Ben-Gurion University of the Negev
The Faculty of Natural Sciences
The Department of **Computer Science**

A Study of Privacy and Compression in Learning Theory

Menachem Sadigurschi

Thesis submitted in partial fulfillment
of the requirements for the degree of
"DOCTOR OF PHILOSOPHY"

Under the supervision of **Prof. Aryeh Kontorovich** and **Dr. Uri Stemmer**

March 23st, 2023

To my beloved wife and children, thank you for your unwavering support and motivation. Your love and presence in my life have made this period not only bearable but also beautiful.

To my advisors, Prof. Aryeh Kontorovich and Dr. Uri Stemmer, for their invaluable guidance, mentorship, and expertise. Thank you for challenging me to be the best researcher I can be and for sharing your knowledge and insights with me.

Finally, I would like to dedicate this thesis to my grandfathers, Shmuel (Smil) Sadigurschi and Simpson Shimon Kalmus, who sacrificed a lot for their family. Your ability to see the hidden beauty in the world and to rejoice in it has always been an inspiration to me. Guardarei voçês na minha memória para sempre.

Abstract

Machine learning is a rapidly growing field of computer science that has the potential to revolutionize various industries. With the ability to process vast amounts of data and improve performance over time, it has become a powerful tool for solving complex problems and driving innovation, and is being used in a wide range of industry applications.

In this thesis, we examine two interrelated subjects within the realm of machine learning - privacy and compression. The explosion of data generation and usage in recent years has brought about unprecedented opportunities for machine learning and AI-based applications. However, as data becomes increasingly personal and sensitive, the need for protecting individual privacy has become paramount. Differential privacy, a mathematical framework for quantifying the privacy of the computational process of an algorithm with respect to its input, has emerged as a leading technique for addressing this challenge. If we wish to allow the continuation of innovation and progress in the field, and perhaps even expand it, users must be assured that the use of the databases will not allow the privacy of any individual to be compromised. Another aspect whose importance is becoming more apparent, as databases increase, is the issue of compression. In the last few years, we have witnessed the rise of models based on huge amounts of information. Whether these are large natural language models with billions of parameters, visual analysis algorithms or image generators trained using terabytes of images from all over the web, the size of the systems and databases is starting to pose a problem. First, it creates enormous challenges in the computational and engineering aspects involved in training such models. Nowadays, it is evident that there are many tasks in which it is impossible for anyone other than giant technology companies to make progress, since only they can carry out computations of this size. One idea that can offer the possibility of a change in this paradigm is compression. The basis of this old idea is that patterns in the information, such as those identified by machine learning tools, could allow to compress the data into a relatively small number of records that contain all the knowledge needed for the labeling pattern. The same logic applies in the other direction as well. If a small number of such records can be identified, then it is possible to leverage this identification process to produce learning processes. This idea arose naturally over the years in the development of popular algorithms such as the Support Vector Machine

and the Condensed Nearest Neighbor. Although the idea of compression is a significant tool in the toolbox of research and development in the field of machine learning, the exact relationship between the concept of compression and learning includes a number of fundamental unsolved questions. This dissertation investigates the intersection of these three areas by addressing four research questions that aim to deepen our understanding of the connections between privacy, compression, and machine learning. The first topic, that will be the basis of this thesis, stems from the following question:

> **Question:** *To what extent are machine learning and compression conceptually intertwined, both qualitatively and quantitatively?*

The two concepts are known to be highly connected, and for some settings even equivalent. First, we investigate whether this equivalence can be extended to more fundamental cases and, specifically, to real-valued functions also known as regression problems. To tackle the challenge we start by constructing an efficient method to convert any learning algorithm to a compression scheme. We then extend this technique from the basic case of classification, meaning learning binary functions, to the broader one of regression. Thus, we obtain the first general compressed regression result, guaranteeing that the information lost is arbitrarily small.

The field of differential privacy has experienced a boom in recent years, but at the same time, there are fundamental tasks whose understanding is incomplete. In the second part of the dissertation, we explore the line of work related to the question:

> **Question:** *How much data is needed in order to learn from data, while guaranteeing that privacy is not violated?*

This quantification of the data size needed is referred to as the sample complexity of the problem. We examine one such task - learning axis-aligned rectangles. It is known that the dependency on the dimension must be at least linear, but prior works attaining such optimal dependency required the sample complexity to grow logarithmically in the space size. We present a novel algorithm that achieves both, as the data it requires scales linearly in the dimension and asymptotically smaller than the log of the space size. The technique used in order to attain this improvement involves the deletion of "exposed" data-points sequentially, so that the influence of each individual on the final hypothesis is limited inherently in the algorithm design.

In the third part of the dissertation, we investigate the very definition of private learning. The standard definition of learning is aimed at providing accuracy guarantees, under the assumption that the underlying distribution of the data is the worst possible each time. There is a growing voice in the research advocating that this definition is too pessimistic, i.e. it doesn't reflect the actual properties of real-life data. This worst-case paradigm is blamed for being part of the well-known gap between theory and practice in machine learning. Moreover, this pessimistic prospect gets amplified under the privacy requirement, e.g. there are fundamental problems which are simply impossible to learn under the privacy constraint, in contrast to an easy solution without restrictions. We join this line of work, advocate the use of a more flexible model called “Universal Learning”, and investigate its advantages over the classical model. Finally, in the last part of this thesis, we move on to deal with adaptive data analysis. The research in this field revolves around the attempt to produce analysis tools in a formal model for learning that aims to be closer to the process that takes place in laboratories and in various practices of data analysis. Instead of a static model of analysis where one question is asked and one answer received, in the adaptive model many research and statistical queries are asked in such a way that each query arises from the information accumulated during the prior analytical process. In such a model, which is very close to a realistic process, statistical constructions from the usual static model are not valid. Research in this area combines ideas from the privacy and compression literature, since both can be used for designing reliable algorithms under adaptive models. Under this complex yet important setting, we explore the problem of extending the tools and results from the adaptive data analysis literature to the setting of correlated examples. We provide results both for privacy-based and compression-based tools.

Overall, this dissertation aims to deepen the understanding of the intersection between differential privacy, machine learning and compression schemes, by addressing these research questions. By doing so, we hope to contribute to the development of more efficient and privacy-preserving machine learning algorithms.

