Sample-Complexity Optimality Under Local Differential Privacy and Related Models

by

Alex Edmonds

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Department of Computer Science
University of Toronto

Sample-Complexity Optimality Under Local Differential Privacy and Related Models

Alex Edmonds
Doctor of Philosophy

Department of Computer Science
University of Toronto
2023

Abstract

This work explores the sample complexity of estimation and learning tasks under local differential privacy (LDP), an especially strong approach to differential privacy which does not rely on any trusted party to guarantee privacy. We give a new characterization of the sample complexity of answering statistical (linear) queries under non-interactive LDP in terms of the γ_2 norm, showing that a slight generalization of the well-studied factorization mechanism achieves, in polynomial time, nearly optimal sample complexity for each specific workload of queries. These results are obtained by leveraging information theoretic bounds for LDP and applying them together with geometric techniques which allow us to construct hard distributions for answering statistical queries. We extend these ideas to obtain characterizations of both agnostic learning and and agnostic refutation under non-interactive LDP in terms of the γ_2 norm, and derive, by consequence, a sample-complexity equivalence between the two tasks. We also give a characterization of realizable refutation in terms of a factorization norm which we define. A matching upper boud is given for realizable refutation, implying that realizable refutability implies realizable learning.

We explore other techniques for obtaining characterizations of local privacy tasks, including an approach which combines a mutual information bound for differential privacy with an information complexity lower bound borrowed from communication theory.

We also consider relationships between local privacy and data-access models which turn out to be closely related. In particular, we demonstrate an equivalence between the sample complexity of learning under sequential LDP and the sample complexity of learning under single-intrusion pan-privacy. Finally, we demonstrate an equivalence between the sample complexity of agnostic learning under non-interactive LDP and the query complexity of agnostic learning under the adaptive correlational statistical query model.

Acknowledgements

I would like to extend my deeply felt gratitude to my supervisors, Toni (Toniann) Pitassi and Sasho (Aleksandar) Nikolov, for being compassionate and dedicated mentors, always patient, uplifiting and extraordinary inspirations to me. Thank you also, Jonathan Ullman, our invaluble collaborator.

I would like to thank my family. To my dad, Jack Edmonds, my mom, Kathie Cameron, my brother, Jeff Edmonds, you paved the path that I follow. You taught me the beauty of mathematics in its simplicity, its elegance and its depth. Thank you for your love, your enthusiasism and your belief in me. My sister Laura, thank you for being my emotional rock, always ready to listen and offer guidance, and for showing us by your example that the road not taken can make all the difference.

Thank you to the friends and family who will remain unnamed. You know who you are. I would not be who I am without you and I could not have gotten here without your love and support.

Finally, thank you to my peers in the theory group who created an atmosphere of friendship and intellectual creativity, which was such a joy to be part of.

Contents

Chapter 1

Introduction

1.1 Differential privacy

Increasingly, in a wide range of contexts such as medicine, censuses and social media, there exist rich data sets whose usefulness for purposes of data analysis is compromised by the danger of revealing sensitive private information of individuals. Privacy breaches may threaten individuals personally, socially or economically. Moreover, the risk of privacy breaches can even pose a danger to the validity of a study itself if it discourages individuals from answering truthfully or participating in the first place.

Naive approaches to protecting privacy in the context of data analysis often fail. For instance, removing obvious personal identifiers such as name and phone number is not effective, since a sufficient amount of information of almost any kind about an individual is unique to that individual and may be used to identify them. Indeed, linkage attacks exploit this idea by taking advantage of publicly available information about individuals and linking it to their profile – intended to be anonymized – in a data set. An example of this is the Netflix de-anonymization attack [CT19] which recovered the identities behind profiles in an Netflix movie ratings data set by linking them to public Internet Movie Database (IMDb) profiles. Only publishing aggregate information is also not sufficient. For instance, a simple differencing attack may look at a sum on a data set before and after an individual is removed from it. Taking the difference of the two values reveals the amount which the individual contribute to the sum. Meanwhile, even releasing noisy sums or averages can reveal private information if the noise is not appropriately scaled with the queries being answered and the size of the data set [DN03].

Privacy techniques dealing with such attacks on a case-by-case basis may still be susceptible to attacks of unknown kinds. Instead, we want privacy techniques which provide guarantees even against unforeseen attacks. *Differential privacy* (DP) [DMNS06] provides a rigorous mathematical framework for such guarantees. To satisfy differential privacy, the probability of any output of the mechanism should be affected only slightly when the input of a single individual is changed. Formally, for $\varepsilon > 0$, a randomized function $\mathcal{M} : \mathcal{X}^n \to \mathcal{Z}$ is called *ε-differentially private* (ε-DP) if, when $\overline{x}, \overline{x}' \in \overline{x}^n$ disagree on at most one entry, then, for all $S \subseteq \mathcal{Z}$,

$$\mathop{\mathbb{P}}_{\mathcal{M}}[\mathcal{M}(\overline{x}) \in S] \le e^{\varepsilon} \cdot \mathop{\mathbb{P}}_{\mathcal{M}}[\mathcal{M}(\overline{x}') \in S].$$

The symmetry of the definition also implies

$$\underset{\mathcal{M}}{\mathbb{P}}[\mathcal{M}(\overline{x}) \in S] \geq e^{-\varepsilon} \cdot \underset{\mathcal{M}}{\mathbb{P}}[\mathcal{M}(\overline{x}') \in S].$$

By the post-processing property of DP, any function applied on top of an ε-DP function is itself ε-DP. This implies that the probability of any consequence for an individual is affected by only a small amount by their choice to participate in a differentially private protocol. This is the essential promise of differential privacy to the individual participant – "you will be at most negligibly affected, for good or ill, by your choice to participate."

The strategic application of randomness to the output is essential in the construction of differentially private algorithms [DN03]. This introduces an inherent trade-off between accuracy and privacy parameters. A core objective in the study of DP, and in this work in particular, is to elucidate the optimal trade-off between accuracy and privacy parameters for given tasks. A core objective in this work will be to characterize, for various estimation and learning problems, the complexity of the task under local privacy.

In addition to a rich academic literature, differential privacy is now being deployed on a large scale by Apple [App17], Google [EPK14, BEM$^+$17, WZL$^+$19], Uber [JNS18], and the US Census Bureau [DLS$^+$17].

Different models of differential privacy may be distinguished by the type of adversary they protect against. *Central differential privacy* assumes there is a trusted central curator holding the sensitive data of all individuals. This central curator releases an output with appropriate randomness introduced so as to guarantee privacy. Privacy is guaranteed against an external adversary, but not against the curator.

By contrast, this work will focus on *local differential privacy* (LDP) [EGS03a, DMNS06, KLN$^+$08], an especially strong model of privacy which eschews a trusted curator and instead relies only on an untrusted aggregator to collect already private information from individuals and report on it in a useful way. Even the aggregator is unable to compromise privacy. Such protocols are local in the sense that each individual, referred to as a *local agent*, is responsible for applying an ε-DP function to their own data point before sending the result to the aggregator. Local differential privacy provides the strongest possible type of privacy guarantee, since there is no need for the local agent to trust the party collecting their data. This handles an essential concern with regard to the real-world application of the central model, where data stewards themselves may pose the most significant threat to the privacy of individuals. However, while the local model protects against a particularly strong adversarial model, it is also more restrictive than the central model. Approaches to central differential privacy do not always readily translate to the local setting. For this reason, it is essential to understand the complexity of performing relevant tasks under LDP.

We will consider a taxonomy of LDP protocols according to the type of *interactivity* involved. *Non-interactive* LDP allows each local agent to speak once, and their output is not allowed to depend on the outputs of the other local agents. *Sequential* LDP allows each local agent to speak once in sequence, and their output is allowed to depend on the previous outputs of the other local agents. Finally, general *interactive* LDP protocols also proceed in a series of sequential rounds. Each agent speaks once per round and their output is allowed to depend on any of the messages already sent. The entire transcript of the local agents' communication should be differentially private as a function of the inputs. We are interested in understanding the relative power of different types of interactivity.

We also extend ideas we apply to local differential privacy to characterize the complexity of learning and estimation tasks to models where the underlying distribution is accessed in other ways. These other models include: pan-privacy, where a streaming algorithm is guaranteed to preserve privacy even in the case where its internal state is compromised; the correlational statistical query model, where statistical queries on an underlying distribution of labelled samples each ask the correlation of the labels with a real function.

1.2 Results

1.2.1 Sample-complexity lower bounds against LDP from information-complexity lower bounds

The first lower bound we present is due to collaborator Jonathan Ullman, and presented here with his permission. The lower bound is against general interactive LDP, and takes advantage of known information-complexity lower bounds for 2-party communication. This is leveraged against a generic mutual-information upper bound for LDP, which bounds the mutual information between the transcript of the protocol and the input data set [MMP+10, SZ20]. This technique will be applied to the selection problem, where each local agent i holds a data point x_i drawn independently from an underlying distribution λ on $\{0,1\}^d$ and the goal of the protocol is to identify a coordinate $j \in [d]$ which satisfies

$$\mu_j \geq \max_{k \in [d]} \mu_k - \alpha,$$

where $\alpha > 0$ is the accuracy parameter and each μ_k denotes the expectation $\mathop{\mathbb{E}}_{x \sim \lambda}[x_k]$ of coordinate k. To derive our lower bound for selection, we show that an arbitrary LDP protocol for selection can be translated into a 2-party communication protocol for set disjointness. This allows us to apply the known information complexity lower bound for 2-party set-disjointness [BJKS04].

Although our lower bound for selection under LDP has been shown via other techniques [DJW18], our reduction of LDP sample-complexity lower bounds to information complexity lower bounds is presented here for the sake of both its novelty and simplicity.

1.2.2 Characterization of statistical query release under non-interactive LDP

One of the principal tasks we consider for LDP in this work is answering statistical queries. The simplest example of a statistical query is “What fraction of individuals in the data set have property P?” More generally, a statistical query is given by some $q : \mathcal{X} \to \mathbb{R}$ and answering it means approximating its average $\frac{1}{n}\sum_{i \in [n]} q(x_i)$ on the data set $\overline{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$. In the context of LDP, each x_i is held by a distinct local agent. We are often interested in answering each query in a *workload* $Q = \{q_1, \ldots, q_k\}$ of statistical queries. For instance, when $\mathcal{X} = [T]$, then the *threshold queries* are given, for $t \in [T]$, by

$$q_t(x) = \mathbb{I}[x \leq t]. \tag{1.1}$$

Answering the workload of threshold queries corresponds to computing the CDF of our data set. Indeed, answering a workload of statistical queries captures a variety of statistical tasks: computing histograms and PDFs, answering range queries and computing CDFs, estimating mean, computing correlations and higher-order marginals, and estimating the risk of a classifier. Our interest is in characterizing, for an arbitrary given workload, the optimal sample complexity required to achieve given accuracy and privacy parameters. We give a new algorithm for the problem of answering a workload of statistical queries and we give a lower bound which shows that this algorithm is nearly the best possible in terms of sample complexity.

Previously, the optimal sample complexity for answering a worst-case workload of statistical queries had been well understood [DN03], with known bounds being essentially tight as a function of the data set size, the data domain, and the size of the workload. Such techniques effectively treat each query independently. However, with certain workloads, such as those corresponding to computing PDFs or CDFs, information used to answer one query may be useful in answering another. Such structured workloads may be answered with considerably fewer samples. This work demonstrates that factorization techniques give sample complexity guarantees which are nearly optimal relative to an arbitrary given query workload.

It is useful to represent the workload Q by a matrix $W \in \mathbb{R}^{Q \times \mathcal{X}}$ with entries $w_{q,x} = q(x)$. Representing the data set $\overline{x} = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ by its histogram vector $h \in \mathbb{Z}_{\geq 0}^{\mathcal{X}}$ with entries given by $h_x = |\{i \, : \, x_i = x\}|$, then the vector of query answers is given by $\frac{1}{n} Wh$. Standard techniques for LDP would have each agent i, holding a sample x_i, report the vector $(q_1(x_i), \ldots, q_k(x_i))$ with added noise drawn from an appropriate distribution. Aggregating the outputs of all agents by taking the average gives $\frac{1}{n}Wh + G$ where G is a random noise vector with mean zero. For a worst-case workload Q, this approach, requiring

$$n = O\left(\frac{\|W\|_{1\to 2} \log |Q|}{\alpha^2 \varepsilon^2}\right)$$

samples to approximate each query within α under ε-LDP, will have optimal sample complexity. The quantity $\|W\|_{1\to 2}$ in the upper bound corresponds to the ℓ_2-sensitivity of the workload, given by the maximum ℓ_2 norm of a column of W. However, for certain workloads, it is possible to do better. For example, applying the previous mechanism to threshold queries as given by (1.1) would result in a sample guarantee of

$$n = O\left(\frac{\sqrt{T} \log T}{\alpha^2 \varepsilon^2}\right).$$

However, rather than treat each query independently, we can recycle the same information to answer multiple queries. Specifically, consider the interval queries given by

$$q_{s:t}(x) = \mathbb{I}[s \leq x \leq t].$$

Instead of answering the threshold queries directly, we can instead answer the following *dyadic* *queries* (when T is a power of 2):

$$\begin{aligned}
&q_{1:T/2}, q_{T/2+1,T}\\
&q_{1:T/4}, q_{T/4+1,2T/4}, \cdots q_{3T/4+1,T}\\
&\vdots\\
&q_{1:1}, q_{2,2}, \cdots q_{T,T}
\end{aligned}$$

Each of our threshold queries may be expressed as a linear combination of these dyadic queries. For instance, $q_7 = q_{1:4} + q_{5:6} + q_{7:7}$. Answering the dyadic queries and using the answers to reconstruct the answer to our threshold queries gives an algorithm where the number of samples required is bounded by

$$n = O\left(\frac{\log^3 T}{\alpha^2 \varepsilon^2}\right).$$

Generalizing this approach to an arbitrary workload Q with workload matrix W, we consider a factorization $W = RA$ of the query matrix. The matrix A itself represents a workload of queries so, by the same technique, we may obtain $Ah + G$. Multiplying the result by R reconstructs an approximation $RAh + RG = Wh + RG$ to the original workload. The number of samples which this procedure requires to achieve given accuracy and privacy parameters is affected by the choice of A since it represents the initial workload being answered. The number of samples required is also affected by the choice of R, since it determines how the noise G is scaled. Optimizing the choices of R and A gives sample complexity bounds in terms of $\gamma_2(W)$, a well-studied matrix norm which will be defined later on. This *factorization* technique, discovered in [LHR+10], is given a slight generalization here called *approximate factorization* which considers a matrix $\widetilde{W}$ close to W in ℓ_∞ norm and applies factorization to $\widetilde{W}$ rather than W. Optimizing the choice of $\widetilde{W}$ gives a sample complexity upper bound in terms of the quantity $\gamma_2(W, \alpha)$, the *approximate* γ_2 *norm*, defined later. The primary contribution of this work in the context of answering statistical queries is the derivation of lower bound as a function $\gamma_2(W, \alpha)$ which, for any given workload, is nearly tight. The lower bound implies that the approximate factorization mechanism is nearly optimal in sample complexity for the given workload. We apply our result to obtain new lower bounds for the following well studied families of queries:

1. *Threshold queries*, which are equivalent to computing the CDF of the data;

2. *Parity queries*, which capture covariance and higher-order moments of the data;

3. *Marginal queries*, also known as conjunctions, which capture the marginal distribution on subsets of the attributes.

This lower bound is obtained via information theoretic techniques. In particular, we introduce a new KL-divergence bound for non-interactive LDP, similar to one for sequential LDP previously discovered in [DJW13]. While our KL-divergence bound is restricted to non-interactive LDP, it is more general for that setting in a way that our proof depends upon.

The research which appears in this chapter is joint work with Aleksandar Nikolov and Jonathan Ullman, and was originally published in [ENU19].

1.2.3 Characterization of agnostic learning under non-interactive LDP

For the *agnostic learning* problem, each sample (a_i, b_i) is drawn independently from an arbitrary unknown distribution λ on $\mathcal{U} \times \{-1, 1\}$. In the context of LDP, each sample (a_i, b_i) is held by a local agent. The goal is to produce some hypothesis $h \subseteq \{-1, 1\}^{\mathcal{X}}$ with small loss, defined by

$$L_\lambda(h) = \underset{(x,b)\sim\lambda}{\mathbb{P}}[h(x) \neq b].$$

In particular, the hypothesis should have loss nearly as small as any concept in the concept class, so that

$$L_\lambda(h) \leq \min_{c\in\mathcal{C}} L_\lambda(c) + \alpha,$$

where $\alpha > 0$ is the accuracy parameter. The empirical loss of h on a data set $\overline{x} \in (\mathcal{X} \times \{-1, 1\})^n$ is defined by

$$L_{\overline{x}}(h) = \underset{(x,b)\sim\mathrm{Unif}(\overline{x})}{\mathbb{P}}[h(x) \neq b]$$

where Unif($\overline{x}$) is the uniform distribution on $\overline{x}$. We may express $L_{\overline{x}}(h)$ as a function of the statistical query given by $q_h(x, b) = y \cdot h(x)$. Specifically,

$$L_{\overline{x}}(h) = \frac{1}{2} - \frac{1}{2} \cdot \underset{(x,b)\sim\lambda}{\mathbb{E}}[q_h(x, b)].$$

This provides a straightforward approach to learning by answering the statistical query q_c for each $c \in \mathcal{C}$. Assuming uniform convergence, the empirical loss will be close to the true loss and so, by returning a concept c which nearly maximizes the answer to q_c, we learn the concept class. The workload $Q_{\mathcal{C}} = \{q_c\}_{c\in\mathcal{C}}$ may be answered by applying the approximate factorization mechanism, giving an upper bound in terms of the quantity $\gamma_2(W, \alpha)$ where $W \in \mathbb{R}^{\mathcal{C}\times\mathcal{X}}$ has entries $w_{c,x} = c(x)$.

Ultimately, we will show that this approach to agnostic learning is nearly optimal under LDP. However, the lower bound against statistical query release does not immediately apply to the agnostic learning problem. It does apply to the *refutation* version of the agnostic learning problem where the learner is required only to distinguish between the cases

$$\min_{c\in\mathcal{C}} L_\lambda(c) \leq \frac{1}{2} - \alpha$$

versus

$$\forall c \in \mathcal{C},\ L_\lambda(c) = \frac{1}{2}.$$

In most settings other than our own, agnostic learning enables solving the refutation task, by estimating the loss of the learned concept. In our setting however, estimating the loss involves introducing a round of interactivity into our algorithm.

We introduce additional machinery which deals with this issue, and obtain nearly tight upper and lower bounds for the general agnostic learning problem in terms of the approximate γ_2 norm. This result implies an equivalence between the sample complexities of agnostic learning and agnostic refutation under non-interactive LDP.

The upper bounds for both learning and refutation are joint work with Aleksandar Nikolov and Jonathan Ullman, originally published in [ENU19]. The lower bound against learning is joint work with Aleksandar Nikolov and Toniann Pitassi, originally published in [ENP22].

1.2.4 Characterization of realizable refutation under non-interactive LDP

Realizable learning is a special case of the agnostic learning problem where the underlying distribution λ on $\mathcal{X} \times \{-1, 1\}$ is guaranteed to be *realized* by some concept $c \in \mathcal{C}$, specifically $L_\lambda(c) = 0$. Since it is a special case, our previous learning algorithm may be applied. However, the 'hard distributions' used in the proof of that lower bound are not necessarily realizable.

Previously, lower bounds were obtained against realizable learning of linear separators, in a work which exploited the fact that learnability of a concept class by an SQ algorithm implied an upper bound on its margin complexity [DF19]. This suggested the possibility that margin complexity might characterize realizable learnability under non-interactive LDP, but this was later falsified by [DF20] which showed that even a large-margin linear classifier could not be learned efficiently under LDP without interactivity.

Our work will fill in this gap by introducing a measure of the complexity of a concept class, which is used to give an upper bound on the sample complexity of learning the concept class $\mathcal{C}$ in the realizable case under non-interactive LDP. The LDP protocol which witnesses the upper bound is also able to recognize when none of the concepts fit the underlying distribution. In particular, it also solves *realizable refutation* where the algorithm is required to distinguish between the cases $L_\lambda(c) = 0$ versus $\min_{c \in \mathcal{C}} L_\lambda(c) \geq \alpha$. By applying techniques analogous to those we use to get lower bounds against agnostic learning, we give a nearly tight lower bound for realizable refutation under non-interactive LDP. The problem of getting a lower bound against non-interactive realizable learners when they are not required to solve the refutation problem is left open.

This chapter is joint work with Aleksandar Nikolov and Toniann Pitassi, originally published in [ENP22].

1.2.5 Characterization of CSQ learning in terms of γ_2 norm

A classic work [KLN08] on differentially private learning established the close connection between learning under the *statistical query (SQ) model* of machine learning and learning under LDP. Under the statistical query model, instead of accessing the underlying distribution λ on $\mathcal{X} \times \{-1, 1\}$ via i.i.d. samples, the algorithm may pose queries where each is determined by some function $q : \mathcal{X} \times \{-1, 1\} \to [-1, 1]$. In response, it receives $r \in [-1, 1]$ which satisfies

$$\left| \mathop{\mathbb{E}}_{(x,b)\sim\lambda} [q(w)] - r \right| \leq \tau,$$

with $\tau > 0$ being the tolerance parameter. An algorithm in the SQ model is called *adaptive* when the choice of a query is allowed to depend on the answers to earlier queries posed, and *non-adaptive* otherwise. The SQ model may be viewed as a special case of learning with i.i.d. samples since i.i.d. samples may be used to answer such queries with high probability. In [KLN+08], an equivalence was shown between query complexity under the SQ model and sample complexity under LDP. Specifically, the equivalence was shown to hold between non-interactive LDP and non-adaptive SQ, as well as between sequential LDP and adaptive SQ. In this work, we investigate a special case of the SQ model, called the *correlational statistical query (CSQ) model*, where each query

$q : \mathcal{U} \times \{-1, 1\} \to [-1, 1]$ takes the form $q(x, b) = b \cdot f(x)$ for some function $f : \mathcal{X} \to [-1, 1]$. We show that the approximate factorization algorithm, as well as the lower bound in terms of the approximate γ_2 norm, may be adapted to the CSQ setting. This gives a characterization of learning under the CSQ model in terms of the approximate γ_2 norm and allows us to draw an equivalence for learning between query complexity under adaptive CSQ and sample complexity under non-interactive LDP. To obtain the CSQ lower bound, we take advantage of the CSQ lower bound of [MS20a], expressed in terms of a 'correlational variance' quantity.

The research represented by this chapter has not been published elsewhere and is the result of joint work with Aleksandar Nikolov and Toniann Pitassi.

1.2.6 Equivalence between sequential LDP and single-intrusion pan-privacy

Pan-privacy is a model of differential privacy which protects against a stronger adversarial model than does central DP, though it still relies on a trusted central data curator, unlike with LDP. A pan-private mechanism is an online algorithm which processes the data points in sequence. The *single-intrusion* variant we will consider assumes that the internal state of the mechanism may be compromised at any one moment in time. Privacy is guaranteed against an adversary who has access to both the output of the algorithm and this single internal state. In [CU21], an information theoretic bound was derived which holds for single-intrusion pan-private algorithms and enables the derivation of sample-complexity lower bounds against this model in terms of a variant of *statistical query (SQ) dimension*, as formulated in [Fel17]. Sample complexity upper bounds for private algorithms may be obtained in terms of SQ dimension by following multiplicative-weight techniques as applied in [Fel17] to the statistical query model of machine learning. In this way, we obtain a characterization of sample complexity for pan-private learning in terms of statistical query dimension. By relating it to the characterization in terms of SQ dimension for learning in the statistical query model of machine learning, we are able to show an equivalence between sample complexity under single-intrusion pan-privacy and query complexity for learning under the statistical query model. Similarly, we may show an equivalence between sample complexities for learning under single-intrusion pan-privacy and sequential LDP. In particular, it is shown that an arbitrary concept class is efficiently learnable under single intrusion pan-privacy if and only if it is efficiently learnable under sequential LDP.

The research presented in this chapter has not been published elsewhere and is the result of joint work with Aleksandar Nikolov and Toniann Pitassi.

1.3 Summary of known separations and equivalences

For reference, we include Table 1.1 and Table 1.2 summarizing known equivalences and separations for differential privacy.

Non-interactive LDP vs. sequentially interactive LDP	Exponential sample-complexity separation for distribution dependent learning where the marginal on unlabelled samples is known [KLN+11]. Exponential sample-complexity separation for agnostic PAC learning of linear separators [DF19] and also for large-margin linear separators [DF20].
Sequentially interactive LDP vs. fully interactive LDP	By [JMR19], there exists an exponential sample-complexity separation between sequentially interactive LDP with i.i.d. inputs and fully interactive LDP with i.i.d. inputs for solving a certain combinatorial problem. It is not known whether such a separation exists for problems such as PAC learning.

Table 1.1: Known separations between models

Non-interactive LDP vs. statistical query model	By [KLN+11], non-interactive LDP protocols operating on a data set drawn i.i.d. from an unkown distribution are equivalent to statistical query algorithms operating on the same unknown distribution, in the sense that the distribution of the output of the simulation and the distribution of the output of the original are close in total variation. Polynomial interdepdence between number of statistical queries and inverse tolerance on one hand, versus data set size under LDP and inverse privacy parameter on the other.
Non-interactive LDP vs. statistical query model	By [KLN+11], equivalent in same sense as non-interactive LDP and statistical query model.
Sequentially interactive LDP vs. compositional LDP	By [JMNR19], equivalent when inputs are drawn i.i.d. from an unknown distribution. Simulation of compositional protocol by sequential protocol requires, with high probability, at most a constant factor more samples.
Sequentially interactive LDP vs. 2-intrusion pan-privacy	By [AJM20], any 2-intrusion ε-pan-private protocol may be translated into a sequentially interactive ε-LDP protocol which generates its transcript distribution using an input of the same size, as well as vice versa. No assumptions required of input distribution.
Sequentially interactive LDP vs. single-intrusion pan-privacy	In Chapter 8, realizable learning under both models is characterized by $\mathrm{SQD}^{\mathrm{R}}(\mathcal{C}, \alpha)$ (Definition 55). This leads polynomial equivalence between sample complexities for realizable learning under these two models (Corollary 54). This equivalence may also be extended to agnostic learning (Section 8.6).
Non-interactive LDP vs. correlational statistical query model	In Chapter 6, agnostic learning under non-interactive LDP is characterized in terms of the approximate γ_2 norm. In Chapter 9, agnostic learning under the correlational statistical query model is also characterized in terms of the approximate γ_2 norm. Combining these characterizations leads to a polynomial equivalence between the two models.

Table 1.2: Known equivalences between models

Chapter 2

Preliminaries

In this chapter we introduce basic notation and definitions which will be used throughout the rest of this work.

2.1 Norms

For a set $\mathcal{S}$, the ℓ_1, ℓ_2 and ℓ_∞ norms on $\mathbb{R}^{\mathcal{S}}$ are given respectively by

$$\|a\|_1 = \sum_{v \in \mathcal{S}} |a_v|, \quad \|a\|_2 = \sqrt{\sum_{v \in \mathcal{S}} (a_v)^2}, \quad \|a\|_\infty = \max_{v \in \mathcal{S}} |a_v|.$$

Given a probability distribution π on $\mathcal{S}$, we consider the norms $\|\cdot\|_{L_1(\pi)}$ and $\|\cdot\|_{L_2(\pi)}$ on $\mathbb{R}^{\mathcal{S}}$, given by

$$\|a\|_{L_1(\pi)} = \sum_{v \in \mathcal{S}} \pi(v)|a_v|, \quad \|a\|_{L_2(\pi)} = \sqrt{\sum_{v \in \mathcal{S}} \pi(v)(a_v)^2}.$$

We also take advantage of a number of matrix norms. For norms $\|\cdot\|_\zeta$ and $\|\cdot\|_\xi$ on $\mathbb{R}^{\mathcal{S}}$ and $\mathbb{R}^{\mathcal{S}'}$ respectively, we consider the *matrix operator norm* of $M \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}'}$ given by

$$\|M\|_{\zeta \to \xi} = \max_{x \in \mathbb{R}^{\mathcal{S}} \setminus \{0\}} \frac{\|Mx\|_\xi}{\|x\|_\zeta}.$$

For the special case of $\|M\|_{\ell_s \to \ell_t}$, we will simply write $\|M\|_{s \to t}$. Of particular importance are $\|M\|_{1 \to \infty}$ which corresponds to the largest entries of M, $\|M\|_{1 \to 2}$, which corresponds to the maximum ℓ_2-norm of a column of M, and $\|M\|_{2 \to \infty}$, which corresponds to the maximum ℓ_2-norm of a row of M.

The *inner product* of two matrices M and N in $\mathbb{R}^{\mathcal{S} \times \mathcal{S}'}$ is defined by $M \bullet N = \mathrm{Tr}(M^\top N) = \sum_{u \in \mathcal{S}, v \in \mathcal{S}'} m_{u,v} n_{u,v}$.

Lastly, the γ_2 *norm*, a type of *factorization norm* which will play a central role in chapters 5, 6 and 7, is given for $M \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}'}$ by

$$\gamma_2(M) = \min\{\|R\|_{2 \to \infty} \|A\|_{1 \to 2} : RA = M\}.$$

2.2 Differential privacy

2.2.1 Central model

Let $\mathcal{X}$ denote the *data universe*. A generic element from $\mathcal{X}$ will be denoted by x. We consider *data sets* of the form $\overline{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, each of which is identified with its *histogram* $h \in \mathbb{N}_{\geq 0}^{\mathcal{X}}$ where, for every $x \in \mathcal{X}$, $h_x = |\{i : x_i = x\}|$, so that $\|h\|_1 = n$. To refer to a data set, we use $\overline{x}$ and h interchangeably. A pair of data sets $\overline{x} = (x_1, \ldots, x_i, \ldots, x_n)$ and $\overline{x}' = (x_1, \ldots, x_i', \ldots, x_n)$ are called *adjacent* if X' is obtained from X by replacing an element x_i of $\overline{x}$ with a new universe element $x_i' \in \mathcal{X}$.

For parameters $\varepsilon, \delta > 0$, an (ε, δ)*-differentially private ((ε, δ)-DP) mechanism* [DMNS06] is a random function $\mathcal{M} : \mathcal{X}^n \to \mathcal{Z}$ which, for all adjacent data sets $\overline{x}$ and $\overline{x}'$, for all outcomes $S \subseteq \mathcal{Z}$, satisfies

$$\Pr_{\mathcal{M}}[\mathcal{M}(X) \in S] \leq e^{\varepsilon} \Pr_{\mathcal{M}}[\mathcal{M}(X') \in S] + \delta.$$

A mechanism which is $(\varepsilon, 0)$-differentially private will be referred to as being simply ε*-differentially private (ε-DP for short).*

2.2.2 Local model

The main focus of this work is on the local model of differential privacy, where privacy is guaranteed against even the central party responsible for aggregating the data of individuals [EGS03b, DMNS06, KLN+11].

A (ε, δ)-differentially private mechanisms $\mathcal{M}_i : \mathcal{X} \to \mathcal{Y}$ which take a singleton data set $\overline{x} = \{x\}$ as input is referred to as a *local randomizer.* A sequence of (ε, δ)-differentially private local randomizers $\mathcal{M}_1, \ldots, \mathcal{M}_n$ – perhaps chosen at random from an arbitrary joint distribution (independently from the input) – together with a *post-processing function* $\mathcal{A} : \mathcal{Y}^n \to \mathcal{Z}$ specify a *non-interactive locally* (ε, δ)*-differentially private mechanism* $\mathcal{M} : \mathcal{X}^n \to \mathcal{Z}$. When the local protocol $\mathcal{M}$ is applied to a data set $\overline{x}$, we refer to

$$\mathcal{T}_{\mathcal{M}}(\overline{x}) = (\mathcal{M}_1(x_1), \ldots, \mathcal{M}_n(x_n))$$

as the *transcript* of the protocol. Then the final output of the protocol is given by $\mathcal{M}(\overline{x}) = \mathcal{A}(\mathcal{T}_{\mathcal{M}}(\overline{x}))$.

Local privacy may be defined more generally to allow for various forms of interactivity. In general, a locally differentially private protocol $\mathcal{M}$ is a distributed communication protocol which proceeds in rounds. In round $t \in [S]$, each agent $i \in [n]$ communicates in sequence a random message $y_{t,i} \in \mathcal{Y}$, which is seen by the other agents. In particular, each $y_{t,i}$ is allowed to depend on the input x_i to agent i as well as any of earlier messages $y_{t',i'}$ where $t' < t$ or where both $t' = t$ and $i' < i$. The transcript of the protocol is then defined as the sequence of all such messages, namely

$$\mathcal{T}_{\mathcal{M}}(\overline{x}) = (y_{1,1}, \ldots, y_{1,n} \ ; \ y_{2,1}, \ldots, y_{2,n} \ ; \ \ldots ; \ y_{S,1}, \ldots, y_{S,n}) \in \mathcal{Y}^{S \times n}.$$

A post-processing function $\mathcal{A} : \mathcal{Y}^{S \times n} \to \mathcal{Z}$ is applied to the transcript to obtain the protocol's final output

$$\mathcal{M}(\overline{x}) = \mathcal{A}(\mathcal{T}_{\mathcal{M}}(\overline{x})).$$

We say that $\mathcal{M}$ is *locally* (ε, δ)*-differentially private (ε-LDP)* if the function $\mathcal{T}_{\mathcal{M}} : \mathcal{X}^n \to \mathcal{Y}^{S \times n}$ is itself (ε, δ)-DP. As with central DP, we refer to a mechanism which is locally $(\varepsilon, 0)$-differentially private as *locally ε-differentially private $\varepsilon - LDP$.*

In [BNS18], a reduction from (ε, δ)-LDP to ε-LDP is given, which shows that (ε, δ)-LDP provides essentially no additional power over ε-LDP. For this reason, this work will focus almost exclusively on "pure" ε-LDP. For conciseness, we give the rest of our definitions only for ε-LDP rather than (ε, δ)-LDP.

A *sequential* ε-LDP protocol $\mathcal{M}$ is one consisting of a single round, where each agent i communicates only a single message y_i which depends on their input x_i and the messages $y_1, \ldots, y_{i-1}$ already sent. In particular, we may view each agent i as being assigned a local randomizer $\mathcal{M}_i$ which depends on the messages $y_1, \ldots, y_{i-1}$ already sent, with local agent i reporting $y_i = \mathcal{M}_i(x_i)$.

A *compositional* ε-LDP protocol $\mathcal{M}$ allows for multiple rounds. In each round $t \in [S]$, each agent i is assigned a $\varepsilon_{t,i}$-DP local randomizer $\mathcal{M}_{t,i}$ which depends on the messages previously sent. They report $y_{t,i} = \mathcal{M}_{t,i}(x_i)$. If $\max_{i \in [n]} \sum_{t \in [S]} \varepsilon_{t,i} \leq \varepsilon$, then $\mathcal{M}$, which is guaranteed to be ε-LDP by basic composition properties, is called a *compositional* ε-LDP protocol.

In [JMNR19], it is shown that the sequential and compositional models of local differential privacy are equivalent when the data points are drawn i.i.d., in the sense that a compositional local protocol can then be simulated with a sequential local protocol using, with high probability, at most a constant factor more samples.

In [JMR19], it is shown that a certain combinatorial problem requires exponentially more samples to be solved with a compositional LDP protocol instead of with one which is fully interactive. It is not known whether such a separation exists for natural statistical and learning tasks.

2.3 Answering statistical queries

A *statistical query* is specified by a bounded function $q : \mathcal{X} \to \{-1, 1\}$. Abusing notation slightly, its answer on a data set $\overline{x} \in \mathcal{X}^n$ is given by $q(\overline{x}) = \frac{1}{n} \sum_{i=1}^{n} q(x_i)$. We also extend this notation to distributions: if μ is a distribution on $\mathcal{X}$, then we write $q(\mu)$ for $\underset{x \sim \mu}{\mathbb{E}}[q(x)]$. A *workload* is a set of statistical queries $Q = \{q_1, \ldots, q_k\}$, and $Q(\overline{x}) = (q_1(\overline{x}), \ldots, q_k(\overline{x}))$ is used to denote their answers on the data set, $Q(\mu) = (q_1(\mu), \ldots, q_k(\mu))$ their answers on μ. We will often represent Q by its *workload matrix* $W \in \mathbb{R}^{Q \times \mathcal{X}}$ with entries $w_{q,x} = q(x)$. In this notation, the answers to the queries on $\overline{x} \in \mathcal{X}^n$ are given by $\frac{1}{n} W h$ where $h \in \mathbb{N}^{\mathcal{X}}$ is the histogram of $\overline{x}$, defined by $h_x = |\{i : x_i = x\}|$ We will often use Q and W interchangeably.

The ℓ_∞ *error* of a mechanism $\mathcal{M} : \mathcal{X}^* \to \mathcal{Z}$ on the query workload Q is given by

$$\mathrm{err}^{\ell_\infty}(\mathcal{M}, Q, n) = \max_{\overline{x} \in \mathcal{X}^n} \mathop{\mathbb{E}}_{\mathcal{M}} [\|\mathcal{M}(\overline{x}) - Q(\overline{x})\|_\infty].$$

We can then define the *sample complexity* of $\mathcal{M}$ for a given ℓ_∞ error α by

$$\mathrm{sc}^{\ell_\infty}(\mathcal{M}, Q, \alpha) = \min\{n \in \mathbb{N} : \mathrm{err}^{\ell_\infty}(\mathcal{M}, Q, n) \leq \alpha\}.$$

Having defined error and sample complexity for a fixed mechanism, we can define the optimal error and sample complexity under local differential privacy by

$$\begin{aligned}\mathrm{err}^{\ell_\infty}_{\varepsilon\text{-LDP}}(Q,n) &= \min_{\mathcal{M}\text{ is }\varepsilon\text{-LDP}} \mathrm{err}^{\ell_\infty}(\mathcal{M},Q,n),\\ \mathrm{sc}^{\ell_\infty}_{\varepsilon\text{-LDP}}(Q,\alpha) &= \min_{n\in\mathbb{N}}\{n\in\mathbb{N} : \mathrm{err}^{\ell_\infty}_{\varepsilon\text{-LDP}}(Q,n)\le\alpha\}\end{aligned}$$

We may also define *distributional* versions of some of these quantities by

$$\begin{aligned}\text{dist-err}^{\ell_\infty}(\mathcal{M},Q,n) &= \max_{\mu} \mathop{\mathbb{E}}_{\substack{\overline{x}\sim\mu^n\\ \mathcal{M}}}[\|\mathcal{M}(\overline{x})-Q(\mu)\|_\infty],\\ \text{dist-sc}^{\ell_\infty}(\mathcal{M},Q,\alpha) &= \min\{n : \text{dist-err}^{\ell_\infty}(\mathcal{M},Q,n)\le\alpha\},\\ \text{dist-sc}^{\ell_\infty}_{\varepsilon\text{-LDP}}(Q,\alpha) &= \min_{\mathcal{M}\text{ is }\varepsilon\text{-LDP}} \text{dist-sc}^{\ell_\infty}_{\varepsilon\text{-LDP}}(\mathcal{M},Q,n).\end{aligned}$$

Analogous quantities can be defined for both non-interactive LDP and sequential LDP. We label these $\mathrm{err}^{\ell_\infty}_{\varepsilon\text{-NILDP}}(Q,n)$ and $\mathrm{sc}^{\ell_\infty}_{\varepsilon\text{-NILDP}}(Q,\alpha)$ for non-interactive ε-LDP. The notation $\mathrm{err}^{\ell_\infty}_{\varepsilon\text{-SeqLDP}}(Q,n)$ and $\mathrm{sc}^{\ell_\infty}_{\varepsilon\text{-SeqLDP}}(Q,\alpha)$ is used for sequential ε-LDP.

2.4 Learning

A concept $c:\mathcal{U}\to\{-1,1\}$ from a concept class $\mathcal{C}\subseteq\{-1,1\}^{\mathcal{U}}$ identifies each sample $a\in\mathcal{U}$ with a label $c(a)$. The *empirical loss* of the *hypothesis* $h:\mathcal{U}\to\{-1,1\}$ on a data set $\overline{x}=((a_1,b_1),\ldots,(a_n,b_n))\in(\mathcal{U}\times\{-1,1\})^n$ is given by

$$L_{\overline{x}}(h)=\frac{1}{n}\sum_{i=1}^{n}(\mathbb{I}[h(a_i)\neq b_i])$$

For a distribution μ on $\mathcal{U}\times\{-1,1\}$, the *population loss* of h on μ is given by $L_\mu(h)=\mathop{\mathbb{P}}_{(a,b)\sim\mu}[h(a)\neq b]$.

We will say that a mechanism $\mathcal{M}:(\mathcal{U}\times\{-1,1\})^n\to\{-1,1\}^{\mathcal{U}}$ *(α,β)-learns $\mathcal{C}$ agnostically* with n samples if, for any distribution μ over $\mathcal{U}\times\{-1,1\}$, given as input a random data set $\overline{x}$ drawn i.i.d. from μ, the mechanism returns some hypothesis $h:\mathcal{U}\to\{-1,1\}$ which satisfies

$$\mathop{\mathbb{P}}_{\overline{x},\mathcal{M}}\left[L_\mu(h)\le\min_{c\in\mathcal{C}}L_\mu(c)+\alpha\right]\ge 1-\beta.$$

Realizable learning is an important special case of agnostic learning where the underlying distribution agrees with some concept. We say that $\mathcal{M}:(\mathcal{U}\times\{-1,1\})^n\to\{-1,1\}^{\mathcal{U}}$ *(α,β)-learns $\mathcal{C}$ realizably* with n samples if, whenever μ is a distribution over $\mathcal{U}\times\{-1,1\}$ which satisfies $\mathop{\mathbb{P}}_{(a,b)\sim\mu}[c(a)=b]=1$ for some unknown $c\in\mathcal{C}$, then, given a random data set $\overline{x}$ drawn i.i.d. from μ, the mechanism returns a *hypothesis* $h\in\mathcal{U}\to\{-1,1\}$ which satisfies

$$\mathop{\mathbb{P}}_{\overline{x},\mathcal{M}}[L_\mu(h)\le\alpha]\ge 1-\beta.$$

The problem of refutation asks whether the underlying distribution is well approximated by the concept class. In particular, for $\theta\in[0,1]$, we will say that $\mathcal{M}_\theta:(\mathcal{U}\times\{-1,1\})^n\to\{-1,1\}$ *(α,β)-refutes $\mathcal{C}$ for threshold θ* if the following two conditions are met:

1. When μ is a distribution on $\mathcal{U} \times \{-1, 1\}$ which satisfies $L_\mu(c) \le \theta$ for some $c \in \mathcal{C}$,

$$\mathop{\mathbb{P}}_{\overline{x}, \mathcal{M}} [\mathcal{M}(\overline{x}) = 1] \ge 1 - \beta;$$

2. When μ is a distribution on $\mathcal{U} \times \{-1, 1\}$ which, for all $h \in \{-1, 1\}^{\mathcal{U}}$, satisfies $L_\mu(h) > \theta + \alpha$, then

$$\mathop{\mathbb{P}}_{\overline{x}, \mathcal{M}} [\mathcal{M}(\overline{x}) = -1] \ge 1 - \beta.$$

We say that $\{\mathcal{M}_\theta\}_{\theta \in [0,1]}$ (α, β)-refutes $\mathcal{C}$ agnostically if, for all $\theta \in [0, 1]$, $\mathcal{M}_\theta$ (α, β)-refutes $\mathcal{C}$ agnostically for threshold θ.

Realizable refutation is a special case of agnostic refutation where the goal is to recognize whether the underlying distribution is labeled by a concept from the concept class. We say that $\mathcal{M}_\theta : (\mathcal{U} \times \{-1, 1\})^n \to \{-1, 1\}$ (α, β)-*refutes* $\mathcal{C}$ *realizably* if it (α, β)-refutes $\mathcal{C}$ for threshold 0.

