$$

By linearity, $\langle \overline{y}, x \rangle \ge an$ and

$$\|\overline{y}\|_2^2 = \frac{1}{k^2} \sum_{j,j'=1}^k \langle y^j, y^{j'} \rangle \le \frac{1}{k^2} \left(kn + (k^2-k)bn\right) \le \left(\frac{1}{k} + b\right) n.$$

Define a random variable $W \in [-1,1]$ to be $x_i \overline{y}_i$ for a uniformly random $i \in [n]$. Then

$$\mathbb{E}[W] = \frac{1}{n}\langle x, \overline{y} \rangle \ge a \quad \text{and} \quad \mathbb{E}\left[W^2\right] = \frac{1}{n} \sum_{i=1}^n x_i^2 \overline{y}_i^2 = \frac{1}{n}\|\overline{y}\|_2^2 \le \frac{1}{k} + b$$

By Chebyshev's inequality,

$$\mathbb{P}[W \le 0] \le \mathbb{P}\left[|W - \mathbb{E}[W]| \ge a\right] \le \frac{\mathrm{Var}[W]}{a^2} = \frac{\mathbb{E}[W^2] - \mathbb{E}[W]^2}{a^2} \le \frac{\frac{1}{k} + b - a^2}{a^2}.$$

Meanwhile,

$$\mathbb{P}[W \le 0] = \frac{1}{n} \sum_{i=1}^n \mathbb{I}[x_i \overline{y}_i \le 0] \ge \frac{1}{n} \sum_{i=1}^n \mathbb{I}[\mathrm{sign}(\overline{y}_i) \ne x_i] = \frac{1}{2} - \frac{1}{2n}\langle \mathrm{sign}(\overline{y}), x \rangle.$$

Thus we conclude

$$\langle \mathrm{sign}(\overline{y}), x \rangle \ge n - 2n\mathbb{P}[W \le 0] \ge n - 2n\left(\frac{\frac{1}{k} + b - a^2}{a^2}\right)$$

To complete the proof, we rearrange terms and note that $\text{sign}(\overline{y}) = \text{sign}(\sum_{j=1}^{k} y^j)$.

□

Now we are ready to prove our lower bound for algorithms that answer adaptively-chosen correlated vector queries.

Proof of Theorem 7.4.2. We will show that the output $y^1, \ldots, y^k$ of any algorithm $\mathcal{W}$ that takes a dataset $x \in \{\pm 1\}^n$ and answers $k = 100/\alpha^2$ adaptively-chosen correlated vector queries can be used to find a vector $\tilde{x} \in \{\pm 1\}^n$ such that $\langle \tilde{x}, x \rangle > n/2$. In light of Lemma 7.4.3, this vector will simply be $\tilde{x} = \text{sign}(\sum_{j=1}^{k} y^j)$. We will then invoke the following elementary fact that differentially private algorithms do not admit this sort of reconstruction of their input dataset.

Fact 7.4.4. *For every sufficiently large $n \in \mathbb{N}$, there is no $(1, 1/20)$-differentially private algorithm $\mathcal{W} : \{\pm 1\}^n \to \{\pm 1\}^n$ such that for every $x \in \{\pm 1\}^n$, with probability at least $99/100$, $\langle \mathcal{W}(x), x \rangle > n/2$.*

The attack works as follows. For $j = 1, \ldots, k$, define the set $V_j = \{y^1, \ldots, y^{j-1}\}$ and ask the query $q_{V_j}(x) \in Q_{\text{corr}}^{n,\alpha,k}$ to obtain some vector y^j. Since $\mathcal{W}$ is assumed to be accurate for k adaptively-chosen queries, with probability $99/100$, we obtain vectors $y^1, \ldots, y^k \in \{\pm 1\}^n$ such that

$$\begin{aligned} \forall 1 \le j \le k \quad \langle y^j, x \rangle &\ge \langle \alpha x, x \rangle - |\langle y - \alpha x, x \rangle| \\ &\ge \alpha n - \frac{\alpha^2 n}{100} \\ &\ge an, \end{aligned}$$

$$\begin{aligned}
\forall 1 \le j < j' \le k \quad |\langle y^j, y^{j'} \rangle| &\le |\langle \alpha x, y^j \rangle| + |\langle y^{j'} - \alpha x, y^j \rangle| \\
&\le \alpha |\langle y^j, x \rangle| + \frac{\alpha^2 n}{100} \\
&\le \alpha \left(|\langle \alpha x, x \rangle| + |\langle y^j - \alpha x, x \rangle| \right) + \frac{\alpha^2 n}{100} \\
&\le \alpha^2 n + \frac{\alpha^3 n}{100} + \frac{\alpha^2 n}{100} \\
&\le \frac{51}{50} \alpha^2 n \\
&= bn,
\end{aligned}$$

where $a = 99\alpha/100$ and $b = 51\alpha^2/50$. Thus, by Lemma 7.4.3, if $\tilde{x} = \text{sign}(\sum_{j=1}^{k} y^j)$, and $k = 100/\alpha^2$, we have

$$\begin{aligned}
\langle \tilde{x}, x \rangle &\ge \left(1 - \frac{2}{a^2 k} - \frac{2(b - a^2)}{a^2} \right) n \\
&= \left(1 - \frac{2}{(99\alpha/100)^2 k} - \frac{2(51\alpha^2/50 - (99\alpha/100)^2)}{(99\alpha/100)^2} \right) n \\
&= \left(1 - \frac{2(100/99)^2}{100} - 2 \left(\frac{(51/50) - (99/100)^2}{(99/100)^2} \right) \right) n \\
&\ge 0.89n \ge n/2.
\end{aligned}$$

By Fact 7.4.4, this proves that $\mathcal{W}$ cannot be $(1, 1/20)$-differentially private. □

7.5 Threshold Queries

First we define threshold queries, which are a family of statistical queries.

Definition 7.5.1. *Let* Thresh_X *denote the class of threshold queries over a totally ordered domain* X*. That is,* $\mathsf{Thresh}_X = \{c_x : x \in X\}$ *where* $c_x : X \to \{0,1\}$ *is defined by* $c_x(y) = 1$ *iff* $y \le x$*.*

7.5.1 Separation for Pure Differential Privacy

In this section, we show that the sample complexity of answering adaptively-chosen thresholds can be exponentially larger than that of answering thresholds offline.

Proposition 7.5.2 ([DNPR10, CSS11, DNRR15]). *Let X be any totally ordered domain. Then there exists a $(\varepsilon, 0)$-differentially private mechanism M that, given $x \in X^n$, gives α-accurate answers to k offline queries from Thresh_X for*

$$n = O\left(\min\left\{\frac{\log k + \log^2(1/\alpha)}{\alpha\varepsilon}, \frac{\log^2 k}{\alpha\varepsilon}\right\}\right)$$

On the other hand, we show that answering k adaptively-chosen threshold queries can require sample complexity as large as $\Omega(k)$ – an exponential gap. Note that this matches the upper bound given by the Laplace mechanism [DMNS06].

Proposition 7.5.3. *Answering k adaptively-chosen threshold queries on $[2^{k-1}]$ to accuracy α subject to ε-differential privacy requires sample complexity $n = \Omega(k/\alpha\varepsilon)$.*

The idea for the lower bound is that an analyst may adaptively choose k threshold queries to binary search for an "approximate median" of the dataset. However, a packing argument shows that locating an approximate median requires sample complexity $\Omega(k)$.

Definition 7.5.4 (Approximate Median). *Let X be a totally ordered domain, $\alpha > 0$, and $x \in X^n$. We call $y \in X$ an* α-approximate median *of x if*

$$\frac{1}{n}\left|\{i \in [n] : x_i \leq y\}\right| \geq \frac{1}{2} - \alpha \qquad \text{and} \qquad \frac{1}{n}\left|\{i \in [n] : x_i \geq y\}\right| \geq \frac{1}{2} - \alpha.$$

Proposition 7.5.3 is obtained by combining Lemmas 7.5.5 and 7.5.6 below.