Contents

List of Algorithms

Chapter 1

Introduction

In the last decades we have experienced continuous growth in all areas of research in computer science. Many tools and algorithms have been developed and are still being developed to analyze inputs of different types and to understand what is possible and what is impossible in each and every field. At the same time, the world of static research is also experiencing a significant boom and concepts such as "statistical significance", "correlation" and "linear regression" have become a must in the toolbox of natural and social science researchers. Another new field in which interest has grown over the years is "signal processing", the research field that integrates the two worlds of static and dynamic research. In this field, the goal is to analyze a signal, which is a function of time, and to understand its properties and characteristics. The signal can be of any type, such as a sound signal, a video signal, a signal that represents the temperature in a given area, and more. With the development of wide range of communication technologies the need for signal processing has become more and more important. On this background the field of "machine learning" has developed.

Machine learning is a field of computer science that deals with the development of algorithms that can learn from data and make predictions on new data. The main goal is to try and leverage information in order to produce systems and software capable of performing diverse tasks. This field has experienced tremendous growth in recent years. Systems that grow out of machine learning find many and quite varied uses: from analyzing medical tests, to chatbots and automatic art generators, to autonomous vehicles. Most of the involvement in the field, both applied and research, revolves around the construction of sophisticated models and new tools that will allow the industry to continue fast-forward toward more goals and peaks. At the same time, there is great interest in understanding the limitations of the core concepts and possibilities inherent in each tool and in each situation. To this end, we must create a precise mathematical system that will define the situations and challenges we face and enable their systematic and meticulous research.

The field of learning theory is a fundamental pillar of machine learning, providing a rigorous framework for understanding the mechanisms of learning algorithms and their ability to acquire knowledge from data. To fully explore the concepts and challenges of learning theory, it is essential to establish a solid theoretical foundation. Key concepts include hypothesis spaces, empirical risk minimization, and the bias-variance trade-off. Theoretical properties such as convergence guarantees, sample complexity, and computational complexity provide insights into the performance and limitations of learning algorithms.

In the context of learning theory, the notion of compression has gained considerable attention. Compression refers to the ability to succinctly represent and describe patterns or regularities in data. A system that exhibits compression can convey its results more concisely than merely detailing the results themselves. The connection between compression and learning has been established through theories and results developed in statistical learning.

By considering these various aspects, including the foundations, challenges, and connections to compression, we can gain a deeper understanding of the limitations and possibilities of learning algorithms. This understanding forms the basis for our investigation into the specific questions and goals outlined in this thesis.

First, the question *"What is learning?"* must be answered - or, in a more detailed way, *"What requirements must an algorithm meet in order to be considered a learning algorithm?"*. Intuitively, the idea is that the algorithm should improve as more information is given. However, the definition of "improvement", situations in which the algorithm is expected to operate - plus other significant issues - needs to be well-defined. Once an appropriate definition is chosen, the next critical question is *"Which problems or situations are learnable and which are not?"*. Naturally, this question is not easy to answer, and many tools are required to shed light on this challenge.

Addressing the question of learnability necessitates the development of rigorous mathematical frameworks. One such framework is based on the notion of uniform convergence. The principle of uniform convergence asserts that as the size of the training dataset grows, the algorithm's performance on unseen data should converge to its expected performance. In other words, the algorithm should generalize well to new instances beyond the training set. The study of uniform convergence provides insights into the trade-offs between the complexity of a learning algorithm, the size of the training dataset, and the algorithm's ability to generalize accurately.

Another approach to the learnability question is to formulate learning problems as optimization problems. Convex optimization provides a powerful mathematical framework for solving optimization problems where the objective function and constraints exhibit convexity. By casting learning problems as optimization problems, researchers can leverage this rich mathematical theory and develop learning algorithms, allowing for efficient computation and convergence guarantees to globally optimal solutions. The interplay between convex optimization and machine learning has led through the years to significant advancements in developing robust and scalable learning algorithms.

However, even with tools such as those of uniform convergence and convex optimization at our disposal, determining the learnability of a given problem remains a formidable challenge. It is often influenced by the inherent complexity of the problem domain, the quality and quantity of available data, and the expressiveness of the learning algorithm. This challenge has prompted researchers to explore additional avenues for understanding the limits and capabilities of machine learning systems.

One such avenue is rooted in the principle of Occam's Razor, one of the most famous and classic principles in theories dealing with learning and drawing conclusions based on information. This principle can be found in Aristotle's writings from two thousand years ago, stating: *"We may assume the superiority ceteris paribus of the demonstration which derives from fewer postulates or hypotheses"* (Aristotle, 1995). This idea was later popularized by William of Ockham who phrased it as *"Plurality should not be posited without necessity"* (Duignan, 1998), defines a theory or pattern deduced from data as "good" if it is simple. If we wish to measure simplicity in a more precise manner, a possible option is explanation length, i.e. the shorter, the better. In the language of computer science, we could say that a system that has a pattern or regularity is one whose results can be compressed and described in a shorter way than detailing the results themselves. From this idea, and in light of theories and results developed in the field of statistical learning, it has become evident that this type of compression is deeply connected to learning.

Yet, in order to study this connection, the notions of learning and of compression must be properly defined. Since the foundation of the learning theory field, several notions of learning were proposed in an attempt to capture the characteristics of learning. One of the main notions at the core of learning theory research, is that of Probably Approximately Correct (PAC) learning. We can informally describe a PAC-learning algorithm as acting on given data and outputting a hypothesis which will accurately predict, with high probability, the label of almost any newly sampled data point.

One of the main problems in learning theory is characterizing *sample complexity*, which is the amount of data required in order to guarantee PAC-learning for a given class of functions. It is known that the sample complexity of learning a class of binary functions is proportional to its VC dimension (which we will define at Definition 3.1.5). For classes of real-valued functions, an analogous result was proven using the notion of *Fat-Shattering* dimension (Alon et al. (1997)). Nevertheless, various other notions of "learnability" have been found beneficial and insightful. One of them is, indeed, the idea of compression.

1.1 Compression

As progressively more novel learning algorithms have been designed, one of the common aspects of note is that at the core lies a particular kind of data labeling compression: the principle of finding "representative" subsets of the data as part of a more general *Occam learning* paradigm. Most notable is the SVM algorithm, which derives its name from the set of supporting vectors that uniquely defines the linear separator returned by the algorithm (Cortes and Vapnik, 1995).

Following this path, Littlestone and Warmuth (1986) established a formal framework for discussion of *sample compression schemes* from the learning point of view. In addition, they showed that for the case of binary-labeled classes compression implies learnability [1].