2.5 Statistical query model

A *statistical query (SQ) algorithm* operates on an underlying distribution μ on $\mathcal{X}$ by accessing it via a *statistical query oracle* SQ_μ which, when queried with some $q : \mathcal{X} \to [-1, +1]$, returns some $r \in [-1, 1]$ such that $|r - q(\mu)| \le \tau$, with τ denoting the *tolerance* parameter given to the statistical query oracle. In a sense, the statistical query model is a special case of the setting where we access μ via i.i.d. samples, since, with a sufficiently large data set $\overline{x}$ drawn i.i.d. from μ we may use $q(\overline{x})$ to answer the statistical query q with small probability of failure. Note that the statistical query model allows queries to be answered adversarially by the oracle. A statistical query algorithm which poses all of its statistical queries before viewing the answers to any is referred to as *non-adaptive.* A statistical query algorithm which is allowed to depend its choice of queries on the answers to earlier posed queries is referred to as *adaptive.*

In [KLN+11], a close relationship is shown between compositionally interactive LDP and the statistical query model. In particular, we may translate between LDP protocols and statistical query protocols with only polynomial blow-ups in complexity. Non-interactive LDP protocols correspond specifically to non-adaptive SQ algorithms, and compositionally interactive to adaptive. It is not known whether the relationship holds for fully interactive local protocols.

The statistical query model is often considered in a learning context where μ is a distribution on labelled samples from $\mathcal{X} = \mathcal{U} \times \{-1, 1\}$. In this setting, we sometimes narrow our attention to *correlational statistical query (CSQ) algorithms*, namely those where each of the posed queries $q : \mathcal{U} \times \{-1, 1\} \to [-1, 1]$ is given by some function $f : \mathcal{U} \to [-1, 1]$, so that

$$q(a, b) = f(a) \cdot b.$$

See Chapter 9 for further discussion.

2.6 Pan-privacy

Pan-private mechanisms protect against a stronger adversarial model than does central DP, though they still rely on a trusted curator, and hence do not go as far as the local model in this way. A pan-private mechanism is a streaming algorithms which processes the data set in sequence, and guarantees privacy even if the internal state of the algorithm is revealed to the adversary.

Being a streaming algorithm, $\mathcal{M} : \mathcal{X}^n \to \mathcal{Z}$ may be defined inductively in terms of $\mathcal{M}_1 : \mathcal{X} \to \mathcal{Y}$ together with a sequence, for $i \geq 2$, of $\mathcal{M}_i : (\mathcal{X} \times \mathcal{Y}) \to \mathcal{Y}$, as well as a post-processing function $\mathcal{A} : \mathcal{Y} \to \mathcal{Z}$. Then, on a dataset $\overline{x} = (x_1, \ldots, x_n)$, the first *internal state* is given by $I_1(x_1) = \mathcal{M}_1(x_1)$, while each iteration $i \in \{2, \ldots, n\}$ produces the internal state $I_i(x_1, \ldots, x_i) = \mathcal{M}_i(x_i, I_{i-1}(x_1, \ldots, x_{i-1}))$. The final output is given by $\mathcal{M}(\overline{x}) = \mathcal{A}(I_n(x_1, \ldots, x_n))$.

We call $\mathcal{M}$ ε-pan-private against a single intrusion if the function which jointly releases $\mathcal{M}(\overline{x})$ and $I_i(x_1, \ldots, x_i)$, for some arbitrary $i \in [n]$ is itself ε-DP. In other words, for adjacent datasets $\overline{x} = (x_1, \ldots, x_n)$ and $\overline{x}' = (x'_1, \ldots, x'_n)$, it holds for all $S \subseteq \mathcal{Y} \times \mathcal{Z}$ that

$$\Pr_{\mathcal{M}}[(I_i(x_1, \ldots, x_i), \mathcal{M}(\overline{x})) \in S] \leq e^{\varepsilon} \Pr_{\mathcal{M}}[(I_i(x'_1, \ldots, x'_i), \mathcal{M}(\overline{x}')) \in S].$$

We may also consider for $r \geq 2$, ε-pan-privacy against r intrusions, in which case differential privacy must be satisfied in the scenario where $\mathcal{M}(x_1, \ldots, x_i)$ is jointly released with $I_i(x_1, \ldots, x_i)$ for r arbitrary choices of $i \in [n]$.

Chapter 3

Information-theoretic bounds

3.1 Overview

In this chapter, we present a number of information theoretic bounds applicable to the derivation of sample complexity lower bounds against LDP. We present the mutual information bound for DP mechanisms, Lemma 1, of [MMP+10], generalized in [SZ20], which in Chapter 4 will be applied to transform lower bounds on the information complexity of communication problems into sample-complexity lower bounds for statistical problems under LDP. We also introduce our KL-divergence bound, Lemma 2, which will be one of our primary tools in deriving lower bounds against estimation and learning problems under non-interactive LDP. This result is closely related to Lemma 3 of [DR18], which gives a weaker bound, but holds in the setting of sequential LDP. In Chapter 8, we will apply this bound when showing that sequential LDP and single-intrusion pan-privacy are characterized by the same variant of statistical-query dimension. We present a third bound which takes a similar form to Lemma 2 and Lemma 3, though it is an upper bound on mutual information rather than KL-divergence. In Appendix B.3, it is applied towards getting a tight lower bound against answering parity statistical queries under non-interactive LDP, a generalization of the selection lower bound of [Ull18].

3.2 Mutual information bounds for LDP

Mutual information receives much attention in the context of communication theory where mutual information lower bounds directly yield communication cost lower bounds. We will see how such mutual information lower bounds are easily translated into sample complexity complexity bounds against LDP.

In a *communication protocol*, a dataset $\overline{x} \in \mathcal{X}^n$ is distributed among multiple agents who communicate in turn. On each turn, the message sent is allowed to depend on previously sent messages as well as the input of the agent sending the message. The transcript of these messages $\mathcal{T}_{\mathcal{M}} : \mathcal{X}^n \to \{0,1\}^*$ is a random function on top of which a post-processing function $\mathcal{A} : \{0,1\}^* \to \mathcal{Z}$ is applied to obtain the final output $\mathcal{M}(\overline{x}) = \mathcal{A}(\mathcal{T}_{\mathcal{M}}(\overline{x}))$. In this way, an ε-LDP protocol may be viewed as a communication protocol with the additional constraint that the transcript of the protocol is ε-DP.

Communication complexity is typically studied in the context of a binary decision problem, given by a binary function $f : \mathcal{X}^n \to \{0,1\}$, with the goal being to output $f(\overline{x})$. The minimum mutual information $I\left(\ \overline{X}\ ;\ \mathcal{T}_{\mathcal{M}}(\overline{X})\ \right)$ required of a protocol $\mathcal{M}$ operating on a random dataset $\overline{X}$ which decides the problem with high probability is known as its *information complexity.* Lower bounds on information complexity are of interest in communication theory primarily because $I\left(\ \overline{X}\ ;\ \mathcal{T}_{\mathcal{M}}(\overline{X})\ \right)$ provides a lower bound on the *entropy* $H\left(\mathcal{T}_{\mathcal{M}}(\overline{X})\right)$, which is itself a lower bound on the length $|\mathcal{T}_{\mathcal{M}}(\overline{X})|$ of the transcript. From our point of view, information complexity lower bounds are useful because they may be combined with the mutual information upper bounds which follow from differential privacy.

A precursor to the following mutual information upper bound is found for 2-party differential privacy in [MMP+10]. The general bound for all DP mechanisms and specifically for the transcripts of arbitrary LDP protocols is due to [SZ20]. The stronger bounds in the case of i.i.d. inputs are also due to [SZ20]. It appears that the stronger bound in the case of sequential or compositional LDP, without assumptions on the data distribution, has not been previously published, so a proof of that portion of the lemma – which arose out of collaboration with Aleksandar Nikolov, Toniann Pitassi, and Jonathan Ullman – is given in Appendix A.1.

Lemma 1 (2-party case in [MMP+10]; weak bound for general case and strong bound for i.i.d. case in [SZ20]; original result for non-i.i.d. sequential case)**.** *Let $\varepsilon = O(1)$. Suppose $\mathcal{M} : \mathcal{X}^n \to \mathcal{Z}$ is ε-DP. Let $\overline{X} \in \mathcal{X}^n$ be a random dataset drawn from an arbitrary distribution. Then,*

$$I\left(\ \overline{X}\ ;\ \mathcal{M}(\overline{X})\ \right) = O(n \cdot \varepsilon).$$

When $\overline{X}$ is i.i.d., then

$$I\left(\ \overline{X}\ ;\ \mathcal{M}(\overline{X})\ \right) = O(n \cdot \varepsilon^2).$$

In the context of LDP, with $\mathcal{T}_{\mathcal{M}} : \mathcal{X}^n \to \mathcal{Z}$ being the transcript of an ε-LDP protocol $\mathcal{M}$ which takes $\overline{X}$ as input, then, since $\mathcal{T}_{\mathcal{M}}$ is itself ε-DP, this result implies

$$I\left(\ \overline{X}\ ;\ \mathcal{T}_{\mathcal{M}}(\overline{X})\ \right) = O(n \cdot \varepsilon).$$

and, when $\overline{X}$ is i.i.d.,

$$I\left(\ \overline{X}\ ;\ \mathcal{T}_{\mathcal{M}}(\overline{x})\ \right) = O(n \cdot \varepsilon^2). \tag{3.1}$$

When $\mathcal{M}$ is a sequential LDP protocol, or merely compositional, then we do not need the requirement that $\overline{X}$ is i.i.d., and (3.1) holds nevertheless.

By rearranging, this result allows us to lower-bound the sample complexity of an ε-LDP protocol in terms of the mutual information between the input and the transcript. In particular, we have

$$n = \Omega\left(\frac{I\left(\ \overline{X}\ ;\ \mathcal{M}(\overline{X})\ \right)}{\varepsilon^2}\right)$$

when $\overline{X}$ is i.i.d., and otherwise

$$n = \Omega\left(\frac{I\left(\ \overline{X}\ ;\ \mathcal{M}(\overline{X})\ \right)}{\varepsilon}\right).$$

In either case, a lower bound on $I\left(\ \overline{X}\ ;\ \mathcal{M}(\overline{X})\ \right)$ gives us a lower bound on the number of samples required.

3.3 KL-divergence bounds for LDP

KL-divergence upper bounds play an important role in obtaining lower bounds against LDP. Such techniques rely on the construction of families $\{\lambda_1, \ldots, \lambda_k\}$ and $\{\mu_1, \ldots, \mu_k\}$ of 'hard' input distributions, together with a parameter distribution π on $[k]$. For any $v \in [k]$, let λ_v^n be the distribution which draws n independent samples from λ_v. Let λ_π^n be the mixture $\sum_{v=1}^k \pi(v)\lambda_v^n$. Define μ_v^n and μ_π^n analogously. It is worth emphasizing that λ_π^n and μ_π^n are mixtures of product distributions, and not product distributions themselves. Consider our following result.

Lemma 2 (KL-divergence bound for non-interactive LDP)**.** *Let $\varepsilon = O(1)$, and let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Z}$ be a non-interactive ε-LDP protocol with transcript $\mathcal{T}_\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}^n$. Then, for distributions $\lambda_1, \ldots, \lambda_k$ and $\mu_1, \ldots, \lambda_k$ on $\mathcal{X}$, together with a distribution π over $[k]$,*

$$\begin{aligned}
\mathrm{D}_{KL}(\mathcal{T}_\mathcal{M}(\lambda_\pi^n) \| \mathcal{T}_\mathcal{M}(\mu_\pi^n)) &\le \mathop{\mathbb{E}}_{V\sim\pi}[\mathrm{D}_{KL}(\mathcal{T}_\mathcal{M}(\lambda_V^n) \| \mathcal{T}_\mathcal{M}(\mu_V^n))] \\
&\le O(n\varepsilon^2) \cdot \max_{f \in \mathbb{R}^\mathcal{X} : \|f\|_\infty \le 1} \mathop{\mathbb{E}}_{V\sim\pi}\left[\left(\mathop{\mathbb{E}}_{X\sim\mu_V}[f_X] - \mathop{\mathbb{E}}_{X\sim\lambda_V}[f_X]\right)^2\right].
\end{aligned}$$

In matrix notation, define the matrix $M \in \mathbb{R}^{[K]\times\mathcal{X}}$ by $m_{v,x} = (\mu_v(x) - \lambda_v(x))$. Then,

$$\begin{aligned}
\mathrm{D}_{KL}(\mathcal{T}_\mathcal{M}(\lambda_\pi^n) \| \mathcal{T}_\mathcal{M}(\mu_\pi^n)) &\le \mathop{\mathbb{E}}_{V\sim\pi}[\mathrm{D}_{KL}(\mathcal{T}_\mathcal{M}(\lambda_V^n) \| \mathcal{T}_\mathcal{M}(\mu_V^n))] \\
&\le O(n\varepsilon^2) \cdot \|M\|^2_{\ell_\infty \to L_2(\pi)}.
\end{aligned}$$

Proof. Let $\mathcal{M}_i : \mathcal{X}^n \to \mathcal{Y}$ be the local randomizer which agent i applies to their data point in the execution of the protocol $\mathcal{M}$. We have

$$\begin{aligned}
\mathop{\mathbb{E}}_{V\sim\pi}[\mathrm{D}_{\mathrm{KL}}(\mathcal{T}_\mathcal{M}(\lambda_V^n) \| \mathcal{T}_\mathcal{M}(\mu_V^n))] &= \mathop{\mathbb{E}}_{V\sim\pi}\left[\sum_{i=1}^n \mathrm{D}_{\mathrm{KL}}(\mathcal{M}_i(\lambda_V) \| \mathcal{M}_i(\mu_V))\right] \qquad (3.2) \\
&= \sum_{i=1}^n \mathop{\mathbb{E}}_{V\sim\pi}[\mathrm{D}_{\mathrm{KL}}(\mathcal{M}_i(\lambda_V) \| \mathcal{M}_i(\mu_V))] \\
&\le \sum_{i=1}^n \mathop{\mathbb{E}}_{V\sim\pi}[\mathrm{D}_{\chi^2}(\mathcal{M}_i(\lambda_V) \| \mathcal{M}_i(\mu_V))] \qquad (3.3)
\end{aligned}$$

where (3.2) is by independence, and (3.3) by the fact that χ^2-divergence is always an upper bound on KL-divergence [GS02].

Remark. If we were in the sequential setting, where the output of agent i can depend on the outputs $\mathcal{T}_\mathcal{M}^{<i}$ of the previous $i-1$ agents, then, instead of independence, we could use the chain rule for KL-divergence. In line (3.2) of the inequality above, this would result in conditioning each KL-divergence term on $\mathcal{T}_\mathcal{M}^{<i}$ and taking the expectation of the summands with respect to $\mathcal{T}_\mathcal{M}^{<i}(\lambda_V^n)$ as well. Unfortunately, then the expectation with respect to $V \sim \pi$ does not appear next to the KL-divergence expressions and we cannot change the order of expectation since $\mathcal{T}_\mathcal{M}^{<i}(\lambda_V^n)$ depends on V. We will see that it is possible to obtain a weak version of this result, Lemma 3, which holds

for sequential local protocols but requires all the distributions $\{\mu_v\}_{v\in[k]}$ to be identical, allowing variety only in $\{\lambda_v\}_{v\in[k]}$.

It remains to show

$$\begin{aligned}&\mathop{\mathbb{E}}_{V\sim\pi}\left[\mathrm{D}_{\chi^2}(\mathcal{M}_i(\lambda_V)\|\mathcal{M}_i(\mu_V))\right]\\&= O(\varepsilon^2)\cdot\max_{f\in\mathbb{R}^{\mathcal{X}}:\|f\|_\infty\le 1}\mathop{\mathbb{E}}_{V\sim\pi}\left[\left(\mathop{\mathbb{E}}_{X\sim\lambda_V}[f_X]-\mathop{\mathbb{E}}_{X\sim\mu_V}[f_X]\right)^2\right].\end{aligned}\tag{3.4}$$

To that end, fix some i and, for $x\in\mathcal{X}$, $y\in\mathcal{Y}$, let $r(y|x)$ denote $\Pr_{\mathcal{M}_i}(\mathcal{M}_i(x)=y)$. Also, let $a_v(y)=\mathop{\mathbb{E}}_{X\sim\lambda_v}[r(y|X)]$ and let $b_v(y)=\mathop{\mathbb{E}}_{X\sim\mu_v}[r(y|X)]$. Let us assume, for notational simplicity, but without loss of generality, that the range $\mathcal{Y}$ of $\mathcal{M}_i$ is finite. Then, by applying the definition of χ^2-divergence, the left-hand side of (3.4) may be rewritten as

$$\begin{aligned}&\mathop{\mathbb{E}}_{V\sim\pi}\left[\mathop{\mathbb{E}}_{Y\sim\mathcal{M}_i(\mu_V)}\left[\left(\frac{b_V(Y)-a_V(Y)}{b_V(Y)}\right)^2\right]\right]\\&=\mathop{\mathbb{E}}_{V\sim\pi}\left[\sum_{y\in\mathcal{Z}}\left(\frac{b_V(y)-a_V(y)}{b_V(y)}\right)^2\cdot b_V(y)\right]\end{aligned}$$

Let μ_0 be the uniform distribution on $\mathcal{X}$ (though any other distribution will work). Let $u(y)=\mathop{\mathbb{E}}_{X\sim\mu_0}[r(y|X)]$. Since privacy implies $\frac{u(y)}{b_v(y)}\le e^\varepsilon$, we may obtain an upper bound on the right-hand side of (3.5) as follows.

$$\mathop{\mathbb{E}}_{V\sim\pi}\left[\sum_{y\in\mathcal{Z}}\left(\frac{b_V(y)-a_V(y)}{b_V(y)}\right)^2\cdot b_V(y)\right]\tag{3.5}$$

$$\begin{aligned}&=\mathop{\mathbb{E}}_{V\sim\pi}\left[\sum_{y\in\mathcal{Z}}\left(\frac{b_V(y)-a_V(y)}{u(y)}\right)^2\cdot\frac{u(y)^2}{b_V(y)}\right]\\&\le e^\varepsilon\cdot\mathop{\mathbb{E}}_{V\sim\pi}\left[\sum_{y\in\mathcal{Z}}\left(\frac{b_V(y)-a_V(y)}{u(y)}\right)^2\cdot u(y)\right]\\&=e^\varepsilon\cdot\mathop{\mathbb{E}}_{V\sim\pi}\left[\mathop{\mathbb{E}}_{Y\sim\mathcal{M}_i(\mu_0)}\left[\left(\frac{b_V(Y)-a_V(Y)}{u(Y)}\right)^2\right]\right]\end{aligned}$$

$$=e^\varepsilon\cdot\mathop{\mathbb{E}}_{Y\sim\mathcal{M}_i(\mu_0)}\left[\mathop{\mathbb{E}}_{V\sim\pi}\left[\left(\frac{b_V(Y)-a_V(Y)}{u(Y)}\right)^2\right]\right]\tag{3.6}$$

By taking $f^y\in\mathbb{R}^{\mathcal{X}}$ to be given by $f^y_x=\frac{r(y|x)}{u(y)}-1$, then we obtain

$$\frac{b_V(y)-a_V(y)}{u(y)}=\mathop{\mathbb{E}}_{X\sim\mu_V}[f^y_X]-\mathop{\mathbb{E}}_{X\sim\lambda_V}[f^y_X].$$

Furthermore, $\frac{r(y|x)}{u(y)}\le e^\varepsilon$ is implied by privacy, from which it follows that $\|f^y\|_\infty\le e^\varepsilon-1$. This gives us the following bound on (3.6).

$$\begin{aligned}
&e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{Y\sim\mathcal{M}_i(\mu_0)}\left[\mathop{\mathbb{E}}_{V\sim\pi}\left[\left(\frac{b_V(Y)-a_V(Y)}{u(Y)}\right)^2\right]\right] \\
&= e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{Y\sim\mathcal{M}_i(\mu_0)}\left[\mathop{\mathbb{E}}_{V\sim\pi}\left[\left(\mathop{\mathbb{E}}_{X\sim\mu_V}[f^Y_X] - \mathop{\mathbb{E}}_{X\sim\lambda_V}[f^Y_X]\right)^2\right]\right] \\
&\le e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{Y\sim\mathcal{M}_i(\mu_0)}\left[\sup_{\|f\|_\infty \le e^{\varepsilon}-1}\mathop{\mathbb{E}}_{V\sim\pi}\left[\left(\mathop{\mathbb{E}}_{X\sim\mu_V}[f_X] - \mathop{\mathbb{E}}_{X\sim\lambda_V}[f_X]\right)^2\right]\right] \\
&\le e^{\varepsilon}(e^{\varepsilon}-1)^2 \cdot \sup_{\|f\|_\infty\le 1}\mathop{\mathbb{E}}_{V\sim\pi}\left[\left(\mathop{\mathbb{E}}_{X\sim\mu_V}[f_X] - \mathop{\mathbb{E}}_{X\sim\lambda_V}[f_X]\right)^2\right].
\end{aligned}$$

Using the fact that $e^{\varepsilon}(e^{\varepsilon}-1)^2 = O(\varepsilon^2)$, then putting everything together, we get

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(\mathcal{T}_{\mathcal{M}}(\lambda^n_\pi)\|\mathcal{T}_{\mathcal{M}}(\mu^n_\pi)) \\
&\le O(n\varepsilon^2) \cdot \max_{f\in\mathbb{R}^{\mathcal{X}}:\|f\|_\infty\le 1}\mathop{\mathbb{E}}_{V\sim\pi}\left[\left(\mathop{\mathbb{E}}_{X\sim\lambda_V}[f_X] - \mathop{\mathbb{E}}_{X\sim\mu_V}[f_X]\right)^2\right].
\end{aligned}$$

To obtain our result in terms of the matrix $M \in \mathbb{R}^{[K]\times\mathcal{X}}$ given by $m_{v,x} = (\lambda_v(x) - \mu_v(x))$, we note that the entries of Mf, indexed by $v \in [K]$, are given by

$$\begin{aligned}
(Mf)_v = \sum_{x\in\mathcal{X}} f_x m_{v,x} &= \sum_{x\in\mathcal{X}}(\lambda_v(x)f_x - \mu_v(x)f_x) \\
&= \mathop{\mathbb{E}}_{x\sim\lambda_v}[f_x] - \mathop{\mathbb{E}}_{x\sim\mu_v}[f_x].
\end{aligned}$$

Hence

$$\|Mf\|^2_{L_2(\pi)} = \sum_{v\in[K]}\pi(v)(Mf)_v^2 = \mathop{\mathbb{E}}_{V\sim\pi}\left[\left(\mathop{\mathbb{E}}_{X\sim\lambda_V}[f_X] - \mathop{\mathbb{E}}_{X\sim\mu_V}[f_X]\right)^2\right].$$

Finally, we have

$$\begin{aligned}
\|M\|^2_{\ell_\infty\to L_2(\pi)} &= \sup_{f\in\mathbb{R}^{\mathcal{X}}:\|f\|_\infty\le 1}\|Mf\|^2_{L_2(\pi)} \\
&= \max_{f\in\mathbb{R}^{\mathcal{X}}:\|f\|_\infty\le 1}\mathop{\mathbb{E}}_{V\sim\pi}\left[\left(\mathop{\mathbb{E}}_{X\sim\lambda_V}[f_X] - \mathop{\mathbb{E}}_{X\sim\mu_V}[f_X]\right)^2\right],
\end{aligned}$$

and this completes the proof. □

Suppose we could distinguish between $\mathcal{M}(\lambda^n_\pi)$ and $\mathcal{M}(\mu^n_\pi)$ with success probability $\frac{1}{2} + \Omega(1)$, whereby $d_{\mathrm{TV}}(\mathcal{M}(\lambda^n_\pi), \mathcal{M}(\mu^n_\pi))^2 \ge \Omega(1)$. Then, bounding total variation in terms of KL-divergence and using the fact that post-processing can only decrease total variation, we would have

$$\mathrm{D}_{\mathrm{KL}}(\mathcal{T}_{\mathcal{M}}(\lambda^n_\pi)\|\mathcal{T}_{\mathcal{M}}(\mu^n_\pi)) \ge d_{\mathrm{TV}}(\mathcal{T}_{\mathcal{M}}(\lambda^n_\pi), \mathcal{T}_{\mathcal{M}}(\mu^n_\pi))^2 \ge d_{\mathrm{TV}}(\mathcal{M}(\lambda^n_\pi), \mathcal{M}(\mu^n_\pi))^2 \ge \Omega(1).$$

This implies

$$n = \Omega\left(\frac{1}{\varepsilon^2 \cdot \|M\|^2_{\ell_\infty\to L_2(\pi)}}\right).$$

In this way, there are two goals which which we want to meet simultaneously in the construction of our hard distributions. First, the task which our protocol performs – say, the learning or estimation task – should allow us to distinguish between λ_π^n and μ_π^n. The second goal is to design $\lambda_1, \dots, \lambda_k$ and $\mu_1, \dots, \lambda_k$, together with π, so as to make $\|M\|_{\ell_\infty \to L_2(\pi)}^2$ as small as possible.

The approach just described relies on the bound on $\mathrm{D}_{\mathrm{KL}}(\mathcal{T}_\mathcal{M}(\mu_\pi^n) \| \mathcal{T}_\mathcal{M}(\lambda_\pi^n))$. However, the bound on $\mathop{\mathbb{E}}_{V \sim \pi}[\mathrm{D}_{\mathrm{KL}}(\mathcal{T}_\mathcal{M}(\mu_V^n) \| \mathcal{T}_\mathcal{M}(\lambda_V^n))]$ is more useful in some cases as we will see later in our lower bound against estimating linear queries. To apply this bound, it suffices to demonstrate that, for each $v \in [k]$, we can distinguish between $\mathcal{M}(\lambda_v^n)$ and $\mathcal{M}(\mu_v^n)$ with probability $\frac{1}{2} + \Omega(1)$, whereby $\mathrm{D}_{\mathrm{KL}}(\mathcal{T}_\mathcal{M}(\mu_v^n) \| \mathcal{T}_\mathcal{M}(\lambda_v^n)) = \Omega(1)$ for each $v \in [k]$, and hence $\mathop{\mathbb{E}}_{V \sim \pi}[\mathrm{D}_{\mathrm{KL}}(\mathcal{T}_\mathcal{M}(\mu_V^n) \| \mathcal{T}_\mathcal{M}(\lambda_V^n))] = \Omega(1)$.

This result was preceded by a similar KL-divergence bound for sequential rather than non-interactive LDP which handles the case where $\{\lambda_1, \dots, \lambda_k\}$ is being distinguished from a single reference distribution μ instead of a second family of distributions.

Lemma 3 (KL-divergence bound for sequential LDP, [DR18]). *Let $\varepsilon = O(1)$, and let $\mathcal{M}$ be a sequential ε-LDP protocol with transcript $\mathcal{T}_\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}^n$. Then, for distributions $\lambda_1, \dots, \lambda_k$ and μ on $\mathcal{X}$, together with a distribution π over $[k]$,*

$$\mathrm{D}_{KL}(\mathcal{T}_\mathcal{M}(\mu^n) \| \mathcal{T}_\mathcal{M}(\lambda_\pi^n)) \le O(n\varepsilon^2) \cdot \max_{f \in \mathbb{R}^\mathcal{X} : \|f\|_\infty \le 1} \mathop{\mathbb{E}}_{V \sim \pi} \left[\left(\mathop{\mathbb{E}}_{X \sim \lambda_V}[f_X] - \mathop{\mathbb{E}}_{X \sim \mu}[f_X] \right)^2 \right].$$

In matrix notation, define the matrix $M \in \mathbb{R}^{[K] \times \mathcal{X}}$ by $m_{v,x} = \lambda_v(x) - \mu(x)$. Then,

$$\mathrm{D}_{KL}(\mathcal{T}_\mathcal{M}(\mu^n) \| \mathcal{T}_\mathcal{M}(\lambda_\pi^n)) \le O(n\varepsilon^2) \cdot \|M\|_{\ell_\infty \to L_2(\pi)}^2.$$

Similar to Lemma 2, application of Lemma 3 typically involves showing, for all $v \in [k]$, $\mathcal{M}(\lambda_v^n)$ can be distinguished from $\mathcal{M}(\mu^n)$. This also implies $\mathcal{M}(\lambda_\pi^n)$ can also be distinguished from $\mathcal{M}(\mu^n)$.

Clearly, when restricted to the non-interactive setting, Lemma 3 is weaker than Lemma 2 When we have the restriction of both non-interactivity and a single reference distribution $\mu = \sum_{v \in [k]} \pi(v) \lambda_v$ which is all λ_v mixed according to π, then we may obtain a bound on mutual information, Lemma 4, which will sometimes enable us to get quantitatively stronger sample-complexity lower bounds than those given by our Lemma 2. Note that λ_π^n is a mixture of i.i.d. distributions, whereas μ^n corresponds to i.i.d. samples drawn from a mixture.

Lemma 4 (Theorem 2 of [DJW18]). *Let $\varepsilon = O(1)$, and let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Z}$ be a non-interactive ε-LDP protocol with transcript $\mathcal{T}_\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}^n$. Consider distributions $\lambda_1, \dots, \lambda_k$ on $\mathcal{X}$, together with a distribution π over $[k]$. Then let $\mu = \sum_{v \in [k]} \pi(v) \lambda_v$. For a random dataset $\overline{X} \sim \lambda_v^n$ and a random parameter $V \sim \pi$, then*

$$I\left(\ \mathcal{T}_\mathcal{M}(X)\ ;\ V\ \right) \le O(n\varepsilon^2) \cdot \max_{f \in \mathbb{R}^\mathcal{X} : \|f\|_\infty \le 1} \mathop{\mathbb{E}}_{V \sim \pi} \left[\left(\mathop{\mathbb{E}}_{X \sim \lambda_V}[f_X] - \mathop{\mathbb{E}}_{X \sim \mu}[f_X] \right)^2 \right].$$

In matrix notation, define the matrix $M \in \mathbb{R}^{[K] \times \mathcal{X}}$ by $m_{v,x} = \lambda_v(x) - \mu(x)$. Then,

$$I\left(\ \mathcal{T}_\mathcal{M}(\overline{X})\ ;\ V\ \right) \le O(n\varepsilon^2) \cdot \|M\|_{\ell_\infty \to L_2(\pi)}^2.$$

The advantage of this bound is that it allows us to invoke Fano's inequality which says that, if

$\mathcal{A}(\mathcal{T}_{\mathcal{M}}(\overline{x}))$ is a predictor for V, then

$$\Pr[\mathcal{A}(\mathcal{T}_{\mathcal{M}}(\overline{X})) = V] \le I\left(\ \mathcal{A}(\mathcal{T}_{\mathcal{M}}(\overline{X}))\ ;\ V\ \right) / \log k.$$

Since post-processing $\mathcal{T}_{\mathcal{M}}(\overline{X})$ by $\mathcal{A}$ can only decrease the mutual information with V, this implies

$$\Pr[\mathcal{A}(\mathcal{T}_{\mathcal{M}}(\overline{X})) = V] \le I\left(\ \mathcal{T}_{\mathcal{M}}(\overline{X})\ ;\ V\ \right) / \log k.$$

Supposing that $\mathcal{A}(\mathcal{T}_{\mathcal{M}}(\overline{X})) = V$ holds with probability $\Omega(1)$, then, by rearranging and applying Lemma 4, we obtain

$$n = \Omega\left(\frac{\log k}{\varepsilon^2 \cdot \|M\|^2_{\ell_\infty \to L_2(\pi)}}\right).$$

The factor of $\log k$ in the numerator is essential for obtaining tight lower bounds in some settings. In particular, Appendix B.3 uses this result to get a lower bound against answering parities, a generalization of the unpublished result of [Ull18].

3.4 Total variation bound for pan-privacy

Information theoretic bounds can also play an important role in obtaining lower bounds against pan-privacy. The following lemma generalizes the total variation bound of [CU21] to allow for an arbitrary reference distribution μ, rather than one which is the mixture of the distributions $\{\lambda_v\}_{v\in\mathcal{V}}$, as in that work. The proof is deferred to Appendix A.2.

Lemma 5 (Generalization of [CU21])**.** *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Z}$ be a single-intrusion ε-pan-private protocol. Let $\{\lambda_v\}_{v\in\mathcal{V}}$ consist of distributions on $\mathcal{X}$. Let μ also be a distribution on $\mathcal{X}$. Let π be a distribution on $\mathcal{V}$. Then,*

$$d_{\mathrm{TV}}\left(\mathcal{M}(\lambda_\pi^n), \mathcal{M}(\mu^n)\right) \le O(n\varepsilon) \cdot \max_{f\in\mathbb{R}^{\mathcal{X}}:\|f\|_\infty\le 1} \mathop{\mathbb{E}}_{V\sim\pi}\left[\left(\mathop{\mathbb{E}}_{X\sim\lambda_V}[f_X] - \mathop{\mathbb{E}}_{X\sim\mu}[f_X]\right)^2\right].$$

Note that the expression

$$\max_{f\in\mathbb{R}^{\mathcal{X}}:\|f\|_\infty\le 1} \mathop{\mathbb{E}}_{V\sim\pi}\left[\left(\mathop{\mathbb{E}}_{X\sim\lambda_V}[f_X] - \mathop{\mathbb{E}}_{X\sim\mu}[f_X]\right)^2\right] \tag{3.7}$$

which appears on the right-hand side of the bound on total variation in Lemma 5 is identical to that which appears in the bound on KL-divergence in Lemma 3. As before, if the mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{Z}$ is able to distinguish, with probability $\frac{1}{2} + \Omega(1)$, between a dataset drawn from λ_π^n versus one drawn from μ^n, then $d_{\mathrm{TV}}\left(\mathcal{M}(\lambda_\pi^n), \mathcal{M}(\mu^n)\right) = \Omega(1)$. Once again, by rearranging the bound of Lemma 5, this implies a sample-complexity lower bound in terms of the quantity (3.7).

Chapter 4

Lower bounds against LDP from information-complexity lower bounds

4.1 Overview

In this chapter, we demonstrate how information complexity lower bounds from communication theory readily yield sample complexity lower bounds against interactive LDP. In particular, the lower bound of Theorem 6 against the selection problem, due to collaborator Jonathan Ullman and presented here with his permission, is obtained by combining information complexity lower bounds with the generic mutual-information upper bound given in Lemma 1, which holds for all LDP protocols.

Theorem 6. *Let $\alpha \in (0, 1/8]$ and let $d \in \mathbb{Z}, d \geq 3$. Let $\mathcal{X} = \{0,1\}^d$. Suppose $\mathcal{M} : \mathcal{X}^n \to [d]$ is an ε-LDP protocol such that, given a dataset $\overline{X} \in \mathcal{X}^n$ drawn i.i.d. from an unknown distribution μ on $\{0,1\}^d$ with mean vector $(m_1, \ldots, m_d)$, the output $Z = \mathcal{M}(\overline{X})$ of the protocol satisfies*

$$m_Z \geq \max_{j \in [d]} m_j - \alpha$$

with probability $\Omega(1)$. Then the number of samples required is bounded below by

$$n = \Omega\left(\frac{d}{\alpha\varepsilon}\right).$$

If, in addition, $\mathcal{M}$ is sequential or compositional, then

$$n = \Omega\left(\frac{d}{\alpha^2\varepsilon^2}\right). \tag{4.1}$$

We present a matching upper bound of our own, Theorem 7, which applies the bandit strategy of [AB10] to construct an optimal sequential LDP protocol for selection.

Theorem 7. *Let $\alpha, \beta \in (0, 1]$, $1 - \beta = \Omega(1)$. Let $\mathcal{X} = \{0, 1\}^d$. There is a sequential ε-LDP protocol $\mathcal{M} : \mathcal{X}^n \to [d]$ such that, given a dataset $\overline{X} \in \mathcal{X}^n$, of size*

$$n = O\left(\frac{d \log(1/\beta)}{\alpha^2 \varepsilon^2}\right)$$

drawn i.i.d. from an unknown distribution μ on $\{0, 1\}^d$ with mean vector $(m_1, \ldots, m_d)$, the output $Z = \mathcal{M}(\overline{X})$ of the protocol satisfies

$$m_Z \geq \max_{j \in [d]} m_j - \alpha$$

with probability at least $1 - \beta$.

4.2 Interactive selection lower bound

For a decision problem in the distributed setting which takes a random dataset $\overline{X} \in \mathcal{X}^n$ as input, the minimum mutual information $I\left(\ \overline{X}\ ;\ \mathcal{T}_{\mathcal{M}}(\overline{X})\ \right)$ required of a protocol $\mathcal{M}$ which decides the problem with high probability is known as its *information complexity.* Lower bounds on information complexity are of interest in communication theory because $I\left(\ \overline{X}\ ;\ \mathcal{T}_{\mathcal{M}}(\overline{X})\ \right)$ provides a lower bound on the *entropy* $H\left(\mathcal{T}_{\mathcal{M}}(\overline{X})\right)$, which is itself a lower bound on the length $|\mathcal{T}_{\mathcal{M}}(\overline{X})|$ of the transcript. From our point of view, information complexity lower bounds are useful because they may be combined with the mutual information upper bound of Lemma 1 which says that, when $\mathcal{T}_{\mathcal{M}}$ is the transcript of an ε-LDP protocol, then $I\left(\ \overline{X}\ ;\ \mathcal{T}_{\mathcal{M}}(\overline{X})\ \right) = O(n \cdot \varepsilon)$. In other words,

$$n = \Omega\left(\frac{1}{\varepsilon \cdot I\left(\ \overline{X}\ ;\ \mathcal{T}_{\mathcal{M}}(\overline{X})\ \right)}\right).$$

The task for which we obtain lower bounds against LDP is the selection problem. For this problem, each local agent i holds a data point X_i drawn independently from an underlying distribution λ on $\{0, 1\}^d$ with mean vector $(m_1, \ldots, m_d)$, and the goal of the protocol is to identify a coordinate $Z \in [d]$ which satisfies

$$m_Z \geq \max_{j \in [d]} m_j - \alpha \tag{4.2}$$

where $\alpha > 0$ is the accuracy parameter.

Meanwhile, communication theory's *two-party set disjointness problem* gives players Alice and Bob respective binary strings $A \in \{0, 1\}^d$ and $B \in \{0, 1\}^d$, each being the indicator vector of a subset of $[d]$. The players exchange bits according to a communication protocol to decide whether the sets are disjoint. In other words, they should return `True` if and only if $\not\exists j \in [d],\ A_j \wedge B_j$. Deriving our lower bound for selection involves showing that an LDP protocol for selection gives us a protocol for set disjointness with roughly the same information cost.

Proof of Theorem 6. Suppose we have an ε-LDP protocol $\mathcal{M}$ for n agents which solves selection with accuracy α. Let us construct from $\mathcal{M}$ a 2-party communication protocol $\mathcal{M}'$ for set-disjointness. Each agent in the local protocol is assigned uniformly at random to one of the players Alice and Bob who will communicate by following the local protocol. In particular, if agent i is assigned to Alice, then Alice simulates agent i with input A. Similarly, Bob simulates his assigned agents with his own input B. In this way, the local protocol being simulated receives inputs drawn i.i.d. from the two-point uniform distribution on $\{A, B\}$, which we denote by λ. Let μ be the mean of λ. Then, since $A_j \wedge B_j$ is equivalent to $\mu_j = 1$ while $\neg(A_j \wedge B_j)$ is equivalent to $\mu_j \leq 1/2$, it suffices to achieve (4.2) with $\alpha = 1/8$. Knowing j, then Alice and Bob may exchange A_j and B_j and output `True` precisely when $\neg(A_j \wedge B_j)$.

In this construction, Alice's and Bob's transcript $\mathcal{T}'_{\mathcal{M}}$ follows the transcript $\mathcal{T}_{\mathcal{M}}$ of the local protocol and then exchanges two additional bits, A_j and B_j. If we suppose that $A, B \in \{0,1\}^d$ are jointly distributed random variables, the local protocol is being simulated on the dataset $\overline{X}$ consisting of rows drawn i.i.d. from the uniform distribution on $\{a, b\}$, conditional on $(A, B) = (a, b)$, then

$$\begin{aligned} I\left(\ \mathcal{T}'_{\mathcal{M}}(A,B)\ ;\ (A,B)\ \right) &\leq I\left(\ \mathcal{T}_{\mathcal{M}}(\overline{X})\ ;\ (A,B)\ \right) + 2 \\ &\leq I\left(\ \mathcal{T}_{\mathcal{M}}(\overline{X})\ ;\ \overline{X}\ \right) + 2. \end{aligned} \tag{4.3}$$

The first inequality from the fact that, since $\mathcal{T}'_{\mathcal{M}}(A,B)$ is $\mathcal{T}_{\mathcal{M}}(A,B)$ together with two additional bits, their entropies satisfy the inequality $H(\mathcal{T}'_{\mathcal{M}}(A,B)) \leq H(\mathcal{T}_{\mathcal{M}}(A,B)) + 2$. The second inequality follows from the data processing inequality for mutual information, given that $\mathcal{T}_{\mathcal{M}}(\overline{X})$ is independent of (S, T), conditional on $\overline{X}$.

If $\mathcal{M}$ is a compositional protocol, then Lemma 1 implies

$$I\left(\ \overline{X}\ ;\ \mathcal{T}_{\mathcal{M}}(\overline{X})\ \right) = O(n \cdot \varepsilon^2). \tag{4.4}$$

If $\mathcal{M}$ is not compositional, then Lemma 1 says instead

$$I\left(\ \overline{X}\ ;\ \mathcal{T}_{\mathcal{M}}(\overline{X})\ \right) = O(n \cdot \varepsilon). \tag{4.5}$$

Meanwhile, by the information-complexity lower bound for set disjointness [BJKS04], there exist jointly distributed random variables A and B from $\{0,1\}^d$ such that

$$I\left(\ (A,B)\ ;\ \mathcal{T}'_{\mathcal{M}}(A,B)\ \right) = \Omega(d) \tag{4.6}$$

for all 2-party communication protocols $\mathcal{M}'$ which solve set-disjointness.

Combining (4.3), and (4.6) with (4.4) dependening on whether $\mathcal{M}$ is compositional, gives a sample-complexity lower bound of

$$n = \Omega\left(\frac{d}{\varepsilon^2}\right)$$

for solving selection with accuracy $\alpha = 1/8$ in the compositional case. If $\mathcal{M}$ is not compositional, the using (4.5) in place of (4.4) gives instead

$$n = \Omega\left(\frac{d}{\varepsilon}\right).$$

This argument may be generalized to obtain a lower bound which scales with the accuracy parameter. To do so, we use a simple trick which allows accuracy to be traded for privacy. Suppose we have an ε-LDP protocol $\mathcal{M} : \mathcal{X}^n \to [d]$ which solves selection with accuracy $\alpha \in (0, 1/8]$. From $\mathcal{M}$, we construct the protocol $\mathcal{M}' : \mathcal{X}^n \to [d]$ in the following way. Each agent $i \in [n]$, independently with probability $1 - \eta$, follows the original protocol as if their data point was 0; with probability η, they follow the protocol as usual with their data point X_i. By Lemma 5.5 of [BS15] the protocol $\mathcal{M}'$ is $\eta\varepsilon$-LDP. We can also see that, if $\overline{X}'$ is a data set obtained from $\overline{X}$ by replacing each data point with 0 indepedently with probability $1-\eta$, then $\overline{X}'$ effectively consists of i.i.d. samples from $\{0,1\}^d$ with mean $(m'_1,\dots,m'_d) = (\eta m_1,\dots,\eta m_d)$. We have $\mathcal{M}'(\overline{x}) = \mathcal{M}(\overline{x}')$. Since $\mathcal{M}$ solves the selection problem with accuracy α, then, with probability $1-\beta$, its output $Z \in [d]$ satisfies

$$m'_Z \geq \max_{j\in[d]} m'_j - \alpha.$$

In other words,

$$\eta m_Z \geq \max_{j\in[d]} \eta m_j - \alpha$$

or

$$m_Z \geq \max_{j\in[d]} m_j - \alpha/\eta.$$

Hence, $\mathcal{M}'$ solves selection with accuracy α/η. By taking $\eta = 8\alpha$, then $\mathcal{M}'$ is an $8\alpha\varepsilon$-LDP protocol which solves selection with accuracy $1/8$. Applying our previous lower bound to $\mathcal{M}'$ gives

$$n = \Omega\left(\frac{d}{\alpha^2\varepsilon^2}\right)$$

when $\mathcal{M}$ is compositional. If $\mathcal{M}$ is not compositional, we obtain instead the lower bound

$$n = \Omega\left(\frac{d}{\alpha\varepsilon}\right).$$

□

4.3 Sequential selection upper bound

We give an upper bound, Theorem 7, for the selection problem under sequential ε-LDP, which agrees with the lower bound (4.1). This result relies on the multi-armed bandit strategy of [AB10]. In particular we use their 'pseudo-regret' bound for what they refer to as the 'stochastic adversary.'