Lemma 7.5.5. *Suppose M answers $k = \lceil 1 + \log_2 T \rceil$ adaptively-chosen queries from $\mathsf{Thresh}_{[T]}$ with ε-differential privacy and (α, β)-accuracy. Then there exists an ε-differentially*

private $M' : [T]^n \to [T]$ *that computes an* α*-approximate median with probability at least* $1 - \beta$.

Proof. The algorithm M', formalised in Figure 7.6, uses M to perform a binary search.

Input: $x \in X^n$.
M is given x.
Initialize $\ell_1 = 0$, $u_1 = T$, and $j = 1$.
While $u_j - \ell_j > 1$ repeat:
  Let $m_j = \lceil (u_j + \ell_j)/2 \rceil$.
  Give M the query $c_{m_j} \in \mathsf{Thresh}_{[T]}$ and obtain the answer $a_j \in [0, 1]$.
  If $a_j \geq \frac{1}{2}$, set $(\ell_{j+1}, u_{j+1}) = (\ell_j, m_j)$; otherwise set $(\ell_{j+1}, u_{j+1}) = (m_j, u_j)$.
  Increment j.
Output u_j.

Figure 7.6: $M' : X^n \to X$

We have $u_1 - \ell_1 = T$ and, after every query j, $u_{j+1} - \ell_{j+1} \leq \lceil (u_j - \ell_j)/2 \rceil$. Since the process stops when $u_j - \ell_j = 1$, it is easy to verify that M' makes at most $\lceil 1 + \log_2(T - 1) \rceil$ queries to M.

Suppose all of the answers given by M are α-accurate. This happens with probability at least $1 - \beta$. We will show that, given this, M' outputs an α-approximate median, which completes the proof.

We claim that $c_{u_j}(x) \geq \frac{1}{2} - \alpha$ for all j. This is easily shown by induction. The base case is $c_T(x) = 1 \geq \frac{1}{2} - \alpha$. At each step either $u_{j+1} = u_j$ (in which case the induction hypothesis can be applied) or $u_{j+1} = m_j$; in the latter case our accuracy assumption gives

$$c_{u_{j+1}}(x) = c_{m_j}(x) \geq a_j - \alpha \geq \frac{1}{2} - \alpha.$$

We also claim that $c_{\ell_j}(x) < \frac{1}{2} + \alpha$ for all j. This follows from a similar induction and completes the proof. □

Lemma 7.5.6. *Let $M : [T]^n \to [T]$ be an ε-differentially private algorithm that computes an α-approximate median with confidence $1 - \beta$. Then*

$$n \geq \Omega\left(\frac{\log T + \log(1/\beta)}{\alpha\varepsilon}\right).$$

Proof. Let $m = \lceil(\frac{1}{2} - \alpha)n\rceil - 1$. For each $t \in [T]$, let $x^t \in [T]^n$ denote the dataset containing m copies of 1, m copies of T, and $n - 2m$ copies of t. Then for each $t \in [T]$,

$$\mathbb{P}\left[M(x^t) = t\right] \geq 1 - \beta.$$

On the other hand, by the pigeonhole principle, there must exist $t_* \in [T-1]$ such that

$$\mathbb{P}\left[M(x^T) = t_*\right] \leq \frac{\mathbb{P}\left[M(x^T) \in [T-1]\right]}{T-1} \leq \frac{\beta}{T-1}.$$

The inputs x^T and x^{t_*} differ in at most $n - 2m \leq 2\alpha n + 2$ entries. By group privacy,

$$1 - \beta \leq \mathbb{P}\left[M(x^{t_*}) = t_*\right] \leq e^{\varepsilon(2\alpha n+2)}\mathbb{P}\left[M(x^T) = t_*\right] \leq e^{\varepsilon(2\alpha n+2)}\frac{\beta}{T-1}.$$

Rearranging these inequalities gives

$$O(\varepsilon\alpha n) \geq \varepsilon(2\alpha n + 2) \geq \log\left(\frac{(1-\beta)(T-1)}{\beta}\right) \geq \Omega(\log(T/\beta)),$$

which yields the result. □

Remark 7.5.7. *Proposition 7.5.3 can be extended to* online non-adaptive *queries, which yields a separation between the online non-adaptive and offline models for pure differential privacy and threshold queries.*

The key observation behind remark 7.5.7 is that, while Lemma 7.5.5 in general requires making adaptive queries, for the inputs $x^t \in [T]^n$ ($t \in [T]$) used in Lemma 7.5.6 the queries are "predictable." In particular, on input x^t, the algorithm M' from the proof of Lemma 7.5.5 will (with probability at least $1 - \beta$) always make

the same sequence queries. This allows the queries to be specified in advance in a non-adaptive manner. More precisely, we can produce an algorithm M'_t that produces non-adaptive online queries by simulating M' on input x^t and using those queries. Given the answers to these online non-adaptive queries, M'_t can either accept or reject its input depending on whether the answers are consistent with the input x^t; M'_t will accept x^t with high probability and reject $x^{t'}$ for $t' \neq t$ with high probability. The proof of Lemma 7.5.6 can be carried out using M'_{t_*} instead of M' at the end.

7.5.2 The BetweenThresholds Algorithm

The key technical novelty behind our algorithm for answering adaptively-chosen threshold queries is a refinement of the "Above Threshold" algorithm [DR14, §3.6], which underlies the ubiquitous "sparse vector" technique [DNR+09, RR10, DNPR10, HR10].

The sparse vector technique addresses a setting where we have a stream of k (adaptively-chosen) low-sensitivity queries and a threshold parameter t. Instead of answering all k queries accurately, we are interested in answering only the ones that are above the threshold t – for the remaining queries, we only require a signal that they are below the threshold. Intuitively, one would expect to only pay in privacy for the queries that are actually above the threshold. And indeed, one can get away with sample complexity proportional to the number of queries that are above the threshold, and to the *logarithm* of the total number of queries.

We extend the sparse vector technique to settings where we demand slightly more information about each query beyond whether it is below a single threshold. In particular, we set two thresholds $t_\ell < t_u$, and for each query, release a signal as to whether the query is below the lower threshold, above the upper threshold, or

between the two thresholds.

As long as the thresholds are sufficiently far apart, whether (the noisy answer to) a query is below the lower threshold or above the upper threshold is *stable*, in that it is extremely unlikely to change on neighboring datasets. As a result, we obtain an (ε, δ)-differentially private algorithm that achieves the same accuracy guarantees as the traditional sparse vector technique, i.e. sample complexity proportional to $\log k$.

Our algorithm is summarised by the following theorem.[6]

Theorem 7.5.8. *Let $\alpha, \beta, \varepsilon, \delta, t \in (0,1)$ and $n, k \in \mathbb{N}$ satisfy*

$$n \geq \frac{1}{\alpha\varepsilon} \max\left\{12\log(30/\varepsilon\delta), 16\log((k+1)/\beta)\right\}.$$

Then there exists a (ε, δ)-differentially private algorithm that takes as input $x \in X^n$ and answers a sequence of adaptively-chosen queries $q_1, \cdots, q_k : X^n \to [0,1]$ of sensitivity $1/n$ with $a_1, \cdots, a_{\leq k} \in \{\mathrm{L}, \mathrm{R}, \top\}$ such that, with probability at least $1-\beta$,

- $a_j = \mathrm{L} \implies q_j(x) \leq t$,
- $a_j = \mathrm{R} \implies q_j(x) \geq t$, *and*
- $a_j = \top \implies t - \alpha \leq q_j(x) \leq t + \alpha$.

The algorithm may halt before answering all k queries; however, it only halts after outputting $\top$.

Our algorithm is given in Figure 7.7. The analysis is split into Lemmas 7.5.9 and 7.5.10.