A fundamental question posed by Littlestone and Warmuth (1986) in the same paper concerns the reverse implication: Can every learner be converted into a sample compression scheme? Or, in a more quantitative formulation: Does every learnable class admit a constant-size sample compression scheme? A series of partial results (Floyd (1989); Helmbold et al. (1992); Floyd and Warmuth (1995); Ben-David and Litman (1998); Kuzmin and Warmuth (2007);Rubinstein et al. (2009); Rubinstein and Rubinstein (2012); Chernikov and Simon (2013); Livni and Simon (2013); Moran et al. (2017)) culminated in Moran and Yehudayoff (2016), which resolved the latter question. [2]

The usefulness of this link is that, while learning is a statistical notion, compression is a combinatorial one. Thus, linking the two by such an equivalence could help move questions about learning to the combinatorial world, opening the research to other directions and to a wide range of tools previously not relevant to this area.

In the same way, the connection between compression and learnability can be investigated in various settings and regimes. In recent years, it has been proven to be an extremely useful tool for constructing learning algorithms for scenarios far from the classical PAC model, such as adversarial learning (Montasser et al., 2019a) and parametric distribution learning Ashtiani et al. (2020). But, at the same time, the connection in the other direction - converting learning algorithms and learnable problems into compressing schemes - was left almost untouched. Since the main binary case had been in the center of interest for so many years, and as this very setting had an equivalent open problem in category theory, natural extensions and variations had been almost not studied at all. Moreover, the tools used in order to attack and eventually solve the conjecture seem to rely on the binary nature of the problem. This leads to our starting point for this part of the thesis,

[1] Lately there is growing interest in the properties and the generalization bounds of compressing-based learning algorithms, see for example Gottlieb et al. (2016); Graepel et al. (2005); Cummings et al. (2016)

[2] The refined conjecture of Littlestone and Warmuth (1986), that any concept class $\mathcal{C}$ with VC-dimension $d_{\mathcal{C}}$ admits a compression scheme of size $\mathcal{O}\left(d_{\mathcal{C}}\right)$, remains open.

which is the following:

Question 1. Are learning and compression equivalent definitions also for regression problems?

Our Contribution

Our first contribution was to extend Moran and Yehudayoff's fundamental result, relating compression and learning to the case of real-valued function classes. We begin with an algorithmically efficient version of the learner-to-compression scheme conversion in Moran and Yehudayoff (2016). Namely, our compression scheme size is linear in the dimension and the dual-dimension of the class. Furthermore, the compression scheme is computable in linear time in the initial sample size. More formally:

Theorem 1.1.1 (Efficient compression for classification, informal)**.** *Let $\mathcal{C}$ be a concept class over some instance space $\mathcal{X}$ with VC-dimension $d_{\mathcal{C}}$, dual VC-dimension $d_{\mathcal{C}}^*$, and suppose that $\mathcal{A}$ is a (proper, consistent) PAC-learner for $\mathcal{C}$. There is a randomized sample compression scheme for $\mathcal{C}$ of size $O(k \log k)$, where $k = O(d_{\mathcal{C}} d_{\mathcal{C}}^*)$. Furthermore, on a sample of any size m, the compression set may be computed in expected time $\mathcal{O}(m \log m)$.* [3]

For comparison, a naive implementation of the Moran and Yehudayoff (2016) existence proof yields a runtime of order $m^{cd_{\mathcal{C}}} + m^{cd_{\mathcal{C}}^*}$ (for some universal constants c, c'), which can be doubly exponential when $d_{\mathcal{C}}^* = 2^{d}_{\mathcal{C}}$; this is without taking into account the cost of computing the minimax distribution on the $m^{cd_{\mathcal{C}}} \times m$ game matrix.

Next, we extend the result in Theorem 1.1.1 from classification to regression. We provide an efficient compression scheme for real-valued functions, which is computable in linear time in the initial sample size. The size of the compression is linear in the fat-shattering dimension and the dual-dimension of the class. More formally:

Theorem 1.1.2 (Efficient compression for regression, informal)**.** *Let $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ be a function class with t-fat-shattering dimension $Fat_t(\mathcal{F})$, dual t-fat-shattering dimension $Fat_t^*(\mathcal{F})$, and suppose that $\mathcal{A}$ is an ERM (i.e., proper, almost consistent) learner for $\mathcal{F}$. There is a randomized uniformly ε-approximate sample compression scheme for $\mathcal{F}$ of size $\mathcal{O}(k\tilde{m}\log(k\tilde{m}))$, where $\tilde{m} = \mathcal{O}\big(Fat_{c\varepsilon}(\mathcal{F})\log(1/\varepsilon)\big)$ and $k = \mathcal{O}\big(Fat_{c\varepsilon}^*(\mathcal{F})\log(Fat_{c\varepsilon}^*(\mathcal{F})/\varepsilon)\big)$. Furthermore, on a sample of any size m, the compression set may be computed in expected time $\mathcal{O}(m\log(m)+k)$.* [4]

A key component in the above result is our construction of a generic weak-learner. We use the definition of a weak-learner from Simon (1997), which is a different notion than

[3] For clarity, the linear dependency between the runtime of our algorithm and algorithm $\mathcal{A}$ has been omitted.

[4] As in Theorem1.1.1, the linear dependency between the run-time of our algorithm and algorithm $\mathcal{A}$ has been omitted for clarity.

the standard notion of a weak-learner. While the standard notion is defined in therms of average error, Simon's definition is such that the bound on the error is required to be bounded for most of the space. This is a stronger requirement, and is necessary for our construction. Using Simon's definition, we show that a weak-learner can be constructed for every function class with bounded fat-shattering dimension.

Definition 1.1.3. *For $\eta \in [0,1]$ and $\gamma \in [0,1/2]$, we say that $f : \mathcal{X} \to \mathbb{R}$ is an (η,γ)-*weak hypothesis *(with respect to distribution D and target $f^* \in \mathcal{F}$) if*

$$\Pr_{X \sim D}(|f(X) - f^*(X)| > \eta) \le \frac{1}{2} - \gamma.$$

Theorem 1.1.4 (Generic weak learner)**.** *Let $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ be a function class with t-fat-shattering dimension $Fat_t(\mathcal{F})$. For some universal numerical constants $c_1, c_2, c_3 \in (0,\infty)$, for any $\eta, \delta \in (0,1)$ and $\gamma \in (0,1/4)$, any $f^* \in \mathcal{F}$, and any distribution D, letting $X_1, \dots, X_m$ be drawn iid from D, where*

$$m = \left\lceil c_1 \left(Fat_{c_2\eta}(\mathcal{F}) \ln\left(\frac{c_3}{\eta}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right\rceil,$$

with probability at least $1-\delta$, every $f \in \mathcal{F}$ with $\max_{i\in[m]} |f(X_i) - f^*(X_i)| \le \alpha\eta$ for $\alpha \in [0,1)$, is an (η,γ)-weak hypothesis with respect to D and f^*.