Proof of Theorem 7. Let λ be a distribution over $\{0,1\}^d$ with mean vector $(m_1,\dots,m_d)$. Consider the dataset $\overline{X} = (X_1,\dots,X_n)$ with entries drawn i.i.d. from λ. For $i\in[n], j\in[d]$, let $W_{i,j}$ be the result of applying *randomized response* to $X_{i,j}$ according to

$$\begin{bmatrix}\Pr(W_{i,j}=0 \mid X_{i,j}=0) & \Pr(W_{i,j}=+1 \mid X_{i,j}=0)\\ \Pr(W_{i,j}=0 \mid X_{i,j}=1) & \Pr(W_{i,j}=1 \mid X_{i,j}=1)\end{bmatrix} = \begin{bmatrix}\frac{e^\varepsilon}{1+e^\varepsilon} & \frac{1}{1+e^\varepsilon}\\ \frac{1}{1+e^\varepsilon} & \frac{e^\varepsilon}{1+e^\varepsilon}\end{bmatrix}.$$

For any single chosen value of $j\in[d]$, releasing $W_{i,j}$ is ε-DP as a function of the data point X_i.

We consider a local protocol which, on step i, selects an index $J_i \in [d]$ and has agent i report W_{i,J_i}. In particular, we apply the result of [AB10] to obtain a strategy for the choices J_i satisfying

$$\max_{j_0\in[d]}\left\{\mathbb{E}\left[\sum_{i\in[n]} W^{(i)}_{j_0} - \sum_{i\in[n]} W^{(i)}_{J_i}\right]\right\} \leq 25\sqrt{nd} \tag{4.7}$$

where expectation is taken with respect to both the randomness of each J_i as well as of each $W_{i,j}$. This inequality bounds the *pseudo-regret*, namely the difference between the expected reward of our choices J_t and the expected reward of a single choice of $j_0 \in [d]$ which performs as well as possible over all rounds.

Note that

$$\mathbb{E}\left[2W_{i,j} - 1\right] = \left(\frac{e^{\varepsilon} - 1}{e^{\varepsilon} + 1}\right) \cdot \mathbb{E}\left[2X_{i,j} - 1\right]$$

$$\mathbb{E}\left[2W_{i,j} - 1\right] = \left(\frac{e^{\varepsilon} - 1}{e^{\varepsilon} + 1}\right) \cdot \mathbb{E}\left[2X_{i,J_i} - 1\right].$$

so, substituting into (4.7), we get

$$\max_{j_0 \in [d]} \left\{ \mathbb{E}\left[\sum_{i \in [n]} X_{i,j_0} - \sum_{i \in [n]} X_{i,J_i} \right] \right\} \leq 25 \left(\frac{e^{\varepsilon} + 1}{e^{\varepsilon} - 1}\right) \sqrt{nd}.$$

In other words,

$$\max_{j_0 \in [d]} \left\{ \sum_{i \in [n]} \left(m_{j_0} - \mathop{\mathbb{E}}_{J_i}\left[m_{J_i}\right] \right) \right\} \leq 25 \left(\frac{e^{\varepsilon} + 1}{e^{\varepsilon} - 1}\right) \sqrt{nd}. \tag{4.8}$$

When

$$\max_{j_0 \in [n]} \left\{ \frac{1}{n} \sum_{i \in [n]} \mathbb{I}\left[m_{j_0} - \mathop{\mathbb{E}}_{J_i}\left[m_{J_i}\right] \geq \kappa \right] \right\} \geq \beta,$$

then

$$\max_{j_0 \in [d]} \left\{ \sum_{i \in [n]} \left(m_{j_0} - \mathop{\mathbb{E}}_{J_i}\left[m_{J_i}\right] \right) \right\} \geq n\beta\kappa.$$

By (4.8), this implies $n\kappa\beta < 25\left(\frac{e^{\varepsilon}+1}{e^{\varepsilon}-1}\sqrt{nd}\right)$, or equivalently

$$n \leq \frac{625\left(\frac{e^{\varepsilon}+1}{e^{\varepsilon}-1}\right)^2 d}{\kappa^2\beta^2}.$$

Conversely, when

$$n > \frac{625\left(\frac{e^{\varepsilon}+1}{e^{\varepsilon}-1}\right)^2 d}{\kappa^2\beta^2}, \tag{4.9}$$

which we assume from now on, then

$$\max_{j_0 \in [n]} \left\{ \frac{1}{n} \sum_{i \in [n]} \mathbb{I}\left[m_{j_0} - \mathop{\mathbb{E}}_{J_i}\left[m_{J_i}\right] \geq \kappa \right] \right\} < \beta.$$

Let $j_{\max}$ be the value of j_0 which witness the maximum above. For values $i \in [n]$ where

$$m_{j_{\max}} - \mathop{\mathbb{E}}_{J_i}\left[m_{J_i}\right] < \kappa,$$

then, by the non-negativity of $m_{j_{\max}} - m_j$ for all $j \in [d]$, we may apply Markov's inequality to obtain

$$\mathop{\mathbb{P}}_{J_i}\left[m_{j_{\max}} - m_{J_i} \geq k\kappa,\right] \leq \frac{1}{k}$$

for $k \geq 1$. Taking $k = 1/\beta$ and $\kappa = \alpha\beta$ gives

$$\mathop{\mathbb{P}}_{J_i}\left[m_{j_{\max}} - m_{J_i} \geq \alpha\right] \leq \beta$$

This motivates us to consider the algorithm which, after running the bandits process, selects $T \in [n]$ uniformly at random, and then returns J_T. Our previous calculations show

$$\mathop{\mathbb{P}}_{T,J_T}\left[m_{j_{\max}} - m_{J_T} < \alpha\right] \geq (1-\beta)^2 \geq 1 - 2\beta.$$

By (4.9), this is achieved with the number of samples being at most

$$n = O\left(\frac{625\left(\frac{e^\varepsilon+1}{e^\varepsilon-1}\right)^2 d}{\alpha^2\beta^2}\right) = O\left(\frac{d}{\alpha^2\varepsilon^2\beta^2}\right).$$

We can improve the dependence on β by applying this selection process $r = O(\log(1/\beta'))$ times with $\beta = \frac{1}{2}$. A Chernoff bound allows us to guarantee with probability at least $1 - \beta'$ that at least one repitition of the selection algorithm returns an index $j \in [d]$ which satisfies $m_{j_{\max}} - m_j < \alpha$. We may estimate m_j for each of the r choices of j within $\alpha/2$ with probability at least $1 - \beta'$ using

$$O\left(\frac{r\log(1/\beta)}{\varepsilon^2\alpha^2}\right) = \widetilde{O}\left(\frac{\log(\frac{1}{\beta})}{\varepsilon^2\alpha^2}\right)$$

samples. This will allow us to identify some $j \in [d]$ which satisfies $m_{j_{\max}} - m_j < 2\alpha$.

In total, this algorithm requires at most

$$n' = O\left(\frac{d\log(1/\beta')}{\alpha^2\varepsilon^2}\right) + \widetilde{O}\left(\frac{\log(\frac{1}{\beta})}{\varepsilon^2\alpha^2}\right) = O\left(\frac{d\log(1/\beta')}{\alpha^2\varepsilon^2}\right)$$

samples.

□

Chapter 5

Characterization of statistical query release under non-interactive LDP

5.1 Overview

In this chapter, we characterize the number of samples required to answer a workload of *statistical queries* (also called *linear queries)* [Kea93] under non-interactive local differential privacy.

The simplest example of a statistical query is "What fraction of individuals in the data have property P?" Workloads of statistical queries capture a variety of statistical tasks: computing histograms and PDFs, answering range queries and computing CDFs, estimating the mean, computing correlations and higher-order marginals, and estimating the risk of a classifier.

The power of differentially private algorithms for answering a *worst-case* workload of statistical queries is well understood [BUV14, DJW18], and known bounds are essentially tight as a function of the data set size, the data domain, and the size of the workload. However, many workloads, such as those corresponding to computing PDFs or CDFs, have additional structure that makes it possible to answer them with less error than these worst-case workloads. Thus, a central question is

> *Can we characterize the amount of error required to estimate a given workload of statistical queries subject to differential privacy in terms of natural properties of the workload, and can we achieve this error via computationally efficient algorithms?*

In the central model, there has been dramatic progress on this question [HT10, BDKT12, NTZ16, Nik15, BBNS19], giving approximate characterizations for every workload of statistical queries. We extend this line of work by giving the first approximate characterization for the *non-interactive local model of differential privacy* [DMNS06, KLN+08]. This result is also much sharper than analogous results for the central model of differential privacy. In particular, we give a generalization of the natural and well studied *factorization mechanism.* We show our variant, called *approximate factorization mechanism* to be nearly optimal. Factorization mechanisms capture a number of special-purpose mechanisms from the theory literature [BCD+07, DNPR10, CSS11, TUV12, CTUW14], were involved in previous characterizations, and also roughly capture the *matrix mechanisms* [LHR+10, MMHM18] from the databases literature, which have been developed into practical algorithms for US Census Data.

To answer an arbitrary workload $Q \subseteq [-1,1]^{\mathcal{X}}$, represented by its workload matrix $W \in [-1,1]^{Q\times\mathcal{X}}$ with entries $w_{q,x} = q(x)$, we consider a factorization $W = RA$ of the workload matrix. The local randomizer of [BBNS19] satisfies 'pure' ε-DP while granting similar concentration properties to the Gaussian mechanism, which satisfies merely 'approximate' (ε,δ)-LDP. Using this local randomizer to answer the workload of queries given by the matrix A allows us to obtain $Ah+G$ under non-interactive LDP, where G is some random noise. Multiplying by R reconstructs an approximation $RAh + RG = Wh + RG$ to the original workload. The number of samples which this procedure requires to achieve given accuracy and privacy parameters is affected by the choice of A since it represents the initial workload being answered. The number of required samples also depends on the choice of R, since it determines how the noise G is scaled. In particular, we will see that this approach allows a workload Q of size $k = |Q|$ to be answered under ε-LDP with accuracy α using n samples where

$$n = O\left(\frac{\|R\|_{2\to\infty}^2 \|A\|_{1\to 2}^2 \log k}{\varepsilon^2\alpha^2}\right).$$

Optimizing for the choice of R and A gives the sample complexity bound

$$n = O\left(\frac{\gamma_2(W)^2 \log k}{\varepsilon^2\alpha^2}\right)$$

where

$$\gamma_2(W) = \min\{\|R\|_{2\to\infty}\|A\|_{1\to 2} : W = RA\}.$$

This approach, called the *factorization mechanism*, was discovered in [LHR$^+$10]. We give it a slight generalization here. In particular, for any matrix $\widetilde{W}$ satisfying $\|\widetilde{W} - W| \leq \alpha$, answering $\widetilde{W}$ with accuracy α gives answers of accuracy 2α for W. Our *approximate factorization mechanism* optimizes for the choice of $\widetilde{W}$ and applies the *factorization mechanism* to $\widetilde{W}$ rather than W, to give a sample-complexity upper bound in terms of

$$\gamma_2(W,\alpha) = \min\{\gamma_2(\widetilde{W}) : \|W - \widetilde{W}\|_{1\to\infty} \leq \alpha/2\}.$$

This leads to the following result.

Theorem 8 (Approximate Factorization Mechanism)**.** *There exists an ε-LDP mechanism $\mathcal{M}_{\gamma_2}$ such that, for any workload $Q \subseteq [-1,1]^{\mathcal{X}}$ of size $|Q| = k$, with workload matrix $W \in [-1,1]^{Q\times\mathcal{X}}$, we have*

$$\mathrm{sc}(\mathcal{M}^{loc}_{\gamma_2,\alpha}, Q, \alpha) = O\left(\frac{\gamma_2(W,\alpha/2)^2 \log k}{\varepsilon^2\alpha^2}\right),$$

and the mechanism runs in time which is polynomial in n, k, and $|\mathcal{X}|$.

The focus of this chapter will be in showing that this mechanism is nearly optimal among all non-interactive locally differentially private mechanisms. The following result an informal version of our main lower bound which hides some technicalities.

Theorem 9 (Informal)**.** *Let $\alpha, \varepsilon, > 0$ be smaller than some absolute constants and let Q be a* *workload of statistical queries with workload matrix W. Then, for some $\alpha' = \Omega(\alpha/\log(1/\alpha))$,*

$$\mathrm{sc}^{\ell_\infty}_{\varepsilon\text{-LDP}}(Q, \alpha') = \Omega\left(\frac{\gamma_2(W, \alpha/2)^2}{\varepsilon^2\alpha^2}\right).$$

To interpret the theorem, it helps to start by imagining that $\gamma_2(W, \alpha'/2) = \gamma_2(W, \alpha/2)$, in which case the theorem would show that the sample complexity of answering queries up to error α' is

$$\Omega\left(\frac{\gamma_2(W, \alpha'/2)^2}{\varepsilon^2\alpha^2}\right),$$

which differs from the sample complexity of the approximate factorization mechanism, given in Theorem 23, by a factor of just $O(\log(1/\alpha')^2 \log|Q|)$. The fact that we take $\alpha' < \alpha$ means that $\gamma_2(W, \alpha/2)$ can be much smaller than $\gamma_2(W, \alpha'/2)$.[1] Nevertheless, for many natural families of queries and choices of α, $\gamma_2(W, \alpha/2)$ will be relatively stable to small changes in α, in which case our lower bound will be tight up to this $O(\log(1/\alpha)^2 \log|Q|)$ factor. In contrast, existing characterizations for the central model [HT10, BDKT12, NTZ16, Nik15, BBNS19] lose a $\mathrm{poly}(1/\alpha)$ factor, or else they lose a $\mathrm{polylog}|\mathcal{X}|$ factor that is typically large.

Remark 10. *Our proof of Theorem 9, in fact, shows that the lower bound holds in the distributional setting where $\overline{X}$ is sampled i.i.d. from an unknown distribution μ, and the goal is to estimate the quantity $q(\mu) = \mathop{\mathbb{E}}_{X\sim\mu}[q(X)]$ for every query $q \in Q$ up to error at most α.*

Using Theorem 9, we obtain new lower bounds for three well studied families of queries:

1. *Threshold queries,* which are also known as range queries, and equivalent to computing the CDF of the data.

2. *Parity queries,* which capture the covariance and higher-order moments of the data.

3. *Marginal queries,* also known as conjunctions, which capture the marginal distribution on subsets of the attributes.

The research represented by this chapter was originally originally published in [ENU19], and is joint work with Aleksandar Nikolov and Jonathan Ullman.

5.2 Upper bound

Here we give details of the approximate factorization mechanism, which was sketched in the introduction. We derive sample-complexity characterizations in terms of the factorization norm

$$\widehat{\gamma}_2(W, \alpha) = \min\{\gamma_2(\widehat{W}, \alpha) \ : \ \widehat{W} = W + c\mathbf{1}^T, c \in \mathbb{R}^k\}, \tag{5.1}$$

where $\mathbf{1}^T$ is the horizontal all-ones vector of length k so that $\widehat{W} = W + c\mathbf{1}^T$ is obtained by shifting each row q of W by a distinct constant c_q. The factorization norm $\gamma_2(\widehat{W}, \alpha)$ is given by

$$\gamma_2(\widehat{W}, \alpha) = \min\{\gamma_2(\widetilde{W}) : \|\widehat{W} - \widetilde{W}\|_{1\to\infty} \leq \alpha\}, \tag{5.2}$$

[1] For example, if every entry of W is at most α in absolute value, then $\gamma_2(W, \alpha) = 0$ whereas $\gamma_2(W, \alpha')$ can be arbitrarily large for $\alpha' < \alpha$, but this behavior typically does not happen for "non-trivial" values of α.

where

$$\gamma_2(\widetilde{W}) = \min\{\|R\|_{2\to\infty}\|A\|_{1\to 2} \ : \ \widetilde{W} = RA\}. \tag{5.3}$$

Matrices $\widehat{W}, \widetilde{W}$, R, and A achieving the minimum to any degree of accuracy can be computed in polynomial time via semidefinite programming, as shown in [LS09]. Although we initially derive upper and lower bounds for answering the queries associated with W in terms of $\widehat{\gamma}_2(W,\alpha)$, we will ultimately translate these upper and lower bounds so that they are expressed in terms of $\gamma_2(W,\alpha)$. This is achieved by noting that $\widehat{\gamma}_2(W,\alpha)$ and $\gamma_2(W,\alpha)$ are approximately equal when the queries associated with W are bound, which is to say $W \in [-1,1]^{Q,\mathcal{X}}$.

Our main positive result shows that the sample complexity of the corresponding approximate factorization mechanism is bounded above by the approximate γ_2 norm. As sketched in the introduction, this can be achieved via a local version of the Gaussian noise mechanism, which can then be transformed into a purely private mechanism using the results of [BNS18]. This gives a slightly suboptimal bound however, so instead we use the local randomizer from [BBNS19], which is a variant of a local randomizer from [DJW18]. The relevant properties of this local randomizer are captured by the next lemma. We recall that a random variable Z over $\mathbb{R}$ is σ-subgaussian if $\mathbb{E}\exp(Z^2/\sigma^2) \le 2$, and a random variable Z over $\mathbb{R}^d$ is σ-subgaussian if $\theta^\top Z$ is σ-subgaussian for every vector θ such that $\|\theta\|_2 = 1$.

Lemma 11 ([BBNS19])**.** *There exists an ε-DP mechanism $\mathcal{M}$ which takes as input a single data point $x \in \mathbb{R}^d$ such that $\|x\|_2 \le 1$, and outputs a random $Y_x := \mathcal{M}(x) \in \mathbb{R}^d$ such that*

1. *Y_x can be sampled in time polynomial in d on input x,*
2. *$\mathbb{E}[Y_x] = x$,*
3. *$Y_x - x$ is σ-subgaussian with $\sigma = O(\varepsilon^{-1})$.*

Given this local randomizer, and approximate factorizations, we are ready to prove our upper bound.

Theorem 12 (Approximate Factorization Mechanism)**.** *There exists an ε-LDP mechanism $\mathcal{M}_{\widehat{\gamma}_2}$ such that, for any k statistical queries Q with workload matrix W, we have*

$$\mathrm{sc}^{\ell_\infty}(\mathcal{M}_{\widehat{\gamma}_2}, Q, \alpha) = O\left(\frac{\widehat{\gamma}_2(W,\alpha/2)^2 \log k}{\varepsilon^2\alpha^2}\right),$$

and the mechanism runs in time polynomial in n, k, and $|\mathcal{X}|$.

Proof. Let $\widehat{W}$, $\widetilde{W}$, R and A witness (5.1), (5.2) and (5.3) so that $\widehat{W} = W + c\mathbf{1}^T$ for some $c \in \mathbb{R}^k$, $\widetilde{W} \le \alpha/2$, $W = RA$, and $\|R\|_{2\to\infty}\|A\|_{1\to 2} = \widehat{\gamma}_2(W,\alpha)$. Without loss of generality, we may assume $\|A\|_{1\to 2} = 1$ and $\|R\|_{2\to\infty} = \widehat{\gamma}_2(W,\alpha)$ since tR and A/t are valid witnesses whenever R and A are. Moreover, $\widehat{W}$, $\widetilde{W}$, R and A can be computed in polynomial time via semidefinite programming, as noted above.

Let $\overline{x} = (x_1, \ldots, x_n)$ be the data set, and let h be its histogram. Each agent i holds a data point x_i, and the histogram for the single point data set containing x_i is $h_i := e_{x_i}$, i.e. the standard basis vector of $\mathbb{R}^{\mathcal{X}}$ corresponding to x_i. Agent i releases $\mathcal{M}_i(x_i)$ according to the local randomizer of

Lemma 11. Then $Y = \frac{1}{n}\sum_{i=1}^{n} \mathcal{M}_i(x_i)$ has expectation

$$\frac{1}{n}\sum_{i=1}^{n} Ah_i = \frac{1}{n}A\sum_{i=1}^{n} h_i = \frac{1}{n}Ah.$$

Moreover, since $Y - \frac{1}{n}Ah = \frac{1}{n}\sum_{i=1}^{n}(Y_{Ah_i} - Ah_i)$ is the average of n independent σ-subgaussian random variables, where $\sigma = O(\varepsilon^{-1})$ is as in Lemma 11, $Y - \frac{1}{n}Ah$ is $O\left(\frac{\sigma}{\sqrt{n}}\right)$-subgaussian (see e.g. [Ver18, Proposition 2.6.1.]).

Post-processing Y by our reconstruction matrix R gives an approximation of $Wh = RAh$. In particular,

$$\mathbb{E}\left[RY\right] = RAh = \frac{1}{n}\widetilde{W}h.$$

Since $\|\widetilde{W} - \widehat{W}\|_\infty \le \alpha/2$, then

$$\left\|\mathbb{E}[RY] - \frac{1}{n}\widehat{W}h\right\|_\infty = \frac{1}{n}\|(\widetilde{W} - \widehat{W})h\|_\infty \le \frac{1}{n}\|\widetilde{W} - \widehat{W}\|_{1\to\infty}\|h\|_1 \le \frac{\alpha}{2}. \tag{5.4}$$

Every coordinate of $R(Y - \frac{1}{n}Ah) = RY - \mathbb{E}[RY]$ is the inner product of $Y - \frac{1}{n}Ah$ and a row of R, the latter having ℓ_2 norm at most $\|R\|_{2\to\infty}$. Since $Y - \frac{1}{n}Ah$ is $O\left(\frac{\sigma}{\sqrt{n}}\right)$-subgaussian, every coordinate $(RY - \mathbb{E}[RY])_q$ for every $q \in Q$, is $O\left(\frac{\sigma\|R\|_{2\to\infty}}{\sqrt{n}}\right)$-subgaussian. It is then a standard fact (see e.g. [Ver18, Exercise 2.5.10]) that

$$\mathbb{E}\|RY - \mathbb{E}[RY]\|_\infty = O\left(\frac{\sigma\|R\|_{2\to\infty}\sqrt{\log k}}{\sqrt{n}}\right) = O\left(\frac{\widehat{\gamma}_2(W,\alpha)\sqrt{\log k}}{\varepsilon\sqrt{n}}\right).$$

Combining with (5.4), and applying the triangle inequality, we get

$$\begin{aligned}\mathbb{E}\|RY - \widehat{W}h\|_\infty &\le \|\mathbb{E}[RY] - \tfrac{1}{n}\widehat{W}h\|_\infty + \mathbb{E}\|RY - \mathbb{E}[RY]\|_\infty \\ &= \frac{\alpha}{2} + O\left(\frac{\widehat{\gamma}_2(W,\alpha/2)\sqrt{\log k}}{\varepsilon\sqrt{n}}\right).\end{aligned}$$

The number of samples n is chosen so that the second term is at most $\frac{\alpha}{2}$.

The final output of our mechanism is obtained by post-processing RY to reverse the row translations by which $\widehat{W}$ was obtained from W. In particular, define

$$\mathcal{M}(\overline{x}) = RY - c\mathbf{1}^T.$$

Then,

$$\mathbb{E}\|\mathcal{M}(\overline{x}) - Wh\|_\infty = \mathbb{E}\|(RY - c\mathbf{1}^T) - (\widehat{W}h - c\mathbf{1}^T)\|_\infty = \mathbb{E}\|RY - \widehat{W}h\|_\infty \le \alpha.$$

□

5.3 Lower bound

In this section, we present our lower bound against answering statistical queries under non-interactive LDP. For notational convenience, we will assume that the queries in our workload are enumerated, so that $Q = \{q_1, \ldots, q_k\}$. Our lower bound will rely on constructing, for each query q_v, a pair of 'hard' distributions λ_v and μ_v on $\mathcal{X}$. Together with these, we construct a parameter distribution π on $[k]$.

5.3.1 Application KL-divergence bound

We approach the task of showing that $\lambda_1, \ldots, \lambda_k$ and $\mu_1, \ldots, \mu_k$ are hard distributions for Q in two steps. First, we wish to argue that being able to estimate Q on the distributions $\lambda_1, \ldots, \lambda_k$ and $\mu_1, \ldots, \mu_k$ enables us, for each $v \in [k]$, to distinguish between λ_v^n and μ_v^n. Second, we show a lower bound on the number of samples required of a non-interactive LDP mechanism which is able to perform such a distinguishing task. The second of these objectives is shown by way of our KL-divergence bound, Lemma 2.

Being able to distinguish, for all $v \in [k]$, between $\mathcal{T}_{\mathcal{M}}(\lambda_v^n)$ and $\mathcal{T}_{\mathcal{M}}(\mu_v^n)$ with probability $\frac{1}{2} + \Omega(1)$ implies

$$d_{\mathrm{TV}}(\lambda_v^n, \mu_v^n) = \Omega(1),$$

from which it follows by Pinsker's inequality that

$$\mathrm{D}_{\mathrm{KL}}(\mathcal{T}_{\mathcal{M}}(\lambda_v^n) \| \mathcal{T}_{\mathcal{M}}(\mu_v^n)) \geq \Omega(1).$$

Together with Lemma 2, this would imply

$$n = \Omega\left(\frac{1}{\varepsilon^2 \cdot \|M\|^2_{\ell_\infty \to L_2(\pi)}}\right)$$

where $M \in \mathcal{C} \times \mathcal{X}$ is the matrix with entries $m_{c,x} = \lambda_c(x) - \mu_c(x)$. For this reason, our goal will be to define our distributions so that that $\|M\|^2_{\ell_\infty \to L_2(\pi)}$ is small while still meeting the requirement that estimating the queries Q allows us to distinguish between λ_v^n and μ_v^n for all $v \in [k]$.

It is worth noting that Lemma 2 is not known to hold when the protocol is allowed to be sequential. Indeed, this is the bottleneck in generalizing our lower bound to the case of sequential local privacy. If Lemma 2 were to hold for sequential LDP, then our lower bound would apply to that setting. Alternatively, it may be possible to generalize our result to the interactive setting by taking advantage of the closely related KL-divergence bound of Lemma 3. Applying this result would rely on modifying our construction of the hard distributions to make each of the distributions μ_q identical, while the distributions λ_q would still be allowed to be distinct. See the proof of Lemma 2 for further discussion of the technical barriers involved in generalizing this result of to the interactive setting.

5.3.2 Duality for $\widehat{\gamma}_2(W, \alpha)$

Recall that our goal is to prove a lower bound on the sample complexity of mechanisms in the local model in terms of the approximate γ_2 norm. We will do so via Lemma 2, and the distributions $\{\lambda_1, \ldots, \lambda_k\}$ and $\{\mu_1, \ldots, \mu_k\}$ will serve as a certificate of a lower bound on the sample complexity. On the other hand, convex duality can certify a lower bound on the approximate γ_2 norm. In the proof of our lower bounds, we will show that these dual certificates for which the approximate γ_2 norm is large can be turned into hard families of distributions to use in Lemma 2.

The key duality statement follows. Its derivation will closely follow the derivation of the dual of $\gamma_2(W,\alpha)$ which was given in [LS09] for the special case when W has entries in $\{-1,1\}$.[2] See Lemma 32 for the dual of $\gamma_2(W,\alpha)$.

Lemma 13. *For $W \in \mathbb{R}^{k\times T}$ and $\alpha > 0$,*

$$\gamma_2(W,\alpha) = \max\left\{\frac{W\bullet U - \alpha\|U\|_1}{\gamma_2^*(U)} \ : \ U\in\mathbb{R}^{k\times T},\ U\neq 0,\ \forall q\in Q,\ \sum_{x\in\mathcal{X}} u_{q,x} = 0\right\},$$

where γ_2^* is the dual norm to γ_2 given by

$$\begin{aligned}\gamma_2^*(U) &= \max\{U\bullet V : V\in\mathbb{R}^{k\times T},\ \gamma_2(V)\le 1\}\\ &= \max_{\substack{a_1,\dots,a_k\\ b_1,\dots,b_T}} \sum_{i=1}^{k}\sum_{j=1}^{T} u_{i,j} a_i^\top b_j,\end{aligned}$$

where $a_1,\dots,a_k$ and $b_1,\dots,b_T$ range over vectors with unit ℓ_2 norm in $\mathbb{R}^{k+T}$.

Proof of Lemma 13. Consider an arbitrary non-zero matrix $U\in\mathbb{R}^{k\times T}$ satisfying

$$\sum_{x\in[T]} u_{v,x} = 0.$$

By definition of $\widehat{\gamma}_2(W,\alpha)$, there exists some $c\in\mathbb{R}^k$ such that $\widehat{W} = W + c\mathbf{1}^T$ achieves $\widehat{\gamma}_2(W,\alpha) = \gamma_2(\widehat{W},\alpha)$, and there also exists a matrix $\widetilde{W}$ satisfying $\|\widehat{W}-\widetilde{W}\|_{1\to\infty}\le\alpha$ such that $\gamma_2(\widehat{W},\alpha) = \gamma_2(\widetilde{W})$. It follows that

$$\begin{align} W\bullet U &= \widehat{W}\bullet U - c\mathbf{1}^T\bullet U \notag\\ &= \widehat{W}\bullet U \tag{5.5}\\ &= \widetilde{W}\bullet U + (\widehat{W}-\widetilde{W})\bullet U \notag\\ &\le \gamma_2(\widetilde{W})\gamma_2^*(U) + \|\widehat{W}-\widetilde{W}\|_{1\to\infty}\|U\|_1 \tag{5.6}\\ &\le \gamma_2(\widehat{W},\alpha)\gamma_2^*(U) + \alpha\|U\|_1 \notag\\ &= \widehat{\gamma}_2(W,\alpha)\gamma_2^*(U) + \alpha\|U\|_1. \notag\end{align}$$

Equality (5.5) follows from the fact that the each row of U sums to zero since this implies $c\mathbf{1}^T\bullet U = 0$. Inequality (5.6) follows by the trivial case of Hölder's inequality, and the definition of γ_2^*. Rearranging gives

$$\widehat{\gamma}_2(W,\alpha) \ge \frac{W\bullet U - \alpha\|U\|_1}{\gamma_2^*(U)}.$$

Taking the supremum over all choices of U gives

$$\widehat{\gamma}_2(W,\alpha) \ge \sup\left\{\frac{W\bullet U - \alpha\|U\|_1}{\gamma_2^*(U)} : U\in\mathbb{R}^{k\times T},\ U\neq 0,\ \sum_{x\in[T]} u_{v,x} = 0\right\}.$$

Let us now show that this inequality in the other direction. We do so by showing that, for $t\ge 0$,

[2] Note that in [LS09], Linial and Shraibman use the notation $\gamma_2^\alpha(W) = \inf\{\gamma_2(\widetilde{W}) : 1\le \widetilde{w}_{ij}w_{ij}\le\alpha \ \ \forall i,j\}$. For sign matrices W, this is equal to $\frac{\alpha+1}{2}\gamma_2(W,(\alpha-1)/(\alpha+1))$ in our notation.

if $\widehat{\gamma}_2(W,\alpha) > t$, then there exists some non-zero $U \in \mathbb{R}^{k\times T}$ with entries satisfying $\sum_{x\in[T]} u_{v,x} = 0$ such that $\frac{W\bullet U-\alpha\|U\|_1}{\gamma_2^*(U)} > t$. Let

$$S = \{B \in \mathbb{R}^{k\times T} : \gamma_2(B) \le t\}$$

and

$$T = \{B \in \mathbb{R}^{k\times T} : \exists c \in \mathbb{R}^k,\ \|W - (B - c\mathbf{1}^T)\|_{1\to\infty} \le \alpha\}.$$

Then $\gamma_2(W,\alpha) > t$ equivalently means $S \cap T = \emptyset$. Since both S and T are convex and compact, and $S\cap T = \emptyset$, the hyperplane separator theorem [Roc97, Corollary 11.4.2] implies that there is a hyperplane separating them, i.e. there is a matrix $U \in \mathbb{R}^{k\times T} \setminus \{0\}$ such that

$$\max\{B\bullet U : B \in S\} < \min\{B \bullet U : B \in T\}. \tag{5.7}$$

The left-hand side equals $t\gamma_2^*(U)$, by definition. The right-hand side equals

$$\begin{aligned}
&\min\{B\bullet U : B\in T\}\\
&= \min\{(B' + c\mathbf{1}^T)\bullet U : c\in\mathbb{R}^k,\ B'\in\mathbb{R}^{k\times T},\ \|W-B'\|_{1\to\infty}\le\alpha\}\\
&= \min\{W\bullet U - (W-B')\bullet U + c\mathbf{1}^T\bullet U : c\in\mathbb{R}^k,\ B'\in\mathbb{R}^{k\times T},\ \|W-B'\|_{1\to\infty}\le\alpha\}\\
&= W\bullet U - \max\{(W-B')\bullet U : B'\in\mathbb{R}^{k\times T},\ \|W-B'\|_{1\to\infty}\le\alpha\} + \min\{c\mathbf{1}^T\bullet U : c\in\mathbb{R}^k\}\\
&= W\bullet U - \alpha\|U\|_1 + \min\{c\mathbf{1}^T\bullet U : c\in\mathbb{R}^k\}
\end{aligned}$$

where the last equality again uses the trivial case of Hölder's inequality. In short, (C.1) implies

$$t\gamma_2^*(U) < W\bullet U - \alpha\|U\|_1 + \min\{c\mathbf{1}^T\bullet U : c\in\mathbb{R}^k\}.$$

This inequality implies that the entries of U satisfy $\sum_{x\in[T]} u_{v,x} = 0$ since otherwise

$$\min\{c\mathbf{1}^T\bullet U : c\in\mathbb{R}^k\} = -\infty,$$

which would contradict the fact that the left-hand side is non-negative. It follows that $c\mathbf{1}^T\bullet U = 0$ for all $c\in\mathbb{R}^k$, and hence

$$t\gamma_2^*(U) < W\bullet U - \alpha\|U\|_1.$$

Therefore, (C.1) is equivalent to $t\gamma_2^*(U) < W\bullet U - \|U\|_1$. We have shown, for all $t\ge 0$, that, whenever $\widehat{\gamma}_2(W,\alpha) > t$, then there exists some non-zero $U\in\mathbb{R}^{k\times T}$ with entries satisfying $\sum_{x\in[T]} u_{v,x} = 0$ such that $\frac{W\bullet U-\alpha\|U\|_1}{\gamma_2^*(U)} > t$. Since $\widehat{\gamma}_2(W,\alpha)\ge 0$, this implies

$$\widehat{\gamma}_2(W,\alpha) \le \sup\left\{\frac{W\bullet U - \alpha\|U\|_1}{\gamma_2^*(U)} : U\in\mathbb{R}^{k\times T},\ U\ne 0,\ \sum_{x\in[T]} u_{v,x} = 0\right\}.$$

□

The expression

$$\gamma_2^*(U) = \max\sum_{i=1}^{k}\sum_{j=1}^{T} u_{i,j} a_i^\top b_j,$$

with the max over unit vectors $a_1, \dots a_k$ and $b_1, \dots, b_T$ can be easily formulated as a semidefinite program, and, in fact, is exactly the semidefinite program that appears in Grothendieck's inequality (see, e.g., [KN12, Pis12]). It is straightforward to check (just take all the a_i and b_j co-linear) that

$$\gamma_2^*(U) \geq \max\{y^\top U z : y \in \{-1,1\}^m, z \in \{-1,1\}^N\} = \|U\|_{\infty\to 1}. \tag{5.8}$$

Moreover, Grothendieck showed that this inequality is always tight up to a universal constant [Gro53], although this fact will not be used here. Instead, we will need the following lemma, which can be derived from SDP duality, and is also due to Grothendieck. For a proof using the Hahn-Banach theorem, see [Pis12].

Lemma 14 ([Gro53])**.** *For any $k \times T$ matrix U, $\gamma_2^*(U) \leq t$ if and only if there exist diagonal matrices $P \in \mathbb{R}^{k\times k}$ and $Q \in \mathbb{R}^{T\times T}$, and a matrix $\widetilde{U} \in \mathbb{R}^{k\times T}$ such that $\mathrm{Tr}(P^2) = \mathrm{Tr}(Q^2) = 1$, $U = P\widetilde{U}Q$, and $\|\widetilde{U}\|_{2\to 2} \leq t$.*

By (5.8), the $\gamma_2^*(\cdot)$ norm is an upper bound on the $\|\cdot\|_{\infty\to 1}$ norm. We use Lemma 14 to show a similar upper bound on the $\|\cdot\|_{\infty\to 2}$, which allows projecting out some of the rows of the matrix, but is quantitatively stronger. The reason we are interested in the $\|\cdot\|_{\infty\to 2}$ norm is that this is the norm that appears in the statement of Lemma 2.

Lemma 15. *For any matrix $U \in \mathbb{R}^{k\times T}$, there exists a set $S \subseteq [k]$ of size $|S| \geq \frac{k}{2}$ such that $\sqrt{\frac{k}{2}}\|\pi_S U\|_{\infty\to 2} \leq \gamma_2^*(U)$, where π_S is the projection onto the subspace $\mathbb{R}^S$.*

The next lemma slightly strengthens Lemma 15 to allow for weights on the rows of the matrix. This is the key fact about the γ_2^* norm that we need for our lower bounds.

Lemma 16. *Let U and M be $k \times T$ matrices, and let π be a probability distribution on $[k]$ where, for any $i \in [k], j \in [T]$, we have $u_{i,j} = \pi(i) m_{i,j}$. Then there exists a probability distribution $\widehat{\pi}$ on $[k]$, with support contained in the support of π, such that $\|M\|_{\ell_\infty \to L_2(\widehat{\pi})} \leq 4\gamma_2^*(U)$.*

Lemmas 15 and 16 are proved in Appendix B.1.

5.3.3 Construction of hard distributions based on dual solution

In this section we put together the different tools we have already set up – the KL-divergence lower bound, and the duality of the approximate γ_2 norm – in order to prove our main lower bound, Theorem 9.

For this section, it is convenient to consider the enumeration $q_1, \dots, q_k$ of the queries of a workload Q with workload matrix $W \in \mathbb{R}^{[k]\times\mathcal{X}}$. Let $U \in \mathbb{R}^{k\times T}$ be the dual witness to the lower bound on $\gamma_2(W, \alpha)$, as given by Lemma 13, so that

$$\widehat{\gamma}_2(W, \alpha) = \frac{W \bullet U - \alpha\|U\|_1}{\gamma_2^*(U)} \tag{5.9}$$

while each row $v \in [k]$ of U has entries with sum $\sum_{x\in\mathcal{X}} u_{v,x} = 0$. By dividing each entry of U by $\|U\|_1$ if necessary, then we may assume without loss of generality that $\|U\|_1 = 1$. In this case,

$$\widehat{\gamma}_2(W, \alpha) = \frac{W \bullet U - \alpha}{\gamma_2^*(U)}. \tag{5.10}$$

Let us make a first attempt at constructing our collection of 'hard' distributions $\lambda_1, \ldots, \lambda_k$ and $\mu_1, \ldots, \mu_k$ for Q. Since $\|U\|_1 = 1$, then

$$\pi(v) = \sum_{x \in \mathcal{X}} |u_{v,x}| \tag{5.11}$$

defines a valid probability distribution over $[k]$. For each $v \in [k]$, we then define a pair of distributions λ_v and μ_v given, for $x \in \mathcal{X}$, by

$$\lambda_v(x) = \begin{cases} 2|u_{v,x}|/\pi(v) & \text{if } u_{v,x} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_v(x) = \begin{cases} 2|u_{v,x}|/\pi(v) & \text{if } u_{v,x} \leq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.12}$$

In particular, each λ_v is defined by the positive entries of row v of U while each μ_v is defined by the negative entries of row v of U. Since $\|U\|_1 = 1$ while each row of U sums to zero, these define valid probability distributions on $\mathcal{X}$.

Moreover,

$$W \bullet U = \mathop{\mathbb{E}}_{V \sim \pi} [q_V(\lambda_V) - q_V(\mu_V)].$$

If $\widehat{\gamma}_2$ is positive, then the numerator in its dual formulation (5.10) must be positive. Hence $W \bullet U > 0$ and thereby

$$\mathop{\mathbb{E}}_{V \sim \pi} [q_V(\lambda_V) - q_V(\mu_V)] \geq \alpha.$$

Suppose, instead of holding on average over $V \sim \pi$, this inequality were to hold in the worst case over all $v \in [k]$, so that

$$q_v(\lambda_v) - q_v(\mu_v) \geq \alpha.$$

This implies that a mechanism $\mathcal{M}$ answering q_v with accuracy $\alpha/4$ would have to output different answers for λ_v versus μ_v. Doing so with probability $\Omega(1)$ would imply

$$\mathrm{D_{KL}}(\mathcal{T}_{\mathcal{M}}(\lambda_v^n) \| \mathcal{T}_{\mathcal{M}}(\mu_v^n)) \geq \Omega(1).$$

Taking the expectation with respect to π gives

$$\mathop{\mathbb{E}}_{V \sim \pi} [\mathrm{D_{KL}}(\mathcal{T}_{\mathcal{M}}(\lambda_V^n) \| \mathcal{T}_{\mathcal{M}}(\mu_V^n))] \geq \Omega(1).$$

This inequality may be combined with the KL-divergence bound of Lemma 2 to obtain a sample-complexity lower bound. The following result modifies our distributions in a way that resolves this issue. This result also introduces scaling by $W \bullet U/\alpha$ which ultimately will have the effect of removing the dependence on $W \bullet U/\alpha$ in our lower bound.

Lemma 17. *Let Q be a collection of queries with workload matrix $W \in \mathbb{R}^{[k] \times \mathcal{X}}$. Let $U \in \mathbb{R}^{[k] \times \mathcal{X}}$ be the dual witness so that (6.2) is satisfied. Then there exist probability distributions $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k$ and $\widetilde{\mu}_1, \ldots, \widetilde{\mu}_k$ over $\mathcal{X}$, and a distribution $\widetilde{\pi}$ over $[k]$ such that:*

1. *for all v in the support of $\widetilde{\pi}$, $q_v(\widetilde{\lambda}_v) - q_v(\widetilde{\mu}_v) \geq \frac{\alpha}{O(\log(1/\alpha))}$;*
2. *the matrix $\widetilde{U} \in \mathbb{R}^{[Q]\times\mathcal{X}}$ with entries $\widetilde{u}_{v,x} = \widetilde{\pi}(v)(\widetilde{\lambda}_v(x) - \widetilde{\mu}_v(x))$ satisfies $\gamma_2^*(\widetilde{U}) \leq \frac{\alpha\gamma_2^*(U)}{W \bullet U}$.*

The proof of Lemma 17 will take advantage of the following exponential binning lemma.