Lemma 7.5.9 (Privacy for BetweenThresholds)**.** *Let $\varepsilon, \delta \in (0,1)$ and $n \in \mathbb{N}$. Then* BetweenThresholds *(Figure 7.7) is (ε, δ)-differentially private for any adaptively-chosen*

[6] In Theorem 7.5.8, only one threshold is allowed. However, our algorithm is more general and permits the setting of two thresholds. We have chosen this statement for simplicity.

Input: $x \in X^n$.
Parameters: $\varepsilon, t_\ell, t_u \in (0,1)$ and $n, k \in \mathbb{N}$.
Sample $\mu \sim \mathrm{Lap}(2/\varepsilon n)$ and initialize noisy thresholds $\hat{t}_\ell = t_\ell + \mu$ and $\hat{t}_u = t_u - \mu$.
For $j = 1, 2, \cdots, k$:
  Receive query $q_j : X^n \to [0,1]$.
  Set $c_j = q_j(x) + \nu_j$ where $\nu_j \sim \mathrm{Lap}(6/\varepsilon n)$.
  If $c_j < \hat{t}_\ell$, output L and continue.
  If $c_j > \hat{t}_u$, output R and continue.
  If $c_j \in [\hat{t}_\ell, \hat{t}_u]$, output $\top$ and halt.

Figure 7.7: BetweenThresholds

sequence of queries as long as the gap between the thresholds t_ℓ, t_u satisfies

$$t_u - t_\ell \geq \frac{12}{\varepsilon n} \left(\log(10/\varepsilon) + \log(1/\delta) + 1 \right).$$

Lemma 7.5.10 (Accuracy for BetweenThresholds)**.** *Let $\alpha, \beta, \varepsilon, t_\ell, t_u \in (0,1)$ and $n, k \in \mathbb{N}$ satisfy*

$$n \geq \frac{8}{\alpha\varepsilon} \left(\log(k+1) + \log(1/\beta) \right).$$

Then, for any input $x \in X^n$ and any adaptively-chosen sequence of queries $q_1, q_2, \cdots, q_k$, the answers $a_1, a_2, \cdots a_{\leq k}$ produced by BetweenThresholds *(Figure 7.7) on input x satisfy the following with probability at least $1 - \beta$. For any $j \in [k]$ such that a_j is returned before* BetweenThresholds *halts,*

- $a_j = \mathrm{L} \implies q_j(x) \leq t_\ell + \alpha$,
- $a_j = \mathrm{R} \implies q_j(x) \geq t_u - \alpha$, *and*
- $a_j = \top \implies t_\ell - \alpha \leq q_j(x) \leq t_u + \alpha$.

Combining Lemmas 7.5.9 and 7.5.10 and setting $t_\ell = t - \alpha/2$ and $t_u = t + \alpha/2$ yields Theorem 7.5.8.

Proof of Lemma 7.5.9. Our analysis is an adaptation of Dwork and Roth's [DR14, §3.6] analysis of the AboveThreshold algorithm. Recall that a transcript of the execution of BetweenThresholds is given by $a \in \{\mathrm{L}, \mathrm{R}, \top\}^*$. Let $\mathcal{M} : X^n \to \{\mathrm{L}, \mathrm{R}, \top\}^*$ denote the function that simulates BetweenThresholds interacting with a given adaptive adversary (cf. Figure 7.3) and returns the transcript.

Let $S \subset \{\mathrm{L}, \mathrm{R}, \top\}^*$ be a set of transcripts. Our goal is to show that for adjacent datasets $x \sim x'$,

$$\mathbb{P}\left[\mathcal{M}(x) \in S\right] \leq e^{\varepsilon}\mathbb{P}\left[\mathcal{M}(x') \in S\right] + \delta.$$

Let

$$z^* = \frac{1}{2}(t_u - t_\ell) - \frac{6}{\varepsilon n}\log(10/\varepsilon) - 1/n \geq \frac{2}{\varepsilon n}\log(1/\delta).$$

Our strategy will be to show that as long as the noise value μ is under control, in particular if $\mu \leq z^*$, then the algorithm behaves in essentially the same way as the standard AboveThreshold algorithm. Meanwhile, the event $\mu > z^*$ which corresponds to the (catastrophic) event where the upper and lower thresholds are too close or overlap, happens with probability at most δ.

The following claim reduces the privacy analysis to examining the probability of obtaining any single transcript a:

Claim 7.5.11. *Suppose that for any transcript $a \in \{\mathrm{L}, \mathrm{R}, \top\}^*$, and any $z \leq z^*$, that*

$$\mathbb{P}\left[\mathcal{M}(x) = a | \mu = z\right] \leq e^{\varepsilon/2}\mathbb{P}\left[\mathcal{M}(x') = a | \mu = z + 1/n\right].$$

Then $\mathcal{M}$ is (ε, δ)-differentially private.

Proof. By properties of the Laplace distribution, since $\mu \sim \mathrm{Lap}(2/\varepsilon n)$, for any $z \in \mathbb{R}$, we have

$$\mathbb{P}\left[\mu = z\right] \leq e^{\varepsilon/2}\mathbb{P}\left[\mu = z + 1/n\right],$$

and

$$\mathbb{P}\left[\mu > z^*\right] = \frac{1}{2} e^{-\varepsilon n z^*/2} \leq \delta.$$

Fix a set of transcripts S. Combining these properties allows us to write

$$\begin{aligned}
\mathbb{P}\left[\mathcal{M}(x) \in S\right] &= \int_{\mathbb{R}} \mathbb{P}\left[\mathcal{M}(x) \in S | \mu = z\right] \mathbb{P}\left[\mu = z\right] \mathrm{d}z \\
&\leq \left(\int_{-\infty}^{z^*} \mathbb{P}\left[\mathcal{M}(x) \in S | \mu = z\right] \mathbb{P}\left[\mu = z\right] \mathrm{d}z \right) + \mathbb{P}\left[\mu > z^*\right] \\
&\leq \left(e^{\varepsilon/2} \int_{-\infty}^{z^*} \mathbb{P}\left[\mathcal{M}(x') \in S | \mu = z + 1/n\right] \mathbb{P}\left[\mu = z\right] \mathrm{d}z \right) + \delta \\
&\leq \left(e^{\varepsilon} \int_{-\infty}^{z^*} \mathbb{P}\left[\mathcal{M}(x') \in S | \mu = z + 1/n\right] \mathbb{P}\left[\mu = z + 1/n\right] \mathrm{d}z \right) + \delta \\
&\leq e^{\varepsilon} \mathbb{P}\left[\mathcal{M}(x') \in S\right] + \delta
\end{aligned}$$

□

Returning to the proof of Lemma 7.5.9, fix a transcript $a \in \{\mathrm{L}, \mathrm{R}, \top\}^*$. Our goal is now to show that $\mathcal{M}$ satisfies the hypotheses of Claim 7.5.11, namely that for any $z \leq z^*$,

$$\mathbb{P}\left[\mathcal{M}(x) = a | \mu = z\right] \leq e^{\varepsilon/2} \mathbb{P}\left[\mathcal{M}(x') = a | \mu = z + 1/n\right]. \tag{7.4}$$

For some $k \geq 1$, we can write the transcript a as $(a_1, a_2, \ldots, a_k)$, where $a_j \in \{\mathrm{L}, \mathrm{R}\}$ for each $j < k$, and $a_k = \top$.

For convenience, let $A = \mathcal{M}(x)$ and $A' = \mathcal{M}(x')$. We may decompose

$$\begin{aligned}
\mathbb{P}\left[\mathcal{M}(x) = a | \mu = z\right] &= \mathbb{P}\left[(\forall j < k, A_j = a_j) \ \wedge \ q_k(x) + \nu_k \in [\hat{t}_\ell, \hat{t}_u] | \mu = z\right] \\
&= \mathbb{P}\left[(\forall j < k, A_j = a_j) | \mu = z\right] \\
&\quad \cdot \ \mathbb{P}\left[q_k(x) + \nu_k \in [\hat{t}_\ell, \hat{t}_u] | \mu = z \wedge (\forall j < k, A_j = a_j)\right].
\end{aligned} \tag{7.5}$$

We upper bound each factor on the right-hand side separately.