As one can see, our results allow us to use any hypothesis $f \in \mathcal{F}$ with $\max_{i\in[m]} |f(X_i) - f^*(X_i)|$ bounded below η: for instance, bounded by $\eta/2$.

This result sheds new light on an open question of Simon (1997). Moreover, the ideas used in our construction proved fruitful in new settings, robust learning Montasser et al. (2019b). In order to demonstrate the efficacy of the above results, we show applications to two regression problems: learning Lipschitz and bounded-variation functions.

Another direction of learning theory and its characteristics which has emerged in the last two decades is trustworthy machine learning. This includes various aspects and implications of using the ideas and tools of learning theory, such as the robust-learning mentioned above (see Attias et al. (2019); Madry et al. (2017); Goodfellow et al. (2014)), fairness (see Dwork et al. (2012); Kearns et al. (2018)) and, most notably, privacy.

In light of the above, compression schemes evolved into a crucial concept on which classical vanilla learning theory diverges in an essential way from the area of learning under privacy concerns. The work in this chapter is joint with Steve Hanneke and Aryeh Kontorovich (ALT 2019b).

1.2 Privacy

As progressively more technology products become based upon machine learning tools, and more branches of science turn to a more "evidence based" methodology, the use of data becomes increasingly dominant. A lot of these data sets consist of personal information, such as medical records, customer preferences, music listening history or user behavior within sites. Such personal data is being collected by more companies in order to gain insight and improve products, or to conduct scientific studies.

The importance of preserving the privacy of such data is common knowledge. It has induced laws restricting the information that can be gathered by companies, and regulations regarding who may use it in research, and in what manner. Privacy is important when treating personal data, but crucial when a leak could cause actual harm or embarrassment to an individual. The straightforward definition of data leakage is unauthorized exposure of the data itself. This type of concern is at the core of cryptography and security research. A different type of data leakage, which is less intuitive, comes from releasing information about the data. Partial information about data, summary statistics or drawn conclusions are often regarded as safe for public release. To understand why this might reveal sensitive information, consider the case of language models. Huge language models have become essential in the recent remarkable progress in the field of natural language processing. It is known that models of this size do "memorize" part of their training data Liu et al. (2020); Arpit et al. (2017). Although this memorized data seems to be concealed within the model, which often serves as a black box, it was shown that clever inference attacks can recover properties of the training data, such as the membership of sentences or recovery of strings contained in the data (Shokri et al., 2016; Carlini et al., 2020). This might be a serious problem if attackers could use those models to infer private textual information such as social numbers, medical status, and other personal data.

To ensure that the result of such statistical and computational analysis will preserve privacy even against unknown future attacks, a mathematical definition and guarantee of privacy is desirable. In this thesis, we will focus on such a privacy notion called *Differential Privacy* Dwork et al. (2006c).

Consider a data set S consisting of n rows, each row representing the information about a specific individual. A (randomized) data analysis mechanism $\mathcal{M}$ will be regarded as privacy-preserving if its output distribution will not be significantly affected by any individual row. This intuitively guarantees that whatever can be learned about an individual by the mechanism output can be learned if the personal data is arbitrarily modified.

Definition 1.2.1 (Dwork et al. (2006c); Dwork (2008); Dwork et al. (2006a)). *A randomized algorithm $\mathcal{M}$ is* (ε, δ)-differentially private *if for every two databases S, S' that* *differ on one row (such databases are called* neighboring*), and every set of outcomes F, we have* $\Pr[\mathcal{M}(S) \in F] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(S') \in F] + \delta$*. The definition is referred to as* pure *differential privacy when $\delta = 0$, and* approximate *differential privacy when $\delta > 0$.*

Note that differential privacy is a property of the mechanism and not of its outcome.

Differential privacy began with the revolutionary work of Dinur, Dwork, Nissim, McSherry and Smith (Dwork et al., 2006b; Dinur and Nissim, 2003), who presented for the first time a mathematical formulation of algorithmic information privacy. Their model is based on the assumption that we trust the data curator or the server administrator who collects and holds the data, but not the rest of the public and not even the researchers or analysts who make professional use of the database. Still, we wish to provide meaningful statistics and algorithmic tools based on the gathered data. The main idea, to prevent the violation of the privacy of the information in the database, was to introduce additional randomness into the analytic process, usually by injecting limited noise into the calculations. This addition would mask the influence of any individual in the database on the results. Although this noise often affects accuracy, the hope is that since it is bounded, this defect in accuracy will vanish as the sample size increases. Several studies in recent years have shown that for many basic tasks, this is indeed the case. On the other hand, other tasks turned out to be problematic in the sense that there is a gap, sometimes unbridgeable, between the classical level of accuracy and that which can be achieved under private requirements.

Following this line, over the last decade we have witnessed an explosion of research in differential privacy, by now also being largely employed by major corporates such as DeepMind Balle et al. (2022a), Alphabet Erlingsson et al. (2014a) and IBM Holohan et al. (2019), and it was even embedded in the query system of the 2020 United States census Haney et al. (2021). for privacy-preserving data analysis. In particular, there has been a lot of interest in designing *private learning algorithms*, which guarantee differential privacy for their training data. Intuitively, this guarantees that the outcome of the learner (the identified hypothesis) leaks very little information in any particular point from the training set. Works in this area include (Kasiviswanathan et al., 2011; Beimel et al., 2014, 2019b, 2016a, 2020; Bun et al., 2015; Feldman and Xiao, 2015; Bun et al., 2019a; Beimel et al., 2019a; Kaplan et al., 2019, 2020a; Alon et al., 2020; Kaplan et al., 2020b; Bun et al., 2020b; Alon et al., 2019), and much more. At the same time, the boundaries and limitations of private learning were studied thoroughly. The main question is to characterize the *sample complexity* of private learning; more specifically, the cost of requiring learners to preserve privacy, i.e. by how much the sample complexity increases as a result of this requirement.

Several works demonstrated that, under pure differential privacy, some learning problems which non-privately can be learned with constant sample complexity require $\mathcal{O}\left(\log(|\mathcal{X}|)\right)$, when $\mathcal{X}$ is the domain from which data points are sampled.