Lemma 18. *Suppose that $a_1, \dots, a_k \in [0,1]$ and that π is a probability distribution over $[k]$. Then for any $\beta \in (0,1]$, there exists a set $S \subseteq [k]$ such that $\pi(S) \cdot \min_{v\in S} a_v \geq \frac{\sum_{v=1}^k \pi(v)a_v - \beta}{O(\log(1/\beta))}$.*

Proof. Let $S_\ell = \{v : 2^{-\ell-1} < a_v \leq 2^{-\ell}\}$ for $\ell \in \{0, \dots, L\}$, where $L = \log_2(1/\beta) - 1$, and let $S_\infty = \{v : a_v \leq \beta\}$. Then, because $\sum_{v\in S_\infty} \pi(v)a_v \leq \beta$, we have

$$\sum_{\ell=1}^{L} \sum_{v\in S_\ell} \pi(v)a_v \geq \sum_{v=1}^{k} \pi(v)a_v - \beta.$$

Therefore, there exists ℓ such that

$$\sum_{v\in S_\ell} \pi(v)a_v \geq \frac{\sum_{v=1}^m \pi(v)a_v - \beta}{L}.$$

The lemma now follows by taking $S = S_\ell$, since $\min_{v\in S_\ell} a_v \geq \frac{1}{2}\max_{v\in S_\ell} a_v$. □

Proof of Lemma 17. Let $\lambda_1, \dots, \lambda_k$, $\mu_1, \dots, \mu_k$, and π be as given by equations (5.11) - (5.12). We may assume that the dot product of row v of W with row v of U is non negative. Otherwise, replacing row v of U with its negation would increase the expression for the dual formulation (5.10) of $\widehat{\gamma}_2$. This implies $q_v(\lambda_v) - q_v(\mu_v) \geq 0$ for all $v \in [k]$. Hence, we may apply Lemma 18 with $a_v = q_v(\lambda_v) - q_v(\mu_v)$ and $\beta = \alpha/4$ to obtain a subset $S \subseteq [k]$ for which

$$\pi(S) \cdot \min_{v\in S} [q_v(\lambda_v) - q_v(\mu_v)] \geq \frac{\mathbb{E}_{V\sim\pi}[q_v(\lambda_v) - q_v(\mu_v)] - \alpha/4}{O(\log(1/\alpha))} = \frac{W \bullet U - \alpha/4}{O(\log(1/\alpha))}.$$

Now define $\widetilde{\pi}$ as π conditional on S. In particular,

$$\widetilde{\pi}(v) = \begin{cases} \pi(v)/\pi(S), & \text{if } v \in S \\ 0, & \text{otherwise.} \end{cases} \tag{5.13}$$

Let $\tau = \frac{\alpha}{W\bullet U}$. Then, for all $v \in [k]$, define $\widetilde{\lambda}_v = \lambda_v$ and $\widetilde{\mu}_v = \tau\pi(S)\mu_v + (1-\tau\pi(S))\lambda_v$. This implies

$$\widetilde{\lambda}_v - \widetilde{\mu}_v = \tau\pi(S) \cdot \left(\widetilde{\lambda}_v - \widetilde{\mu}_v\right). \tag{5.14}$$

Hence,

$$\min_{v\in S} \left[q_v(\widetilde{\lambda}_v) - q_v(\widetilde{\mu}_v)\right] = \tau\pi(S) \cdot \min_{v\in S} [q_v(\lambda_v) - q_v(\mu_v)] \geq \tau \cdot \frac{W \bullet U - \alpha/4}{O(\log(1/\alpha))} \geq \frac{\alpha}{O(\log(1/\alpha))},$$

with the last inequality using the definition of τ and the fact that $W \bullet U > \alpha$. Furthermore, by

(5.13) and (5.14), the entries of the matrix $\widetilde{U}$ defined by $\widetilde{u}_{v,x} = \widetilde{\pi}(v)(\widetilde{\lambda}_v(x) - \widetilde{\mu}_v(x))$ satisfy

$$\widetilde{u}_{v,x} = \begin{cases} \tau u_{v,x}, & \text{if } v \in S \\ 0, & \text{otherwise.} \end{cases}$$

In other words, $\widetilde{U}$ is obtained from U by replacing some of its rows with the zero-vector. It is easy to see from the definition of γ_2^* that this implies $\gamma_2^*(\widetilde{U}) \leq \tau\gamma_2^*(U)$. □

Consider now the matrix $\widetilde{M} \in \mathbb{R}^{[k]\times\mathcal{X}}$ with entries $\widetilde{m}_{v,x} = \widetilde{\lambda}_v(x) - \widetilde{\mu}_v(x)$. Since $\widetilde{M}$ is obtained from the matrix $\widetilde{U}$ of Lemma 17 by scaling each row v of $\widetilde{U}$ by $\frac{1}{\pi(v)}$, it follows by (5.8) that

$$\|\widetilde{M}\|_{\ell_\infty \to L_1(\widetilde{\pi})} = \|\widetilde{U}\|_{\infty\to 1} \leq \gamma_2^*(\widetilde{U}) \leq \frac{\alpha\gamma_2^*(U)}{W \bullet U}.$$

This is not quite the quantity

$$\|\widetilde{M}\|^2_{\ell_\infty \to L_2(\widetilde{\pi})} = \max_{f\in\mathbb{R}^{\mathcal{X}}:\|f\|_\infty \leq 1} \mathop{\mathbb{E}}_{V\sim\pi}\left[\left(\mathbb{E}_{x\sim\widetilde{\lambda}_V}[f_x] - \mathbb{E}_{x\sim\widetilde{\mu}_V}[f_x]\right)^2\right]$$

which Lemma 2 would have us bound. For comparison, note

$$\|\widetilde{M}\|_{\ell_\infty \to L_1(\widetilde{\pi})} = \max_{f\in\mathbb{R}^{\mathcal{X}}:\|f\|_\infty \leq 1} \mathop{\mathbb{E}}_{V\sim\pi}\left[\left|\mathop{\mathbb{E}}_{x\sim\widetilde{\lambda}_V}[f_x]] - \mathop{\mathbb{E}}_{x\sim\widetilde{\mu}_V}[f_x]\right|\right].$$

Since the trivial case of Holder's inequality implies that the $L_1(\widetilde{\pi})$ norm is always bounded above by the $L_2(\widetilde{\pi})$ norm, it holds that $\|\widetilde{M}\|_{\ell_\infty\to L_1(\widetilde{\pi})} \leq \|\widetilde{M}\|_{\ell_\infty\to L_2(\widetilde{\pi})}$. However, this inequality goes in the wrong direction for our requirements. This issue is remedied by taking advantage of Lemma 16.

Lemma 19. *Let Q be a collection of queries with workload matrix $W \in \mathbb{R}^{[k]\times\mathcal{X}}$. Let $U \in \mathbb{R}^{[k]\times\mathcal{X}}$ be the dual witness so that (6.2) is satisfied. Then there exist probability distributions $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k$ and $\widetilde{\mu}_1, \ldots, \widetilde{\mu}_k$ over $\mathcal{X}$, and a distribution $\widehat{\pi}$ over $[k]$ such that:*

1. *$\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k, \widetilde{\mu}_1, \ldots, \widetilde{\mu}_k$ and $\widehat{\pi}$ satisfy criteria 1 of Lemma 17;*
2. *the matrix $\widetilde{M}$ with entries $\widetilde{m}_{v,x} = \widetilde{\lambda}_v(x) - \widetilde{\mu}_v(x)$ satisfies*

$$\|\widetilde{M}\|_{\ell_\infty \to L_2(\widehat{\pi})} \leq \frac{4\alpha \cdot \gamma_2^*(U)}{W \bullet U}.$$

Proof. Let $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k, \widetilde{\mu}_1, \ldots, \widetilde{\mu}_k$ and $\widetilde{\pi}$ be the distributions guaranteed to exist by Lemma 17, and let $\widetilde{U} \in \mathbb{R}^{[k]\times\mathcal{X}}$ be the corresponding matrix with entries $\widetilde{u}_{v,x} = \widetilde{\pi}(v)(\widetilde{\lambda}_v(x) - \widetilde{\mu}_v(x))$. The entries of the matrix $\widetilde{M}$ satisfy $\pi(v)\widetilde{m}_{v,x} = \widetilde{u}_{v,x}$, so we may apply Lemma 16 to obtain a distribution $\widehat{\pi}$ such that

$$\|\widetilde{M}\|_{\ell_\infty \to L_2(\widehat{\pi})} \leq 4\gamma_2^*(\widetilde{U}) \leq \frac{4\alpha \cdot \gamma_2^*(U)}{W \bullet U}.$$

Lemma 16 further guarantees that the support of $\widehat{\pi}$ lies within the support of $\widetilde{\pi}$, which together with the properties of the distributions $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k, \widetilde{\mu}_1, \ldots, \widetilde{\mu}_k$ and $\widetilde{\pi}$ gives the first condition of our lemma. □

5.3.4 Lower bound derivation

At last, we have all the components needed to prove our lower bound. First, we give a distributional lower bound against the problem of estimating the true mean $Q(\mu)$. See Section 2.3 for relevant definitions.

Theorem 20. *Let $\alpha, \varepsilon \in (0,1]$. Let Q be a workload of statistical queries with workload matrix $W \in \mathbb{R}^{[k]\times\mathcal{X}}$. Then, for some $\alpha' = \Omega(\alpha/\log(1/\alpha))$, we have*

$$\text{dist-sc}^{\ell_\infty}_{\varepsilon\text{-NILDP}}(Q, \alpha') = \Omega\left(\frac{\widehat{\gamma}_2(W,\alpha)^2}{\varepsilon^2\alpha^2}\right).$$

Proof. Let $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k$, $\widetilde{\mu}_1, \ldots, \widetilde{\mu}_k$ and $\widehat{\pi}$ be the distributions, and $\widetilde{M} \in \mathbb{R}^{[k]\times\mathcal{X}}$ the matrix, guaranteed to exist by Lemma 17.

Lemma 17 guarantees that for all v in the support of $\widehat{\pi}$,

$$q_v(\widetilde{\lambda}_v) - q_v(\widetilde{\mu}_v) \geq \frac{W \bullet U - \alpha/4}{O(\log(1/\alpha))}.$$

Let

$$\alpha' = \frac{1}{16}\min_v \left[q_v(\widetilde{\lambda}_v) - q_v(\widetilde{\mu}_v)\right].$$

where the minimum is taken over those $v \in [k]$ which lie in the support of $\widehat{\pi}$. Suppose now that $\mathcal{M}$ is a non-interactive LDP protocol satisfying $\text{dist-err}^{\ell_\infty}(\mathcal{M}, Q, n) \leq \alpha'$.

For v which lies in the support of $\widehat{\pi}$, let $S_v = \{z \in \mathbb{R}^k \;:\; \|z - Q(\widetilde{\lambda}_v)\|_\infty \leq 4\alpha'\}$ and let $T_v = \{z \in \mathbb{R}^k \;:\; \|z - Q(\widetilde{\mu}_v)\|_\infty \leq 4\alpha'\}$. By the definition of α', together with the fact that α' is positive, it follows that $S_v \cap T_v = \emptyset$. Furthermore, since $\text{dist-err}^{\ell_\infty}(\mathcal{M}, Q, n) \leq \alpha'$, we may apply Markov's inequality to obtain

$$\mathop{\mathbb{P}}_{\overline{X}\sim\widetilde{\lambda}_v^n}\left[\mathcal{M}(\overline{X}) \in S_v\right] \geq \frac{3}{4}$$

and

$$\mathop{\mathbb{P}}_{\overline{X}\sim\widetilde{\mu}_v^n}\left[\mathcal{M}(\overline{X}) \in T_v\right] \geq \frac{3}{4}.$$

The last inequality, together with the disjointness of S_v and T_v, gives

$$\mathop{\mathbb{P}}_{\overline{X}\sim\widetilde{\mu}_v^n}\left[\mathcal{M}(\overline{X}) \in S_v\right] \leq \mathop{\mathbb{P}}_{\overline{X}\sim\widetilde{\mu}_v^n}\left[\mathcal{M}(\overline{X}) \in T_v^c\right] \leq \frac{1}{4}.$$

By definition, the total variation between the distribution of $\mathcal{M}(\overline{X})$ when $\overline{X} \sim \widetilde{\lambda}_v^n$ and the distribution of $\mathcal{M}(\overline{X})$ when $\overline{X} \sim \widetilde{\mu}_v^n$ is at least

$$\left|\mathop{\mathbb{P}}_{\overline{X}\sim\widetilde{\lambda}_v^n}\left[\mathcal{M}(\overline{X}) \in S_v\right] - \mathop{\mathbb{P}}_{\overline{X}\sim\widetilde{\mu}_v^n}\left[\mathcal{M}(\overline{X}) \in S_v\right]\right| \geq \frac{1}{2}.$$

Bounding KL-divergence below by total variation, this gives

$$\mathrm{D_{KL}}(\mathcal{T}_{\mathcal{M}}(\widetilde{\lambda}_v^n) \| \mathcal{T}_{\mathcal{M}}(\widetilde{\mu}_v^n)) \geq \frac{1}{2}. \tag{5.15}$$

Lemma 19 further guarantees that the the matrix $\widetilde{M}$ with entries $\widetilde{m}_{v,x} = \widetilde{\lambda}_v(x) - \widetilde{\mu}_v(x)$ satisfies

$$\|\widetilde{M}\|_{\ell_\infty \to L_2(\widehat{\pi})} \leq \frac{4\alpha \cdot \gamma_2^*(U)}{W \bullet U} = \frac{4\alpha}{W \bullet U} \cdot \frac{W \bullet U - \alpha}{\widehat{\gamma}_2(W,\alpha)} \leq \frac{4\alpha}{\widehat{\gamma}_2(W,\alpha)}.$$

By Lemma 2, this implies

$$\mathop{\mathbb{E}}_{V \sim \widehat{\pi}} \left[\mathrm{D_{KL}}(\mathcal{T}_{\mathcal{M}}(\widetilde{\lambda}_V^n) \| \mathcal{T}_{\mathcal{M}}(\widetilde{\mu}_V^n)) \right] \leq O(n\varepsilon^2) \cdot \left(\frac{\alpha}{\widehat{\gamma}_2(W,\alpha)} \right)^2. \tag{5.16}$$

By (5.15), the expression under the expectation is bounded below by a constant. Since this is true of all v in the support of $\widehat{\pi}$, the bound (7.7) implies

$$n = \Omega\left(\frac{\widehat{\gamma}_2(W,\alpha)}{\varepsilon^2 \alpha^2} \right).$$

□

We can use uniform convergence to get a lower bound for the case of empirical error.

Theorem 21. *Let $\alpha, \varepsilon \in (0,1]$. Let Q be a workload of statistical queries with workload matrix $W \in \mathbb{R}^{[k] \times \mathcal{X}}$. Then, for some $\alpha' = \Omega(\alpha/\log(1/\alpha))$, if $\frac{\widehat{\gamma}_2(W,\alpha)^2}{\varepsilon^2\alpha^2} \geq \frac{C \log 2k}{(\alpha')^2}$ for a large enough constant C, we have*

$$\mathrm{sc}^{\ell_\infty}_{\varepsilon\text{-NILDP}}(Q, \alpha') = \Omega\left(\frac{\widehat{\gamma}_2(W,\alpha)^2}{\varepsilon^2\alpha^2} \right).$$

Proof. Consider running an ε-LDP protocol $\mathcal{M}$ on $n = \max\left\{ \mathrm{sc}^{\ell_\infty}(\mathcal{M}, Q, \alpha'), \frac{C' \log 2k}{(\alpha')^2} \right\}$ samples drawn i.i.d. from some distribution μ on $\mathcal{X}$. By classical uniform convergence results,

$$\mathop{\mathbb{E}}_{\overline{x} \sim \mu^n} \left[\|Q(\overline{x}) - Q(\mu)\|_\infty \right] \leq \alpha'.$$

Together with the mechanism's accuracy guarantee, this implies

$$\mathop{\mathbb{E}}_{\substack{\overline{x} \sim \mu^n \\ \mathcal{M}}} \left[\|\mathcal{M}(\overline{x}) - Q(\mu)\|_\infty \right] \leq 2\alpha'.$$

Applying Theorem 20, this implies that we can choose $\alpha' = \Omega(\alpha/\log(1/\alpha))$ so that we get the lower bound $n = \Omega\left(\frac{\widehat{\gamma}_2(W,\alpha)^2}{\varepsilon^2\alpha^2} \right)$. Then $\max\left\{ \mathrm{sc}^{\ell_\infty}(\mathcal{M}, Q, \alpha'), \frac{C' \log 2k}{(\alpha')^2} \right\} = \Omega\left(\frac{\widehat{\gamma}_2(W,\alpha)^2}{\varepsilon^2\alpha^2} \right)$, which implies the theorem by the assumption on $\frac{\widehat{\gamma}_2(W,\alpha)^2}{\varepsilon^2\alpha^2}$.

□

For the special case of anwering parities, we give a tight characterization which follows from a more direct proof in Appendix B.3.

5.3.5 $\gamma_2(W,\alpha)$ versus $\widehat{\gamma}_2(W,\alpha)$

The previous section stated and derived our upper and lower bounds for answering a workload statistical queries given by the matrix $W \in \mathbb{R}^{Q \times \mathcal{X}}$ under non-interactive LDP in terms of $\widehat{\gamma}_2(W,\alpha)$. However, $\gamma_2(W,\alpha)$ is a more standard notion, and is useful for applications since it may be computed more directly. To be able to state our results in terms of $\gamma_2(W,\alpha)$, we derive the relationship between the two quantities as given by the following result. The proof may be found in Appendix B.2.1.

Lemma 22. *Let $W \in [-1,1]^{Q\times\mathcal{X}}$ be workload matrix. Let $\alpha \in [0,1]$. Then*

$$\widehat{\gamma}_2(W,\alpha) \le \gamma_2(W,\alpha) \le 2\widehat{\gamma}_2(W,\alpha) + 1 + \alpha$$

Now we may restate Theorem 12 and Theorem 21 in terms of $\gamma_2(W,\alpha)$.

Theorem 23 (Approximate Factorization Mechanism)**.** *There exists an ε-LDP mechanism $\mathcal{M}_{\gamma_2}$ such that, for any k statistical queries Q with workload matrix W, we have*

$$\mathrm{sc}^{\ell_\infty}(\mathcal{M}_{\gamma_2}, Q, \alpha/2) = O\left(\frac{\gamma_2(W,\alpha/2)^2\log k}{\varepsilon^2\alpha^2}\right),$$

and the mechanism runs in time polynomial in n, k, and $|\mathcal{X}|$.

Theorem 24. *Let $\alpha,\varepsilon \in (0,1]$. Let Q be a workload of statistical queries with workload matrix $W \in \mathbb{R}^{[k]\times\mathcal{X}}$. Then, for some $\alpha' = \Omega(\alpha/\log(1/\alpha))$, if $\frac{\gamma_2(W,\alpha)^2}{\varepsilon^2\alpha^2} \ge \frac{C\log 2k}{(\alpha')^2}$ for a large enough constant C, we have*

$$\mathrm{sc}^{\ell_\infty}_{\varepsilon\text{-NILDP}}(Q,\alpha') = \Omega\left(\frac{\gamma_2(W,\alpha)^2}{\varepsilon^2\alpha^2}\right).$$

5.3.6 Applications of the lower bounds

In this subsection we apply Theorem 21 to several workloads of interest, and, using known bounds on the approximate γ_2 norm, prove new lower bounds on the sample complexity of these workloads.

We start with the threshold queries Q_T^{cdf} for some $T \in \mathbb{N}$, consisting of queries over the domain $[T]$ given by $q_t(x) = \mathbb{I}\{x \le t\}$. We see that the corresponding workload matrix W is a lower triangular matrix, with entries equal to 1 on and below the main diagonal. Let us consider a different matrix $W' = 2W - J$, where J is the all-ones $T\times T$ matrix. Forster et al. [FSSS03] showed a lower bound on the margin complexity of W', which implies that for any $\widehat{W}$ where $\widehat{w}_{t,x}w'_{t,x} \ge 1$ holds for all $t,x \in [T]$, we have

$$\gamma_2(\widehat{W}) = \Omega(\log T). \tag{5.17}$$

Note that if $\widetilde{W}$ satisfies $\|\widetilde{W} - W'\|_{1\to\infty} \le \frac{1}{2}$, then we can take $\widehat{W} = 2\widetilde{W}$, and (5.17) implies $\gamma_2(W',1/2) = \Omega(\log T)$. Finally, homogeneity and the triangle inequality for γ_2, and $\gamma_2(J) = 1$ imply that $\gamma_2(W,1/2) \ge \frac{1}{2}\gamma_2(W',1/2) - \frac{1}{2} = \Omega(\log T)$. Together with Theorem 24, this gives the following result, which should be compared to the upper bound of $O(\log^3 T)$ that can be obtained from the local analogue of the *binary tree mechanism* [DNPR10, CSS11]. Ours is the first lower bound to go beyond the $\Omega(\log T)$ lower bound for this problem, which follows via a packing argument.

Corollary 25 (Thresholds / CDFs)**.** *Let Q_T^{cdf} be the family of statistical queries over the domain $\mathcal{X} = [T]$ that, for every $1 \le t \le T$, contains the statistical query $q_t(x) = \mathbb{I}\{x \le t\}$. Then for every $T \in \mathbb{N}$ and ε, α smaller than an absolute constant,*

$$\mathrm{sc}^{\ell_\infty}_{\varepsilon\text{-NILDP}}(Q_T^{cdf},\alpha) = \Omega\left(\log^2 T\right).$$

Next, we consider the parity queries $Q^{\text{parity}}_{d,w}$. Note that the workload matrix W of these queries is a submatrix consisting of $\binom{d}{w}$ rows of the $2^d \times 2^d$ Hadamard matrix. Let $s = 2^d\binom{d}{w}$ be the number of entries in W. To prove a lower bound on $\gamma_2(W, \alpha)$, we can use Lemma 13 with $U = W$. The rows of a Hadamard matrix are pairwise orthogonal and have ℓ_2 norm $2^{d/2}$, and, so, Lemma 14, used with P and Q set to appropriately scaled copies of the identity matrices of the respective dimensions, implies that $\gamma_2^*(U) \leq \sqrt{s2^d}$. Moreover, $W \bullet U = \|U\|_1 = s$, and, by Lemma 13, we have

$$\gamma_2(W, 1/2) \geq \frac{\sqrt{s}}{2^{(d/2)+1}} = \Omega\left(\binom{d}{w}^{1/2}\right).$$

This gives the following result, which implies that adding independent Gaussian noise to each query is optimal up to a $O(w\log(d/w))$ factor. Appendix B.3 gives a more direct proof of a lower bound for parities which is tight up to constant factors.

Corollary 26 (Parities)**.** *Let $Q^{\text{parity}}_{d,w}$ be the family of statistical queries over the domain $\mathcal{X} = \{\pm 1\}^d$ that, for every $S \subseteq [d], |S| \leq w$, contains the statistical query $q_S(x) = \prod_{j\in S} x_j$. Then for every $k \leq d \in \mathbb{N}$ and ε, α smaller than an absolute constant,*

$$\mathrm{sc}^{\ell_\infty}_{\varepsilon\text{-NILDP}}(Q^{\text{parity}}_{d,w}, \alpha) = \Omega((d/w)^w).$$

Finally, we treat marginal queries, which have been well studied in differential privacy [BCD+07, KRSU10, GHRU11, HRS12, TUV12, CTUW14, DNT15]. We define $Q^{\text{marginal}}_{d,w}$ to consist of the queries $q_{S,y}(\overline{x}) = \frac{1}{n}\sum_{i=1}^n \prod_{j\in S} \mathbb{I}[x_{i,j} = y_j]$, with S ranging over subsets of $[d]$ of size at most w, and y ranging over $\{0,1\}^d$. To prove a lower bound for $Q^{\text{marginal}}_{d,w}$, we use the pattern matrix method of Sherstov [She11]. We will omit a full definition of a pattern matrix here, and refer the reader to Sherstov's paper. Instead, we remark that, denoting by f the `AND` function on w bits, a (d, w, f)-pattern matrix W' is a $((2d)^w/w^w) \times 2^d$ submatrix of the workload matrix W for $Q^{\text{marginal}}_{d,w}$. Let $s = 2^d\frac{(2d)^w}{w^w}$ be the number of entries in W'. By Theorem 8.1. in [She11], we have that, for any $\alpha \leq \frac{1}{6}$,

$$\min\left\{\frac{1}{\sqrt{s}}\|\widetilde{W}\|_{tr} : \|\widetilde{W} - W'\|_{1\to\infty} \leq \alpha\right\} = \Omega\left(\frac{d}{w}\right)^{\deg_{1/3}(f)/2},$$

where $\|\widetilde{W}\|_{tr}$ is the trace norm, i.e., the sum of singular values of $\widetilde{W}$, and $\deg_{1/3}(f)$ is the $(1/3)$-approximate degree of f, which is known to be $\Omega(\sqrt{w})$ [NS94]. Since $\frac{1}{\sqrt{s}}\|\widetilde{W}\|_{tr}$ is a lower bound on $\gamma_2(\widetilde{W})$ (see [LMSS07, Lemma 3.4]), this implies

$$\gamma_2(W, 1/6) \geq \frac{1}{\sqrt{s}}\|\widetilde{W}\|_{tr} = \Omega\left(\frac{d}{w}\right)^{\Omega(\sqrt{w})},$$

giving us the following lower bound. Corollary 27 shows that a natural local analogue of the algorithm of [TUV12] is optimal for answering marginal queries up to the hidden constant factor in the exponent.

Corollary 27 (Marginals)**.** *Let $Q^{\text{marginal}}_{d,w}$ be the family of statistical queries ver the domain $\mathcal{X} = \{0,1\}^d$ that, for every $S \subseteq [d], |S| \leq w$, contains the statistical query $q_S(x) = \prod_{j\in S} x_j$. Then for* *every $k \leq d \in \mathbb{N}$ and ε, α smaller than an absolute constant,*

$$\mathrm{sc}^{\ell_\infty}_{\varepsilon\text{-NILDP}}(Q^{\mathrm{marginal}}_{d,w}, \alpha) = (d/w)^{\Omega(\sqrt{w})}.$$

5.4 Open problems

It is not known whether the lower bound of Theorem 20 for statistical query release in terms of $\gamma_2(W, \alpha)$ holds for interactive LDP. We leave it as an open problem whether the lower bound can be generalized to allow for sequential interactivity. Our arguments fail to generalize to the interactive setting due to their reliance on the KL-divergence bound of Lemma 2, which is only known to hold for non-interactive LDP. This motivates the related open problem of whether there is a generalization of Lemma 2 to the interactive setting,

Chapter 6

Characterization of agnostic learning under non-interactive LDP

6.1 Overview

In this chapter, we study two related basic statistical tasks, *agnostic learning* and *agnostic refutation*, in the setting of non-interactive LDP. For both tasks, we have an unknown distribution λ on labeled data points in the universe $\mathcal{U} \times \{-1, 1\}$, and we receive samples from λ. We are also given a concept class $\mathcal{C} \subseteq \{-1, 1\}^{\mathcal{U}}$, which, hopefully, is capable of capturing the labels given by λ. We define our two tasks as follows.

- *Learning* requires finding a concept that best fits λ. In particular, using the binary loss function $L_\lambda(h) = \underset{(A,B)\sim\lambda}{\mathbb{E}} [\mathbb{I}[h(A) \neq B]]$, the goal of agnostic learning with accuracy α is to produce some $h : \mathcal{U} \to \{-1, 1\}$ which, with probability $1 - \beta$, satisfies $L_\lambda(h) \leq \min_{c\in\mathcal{C}} L_\lambda(c) + \alpha$. If an algorithm solves this problem for any distribution λ, then we say it (α, β)-learns $\mathcal{C}$ agnostically.
- *Refutation* requires distinguishing between data distributions λ that are well correlated with some concept $c \in \mathcal{C}$, vs. data distributions where the labels are random. I.e., the goal of agnostic refutation with accuracy α is to distinguish, with probability $1 - \beta$, between the following two cases: (i) $\min_{c\in\mathcal{C}} L_\lambda(c) \leq \frac{1}{2} - \alpha$; versus, (ii) for all $h : \mathcal{U} \to \{-1, 1\}$, $L_\lambda(h) = \frac{1}{2}$. If an algorithm solves this problem for any distribution λ, then we say it (α, β)-refutes $\mathcal{C}$ for threshold $\frac{1}{2}$.

The definition of agnostic learning above is classical. Refutation is a more recent notion, and was studied by [KL18] (and in the realizable setting by [Vad17]), where it was shown that computationally efficient refutation is equivalent to computationally efficient agnostic learning. Refutation captures a subproblem of evaluating the choice of model in supervised learning, that is to say, of estimating the best achievable loss $\min_{c\in\mathcal{C}} L_\lambda(c)$ by the concept class $\mathcal{C}$. While agnostic learning is well-defined for any concept class, it is less meaningful when the best achievable loss is large, which may be an indication that we should be learning a different concept class. For this reason, we would ideally like our learning algorithm to also tell us what loss it is able to achieve. Refutation is a more basic version of this problem, in which we merely want to distinguish data distributions with labels which are well approximated by our concept class from distributions with random labels, for which no model can achieve good results. Being able to solve the refutation problem is at least as hard as estimating $\min_{c\in\mathcal{C}} L_\lambda(c)$.

Our main goal is to characterize, for any given concept class $\mathcal{C}$, the sample complexity of learning and refutation under the constraints of non-interactive LDP. Moreover, we aim to understand how these two problems are related to each other.

In many settings, it is trivial to take an algorithm for learning and use it to obtain an algorithm for refutation, by executing the learning algorithm for accuracy $\alpha/4$, and estimating the loss of the returned hypothesis within $\alpha/4$. A converse of this simple reduction was established by [KL18], and by [Vad17]. Unfortunately, neither of these reductions applies to the setting of non-interactive LDP, since they rely on interacting with the distribution λ adaptively.

This leaves open the question of whether learning and refutation in the non-interactive LDP setting are equivalent with respect to sample complexity. In the forward (typically easier) direction, can non-interactive learning algorithms solve non-interactive refutation without a significant increase in sample complexity, and conversely, can the reductions from refutation to learning from [KL18] and [Vad17] be extended to the setting of non-interactive protocols?

We note that, by the equivalence proved in [KLN+11] between LDP and the statistical query (SQ) model of [Kea93], this also means that the relationship between the query complexity of non-adaptive SQ learning versus refutation is open. Similarly, all our results extend to the non-adaptive SQ model. Adaptive SQ learning has been characterized by [Fel17], and this in turn implies a characterization for sequential LDP (LDP protocols in which each participant sends one message, which can depend on the messages of previous participants).

Recall that Theorem 23 of the previous chapter guarantees the existence of non-interactive LDP protocol for estimating a workload of statistical queries. This provides one approach to solving either the agnostic learning or the agnostic refutation problem. Indeed, we may consider, for each concept $c \in \mathcal{C}$, the corresponding "correlational query" $q_c : (\mathcal{U} \times \{-1, 1\}) \to \{-1, 1\}$ given by

$$q_c(a, b) = b \cdot c(a).$$

Then the loss of the concept c on a data set $\overline{X} = ((a_1, b_1), \ldots, (a_n, b_n))$, denoted $L_{\overline{X}}(c)$, is given by

$$L_{\overline{X}}(c) = \frac{1}{2} - \frac{1}{2} q_c(\overline{X}).$$

In this way, estimating $L_{\overline{X}}(c)$ is equivalent to estimating $q_c(\overline{X})$. More generally, if we consider the query workload Q consisting of all such queries q_c obtained from some concept c of $\mathcal{C}$ as such, then estimating $Q(\overline{X})$ is equivalent to estimating $(L_{\overline{X}}(c))_{c\in\mathcal{C}}$ for each $c \in \mathcal{C}$. When $\overline{X}$ consists of sufficiently many samples drawn i.i.d. from some distribution λ, then $L_{\overline{X}}(c)$ provides a good approximation of $L_\lambda(c)$. Given this information for each $c \in \mathcal{C}$, we are able to solve either the learning or refutation problem, yielding the following result, bounding the sample complexity for these tasks in terms of the approximate γ_2 norm of the *concept matrix* $W_{\mathcal{C}}$.

Definition 28. *Let $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ be a concept class. The concept matrix $W_{\mathcal{C}} \in \{-1,1\}^{\mathcal{C}\times\mathcal{U}}$ of $\mathcal{C}$ is the matrix with entries given by $w_{c,a} = c(a)$.*

Theorem 29. *Let $\mathcal{C} \subseteq \{\pm 1\}^{\mathcal{U}}$ be a finite concept class with concept matrix $W_{\mathcal{C}} \in \{\pm 1\}^{\mathcal{C}\times\mathcal{U}}$ Let $\varepsilon, \alpha, \beta > 0$. Then, to either (α,β)-learn $\mathcal{C}$ agnostically, or (α,β)-refute $\mathcal{C}$ agnostically under non-interactive ε-LDP, it suffices to have a data set of size*

$$n = O\left(\frac{\gamma_2(W_{\mathcal{C}},\alpha/2)^2 \cdot \log(|\mathcal{C}|/\beta)}{\varepsilon^2\alpha^2}\right).$$

However, the lower bound against statistical query release in the previous chapter does not apply to either the refutation or learning problem since the hard task against which that lower bound was obtained, when applied to the workload of correlational queries corresponding to $\mathcal{C}$, would require estimating the minimum loss of a concept in the concept class on the underlying distribution. One of the primary goals of this chapter is to develop the machinery we used to give lower bounds against statistical query release towards giving the following lower bounds against agnostic refutation and learning.

Theorem 30. *Let $\alpha, \varepsilon \in (0,1]$. Let $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ be a concept class with concept matrix $W_{\mathcal{C}} \in \{-1,1\}^{\mathcal{C}\times\mathcal{U}}$ Then, for some $\alpha' = \Omega(\alpha/\log(1/\alpha))$, if $\mathcal{M}$ is a non-interactive LDP protocol which $(\alpha', \frac{1}{2} + \Omega(1))$-refutes $\mathcal{C}$ agnostically, we have*

$$n = \Omega\left(\frac{\gamma_2(W_{\mathcal{C}},\alpha/2)^2}{\varepsilon^2\alpha^2}\right).$$

Theorem 31. *Let $\varepsilon, \alpha \in (0,1]$. Let $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ be a concept class with concept matrix $W_{\mathcal{C}} \in \{-1,1\}^{\mathcal{C}\times\mathcal{U}}$. Then, for some $\alpha' = \Omega\left(\frac{\alpha}{\log(1/\alpha)}\right)$, under non-interactive ε-LDP, the number of samples required to $\left(\alpha', \frac{1}{2} + \Omega(1)\right)$-learn $\mathcal{C}$ agnostically is at least*

$$n = \Omega\left(\frac{(\gamma_2(W,\alpha) - 1)^2}{\varepsilon^2\alpha^2}\right)$$

Theorem 30 as well as the refutation lower bound of Theorem 31 were originally published in [ENU19], joint work with Aleksandar Nikolov and Jonathan Ullman. The learning lower bound of Theorem 31 was originally published in [ENP22], and is joint work with Aleksandar Nikolov and Toniann Pitassi.

6.2 Agnostic learning and refutation upper bounds

Let $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ be a concept class. For a concept $c \in \mathcal{C}$, we consider the corresponding *correlational query* $q_c : (\mathcal{U} \times \{-1,1\}) \to \{-1,1\}$ given by

$$q_c(a,b) = b \cdot c(a).$$

Then the *loss* of the concept $c \in \mathcal{C}$ on a distribution λ is

$$L_\lambda(c) = \mathop{\mathbb{E}}_{(a,b)\sim}\left[\mathbb{I}[c(a_i) \neq b_i]\right] = \mathbb{E}\left[\frac{1}{2} - \frac{b\cdot c(a)}{2}\right] = \frac{1}{2} - \frac{1}{2}q_c(\lambda).$$

In this way, estimating $L_\lambda(c)$ is equivalent to estimating $q_c(\lambda)$. Indeed, we may consider the query workload $Q_\mathcal{C}$ consisting of all such queries q_c, $c \in \mathcal{C}$. In this way, estimating $Q(\lambda)$ is equivalent to estimating the loss of each concept. Recall that Theorem 23 allows us to answer a workload of statistical queries on a data set $\overline{X} \sim \lambda^n$. When $\overline{X}$ consists of at least $\frac{C \log 2|Q|}{(\alpha')^2}$ samples for some universal constant C, classical uniform convergence results guarantee that $\mathbb{E}\left[Q(\overline{X}) - Q(\lambda)\right] \leq \alpha'$. Together with a Hoeffding bound, this implies the upper bound on learning and refutation given by Theorem 29.

6.3 Agnostic refutation lower bound

In this section we present our lower bound against agnostic refutation under non-interactive LDP. We first present the local approximate factorization mechanism. The lower-bound techniques applied will be similar to those of the previous chapter, invoking the same KL-divergence bound, Lemma 2, and using a dual formulation of the approximate γ_2 norm to construct our 'hard' distributions.

6.3.1 Duality for $\gamma_2(W, \alpha)$

The dual formulation of $\gamma_2(W, \alpha)$ we rely on is given in Lemma 32. This dual formulation was also given in [LS09] for the special case when W has entries in $\{-1, 1\}$. For completeness, here we rederive it in Appendix C.1 by directly applying the hyperplane separator theorem.

Lemma 32. *For any $W \in \mathbb{R}^{k \times T}$ and $\alpha > 0$,*

$$\gamma_2(W, \alpha) = \max\left\{\frac{W \bullet U - \alpha\|U\|_1}{\gamma_2^*(U)} \ : \ U \in \mathbb{R}^{k\times T}, \ U \neq 0\right\}, \tag{6.1}$$

where γ_2^* is the dual norm to γ_2 given by

$$\begin{aligned}\gamma_2^*(U) &= \max\{U \bullet V : V \in \mathbb{R}^{k\times T}, \ \gamma_2(V) \leq 1\} \\ &= \max_{\substack{a_1,\ldots,a_k \\ b_1,\ldots,b_T}} \sum_{i=1}^{k}\sum_{j=1}^{T} u_{i,j} a_i^\top b_j,\end{aligned}$$

where $a_1, \ldots, a_k$ and $b_1, \ldots, b_T$ range over vectors with unit ℓ_2 norm in $\mathbb{R}^{k+T}$.

6.3.2 Symmetric workloads

The query workload $Q_\mathcal{C}$ has additional structure which we will take advantage of. Consider the following definition.

Definition 33. *Let Q be a workload of statistical queries with workload matrix $W \in \mathbb{R}^{Q\times\mathcal{X}}$. Suppose there exists a partition of $\mathcal{X}$ into sets $\mathcal{X}^+$ and $\mathcal{X}^-$, $|\mathcal{X}^+| = |\mathcal{X}^-|$, where each element x of $\mathcal{X}^+$ is identified with a distinct element of $\mathcal{X}^-$, denoted $-x$, such that, for all $q \in Q$, for all $x \in \mathcal{X}$, $q(-x) = -q(x)$. In other words, W can be expressed as (W^+, W^-), where $W^+ \in \mathbb{R}^{Q\times\mathcal{X}^+}$ and $W^- \in \mathbb{R}^{Q\times\mathcal{X}^-}$ are the restrictions of W to $Q \times \mathcal{X}^+$ and $Q \times \mathcal{X}^-$ respectively, with each entry $w^+_{q,x}$ of W^+ and the corresponding entry $w^-_{q,-x}$ of W^- satisfying $w^-_{q,x} = -w^+_{q,-x}$. Also write Q^+ to denote the collection of queries with workload matrix W^+ so that the queries $q^+ : \mathcal{X}^+ \to \mathbb{R}$ of Q^+ are* *obtained by restricting queries $q : \mathcal{X} \to \mathbb{R}$ of Q to the input space $\mathcal{X}^+$; define Q^- analogously. Then Q, and also W, are called symmetric.*

By taking $\mathcal{X}^+ = \mathcal{U} \times \{1\}$ and $\mathcal{X}^- = \mathcal{U} \times \{-1\}$, we can see that $Q_\mathcal{C}$ is a symmetric workload. In particular, given an element $x = (a, 1)$ of $\mathcal{X}^+$, we let $-x = (a, -1)$. Each query q_c is symmetric since, for all $x = (a, 1) \in \mathcal{X}^+$,

$$q_c(-x) = -c(a) = -q_c(x).$$

Lemma 34 allows us to relate $\gamma_2(W)$ and $\gamma_2(W^+)$ and their witnesses. Its proof is also given in Appendix C.2.

Lemma 34. *Let $\alpha > 0$ and let $W \in \mathbb{R}^{Q\times\mathcal{X}}$ be a symmetric workload matrix with $\mathcal{X}^+$ and W^+ as given by Definition 33. Then it holds that $\gamma_2(W) = \gamma_2(W^+)$ and $\gamma_2(W, \alpha) = \gamma_2(W^+, \alpha)$. Moreover, if, for some $U^+ \in \mathbb{R}^{Q\times\mathcal{X}^+}$,*

$$\gamma_2(W^+, \alpha) = \frac{W^+ \bullet U^+ - \alpha\|U^+\|_1}{\gamma_2^*(U^+)},$$

then

$$\gamma_2(W, \alpha) = \frac{W \bullet U - \alpha\|U\|_1}{\gamma_2^*(U)},$$

where $U = \frac{1}{2}(U^+, U^-)$ is a matrix in $\mathbb{R}^{Q\times\mathcal{X}}$ such that the submatrix U^- is indexed by $\mathcal{X}^-$ and has entries $u^-_{q,-x} = -u^+_{q,x}$ for all $x \in \mathcal{X}^+$ and $q \in Q$.

6.3.3 Construction of hard distributions based on dual solution

In this section, we construct our hard distributions based on the dual witness to the approximate γ_2 norm. This section will consider a generic symmetric query workload Q with workload matrix W rather than focus on the special case of learning a concept class. Indeed, these techniques may be used to get lower bounds against answering an arbitrary symmetric workload of queries, though such a result is already subsumed by the lower bound of the previous chapter. It will be notationally convenient in this section to consider the enumeration $q_1, \ldots, q_k$ of the queries in Q and view W as a matrix in $\mathbb{R}^{[k]\times\mathcal{X}}$.

Let $U \in \mathbb{R}^{[k]\times\mathcal{X}}$ be the dual witness to $\gamma_2(W, \alpha)$, as given by Lemma 32, so that

$$\gamma_2(W, \alpha) = \frac{W \bullet U - \alpha\|U\|_1}{\gamma_2^*(U)}. \tag{6.2}$$

By Lemma 34, we may assume without loss of generality that U is of the form (U^+, U^-) where each entry of U^- is the additive inverse of the corresponding entry of U^+. Furthermore, by dividing each entry of U by $\|U\|_1$ if necessary, we may assume without loss of generality that $\|U\|_1 = 1$. In this case,

$$\gamma_2(W, \alpha) = \frac{W \bullet U - \alpha}{\gamma_2^*(U)}.$$

Let us make a first attempt at constructing our collection of "hard" distributions $\lambda_1, \ldots, \lambda_k$ and $\mu_1, \ldots, \mu_k$ for Q. Since $\|U\|_1 = 1$, then

$$\pi(v) = \sum_{x\in\mathcal{X}} |u_{v,x}|$$

defines a valid probability distribution over $[k]$. For each $v \in [k]$, we then define a pair of distributions λ_v and μ_v given by

$$\forall x \in \mathcal{X}^+ : \ \lambda_v(x) = \lambda_v(-x) = |u_{v,x}|/\pi(v)$$

$$\forall x \in \mathcal{X}^+ : \ \mu_v(x) = \begin{cases} 2|u_{v,x}|/\pi(v) & \text{if } u_{v,x} \geq 0 \\ 0 & \text{if } u_{v,x} < 0 \end{cases}$$

$$\mu_v(-x) = \begin{cases} 0 & \text{if } u_{v,x} \geq 0 \\ 2|u_{v,x}|/\pi(v) & \text{if } u_{v,x} < 0 \end{cases}$$

Then, for all $v \in [k]$, for all $x \in \mathcal{X}^+$, $\lambda_v(x) = \lambda_v(-x)$. At the same time, it holds for all $v \in [k]$ that

$$q_v(\mu_v) = \sum_{x \in \mathcal{X}} q_v(x)\mu_v(x) = \sum_{x \in \mathcal{X}^+} q_v(x)(\mu_v(x) - \mu_v(-x)).$$

Hence,

$$\mathop{\mathbb{E}}_{V \sim \pi}[q_V(\mu_V)] = 2W^+ \bullet U^+ = W \bullet U.$$

Since $W \bullet U = \gamma_2^*(U)\gamma_2(W, \alpha) + \alpha \geq \alpha$ by Lemma 32, then

$$\mathbb{E}_{V \sim \pi}[q_V(\mu_V)] \geq \alpha.$$

However, we want the inequality

$$q_v(\mu_v) \geq \alpha$$

to hold for all $v \in [k]$, as opposed to merely on average. The following result modifies our distributions in a way that resolves this issue.

Lemma 35. *Let Q be a collection of symmetric queries with workload matrix $W \in \mathbb{R}^{[k] \times \mathcal{X}}$. Let $U \in \mathbb{R}^{[k] \times \mathcal{X}}$ be the dual witness so that (6.2) is satisfied. Then there exist probability distributions $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k$ and $\widetilde{\mu}_1, \ldots, \widetilde{\mu}_k$ over $\mathcal{X}$, and a distribution $\widetilde{\pi}$ over $[k]$ such that:*

1. *for all $v \in [k]$, for all $x \in \mathcal{X}$ $\lambda_v(x) = \lambda_v(-x)$;*
2. *for all v in the support of $\widetilde{\pi}$, $q_v(\widetilde{\mu}_v) \geq \frac{\alpha}{O(\log(1/\alpha))}$;*
3. *the matrix $\widetilde{U} \in \mathbb{R}^{[Q] \times \mathcal{X}}$ with entries $\widetilde{u}_{v,x} = \widetilde{\pi}(v)(\widetilde{\lambda}_v(x) - \widetilde{\mu}_v(x))$ satisfies $\gamma_2^*(\widetilde{U}) \leq \frac{\alpha\gamma_2^*(U)}{W \bullet U}$.*

Proof. Let $\lambda_1, \ldots, \lambda_k$, $\mu_1, \ldots, \mu_k$, and π be as given by equations (6.3.3) - (6.3.3). Since $q_v(\mu_v) > 0$ for all v, we may apply Lemma 18 with $a_v = q_v(\mu_v)$ and $\beta = \alpha/4$ to obtain a subset $S \subseteq [k]$ for which

$$\pi(S) \cdot \min_{v \in S} q_v(\mu_v) \geq \frac{\mathop{\mathbb{E}}_{V \sim \pi}[q_V(\mu_V)] - \alpha/4}{O(\log(1/\alpha))} = \frac{W \bullet U - \alpha/4}{O(\log(1/\alpha))}.$$

Now define $\widetilde{\pi}$ as π conditional on S. In particular,

$$\widetilde{\pi}(v) = \begin{cases} \pi(v)/\pi(S), & \text{if } v \in S \\ 0, & \text{otherwise.} \end{cases}$$

Let $\tau = \frac{\alpha}{W \bullet U}$. Then, for all $v \in [k]$, define $\widetilde{\lambda}_v = \lambda_v$ and $\widetilde{\mu}_v = \tau\pi(S)\mu_v + (1 - \tau\pi(S))\lambda_v$. This implies

$$\forall v \in [k],\ \forall x \in \mathcal{X}^+ : \quad \lambda_v(x) = \lambda_v(-x)$$

$$\forall v \in [k] : \quad q_v(\widetilde{\mu}_v) = \tau\pi(S)q_v(\mu_v) \geq \frac{\alpha}{W \bullet U} \cdot \frac{W \bullet U - \alpha/4}{O(\log(1/\alpha))} \geq \frac{\alpha}{O(\log(1/\alpha))} \tag{6.3}$$

$$\forall v \in [k] : \quad \widetilde{\mu}_v - \widetilde{\lambda}_v = \tau \cdot \pi(S)(\mu_v - \lambda_v)$$

where (6.3) uses the fact that $W \bullet U \geq \alpha$.