Claim 7.5.12.

$$\mathbb{P}\left[(\forall i < k, A_i = a_i) | \mu = z\right] \leq \mathbb{P}\left[(\forall i < k, A'_i = a_i) | \mu = z + 1/n\right]$$

Proof. For fixed z, let $A_z(x)$ denote the set of noise vectors $(\nu_1, \ldots, \nu_{k-1})$ for which $(A_1, \ldots, A_{k-1}) = (a_1, \ldots, a_{k-1})$ when $\nu = z$. We claim that as long as $z \leq z^*$, then $A_z(x) \subseteq A_{z+1/n}(x')$. To argue this, let $(\nu_1, \ldots, \nu_{k-1}) \in A_z(x)$. Fix an index $j \in \{1, \ldots, k-1\}$ and suppose $a_j = \mathrm{L}$. Then $q_j(x) + \nu_j < t_\ell + z$, but since q_j has sensitivity $1/n$, we also have $q_j(x') + \nu_j < t_\ell + (z + 1/n)$. Likewise, if $a_j = \mathrm{R}$, then $q_j(x) + \nu_j > t_u - z$, so

$$q_j(x') + \nu_j > t_u - z - 1/n \geq t_\ell + (z + 1/n)$$

as long as $z \leq z^* \leq \frac{1}{2}(t_u - t_\ell) - 1/n$. (This ensures that $\mathcal{M}(x')$ does not output L on the first branch of the "if" statement, and proceeds to output R.)

Since $A_z(x) \subseteq A_{z+1/n}(x')$, this proves that

$$\begin{aligned}
\mathbb{P}\left[(\forall i < k, A_i = a_i) | \mu = z\right] &= \mathbb{P}\left[(\nu_1, \ldots, \nu_{k-1}) \in A_z(x)\right] \\
&\leq \mathbb{P}\left[(\nu_1, \ldots, \nu_{k-1}) \in A_{z+1/n}(x')\right] \\
&= \mathbb{P}\left[(\forall i < k, A'_i = a_i) | \mu = z + 1/n\right].
\end{aligned}$$

□

Given Claim 7.5.12, all that is needed to prove (7.4) and, thereby, prove Lemma 7.5.9 is to bound the second factor in (7.5) — that is, we must only show that

$$\begin{aligned}
&\mathbb{P}\left[q_k(x) + \nu_k \in [\hat{t}_\ell, \hat{t}_u] | \mu = z \wedge (\forall j < k, A_j = a_j)\right] \\
&\leq e^{\varepsilon/2}\mathbb{P}\left[q_k(x') + \nu_k \in [\hat{t}_\ell, \hat{t}_u] | \mu = z + 1/n \wedge (\forall j < k, A'_j = a_j)\right].
\end{aligned}$$

Let $\Delta = (q_k(x') - q_k(x)) \in [-1/n, 1/n]$. Then

$$\begin{aligned}
&\mathbb{P}\left[q_k(x) + \nu_k \in [\hat{t}_\ell, \hat{t}_u] | \mu = z \wedge (\forall j < k, A_j = a_j)\right] \\
&= \mathbb{P}\left[t_\ell + z \leq q_k(x) + \nu_k \leq t_u - z\right] \\
&= \mathbb{P}\left[t_\ell + z + \Delta \leq q_k(x') + \nu_k \leq t_u - z + \Delta\right] \\
&= \mathbb{P}\left[t_\ell + (z + 1/n) + (\Delta - 1/n) \leq q_k(x') + \nu_k \leq t_u - (z + 1/n) + (\Delta + 1/n)\right] \\
&= \mathbb{P}\left[q_k(x') + \nu_k \in [\hat{t}_\ell + \Delta - 1/n, \hat{t}_u + \Delta + 1/n] | \mu = z + 1/n\right] \\
&\leq e^{\varepsilon/2}\mathbb{P}\left[q_k(x') + \nu_k \in [\hat{t}_\ell, \hat{t}_u] | \mu = z + 1/n\right] \\
&= e^{\varepsilon/2}\mathbb{P}\left[q_k(x') + \nu_k \in [\hat{t}_\ell, \hat{t}_u] | \mu = z + 1/n \wedge (\forall j < k, A'_j = a_j)\right]
\end{aligned}$$

where the last inequality follows from Claim 7.5.13 below (setting $\eta = 2/n$, $\lambda = 6/\varepsilon n$, $[a, b] = [\hat{t}_\ell, \hat{t}_u]$, and $[a', b'] = [\hat{t}_\ell + \Delta - 1/n, \hat{t}_u + \Delta + 1/n]$) and the fact that $z \leq z^* = \frac{1}{2}(t_u - t_\ell) - \frac{6}{\varepsilon n}\log(10/\varepsilon) - 1/n$ implies

$$b - a = \hat{t}_u - \hat{t}_\ell = t_u - t_\ell - 2\mu \geq \frac{12}{\varepsilon n}\log\left(\frac{10}{\varepsilon}\right) \geq 2\lambda \log\left(\frac{1}{1 - e^{-\varepsilon/6}}\right)$$

whenever $0 \leq \varepsilon \leq 1$.

Claim 7.5.13. *Let $\nu \sim Lap(\lambda)$ and let $[a, b], [a', b'] \subset \mathbb{R}$ be intervals satisfying $[a, b] \subset [a', b']$. If $\eta \geq (b' - a') - (b - a)$, then*

$$\mathbb{P}\left[\nu \in [a', b']\right] \leq \frac{e^{\eta/\lambda}}{1 - e^{-(b-a)/2\lambda}} \cdot \mathbb{P}\left[\nu \in [a, b]\right].$$

Proof. Recall that the probability density function of the Laplace distribution is given by $f_\lambda(x) = \frac{1}{2\lambda}e^{-|x|/\lambda}$. There are four cases to consider: In the first case, $a < b \leq 0$. In the second case, $a < 0 < b$ with $|a| \leq |b|$. In the third case, $0 \leq a < b$. Finally, in the fourth case, $a < 0 < b$ with $|a| \geq |b|$. Since the Laplace distribution is symmetric,

it suffices to analyse the first two cases.

Case 1: Suppose $a < b \leq 0$. Then

$$\begin{aligned}
\mathbb{P}\left[\nu \in [a', b']\right] &\leq \mathbb{P}\left[\nu \in [a, b]\right] + \int_b^{b+\eta} \frac{1}{2\lambda} e^{x/\lambda} \mathrm{d}x \\
&= \frac{1}{2}(e^{(b+\eta)/\lambda} - e^{a/\lambda}) \\
&= \frac{1}{2} \cdot \left(\frac{e^{\eta/\lambda} - e^{(a-b)/\lambda}}{1 - e^{(a-b)/\lambda}}\right) \cdot (e^{b/\lambda} - e^{a/\lambda}) \\
&= \left(\frac{e^{\eta/\lambda} - e^{-(b-a)/\lambda}}{1 - e^{-(b-a)/\lambda}}\right) \cdot \mathbb{P}\left[\nu \in [a, b]\right].
\end{aligned}$$

Case 2: Suppose $a < 0 < b$ and $|a| \leq |b|$. Note that this implies $b \geq (b-a)/2$. Then

$$\begin{aligned}
\mathbb{P}\left[\nu \in [a', b']\right] &\leq \mathbb{P}\left[\nu \in [a, b]\right] + \eta \cdot \frac{1}{2\lambda} e^{a/\lambda} \\
&\leq \mathbb{P}\left[\nu \in [a, b]\right] \left(1 + \frac{\eta}{2\lambda} \frac{e^{a/\lambda}}{\mathbb{P}\left[\nu \in [0, b]\right]}\right) \\
&= \mathbb{P}\left[\nu \in [a, b]\right] \frac{1 - e^{-b/\lambda} + \frac{\eta}{\lambda} e^{a/\lambda}}{1 - e^{-b/\lambda}} \\
&\leq \mathbb{P}\left[\nu \in [a, b]\right] \frac{1 + \eta/\lambda}{1 - e^{-b/\lambda}} \\
&\leq \mathbb{P}\left[\nu \in [a, b]\right] \frac{e^{\eta/\lambda}}{1 - e^{-(b-a)/2\lambda}}.
\end{aligned}$$