More generally, Bun et al. showed that the complexity of private learning grows at most double exponentially with respect to the Littlestone, dimension and, specifically, the ability to learn privately is equivalent to the class having a finite Littlestone dimension. This characterization widens the gap, as the Littlestone dimension can grow in an unbounded manner relative to the VC dimension of the same class. In particular, there are known classes whose VC dimension is finite and even constant, but Littlestone dimension is infinite.

One of the most fundamental learning problems in which such a gap emerges is the class of thresholds or, more generally, the class of axis aligned rectangles.

1.2.1 Learning axis aligned rectangles

In this thesis, we revisit this fundamental open question of the sample complexity of learning axis-aligned rectangles with privacy. Non-privately, learning axis-aligned rectangles is one of the most simple and basic learning tasks which can be solved using compression ideas. As it is easily and intuitively compressible, it is often given as *the first* example for PAC learning in courses or books. Nevertheless, under privacy constraints the problem is not only impossible for infinite domains but also for finite domains much more work is needed in order to solve it. More formally, recall that the VC dimension of the class of all axis-aligned rectangles over $\mathbb{R}^d$ is $O(d)$, and hence a sample of size $O(d)$ suffices to learn axis-aligned rectangles non-privately (we omit throughout the introduction the dependency of the sample complexity in the accuracy, confidence, and privacy parameters). In contrast, it turns out that, with differential privacy, learning axis-aligned rectangles over $\mathbb{R}^d$ is impossible, even when $d = 1$ (Feldman and Xiao, 2015; Bun et al., 2015; Alon et al., 2019). In more detail, let $\mathcal{X} = \{1, 2, \dots\}$ be a finite (one dimensional) grid, and consider the task of learning axis-aligned rectangles over the finite d-dimensional grid $\mathcal{X}^d \subseteq \mathbb{R}^d$. In other words, consider the task of learning axis-aligned rectangles under the promise that the underlying distribution is supported on (a subset of) the finite grid $\mathcal{X}^d$.

For pure private learning, Feldman and Xiao (2015) showed a lower bound of $\Omega\left(d \cdot \log |\mathcal{X}|\right)$ on the sample complexity of this task. This lower bound is tight, as a pure-private learner with sample complexity $\Theta\left(d \cdot \log |\mathcal{X}|\right)$ can be obtained using the generic upper bound of Kasiviswanathan et al. (2011). This should be contrasted with the non-private sample complexity, which is independent of $|\mathcal{X}|$.

For approximate-private learning, Beimel et al. (2016a) showed that the dependency of the sample complexity in $|\mathcal{X}|$ can be significantly reduced. This, however, came at the cost of increasing the dependency in the dimension d. Specifically, the private learner of Beimel et al. (2016a) has sample complexity $\tilde{O}\left(d^3 \cdot 8^{\log^* |\mathcal{X}|}\right)$. We mention that a dependency on $\log^* |\mathcal{X}|$ is known to be necessary (Bun et al., 2015; Alon et al., 2019). Recently, Beimel et al. (2019a) and Kaplan et al. (2020b) studied the related problem of privately learning *halfspaces* over a finite grid $\mathcal{X}^d$, and presented algorithms with sample complexity $\tilde{O}\left(d^{2.5} \cdot 8^{\log^* |\mathcal{X}|}\right)$. Their algorithms can be used to privately learn axis-aligned rectangles over $\mathcal{X}^d$ with sample complexity $\tilde{O}\left(d^{1.5} \cdot 8^{\log^* |\mathcal{X}|}\right)$. This can be further improved using the recent results of Kaplan et al. (2020a), and obtain a differentially private algorithm for learning axis-aligned rectangles over $\mathcal{X}^d$ with sample complexity $\tilde{O}\left(d^{1.5} \cdot (\log^* |\mathcal{X}|)^{1.5}\right)$. We consider this bound to be the baseline for our work, and we will elaborate on it later on Section 2.

To summarize, our current understanding of the task of privately learning axis-aligned rectangles over $\mathcal{X}^d$ gives us two kinds of upper bounds in the sample complexity: Either $d \cdot \log |\mathcal{X}|$ or $d^{1.5} \cdot (\log^* |\mathcal{X}|)^{1.5}$. That is, current algorithms either require sample complexity that scales with $\log |\mathcal{X}|$, or else it scales super linearly in the dimension d. Our starting point for this part of our work was to find a differentially private algorithm for learning axis-aligned rectangles with sample complexity that scales linearly in d and is asymptotically smaller than $\log |\mathcal{X}|$.

This naturally leads to the following question.

Question 2. Is there a differentially private algorithm for learning axis-aligned rectangles with sample complexity that scales linearly in d and asymptotically smaller than $\log |\mathcal{X}|$?

Our Contribution

We answer question 2 in the affirmative, and present the following theorem.

Theorem 1.2.2 (informal)**.** *There exists a differentially private algorithm for learning axis-aligned rectangles over $\mathcal{X}^d$ with sample complexity $\tilde{O}\left(d \cdot (\log^* |\mathcal{X}|)^{1.5}\right)$.*

We do so by presenting a novel private algorithm for the problem which achieves this sample complexity.

In order to attain this improvement, a new algorithmic technique had to be developed. We elaborate on this in the main part of the thesis, we now present an intuitive simplified version of the technique. The main idea includes the deletion of "exposed" data-points on the go, in a manner designed to avoid the cost of the adaptive composition theorems (see Chapter 2), as each iteration can't affect other iterations, and by that avoid the super-linear growth in complexity suffered by former solutions. At each iteration, one axis is examined and a set of "candidate edge points" chosen, from which one point is picked using an interior-point solver. The core of this technique may be of individual interest, introducing a new method for constructing statistically-efficient private algorithms.

Nevertheless, this idea alone, removing "exposed"data points on the go, is not enough. The failure point is that by deleting *one* point from the data, we can create a "domino effect" that affects (one by one) many of the candidate sets throughout execution. Recall that differential privacy requires analysis of runtime upon neighboring datasets (see Definition 1.2.1 above). Consider two neighboring datasets S and $S' = S \cup \{(x', y')\}$ for some labeled point $(x', y') \in \mathcal{X}^d \times \{0, 1\}$. Suppose that during the execution on S', x' gets picked as a "candidate point", thus the additional point x' participates "only" in the first iteration of the algorithm, and is afterwards deleted. However, since the size of the candidate sets is fixed, during the execution on S (without the point x') it holds that *a different point* z gets included instead of x', and this point z is then deleted from S (but it is not deleted from S' during the execution on S'). Therefore, also during the second iteration, we have that S and S' are not identical (they still differ on one point) and this domino effect can continue throughout the execution. That is, a single data point can affect many of the executions of the algorithm, and we would still need to pay in composition to argue privacy.