By the last of these facts, together with the definition of $\widetilde{\pi}$, it follows that the entries $\widetilde{u}_{v,x} = \widetilde{\pi}(v)(\widetilde{\lambda}_v(x) - \widetilde{\mu}_v(x))$ of the matrix $\widetilde{U}$ satisfy

$$\widetilde{u}_{v,x} = \begin{cases} \tau u_{v,x}, & \text{if } v \in S \\ 0, & \text{otherwise.} \end{cases}$$

In other words, $\widetilde{U}$ is obtained from U by replacing some of its rows with the zero-vector. It is easy to see from the definition of γ_2^* that this implies $\gamma_2^*(\widetilde{U}) \leq \tau\gamma_2^*(U)$. □

Consider now the matrix $\widetilde{M} \in \mathbb{R}^{[k]\times\mathcal{X}}$ with entries $\widetilde{m}_{v,x} = \widetilde{\lambda}_v(x) - \widetilde{\mu}_v(x)$. Since $\widetilde{M}$ is obtained from the matrix $\widetilde{U}$ of Lemma 35 by scaling each row v of $\widetilde{U}$ by $\frac{1}{2\pi(v)}$, it follows that

$$\|\widetilde{M}\|_{\ell_\infty \to L_1(\widetilde{\pi})} = \frac{1}{2}\|\widetilde{U}\|_{\infty\to 1} \leq \gamma_2^*(\widetilde{U}) \leq \tau\gamma_2^*(U) = \frac{\alpha}{W \bullet U} \cdot \frac{W \bullet U - \alpha/4}{O(\log(1/\alpha))} \geq \frac{\alpha}{O(\log(1/\alpha))}.$$

This is not quite the quantity

$$\|\widetilde{M}\|^2_{\ell_\infty \to L_2(\widetilde{\pi})} = \max_{f \in \mathbb{R}^{\mathcal{X}} : \|f\|_\infty \leq 1} \mathop{\mathbb{E}}_{V \sim \pi}\left[\left(\mathbb{E}_{x\sim\widetilde{\lambda}_V}[f_x] - \mathbb{E}_{x\sim\widetilde{\mu}_V}[f_x]\right)^2\right]$$

which Lemma 2 would have us bound. For comparison, note

$$\|\widetilde{M}\|_{\ell_\infty \to L_1(\widetilde{\pi})} = \max_{f \in \mathbb{R}^{\mathcal{X}} : \|f\|_\infty \leq 1} \mathop{\mathbb{E}}_{V \sim \pi}\left[\left|\mathop{\mathbb{E}}_{x\sim\widetilde{\lambda}_V}[f_x]] - \mathop{\mathbb{E}}_{x\sim\widetilde{\mu}_V}[f_x]\right|\right].$$

Since the trivial case of Holder's inequality implies that the $L_1(\widetilde{\pi})$ norm is always bounded above by the $L_2(\widetilde{\pi})$ norm, it holds that $\|\widetilde{M}\|_{\ell_\infty \to L_1(\widetilde{\pi})} \leq \|\widetilde{M}\|_{\ell_\infty \to L_2(\widetilde{\pi})}$. However, this inequality goes in the wrong direction for our requirements. This issue is remedied by taking advantage of Lemma 16.

Lemma 36. *Let Q be a collection of symmetric queries with workload matrix $W \in \mathbb{R}^{[k]\times\mathcal{X}}$. Let $U \in \mathbb{R}^{[k]\times\mathcal{X}}$ be the dual witness so that (6.2) is satisfied. Then there exist probability distributions $\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_k$ and $\widetilde{\mu}_1, \dots, \widetilde{\mu}_k$ over $\mathcal{X}$, and a distribution $\widetilde{\pi}$ over $[k]$ such that:*

1. *$\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_k, \widetilde{\mu}_1, \dots, \widetilde{\mu}_k$ and $\widetilde{\pi}$ satisfy criteria 1. and 2. of Lemma 35;*
2. *the matrix $\widetilde{M}$ with entries $\widetilde{m}_{v,x} = \widetilde{\lambda}_v(x) - \widetilde{\mu}_v(x)$ satisfies*

$$\|\widetilde{M}\|_{\ell_\infty \to L_2(\widetilde{\pi})} \leq \frac{4\alpha}{\gamma_2(W, \alpha)}.$$

Proof. Let $\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_k, \widetilde{\mu}_1, \dots, \widetilde{\mu}_k$ and $\widetilde{\pi}$ be the distributions guaranteed to exist by Lemma 35, and let $\widetilde{U} \in \mathbb{R}^{[k]\times\mathcal{X}}$ be the corresponding matrix with entries $\widetilde{u}_{v,x} = \widetilde{\pi}(v)(\widetilde{\lambda}_v(x) - \widetilde{\mu}_v(x))$. The entries of the matrix $\widetilde{M}$ satisfy $\pi(v)\widetilde{m}_{v,x} = \widetilde{u}_{v,x}$, so we may apply Lemma 16 to obtain a distribution $\widehat{\pi}$ such that

$$\|\widetilde{M}\|_{\ell_\infty \to L_2(\widehat{\pi})} \leq 4\gamma_2^*(\widetilde{U}) \leq \frac{4\alpha\gamma_2^*(U)}{W \bullet U} = \frac{4\alpha(W \bullet U - \alpha)}{(W \bullet U)\gamma_2(W,\alpha)} \leq \frac{4\alpha}{\gamma_2(W,\alpha)}.$$

Lemma 16 further guarantees that the support of $\widehat{\pi}$ lies within the support of $\widetilde{\pi}$, which together with the properties of the distributions $\widetilde{\lambda}_1,\ldots,\widetilde{\lambda}_k,\widetilde{\mu}_1,\ldots,\widetilde{\mu}_k$ and $\widetilde{\pi}$ gives the first condition of our lemma. □

6.3.4 Lower bound derivation

At last, we have all the components needed to prove our lower bound against agnostic refutation, Theorem 30.

Proof of Theorem 30. As described in Section 6.3.2, take $\mathcal{X}^+ = \mathcal{U} \times \{1\}$, $\mathcal{X}^- = \mathcal{U} \times \{-1\}$, and $\mathcal{X} = \mathcal{X}^- \cup \mathcal{X}^+$, associating each element $x = (a,1) \in \mathcal{X}^+$ with $-x = (a,-1) \in \mathcal{X}^-$. Then, $Q_{\mathcal{C}} = \{q_c\}_{c\in\mathcal{C}}$, consisting of queries $q_c(a,b) = b \cdot c(a)$ is a symmetric workload. Let $\overline{W}_{\mathcal{C}} \in \{-1,1\}^{\mathcal{C}\times\mathcal{X}}$ denote the corresponding workload matrix.

Apply Lemma 36 to the queries $Q_{\mathcal{C}}$ to obtain collections $\{\widetilde{\lambda}_c\}_{c\in\mathcal{C}}$, and $\{\widetilde{\mu}_c\}_{c\in\mathcal{C}}$ of distributions on $\mathcal{X}$ as well as a distribution $\widehat{\pi}$ on $\mathcal{C}$. The matrix $\widetilde{M} \in \{-1,1\}^{\mathcal{C}\times\mathcal{X}}$ with entries $\widetilde{m}_{c,x} = \widetilde{\lambda}_c(x) - \widetilde{\mu}_c(x)$ satisfies

$$\|\widetilde{M}\|_{\ell_\infty \to L_2(\widehat{\pi})} \leq \frac{4\alpha}{\gamma_2(\overline{W}_{\mathcal{C}},\alpha)}.$$

Equivalently,

$$\max_{f\in\mathbb{R}^{\mathcal{X}}:\|f\|_\infty\leq 1} \mathop{\mathbb{E}}_{c\sim\widehat{\pi}}\left[\left(\mathop{\mathbb{E}}_{X\sim\widetilde{\lambda}_c}[f_X] - \mathop{\mathbb{E}}_{X\sim\widetilde{\mu}_c}[f_X]\right)^2\right] \leq \left(\frac{4\alpha}{\gamma_2(\overline{W}_{\mathcal{C}},\alpha)}\right)^2.$$

By Lemma 2, this implies

$$\mathrm{D}_{\mathrm{KL}}(\mathcal{T}_{\mathcal{M}}(\widetilde{\lambda}_{\widehat{\pi}}^n)\|\mathcal{T}_{\mathcal{M}}(\widetilde{\mu}_{\widehat{\pi}}^n)) \leq O(n\varepsilon^2)\cdot\left(\frac{4\alpha}{\gamma_2(\overline{W}_{\mathcal{C}},\alpha)}\right)^2. \tag{6.4}$$

Lemma 36 guarantees further that there exists some

$$\alpha' \geq \frac{\alpha}{O(\log(1/\alpha))}$$

such that, for all $c \in \mathcal{C}$, $q_c(\widetilde{\pi}) \geq \alpha'$. This implies $L_{\mu_c}(c) \leq \frac{1}{2} - \frac{1}{2}\alpha'$. At the same time, $\lambda_c(a,+1) = \mu_c(a,-1)$ for all $a \in \mathcal{U}$. Hence, if a non-interactive LDP protocol $\mathcal{M}$ $(\alpha', \frac{1}{2}+\Omega(1))$-refutes $\mathcal{C}$ agnostically, then it distinguishes between λ_c and μ_c with probability $\frac{1}{2}+\Omega(1)$. Hence,

$$d_{\mathrm{TV}}(\mathcal{T}_{\mathcal{M}}(\lambda_v^n)\|\mathcal{T}_{\mathcal{M}}(\mu_v^n)) = \Omega(1).$$

By Pinsker's inequality,

$$\mathrm{D}_{\mathrm{KL}}\left(\mathcal{T}_{\mathcal{M}}(\lambda_v^n) \parallel \mathcal{T}_{\mathcal{M}}(\mu_v^n)\right) = \Omega(1).$$

Combining this lower bound on KL-divergence with the upper bound given by (7.7), we obtain

$$n = \Omega\left(\frac{\gamma_2(\overline{W}_{\mathcal{C}}, \alpha)^2}{\varepsilon^2 \alpha^2}\right).$$

Finally, by application of Lemma 34, we have $\gamma_2(W, \alpha) = \gamma_2(\overline{W}_{\mathcal{C}}, \alpha)$.

□

6.4 Agnostic learning lower bound

In this section, we present our results for agnostic learning under non-interactive LDP. The lower bound will use similar machinery to that used to obtain our lower bounds against statistical query release and against agnostic refutation, relying on the KL-divergence bound of Lemma 2 and using the dual to the measure of the concept class's complexity to construct hard distributions for agnostic learning of the given concept class. One novel aspect of the argument will be the consideration of the concept class's "difference matrix," defined in the next section.

6.4.1 Difference matrix

Theorem 31 is given in terms of the concept matrix associated with the concept class; however, our proof of this result will focus instead on the difference matrix associated with the concept class, defined below.

Definition 37. *The difference matrix of a concept class $\mathcal{C} : \mathcal{U} \to \{-1, 1\}$ is the matrix $D_{\mathcal{C}} \in \{-1, 1\}^{\mathcal{C}^2 \times \mathcal{U}}$ with entries given, for $c, c' \in \mathcal{C}, a \in \mathcal{U}$, by*

$$d_{(c,c'),a} = \frac{1}{2}(c(a) - c'(a)) = \begin{cases} 0 & \text{if } c(a) = c'(a) \\ -1 & \text{if } c(a) = -1, c'(a) = +1 \\ +1 & \text{if } c(a) = +1, c'(a) = -1. \end{cases} \tag{6.5}$$

The difference matrix is one of the key ideas that enables the proof of Theorem 31. We will use a dual formulation of $\gamma_2(D_{\mathcal{C}}, \alpha)$ to construct pairs of hard distributions for our lower bound, each pair corresponding to a pair of concepts $c, c' \in \mathcal{C}$. The structure of the difference matrix will help us ensure that no correct agnostic learning algorithm can output, with high probability, the same hypothesis for both distributions in a pair. It is not apparent how to guarantee this property when working directly with the concept matrix $W_{\mathcal{C}}$.

To the motivate the definition of the difference matrix, consider the symmetrization $\overline{D} \in \{-1, 1\}^{\mathcal{C} \times (\mathcal{U} \times \{-1,1\})}$ of $D_{\mathcal{C}}$ with entries given by $\overline{d}_{c,(a,b)} = b \cdot \overline{d}_{a,b}$. By answering the workload of queries $\{q_{c,c'}\}_{c,c' \in \mathcal{C}}$ represented by $\overline{D}$ we obtain, for each $c, c' \in \mathcal{C}$, an estimate of

$$q_{c,c'}(\lambda) = \mathop{\mathbb{E}}_{(A,B)\sim\lambda}\left[B \cdot d_{(c,c'),A}\right] = \mathop{\mathbb{E}}_{(A,B)\sim\lambda}\left[\frac{b}{2} \cdot (c(A) - c'(A))\right] = L_\lambda(c) - L_\lambda(c').$$

In other words, answering these queries gives the difference in loss for each pair of concepts. This provides sufficient information to agnostically learn the concept class. Estimating these queries with the approximate factorization mechanism for statistical query release of [ENU19] gives an upper bound for agnostic learning under in terms of $\gamma_2(D_{\mathcal{C}},\alpha)$. Indeed, this provides an alternate approach to obtaining a non-interactive LDP protocol for agnostic learning, with the sample complexity of the protocol dependent on $\gamma_2(D_{\mathcal{C}},\alpha)$. However, the following lemma shows that $\gamma_2(D_{\mathcal{C}},\alpha)$ and $\gamma_2(W_{\mathcal{C}},\alpha)$ are essentially the same. (See Appendix C.3 for the proof.)

Lemma 38. *Let $\mathcal{C}$ be a concept class with concept matrix $W_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}\times\mathcal{U}}$ and difference matrix $D_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}^2\times\mathcal{U}}$. Then $\gamma_2(D_{\mathcal{C}},\alpha) \le \gamma_2(W,\alpha)$. Conversely, $\gamma_2(W,\alpha) \le 2\gamma_2(D_{\mathcal{C}},\alpha/2)+1$, and if $\mathcal{C}$ is closed under negation then $\gamma_2(W,\alpha) \le \gamma_2(D_{\mathcal{C}},\alpha)$.*

Despite this equivalence, the dual witness to $\gamma_2(D_{\mathcal{C}},\alpha)$ will be especially useful in the construction of our lower bound. In particular, we obtain our lower bound first in terms of $\gamma_2(D_{\mathcal{C}},\alpha)$, giving Lemma 39, and we will then apply Lemma 38 to get a lower bound in terms of $\gamma_2(W,\alpha)$.

Lemma 39. *Let $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ be a concept class with concept matrix $D_{\mathcal{C}} \in \{-1,1\}^{\mathcal{C}\times\mathcal{U}}$. Then Theorem 31 holds with $W_{\mathcal{C}}$ replaced by the difference matrix, $D_{\mathcal{C}}$.*

The rest of this section is devoted to the proof of Lemma 39.

6.4.2 Duality and hard distributions

For the construction of hard families of distributions, it will be convenient to make use of the dual formulation given by Lemma 32, applying it to the difference matrix instead of the concept matrix. In particular, for an arbitrary concept class $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ with difference matrix $D_{\mathcal{C}}$, let $U \in \mathbb{R}^{\mathcal{C}^2\times\mathcal{U}}$ be the dual witness to $\gamma_2(D_{\mathcal{C}},\alpha)$ so that

$$\gamma_2(D_{\mathcal{C}},\alpha) = \frac{D_{\mathcal{C}} \bullet U - \alpha\|U\|_1}{\gamma_2^*(U)}. \tag{6.6}$$

By normalizing U, we may assume, without loss of generality, that $\|U\|_1 = 1$. Moreover, we can assume that, for any $c, c' \in \mathcal{C}$, $\sum_{a\in\mathcal{U}} d_{(c,c'),a}u_{(c,c'),a} \ge 0$. Otherwise, U cannot achieve the maximum of (6.1), since we can multiply the row of U indexed by (c,c') by -1, which increases $D_{\mathcal{C}} \bullet U$ and does not change $\|U\|_1$ or $\gamma_2^*(U)$.

We will consider the matrices $U^+, U^- \in \mathbb{R}^{\mathcal{C}^2\times\mathcal{U}}$ with non-negative entries which satisfy $U = U^+ - U^-$, so that U^+ and U^- correspond to the positive and negative entries of U respectively. We define the distribution π on $\mathcal{C}^2$ by

$$\pi(c,c') = \sum_{a\in\mathcal{U}} u_{(c,c'),a}. \tag{6.7}$$

Then, for $c, c' \in \mathcal{C}$, consider the distribution $\lambda_{c,c'}$ on $\mathcal{U}\times\{-1,1\}$ given by

$$\lambda_{c,c'}(a,1) = \frac{u^+_{(c,c'),a}}{\pi(c,c')}, \qquad \lambda_{c,c'}(a,-1) = \frac{u^-_{(c,c'),a}}{\pi(c,c')} \tag{6.8}$$

Similarly, let $\mu_{c,c'}$ be the distribution on $\mathcal{U}\times\{-1,1\}$ given by

$$\mu_{c,c'}(a,1) = \frac{u^-_{(c,c'),a}}{\pi(c,c')}, \qquad \mu_{c,c'}(a,-1) = \frac{u^+_{(c,c'),a}}{\pi(c,c')}. \tag{6.9}$$

Since U has unit ℓ_1 norm, the above distributions are well-defined. Note that $\lambda_{c,c'}$ and $\mu_{c,c'}$ have the same marginal on $\mathcal{U}$ which we denote $\kappa_{(c,c')}$. In particular, $\kappa_{(c,c')}(a) = \frac{|u_{(c,c'),a}|}{\pi(c,c')}$. Meanwhile, $\lambda_{c,c'}$

always gives a the label $b = \mathrm{sign}(u_{(c,c'),a})$, while $\mu_{c,c'}$ always gives a the label $b = -\mathrm{sign}(u_{(c,c'),a})$. It will be useful to have notation for one of these labelling functions, so define $s_{c,c'} : \mathcal{U} \to \{-1, 1\}$ by $s_{c,c'}(a) = \mathrm{sign}(u_{(c,c'),a})$.

Consider the following relationship between U and the distributions we have constructed.

$$u_{(c,c'),a} = \pi(c,c') \left(\lambda_{c,c'}(a,1) - \mu_{c,c'}(a,1)\right) = \pi(c,c')\kappa_{c,c'}(a)s_{c,c'}(a)$$

Note that

$$\begin{aligned}\sum_{a\in\mathcal{U}} d_{(c,c'),a}u_{(c,c'),a} &= \pi(c,c') \cdot \sum_{a\in\mathcal{U}} \frac{1}{2} \cdot \kappa_{c,c'}(a) \cdot [c(a)s_{c,c'}(a) - c'(a)s_{c,c'}(a)] \\ &= \pi(c,c') \cdot \left(L_{\lambda_{c,c'}}(c) - L_{\lambda_{c,c'}}(c')\right). \qquad (6.10)\end{aligned}$$

Similarly,

$$\sum_{a\in\mathcal{U}} d_{(c,c'),a}u_{(c,c'),a} = \pi(c,c') \cdot \left(L_{\mu_{c,c'}}(c') - L_{\mu_{c,c'}}(c)\right). \qquad (6.11)$$

Hence,

$$D_{\mathcal{C}} \bullet U = \mathop{\mathbb{E}}_{(c,c')\sim\pi}\left[\left(L_{\lambda_{c,c'}}(c') - L_{\lambda_{c,c'}}(c)\right)\right] = \mathop{\mathbb{E}}_{(c,c')\sim\pi}\left[\left(L_{\mu_{c,c'}}(c) - L_{\mu_{c,c'}}(c')\right)\right].$$

Whenever $\mathcal{C}$ contains at least two distinct concepts, $\gamma_2(D_{\mathcal{C}}, \alpha) > 0$, and then (6.6) implies $D_{\mathcal{C}} \bullet U > \alpha$. By the equations above, this implies that, on average with respect to $(c,c') \sim \pi$, the loss of c is greater by α than the loss of c' on $\lambda_{c,c'}$. Likewise, on average, the loss of c' is greater by α than the loss of c on $\mu_{c,c'}$. We will see that, if we can obtain these properties in the worst case over all (c,c'), rather than only on average, then no hypothesis can fit both $\lambda_{c,c'}$ and $\mu_{c,c'}$ for any $c, c' \in \mathcal{C}$. By applying exponential binning, we give such worst-case bounds.

Lemma 40. *Let $\mathcal{C}$ be a concept class. Let $D_{\mathcal{C}}$ be the matrix given by (6.5). Let $U \in \mathbb{R}^{\mathcal{C}^2\times\mathcal{U}}$, $\|U\|_1 = 1$, satisfy (6.6). Then there exist probability distributions $\widetilde{\lambda}_{c,c'}$ and $\widetilde{\mu}_{c,c'}$ over $\mathcal{U} \times \{-1,1\}$, and a distribution $\widetilde{\pi}$ over $\mathcal{C}^2$ such that:*

1. *For all (c,c') in the support of $\widetilde{\pi}$, $L_{\widetilde{\lambda}_{c,c'}}(c) - L_{\widetilde{\lambda}_{c,c'}}(c') \geq \frac{\alpha}{O(\log(1/\alpha))}$.*

2. *For all (c,c') in the support of $\widetilde{\pi}$, $L_{\widetilde{\mu}_{c,c'}}(c') - L_{\widetilde{\mu}_{c,c'}}(c) \geq \frac{\alpha}{O(\log(1/\alpha))}$.*

3. *The matrix $\widetilde{U} \in \mathbb{R}^{\mathcal{C}^2\times\mathcal{U}}$ with entries $\widetilde{u}_{(c,c'),a} = \widetilde{\pi}(c,c')(\widetilde{\lambda}_{c,c'}(a) - \widetilde{\mu}_{c,c'}(a))$ satisfies $\gamma_2^*(\widetilde{U}) \leq \frac{\alpha\gamma_2^*(U)}{D_{\mathcal{C}}\bullet U}$.*

Proof. Let π, together with $\lambda_{c,c'}$ and $\mu_{c,c'}$, be defined as in (6.7), (6.8) and (6.9). We will apply Lemma 18 to the values given, for $c, c' \in \mathcal{C}$, by

$$a_{c,c'} = L_{\lambda_{c,c'}}(c) - L_{\lambda_{c,c'}}(c) = L_{\mu_{c,c'}}(c) - L_{\mu_{c,c'}}(c).$$

Recall that we may assume, that for all $c, c' \in \mathcal{C}$ the following inequality holds

$$\sum_{a\in\mathcal{U}} d_{(c,c'),a}u_{(c,c'),a} \geq 0.$$

Together with (6.10), this gives $a_{c,c'} \geq 0$ for all $c, c' \in \mathcal{C}$.

By Lemma 18, there exists some $S \subset \mathcal{C}^2$ such that

$$\pi(S) \cdot \min_{(c,c')\in S} \left(L_{\lambda_{c,c'}}(c) - L_{\lambda_{c,c'}}(c')\right) \geq \frac{\underset{(c,c')\sim\pi}{\mathbb{E}}\left[L_{\lambda_{c,c'}}(c) - L_{\lambda_{c,c'}}(c')\right] - \alpha/4}{O(\log(1/\alpha))} = \frac{D_{\mathcal{C}} \bullet U - \alpha/4}{O(\log(1/\alpha))}.$$

By applying (6.11), we get the similar

$$\pi(S) \cdot \min_{(c,c')\in S} \left(L_{\mu_{c,c'}}(c') - L_{\mu_{c,c'}}(c)\right) \geq \frac{D_{\mathcal{C}} \bullet U - \alpha/4}{O(\log(1/\alpha))}.$$

Let $\widetilde{\pi}$ be defined by

$$\widetilde{\pi}(c,c') = \begin{cases} \pi(c,c')/\pi(S), & \text{if } c,c' \in S \\ 0, & \text{otherwise.} \end{cases}$$

Let also $\tau = \frac{\alpha}{D_{\mathcal{C}} \bullet U} \in (0,1)$. For $(c,c') \in S$, let $\widetilde{\lambda}_{c,c'} = \lambda_{c,c'}$ and $\widetilde{\mu}_{c,c'} = (1 - \tau\pi(S))\lambda_{c,c'} + \tau\pi(S)\mu_{c,c'}$. It holds then, for $(c,c') \in S$, that

$$\widetilde{\lambda}_{c,c'} - \widetilde{\mu}_{c,c'} = \tau \cdot \pi(S) \cdot (\lambda_{c,c'} - \mu_{c,c'})$$

Hence, the matrix $\widetilde{U} \in \mathbb{R}^{\mathcal{C}^2 \times \mathcal{U}}$ with entries defined by

$$\widetilde{u}_{v,a} = \widetilde{\pi}(v) \cdot (\widetilde{\lambda}_v(a,1) - \widetilde{\mu}_v(a,1)) = -\widetilde{\pi}(v) \cdot (\widetilde{\lambda}_v(a,-1) - \widetilde{\mu}_v(a,-1))$$

satisfies

$$\widetilde{u}_{(c,c'),a} = \begin{cases} \tau u_{(c,c'),a}, & \text{if } (c,c') \in S \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to see from the definition of γ_2^* that this implies $\gamma_2^*(\widetilde{U}) \leq \tau\gamma_2^*(U) = \frac{\alpha\gamma_2^*(U)}{D_{\mathcal{C}} \bullet U}$. □

We will also want to bound the operator norm, which appears in Lemma 2, in terms of $\gamma_2^*(U)$. To do so, we use Lemma 16 which relates γ_2^* and the $\infty \to 2$ operator norm.

At the same time, we want to obtain a lower bound on $d_{\mathrm{TV}}(\mathcal{T}_{\mathcal{M}}(\lambda_{c,c'}^n) \| \mathcal{T}_{\mathcal{M}}(\mu_{c,c'}^n))$ when $\mathcal{M}$ is a learning algorithm for $\mathcal{C}$. For this purpose, we apply the following lemma, whose proof is deferred to Appendix C.4. The main observation in the proof is, for any hypothesis $h : \mathcal{C} \to \{-1,1\}$, and any distributions λ and μ satisfying the conditions of the lemma, we have $L_\lambda(h) + L_\mu(h) = 1$.

Lemma 41. *Let λ and μ be distributions on $\mathcal{U} \times \{-1,1\}$. Assume that λ and μ have the same marginal on $\mathcal{U}$. Also assume that λ is labelled by some $s : \mathcal{U} \to \{-1,1\}$ while μ is labelled by $-s$. Finally, assume that for some $c, c' \in \mathcal{C}$, $L_\mu(c') - L_\mu(c) > \alpha$. If $h : \mathcal{U} \to \{-1,1\}$ satisfies $L_\lambda(h) \leq L_\lambda(c') + \alpha/4$, then $L_\mu(h) > L_\mu(c) + 3\alpha/4$. Hence, if $\mathcal{M}$ is an algorithm which $(\alpha/4, \beta)$-learns $\mathcal{C}$ from n samples, then $d_{\mathrm{TV}}(\mathcal{M}(\lambda^n), \mathcal{M}(\mu^n)) \geq 1 - 2\beta$.*

6.4.3 Lower bound derivation

Finally, with these results at our disposal, we may obtain the lower bound of Lemma 39.

Proof of Lemma 39. Let $U \in \mathbb{R}^{\mathcal{C}^2 \times \mathcal{U}}$, $\|U\|_1 = 1$, satisfy (6.6). Let $\widetilde{\pi}$, together with $\widetilde{\lambda}_{c,c'}$ and $\widetilde{\mu}_{c,c'}$ be the distributions guaranteed to exist by Lemma 35 and let $\widetilde{U} \in \mathbb{R}^{\mathcal{C}^2 \times \mathcal{U}}$ be the corresponding matrix with entries

$$\widetilde{u}_{(c,c'),a} = \widetilde{\pi}(c,c')\left(\widetilde{\lambda}_{c,c'}(a,1) - \widetilde{\mu}_{c,c'}(a,1)\right) = -\widetilde{\pi}(c,c')\left(\widetilde{\lambda}_{c,c'}(a,-1) - \widetilde{\mu}_{c,c'}(a,-1)\right)$$

Let M be the matrix with entries $m_{(c,c'),a} = \widetilde{u}_{(c,c'),a}/\widetilde{\pi}(c,c')$. By Lemma 16, there exists some distribution $\hat{\pi}$ with support contained in that of $\widetilde{\pi}$ such that

$$\|M\|_{\ell_\infty \to L_2(\hat{\pi})} \leq 4\gamma_2^*(\widetilde{U}) \leq \frac{4\alpha\gamma_2^*(U)}{D_{\mathcal{C}} \bullet U},$$

where the last inequality follows from Lemma 35. Combining Lemma 2 with the dual formulation (6.6) gives

$$\begin{aligned}\mathop{\mathbb{E}}_{(C,C')\sim\hat{\pi}}\left[\mathrm{D_{KL}}(\mathcal{T}_{\mathcal{M}}(\widetilde{\lambda}^n_{C,C'})\|\mathcal{T}_{\mathcal{M}}(\widetilde{\mu}^n_{c,c'}))\right] &\leq O(n\varepsilon^2)\cdot\|M\|^2_{\ell_\infty\to L_2(\hat{\pi})} \\ &\leq O(n\varepsilon^2)\cdot\left(\frac{\alpha\gamma_2^*(U)}{D_{\mathcal{C}} \bullet U}\right)^2 = O(n\varepsilon^2)\cdot\left(\frac{\alpha}{\gamma_2(D_{\mathcal{C}},\alpha)}\right)^2.\end{aligned}$$

Now let

$$\alpha' = \frac{1}{4}\left(\min_{c,c'} L_{\mu_{c,c'}}(c') - L_{\mu_{c,c'}}(c)\right) \geq \frac{\alpha}{O(\log(1/\alpha))},$$

where the last inequality is by Lemma 35. By Lemma 41, if $\mathcal{M}$ $(\alpha', 1+\Omega(1))$-learns $\mathcal{C}$, then $\mathop{\mathbb{E}}_{(C,C')\sim\widetilde{\pi}}\left[\mathrm{D_{KL}}(\mathcal{T}_{\mathcal{M}}(\widetilde{\lambda}^n_{c,c'})\|\mathcal{T}_{\mathcal{M}}(\widetilde{\mu}^n_{c,c'}))\right] = \Omega(1)$. This implies $n = \Omega\left(\frac{\gamma_2(D_{\mathcal{C}},\alpha)^2}{\varepsilon^2\alpha^2}\right)$, as was to be proved. □

6.5 Open problems

This work, together with [ENU19], largely completes the picture of agnostic refutability and learnability under non-interactive LDP.

As with the lower bound for statistical query release in Chapter 5, the lower bounds in this chapter for agnostic learning and refutation, do not hold against interactive LDP. We leave it as an open problem whether these lower bounds can be generalized to allow for sequential interactivity.

Further, the relationships obtained between the sample complexities of refutability and learnability in this work are indirect, via characterizations of these tasks in terms of the approximate γ_2 norm. For example, although realizable refutability implies realizable learnability under non-interactive LDP, it remains open to show how to construct a non-interactive LDP protocol for learnability directly from one for refutability.

Chapter 7

Characterization of realizable refutation under non-interactive LDP

7.1 Overview

Realizable learning is a special case of agnostic learning, where the underlying distribution λ on $\mathcal{U} \times \{-1, 1\}$ is guaranteed to be labelled by a concept $c \in \mathcal{C}$, i.e., $L_\lambda(c) = 0$. Correspondingly, we say that an algorithm $\mathcal{M} : (\mathcal{U} \times \{-1, 1\})^n \to \{-1, 1\}$ (α, β)-*refutes* $\mathcal{C}$ *realizably* if, with probability $1 - \beta$, it distinguishes the case where there exists some concept which achieves zero loss from the case where all concepts have loss at least α.

The lower bounds of the previous chapter do not apply to realizable refutation which requires additional structure of the hard distributions. Meanwhile, though the upper bounds of the previous chapter can be applied to the realizable problems, it is possible that they can be improved by taking advantage of the additional structure. Indeed, some concept classes may require exponentially fewer samples to learn in the realizable case than to learn agnostically. For example, the class of conjunctions over $\mathcal{U} = \{0, 1\}^d$ can be learned with polynomial in d query complexity using a non-adaptive SQ algorithm [Kea93], and, therefore, also with polynomial sample complexity by a non-interactive LDP algorithm. The γ_2 norm of the matrix associated with this class is, however, exponential in d, as shown in Chapter 5. Therefore, conjunctions require exponential sample complexity to learn agnostically under non-interactive LDP.

While we are not able to characterize realizable learning, we give a characterization of a realizable refutation, and show that realizable learning is no harder than realizable refutation.

[DF19] showed that (for $\mathcal{C}$ closed under negation) the sample complexity of realizable learning under non-interactive LDP is bounded from below by the margin complexity of $\mathcal{C}$. They left open the question whether one can prove a matching upper bound. This question was resolved in the negative by [DF20]. The problem of characterizing the sample complexity of realizable learning under non-interactive LDP thus remains open.

In this work, we give a non-interactive LDP protocol which may be applied towards both realizable learning and realizable refutation. This gives a sample complexity upper bound for these problems in terms of a new efficiently computable quantity $\eta(\mathcal{C}, \alpha)$ which we define. This quantity combines elements of both the γ_2 norm and of margin complexity, and is sandwiched between them. Further, we derive a lower bound for realizable refutation in terms of $\eta(\mathcal{C}, \alpha)$, showing that our protocol is nearly optimal for realizable refutation, and that the sample complexity of realizable refutation is an upper bound on the sample complexity of realizable learning under non-interactive LDP. Our main theorem for realizable learning is stated next. See Section 7.2 for the definition of $\eta(\mathcal{C}, \alpha)$.

Theorem 42. *Let $\alpha, \varepsilon, \beta > 0$. Let $\mathcal{C} \subseteq \{\pm 1\}^{\mathcal{U}}$ be a finite concept class. Let $\varepsilon > 0$, $\alpha, \beta \in (0, 1/2]$. Then, to either (α, β)-learn $\mathcal{C}$ realizably, or (α, β)-refute $\mathcal{C}$ realizably under non-interactive ε-LDP, it suffices to have a sample of size*

$$n = O\left(\frac{\eta(\mathcal{C}, \alpha/2)^2 \cdot \log(|\mathcal{C}|/\beta)}{\varepsilon^2 \alpha^2}\right).$$

Theorem 43. *Let $\alpha, \varepsilon > 0$. For some $\alpha' = \Omega\left(\frac{\alpha}{\log(1/\alpha)}\right)$, the number of samples required to $(\alpha', \Omega(1))$-refute $\mathcal{C}$ realizably under non-interactive ε-LDP is at least*

$$n = \Omega\left(\frac{\eta(\mathcal{C}, \alpha/2)^2}{\varepsilon^2 \alpha^2}\right).$$

The research represented by this chapter was originally originally published in [ENP22], and is joint work with Aleksandar Nikolov and Toniann Pitassi.

7.2 Upper bound

In this section, we present our algorithm for realizable learning and refutation for non-interactive LDP. For a concept class $\mathcal{C} : \mathcal{U} \to \{-1, 1\}$, we define a quantity $\eta(\mathcal{C}, \alpha)$ and argue that it gives an upper bound on the sample complexity for realizable learning of $\mathcal{C}$.

Definition 44. *Let $\mathcal{C} : \mathcal{U} \to \{-1, 1\}$ be a concept class. Let*

$$K_{\mathcal{C}} = \Big\{W \in \mathbb{R}^{\mathcal{C} \times (\mathcal{U} \times \{-1,1\})} \ : \ |w_{c,(a,c(a))}| \leq \alpha \text{ and } w_{c,(a,-c(a))} \geq 1 \ \forall c \in \mathcal{C}, a \in \mathcal{U}\Big\}. \tag{7.1}$$

Let

$$K'_{\mathcal{C}} = \Big\{\widetilde{W} \in \mathbb{R}^{\mathcal{C} \times (\mathcal{U} \times \{-1,1\})} \ : \ \exists W \in K_{\mathcal{C}}, \ \exists \theta \in \mathbb{R}^{\mathcal{C}}, \ \widetilde{W} = W + \theta \mathbf{1}^T\Big\}, \tag{7.2}$$

where $\mathbf{1}^T$ is the all-ones row vector indexed over $\mathcal{C}$, so that $\widetilde{W} = W + \theta \mathbf{1}^T$ is the matrix obtained by shifting each row c of W in each entry by θ_c.

Then define

$$\eta(\mathcal{C}, \alpha) = \min\Big\{\gamma_2(\widetilde{W}) : \widetilde{W} \in K'_{\mathcal{C}}\Big\}.$$

The idea is that each row of W defines a statistical query corresponding to a concept, $q_c(a, b) = w_{c,(a,b)}$. The statistical query corresponding to the true concept that was used to label the data will have a small value, whereas any query corresponding to a concept with large loss will have a large value. The next theorem formalizes this argument. Intuitively, this works because W assigns a penalty for each labeled sample, and the penalty is at most α for correctly labeled samples, and at least 1 for incorrectly labeled ones. Moreover, if $\widetilde{W}$ is obtained from $W \in K_{\mathcal{C}}$ by translating every row c of W by some $\theta_c \in \mathbb{R}$ in each dimension, then answering the queries given by $\widetilde{W}$ allows us to answer the queries given by W by just shifting the query answers. Applying these ideas gives us Theorem 42.

Proof of Theorem 42. As per Definition 44, let $\widetilde{W} \in K'_{\mathcal{C}}$ be the matrix that witnesses $\eta(\mathcal{C}, \alpha)$ and let $W \in K_{\mathcal{C}}$ and $\theta \in \mathbb{R}^{\mathcal{C}}$ be the matrix and vector which witness $\widetilde{W} \in K'_{\mathcal{C}}$. If we can answer the statistical queries given by $\widetilde{W}$, then we can answer the queries given by W with the same accuracy by subtracting θ_c from the answer to the query for concept c.

By the definition of W, if, for some $c \in \mathcal{C}$, λ is supported on on those $(a, b) \in \mathcal{U} \times \{-1, 1\}$ which satisfy $c(a) = b$, then the value of the query corresponding to c is bounded as

$$\mathop{\mathbb{E}}_{(A,B)\sim\lambda}\left[w_{c,(A,B)}\right] = \mathop{\mathbb{E}}_{(A,B)\sim\lambda}\left[w_{c,(A,c(A))}\right] \le \alpha.$$

Meanwhile, for an arbitrary distribution λ on $\mathcal{U} \times \{-1, 1\}$, the value of the query corresponding to $c \in \mathcal{C}$ may be bounded as

$$\mathop{\mathbb{E}}_{(A,B)\sim\lambda}\left[w_{c,(A,B)}\right] \ge \mathop{\mathbb{P}}_{(A,B)\sim\lambda}\left[b \ne c(A)\right] - \alpha \cdot \mathop{\mathbb{P}}_{(A,B)\sim\lambda}\left[B = c(A)\right] \ge L_\lambda(c) - \alpha.$$

In particular, if $L_\lambda(c) \ge 3\alpha$, then $\mathop{\mathbb{E}}_{(A,B)\sim\lambda}\left[w_{c,(A,B)}\right] \ge 2\alpha$.

It follows that, by approximating the statistical queries given by W with worst-case error $\frac{\alpha}{4}$, we can distinguish the case where λ agrees with some $c \in \mathcal{C}$ from the case where, for all concepts $c \in \mathcal{C}$, $L_\lambda(c) \ge 3\alpha$. In the former case, returning some $c' \in \mathcal{C}$ where our estimate of $\mathop{\mathbb{E}}_{(A,B)\sim\lambda}\left[w_{c,(A,B)}\right]$ is strictly less than 2α guarantees $L_\lambda(c) < 3\alpha$.

To complete the proof, it suffices to apply the upper bound from [ENU19] which says that, to answer the collection of statistical queries given by $\widetilde{W}$ under non-interactive ε-LDP, with accuracy $\alpha/4$ and probability of failure at most β, the number of samples required is at most

$$O\left(\frac{\gamma_2(\widetilde{W})\log(|\mathcal{C}|/\beta)}{\varepsilon^2\alpha^2}\right) = O\left(\frac{\eta(\mathcal{C}, \alpha)\log(|\mathcal{C}|/\beta)}{\varepsilon^2\alpha^2}\right). \qquad \square$$

7.3 Lower bound

Our lower bound will follow a similar strategy as in the agnostic case. However, our construction of hard distributions will be tailored to $\eta(\mathcal{C}, \alpha)$ and its dual.

7.3.1 Duality

We will again use convex duality in our lower bound. We will express $\eta(\mathcal{C}, \alpha)$ as a maximum over dual matrices U, and we will use an optimal U to construct 'hard distributions' for realizable refutation. To this end, consider the following duality lemma.

Lemma 45. *For any concept class $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ and any α,*

$$\eta(\mathcal{C},\alpha) = \max_{U \in S_{\mathcal{C}}} \frac{\sum_{c\in\mathcal{C},a\in\mathcal{U}}(u_{c,(a,-c(a))} - \alpha|u_{c,(a,c(a))}|)}{\gamma_2^*(U)}, \tag{7.3}$$

where we define

$$S_{\mathcal{C}} := \Bigg\{ U \in \mathbb{R}^{\mathcal{C}\times(\mathcal{U}\times\{-1,1\})} \ : \ \forall c \in \mathcal{C},\ \sum_{a\in\mathcal{U}}(u_{c,(a,c(a))} + u_{c,(a,-c(a))}) = 0$$
$$\text{and, } \forall c \in \mathcal{C}, \forall a \in \mathcal{U},\ u_{c,(a,-c(a))} \geq 0 \Bigg\}.$$

Proof. Let $L_{\mathcal{C}} = \{G \in \mathbb{R}^{\mathcal{C}\times(\mathcal{U}\times\{-1,1\})} : \gamma_2(G) \leq t\}$. Let $K_{\mathcal{C}}$ and $K'_{\mathcal{C}}$ be as defined by equations (7.1) and (7.2). By definition, $\eta(\mathcal{C},\alpha) > t$ if and only if $L_{\mathcal{C}}$ and $K'_{\mathcal{C}}$ are disjoint.