□

□

Proof of Lemma 7.5.10. We claim that it suffices to show that with probability at least $1 - \beta$ we have

$$\forall 1 \leq j \leq k \qquad |\nu_j| + |\mu| \leq \alpha.$$

To see this, suppose $|\nu_j| + |\mu| \leq \alpha$ for every j. Then, if $a_j = \mathrm{L}$, we have

$$c_j = q_j(x) + \nu_j < \hat{t}_\ell = t_\ell + \mu, \qquad \text{whence} \qquad q_j(x) < t_\ell + |\mu| + |\nu_j| \leq t_\ell + \alpha.$$

Similarly, if $a_j = \mathrm{R}$, then

$$c_j = q_j(x) + \nu_j > \hat{t}_u = t_u - \mu, \qquad \text{whence} \qquad q_j(x) > t_u - (|\mu| + |\nu_j|) \geq t_u - \alpha.$$

Finally, if $a_j = \top$, then

$$c_j = q_j(x) + \nu_j \in [\hat{t}_\ell, \hat{t}_u] = [t_\ell + \mu, t_u - \mu], \qquad \text{whence} \qquad t_\ell - \alpha \leq q_j(x) \leq t_u + \alpha.$$

We now show that indeed $|\nu_j| + |\mu| \leq \alpha$ for every j with high probability. By tail bounds for the Laplace distribution,

$$\mathbb{P}\left[|\mu| > \alpha/4\right] = \exp\left(-\frac{\varepsilon\alpha n}{8}\right) \qquad \text{and} \qquad \mathbb{P}\left[|\nu_j| > 3\alpha/4\right] = \exp\left(-\frac{\varepsilon\alpha n}{8}\right)$$

for all j. By a union bound,

$$\mathbb{P}\left[|\mu| > \alpha/4 \quad \vee \quad \exists j \in [k] \;\; |\nu_j| > 3\alpha/4\right] \leq (k+1) \cdot \exp\left(-\frac{\varepsilon\alpha n}{8}\right) \leq \beta,$$

as required. □

7.5.3 The Online Interior Point Problem

Our algorithm extends a result of [BNSV15] showing how to reduce the problem of privately releasing thresholds to the much simpler *interior point problem*. By analogy, our algorithm for answering adaptively-chosen thresholds relies on solving multiple instances of an online variant of the interior point problem in parallel. In this section, we present the OIP problem and give an (ε, δ)-differentially private solution that can handle k adaptively-chosen queries with sample complexity $O(\log k)$. Our OIP algorithm is a direct application of the BetweenThresholds algorithm from Section

7.5.2.

Definition 7.5.14 (Online Interior Point Problem)**.** *An algorithm M solves the* Online Interior Point (OIP) Problem *for k queries with confidence β if, when given as input any private dataset $x \in [0,1]^n$ and any adaptively-chosen sequence of real numbers $y_1, \cdots, y_k \in [0,1]$, with probability at least $1-\beta$ it produces a sequence of answers $a_1, \cdots, a_k \in \{\mathrm{L}, \mathrm{R}\}$ such that*

$$\forall j \in \{1, 2, \cdots, k\} \qquad y_j < \min_{i \in [n]} x_i \implies a_j = \mathrm{L}, \qquad y_j \geq \max_{i \in [n]} x_i \implies a_j = \mathrm{R}\,.$$

(If $\min_{i\in[n]} x_i \leq y_j < \max_{i\in[n]} x_i$, then M may output either symbol L *or* R*.)*

Input: Dataset $x \in [0,1]^n$.
Initialise a BetweenThresholds instance (Figure 7.7) $\mathcal{B}$ on dataset x with thresholds $t_\ell = \frac{1}{3}$, $t_u = \frac{2}{3}$.
For $j = 1, 2, \cdots, k$:
  Receive query $y_j \in [0,1]$.
  If $\mathcal{B}$ already halted on some query q_{y^*}, output L if $y_j < y^*$ and output R if $y_j \geq y^*$.
  Otherwise, give $\mathcal{B}$ the query $c_{y_j} \in \mathsf{Thresh}_{[0,1]}$.
  If $\mathcal{B}$ returns $\top$, output R. Otherwise, output the answer produced by $\mathcal{B}$.

Figure 7.8: *Online Interior Point Algorithm*

Proposition 7.5.15. *The algorithm in Figure 7.8 is (ε, δ)-differentially private and solves the OIP Problem with confidence β as long as*

$$n \geq \frac{36}{\varepsilon}\left(\log(k+1) + \log(1/\beta) + \log(10/\varepsilon) + \log(1/\delta) + 1\right).$$

Proof. Privacy follows immediately from Lemma 7.5.9, since Algorithm 7.8 is obtained by post-processing Algorithm 7.7, run using thresholds with a gap of size $1/3$.

To argue utility, let $\alpha = 1/3$ so that

$$n \geq \frac{8}{\varepsilon\alpha}(\log(k+1) + \log(1/\beta)).$$

By Lemma 7.5.10, with probability at least $1 - \beta$, the following events occur:

- If the BetweenThresholds instance $\mathcal{B}$ halts when it is queried on c_{y^*}, then $\min_{i\in[n]} x_i \leq y^* < \max_{i\in[n]} x_i$.
- If $\mathcal{B}$ has not yet halted and $y_j < \min_{i\in[n]} x_i$, its answer to c_{y_j} is L.
- If $\mathcal{B}$ has not yet halted and $y_j \geq \max_{i\in[n]} x_i$, its answer to c_{y_j} is R.

Thus, if $\mathcal{B}$ has not yet halted, the answers provided are accurate answers for the OIP Problem. On the other hand, when $\mathcal{B}$ halts, it has successfully identified an "interior point" of the dataset x, i.e. a y^* such that $\min_{i\in[n]} x_i \leq y^* < \max_{i\in[n]} x_i$. Thus, for any subsequent query y, we have that

$$y < \min_{i\in[n]} x_i \implies y < y^*,$$

so Algorithm 7.8 correctly outputs L. Similarly,

$$y \geq \max_{i\in[n]} x_i \implies y \geq y^*,$$

so Algorithm 7.8 correctly outputs R on such a query. □

7.5.4 Releasing Adaptive Thresholds with Approximate Differential Privacy

We are now ready to state our reduction from releasing thresholds to solving the OIP Problem.

Theorem 7.5.16. *If there exists an (ε, δ)-differentially private algorithm solving the OIP problem for k queries with confidence $\alpha\beta/8$ and sample complexity n', then there is a $(4\varepsilon, (1+e^{\varepsilon})\delta)$-differentially private algorithm for releasing k threshold queries with (α, β)-accuracy and sample complexity*

$$n = \max\left\{\frac{6n'}{\alpha}, \frac{24\log^{2.5}(4/\alpha)\cdot\log(2/\beta)}{\alpha\varepsilon}\right\}.$$

Combining this reduction with our algorithm for the OIP Problem (Proposition 7.5.15) yields:

Corollary 7.5.17. *There is an (ε, δ)-differentially private algorithm for releasing k adaptively-chosen threshold queries with (α, β)-accuracy for*

$$n = O\left(\frac{\log k + \log^{2.5}(1/\alpha) + \log(1/\beta\varepsilon\delta)}{\alpha\varepsilon}\right).$$

Proof of Theorem 7.5.16. Our algorithm and its analysis follow the reduction of Bun et al. [BNSV15] for reducing the (offline) query release problem for thresholds to the offline interior point problem.