Our solution is based on two modifications to this approach. First, we add noise to the size of the candidate sets, but we only use the n "inner" points from these sets. Second, we delete elements from S not based on them being inside those sets, but only based on the (privately computed) interval stretching from the first chosen edge point to the second (at a given axis). This allows us to indeed separate the privacy analysis to each axis in its own, without any influence between iteration, hence avoiding the need for applying composition theorems and achieving an improved sample complexity. The work in this chapter is joint with Uri Stemmer (NeurIPS 2021).

1.2.2 Universal Private Learning

Continuing this line of research, we understand that private learning is inherently harder than classical. This is true theoretically, as just explained, but also practically, as designing and implementing differential private algorithms proved to be challenging. Despite tremendous efforts, in a lot of cases the state-of-the-art is still far from satisfactory. For example, the recent deployment of differential privacy by the US Census only guarantees a privacy parameter of $\varepsilon = 19.61$ (Bureau, 2021a), which translates to a relatively weak privacy guarantee.[5] One possible explanation for these challenges is that most of the works on DP learning are inspired and explained by worst-case mathematical models such as the theory of PAC Learning (Valiant, 1984), which is based on a *distribution-free perspective*. While it gives rise to a clean and compelling mathematical picture, one may argue that the PAC model fails to capture at a fundamental level the true behavior of many practical learning problems (regardless of privacy consideration). A key criticism

[5] Observe that smaller privacy parameters ε, δ translate to stronger privacy guarantees in Definition 1.2.1, in the sense that a single input point would have a smaller effect on the outcome distribution. On the flip side, reducing the privacy parameters is typically obtained by adding more noise and uncertainty to the computation, which often translates to a loss in accuracy.

of the PAC model is that the distribution-independent definition of learnability is too pessimistic: real-world data is rarely worst-case, and experiments show that practical learning rates can be much faster than predicted by PAC theory (Cohn and Tesauro, 1990b, 1992b). It therefore appears that the worst-case nature of the PAC model hides key features that are observed in practice. Furthermore, these shortcomings seem to be amplified in the context of *private* PAC learning. Those challenges, which are strengthened by theoretical results (e.g. the sample complexity gap mentioned in Section 1.2.1) seem to reflect the worst-case distribution-free nature of the PAC model rather than fundamental limitations of DP learning. This thesis, therefore, advocates the study of distribution-dependent private-learning, as this can lead to a more optimistic (and realistic) landscape of differentially private learning. We investigate a distribution dependent model known as *universal learning* and ask the following fundamental question:

Question 3. For which problems or classes there exists a learning rule, such that for every distribution, the rule's learning rate converges with the best risk in class as the number of examples tends to infinity.

As before, we might need to avoid natural universal learners which are compression based, such as K-nearest neighbors Devroye et al. (2013) and OptiNet (Hanneke et al., 2019a), as their behavior is inherently influenced by the sample at hand and thus might be too sensitive to preserve privacy.

Our Contribution

We uncover the following general result:

Theorem 1.2.3. *For every $d \in \mathbb{N}$ and every $\varepsilon \leq 1$ there is an $(\varepsilon, 0)$-differentially private universal consistent (UC) learner over $\mathbb{R}^d$.*

Recall that, as we mentioned, learning one-dimensional linear classifiers over $\mathbb{R}$ with differential privacy is impossible in the PAC model. Theorem 1.2.3 circumvents this impossibility result: not only are one-dimensional linear classifiers learnable in the UC model, but in fact *every* class (over $\mathbb{R}^d$) is learnable in this setting, and furthermore, there is a single (universal consistent) algorithm that learns them all (w.r.t. any distribution).

To obtain Theorem 1.2.3 we design a simple variant for the classical *histogram rule* (Glick, 1973; Gordon and Olshen, 1978, 1980; Devroye et al., 2013) that partitions $\mathbb{R}^d$ into cubes of the same size (where the size decreases with the sample size n), and makes a decision according to the majority vote within each cube. This algorithm is particularly suitable for differential privacy, and can be made private simply by adding noise to the votes within each cube. In the analysis, we show that this does not break the universal consistency of the histogram rule. Furthermore, We generalize those results to any separable metric spaces with bounded doubling dimension.

Remark 1.2.4. *For simplicity, in Theorem 1.2.3 we fixed ε to be a constant (independent of the sample size n). Our algorithm trivially extends to a setting where ε decreases with the sample size.*

We extend our results to the more general setting of *density estimation* in the UC model (with respect to the total variation metric). That is, we seek a differentially private algorithm which upon receiving a sample, outputs a density function which should be close to the true unknown density function from which the sample was drawn. We say that an algorithm is universal consistent if the *Total Variation distance* between the output density and the true density converges to zero as the sample size grows to infinity. Formally:

Definition 1.2.5 (Universal consistent density estimation, informal (Devroye and Györfi, 1985))**.** *Let $\mathcal{X}$ be a domain and let $\mathcal{A}$ be an algorithm whose output is a density function over $\mathcal{X}$. Algorithm $\mathcal{A}$ is a* universal consistent (UC) density estimator *over $\mathcal{X}$ if for every α, β and for every distribution μ over $\mathcal{X}$ there is a constant $n = n(\alpha, \beta, \mu)$ such that $\Pr_{\substack{S\sim\mu^n \\ f\leftarrow\mathcal{A}(S)}} [\|f - \mu\|_{\mathsf{TV}} > \alpha] < \beta$.*

Unlike our private UC learner, which satisfies differential privacy with $\delta = 0$ (this is sometimes referred to a *pure* differential privacy), our private UC density estimator only satisfies differential privacy with $\delta > 0$. When using $\delta > 0$, it is commonly agreed that the definition of differential privacy only provides meaningful guarantees as long as $\delta \ll 1/n$. Therefore, unlike with our private UC learner, we must let δ decay with the sample size n. Our result is a universal consistent differentially private algorithm for density estimation for which δ decays exponentially in the sample size.