Given some $U \in \mathbb{R}^{\mathcal{C}\times(\mathcal{U}\times\{-1,1\})}$, we are interested in the quantities $\max\{U \cdot G : G \in L_{\mathcal{C}}\}$ and $\min\{U \cdot G : G \in K'_{\mathcal{C}}\}$. In particular, by the hyperplane separation theorem, since $L_{\mathcal{C}}$ and $K'_{\mathcal{C}}$ are convex and $L_{\mathcal{C}}$ is also compact, they are disjoint exactly when there exists some $U \in \mathbb{R}^{\mathcal{C}\times(\mathcal{U}\times\{-1,1\})}$ such that

$$\max\{U \cdot G : G \in L_{\mathcal{C}}\} < \min\{U \cdot G : G \in K'_{\mathcal{C}}\}.$$

By definition,

$$\max\{U \cdot G : G \in L_{\mathcal{C}}\} = t\gamma_2^*(U).$$

Also,

$$\begin{aligned}
&\min\{U \cdot G : G \in K'_{\mathcal{C}}\} \\
&= \min_{G\in K'_{\mathcal{C}}} \sum_{c\in\mathcal{C},a\in\mathcal{U}} (u_{c,(a,c(a))}g_{c,(a,c(a))} + u_{c,(a,-c(a))}g_{c,(a,-c(a))}) \\
&= \min_{\substack{G\in K_{\mathcal{C}} \\ \theta\in\mathbb{R}^{\mathcal{C}}}} \sum_{c\in\mathcal{C},a\in\mathcal{U}} (u_{c,(a,c(a))} \cdot (g_{c,(a,c(a))} + \theta_c) + u_{c,(a,-c(a))} \cdot (g_{c,(a,-c(a))} + \theta_c)) \\
&= \min_{G\in K_{\mathcal{C}}} \sum_{c\in\mathcal{C},a\in\mathcal{U}} (u_{c,(a,c(a))} \cdot g_{c,(a,c(a))} + u_{c,(a,-c(a))} \cdot g_{c,(a,-c(a))}) \\
&\quad + \min_{\theta\in\mathbb{R}^{\mathcal{C}}} \sum_{c\in\mathcal{C}} \theta_c \cdot \sum_{a\in\mathcal{U}} (u_{c,(a,c(a))} + u_{c,(a,-c(a))})
\end{aligned}$$

If, for some $c \in \mathcal{C}$, it holds that $\sum_{a\in\mathcal{U}}(u_{c,(a,c(a))} + u_{c,(a,-c(a))}) \neq 0$, then

$$\min_{\theta_c\in\mathbb{R}} \theta_c \cdot \sum_{a\in\mathcal{U}} (u_{c,(a,c(a))} + u_{c,(a,-c(a))}) = -\infty.$$

Also, if there exist $c \in \mathcal{C}$ and $x \in \mathcal{U}$ such that $u_{c,(a,-c(a))} < 0$, then

$$\min_{G\in K_{\mathcal{C}}} u_{c,(a,-c(a))}g_{c,(a,-c(a))} = -\infty.$$

However, in the remaining case where U is in the set $S_{\mathcal{C}}$, then

$$\begin{aligned}\min\{U\cdot G : G\in K'_{\mathcal{C}}\} &= \min_{G\in K_{\mathcal{C}}} \sum_{c\in\mathcal{C},a\in\mathcal{U}} (u_{c,(a,c(a))}\cdot g_{c,(a,c(a))} + u_{c,(a,-c(a))}\cdot g_{c,(a,-c(a))}) \\ &= \sum_{c\in\mathcal{C},a\in\mathcal{U}} (-\alpha|u_{c,(a,c(a))}| + u_{c,(a,-c(a))}).\end{aligned}$$

With these facts at our disposal, we obtain

$$\begin{aligned}\eta(\mathcal{C},\alpha) > t &\Leftrightarrow K'_{\mathcal{C}}\cap L_{\mathcal{C}} = \emptyset \\ &\Leftrightarrow \exists U\in\mathbb{R}^{\mathcal{C}\times(\mathcal{U}\times\{\pm 1\})},\ \max\{U\cdot G : G\in L_{\mathcal{C}}\} < \min\{U\cdot G : G\in K_{\mathcal{C}}\} \\ &\Leftrightarrow \exists U\in S_{\mathcal{C}},\ t\gamma_2^*(U) < \sum_{c\in\mathcal{C},a\in\mathcal{U}} (-\alpha|u_{c,(a,c(a))}| + u_{c,(a,-c(a))}) \\ &\Leftrightarrow \max_{U\in S_{\mathcal{C}}} \frac{\sum_{c\in\mathcal{C},a\in\mathcal{U}}(u_{c,(a,-c(a))} - \alpha|u_{c,(a,c(a))}|)}{\gamma_2^*(U)} > t\end{aligned}$$

Since the equivalence holds for all $t\in\mathbb{R}$, it follows that

$$\eta(\mathcal{C},\alpha) = \max_{U\in S_{\mathcal{C}}} \frac{\sum_{c\in\mathcal{C},a\in\mathcal{U}}(u_{c,(a,-c(a))} - \alpha|u_{c,(a,c(a))}|)}{\gamma_2^*(U)}.$$

□

7.3.2 Hard distributions

Let $U\in\mathbb{R}^{\mathcal{C}\times(\mathcal{U}\times\{-1,1\})}$ witness (7.3). By normalizing, we may assume without loss of generality that $\|U\|_1 = 1$. We will consider the matrices $U^+, U^- \in\mathbb{R}^{m\times N}$ with non-negative entries which satisfy $U = U^+ - U^-$ so that U^+ and U^- correspond to the positive and negative entries of U respectively. We define the distribution π on $\mathcal{C}$ given by $\pi(c) = \sum_{(a,b)\in\mathcal{U}\times\{-1,1\}} u_{c,(a,b)}$. Then, for each $c\in\mathcal{C}$, let λ_c and μ_c be the distributions on $\mathcal{U}\times\{-1,1\}$ given by

$$\lambda_c(a,b) = \frac{2u^+_{c,(a,b)}}{\pi(c)} \qquad \text{and} \qquad \mu_c(a,b) = \frac{2u^-_{c,a,b}}{\pi(c)}.$$

Since the rows of U each sum to zero and have unit ℓ_1 norm, the distributions λ_c and μ_c are well-defined. Moreover, since $u_{c,(a,-c(a))}\geq 0$ for all $c\in\mathcal{C}, a\in\mathcal{U}$, the only negative entries of U are those of the form $u_{c,(x,c(x))}$. This implies that the distribution μ_c always labels samples $a\in\mathcal{U}$ by $c(a)$.

7.3.3 Warm-up: single-concept case

Consider the case where $\mathcal{C}$ consists of a single concept c. Since $\eta(\mathcal{C},\alpha) > 0$, then (7.3) implies

$$\sum_{a\in\mathcal{U}} u_{c,(a,-c(a))} > \sum_{a\in\mathcal{U}} \alpha|u_{c,(a,c(a))}|. \tag{7.4}$$

Hence,

$$\mathop{\mathbb{P}}_{(A,B)\sim\lambda_c}[c(A)\neq B] - \mathop{\mathbb{P}}_{(A,B)\sim\mu_c}[c(A)\neq B] > \alpha\cdot\left(\mathop{\mathbb{P}}_{(A,B)\sim\lambda_c}[c(A)=B] + \mathop{\mathbb{P}}_{(A,B)\sim\mu_c}[c(A)=B]\right).$$

Using

$$\mathop{\mathbb{P}}_{(A,B)\sim\mu_c}[c(A)=B]=1 \tag{7.5}$$

and rearranging, this gives

$$L_{\lambda_c}(c)=\mathop{\mathbb{P}}_{(A,B)\sim\lambda_c}[c(A)\neq B]>\frac{2\alpha}{1+\alpha}.$$

In other words, if we can distinguish a distribution on $\mathcal{U}\times\{-1,1\}$ which labels samples according to c from one which disagrees with c with probability greater than $\frac{2\alpha}{1+\alpha}$, then we can distinguish between λ_c and μ_c.

7.3.4 General case

It remains to generalize the lower bound of the previous section to the general case where the concept class is not restricted to a single concept.

The first issue which needs to be addressed in the general case is that (7.4), rather than holding in worst case over all concepts, holds on average. In particular,

$$\sum_{c\in\mathcal{C},a\in\mathcal{U}} u_{c,(a,-c(a))}>\sum_{c\in\mathcal{C},a\in\mathcal{U}}\alpha|u_{c,(a,c(a))}|.$$

Equivalently,

$$\mathop{\mathbb{E}}_{C\sim\pi}[L_\lambda(C)]>\frac{2\alpha}{1+\alpha}. \tag{7.6}$$

This issue is handled by applying the binning result of Lemma 18.

The second issue which needs to be addressed is that, while each $c\in\mathcal{C}$ is guaranteed not to fit the corresponding distribution λ_c, as with (7.5), it may hold that some other $h:\mathcal{U}\to\{-1,1\}$ has small loss on λ_c. This is remedied by mixing a distribution σ_c which agrees with c into the distribution λ_c. This guarantees that every $h:\mathcal{U}\to\{-1,1\}$ has large loss on σ_c.

The first issue is resolved in Lemma 46 by applying the binning result of Lemma 18. The second issue will be resolved in Lemma 47.

Lemma 46. *Suppose there exist families $\{\lambda_c\}_{c\in\mathcal{C}}$ and $\{\mu_c\}_{c\in\mathcal{C}}$ of distributions over $\mathcal{U}$, together with a parameter distribution π over $\mathcal{C}$, such that*

$$\Delta=\mathop{\mathbb{E}}_{C\sim\pi}[L_{\lambda_C}(c)]>\frac{2\alpha}{1+\alpha}$$

while, for all $c\in\mathcal{C}$, $L_{\mu_c}(c)=0$. Further, let $U\in\mathbb{R}^{\mathcal{C}\times\mathcal{U}}$ be the matrix with entries $u_{c,a}=\pi(c)(\lambda_c(a)-\mu_c(a))$.

Then there exist families $\{\widetilde{\lambda}_c\}_{c\in\mathcal{C}}$ and $\{\widetilde{\mu}_c\}_{c\in\mathcal{C}}$ of distributions over $\mathcal{U}\times\{-1,1\}$, together with a parameter distribution $\widetilde{\pi}$ over $\mathcal{C}$, such that, for all c in the support of $\widetilde{\pi}$,

$$L_{\widetilde{\lambda}_c}(c)\geq\Omega\left(\frac{\alpha}{1+\alpha}\Big/\log\left(\frac{1+\alpha}{\alpha}\right)\right),$$

while still $L_{\widetilde{\mu}_c}(c)=0$ for all $c\in\mathcal{C}$. Moreover, the matrix $\widetilde{U}\in\mathbb{R}^{\mathcal{C}\times\mathcal{U}}$ with entries $\widetilde{u}_{c,a}=\widetilde{\pi}(c)(\widetilde{\lambda}_c(a)-$

$\widetilde{\mu}_c(a)$) satisfies

$$\gamma_2^*(\widetilde{U}) \le \frac{2\alpha\gamma_2^*(U)}{(1+\alpha)\Delta}.$$

Proof. Apply Lemma 18, with $a_c = L_{\lambda_c}(c)$ for all $c \in \mathcal{C}$, and $\beta = \frac{\alpha}{1+\alpha} < \frac{\Delta}{2}$, to obtain $S \subseteq \mathcal{C}$ such that

$$\pi(S) \cdot \min_{c \in S} a_c \ge \frac{\Delta - \beta}{O(\log(1/\beta)} \ge \frac{\Delta}{O(\log((1+\alpha)/\alpha)}.$$

Let $\widetilde{\pi}$ be π conditional on membership in S. Thus,

$$\widetilde{\pi}(v) = \begin{cases} \pi(v)/\pi(S), & \text{if } v \in S \\ 0, & \text{otherwise.} \end{cases}$$

Let $\tau = \frac{2\alpha}{(1+\alpha)\Delta} \in (0,1)$. For all $c \in \mathcal{C}$, define $\widetilde{\mu}_c = \mu_c$ and $\widetilde{\lambda}_c = \tau\pi(S)\lambda_c + (1 - \tau\pi(S))\mu_c$. Then, for all c in the support of $\widetilde{\pi}$,

$$L_{\widetilde{\mu}_c}(c) = L_{\mu_c}(c) = 0$$

$$L_{\widetilde{\lambda}_c}(c) = \tau \cdot \pi(S) \cdot L_{\lambda_c}(c) \ge \frac{\frac{\alpha}{1+\alpha}}{O((\log\left(\frac{1+\alpha}{\alpha}\right))}.$$

Moreover, the matrix $\widetilde{U} \in \mathbb{R}^{\mathcal{C}\times\mathcal{U}}$ with entries $\widetilde{u}_{c,a} = \widetilde{\pi}(v)(\widetilde{\lambda}_v(a) - \widetilde{\mu}_v(a))$ is obtained from the matrix τU by replacing some of its rows with the zero-vector. It follows immediately that $\gamma_2^*(\widetilde{U}) \le \tau\gamma_2^*(U) = \frac{2\alpha\gamma_2^*(U)}{(1+\alpha)\Delta}$. □

Lemma 47. *Suppose we have distributions λ_c and μ_c on $\mathcal{U} \times \{-1, 1\}$ for each $c \in \mathcal{C}$ where:*

(a) $L_{\mu_c}(c) = 0$*;*

(b) $L_{\lambda_c}(c) > \alpha$*.*

Then there exist distributions $\widetilde{\lambda}_c$ and $\widetilde{\mu}_c$ for each $c \in \mathcal{C}$ such that:

(c) $L_{\widetilde{\mu}_c}(c) = 0$*;*

(d) $\forall h : \mathcal{U} \to \{-1, 1\},\ \ L_{\widetilde{\lambda}_c}(h) > \frac{\alpha}{2}$*;*

(e) $\widetilde{\lambda}_c - \widetilde{\mu}_c = \frac{1}{2}(\lambda_c - \mu_c)$*.*

Proof. For $c \in \mathcal{C}$, let σ_c be the distribution on $\mathcal{U} \times \{-1, 1\}$ which has the same marginal on $\mathcal{U}$ as does λ_c, and which satisfies $c(a) = b$ for all (a, b) in the support of σ_c. Also, let $\widetilde{\lambda}_c = \frac{1}{2}\lambda_c + \frac{1}{2}\sigma_c$ and $\widetilde{\mu}_c = \frac{1}{2}\mu_c + \frac{1}{2}\sigma_c$. Properties (c) and (e) follow immediately.

To establish property (d), notice first that, for any $a \in \mathcal{U}$ in the support of λ_c,

$$\mathop{\mathbb{P}}_{(A,B)\sim\widetilde{\lambda}_c}[B = c(A) \mid A = a] \ge \frac{1}{2},$$

and also

$$\mathop{\mathbb{P}}_{(A,B)\sim\widetilde{\lambda}_c}[B \ne c(A) \mid A = a] = \frac{1}{2} \cdot \mathop{\mathbb{P}}_{(A,B)\sim\lambda_c}[B \ne c(A) \mid A = a].$$

For any function $h : \mathcal{U} \to \{-1, 1\}$, then

$$\begin{aligned}
L_{\widetilde{\lambda}_c}(h) &= \underset{(A,B)\sim\widetilde{\lambda}_c}{\mathbb{P}}[h(A) \neq B] \\
&= \sum_{(a,b)\in\mathcal{U}\times\{-1,1\}} \underset{(A,B)\sim\widetilde{\lambda}_c}{\mathbb{P}}[A = a] \cdot \underset{(A,B)\sim\widetilde{\lambda}_c}{\mathbb{P}}[h(A) \neq B \mid A = a] \\
&\geq \sum_{(a,b)\in\mathcal{U}\times\{-1,1\}} \underset{(A,B)\sim\widetilde{\lambda}_c}{\mathbb{P}}[A = a] \cdot \min\left\{ \underset{(A,B)\sim\widetilde{\lambda}_c}{\mathbb{P}}[c(A) = B \mid A = a], \underset{(A,B)\sim\widetilde{\lambda}_c}{\mathbb{P}}[c(A) \neq B \mid A = a] \right\} \\
&\geq \sum_{(a,b)\in\mathcal{U}\times\{-1,1\}} \underset{(A,B)\sim\widetilde{\lambda}_c}{\mathbb{P}}[A = a] \cdot \min\left\{ \frac{1}{2}, \frac{1}{2} \cdot \underset{(A,B)\sim\lambda_c}{\mathbb{P}}[c(A) \neq B \mid A = a] \right\} \\
&= \frac{1}{2} \cdot \sum_{(a,b)\in\mathcal{U}\times\{-1,1\}} \underset{(A,B)\sim\widetilde{\lambda}_c}{\mathbb{P}}[A = a] \cdot \underset{(A,B)\sim\lambda_c}{\mathbb{P}}[c(A) \neq B \mid A = a] \\
&= \frac{1}{2} \cdot L_{\lambda_c}(c) \\
&> \frac{\alpha}{2}.
\end{aligned}$$

□

Equipped with Lemmas 46 and 47, we are ready to prove our lower bound against realizable refutation, Theorem 43.

Proof of Theorem 43. We define the parameter distribution π over $\mathcal{C}$, and the distribution families $\{\lambda_c\}_{c\in\mathcal{C}}$ and $\{\mu_c\}_{c\in\mathcal{C}}$ over $\mathcal{U}\times\{-1,1\}$, as in Section 7.3.2. We denote $\Delta = \mathbb{E}_{c\sim\pi}\left[L_{\lambda_c}(c)\right]$. By equation (7.6), together with Lemmas 46 and 47, we obtain modified families of distributions $\{\widetilde{\lambda}_c\}_{c\in\mathcal{C}}$ and $\{\widetilde{\mu}_c\}_{c\in\mathcal{C}}$, together with a parameter distribution $\widetilde{\pi}$ over $\mathcal{C}$, such that, for all c in the support of $\widetilde{\pi}$, and for all functions $h : \mathcal{U} \to \{\pm 1\}$,

$$L_{\widetilde{\lambda}_c}(h) = \Omega\left(\frac{\alpha}{1+\alpha} \Big/ \log\left(\frac{1+\alpha}{\alpha}\right) \right)$$

while $L_{\widetilde{\mu}_c}(c) = 0$ for all $c \in \mathcal{C}$. By Lemmas 16, 46 and 47, we may assume further that the matrix $\widetilde{M} \in \mathbb{R}^{\mathcal{C}\times(\mathcal{U}\times\{\pm1\})}$ with entries $m_{c,(a,b)} = \widetilde{\lambda}_c(a,b) - \widetilde{\mu}_c(a,b)$ satisfies

$$\|\widetilde{M}\|_{\ell_\infty\to L_2(\widetilde{\pi})} \leq \frac{4\alpha\gamma_2^*(U)}{(1+\alpha)\Delta}.$$

Now let $\mathcal{M}$ be an ε-LDP protocol which is able to distinguish a labeling by some $c \in \mathcal{C}$ from a distribution with which every function $h : \mathcal{U} \to \{\pm 1\}$ disagrees with the labels with probability $\Omega\left(\frac{\alpha}{1+\alpha}\Big/\log\left(\frac{1+\alpha}{\alpha}\right)\right)$. If this is true, then, for every $c \in \mathcal{C}$ in the support of $\widetilde{\pi}$,

$$\mathrm{D}_{\mathrm{KL}}(\mathcal{T}_{\mathcal{M}}(\widetilde{\lambda}_c^n) \| \mathcal{T}_{\mathcal{M}}(\widetilde{\mu}_c^n)) = \Omega(1). \tag{7.7}$$

Meanwhile, Lemma 2 guarantees

$$\underset{C\sim\widetilde{\pi}}{\mathbb{E}}\left[\mathrm{D}_{\mathrm{KL}}(\mathcal{T}_{\mathcal{M}}(\widetilde{\lambda}_C^n) \| \mathcal{T}_{\mathcal{M}}(\widetilde{\mu}_C^n))\right] \leq O(n\varepsilon^2) \cdot \|\widetilde{M}\|^2_{\ell_\infty\to L_2(\pi)},$$

whereby we obtain

$$n = \Omega\left(\frac{1}{\varepsilon^2 \cdot \|\widetilde{M}\|^2_{\ell_\infty \to L_2(\pi)}}\right) = \Omega\left(\frac{(1+\alpha)^2\Delta^2}{\varepsilon^2\alpha^2\gamma_2^*(U)^2}\right). \tag{7.8}$$

Now we may use

$$\gamma_2^*(U) = \frac{\sum_{c\in\mathcal{C},a\in\mathcal{U}}(u_{c,(a,-c(a))} - \alpha|u_{c,(a,c(a))}|)}{\eta(\mathcal{C},\alpha)}. \tag{7.9}$$

Note that, for any $c \in \mathcal{C}$, since $L_{\mu_c}(c) = 0$,

$$\begin{aligned}
&\frac{1}{\pi(c)}\sum_{a\in\mathcal{U}} u_{c,(a,-c(a))} - \alpha|u_{c,(a,c(a))}| \\
&= \underset{(A,B)\sim\lambda_c}{\mathbb{P}}[c(A) \neq B] - \underset{(A,B)\sim\mu_c}{\mathbb{P}}[c(A) \neq B] \\
&\qquad - \alpha \cdot \left(\underset{(A,B)\sim\lambda_c}{\mathbb{P}}[c(A) = B] + \underset{(A,B)\sim\mu_c}{\mathbb{P}}[c(A) = B]\right) \\
&= (1+\alpha)\cdot \underset{(A,B)\sim\lambda_c}{\mathbb{P}}[c(A) \neq B] - 2\alpha \\
&= (1+\alpha)\cdot L_{\lambda_c}(c) - 2\alpha.
\end{aligned}$$

Taking expectations over $c \sim \pi$, we have

$$\begin{aligned}
\sum_{c\in\mathcal{C},a\in\mathcal{U}} u_{c,(a,-c(a))} - \alpha|u_{c,(a,c(a))}| &= (1+\alpha)\cdot \underset{c\sim\pi}{\mathbb{E}}[L_{\lambda_c}(c)] - 2\alpha \\
&= (1+\alpha)\cdot\Delta - 2\alpha. \qquad (7.10)
\end{aligned}$$

Putting equations (7.8), (7.9), and (7.10) together, we have

$$n = \Omega\left(\frac{(1+\alpha)^2\Delta^2\eta(\mathcal{C},\alpha)^2}{\varepsilon^2\alpha^2((1+\alpha)\Delta - 2\alpha)^2}\right) = \Omega\left(\frac{\eta(\mathcal{C},\alpha)^2}{\varepsilon^2\alpha^2}\right). \qquad \square$$

As a corollary of Theorem 42 and Theorem 43, it follows that realizable refutability implies realizable learnability. In particular, by Theorem 43, a sample complexity upper bound for realizable refutability of concept class $\mathcal{C}$ gives an upper bound on $\eta(\mathcal{C},\alpha)$. Then the sample complexity upper bound of Theorem 42 in terms of $\eta(\mathcal{C},\alpha)$ gives an upper bound on the number of samples required for realizable learning of $\mathcal{C}$.

Corollary 48. *Let $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ be a concept class. Let $\varepsilon > 0$, $\alpha \in (0,1]$. Then, for some*

$$\alpha' = \Omega\left(\frac{\alpha}{1+\alpha}\Big/\log\left(\frac{1+\alpha}{\alpha}\right)\right),$$

if there exists a mechanism $\mathcal{M}' : (\mathcal{U}\times\{-1,1\})^{n'} \to \{-1,1\}$ which $(\alpha', 1-\Omega(1)$-refutes $\mathcal{C}$ realizably with n' samples, then there exists a mechanism $\mathcal{M} : (\mathcal{U}\times\{-1,1\})^n \to \{-1,1\}^{\mathcal{U}}$ which (α,β)-learns $\mathcal{C}$ realizably with sample size $n = O(n' \cdot \log(|\mathcal{C}|/\beta))$.

7.4 Open problems

For the non-interactive setting, we have characterized realizable refutability, and shown that realizable refutability implies realizable learnability. It is an interesting open problem to determine the converse - whether realizable learning implies refutation. Secondly, for an arbitrary concept class $\mathcal{C}$, can we obtain a characterization of realizable learnability in terms of a quantity which is efficiently computable from the definition of $\mathcal{C}$?

Chapter 8

Equivalence between sequential LDP and single-intrusion pan-privacy

8.1 Overview

Pan-private mechanisms are streaming algorithms which protect against a stronger adversarial model than does central DP, allowing for the possibility that the internal state of the algorithm may be revealed to the adversary and guaranteeing privacy nevertheless. At the same time, pan-private mechanisms do not impose the strong requirement of LDP where privacy is guaranteed even against the central party responsible for aggregrating individuals' data. Pan-private algorithms may be categorized according to the number of intrusions which can be tolerated without compromising privacy. In this chapter, we focus on single-intrusion pan-privacy, where privacy is guaranteed against an adversary who observes the internal state of the mechanism at only a single point in time during the mechanism's execution.

Part of the initial motivation in the study of pan-privacy was the hope that it would allow for some of the flexibility of the central model, while also protecting against a stronger adversarial model. However, in contrast to central differential privacy which, for various learning problems, enables exponential improvements in sample complexity relative to the local model [KLN$^+$11, DF19] it has been previously shown that polynomial improvements are the most one can hope for in the case of r-intrusion pan-privacy relative to sequential local privacy when $r \geq 2$. Indeed, any 2-intrusion pan-private mechanism may be translated into a sequential local protocol with at most polynomial blow-up in the number of samples required [AJM20]. This result is obtained by the observation that if the internal state can be observed at two consecutive moments in time, then it must be essentially behave as a local randomizer.

We derive a result which applies specifically to realizable and agnostic learning as opposed to giving a general simulation. However, for these specific tasks, we obtain a surprising sample-complexity equivalence between even single-intrusion pan-privacy and sequential LDP, showing that the former can offer only limited sample-complexity advantages. We show, for $\alpha \in (0, 1]$, given a single-intrusion ε-pan-private mechanism which realizably $(\alpha, \frac{1}{2} + \Omega(1))$-learns the concept class $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ with n samples, there exists a sequential ε-LDP protocol which realizably $(2\alpha, \beta)$-learns $\mathcal{C}$ with $n' = \mathrm{poly}(n, \log|\mathcal{U}|, 1/\alpha)$ samples (Theorem 54).

This result is obtained by taking advantage of statistical query dimension (SQ dimension) which is typically used to characterize the query complexity of learning under the statistical query model. In particular, we rely on variants of statistical query dimension as presented in [Fel17], adapting the definitions in that work slightly to suit our purposes. We show that, just as the query complexity of learning via statistical queries is characterized by statistical query dimension, so too is the sample complexity of learning under either sequential LDP or single-intrusion pan-privacy. Lower bounds in terms of statistical query dimension for LDP and single-intrusion pan-privacy are obtained by taking advantage of the respective information theoretic bounds which hold for these models, namely Lemma 3, due to [DR18], and Lemma 5, due to [CSU⁺19]. As it turns out, the quantity which appears in those bounds is closely related to statistical query dimension. Meanwhile, the upper bound in terms of statistical query dimension for LDP is obtained by simulating an SQ algorithm for learning due to [Fel17].

The relationship we derive between the sample complexities of sequential LDP and single-intrusion pan-privacy are obtained by comparing the respective sample-complexity characterizations of learning under these models, rather than by giving a direct reduction.

The research represented by this chapter has not been published elsewhere and is the result of joint work with Aleksandar Nikolov and Toniann Pitassi.

8.2 LDP lower bound for realizable learning in terms of statistical query dimension

Upper and lower bounds for realizable learning may be obtained in terms of the quantity which we will refer to as $\mathrm{SQD}^{\mathrm{R}}(\mathcal{C}, \alpha)$. This quantity is defined in [Fel17] where it is referred to as $\mathrm{cRSD}_{\bar{\kappa}_1}$ rather than $\mathrm{SQD}^{\mathrm{R}}$. In that work, an emphasis is placed on variants of this quantity which use a tail probability in place of an expectation, though an expectation version of statistical query dimension is used in that work to analyze certain decision problems. Using expectation allows the quantity to be more easily related to the information theoretic bounds we use.

Definition 49 (Statistical query dimension for realizable learning, [Fel17])**.** *Consider a concept class $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ and an accuracy parameter $\alpha \in (0,1]$. Let $\{\lambda_v\}_{v \in \mathcal{V}}$ consist of all distributions λ_v on $\mathcal{U} \times \{-1,1\}$ where $\exists c \in \mathcal{C},\ L_{\lambda_v}(c) = 0$. Let $\{\mu_w\}_{w \in \mathcal{W}}$ consist of all distributions μ_w on $\mathcal{U} \times \{-1,1\}$ where $\forall h \in \{-1,1\}^{\mathcal{U}},\ L_{\mu_w}(h) > \alpha$. Then the (realizable) statistical query dimension of $\mathcal{C}$ for parameter α is given by*

$$\mathrm{SQD}^{\mathrm{R}}(\mathcal{C}, \alpha) = \left(\inf_{w \in \mathcal{W}} \sup_{\zeta} \inf_{v \in \mathcal{V}} \mathop{\mathbb{E}}_{F \sim \zeta} \left[\left| \mathop{\mathbb{E}}_{X \sim \lambda_v} [F_X] - \mathop{\mathbb{E}}_{X \sim \mu_w} [F_X] \right| \right] \right)^{-1}$$

where the supremum is taken over the distribution ζ of a random function $F \in \mathbb{R}^{\mathcal{U} \times \{-1,1\}}$ satisfying $\|F\|_\infty \leq 1$.

Lemma 50. *Let $\mathcal{X} = \mathcal{U} \times \{-1,1\}$. Consider a concept class $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ and an accuracy parameter $\alpha \in (0,1]$.*

1. *$\mathcal{C}$ can be realizably $\left(\alpha, \frac{1}{2} - \Omega(1)\right)$-refuted under sequentially interactive ε-LDP with a data set of size n;*
2. *$\mathrm{SQD}^{\mathrm{R}}(\mathcal{C},\alpha) = \omega\left(\frac{1}{\alpha}\right)$ and $\mathcal{C}$ can be realizably $\left(\frac{\alpha}{3}, \frac{1}{2} - \Omega(1)\right)$-learned under sequentially interactive ε-LDP with a data set of size n.*

Then,

$$n' = \Omega\left(\frac{\mathrm{SQD}^{\mathrm{R}}(\mathcal{C},\alpha)^2}{\varepsilon^2}\right).$$

Proof. Any choice of $w \in \mathcal{W}$ and parameter distribution π determine, along with $\{\mu_v\}_{v\in\mathcal{V}}$, the matrix $U^{\pi,w} \in [-1,1]^{\mathcal{V}\times\mathcal{X}}$ with entries given by

$$u^{\pi,w}_{v,x} = \pi(v)(\lambda_v(x) - \mu_w(x)).$$

By Lemma 16 together with Grothendieck's inequality [Gro53], there exists a distribution $\hat{\pi}$ with support contained in that of π such that the matrix $M^w \in [-1,1]^{\mathcal{V}\times\mathcal{X}}$ with entries

$$m^w_{v,x} = \lambda_v(x) - \mu_w(x)$$

satisfies $\|M^w\|_{\ell_\infty \to L_2(\hat{\pi})} \leq K_G \cdot \|U^{\pi,w}\|_{\infty\to 1}$ for some universal constant K_G. In other words,

$$\begin{aligned} &\sup_{f\in\mathbb{R}^{\mathcal{X}}:\|f\|_\infty\leq 1} \mathop{\mathbb{E}}_{V\sim\hat{\pi}}\left[\left(\mathop{\mathbb{E}}_{X\sim\lambda_V}[f_X] - \mathop{\mathbb{E}}_{X\sim\mu_w}[f_X]\right)^2\right] \\ &\leq K_G^2 \cdot \sup_{f\in\mathbb{R}^{\mathcal{X}}:\|f\|_\infty\leq 1} \mathop{\mathbb{E}}_{V\sim\pi}\left[\left|\mathop{\mathbb{E}}_{X\sim\lambda_V}[f_X] - \mathop{\mathbb{E}}_{X\sim\mu_w}[f_X]\right|\right]^2. \end{aligned} \tag{8.1}$$

Moreover, if there exists a sequential protocol $\mathcal{M} : \mathcal{X}^n \to \mathcal{Z}$ which $\left(\alpha, \frac{1}{2} - \Omega(1)\right)$-refutes $\mathcal{C}$ then it may be used to distinguish a data set drawn from λ^n_π versus one drawn from μ^n_w. In this case, the KL-divergence bound of Lemma 3 implies a lower bound of $\Omega\left(\frac{1}{\varepsilon^2 n}\right)$ on the left-hand side of the previous inequality. By consequence,

$$\sup_{f\in\mathbb{R}^{\mathcal{X}}:\|f\|_\infty\leq 1} \mathop{\mathbb{E}}_{V\sim\pi}\left[\left|\mathop{\mathbb{E}}_{X\sim\lambda_V}[f_X] - \mathop{\mathbb{E}}_{X\sim\mu_w}[f_X]\right|\right] = \Omega\left(\frac{1}{\varepsilon\sqrt{n}}\right).$$

Since this holds for any choice of distribution π, we obtain

$$\inf_{\pi} \sup_{f\in\mathbb{R}^{\mathcal{X}}:\|f\|_\infty\leq 1} \mathop{\mathbb{E}}_{V\sim\pi}\left[\left|\mathop{\mathbb{E}}_{X\sim\lambda_V}[f_X] - \mathop{\mathbb{E}}_{X\sim\mu_w}[f_X]\right|\right] = \Omega\left(\frac{1}{\varepsilon\sqrt{n}}\right)$$

where the infimum is taken with respect to all distributions π on $\mathcal{V}$.

By Von Neumann's Minimax Theorem,

$$\sup_{\zeta} \inf_{v\in\mathcal{V}} \mathop{\mathbb{E}}_{V\sim\pi}\left[\left|\mathop{\mathbb{E}}_{X\sim\lambda_V}[f_X] - \mathop{\mathbb{E}}_{X\sim\mu_w}[f_X]\right|\right] = \Omega\left(\frac{1}{\varepsilon\sqrt{n}}\right)$$

where the supremum is taken over the distribution ζ of a random function $F \in \mathbb{R}^{\mathcal{X}}$ which always satisfies $\|F\|_\infty \leq 1$. By Cauchy-Schwarz,

$$\sup_{\zeta} \inf_{v \in \mathcal{V}} \mathop{\mathbb{E}}_{F \sim \zeta} \left[\left| \mathop{\mathbb{E}}_{X \sim \lambda_v} [F_X] - \mathop{\mathbb{E}}_{X \sim \mu_w} [F_X] \right| \right] = \Omega\left(\frac{1}{\varepsilon\sqrt{n}}\right)$$

Since this inequality holds for all choices of $w \in \mathcal{W}$,

$$\inf_{w \in \mathcal{W}} \sup_{\zeta} \inf_{v \in \mathcal{V}} \mathop{\mathbb{E}}_{F \sim \zeta} \left[\left| \mathop{\mathbb{E}}_{X \sim \lambda_v} [F_X] - \mathop{\mathbb{E}}_{X \sim \mu_w} [F_X] \right| \right] = \Omega\left(\frac{1}{\varepsilon\sqrt{n}}\right)$$

The multiplicative inverse of the quantity on the left-hand side is $\mathrm{SQD}^{\mathrm{R}}(\mathcal{C}, \alpha)$. Thus, our lower bound against refutation is obtained by rearranging to isolate n.

When sequential interactivity is allowed, it is straightforward to translate a learning algorithm into a refutation algorithm. In particular, given the output $c \in \mathcal{C}$ of a mechanism which $(\alpha/3, \frac{1}{2} - \Omega(1))$-learns $\mathcal{C}$ in the realizable setting, an additional $O\left(\frac{1}{\varepsilon^2\alpha^2}\right)$ samples suffice to estimate the loss of c on the underlying distribution within $\alpha/3$ with failure probability bounded by an arbitrarily small constant. This allows us to $(\alpha, \frac{1}{2} - \Omega(1))$-refute $\mathcal{C}$ realizably under sequentially interactive ε-LDP. Thus, the number of samples required for $(\alpha, \frac{1}{2} - \Omega(1))$-learning $\mathcal{C}$ under sequentially interactive ε-LDP is at least

$$n = \Omega\left(\frac{\mathrm{SQD}^{\mathrm{R}}(\mathcal{C}, 3\alpha)^2}{\varepsilon^2}\right) - O\left(\frac{1}{\varepsilon^2\alpha^2}\right).$$

When $\mathrm{SQD}^{\mathrm{R}}(\mathcal{C}, 3\alpha) = \omega\left(\frac{1}{\alpha}\right)$, then

$$n = \Omega\left(\frac{\mathrm{SQD}^{\mathrm{R}}(\mathcal{C}, 3\alpha)^2}{\varepsilon^2}\right).$$

□

8.3 Lower bound against single-intrusion pan-privacy for realizable learning in terms of statistical query dimension

Similar to the KL-divergence bound of Lemma 3 which we rely on to give lower bounds against sequential LDP, this section will rely on the total variation bound of Lemma 5, a generalization of the result of [CU21] used to obtain lower bounds against that model for a variety of learning and estimation problems. Following the same approach as in the previous section, we show that Lemma 5 implies a sample-complexity lower bound against single-intrusion pan-privacy for both realizable refutation and realizable learning in terms of statistical query dimension.

Lemma 51. *Let $\mathcal{X} = \mathcal{U} \times \{-1, 1\}$. Consider a concept class $\mathcal{C} \subseteq \{-1, 1\}^{\mathcal{U}}$ and an accuracy parameter $\alpha \in (0, 1]$. Suppose at least one of the following conditions holds:*

1. *$\mathcal{C}$ can be realizably (α, β)-refuted under single-intrusion ε-pan-privacy with a data set of size n;*
2. *$\mathcal{C} = \omega\left(\frac{1}{\alpha}\right)$ and $\mathcal{C}$ can be realizably $\left(\frac{\alpha}{3}, \frac{1}{2} - \Omega(1)\right)$-learned under single-intrusion ε-pan-privacy with a data set of size n.*

Then,

$$n = \Omega\left(\frac{\mathrm{SQD}^{\mathrm{R}}(\mathcal{C}, \alpha)^2}{\varepsilon}\right).$$

Proof. Consider again the problem of realizable refutation where we wish to distinguish between distributions $\{\lambda_v\}_{v \in \mathcal{V}}$ which agree with $\mathcal{C}$ from those distributions $\{\mu_w\}_{w \in \mathcal{W}}$ on which each $h \in \{-1,1\}^{\mathcal{U}}$ has loss at least α. If, for $w \in \mathcal{W}$ and a distribution $\hat{\pi}$ on $\mathcal{V}$, $\mathcal{M} : \mathcal{X}^n \to \{-1,1\}$ is able to realizably $\left(\alpha, \frac{1}{2} - \Omega(1)\right)$-refutes by $\mathcal{C}$, then, by Lemma 5,

$$\max_{f \in \mathbb{R}^{\mathcal{X}} : \|f\|_\infty \leq 1} \mathop{\mathbb{E}}_{V \sim \hat{\pi}} \left[\left(\mathop{\mathbb{E}}_{X \sim \lambda_V}[f_X] - \mathop{\mathbb{E}}_{X \sim \mu_w}[f_X]\right)^2\right] = \Omega\left(\frac{1}{n\varepsilon}\right).$$

The expression on the left-hand side is identical to the left-hand of (8.1). Hence, we may follow the derivation of Lemma 50 to obtain our result.

□

8.4 Upper bounds for realizable learning in terms of SQ dimension

In [Fel17], the following upper bound for learning in the statistical query model is obtained in terms of $\mathrm{SQD}^{\mathrm{R}}$.

Theorem 52 ([Fel17])**.** *Let $\mathcal{X} = \mathcal{U} \times \{-1,1\}$. Let $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ be a concept class. Let $d = \mathrm{SQD}^{\mathrm{R}}(\mathcal{C}, \alpha)$ Then $\mathcal{C}$ can be realizably $(\alpha + 3/d, \beta)$-learned in the statistical query model with*

$$O\left(d^3 \cdot \log(|\mathcal{U}|) \cdot \log(1/\beta)\right)$$

statistical queries of tolerance $1/\sqrt{d}$.

Corollary 53. *Let $\mathcal{X} = \mathcal{U} \times \{-1,1\}$. Let $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{X}}$ be a concept class. Let $\varepsilon > 0$. Let $\alpha, \beta \in (0,1]$. Let $d = (\mathrm{SQD}^{\mathrm{R}}(\mathcal{C}, \alpha))$. Then $\mathcal{C}$ can be realizably $(\alpha + 3/d, \beta)$-learned with a sequentially interactive ε-LDP protocol $\mathcal{M} : \mathcal{X}^n \to \{-1,1\}^{\mathcal{U}}$ which takes as input a data set of size*

$$n = \widetilde{O}\left(\frac{d^4 \cdot \log(|\mathcal{U}|) \cdot \log(1/\beta)}{\varepsilon^2}\right)$$

Proof. By [KLN+11], an adaptive statistical query algorithm, making T statistical queries of tolerance τ to an underlying distribution λ, may be simulated by a sequentially interactive ε-LDP protocol $\mathcal{M} : \mathcal{X}^n \to \{-1,1\}^{\mathcal{U}}$ which takes as input an i.i.d. data set $\overline{X} \sim \lambda^n$ of size $n = O\left(\frac{T \log(T/\beta)}{\varepsilon^2 \tau^2}\right)$. The total variation between the distributions on the outputs of the two algorithms is at most β.

By applying this transformation to the statistical query algorithm given by Theorem 52, we obtain our local protocol.

□

Proof. An adaptive statistical query algorithm, making T statistical queries of tolerance τ to an underlying distribution λ, may be simulated by a single-intrusion ε-pan-private protocol $\mathcal{M} : \mathcal{X}^n \to$

$\{-1,1\}^{\mathcal{U}}$ which takes as input an i.i.d. data set $\overline{X} \sim \lambda^n$ of size $n = O\left(\frac{T\log(T/\beta)}{\varepsilon\tau}\right)$. The total variation between the distributions on the outputs of the two algorithms is at most β.

Apply this transformation to the statistical query algorithm given by Theorem 52 when $\tau = \min(\mathrm{SQD}^{\mathrm{R}}(\mathcal{C},\alpha)^{-1},\alpha)$.

□

8.5 Equivalence of single-intrusion pan-privacy and sequential LDP for realizable learning

Combining the lower bound of Lemma 51 and the upper bound of Corollary 53 allows us to derive the following result, which bounds the number of samples required to learn a concept class $\mathcal{C}$ under single-intrusion pan-privacy in terms of the number of samples required to learn $\mathcal{C}$ under sequential LDP.

Theorem 54. *Let $\beta \in (0,1/2)$ be a constant. Let $\mathcal{X} = \mathcal{U} \times \{-1,1\}$. Let $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ be a concept class. Let $\alpha \in (0,1/2)$. Suppose at least one of the following conditions holds:*

1. *$\mathcal{C}$ can be realizably $(\alpha, \frac{1}{2} - \Omega(1))$-refuted under single-intrusion ε-pan-privacy with a data set of size n;*
2. *$\mathcal{C} = \omega\left(\frac{1}{\alpha}\right)$ and $\mathcal{C}$ can be realizably $\left(\frac{\alpha}{3}, \frac{1}{2} - \Omega(1)\right)$-learned under single-intrusion ε-pan-privacy with a data set of size n.*

Then $\mathcal{C}$ can also be realizably $(\alpha + O(1/\sqrt{\varepsilon n}), \beta)$-learned under sequential ε-LDP with an input of size

$$n' = \widetilde{O}\left(n^2 \cdot \log(|\mathcal{U}|) \cdot \log(1/\beta)\right).$$

Proof. Suppose $\mathcal{C}$ can be realizably (α,β)-learned under single-intrusion ε-pan-privacy with an input of size n. If either conditions 1 or 2 hold, then Lemma 51 implies

$$n = \Omega\left(\frac{\mathrm{SQD}^{\mathrm{R}}(\mathcal{C},\alpha)^2}{\varepsilon}\right).$$

Equivalently,

$$\mathrm{SQD}^{\mathrm{R}}(\mathcal{C},\alpha) = O(\sqrt{\varepsilon n}).$$

Now, by Corollary 53, $\mathcal{C}$ can be realizably $(\alpha + O(1/\sqrt{\varepsilon n}), \beta)$-learned under sequentially interactive LDP given an input of size

$$n' = \widetilde{O}\left(n^2 \cdot \log(|\mathcal{U}|) \cdot \log(1/\beta)\right).$$

□

It is straightforward to transform a sequentially interactive ε-LDP protocol into an r-intrusion ε-pan-private protocol (for arbitrary $r \geq 1$) which simulates its transcript with a data set of the same size, as shown in [AJM20]. This is done by maintaining the transcript of the local protocol as the internal state of the pan-private protocol.

8.6 Characterizations of agnostic learning

While so far we have discussed statistical query dimension only in the context of realizable learning, it is also possible, following [Fel17], to adapt these techniques to characterize agnostic learning. Consider the following definition.

Definition 55 (Statistical query dimension for agnostic learning)**.** *Consider a concept class $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ and an accuracy parameter $\alpha \in (0,1]$. For $\theta \in [0,1]$, let $\{\lambda_v\}_{v\in\mathcal{V}_\theta}$ consist of all distributions λ_v on $\mathcal{U} \times \{-1,1\}$ where $\exists c \in \mathcal{C}$, $L_{\lambda_v}(c) \leq \theta$. Let $\{\mu_w\}_{w\in\mathcal{W}_{\theta+\alpha}}$ consist of all distributions μ_w on $\mathcal{U} \times \{-1,1\}$ where $\forall h \in \{-1,1\}^{\mathcal{U}}$, $L_{\mu_w}(c) > \theta + \alpha$. Then the (agnostic) statistical query dimension of $\mathcal{C}$ for parameter α is given by*

$$\mathrm{SQD}^{\mathrm{A}}(\mathcal{C},\alpha) = \left(\inf_{\theta\in[0,1]} \inf_{w\in\mathcal{W}_{\theta+\alpha}} \sup_{\zeta} \inf_{v\in\mathcal{V}_\theta} \mathop{\mathbb{E}}_{F\sim\zeta}\left[\left| \mathop{\mathbb{E}}_{X\sim\lambda_v}[F_X] - \mathop{\mathbb{E}}_{X\sim\mu_w}[F_X] \right| \right] \right)^{-1}$$

where the supremum is taken over the distribution ζ of a random function $F \in \mathbb{R}^{\mathcal{U}\times\{-1,1\}}$ which always satisfies $\|F\|_\infty \leq 1$.