Let T be an (ε, δ)-differentially private algorithm solving the OIP Problem with confidence $\alpha\beta/8$ and sample complexity n'. Without loss of generality, we may assume that T is differentially private in "add-or-remove-an-item sense"—i.e. if $x \in [0,1]^*$ and x' differs from x up to the addition or removal of a single element, then for every adversary $\mathcal{A}$ and set S of outcomes of the interaction between $\mathcal{A}$ and T, we have $\mathbb{P}\left[\mathsf{Adaptive}_{\mathcal{A}\rightleftarrows T}(x) \in S\right] \leq e^{\varepsilon}\mathbb{P}\left[\mathsf{Adaptive}_{\mathcal{A}\rightleftarrows T}(x') \in S\right] + \delta$. Moreover, T provides accurate answers to the OIP Problem with probability at least $1 - \alpha\beta/8$ whenever its input is of size at least n'. To force an algorithm T to have these properties, we may pad any dataset of size less than n' with an arbitrary fixed element. On the other hand, we may subsample the first n' elements from any dataset with more than this many elements.

Consider the algorithm $\mathsf{AdaptiveThresholds}_T$ in Figures 7.9 and 7.10.

Input: Dataset $x \in [0,1]^n$.
Parameter: $\alpha \in (0,1)$.

Let $(x^{(1)}, \ldots, x^{(M)}) \leftarrow_{\mathrm{R}} \mathsf{Partition}(x_1, \ldots, x_n, \alpha)$.
Initialize an instance of the OIP algorithm $T^{(m)}$ on each chunk $x^{(m)} \in [0,1]^*$, for $m \in [M]$.
For each $j = 1, \cdots, k$:
  Receive query $c_{y_j} \in \mathsf{Thresh}_{[0,1]}$.
  Give query $y_j \in [0,1]$ to every OIP instance $T^{(m)}$, receiving answers $a_j^{(1)}, \cdots, a_j^{(M)} \in \{\mathrm{L}, \mathrm{R}\}$.
  Return $a_j = \frac{1}{M} \cdot \left| \left\{ m \in [M] : a_j^{(m)} = \mathrm{R} \right\} \right|$.

Figure 7.9: $\mathsf{AdaptiveThresholds}_T$

Input: Dataset $x \in [0,1]^n$.
Parameter: $\alpha \in (0,1)$.
Output: (Random) partition $(x^{(1)}, \ldots, x^{(M)}) \in ([0,1]^*)^M$ of x, where $2/\alpha \leq M < 4/\alpha$.

Let $M = 2^{\lceil \log_2(2/\alpha) \rceil}$.
Sort x in nondecreasing order $x_1 \leq x_2 \leq \cdots \leq x_n$.
For each $0 \leq \ell \leq \log_2 M$ and $s \in \{0,1\}^\ell$, sample $\nu_s \sim \mathrm{Lap}((\log_2 M)/\varepsilon)$ independently.
For each $1 \leq m \leq M-1$, let $\eta_m = \sum_{s \in P(m)} \nu_s$, where $P(m)$ is the set of all prefixes of the binary representation of m.
Let $t_0 = 1, t_1 = \left\lfloor \frac{n}{M} + \eta_1 \right\rfloor, \cdots, t_m = \left\lfloor \frac{m \cdot n}{M} + \eta_m \right\rfloor, \cdots, t_M = n+1$.
Let $x^{(m)} = (x_{t_{m-1}}, \ldots, x_{t_m - 1})$ for all $m \in [M]$.

Figure 7.10: Partition

The proof of Theorem 7.5.16 relies on the following two claims about the Partition subroutine, both of which are implicit in the work of Bun et al. [BNSV15, Appendix C] and are based on ideas of Dwork et al. [DNPR10]. Claim 7.5.18 shows that for neighboring databases $x \sim x'$, the behaviors of the Partition subroutine on x and x' are "similar" the following sense: for any fixed partition of x, one is roughly

as likely (over the randomness of the partition algorithm) to obtain a partition of x' that differs on at most two chunks, where the different chunks themselves differ only up to the addition or removal of a single item. This will allow us to show that running M parallel copies of the OIP algorithm on the chunks remains roughly (ε, δ)-differentially private. Claim 7.5.19 shows that, with high probability, each chunk is simultaneously large enough for the corresponding OIP algorithm to succeed, but also small enough so that treating all of the elements in a chunk as if they were the same element still permits us to get α-accurate answers to arbitrary threshold queries.

Claim 7.5.18. *Fix neighboring datasets $x, x' \in [0,1]^n$. Then there exists a (measurable) bijection $\varphi : \mathbb{R}^{2M} \to \mathbb{R}^{2M}$ with the following properties:*

1. *Let $z \in \mathbb{R}^{2M}$ be any noise vector. Let $x^{(1)}, \ldots, x^{(M)}$ denote the partition of x obtained with random noise set to $\nu = z$. Similarly, let $x'^{(1)}, \ldots, x'^{(M)}$ denote the partition of x' obtained under noise $\nu = \varphi(z)$. Then there exist indices i_1, i_2 such that: 1) For $i \in \{i_1, i_2\}$, the chunks $x^{(i)}$ and $x'^{(i)}$ differ up to the addition or removal of at most one item and 2) For every index $i \notin \{i_1, i_2\}$, we have $x^{(i)} = x'^{(i)}$.*

2. *For every noise vector $z \in \mathbb{R}^{2M}$, we have $\mathbb{P}[\nu = \varphi(z)] \leq e^{2\varepsilon}\mathbb{P}[\nu = z]$.*

Claim 7.5.19. *With probability at least $1 - \beta/2$, we have that $|t_m - m \cdot n/M| \leq \alpha n/24$ for all $m \in [M]$.*

Privacy of Algorithm 7.9. We first show how to use Claim 7.5.18 to show that Algorithm 7.9 is differentially private. Fix an adversary A, and let $B = \mathsf{Adaptive}_{A \rightleftarrows \mathsf{AdaptiveThresholds}_T}$ simulate the interaction between A and Algorithm 7.9. Let S be a subset of the range of B. Then, by Property (1) of Claim 7.5.18 and

group privacy, we have that for any $z \in \mathbb{R}^{2M}$:

$$\mathbb{P}\left[B(x) \in S | \nu = z\right] \leq e^{2\varepsilon}\mathbb{P}\left[B(x') \in S | \nu = \varphi(z)\right] + (1 + e^{\varepsilon})\delta.$$

By Property (2) of Claim 7.5.18, we also have $\Pr[\nu = z] \leq e^{2\varepsilon}\Pr[\nu = \varphi(z)]$ for every $z \in \mathbb{R}^{2M}$. Therefore,

$$\begin{aligned}
\mathbb{P}\left[B(x) \in S\right] &= \int_{\mathbb{R}^{2M}} \mathbb{P}\left[B(x) \in S | \nu = z\right] \cdot \mathbb{P}\left[\nu = z\right] \mathrm{d}z \\
&\leq \int_{\mathbb{R}^{2M}} \left(e^{2\varepsilon}\mathbb{P}\left[B(x') \in S | \nu = \varphi(z)\right] + (1 + e^{\varepsilon})\delta\right) \cdot \mathbb{P}\left[\nu = z\right] \mathrm{d}z \\
&\leq (1 + e^{\varepsilon})\delta + \int_{\mathbb{R}^{2M}} e^{2\varepsilon}\mathbb{P}\left[B(x') \in S | \nu = \varphi(z)\right] \cdot e^{2\varepsilon}\mathbb{P}\left[\nu = \varphi(z)\right] \mathrm{d}z \\
&\leq (1 + e^{\varepsilon})\delta + e^{4\varepsilon}\mathbb{P}\left[B(x') \in S\right].
\end{aligned}$$

Hence, B is $(e^{4\varepsilon}, (1 + e^{\varepsilon})\delta)$-differentially private, as claimed.