Theorem 1.2.6. *Let $d \in \mathbb{N}$, let $\varepsilon \leq 1$, and let $\delta : \mathbb{N} \to [0, 1]$ be a function satisfying $\delta(n) = \omega(2^{-\sqrt{n}})$. There is an $(\varepsilon, \delta(n))$-differentially private universal consistent (UC) density estimator over $\mathbb{R}^d$.*

This work is an important first step towards understanding differentially private universal learning.

The work in this chapter is joint with Olivier Bousquet, Haim Kaplan, Aryeh Kontorovich, Yishay Mansour, Shay Moran and Uri Stemmer 2022.

Finally, we turn our attention to the close subject of *adaptive data analysis*.

1.3 Adaptive data analysis

Statistical validity is a well known crucial aspect of modern science. In the past several years, the natural and social science communities have come to realize that such validity was not in fact preserved in numerous peer-reviewed and widely cited studies, leading to many false discoveries. Known as the *replication crisis*, this phenomenon threatens to undermine the very basis for the public's trust in science.

One of the main explanations for the prevalence of false discovery arises from the inherent *adaptivity* in the process of data analysis. To illustrate this issue, consider a data analyst interested in testing a specific research hypothesis. The analyst acquires relevant data, evaluates the hypothesis, and (say) learns that it is false. Based on the findings, the analyst now decides on a second hypothesis to be tested, and evaluates it on the *same data* (acquiring fresh data might be too expensive or even impossible). That is, the analyst chooses the hypotheses *adaptively*, where this choice depends on previous interactions with the data. As a result, the findings are no longer supported by classical statistical theory, which assumes that the tested hypotheses are fixed before the data is gathered, and the analyst runs the risk of overfitting to the data.

In order to tackle this setting, we first make it explicit. We give here the formulation presented by Dwork et al. (2015c). We consider a two-player game between a mechanism $\mathcal{M}$ and an adversary $\mathbb{A}$, defined as follows (see Section 3.1.5 for precise definitions).

1. The adversary $\mathbb{A}$ fixes a measure μ over $\mathcal{X}^n$ (satisfying some conditions).
2. The mechanism $\mathcal{M}$ obtains a sample $S \sim \mu$ containing n (possibly correlated) observations.
3. For k rounds $j = 1, 2, \ldots, k$:
   - The adversary chooses a *query* $h_j : \mathcal{X} \to \{0, 1\}$, possibly as a function of all previous answers given by the mechanism.
   - The mechanism obtains h_j and responds with an answer $z_j \in \mathbb{R}$, which is given to $\mathbb{A}$.

We say that $\mathcal{M}$ is (α, β)-*empirically-accurate* if with probability at least $1 - \beta$ for every j it holds that $|z_j - h_j(S)| \leq \alpha$, where $h_j(S) = \frac{1}{n}\sum_{x\in S} h_j(x)$ is the empirical average of h_j on the sample S. We say that $\mathcal{M}$ is (α, β)-*statistically-accurate* if with probability at least $1 - \beta$ for every j it holds that $|z_j - h_j(\mu)| \leq \alpha$, where $h_j(\mu) =_{T\sim\mu} [h_j(T)] = \mathbb{E}_{T\sim\mu}\left[\frac{1}{n}\sum_{x\in T} h_j(x)\right]$ is the "true" value of the query h_j on the underlying distribution μ. Our goal is to design mechanisms $\mathcal{M}$ providing statistical-accuracy.

Starting from Dwork et al. (2015b,a), it has been demonstrated that various notions of *algorithmic stability*, and in particular *differential privacy*, allow for methods which maintain statistical validity under the adaptive setting. The vast majority of the works in this area, however, strongly rely on the assumption that the data is sampled in an i.i.d. fashion. This scenario excludes some natural and essential problems in learning theory such as Markov chains, active learning, and autoregressive models (Kontorovich and Ramanan, 2008; Kontorovich and Weiss, 2014; Kontorovich and Raginsky, 2017;

Settles, 2009; Hanneke et al., 2014; Sacerdote, 2001).

A notable exception is a stability notion introduced by Bassily and Freund (2016), called *typical-stability.* This beautiful and natural notion has the advantage that, under some conditions on the underlying distribution, it can guarantee statistical validity even for non-i.i.d. settings. However, one downside of the results of Bassily and Freund (2016) is that they do not recover the i.i.d. generalization bounds in the limiting regime where the dependencies decay to zero. In particular, in the i.i.d. setting, it is possible to efficiently answer $\tilde{O}(n^2)$ adaptive queries given a sample of size n. In contrast, the results of Bassily and Freund (2016) only allow to answer $\tilde{O}(n)$ adaptive queries, *even if the dependencies in the data decay to zero.* Motivated by this gap and the above results, we ask the following question:

Question 4. *Can the tools and results from the adaptive data analysis literature be extended to the correlated examples setting, giving a meaningful bound while also recovering the i.i.d. generalization bounds in the limiting regime where the dependencies decay to zero?*

Our Contribution

Our first contribution to this line of work is to extend existing generalization results for differential privacy from the i.i.d. setting to the correlated setting. To that end, we introduce a notion we call *Gibbs dependence* to quantify the dependencies between the covariates of a given joint distribution. We complement this result with a tight negative example. Our second contribution is to extend the connection between transcript-compression and adaptive data analysis also to the non-iid setting. Finally, we demonstrate an application of our results for when the underlying measure can be described as a Markov chain.

Gibbs Dependence We extend the connection between differential privacy and generalization to the case where the observations are correlated. We quantify the correlations in the data using a new notion, called *Gibbs dependence*, which is closely related to the classical *Dobrushin* interdependence coefficient (Kontorovich and Raginsky, 2017; Levin and Peres, 2017). Intuitively, a measure which has ψ-Gibbs dependency is such that knowledge about almost the entire sample does not provide too much information about the remaining portion. Formally,

Definition 1.3.1. *For a probability measure μ over a product space $\mathcal{X}^n$, define*

$$\psi(\mu) = \sup_{x \in \mathcal{X}^n} \mathop{\mathbb{E}}_{i \sim [n]} \left\| \mu_i(\cdot) - \mu_i(\cdot \mid x^{-i}) \right\|_{\mathsf{TV}},$$

where $\mu_i(\cdot)$ is the i^{th} marginal measure and $\mu_i(\cdot \mid x^{-i})$ is the ith marginal measure conditioned on all the coordinates other than i (given some n-tuple x).

Given ψ we say the μ has ψ-Gibbs dependence if $\psi(\mu) \leq \psi$.