Lower bounds against agnostic learning under sequentially interactive LDP and single-intrusion pan-privacy in terms of $\mathrm{SQD}^{\mathrm{A}}(\mathcal{C},\alpha)$ may be obtained by adapting the proofs of Lemma 50 and 51. This is based on the observation that agnostic learning allows us to distinguish between the classes of distributions $\{\lambda_v\}_{v\in\mathcal{V}_\theta}$ and $\{\mu_w\}_{w\in\mathcal{W}_{\theta+\alpha}}$. Meanwhile, in [Fel17], it is observed that the upper bound for realizable learning under the SQ model (Theorem 52) may be adapted to obtain an upper bound for agnostic learning in terms of $\mathrm{SQD}^{\mathrm{A}}(\mathcal{C},\alpha)$. In this way, an analogous result to Theorem 54 may be obtained for agnostic learning.

8.7 Open problems

Though we derive a relationship between the sample complexities of learning under local versus pan-privacy (Corollary 54), we do not directly show how to translate a given pan-private learner into a local one. This is similar to the relationship obtained between the sample complexities of refutability and learnability, discussed in the Open Problems (Section 6.5) of Chapter 6. It would be interesting to see a direct construction which agrees with our result.

Chapter 9

CSQ learning

9.1 Overview

The correlational statistical query (CSQ) model is a special case of the statistical query model where each queried function $q : \mathcal{U} \times \{-1,1\} \to [-1,1]$ is required to be of the form

$$q(a,b) = f(a) \cdot b$$

for some function $f : \mathcal{U} \to [-1,1]$. Many queries of interest can be expressed in this way. In particular, estimating the loss of a hypothesis $h : \mathcal{U} \to \{-1,1\}$ on some underlying distribution λ is equivalent to approximating the correlation of h with λ, since

$$L_\lambda(h) = \frac{1}{2} - \frac{1}{2} \cdot \mathop{\mathbb{E}}_{(A,B)\sim\lambda}[h(A) \cdot B].$$

In this chapter, we demonstrate a close relationship between agnostic learning under the CSQ model and agnostic learning under non-interactive LDP, enabling query complexity lower and upper bounds for the CSQ model to be translated into sample complexity lower and upper bounds for non-interactive LDP. To do so, we show that, just as the approximate γ_2 norm characterizes the sample complexity of agnostic learning under non-interactive LDP, it also characterizes the query complexity of agnostic learning under the CSQ model. We give the following upper bound.

Theorem 56. *There exists a CSQ algorithm such that, for any k statistical queries Q with workload matrix W, the algorithm returns a random $Z \in \mathbb{R}^k$ such that*

$$\mathbb{E}\left[\|Wh - Z\|_\infty\right] \le \alpha$$

with at most

$$J = O\left(\frac{\gamma_2(W, \alpha/2)^2 \log k}{\alpha^2}\right)$$

queries of tolerance $\tau = 1/\sqrt{J}$.

Denoting by $W_\mathcal{C}$ the concept matrix associated with a concept class $\mathcal{C}$ of size k, then Applying this result to the query workload with workload matrix $(-W_\mathcal{C}, W_\mathcal{C})$ gives a CSQ algorithm for learning $\mathcal{C}$

agnostically with at most

$$J = O\left(\frac{\gamma_2(W_{\mathcal{C}}, \alpha/2)^2 \log k}{\alpha^2}\right)$$

queries of tolerance $\tau = 1/\sqrt{J}$.

This may be contrasted with the lower bound which we obtain.

Theorem 57. *Let* $\alpha \in (0,1]$. *Let* $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ *be a concept class with concept matrix* $W \in \{-1,1\}^{\mathcal{C}\times\mathcal{U}}$. *For some* $\alpha' = \Omega(\alpha/\log(\alpha))$, *the number of queries required to learn* $\mathcal{C}$ *agnostically with accuracy* α' *via a CSQ algorithm of tolerance* τ *is at least*

$$\Omega\left(\frac{\tau^2 \cdot \gamma_2(W,\alpha)}{\alpha}\right) - 1$$

so long as $\tau \leq 2\alpha'$. *When* $\tau > 2\alpha'$, *then we cannot learn* $\mathcal{C}$ *agnostically with accuracy* α' *via an SQ algorithm of tolerance* τ, *unless* $\mathcal{C}$ *consists of only two distinct concepts* c_1, c_2 *where, for all* $a \in \mathcal{U}$, $c_1(a) \neq c_2(a)$.

Additionally, the algorithm we give for learning under the CSQ framework is non-adaptive while our lower bound applies even to adaptive CSQ algorithms, implying that adaptivity adds little additional power in the context of agnostic learning under CSQ.

The research represented by this chapter has not been published elsewhere and is the result of joint work with Aleksandar Nikolov and Toniann Pitassi.

9.2 Upper bound

Recall that the approximate factorization norm is given by

$$\gamma_2(W,\alpha) = \min\{\gamma_2(\widetilde{W}) \ : \ \|W - \widetilde{W}\|_{1\to\infty} \leq \alpha/2\}$$

where

$$\gamma_2(\widetilde{W}) = \min\{\|R\|_{2\to\infty}\|A\|_{1\to 2} \ : \ \widetilde{W} = RA\}.$$

Answering the linear queries associated with A allows us to reconstruct the answers to W using R. In Chapter 5, we saw this approach applied under non-interactive LDP. We wish to use the same approach while guaranteeing that only statistical queries are posed to the oracle. Indeed, it is tempting to ask the statistical query oracle for the answers to the queries given by A. After all, they are correlational statistical queries. However, doing so would cause the query complexity to scale with the number of rows in A. Instead, we adapt the local randomizer used in that local protocol to an analogous set of correlational queries. An alternative approach would be to apply the reduction of [KLN+11]. Indeed it is possible to simulate binary symmetric lower randomizers in this way with all necessary queries expressible as correlational statistical queries. However, that approach gives slightly worse bounds.

To prove Theorem 56, we take advantage of the following characterization of subgaussian random variables. See Chapter 2 of [BLM13] for a proof.

Fact 58. *There exist universal constants* C_1, C_2 *such that:*

- *If Z is a mean zero, σ-subgaussian random variable on $\mathbb{R}^m$, then, for every even p, for every $v \in \mathbb{R}^m$, $\|\langle Z, v\rangle\|_p \leq C_1 \sigma \sqrt{p} \|v\|_2$.*

- *If, for every even p, for every $v \in \mathbb{R}^m$, the random variable $Z \in \mathbb{R}^m$ satisfies $\|\langle Z, v\rangle\|_p \leq C_1 \sigma \sqrt{p} \|v\|_2$, then Z is σ-subgaussian.*

An alternative approach to obtaining a similar result is to apply the reduction of [KLN+11] which translates a local protocol into a statistical query algorithm. In the case where the local randomizers are binary and symmetric, the statistical queries required for the reduction can be expressed as correlational queries. However, the result obtained in this way is weaker than Theorem 56.

Proof of Theorem 56. Let λ be the underlying distribution $\mathcal{X}$. Let h denote the probability vector of λ so that h_x is the probability of x under λ.

Similar to our approach in the local model, we take advantage of the factorization $W = RA$. Without loss of generality, we may assume $\|A\|_{1\to 2} = 1$ since we may scale A down by $\|A\|_{1\to 2} = 1$ and R proportionally up. Furthermore, although each column is guaranteed to have ℓ_1 norm at most 1, we may assume, without loss of generality, that each column has ℓ_1 norm exactly 1. Otherwise, we could redefine $A \in \mathbb{R}^{d\times T}$ to be the matrix $A' \in \mathbb{R}^{(d+1)\times T}$ with each column $(a_1, \dots, a_d)$ extended to $(a_1, \dots, a_d, a_{d+1})$ where we define $a_{d+1} = \sqrt{1 - a_1^2 - \cdots - a_d^2}$. At the same time, $R \in \mathbb{R}^{k\times d}$ can be redefined to be the matrix $R' \in \mathbb{R}^{k\times(d+1)}$ whose last row is the zero vector. In this way, the first d entries of the answer to $R'A'$ will provide the answer to $W = RA$.

Our CSQ algorithm draws the variables $Z_1, \dots, Z_J$ from the standard multivariate Guassian distribution $\mathcal{N}(0, I)$ on $\mathbb{R}^d$. For each $t \in [J]$, we ask the query given by $q_{Z_t}(x) = \operatorname{sign}(\langle Z_t, Ae_x\rangle)$, where e_x denotes the standard basis vector of $\mathbb{R}^x$ corresponding to x so that Ae_x is column x of A. For $t \in [J]$, let r_t denote the approximation of $q_{Z_t}(\lambda)$ returned by the oracle so that $|r_t - q_{Z_t}(\lambda)| \leq \tau$. We will use $\frac{\sqrt{\pi} r_t Z_t}{2}$ as a proxy for Ah. In particular, we will see that $\frac{\sqrt{\pi} r_t Z_t}{2}$ has expectation close to Ah and that it is well-concentrated. Taking the average of these values, namely

$$\frac{1}{J} \sum_{t\in[J]} \frac{\sqrt{\pi} r_t Z_t}{2},$$

will give us control over the concentration by our choice of J, to be determined. Using the above average to approximate Ah, our algorithm returns the matrix product

$$R\left(\frac{1}{J} \sum_{t\in[J]} \frac{\sqrt{\pi} r_t Z_t}{2}\right),$$

so as to obtain an approximation of $Wh = RAh$.

Correlationality: To see that each of the queries posed by our algorithm is in fact correlational, note

$$q_{Z_t}(-x) = \operatorname{sign}(\langle Z_t, Ae_{-x}\rangle) = \operatorname{sign}(\langle Z_t, -Ae_x\rangle) = -\operatorname{sign}(\langle Z_t, Ae_x\rangle) = -q_{Z_t}(x),$$

where the second equality uses the fact that, since A is a symmetric matrix, its column x is the additive inverse of its column $-x$.

Accuracy: We have

$$r_t = q_{Z_t}(\lambda) + \rho_t(Z_1, \dots, Z_t),$$

where $\rho_t \in [-\tau, \tau]$ is the error with which the oracle chooses to answer q_{Z_t}. We regard ρ_t as a random variable which can depend on $Z_1, \ldots, Z_t$ since the oracle may respond adversarially. We have

$$R\left(\frac{1}{J}\sum_{t\in[J]}\frac{r_t Z_t}{\sqrt{\pi}}\right) = \frac{1}{J}\sum_{t\in[J]}\frac{q_{Z_t}(\lambda) R Z_t}{\sqrt{\pi}} + \frac{1}{J}\sum_{t\in[J]}\frac{\rho_t R Z_t}{\sqrt{\pi}}$$

Hence, we may bound the error of algorithm as

$$\begin{aligned}&\mathbb{E}\left[\left\|Wh - R\left(\frac{1}{J}\sum_{t\in[J]}\frac{r_t Z_t}{\sqrt{\pi}}\right)\right\|_\infty\right]\\ &\le \mathbb{E}\left[\left\|Wh - R\left(\frac{1}{J}\sum_{t\in[J]}\frac{q_{Z_t}(\lambda) Z_t}{\sqrt{\pi}}\right)\right\|_\infty\right] + \mathbb{E}\left[\left\|R\left(\frac{1}{J}\sum_{t\in[J]}\frac{\rho_t Z_i}{\sqrt{\pi}}\right)\right\|_\infty\right]. \end{aligned} \tag{9.1}$$

We proceed by bounding each of the terms on the right-hand side, starting with the first term. We have

$$\mathop{\mathbb{E}}_{Z_t}[q_{Z_t}(\lambda) Z_t] = \mathop{\mathbb{E}}_{Z_t}\left[\mathop{\mathbb{E}}_{x\sim\lambda}[q_{Z_t}(x) Z_t]\right] = \mathop{\mathbb{E}}_{x\sim\lambda}\left[\mathop{\mathbb{E}}_{Z_t}[q_{Z_t}(x) Z_t]\right].$$

To compute the inner expectation, note that, for all $x \in \mathcal{X}$,

$$\langle q_{Z_t}(x) Z_t, Ae_x\rangle = \langle \operatorname{sign}(\langle Z_t, Ae_x\rangle) Z_t, Ae_x\rangle = \operatorname{sign}(\langle Z_t, Ae_x\rangle)\langle Z_t, Ae_x\rangle = |\langle Z_t, Ae_x\rangle|.$$

Hence,

$$\langle \mathop{\mathbb{E}}_{Z_t}[q_{Z_t}(x) Z_t], Ae_x\rangle = \mathop{\mathbb{E}}_{Z_t}[\langle q_{Z_t}(x) Z_t, Ae_x\rangle] = \mathop{\mathbb{E}}_{Z_t}[|\langle Z_t, Ae_x\rangle|] = \frac{2\|Ae_x\|_2}{\sqrt{\pi}} = \frac{2}{\sqrt{\pi}}.$$

Meanwhile, for any vector $v \perp Ae_x$,

$$\langle \mathop{\mathbb{E}}_{Z_t}[q_{Z_t}(x) Z_t], v\rangle = \mathop{\mathbb{E}}_{Z_t}[q_{Z_t}(x)\langle Z_t, v\rangle] = \mathop{\mathbb{E}}_{Z_t}[\operatorname{sign}(\langle Z_t, Ae_x\rangle)\langle Z_t, v\rangle] = \mathop{\mathbb{E}}_{Z_t}[\operatorname{sign}(\langle Z_t, Ae_x\rangle)]\cdot\mathbb{E}[\langle Z_t, v\rangle] = 0$$

since $\langle Z_t, Ae_x\rangle$ and $\langle Z_t, v\rangle$ are independent. It follows that

$$\mathop{\mathbb{E}}_{Z_t}[q_{Z_t}(x) Z_t] = \frac{2}{\sqrt{\pi}} Ae_x,$$

which implies

$$\mathop{\mathbb{E}}_{Z_t}[q_{Z_t}(\lambda) Z_t] = \mathop{\mathbb{E}}_{x\sim\lambda}[Ae_x] = \frac{2}{\sqrt{\pi}} Ah.$$

Hence,

$$\mathop{\mathbb{E}}_{Z_t}\left[R\left(\frac{1}{J}\sum_{t\in[J]}\frac{\sqrt{\pi} q_{Z_t}(\lambda) Z_t}{2}\right)\right] = RAh = Wh.$$

Now we wish to show that the expression inside the expectation is concentrated around its expectation. In particular we will show that it is subgaussian. To do so, we take advantage of Fact 58. For any $v \in \mathbb{R}^{\mathcal{X}}$ and any even integer $p \ge 2$, we have

$$(\mathbb{E}[\langle q_{Z_t}(\lambda) Z_t, v\rangle^p])^{1/p} \le (\mathbb{E}[|q_{Z_t}(\lambda)|^p \cdot |\langle Z_t, v\rangle|^p])^{1/p} \le (\mathbb{E}[|\langle Z_t, v\rangle|^p])^{1/p} \le C_2\sqrt{p}\|v\|_2$$

so, by Fact 58, $q_{Z_t}(\lambda)Z_t$ is $O(1)$-subgaussian. Thus, $\frac{1}{J}\sum_{t\in[J]}\frac{\sqrt{\pi}q_{Z_t}(\lambda)Z_t}{2}$ is $O\left(\frac{1}{\sqrt{J}}\right)$-subgaussian, since it is the mean of J independent 1-subgaussian random variables. It follows then that each co-ordinate of

$$R\left(\frac{1}{J}\sum_{t\in[J]}\frac{\sqrt{\pi}q_{Z_t}(\lambda)Z_t}{2}\right)$$

is $O\left(\frac{\|R\|_{2\to\infty}}{\sqrt{n}}\right)$-subgaussian. Together with the fact that its mean is Wh, this implies

$$\left\|Wh - R\left(\frac{1}{J}\sum_{t\in[J]}\frac{\sqrt{\pi}q_{Z_t}(\lambda)Z_t}{2}\right)\right\|_\infty = O\left(\frac{\|R\|_{2\to\infty}\sqrt{\log k}}{\sqrt{J}}\right).$$

It remains to bound the last term on the right-hand side of (9.1). Appealing to Fact 58, we show the term to be subgaussian by bounding, for $t \in [J]$ for all even p, and every $v \in \mathbb{R}^{\mathcal{X}}$,

$$\left\|\left\langle\sum_{t\in[J]}\rho_t Z_t, v\right\rangle\right\|_p \le \sum_{t\in[J]}\|\langle\rho_t Z_t, v\rangle\|_p \le \sum_{t\in[J]}\|\rho_t\langle Z_t, v\rangle\|_p \le \sum_{t\in[J]}\tau\,\|\langle Z_t, v\rangle \le C_2 J\tau\|\sqrt{p}\|v\|_2.$$

This proves $\sum_{t\in[J]}\rho_t Z_t$ to be $J\tau$-subgaussian. Hence, $\frac{1}{J}\sum_{t\in[J]}\frac{\rho_t Z_t}{\sqrt{\pi}}$ is $\frac{\tau}{\sqrt{\pi}}$-subgaussian. It follows then that each coordinate of

$$R\left(\frac{1}{J}\sum_{t\in[J]}\frac{\sqrt{\pi}\rho_t Z_t}{2}\right)$$

is $(\sqrt{\pi}\tau\|R\|_{2\to\infty}/2)$-subgaussian. This implies

$$\mathbb{E}\left[\left\|R\left(\frac{1}{J}\sum_{t\in[J]}\frac{\sqrt{\pi}\rho_t Z_t}{2}\right)\right\|_\infty\right] \le O\left(\frac{\sqrt{\pi}\tau\|R\|_{2\to\infty}\sqrt{\log k}}{2}\right).$$

Returning to (9.1), we now have

$$\begin{aligned}&\mathbb{E}\left[\left\|Wh - R\left(\frac{1}{J}\sum_{t\in[J]}\frac{\sqrt{\pi}r_t Z_t}{2}\right)\right\|_\infty\right]\\ &\le O\left(\frac{\sqrt{\pi}\|R\|_{2\to\infty}\sqrt{\log k}}{2}\right) + O\left(\frac{\sqrt{\pi}\tau\|R\|_{2\to\infty}\sqrt{\log k}}{2}\right).\end{aligned}$$

In particular, if $\tau \le 1/\sqrt{J}$, then

$$\begin{aligned}&\mathbb{E}\left[\left\|Wh - R\left(\frac{1}{J}\sum_{t\in[J]}\frac{\sqrt{\pi}r_t Z_t}{2}\right)\right\|_\infty\right]\\ &\le O\left(\frac{\|R\|_{2\to\infty}\sqrt{\log k}}{\sqrt{J}}\right) = O\left(\frac{\gamma_2(W,\alpha)\sqrt{\log k}}{\sqrt{J}}\right).\end{aligned}$$

This allows us to guarantee that the error is bounded by α with at most

$$J = O\left(\frac{\gamma_2(W,\alpha)^2\log k}{\alpha^2}\right)$$

queries.

□

9.3 Lower bound

The correlational variance (Definition 59) of a collection $\{\lambda_v\}_{v\in[k]}$ of distributions on $\mathcal{U} \times \{-1,1\}$ enables lower bounds against those CSQ algorithms which, when acting on some λ_v, identify a function which is correlated with λ_v (Lemma 60).

Definition 59 ([MS20b])**.** *For each $v \in [k]$, let λ_v be a distribution on $\mathcal{U} \times \{-1,1\}$. Let π be a distribution on $[k]$. The correlational variance of $\{\lambda_v\}_{v\in[k]}$ is defined by*

$$\text{c-var}\left(\{\lambda_v\}_{v\in[k]}\right) = \max_{f:\mathcal{U}\to[-1,1]} \mathop{\mathbb{E}}_{V\sim\pi}\left[\left(\mathop{\mathbb{E}}_{(A,B)\sim\lambda_V}[B\cdot f(A)]\right)^2\right].$$

Lemma 60 (Generalization of [MS20b])**.** *For each $v \in [k]$, let λ_v be a distribution on $\mathcal{U} \times \{-1,1\}$. Let π be a distribution on $[k]$. Suppose we have a CSQ algorithm which, for all $v \in [k]$, when accessing λ_v with tolerance τ, outputs a function $h : \mathcal{U} \to \{-1,1\}$ such that*

$$\left|\mathop{\mathbb{E}}_{(A,B)\sim\lambda_v}[B\cdot h(A)]\right| \geq \alpha. \tag{9.2}$$

Then the number of queries posed by the algorithm is at least

$$\tau^2 \cdot \left(\frac{1}{\text{c-var}\left(\{\lambda_v\}_{v\in[k]}\right)} - \frac{1}{\alpha^2}\right).$$

In particular, if $\alpha = \Omega(\tau)$, then the number of queries is at least

$$\frac{\tau^2}{\text{c-var}\left(\{\lambda_v\}_{v\in[k]}\right)} - O(1).$$

We will use Lemma 60 to obtain lower bounds against agnostic learners in the CSQ model. In particular, we will use the fact that (9.2) is implied by

$$L_\lambda(h) \leq \frac{1}{2} - \frac{1}{2}\cdot\alpha$$

since

$$L_{\lambda_v}(h) = \frac{1}{2} - \frac{1}{2}\cdot\left|\mathop{\mathbb{E}}_{(A,B)\sim\lambda_v}[B\cdot h(A)]\right|.$$

Lemma 61 constructs the 'hard' distributions for which we will obtain our lower bound. This construction has parallels to the hard distributions used to obtain lower bounds against agnostic learning under non-interactive LDP, though various adaptations are necessary for the current setting. The proof of this result will rely on technical lemmas which we have used before, specifically the exponential binning of Lemma 18 as well as the relationship between γ_2^* and the $\infty \to 2$ operator norm given by Lemma 16.

Lemma 61. *Let $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{U}}$ be a concept class with concept matrix $W \in \{-1,1\}^{\mathcal{C}\times\mathcal{U}}$. Let $U \in \{-1,1\}^{\mathcal{C}\times\mathcal{U}}$, $\|U\|_1 = 1$, be the dual of W, so that*

$$\gamma_2(W,\alpha) = \frac{W \bullet U - \alpha}{\gamma_2^*(U)}.$$

Then there exists a distribution $\widehat{\pi}$ on $[k]$ as well as, for each $v \in [k]$, a distribution $\widetilde{\lambda}_v$ on $\mathcal{U}\times\{-1,1\}$, such that:

1. *for all v in the support of $\widehat{\pi}$,*

$$\mathop{\mathbb{E}}_{(A,B)\sim\widetilde{\lambda}_v}[B \cdot c_v(A)] \geq \frac{\alpha}{O(\log(1/\alpha))};$$

2. *the matrix $\widetilde{M} \in \mathbb{R}^{\mathcal{C}\times\mathcal{U}}$ with entries $\widetilde{m}_{v,a} = \lambda_v(a,1) - \lambda_v(a,-1)$ satisfies*

$$\|\widetilde{M}\|_{\ell_\infty \to L_2(\widehat{\pi})} \leq \frac{4\alpha\gamma_2^*(U)}{W \bullet U}.$$

Proof. For all $v \in [k]$, let

$$\pi(v) = \sum_{v\in[k]} |u_{v,a}|.$$

For $v \in [k]$, $a \in \mathcal{U}$, let

$$\lambda_v(a,1) = u^+_{v,a}/\pi(v)$$
$$\lambda_v(a,-1) = u^-_{v,a}/\pi(v).$$

Our dual formulation implies $W \bullet U > 0$. Thus,

$$\mathop{\mathbb{E}}_{V\sim\pi}\left[\mathop{\mathbb{E}}_{(A,B)\sim\lambda_V}[B \cdot c_V(A)]\right] = W \bullet U > 0.$$

However, we want a lower bound on $\mathop{\mathbb{E}}_{(A,B)\sim\lambda_V}[B \cdot c_V(A)]$ which holds for all $v \in [k]$.

To remedy this issue, apply Lemma 18 with $a_v = \mathop{\mathbb{E}}_{(A,B)\sim\lambda_v}[B \cdot c_v(A)]$ and $\beta = \alpha/4$ to obtain a set $S \subseteq [k]$ as guaranteed by the lemma.

Define $\widetilde{\pi}$ as π conditional on S. In particular,

$$\widetilde{\pi}(v) = \begin{cases} \pi(v)/\pi(S), & \text{if } v \in S \\ 0, & \text{otherwise.} \end{cases}$$

For $v \in [k]$, let

$$\widetilde{\lambda}_v(a,1) = \frac{1 + \tau\pi(S)\cdot(\lambda_v(a,1) - \lambda_v(a,-1))}{2},$$
$$\widetilde{\lambda}_v(a,-1) = \frac{1 - \tau\pi(S)\cdot(\lambda_v(a,1) - \lambda_v(a,-1))}{2}$$

where $\tau = \frac{\alpha}{W\bullet U}$. In this way,

$$\widetilde{\lambda}_v(a,1) - \widetilde{\lambda}_v(a,-1) = \tau\pi(S)\cdot(\lambda_v(a,1) - \lambda_v(a,-1)).$$

It follows that the matrix $\widetilde{U} \in \mathbb{R}^{\mathcal{C} \times \mathcal{U}}$ with entries $\widetilde{u}_{v,a} = \widetilde{\pi}(v)\left(\lambda_v(a,1) - \lambda_v(a,-1)\right)$ satisfies

$$\widetilde{u}_{v,x} = \begin{cases} \tau u_{v,x}, & \text{if } v \in S \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to see from the definition of γ_2^* that this implies $\gamma_2^*(\widetilde{U}) \leq \gamma_2^*(U)$.

Moreover, for all $v \in [k]$,

$$\mathop{\mathbb{E}}_{(a,b)\sim\widetilde{\lambda}_v}[b \cdot c_v(a)] = \tau\pi(S) \mathop{\mathbb{E}}_{(a,b)\sim\widetilde{\lambda}_v}[b \cdot c_v(a)] \geq \frac{\alpha}{O(\log(1/\alpha))}.$$

Let $\widetilde{M} \in \mathbb{R}^{\mathcal{C} \times \mathcal{U}}$ be the matrix with entries $\widetilde{m}_{v,a} = \lambda_v(a,1) - \lambda_v(a,-1)$. Applying Lemma 16 to $\widetilde{M}$ and $\widetilde{\pi}$, gives a distribution $\widehat{\pi}$ on $[k]$ such that

$$\|\widetilde{M}\|_{\ell_\infty \to L_2(\widehat{\pi})} \leq 4\gamma_2^*(\widetilde{U}) \leq \frac{4\alpha\gamma_2^*(U)}{W \bullet U}.$$

□

Finally, we are equipped to prove the lower bound.

Proof of Theorem 57. Let $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k$ and $\widetilde{\pi}$ be the distributions guaranteed to exist by Lemma 61. The matrix $\widetilde{M}$ has entries $\widetilde{m}_{v,a} = \widetilde{\lambda}_v(a,1) - \widetilde{\lambda}_v(a,-1)$ and satisfies

$$\|\widetilde{M}\|_{\ell_\infty \to L_2(\widehat{\pi})} \leq \gamma_2^*(U) \leq \frac{4\alpha}{\gamma_2(W,\alpha)}.$$

In other words,

$$\max_{f:\mathcal{U}\to[-1,1]} \mathop{\mathbb{E}}_{V\sim\widehat{\pi}}\left[\left(\mathop{\mathbb{E}}_{(A,B)\sim\widetilde{\lambda}_V}[B \cdot f(A)]\right)^2\right] \leq \frac{4\alpha}{\gamma_2(W,\alpha)}.$$

At the same time, we have

$$\mathop{\mathbb{E}}_{(A,B)\sim\widetilde{\lambda}_v}[B \cdot c_v(A)] \geq \alpha', \qquad \text{for some } \alpha' = \Omega\left(\frac{\alpha}{\log(1/\alpha)}\right).$$

Equivalently,

$$L_{\lambda_v}(c_v) \leq \frac{1}{2} - \frac{\alpha'}{2}.$$

If we can learn $\mathcal{C}$ with accuracy $\alpha'/4$ on distribution λ_v, then we can identify some $h : \mathcal{U} \to \{\pm 1\}$ such that

$$L_{\lambda_v}(h) \leq \min_{v\in[k]} L_{\lambda_v}(c_v) + \frac{\alpha'}{4} \leq \frac{1}{2} - \frac{\alpha'}{4}.$$

By Lemma 60, achieving this for each $v \in [k]$, with a CSQ algorithm making queries of tolerance τ, requires the number of queries to be at least

$$\tau^2 \cdot \left(\frac{\gamma_2(W,\alpha)}{\alpha} - \frac{1}{(\alpha')^2}\right).$$

In the case where $\tau \leq \alpha'$, this gives the desired bound.

It remains to consider the trivial case where $\tau > \alpha'$ and the concept class consists of distinct concepts $c_1, c_2 \in \mathcal{C}$ such that, for some $a_1 \in \mathcal{U}$, $c_1(a_1) = c_2(a_1)$. To be distinct, the concepts must also disagree on some $a_2 \in \mathcal{U}$ so that $c_1(a_2) \neq c_2(a_2)$. Let $\rho = \alpha'/2$. Consider the distribution λ which gives the sample $(a_1, c_1(a_1)) = (a_1, c_1(a_1))$ with probability $1 - \rho$ and gives the sample $(a_2, c_1(a_2))$ with probability ρ. Consider also the distribution μ which gives the sample $(a_1, c_1(a_1)) = (a_1, c_1(a_1))$ with probability $1 - \rho$ and gives the sample $(a_2, c_2(a_2))$ with probability ρ. For $q : (\mathcal{U} \times \{-1, 1\}) \to [-1, 1]$, we have $|q(\lambda) - q(\mu)| \leq 2\rho$. Hence, a statistical query oracle of tolerance $\tau > 2\rho = \alpha'$ can give the same answer to the query given by q regardless of whether the underlying distribution is λ or μ. However, any particular hypothesis $h : \mathcal{U} \to \{-1, 1\}$ must have loss at least $\rho/2$ on at least one of λ or μ so a learner of accuracy $\alpha' = \rho/2$ cannot return the same hypothesis on both distributions.

□

9.4 Open problems

It has been a recurring theme in this work (see Section 6.5 and Section 8.7) that the equivalences we show – like the one for learning between the query complexity of CSQ algorithms and the sample complexity of non-interactive LDP – are indirect, reliant on showing that the given tasks are characterized by the same quantity. Instead, it would be nice to see a direct reduction which translates one type of algorithm directly into the other.

It would also be nice to see a tighter characterization of CSQ learning in terms of the approximate γ_2 norm.

Appendix A

Information-theoretic bounds

A.1 Mutual information bound for sequential interactivity

Proof of Lemma 1 for sequential interactivity. Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Z}$ be a sequentially interactive local protocol. Consider a random dataset $\overline{X} = (X_1, \ldots, X_n) \in \mathcal{X}^n$. Each agent i returns some random $Y_i \in \mathcal{Y}$ which depends only on their input X_i and, when $i \geq 2$, the outputs $Y_1, \ldots, Y_{i-1}$ of the previous local agents. Without loss of generality, we may assume that Y_i is independent of $Y_1, \ldots, Y_{i-1}$ conditional on X_i and Y_{i-1}, since, if necessary, we may transform the protocol by having each agent relay the messages of the previous agents along with their own. Since $\mathcal{M}$ is ε-LDP, then

$$\forall i \in [n], S \subseteq \mathcal{Y}, y_{i-1} \in \mathcal{Y}, x_i, x_i' \in \mathcal{X}, \quad \frac{\Pr(Y_i \in S \mid Y_{i-1} = y_{i-1}, X_i = x_i)}{\Pr(Y_i \in S \mid Y_{i-1} = y_{i-1}, X_i = x_i')} \leq e^{\varepsilon}$$

The key observation we rely on is that each Y_i satisfies an analogous privacy criteria when viewed as a function of the entire sequence of samples $\overline{X}$. In particular,

$$\forall i \in \mathbb{N}, S \subseteq \mathcal{Y}, y_{i-1} \in \mathcal{Z}, \overline{x}, \overline{x}' \in \mathcal{X}^n, \quad \frac{\Pr(Y_i \in S \mid Y_{i-1} = y_{i-1}, \overline{X} = \overline{x})}{\Pr(Y_i \in S \mid Y_{i-1} = y_{i-1}, \overline{X} = \overline{x}')} \leq e^{\varepsilon}. \tag{A.1}$$

It is worth emphasizing here that $\overline{x} \in \mathcal{X}^n$ and $\overline{x}' \in \mathcal{X}^n$ are allowed to be arbitrary datasets, not necessarily adjacent. The bound is a consequence of our assumptions regarding conditional independence which imply

$$\Pr(Y_i \in S \mid Y_{i-1} = y_{i-1}, \overline{X} = \overline{x}) = \Pr(Y_i \in S \mid Y_{i-1} = y_{i-1}, X_i = x_i)$$

$$\Pr(Y_i \in S \mid Y_{i-1} = y_{i-1}, \overline{X} = \overline{x}') = \Pr(Y_i \in S \mid Y_{i-1} = y_{i-1}, X_i = x_i').$$

We compute mutual information by application of the chain rule.

$$\begin{aligned} I\left(Y_1, \ldots, Y_n \;;\; \overline{X} \right) &= \sum_{i=1}^{n} I\left(Y_i \;;\; X \mid Y_1, \ldots, Y_{i-1} \right) \\ &= \sum_{i=1}^{n} I\left(Y_i \;;\; X \mid Y_{i-1} \right) \end{aligned}$$

It remains to bound each of the terms in the last expression. Indeed, by (A.1), we have

$$\begin{aligned}
&\sup_{y_i,\overline{x}} \left[\log \frac{\Pr[(Y_i,\overline{X}) = (y_i,\overline{x}) \mid Y_{i-1} = y_{i-1}]}{\Pr[Y_i \otimes \overline{X} = (y_i,\overline{x}) \mid Y_{i-1} = y_{i-1}]}\right] \qquad (A.2)\\
&\leq \sup_{y_i,\overline{x}} \left[\log \frac{\Pr[Y_i = y_i \mid \overline{X} = \overline{x},\ Y_{i-1} = y_{i-1}]}{\Pr[Y_i = y_i \mid Y_{i-1} = y_{i-1}]}\right]\\
&\leq \sup_{y_i,\overline{x}} \left[\log \frac{\Pr[Y_i = y_i \mid \overline{X} = \overline{x}, Y_{i-1} = y_{i-1}]}{\sum_{\overline{x}' \in \mathcal{X}^n} \Pr[Y_i = y_i \mid \overline{X} = \overline{x}', Y_{i-1} = y_{i-1}] \cdot \mathbb{P}\left[\overline{X} = \overline{x}' \mid Y_{i-1} = y_{i-1}\right]}\right]\\
&\leq \varepsilon
\end{aligned}$$

with the supremums taking y_i and $\overline{x}$ over the supports of $(Y_i \mid Y_{i-1} = y_{i-1})$ and $(\overline{X} \mid Y_{i-1} = y_{i-1})$ respectively. Here, $Y_i \otimes \overline{X}$ denotes a random variable from $\mathcal{Y} \times \mathcal{X}^n$ where its component from $\mathcal{Y}$ has the same marginal distribution as Y_i and its component from $\mathcal{X}^n$ has the same marginal distribution as $\overline{X}$, while, at the same, the two components are independent of each other. The quantity (A.2) is sometimes referred to as the "max divergence," in this case between $(Y_i, \overline{X}) = (y_i, \overline{x}) \mid Y_{i-1} = y_{i-1}$ and $Y_i \otimes \overline{X} = (y_i, \overline{x}) \mid Y_{i-1} = y_{i-1}$. It corresponds to a worst-case variant of KL-divergence, the latter being an average-case variant. By Lemma 3.18 of [DR14], an upper bound on max-divergence can surprisingly be translated into a tighter upper bound on KL-divergence. Applying that result, we obtain

$$\begin{aligned}
I\left(\ Y_i\ ;\ \overline{X} \mid Y_{i-1} = y_{i-1}\right) =&\mathrm{D}_{\mathrm{KL}}\left((Y_i, \overline{X} \mid Y_{i_1} = y_{i-1}) \parallel (Y_i \otimes \overline{X} \mid Y_{i-1} = y_{i-1})\right)\\
=&\mathop{\mathbb{E}}_{(y_i,\overline{x}) \sim (Y_i,\overline{X} \ \mid\ Y_{i-1}=y_{i-1})} \left[\log \frac{\Pr[(Y_i,\overline{X}) = (y_i,\overline{x}) \mid Y_{i-1} = y_{i-1}]}{\Pr[Y_i \otimes \overline{X} = (y_i,\overline{x}) \mid Y_{i-1} = y_{i-1}]}\right]\\
\leq&\varepsilon \cdot (e^{\varepsilon} - 1).
\end{aligned}$$

Substituting back into (A.1) and using the fact that, for $\varepsilon = O(1)$, $e^{\varepsilon} - 1 = O(\varepsilon)$, we obtain

$$I\left(\ Y_1, \ldots, Y_n\ ;\ \overline{X}\ \right) \leq n\varepsilon(e^{\varepsilon} - 1) = O(n\varepsilon^2).$$

It is straightforword to extend this analysis to account for compositional LDP. In particular, each round may be analyzed in the same way by conditioning on the transcript of previous rounds.

□

A.2 Total variation bound for single-intrusion pan-private protocols

Proof of Lemma 5. For $i \in \{0, \ldots, n\}$, define $\zeta_i = \mathcal{M}(\mu^i, \lambda_\pi^{n-i})$. Ultimately, we wish to bound

$$d_{\mathrm{TV}}\left(\mathcal{M}(\lambda_\pi^n), \mathcal{M}(\mu^n)\right) = d_{\mathrm{TV}}\left(\zeta_0, \zeta_n\right).$$

To do so, we apply the triangle inequality to obtain

$$d_{\mathrm{TV}}\left(\zeta_0, \zeta_n\right) \leq \sum_{i \in [n]} d_{\mathrm{TV}}(\zeta_{i-1}, \zeta_i) \qquad (A.3)$$

and proceed by bounding each of the terms $d_{\mathrm{TV}}(\zeta_{i-1}, \zeta_i)$.

Let $V \sim \pi$. Conditional on $V = v$, let $X_1, \ldots, X_n \sim \lambda_v^n$. Let $X_1', \ldots, X_n' \sim \mu^n$. Then a sample from ζ_{i-1} is given by

$$Y_i = \mathcal{M}(X_1', \ldots, X_{i-1}', X_i, \ldots, X_n),$$

while a sample from ζ_i is given by

$$Y_i' = \mathcal{M}(X_1', \ldots, X_i', X_{i+1}, \ldots, x_n).$$

Note that Y_i may be viewed as a post-processing of the internal state

$$I_i(X_1', \ldots, X_{i-1}', X_i) = \mathcal{M}_i(X_i, I_{i-1}(X_1', \ldots, X_{i-1}')).$$

Similarly, Y_i' may be viewed as the same post-processing of of the internal state

$$I_i(X_1', \ldots, X_{i-1}', X_i') = \mathcal{M}_i(X_i', I_{i-1}(X_1', \ldots, X_{i-1}')).$$

For notational simplicity, we allow total variation and KL-divergence to take random variables as input, in which case we consider those quantities as being applied to the distributions of those random variables. We have

$$\begin{aligned}
& d_{\mathrm{TV}}(\zeta_{i-1}, \zeta_i)^2 \\
& \leq d_{\mathrm{TV}}((I_i(X_1', \ldots, X_{i-1}', X_i), X_{i+1}, \ldots, X_n), (I_i(X_1', \ldots, X_i'), X_{i+1}, \ldots, X_n))^2, && \text{by post-processing} \\
& \leq \mathrm{D}_{\mathrm{KL}}\left((I_i(X_1', \ldots, X_{i-1}', X_i), X_{i+1}, \ldots, X_n) \parallel (I_i(X_1', \ldots, X_i'), X_{i+1}, \ldots, X_n)\right), && \text{by Pinsker's inequality} \\
& \leq \underset{V \sim \pi}{\mathbb{E}}\left[\mathrm{D}_{\mathrm{KL}}\left(I_i(X_1', \ldots, X_{i-1}', X_i), X_{i+1}, \ldots, X_n \mid V \parallel I_i(X_1', \ldots, X_i'), X_{i+1}, \ldots, X_n \mid V\right)\right], && \text{by convexity} \\
& \leq \underset{V \sim \pi}{\mathbb{E}}\left[\mathrm{D}_{\mathrm{KL}}\left(I_i(X_1', \ldots, X_{i-1}', X_i) \parallel I_i(X_1', \ldots, X_i') \mid V\right)\right], && \text{by independence} \\
& \leq O(\varepsilon^2) \cdot \underset{V \sim \pi}{\mathbb{E}}\left[\left(f(\lambda_V) - f(\mu)\right)^2\right].
\end{aligned}$$

The last inequality follows by viewing I_i as an ε-private local randomizer, with its input being the ith element – either X_i or X_i' – and the random variables $X_1', \ldots, X_i'$ and $X_i, \ldots, X_n$ being thought of as components of its internal randomness. In this way, we may apply the same argument as used in the proof of Lemma 2. Substituting back into (A.3) gives the desired bound.

□

Appendix B

Characterization of linear query release under non-interactive LDP

B.1 Properties of γ_2^*

Proof of Lemma 15. By Lemma 14, there exist diagonal matrices $P \in \mathbb{R}^{k\times k}$ and $Q \in \mathbb{R}^{T\times T}$, satisfying $\mathrm{Tr}(P^2) = \mathrm{Tr}(Q^2) = 1$, and a matrix $\widetilde{U}$ such that $U = P\widetilde{U}Q$ and $\|\widetilde{U}\|_{2\to 2} \le \gamma_2^*(U)$. Define $S = \{i : p_{ii}^2 \le \frac{2}{k}\}$. Then Markov's inequality shows that $|S| \ge \frac{k}{2}$. Furthermore,

$$\gamma_2^*(U) \ge \|\widetilde{U}\|_{2\to 2} \ge \|\Pi_S \widetilde{U}\|_{2\to 2} \ge \sqrt{\frac{k}{2}}\|\Pi_S P\widetilde{U}\|_{2\to 2},$$

where the first inequality follows because multiplying by a projection matrix can only decrease the $\|\cdot\|_{2\to 2}$ norm of the matrix, and the second inequality follows by the definition of S.

To finish the proof, we observe that

$$\|\Pi_S P\widetilde{U}\|_{2\to 2} \ge \|\Pi_S P\widetilde{U}Q\|_{\infty\to 2} = \|\Pi_S U\|_{\infty\to 2}.$$

Indeed, for any $x \in \mathbb{R}^T$, we have

$$\begin{aligned}\Pi_S P\widetilde{U}Qx &\le \|\Pi_S P\widetilde{U}\|_{2\to 2}\|Qx\|_2 \\ &\le \|\Pi_S P\widetilde{U}\|_{2\to 2}\sqrt{\mathrm{Tr}(Q^2)}\|x\|_\infty \\ &= \|\Pi_S P\widetilde{U}\|_{2\to 2}\|x\|_\infty,\end{aligned}$$

where the second inequality follows from Hölder's inequality. □

Proof of Lemma 16. Without loss of generality, we may assume π takes rational values. In particular, let $\widetilde{k} \in \mathbb{Z}$ be such that $\pi(i)\cdot\widetilde{k} \in \mathbb{Z}$ for all $i \in [k]$. Then M and π may be used to define the matrix $\widetilde{M} \in \mathbb{R}^{\widetilde{k}\times\mathcal{X}}$, obtained by taking, for each $i \in [k]$, $\pi(i)\cdot\widetilde{k}$ copies of row i from M.

By Lemma 15, there exists a set $S \subseteq [\widetilde{k}]$, $|S| \ge \frac{\widetilde{k}}{2}$, such that $\sqrt{\frac{\widetilde{k}}{2}}\|\Pi_S\widetilde{M}\|_{\infty\to 2} \le \gamma_2^*(\widetilde{M})$. Use S to define the function $\widetilde{\pi} : [k] \to [0,1]$ where $\widetilde{k}\widetilde{\pi}(i)$ is the number of rows selected from $\widetilde{M}$ by S which correspond to row i from M. Since $|S| \ge \frac{\widetilde{k}}{2}$, then $\sum_{i\in[k]}\widetilde{\pi}(i) \ge \frac{1}{2}$.