Accuracy of Algorithm 7.9. We now show how to use Claim 7.5.19 to show that Algorithm 7.9 produces (α, β)-accurate answers. By a union bound, the following three events occur with probability at least $1 - \beta$:

1. For all $m \in [M]$, $\left|\frac{m}{M} - \frac{t_m}{n}\right| \leq \frac{\alpha}{6}$.
2. Every chunk $x^{(m)}$ has size $|x^{(m)}| = t_m - t_{m-1} \in [\alpha n/6, 2\alpha n/3]$.
3. Every instance of T succeeds.

Now we need to show that if these three events occur, we can produce α-accurate answers to every threshold query $c_{y_1}, \ldots, c_{y_k}$. Write the sorted input database as $x_1 \leq x_2 \leq \cdots \leq x_n$. We consider two cases for the j^{th} query: As our first case, suppose $x_n \leq y_j$. Then for every chunk $x^{(m)}$, we have $\max\{x^{(m)}\} \leq y_j$. Then the success condition of $T^{(m)}$ guarantees that $a_j^{(m)} = \mathrm{R}$. Thus, the answer $a_j = 1$ is (exactly) accurate for the query c_j.

As our second case, let i be the smallest index for which $x_i > y_j$, and suppose the item x_i is in some chunk $x^{(m_i)}$. Note that this means that the true answer to the query c_{y_j} is $(i-1)/n$ and that $t_{m_i-1} \leq i \leq t_{m_i} - 1$. Then again, for every $m < m_i$ we have $\max\{x^{(m)}\} \leq y_j$, so every such $T^{(m)}$ instance yields $a_j^{(m)} = \mathrm{R}$. Thus,

$$a_j = \frac{1}{M} \cdot \left| \left\{ m \in [M] : a_j^{(m)} = \mathrm{R} \right\} \right| \geq \frac{m_i - 1}{M} \geq \frac{t_{m_i}}{n} - \frac{\alpha}{6} - \frac{\alpha}{2} \geq \frac{(i-1)}{n} - \alpha,$$

since $M \geq 2/\alpha$.

On the other hand, for every $m > m_i$, we have $\min\{x^{(m)}\} > y_j$, so every such $T^{(m)}$ instance instead yields $a_j^{(m)} = \mathrm{L}$.

$$a_j \leq \frac{m_i}{M} \leq \frac{t_{m_i}}{n} + \frac{\alpha}{6} \leq \frac{t_{m_i-1} + 2\alpha n/3}{n} + \frac{\alpha}{6} \leq \frac{i}{n} + \frac{2\alpha}{3} + \frac{\alpha}{6} \leq \frac{i-1}{n} + \alpha,$$

since $n \geq 6/\alpha$.

□

Chapter 8

Conclusion

This thesis has presented a number of results about privacy and adaptivity in algorithmic data analysis. The unifying theme of these results is understanding the information-theoretic relationship between the input and output of randomised algorithms. For the upper bounds, differential privacy entails a stability condition that protects privacy and ensures generalisation — changing a single input point should not change the probability distribution of the output much. For the lower bounds, fingerprinting techniques show that the output of any accurate algorithm must "correlate" with its input, which means the output reveals information about the input. Our results illustrate how versatile these two tools are.

Differential privacy was only defined a decade ago, yet a rich literature has developed around it. Many fundamental questions remain to be resolved — even the definition itself is not entirely settled, as Chapter 2 demonstrates. Of particular interest is the development of practical differentially private tools. Our results in Chapters 2 and 4 are geared towards a fine-grained analysis of the power of differential privacy, which we believe will enhance the applicability of differential privacy.

Fingerprinting is discussed in four chapters (§4,§5,§6,§7). These are three different applications with different analyses. First, fingerprinting is used to prove lower bounds for differential privacy; here we have presented the slickest possible proof. Second, fingerprinting is used to analyse practical privacy attacks; here the goal is to extend the fingerprinting analysis to the broadest class of distributions (as we do not have the freedom to choose the distribution with real-world data). Thirdly, fingerprinting codes are constructed in the original cryptographic context; here the challenge is carrying out the analysis when we only have a weak accuracy guarantee. Remarkably, in all three settings we are able to follow the same proof outline. In what other settings can the fingerprinting analysis be deployed?

The connection between differential privacy and generalisation is unexpectedly tight in the adaptive setting. Not only is there a direct connection from differential privacy to generalisation [DFH+15c, §3], but also negative results for privacy can be extended to yield negative results for generalisation in adaptive data analysis [HU14, §6]. This application provides further impetus to differential privacy research and also provides a new perspective on various results about differential privacy, which will hopefully result in more cross-fertilisation of ideas between differential privacy other fields such as machine learning and statistics.

This thesis points out several open problems and we believe that their resolution will provide further insight into the challenges inherent in algorithmic data analysis

References

[AS64] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1964.

[BDMN05] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the SuLQ framework. In *Proceedings of the Twenty-fourth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 13-15, 2005, Baltimore, Maryland, USA*, pages 128–138, 2005.

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[BH15] Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. *CoRR*, abs/1502.04585, 2015.

[BLR13] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. *J. ACM*, 60(2):12, 2013.

[BN08] Dan Boneh and Moni Naor. Traitor tracing with constant size ciphertext. In *CCS*, pages 501–510. ACM, Oct 27–31 2008.

[BNS13] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 16th International Workshop, APPROX 2013, and 17th International Workshop, RANDOM 2013, Berkeley, CA, USA, August 21-23, 2013. Proceedings*, pages 363–378, 2013.

[BNS+16a] Raef Bassily, Kobbi Nissim, Adam Smith, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *ACM Symposium on the Theory of Computing (STOC)*, 2016.

[BNS16b] Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, pages 369–380, New York, NY, USA, 2016. ACM.

[BNSV15] Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 634–649. IEEE, 2015.

[Bon36] Carlo Emilio Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubbl. d. R. Ist. Super. di Sci. Econom. e Commerciali di Firenze.*, 8, 1936.

[BRS+09] Rosemary Braun, William Rowe, Carl Schaefer, Jinghui Zhang, and Kenneth Buetow. Needles in the haystack: identifying individuals present in pooled genomic data. *PLoS genetics*, 5(10):e1000668, 2009.

[BS98] Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44(5):1897–1905, 1998.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *TCC 2016-B*, 2016. http://arxiv.org/abs/1605.02065.

[BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *FOCS*, pages 464–473. IEEE, October 18–21 2014.

[BSU16] Mark Bun, Thomas Steinke, and Jonathan Ullman. Make up your mind: The price of online queries in differential privacy. *CoRR*, abs/1604.04618, 2016.

[BUV14] Mark Bun, Jonathan Ullman, and Salil P. Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 1–10, 2014.

[BZ06] Michael Barbaro and Tom Zeller. A face is exposed for aol searcher no. 4417749. *New York Times*, August 2006.

[CFN94] Benny Chor, Amos Fiat, and Moni Naor. Tracing traitors. In *CRYPTO*, pages 257–270. Springer, August 21-25 1994.

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

[Coo09] John D Cook. Upper and lower bounds for the normal distribution function, 2009.

[CSS11] T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Trans. Inf. Syst. Secur.*, 14(3):26, 2011.

[De12] Anindya De. Lower bounds in differential privacy. In *Proceedings of the 9th International Conference on Theory of Cryptography*, TCC'12, pages 321–338, Berlin, Heidelberg, 2012. Springer-Verlag.

[DFH+15a] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems (NIPS)*, Montreal, December 2015.

[DFH+15b] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, June 2015.

[DFH+15c] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 117–126, 2015.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings*, pages 486–503, 2006.

[DL09] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pages 371–380, 2009.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, pages 265–284, 2006.

[DMT07] 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.

[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA*, pages 202–210, 2003.

[DN04] Cynthia Dwork and Kobbi Nissim. Privacy-preserving datamining on vertically partitioned databases. In *Advances in Cryptology - CRYPTO 2004, 24th Annual International CryptologyConference, Santa Barbara, California, USA, August 15-19, 2004, Proceedings*, pages 528–544, 2004.

[DNPR10] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential privacy under continual observation. In *Symposium on Theory of Computing (STOC)*, pages 715–724. ACM, 2010.

[DNR+09] Cynthia Dwork, Moni Naor, Omer Reingold, Guy N. Rothblum, and Salil P. Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *STOC*, pages 381–390. ACM, May 31 - June 2 2009.

[DNRR15] Cynthia Dwork, Moni Naor, Omer Reingold, and Guy N. Rothblum. Pure differential privacy for rectangle queries via private partitions. In *Advances in Cryptology - ASIACRYPT 2015 - 21st International Conference on the Theory and Application of Cryptology and Information Security, Auckland, New Zealand, November 29 - December 3, 2015, Proceedings, Part II*, pages 735–751, 2015.

[DNT14] Cynthia Dwork, Aleksandar Nikolov, and Kunal Talwar. Efficient algorithms for privately releasing marginals via convex relaxations. In *Symposium on Computational Geometry–SoCG*, 2014.

[DR14] Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9. Foundations and Trends® in Theoretical Computer Sc, 2014.

[DR16] Cynthia Dwork and Guy Rothblum. Concentrated differential privacy. *CoRR*, abs/1603.01887, 2016.

[DRV10] Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *IEEE Symposium on Foundations of Computer Science (FOCS '10)*, pages 51–60. IEEE, 23–26 October 2010.

[DSS+15] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 650–669, Oct 2015.

[DSSU17] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4, 2017.

[DTTZ14] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Symposium on Theory of Computing STOC*, pages 11–20. ACM, May 31–June 3 2014.

[Due10] L. Duembgen. Bounding Standard Gaussian Tail Probabilities. *ArXiv e-prints*, December 2010.

[Dun61] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56:52–64, 1961.

[DW79a] Luc Devroye and Terry J. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979.

[DW79b] Luc Devroye and Terry J. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.

[DY08] Cynthia Dwork and Sergey Yekhanin. New efficient attacks on statistical disclosure control mechanisms. In *Advances in Cryptology - CRYPTO 2008, 28th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 17-21, 2008. Proceedings*, pages 469–480, 2008.

[EN14] Yaniv Erlich and Arvind Narayanan. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, 15(6):409–421, 2014.

[Ete85] N. Etemadi. On some classical results in probability theory. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 47(2):pp. 215–221, 1985.

[FMN13] Nadia Fawaz, S. Muthukrishnan, and Aleksandar Nikolov. Nearly optimal private convolution. In *Algorithms - ESA 2013 - 21st Annual European Symposium, Sophia Antipolis, France, September 2-4, 2013. Proceedings*, 2013.

[FT01] Amos Fiat and Tamir Tassa. Dynamic traitor tracing. *J. Cryptology*, 14(3):211–223, 2001.

[GL14] Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102(6):460, 2014.

[hh] hbp (http://math.stackexchange.com/users/131476/hbp). Hyperbolic trig inequality. Mathematics Stack Exchange. URL:http://math.stackexchange.com/q/1461426 (version: 2015-10-02).

[HR10] Moritz Hardt and Guy Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. pages 61–70, 2010.

[HSR+08] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

[HT10] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, STOC '10, pages 705–714, New York, NY, USA, 2010. ACM.

[HU14] Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *FOCS*. IEEE, October 19-21 2014.

[IGNC12] Hae Kyung Im, Eric R Gamazon, Dan L Nicolae, and Nancy J Cox. On sharing quantitative trait gwas results in an era of multiple-omics data and the limits of genomic privacy. *The American Journal of Human Genetics*, 90(4):591–598, 2012.

[Ioa05] John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):124, August 2005.

[JYW+09] Kevin B Jacobs, Meredith Yeager, Sholom Wacholder, David Craig, Peter Kraft, David J Hunter, Justin Paschal, Teri A Manolio, Margaret Tucker, Robert N Hoover, , Gilles D Thomas, Stephen J Chanock, and Nilanjan Chaterjee. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature genetics*, 41(11):1253–1257, 2009.

[Kea93] Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. In *STOC*, pages 392–401. ACM, May 16-18 1993.

[KOV15] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1376–1385, 2015.

[KR99] Michael J. Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.

[KRS13] Shiva Prasad Kasiviswanathan, Mark Rudelson, and Adam Smith. The power of linear reconstruction attacks. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2013.

[KRSU10] Shiva Prasad Kasiviswanathan, Mark Rudelson, Adam Smith, and Jonathan Ullman. The price of privately releasing contingency tables and the spectra of random matrices with correlated rows. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 775–784, 2010.

[LDR+13] T. Laarhoven, J. Doumen, P. Roelse, B. Skoric, and B. de Weger. Dynamic tardos traitor tracing schemes. *Information Theory, IEEE Transactions on*, 59(7):4230–4242, July 2013.

[McS14] Frank McSherry. Differential privacy for measure concentration. Blog post on "Windows on Theory". http://windowsontheory.org/2014/02/04/differential-privacy-for-measure-concentration/, 2014.

[MMP+10] Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil P. Vadhan. The limits of two-party differential privacy. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 81–90, 2010.

[MN12] S. Muthukrishnan and Aleksandar Nikolov. Optimal private halfspace counting via discrepancy. In *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*, pages 1285–1292, 2012.

[MT07] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS '07. 48th Annual IEEE Symposium on*, pages 94–103, Oct 2007.

[MV16] Jack Murtagh and Salil P. Vadhan. The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography - 13th International Conference, TCC 2016-A, Tel Aviv, Israel, January 10-13, 2016, Proceedings, Part I*, pages 157–175, 2016.

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.

[NTZ13] Aleksandar Nikolov, Kunal Talwar, and Li Zhang. The geometry of differential privacy: the sparse and approximate cases. *STOC*, 2013.

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[Rén61] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561, Berkeley, Calif., 1961. University of California Press.

[Riv12] Omar Rivasplata. Subgaussian random variables: An expository note, 2012. http://www.stat.cmu.edu/ arinaldo/36788/subgaussians.pdf.

[RR10] Aaron Roth and Tim Roughgarden. Interactive privacy via the median mechanism. In *Proc. 42nd Symposium on Theory of Computing (STOC)*, pages 765–774. ACM, 2010.

[SOJH09] Sriram Sankararaman, Guillaume Obozinski, Michael I Jordan, and Eran Halperin. Genomic privacy and limits of individual detection in a pool. *Nature genetics*, 41(9):965–967, 2009.

[SSSS10] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

[SU15a] Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *CoRR*, abs/1501.06095, 2015.

[SU15b] Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 1588–1628, 2015.

[Swe97] Latanya Sweeney. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110, 1997.

[Tar08] Gábor Tardos. Optimal probabilistic fingerprint codes. *J. ACM*, 55(2), 2008.

[Tas05] Tamir Tassa. Low bandwidth dynamic traitor tracing schemes. *J. Cryptology*, 18(2):167–183, 2005.

[Ull13] Jonathan Ullman. Answering n {2+ o (1)} counting queries with differential privacy is hard. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 361–370. ACM, 2013.

[Ull15] Jonathan Ullman. Private multiplicative weights beyond linear queries. In *PODS*. ACM, May 31–June 4 2015.

[vEH14] T. van Erven and P. Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, July 2014.

[VH09] Peter M Visscher and William G Hill. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS genetics*, 5(10):e1000628, 2009.

[War65] Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965. PMID: 12261830.

[WLW+09] Rui Wang, Yong Fuga Li, Xiao Feng Wang, Haixu Tang, and Xiao Yong Zhou. Learning your identity and disease from research papers: information leaks in genome wide association study. In *ACM Conference on Computer and Communications Security*, pages 534–544. ACM, 2009.

[ZPL+11] Xiaoyong Zhou, Bo Peng, Yong Fuga Li, Yangyi Chen, Haixu Tang, and XiaoFeng Wang. To release or not to release: evaluating information leaks in aggregate human-genome data. In *Computer Security–ESORICS 2011*, pages 607–627. Springer, 2011.

This thesis incorporates text from the following papers. [BS16, BNS+16a, SU15a, DSS+15, SU15b, BSU16, DSSU17]