The equality $\sqrt{\widetilde{k}}\|M\|_{\ell_\infty\to L_2(\widetilde{\pi})} = \|\Pi_S\widetilde{M}\|_{\infty\to 2}$ follows because, for all $f \in \mathbb{R}^{\mathcal{X}}$,

$$\|\Pi_S\widetilde{M}f\|_2^2 = \sum_{i\in S}(\widetilde{M}_{v,*}\cdot f)^2 = \widetilde{k}\sum_{v\in[k]}\widetilde{\pi}(v)(M_{v,*}\cdot f)^2 = \widetilde{k}\|Mf\|^2_{L_2(\widetilde{\pi})}$$

where we have denoted row v of $\widetilde{M}$ by $\widetilde{M}_{v,*}$.

To show $\gamma_2^*(\widetilde{M}) \le \widetilde{k}\gamma_2^*(U)$, let $\widetilde{y}_1,\dots,\widetilde{y}_k, z_1,\dots,z_T$ be unit vectors in ℓ_2-norm which satisfy

$$\gamma_2^*(\widetilde{M}) = \sum_{i=1}^{\widetilde{k}}\sum_{j=1}^{T}\widetilde{m}_{i,j}\widetilde{y}_i^\top z_j,$$

as they are guaranteed to exist by Lemma 13. Without loss of generality, we may assume that if two rows i and i' of $\widetilde{M}$ have identical entries, then $\widetilde{y}_i = \widetilde{y}_{i'}$. Now, for all $i \in [k]$, let $y_i = \widetilde{y}_{\widetilde{i}}$, where $\widetilde{i}$ is one of the rows of $\widetilde{M}$ which was copied from row i in M. Then

$$\begin{aligned}\gamma_2^*(\widetilde{M}) &= \sum_{i=1}^{\widetilde{k}}\sum_{j=1}^{T}\widetilde{m}_{i,j}\widetilde{y}_i^\top z_j = \widetilde{k}\sum_{i=1}^{k}\sum_{j=1}^{T}\pi(i)m_{i,j}y_i^\top z_j \\ &= \widetilde{k}\sum_{i=1}^{k}\sum_{j=1}^{T}u_{i,j}y_i^\top z_j \le \widetilde{k}\gamma_2^*(U).\end{aligned}$$

Altogether, this gives

$$\|M\|_{\ell_\infty\to L_2(\widetilde{\pi})} = \sqrt{\frac{2}{\widetilde{k}}}\|\Pi_S\widetilde{M}\|_{\infty\to 2} \le \frac{2}{\widetilde{k}}\gamma_2^*(\widetilde{M}) \le 2\gamma_2^*(U).$$

Finally, normalize $\widetilde{\pi}$ to obtain the probability distribution $\widehat{\pi}$, given by $\widehat{\pi}(i) = \frac{\widetilde{\pi}(i)}{\sum_i \widetilde{\pi}(i)}$. Since $\sum_i \widetilde{\pi}(i) \ge \frac{1}{2}$, then $\widehat{\pi}(i) \le 2\widetilde{\pi}(i)$. This implies $\|M\|_{\ell_\infty\to L_2(\widehat{\pi})} \le 2\|M\|_{\ell_\infty\to L_2(\widetilde{\pi})}$, from which we get $\|M\|_{\ell_\infty\to L_2(\widehat{\pi})} \le 4\gamma_2^*(U)$. Since the rows that were copied from M to obtain $\widetilde{M}$ corresponded to those rows of U which were assigned non-zero probability by π, then, by the definitions of $\widetilde{pi}$ and $\widehat{\pi}$, it follows that the support of $\widehat{\pi}$ is a subset of the support of π. □

B.2 Scaling and subadditivity properties of approximate γ_2

Lemma 62 (Scaling properties of γ_2)**.** *Consider a matrix $W \in \mathbb{R}^{k\times T}$. Let $\alpha \ge 0$. Let $t \in [0,1]$. Then,*

$$t\cdot\gamma_2(W,\alpha) = \gamma_2(tW, t\alpha).$$

When $\alpha = 0$, this gives $t\cdot\gamma_2(W) = \gamma_2(tW)$.

Proof.

$$\begin{aligned}t\cdot\gamma_2(W,\alpha) &= t\cdot\min\{\|R\|_{2\to\infty}\|A\|_{1\to 2} \ : \ \|W - RA\| \le \alpha\} \\ &= \min\{\|R'\|_{2\to\infty}\|A'\|_{1\to 2} \ : \ \|tW - R'A'\| \le t\cdot\alpha\} \\ &= \gamma_2(tW, t\alpha)\end{aligned}$$

□

Lemma 63 (Subadditivity for γ_2)**.** *For matrices $W_1, W_2 \in \mathbb{R}^{k\times T}$,*

$$\gamma_2(W_1 + W_2) \le \gamma_2(W_1) + \gamma_2(W_1).$$

Proof. By definition, there exist matrices R_1, R_2, A_1, A_2 satisfying $W_1 = R_1 A_1$ and $W_2 = R_2 A_2$ such that $\gamma_2(W_1) = \|R_1\|_{2\to\infty}\|A_1\|_{1\to 2}$ and $\gamma_2(W_2) = \|R_2\|_{2\to\infty}\|A_2\|_{1\to 2}$. By rescaling if necessar, we may assume without loss of generality that $\|R_1\|_{2\to\infty} = \|A_1\|_{1\to 2} = \sqrt{\gamma_2(W_1)}$ and $\|R_2\|_{2\to\infty} = \|A_2\|_{1\to 2} = \sqrt{\gamma_2(W_2)}$. Now define

$$R = \begin{bmatrix} R_1 & R_2 \end{bmatrix} \qquad A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$$

Then, $RA = R_1A_1 + R_2A_2 = W_1 + W_2$. Recall also that the 2 → ∞ norm is given by the maximum ℓ_2 norm of a row. Thus,

$$\|R\|^2_{2\to\infty} = \|R_1\|^2_{2\to\infty} + \|R_2\|^2_{2\to\infty} = \gamma_2(W_1) + \gamma_2(W_2).$$

Also, since the 1 → 2 norm is given by the maximum ℓ_2 norm of a column, then

$$\|A\|^2_{1\to 2} = \|A_1\|^2_{1\to 2} + \|A_2\|^2_{1\to 2} = \gamma_2(W_1) + \gamma_2(W_2).$$

Therefore,

$$\|R\|_{2\to\infty}\|A\|_{1\to 2} = \gamma_2(W_1) + \gamma_2(W_2)$$

and hence

$$\gamma_2(W_1 + W_2) \le \gamma_2(W_1) + \gamma_2(W_2).$$

□

Lemma 64 (Subadditivity for approximate γ_2)**.** *Consider matrices $W_1, W_2 \in \mathbb{R}^{k\times T}$. Let $\alpha > 0$. Let $t \in [0, 1]$. Then,*

$$\gamma_2(W_1 + W_2, \alpha) \le \gamma_2(W_1, t\cdot\alpha) + \gamma_2(W_2, (1-t)\cdot\alpha).$$

Proof. By definition,

$$\gamma_2(W_1, t\cdot\alpha) = \gamma_2(W_1 + E_1), \qquad \gamma_2(W_2, (1-t)\cdot\alpha)) = \gamma_2(W_2 + E_2)$$

for some matrices $E_1, E_2 \in \mathbb{R}^{k\times T}$ which satisfy $\|E_1\|_\infty \le t\cdot\alpha$ and $\|E_2\|_\infty \le (1-t)\cdot\alpha$. Thus, by Lemma 63, we have

$$\begin{aligned}\gamma_2(W_1, t\cdot\alpha) + \gamma_2(W_2, (1-t)\cdot\alpha)) &= \gamma_2(W_1 + E_1) + \gamma_2(W_2 + E_2) \\ &\ge \gamma_2(W_1 + W_2 + E_1 + E_2) \\ &\ge \gamma_2(W_1 + W_2, \alpha).\end{aligned}$$

□

B.2.1 $\gamma_2(W,\alpha)$ versus $\widehat{\gamma}_2(W,\alpha)$

Proof of Lemma 22. First, we prove the upper bound on $\widehat{\gamma}_2(W,\alpha)$. Recall that, by definition

$$\widehat{\gamma}_2(W,\alpha) = \min\{\gamma_2(\widehat{W},\alpha) \ : \ \widehat{W} = W + c\mathbf{1}^T, c \in \mathbb{R}^k\}.$$

In particular, if c is the zero vector, then

$$\widehat{\gamma}_2(W,\alpha) \le \gamma_2(W + c\mathbf{1}^T,\alpha) = \gamma_2(W,\alpha).$$

Now, let us prove the lower bound on $\widehat{\gamma}_2(W,\alpha)$. Consider an arbitrary vector $c \in \mathbb{R}^k$. We have

$$\begin{aligned}\gamma_2\left(W + c\mathbf{1}^T,\alpha\right) &\ge \max_{v,x} |w_{v,x} + c_v| - \alpha \\ &\ge \max_{v} |c_v| - \max_{v,x} |w_{v,x}| - \alpha \\ &\ge \max_{v} |c_v| - 1 - \alpha \\ &= \gamma_2\left(c\mathbf{1}^T\right) - 1 - \alpha.\end{aligned}$$

The first inequality uses the property that, for an arbitrary matrix M with entries $m_{v,x}$, $\gamma_2(M) \ge \max_{v,x} |m_{v,x}|$. The equality uses the fact that $\gamma_2\left(c\mathbf{1}^T\right) = \max_v |c_v|$.

Now, the subadditivity property of approximate γ_2, as given by Lemma 63, implies

$$\begin{aligned}\gamma_2(W) &\le \gamma_2(W + c\mathbf{1}^T, \alpha/2) + \gamma_2(c\mathbf{1}^T, \alpha/2) \\ &\le 2\gamma_2(W + c\mathbf{1}^T, \alpha/2) + 1 + \alpha.\end{aligned}$$

Since this holds for all choices of $c \in \mathbb{R}^k$, we obtain $\gamma_2(W) \le 2\widehat{\gamma}_2(W,\alpha/2) + 1 + \alpha$.

□

B.3 Parities

To obtain a tight lower bound on answering parities, we take advantage of mutual information upper bound Lemma 4 from [DJW18], closely related to the KL-divergence bound of Lemma 2. This result extends the unpublished result of [Ull18].

Theorem 65. *Let $d \in \mathbb{N}$, $d > 3$, and let $w \in \mathbb{N}$, $1 \le w < d$. Let $\varepsilon, \alpha > 0$ be bounded above by some arbitrary constant. Let $\mathcal{V} = \{S \subseteq [d], |S| \le w\}$ and let $Q^{parity}_{d,w}$ be the family of statistical queries over the domain $\mathcal{X} = \{-1,1\}^d$ which contains, for every $s \in \mathcal{V}$, the statistical query $q_S(x) = \prod_{j\in S} x_j$. Suppose $\mathcal{M} : \mathcal{X}^n \to \mathcal{V}$ is an ε-differentially private local protocol which satisfies, for all data sets $\overline{x} \in \mathcal{X}^n$,*

$$\forall \overline{x} \in \mathcal{X}^n, \quad \mathop{\mathbb{P}}_{\mathcal{M}}\left[q_{\mathcal{M}(\overline{x})}(\overline{x}) \ge \max_{t\in\mathcal{V}} q_t - \alpha\right] \ge \frac{2}{3}. \qquad \text{(B.1)}$$

Then,

$$n = \Omega\left(\frac{\binom{d}{w} \cdot \log\binom{d}{w}}{\varepsilon^2\alpha^2}\right).$$

Proof. To simplify our argument, we consider a symmetric variant of the problem instead. For

$S \in \mathcal{V}$, $b \in \{-1,1\}$, consider the distribution over the input space $\{-1,1\}^d$ defined by

$$\lambda_{S,b} = \alpha \cdot (\mu \mid q_S(\cdot) = b) + (1-\alpha) \cdot \mu$$

where μ is the uniform distribution over $\{-1,1\}^d$. Let $(\mathcal{S}, B)$ be drawn uniformly at random from $\mathcal{V} \times \{-1,1\}$. Conditional on $(\mathcal{S}, B) = (S, b)$, let $\overline{X} = (X_1, \ldots, X_n) \in \mathcal{X}^n$ consist of i.i.d. samples from $\lambda_{s,b}$. We show a lower bound on the number of samples required to infer $(\mathcal{S}, B)$ from $\overline{x}$ under local differential privacy. In particular, we will assume $\mathcal{M}' : \mathcal{X}^n \to \mathcal{V} \times \{-1,1\}$ is an ε-LDP protocol which satisfies

$$\mathop{\mathbb{P}}_{\mathcal{S},B,\overline{X},\mathcal{M}} \left[\mathcal{M}'(\overline{X}) = (\mathcal{S}, B)\right] \geq \frac{1}{3}. \tag{B.2}$$

Note that the protocol $\mathcal{M}$ which satisfies (B.1) allows us to obtain a protocol ε-LDP $\mathcal{M}'$ which, conditional on $B = 1$, succeeds in identifying $(\mathcal{S}, B)$ with probability 2/3. Since B is uniform on $\{-1,1\}$, this implies also that $\mathcal{M}'$ also satisfies (B.2). Thus it suffices to provide a lower bound on the number of samples required by $\mathcal{M}'$, which we proceed to do now.

By Lemma 1, the transcript of $\mathcal{M}'$ satisfies

$$I\left(\ \mathcal{T}_{\mathcal{M}'}(\overline{x})\ ;\ B, \mathcal{S}\ \right) \leq O(n\varepsilon^2) \cdot \max_{f \in \mathbb{R}^{\mathcal{X}} : \|f\|_\infty \leq 1} \mathop{\mathbb{E}}_{\mathcal{S},B} \left[\left(\mathop{\mathbb{E}}_{x \sim \lambda_{\mathcal{S},B}} [f_x] - \mathop{\mathbb{E}}_{x \sim \lambda} [f_x] \right)^2 \right] \tag{B.3}$$

where λ is the mixture of $\lambda_{s,b}$ with respect to the distribution of $(\mathcal{S}, B)$. By the symmetry of parities, namely $\mathop{\mathbb{P}}_{x \sim \mu}[q_S(x) = 1] = \mathop{\mathbb{P}}_{x \sim \mu}[q_S(x) = -1]$ for all $S \subseteq [d]$, it follows from the definition of $\lambda_{S,b}$ that $\lambda = \mu$. Thus it remains, for arbitrary $f \in \mathbb{R}^{\mathcal{X}}$ satisfying $\|f\|_\infty \leq 1$, to bound

$$\mathop{\mathbb{E}}_{\mathcal{S},B} \left[\left(\mathop{\mathbb{E}}_{X \sim \lambda_{\mathcal{S},B}} [f_X] - \mathop{\mathbb{E}}_{X \sim \mu} [f_X] \right)^2 \right].$$

Consider the Fourier transform given for $S \subseteq [d], |S| \leq w$, by

$$\hat{f}_S = \mathop{\mathbb{E}}_{x \sim \mu} [f_x \cdot q_S(x)].$$

We will take advantage of Parseval's identity which says

$$\sum_{s \subseteq [d]} \hat{f}_S^2 = \mathop{\mathbb{E}}_{X \sim \mu} \left[f_X^2\right].$$

From the definition of $\lambda_{S,b}$, we get

$$\begin{aligned}
&\mathop{\mathbb{E}}_{X\sim\lambda_{S,b}}[f_X] - \mathop{\mathbb{E}}_{X\sim\mu}[f_X] \\
&= \alpha \cdot \mathop{\mathbb{E}}_{X\sim\mu}[f_X \mid q_S(X) = b] - \alpha \cdot \mathop{\mathbb{E}}_{X\sim\mu}[f_X] \\
&= \alpha \cdot \left(\frac{1}{2} \cdot \mathop{\mathbb{E}}_{X\sim\mu}[f_X \mid q_S(X) = b] - \frac{1}{2} \cdot \mathop{\mathbb{E}}_{X\sim\mu}[f_X \mid q_S(X) = -b] \right) && \text{(B.4)} \\
&= \alpha b \cdot \left(\frac{1}{2} \cdot \mathop{\mathbb{E}}_{X\sim\mu}[f_X \cdot q_S(X) \mid q_S(X) = b] + \frac{1}{2} \cdot \mathop{\mathbb{E}}_{X\sim\mu}[f_X \cdot q_S(X) \mid q_S(X) = -b] \right) \\
&= \alpha b \cdot \left(\mathop{\mathbb{E}}_{X\sim\mu}[f_X \cdot q_S(X)] \right) && \text{(B.5)} \\
&= \alpha b \cdot \hat{f}_S
\end{aligned}$$

where (B.4) and (B.5) make use of the fact $\Pr_{x\sim\mu}[q_S(x) = b] = \Pr_{x\sim\mu}[q_S(x) = -b]$. It follows that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{S},B}\left[\left(\mathop{\mathbb{E}}_{X\sim\lambda_{\mathcal{S},B}}[f_X] - \mathop{\mathbb{E}}_{X\sim\mu}[f_X] \right)^2 \right] &= \alpha^2 \cdot \mathop{\mathbb{E}}_{\mathcal{S}}\left[\hat{f}_{\mathcal{S}}^2 \right] \\
&= \frac{\alpha^2}{\binom{d}{w}} \cdot \sum_{\substack{S\subseteq[d] \\ |S|\le w}} \hat{f}_{\mathcal{S}}^2 \\
&\le \frac{\alpha^2}{\binom{d}{w}} \cdot \mathop{\mathbb{E}}_{X\sim\mu}\left[f_X^2 \right] && \text{(B.6)} \\
&\le \frac{\alpha^2}{\binom{d}{w}}
\end{aligned}$$

since (B.6) is a consequence of Parseval's identity. Together with (B.3), this says

$$I\left(\mathcal{T}_{\mathcal{M}'}(\overline{X}) \; ; \; \mathcal{S}, B \right) \le O(n\varepsilon^2) \cdot \frac{\alpha^2}{\binom{d}{w}}. \tag{B.7}$$

The advantage of bounding $I\left(\mathcal{T}_{\mathcal{M}'}(\overline{X}) \; ; \; \mathcal{S}, B \right)$ instead of the KL-divergence quantity of Lemma 2 is that it allows us to apply Fano's inequality. In particular, for the post-processing function $\mathcal{A}'$ associated with $\mathcal{M}'$, we have

$$\mathbb{P}\left[\mathcal{A}'(\mathcal{T}_{\mathcal{M}'}(\overline{X})) \neq (\mathcal{S}, B) \right] \ge \frac{H\left((\mathcal{S}, B) \mid \mathcal{T}_{\mathcal{M}'}(\overline{X}) \right) - 1}{\log(|\mathcal{V} \times \{-1, 1\}| - 1)}.$$

We may apply the identity $I\left(\mathcal{T}_{\mathcal{M}'}(\overline{X}) \; ; \; (\mathcal{S}, B) \right) = H\left((\mathcal{S}, B) \mid - \right) H\left((\mathcal{S}, B) \mid \mathcal{T}_{\mathcal{M}'}(\overline{X}) \right)$ together with $H\left((\mathcal{S}, B) \mid \; = \right) \log(|\mathcal{V} \times \{-1, 1\}|)$ to obtain

$$\mathbb{P}\left[\mathcal{A}'(\mathcal{T}_{\mathcal{M}'}(\overline{X})) = (\mathcal{S}, B) \right] \le \frac{I\left(\mathcal{T}_{\mathcal{M}'}(\overline{X}) \; ; \; (\mathcal{S}, B) \right) + 1}{\log(|\mathcal{V} \times \{-1, 1\}| - 1)}.$$

With our mutual information bound (B.7), this gives

$$\frac{1}{3} \le \mathbb{P}\left[\mathcal{A}'(\mathcal{T}_{\mathcal{M}'}(\overline{X})) = (\mathcal{S}, B) \right] \le \frac{O(n\varepsilon^2\alpha^2) + 1}{\binom{d}{w} \cdot \log(\binom{d}{w} - 1)}.$$

Since $d \geq 3$ together with $1 \leq w < d$ implies $\binom{d}{w} \geq 3$, we obtain

$$n = \Omega\left(\frac{\binom{d}{w} \cdot \log \binom{d}{w}}{\varepsilon^2 \alpha^2}\right).$$

□

Appendix C

Characterization of agnostic learning under non-interactive LDP

C.1 Duality for $\gamma_2(W, \alpha)$ and the Dual Norm

Proof of Lemma 32. Note that $\frac{W \bullet U - \alpha \|U\|_1}{\gamma_2^*(U)}$ is scale-free, and

$$\max_{U \neq 0} \frac{W \bullet U - \alpha \|U\|_1}{\gamma_2^*(U)} = \max\{W \bullet U - \alpha \|U\|_1 : \gamma_2^*(U) = 1\}.$$

Since the set $\{U : \gamma_2^*(U) = 1\}$ is compact, the maximum is achieved. Let us then define $t = \max \frac{W \bullet U - \alpha \|U\|_1}{\gamma_2^*(U)}$ for the rest of the proof.

Let us first check that $t \leq \gamma_2(W, \alpha)$. Let $U \neq 0$, and let $\widetilde{W}$ achieve $\gamma_2(\widetilde{W}) = \gamma_2(W, \alpha)$ and $\|W - \widetilde{W}\|_{1 \to \infty} \leq \alpha$. Then

$$\begin{aligned} W \bullet U &= \widetilde{W} \bullet U + (W - \widetilde{W}) \bullet U \\ &\leq \gamma_2(\widetilde{W}) \gamma_2^*(U) + \|W - \widetilde{W}\|_{1 \to \infty} \|U\|_1 \\ &\leq \gamma_2(W, \alpha) \gamma_2^*(U) + \alpha \|U\|_1. \end{aligned}$$

The first inequality follows by the trivial case of Hölder's inequality, and the definition of γ_2^*. Rearranging shows that $t \leq \gamma_2(W, \alpha)$.

Let us now show the harder direction, $t \geq \gamma_2(W, \alpha)$. Suppose this was false, and we had $\gamma_2(W, \alpha) > t$. We will show this implies that there exists a $U \neq 0$ such that $\frac{W \bullet U - \alpha \|U\|_1}{\gamma_2^*(U)} > t$, a contradiction. Let $S = \{B \in \mathbb{R}^{k \times T} : \gamma_2(B) \leq t\}$ and $T = \{C \in \mathbb{R}^{k \times T} : \|W - C\|_{1 \to \infty} \leq \alpha\}$. Then $\gamma_2(W, \alpha) > t$ equivalently means $S \cap T = \emptyset$. Since both S and T are convex and compact, and $S \cap T = \emptyset$, the hyperplane separator theorem [Roc97, Corollary 11.4.2] implies that there is a hyperplane separating them, i.e. there is a matrix $U \in \mathbb{R}^{k \times T} \setminus \{0\}$ such that

$$\max\{B \bullet U : B \in S\} < \min\{C \bullet U : C \in T\}. \tag{C.1}$$

The left-hand side equals $t\gamma_2^*(U)$, by definition. The right-hand side equals

$$\begin{aligned}&\min\{W \bullet U - (W - C) \bullet U : C \in T\}\\ &= W \bullet U - \max\{(W - C) \bullet U : C \in T\}\\ &= W \bullet U - \max\{E \bullet U : \|E\|_{1\to\infty} \le \alpha\}\\ &= W \bullet U - \alpha\|U\|_1,\end{aligned}$$

where the last equality again uses the trivial case of Hölder's inequality. Therefore, (C.1) is equivalent to $t\gamma_2^*(U) < W \bullet U - \|U\|_1$, which is what we wanted to prove. □

C.2 Symmetrization

Proof of Lemma 34. Since the $\|\cdot\|_{2\to\infty}$ and $\|\cdot\|_{1\to2}$ norms are both non-increasing with respect to taking submatrices, the same holds also for the γ_2 norm, and, therefore, $\gamma_2(W^+) \le \gamma_2(W)$. In the reverse direction, if $R^+A^+ = W^+$ is a factorization achieving $\gamma_2(W)$, then $R^+A = W$, where A is defined by $a_{q,x} = -a_{q,-x} = a^+_{q,x}$ for any $q \in Q$, and any $x \in \mathcal{X}^+$. Clearly, $\|A\|_{1\to2} = \|A^+\|_{1\to2}$, and, therefore, the factorization R^+A certifies $\gamma_2(W) \le \|R^+\|_{2\to\infty}\|A\|_{1\to2} = \gamma_2(W^+)$. The two inequalities imply $\gamma_2(W) = \gamma_2(W^+)$.

Next we show that $\gamma_2(W^+, \alpha) \le \gamma_2(W, \alpha)$. Note that if $\widetilde{W}$ is the approximation of W that achieves $\gamma_2(W, \alpha)$, and $\widetilde{W}^+$ is the submatrix of $\widetilde{W}$ consisting of the columns indexed by $\mathcal{X}^+$, then $\|\widetilde{W}^+ - W^+\|_{1\to\infty} \le \alpha$, and, by the argument above,

$$\gamma_2(W, \alpha) \le \gamma_2(\widetilde{W}^+) \le \gamma_2(\widetilde{W}) = \gamma_2(W, \alpha).$$

To show the reverse inequality $\gamma_2(W, \alpha) \le \gamma_2(W^+, \alpha)$, take an approximation $\widetilde{W}^+$ achieving $\gamma_2(W^+, \alpha)$, and extend it to $\widetilde{W} \in \mathbb{R}^{Q\times\mathcal{X}}$ by defining $\widetilde{w}_{q,x} = -\widetilde{w}_{q,-x} = \widetilde{w}^+_{q,x}$ for all $q \in Q$ and $x \in \mathcal{X}$. Then, clearly, $\|\widetilde{W} - W\|_{1\to\infty} \le \alpha$, and, by the argument above, $\gamma_2(\widetilde{W}) = \gamma_2(\widetilde{W}^+)$. By extension, $\gamma_2(\widetilde{W}, \alpha) = \gamma_2(\widetilde{W}^+, \alpha)$ The remaining part of our claim follows from $W \bullet U = W^+ \bullet U^+$, $\|U\|_1 = \|U^+\|_1$, and $\gamma_2^*(U^+) = \gamma_2(U)$. The only one of these equalities which is non-trivial is the last one. To see why it holds, note that, first, because γ_2^* is the dual norm of γ_2, it is, indeed, a norm, and, by homogeneity and the triangle inequality,

$$\gamma_2^*(U) \le \frac{1}{2}\gamma_2^*(U^+) + \frac{1}{2}\gamma_2^*(U-) = \frac{1}{2}\gamma_2^*(U^+) + \frac{1}{2}\gamma_2^*(-U^+) = \gamma_2^*(U^+).$$

In the other direction, let $V^+ \in \mathbb{R}^{Q\times\mathcal{X}^+}$ be such that $V^+ \bullet U^+ = \gamma_2^*(U^+)$ and $\gamma_2(V^+) = 1$. Then, we can extend V^+ to $V \in \mathbb{R}^{Q\times\mathcal{X}}$ by setting $v_{q,x} = -v_{q,-x} = v^+_{q,x}$ for all $q \in Q$ and $x \in \mathcal{X}^+$. By the discussion above, $\gamma_2(V) = \gamma_2(V^+)$, and, moreover,

$$\gamma_2^*(U) \ge V \bullet U = V^+ \bullet U^+ = \gamma_2^*(U^+).$$

This completes the proof. □

C.3 Equivalence of approximate γ_2 norms of difference and concept matrices

We prove Lemma 38, the equivalence of the approximate γ_2 norm for the concept matrix $W_{\mathcal{C}}$ (Definition 28) and difference matrix $D_{\mathcal{C}}$ of $\mathcal{C}$ (Definition 37), via the following three lemmas.

Lemma 66. *Let $\mathcal{C}$ be a concept class with concept matrix $W \in \mathbb{R}^{\mathcal{C}\times\mathcal{U}}$ and difference matrix $D_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}^2\times\mathcal{U}}$. Then $\gamma_2(D_{\mathcal{C}}, \alpha) \le \gamma_2(W, \alpha)$.*

Lemma 67. *Let $\mathcal{C}$ be a concept class closed under negation. Let $W \in \mathcal{C}\times\mathcal{U}$ be its concept matrix and let $D_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}^2\times\mathcal{U}}$ be its difference matrix. Then $\gamma_2(W, \alpha) \le \gamma_2(D_{\mathcal{C}}, \alpha)$.*

Lemma 68. *Let $\mathcal{C}$ be a concept class. Let $W \in \mathbb{R}^{\mathcal{C}\times\mathcal{U}}$ be its concept matrix and let $D_{\mathcal{C}} \in \mathbb{R}^{\mathcal{C}^2\times\mathcal{U}}$ be its difference matrix. Then $\gamma_2(W, \alpha) \le 2\gamma_2(D_{\mathcal{C}}, \alpha/2) + 1$.*

Proof of Lemma 66. Let $\widetilde{W} \in \mathcal{C}\times\mathcal{U}$ witness $\gamma_2(W,\alpha)$ so that $\|W - \widetilde{W}\|_{1\to\infty} \le \alpha$ and $\gamma_2(W,\alpha) = \gamma_2(\widetilde{W})$.

Let $W' \in \mathbb{R}^{\mathcal{C}^2\times\mathcal{U}}$ be the matrix with entries $w'_{(c,c'),a} = \widetilde{w}_{c,a}$. Similarly, let $W'' \in \mathbb{R}^{\mathcal{C}^2\times\mathcal{U}}$ be the matrix with entries $w''_{(c,c'),a} = \widetilde{w}_{c',a}$. Since W' and W'' are obtained from $\widetilde{W}$ by duplicating rows,

$$\gamma_2(W,\alpha) = \gamma_2(\widetilde{W}) = \gamma_2(W') = \gamma_2(W'').$$

Now consider the matrix $\widetilde{D_{\mathcal{C}}} = \frac{1}{2}(W' - W'')$. By subadditivity and scaling properties,

$$\gamma_2(\widetilde{D_{\mathcal{C}}}) \le \frac{1}{2}(\gamma_2(W') + \gamma_2(W'')) = \gamma_2(\widetilde{W}).$$

Moreover, for all $c, c' \in \mathcal{C}$, $a \in \mathcal{U}$, the entry $\widetilde{d}_{(c,c'),a}$ of $\widetilde{D_{\mathcal{C}}}$ approximates entry $d_{(c,c'),a}$ of $D_{\mathcal{C}}$. Specifically,

$$\begin{aligned}\left|\widetilde{d}_{(c,c'),a} - d_{(c,c'),a}\right| &= \left|\frac{\widetilde{w}_{c,a} - \widetilde{w}_{c',a}}{2} - \frac{c(a) - c'(a)}{2}\right| \\ &\le \left|\frac{\widetilde{w}_{c,a} - c(a)}{2}\right| + \left|\frac{\widetilde{w}_{c',a} - c'(a)}{2}\right| \\ &\le \alpha.\end{aligned}$$

Hence $\|D_{\mathcal{C}} - \widetilde{D_{\mathcal{C}}}\|_{1\to\infty} \le \alpha$. Together with $\gamma_2(\widetilde{D_{\mathcal{C}}}) \le \gamma_2(\widetilde{W})$, this implies

$$\gamma_2(D_{\mathcal{C}},\alpha) \le \gamma_2(\widetilde{D_{\mathcal{C}}}) \le \gamma_2(\widetilde{W}) = \gamma_2(W,\alpha).$$ □

Proof of Lemma 67. Row c of W is identical with row $(c, -c)$ of $D_{\mathcal{C}}$. Hence, W is obtained from $D_{\mathcal{C}}$ by deleting some of its rows. Since the γ_2 norm is non-increasing under taking submatrices, it follows that $\gamma_2(W,\alpha) \le \gamma_2(D_{\mathcal{C}},\alpha)$. □

Proof of Lemma 68. Fix an arbitrary concept $c' \in \mathcal{C}$. Let $D'_{\mathcal{C}}$ be the submatrix of $D_{\mathcal{C}}$ which includes row (c, c') of $D_{\mathcal{C}}$ for each $c \in \mathcal{C}$. Then $D'_{\mathcal{C}} = \frac{1}{2}\left(W - \mathbf{1}(c')^T\right)$ where $\mathbf{1}$ is the all-ones vector of dimension $|\mathcal{C}|$, and we identify c' with a vector in $\mathbb{R}^{\mathcal{U}}$. Expressing our concept matrix as $W =$

$2D'_{\mathcal{C}} + \mathbf{1}(c')^T$, we may apply the subadditivity properties of the approximate γ_2 norm, as given by Lemma 63, to obtain

$$\begin{aligned}\gamma_2(W, \alpha) &= \gamma_2(2D'_{\mathcal{C}} + \mathbf{1}(c')^T, \alpha)\\ &\le \gamma_2(2D'_{\mathcal{C}}, \alpha) + \gamma_2(\mathbf{1}(c')^T, 0)\\ &\le 2\gamma_2(D'_{\mathcal{C}}, \alpha/2) + 1. \qquad \square\end{aligned}$$

C.4 Total variation lower bound

Proof of Lemma 41. The main observation is that, since λ and μ share the same marginal on $\mathcal{U}$ but the labels are given by the functions s and $-s$, for any hypothesis $h : \mathcal{C} \to \{\pm 1\}$ we have $L_\lambda(h) + L_\mu(h) = 1$. Therefore,

$$\begin{aligned}(L_\lambda(h) - L_\lambda(c')) + (L_\mu(h) - L_\mu(c)) &= ((L_\lambda(h) + L_\mu(h)) - (1 - L_\mu(c')) - L_\mu(c)\\ &= L_\mu(c') - L_\mu(c) > \alpha.\end{aligned}$$

This implies that if $L_\lambda(h) - L_\lambda(c') \le \frac{\alpha}{4}$, then $L_\mu(h) - L_\mu(c) > \frac{3\alpha}{4}$, as required.

Suppose now that $\mathcal{M}$ $(\alpha/4, \beta)$-learns $\mathcal{C}$ agnostically with n samples. Let $A \subseteq \{\pm 1\}^{\mathcal{U}}$ be the set of hypotheses with loss at most $L_\lambda(c') + \alpha/4$ on λ^n. As we just showed, every hypothesis in A has loss larger than $L_\mu(c) + 3\alpha/4$ under μ. Since

$$\min_{c'' \in \mathcal{C}} L_\lambda(c'') \le L_\lambda(c'), \qquad \min_{c'' \in \mathcal{C}} L_\mu(c'') \le L_\mu(c),$$

it follows from the definition of agnostic learning that $\mathbb{P}[\mathcal{M}(\lambda^n) \in A] \ge 1 - \beta$, and $\mathbb{P}[\mathcal{M}(\lambda^n) \in A] \le \beta$. Then, by the definition of total variation,

$$d_{\mathrm{TV}}(\mathcal{M}(\lambda^n), \mathcal{M}(\mu^n)) \ge \mathbb{P}[\mathcal{M}(\lambda^n) \in A] - \mathbb{P}[\mathcal{M}(\mu^n) \in A] \ge 1 - 2\beta,$$

completing the proof of the lemma. □

Bibliography

[AB10] Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *J. Mach. Learn. Res.*, 11:2785–2836, 2010.

[AJM20] Kareem Amin, Matthew Joseph, and Jieming Mao. Pan-private uniformity testing. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 183–218. PMLR, 2020.

[App17] Apple Differential Privacy Team. Learning with privacy at scale. Apple Machine Learning Journal, 2017.

[BBNS19] Jaroslaw Blasiok, Mark Bun, Aleksandar Nikolov, and Thomas Steinke. Towards instance-optimal private query release. In *SODA*, pages 2480–2497. SIAM, 2019.

[BCD$^+$07] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the 26th ACM Symposium on Principles of Database Systems*, PODS '07, pages 273–282. ACM, 2007.

[BDKT12] Aditya Bhaskara, Daniel Dadush, Ravishankar Krishnaswamy, and Kunal Talwar. Unconditional differentially private mechanisms for linear queries. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing*, STOC '12, pages 1269–1284, 2012.

[BEM$^+$17] Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th Symposium on Operating Systems Principles*, SOSP '17, pages 441–459. ACM, 2017.

[BJKS04] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68(4):702–732, 2004.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[BNS18] Mark Bun, Jelani Nelson, and Uri Stemmer. Heavy hitters and the structure of local privacy. In *Proceedings of the 37th ACM Symposium on Principles of Database Systems*, PODS'18, pages 435–447. ACM, 2018.

[BS15] Raef Bassily and Adam D. Smith. Local, private, efficient protocols for succinct histograms. In Rocco A. Servedio and Ronitt Rubinfeld, editors, *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 127–135. ACM, 2015.

[BUV14] Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *46th Annual ACM Symposium on the Theory of Computing*, STOC '14, pages 1–10, New York, NY, USA, 2014.

[CSS11] T-H Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Transactions on Information and System Security (TISSEC)*, 14(3):26, 2011.

[CSU$^+$19] Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In *Proceedings of the 38th Annual Conference on the Theory and Applications of Cryptographic Techniques*, EUROCRYPT '19, 2019.

[CT19] Ioannis Caragiannis and Evanthia Tsitsoka. Deanonymizing social networks using structural information. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 1213–1219. ijcai.org, 2019.

[CTUW14] Karthekeyan Chandrasekaran, Justin Thaler, Jonathan Ullman, and Andrew Wan. Faster private release of marginals on small databases. In *Proceedings of the 5th ACM Conference on Innovations in Theoretical Computer Science*, ITCS '14, pages 287–402, Princeton, NJ, 2014. ACM.

[CU21] Albert Cheu and Jonathan R. Ullman. The limits of pan privacy and shuffle privacy for learning and estimation. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 1081–1094. ACM, 2021.

[DF19] Amit Daniely and Vitaly Feldman. Locally private learning without interaction requires separation. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14975–14986, 2019.

[DF20] Yuval Dagan and Vitaly Feldman. Interaction is necessary for distributed learning with privacy or communication constraints. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 450–462. ACM, 2020.

[DJW13] John Duchi, Michael Jordan, and Martin Wainwright. Local privacy and statistical minimax rates. In *IEEE 57th Annual Symposium on Foundations of Computer Science*, FOCS '13, pages 429–438, 2013.

[DJW18] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation. *J. Amer. Statist. Assoc.*, 113(521):182–201, 2018.

[DLS$^+$17] Aref N. Dajani, Amy D. Lauger, Phyllis E. Singer, Daniel Kifer, Jerome P. Reiter, Ashwin Machanavajjhala, Simson L. Garfinkel, Scot A. Dahl, Matthew Graham, Vishesh Karwa, Hang Kim, Philip Lelerc, Ian M. Schmutte, William N. Sexton, Lars Vilhuber, and John M. Abowd. The modernization of statistical disclosure limitation at the U.S. census bureau, 2017. Presented at the September 2017 meeting of the Census Scientific Advisory Committee.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM Symposium on Principles of Database Systems*, PODS '03, pages 202–210. ACM, 2003.

[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.

[DNT15] Cynthia Dwork, Aleksandar Nikolov, and Kunal Talwar. Efficient algorithms for privately releasing marginals via convex relaxations. *Discrete & Computational Geometry*, 53(3):650–673, 2015.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[DR18] John C. Duchi and Feng Ruan. The right complexity measure in locally private estimation: It is not the fisher information. *CoRR*, abs/1806.05756, 2018.

[EGS03a] Alexandre V. Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In Frank Neven, Catriel Beeri, and Tova Milo, editors, *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA*, pages 211–222. ACM, 2003.

[EGS03b] Alexandre V. Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS*, pages 211–222. ACM, 2003.

[ENP22] Alexander Edmonds, Aleksandar Nikolov, and Toniann Pitassi. Learning versus refutation in noninteractive local differential privacy. *CoRR*, abs/2210.15439, 2022.

[ENU19] Alexander Edmonds, Aleksandar Nikolov, and Jonathan Ullman. The power of factorization mechanisms in local and central differential privacy. *CoRR*, abs/1911.08339, 2019.

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security*, CCS'14. ACM, 2014.

[Fel17] Vitaly Feldman. A general characterization of the statistical query complexity. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 785–830. PMLR, 2017.

[FSSS03] Jürgen Forster, Niels Schmitt, Hans Ulrich Simon, and Thorsten Suttorp. Estimating the optimal margins of embeddings in euclidean half spaces. *Machine Learning*, 51(3):263–281, 2003.

[GHRU11] Anupam Gupta, Moritz Hardt, Aaron Roth, and Jonathan Ullman. Privately releasing conjunctions and the statistical query barrier. In *Proceedings of the 43rd ACM Symposium on Theory of Computing*, STOC '11, pages 803–812, San Jose, CA, 2011.

[Gro53] A. Grothendieck. Résumé de la théorie métrique des produits tensoriels topologiques. *Bol. Soc. Mat. São Paulo*, 8:1–79, 1953.

[GS02] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[HRS12] Moritz Hardt, Guy N. Rothblum, and Rocco A. Servedio. Private data release via learning thresholds. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 168–187, 2012.

[HT10] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC*, 2010.

[JMNR19] Matthew Joseph, Jieming Mao, Seth Neel, and Aaron Roth. The role of interactivity in local differential privacy. In David Zuckerman, editor, *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, pages 94–105. IEEE Computer Society, 2019.

[JMR19] Matthew Joseph, Jieming Mao, and Aaron Roth. Exponential separations in local differential privacy through communication complexity. *CoRR*, abs/1907.00813, 2019.

[JNS18] Noah Johnson, Joseph P Near, and Dawn Song. Towards practical differential privacy for sql queries. *Proceedings of the VLDB Endowment*, 11(5):526–539, 2018.

[Kea93] Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. In *STOC*, pages 392–401. ACM, May 16-18 1993.

[KL18] Pravesh K. Kothari and Roi Livni. Improper learning by refuting. In Anna R. Karlin, editor, *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, volume 94 of *LIPIcs*, pages 55:1–55:10. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

[KLN$^+$08] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? In *FOCS*, pages 531–540. IEEE, Oct 25–28 2008.

[KLN$^+$11] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, 2011.

[KN12] Subhash Khot and Assaf Naor. Grothendieck-type inequalities in combinatorial optimization. *Comm. Pure Appl. Math.*, 65(7):992–1035, 2012.

[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 '10, pages 775–784. ACM, 2010.

[LHR$^+$10] Chao Li, Michael Hay, Vibhor Rastogi, Gerome Miklau, and Andrew McGregor. Optimizing linear counting queries under differential privacy. In *Proceedings of the 29th ACM Symposium on Principles of Database Systems*, PODS'10, pages 123–134. ACM, 2010.

[LMSS07] Nati Linial, Shahar Mendelson, Gideon Schechtman, and Adi Shraibman. Complexity measures of sign matrices. *Combinatorica*, 27(4):439–463, 2007.

[LS09] Nati Linial and Adi Shraibman. Lower bounds in communication complexity based on factorization norms. *Random Struct. Algorithms*, 34(3):368–394, 2009.

[MMHM18] Ryan McKenna, Gerome Miklau, Michael Hay, and Ashwin Machanavajjhala. Optimizing error of high-dimensional statistical queries under differential privacy. *Proceedings of the VLDB Endowment*, 11(10):1206–1219, 2018.

[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. IEEE Computer Society, 2010.

[MS20a] Eran Malach and Shai Shalev-Shwartz. When hardness of approximation meets hardness of learning. *CoRR*, abs/2008.08059, 2020.

[MS20b] Eran Malach and Shai Shalev-Shwartz. When hardness of approximation meets hardness of learning. *CoRR*, abs/2008.08059, 2020.

[Nik15] Aleksandar Nikolov. An improved private mechanism for small databases. In *Automata, Languages, and Programming - 42nd International Colloquium, ICALP*, pages 1010–1021, 2015.

[NS94] Noam Nisan and Mario Szegedy. On the degree of boolean functions as real polynomials. *Computational Complexity*, 4:301–313, 1994.

[NTZ16] Aleksandar Nikolov, Kunal Talwar, and Li Zhang. The geometry of differential privacy: The small database and approximate cases. *SIAM J. Comput.*, 45(2):575–616, 2016.

[Pis12] Gilles Pisier. Grothendieck's theorem, past and present. *Bull. Amer. Math. Soc. (N.S.)*, 49(2):237–323, 2012.

[Roc97] R. Tyrrell Rockafellar. *Convex analysis.* Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.

[She11] Alexander A. Sherstov. The pattern matrix method. *SIAM J. Comput.*, 40(6):1969–2000, 2011.

[SZ20] Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 3437–3452. PMLR, 2020.

[TUV12] Justin Thaler, Jonathan Ullman, and Salil P. Vadhan. Faster algorithms for privately releasing marginals. In *39th International Colloquium on Automata, Languages, and Programming* -, ICALP '12, pages 810–821, Warwick, UK, 2012. Springer.

[Ull18] Jonathan Ullman. Tight lower bounds for locally differentially private selection. *CoRR*, abs/1802.02638, 2018.

[Vad17] Salil P. Vadhan. On learning vs. refutation. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 1835–1848. PMLR, 2017.

[Ver18] Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.

[WZL$^+$19] Royce J Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson. Differentially private sql with bounded user contribution. *arXiv preprint arXiv:1909.01917*, 2019.