A naive way for leveraging our notion of Gibbs dependence would be to "union bound" the correlations across the n different coordinates. Specifically, one could show that if μ has Gibbs-dependence ψ then $\|\mu - \mu^*\|_{\mathsf{TV}} \leq n\psi$, where μ^* is the product distribution in which every coordinate is sampled independently of the corresponding marginal distribution in μ. Thus, if $\psi \ll \frac{1}{n}$ then one could argue about generalization w.r.t. μ by applying existing generalization bounds w.r.t. μ^* in the independent case (since in this regime we have $\|\mu - \mu^*\|_{\mathsf{TV}} \ll 1$). This argument, however, only works when the dependencies in μ are *very weak* (i.e., when $\psi \ll \frac{1}{n}$). We contribute to this line of work by showing that differential privacy still provides generalization even if ψ is much larger, e.g., a constant independent of the sample size n. Specifically,

Theorem 1.3.2. *Let $\mathcal{M}$ be an (ε, δ)-differentially-private mechanism which is (α, β)-empirically-accurate for k rounds given n samples. If $n \geq \frac{\log(2k\varepsilon/\delta)}{\varepsilon^2}$, then $\mathcal{M}$ is also $(\alpha + 10\varepsilon + 2\psi, \beta + \frac{\delta}{\varepsilon})$-statistically-accurate.*

Remark 1.3.3. *For the case when ψ is zero, and hence μ is a product measure (see Example 7.4.1 below), Theorem 1.3.2 recovers the results achieved by differential privacy for i.i.d. samples Dwork et al. (2015c); Bassily et al. (2016). Thus, Theorem 1.3.2 generalizes the connection between differential privacy and generalization to the correlated setting.*

Intuitively, the above theorem states that if the underlying distribution has Gibbs-dependence ψ then the additional generalization error incurred by DP algorithms (compared to the iid setting) is at most $O(\psi)$. We complement this result with a tight negative example, showing that there exists a distribution μ with Gibbs-dependence ψ and a DP algorithm $\mathcal{A}$ that obtains generalization error $\Omega(\psi)$. This means that, in terms of the Gibbs-dependence, our result is tight.

By applying Theorem 1.3.2 with a known DP mechanism for answering queries while providing empirical accuracy, we are able to provide a computationally efficient mechanism for answering adaptive queries for the case of non-zero dependencies, which sample complexity depends in the Gibbs dependency of the query class. Formally:

Corollary 1.3.4. *There is a computationally efficient mechanism $\mathcal{M}$ that is $(\alpha + 2\psi, \beta)$-statistically-accurate for k adaptively chosen queries given a sample (an n-tuple) from an underlying measure with Gibbs-dependency ψ satisfying $n \geq \tilde{O}\left(\frac{\sqrt{k}}{\alpha^2} \log \frac{1}{\beta}\right)$.*

This result, again, generalizes the state-of-the-art bounds for the i.i.d. setting, where $\psi = 0$. In particular, Corollary 1.3.4 shows that mild dependencies in the data, say $\psi = \alpha$, come *for free* in terms of the achievable bounds for adaptive data analysis. We emphasize that $\psi = \alpha$ captures non-negligible dependencies. In particular, α could be constant, independent of the sample size n.

Transcript Compression The second direction we examine is that of *transcript compression.*

Compression has been used in the context of adaptive data analysis. Dwork et al. (2015a) used the definition of *bounded description length* (referred to here as *transcript compression*) to present an algorithm that is able to adaptively answer queries when the data is i.i.d. sampled. This notion of compression differs from the one mentioned above; it is more involved but, as shown by Dwork et al. (2015a) and also by our results, it is suited for tasks requiring low sensitivity tools.

Our contribution here is in generalizing this idea by showing that the same definition, when used in the right setting, allows maintaining adaptive accuracy even when the distribution includes dependencies.

Following the approach of Bassily and Freund (2016), we aim to provide the following guarantee: As long as the analyst chooses functions which, in the non-adaptive setting, are concentrated around their expected value, then the answers given by the mechanism should be accurate. Intuitively, the idea is that functions with large variance are hard to approximate even in the non-adaptive setting, and hence, we should not require our mechanism to approximate them well in the adaptive setting.

This is formalized as follows. For every query q and every distribution μ, we write $\gamma(q, \mu, \delta)$ to denote the length of a confidence interval around the expectation of q with confidence level $(1 - \delta)$. That is, $\gamma(q, \mu, \delta)$ is such that when sampling $T \sim \mu$, with probability at least $(1-\delta)$ it holds that $q(T)$ is within $\gamma(q, \mu, \delta)$ from its expectation. We obtain the following theorem (see Section 7.2.7 for a precise statement).

Theorem 1.3.5 (informal)**.** *Fix $\alpha, \delta > 0$. There exists a computationally efficient mechanism with the following properties. The mechanism obtains a sample (an n-tuple) from some unknown underlying distribution μ. Then, for k rounds $i = 1, 2, \ldots, k$, the mechanism obtains a query q_i and responds with an answer a_i such that*

$$\Pr[\exists i \text{ s.t. } |a_i - q_i(\mu)| > \alpha + \gamma(q_i, \mu, \delta)] \leq \delta \cdot k \cdot 2^{k \cdot \log \frac{1}{\alpha}}.$$

Proving that there exists a computationally efficient mechanism for answering adaptive queries, even for the case of non-zero dependencies, which error scales with γ. In particular, as long as the adversary poses queries q_i such that $\gamma(q_i, \mu, \delta) \leq \alpha$, our mechanism guarantees that all of its answers are 2α-accurate, with high probability. The caveat here is that for the probability to be high we must ensure that δ decays fast enough with the number of queries at hand. This is easily obtained in many settings of interest by taking the sample size n to be big enough. For example, for sub-Gaussian or sub-exponential queries, we would get that δ vanishes exponentially with n, and hence, for large enough

n we would get the desired result.

We note that in the case of *non-adaptive* data analysis, learning from non-i.i.d samples is a well-known problem that has been heavily studied in various directions. This includes works on the Markovian criteria Marton (1996); Kontorovich and Raginsky (2017); Wolfer and Kontorovich (2019); Juang and Rabiner (1991), as well as other criteria such as those researched by Daskalakis et al. (2019); Dagan et al. (2019). These lines of work do not transfer, at least not in a way that we are aware of, to the adaptive setting.

Bassily and Freund (2016) also studied the problem of adaptive data analysis with correlated observations; we now elaborate on the differences.

1. **Results regarding transcript compression (Section 7.2)** As we mentioned, Bassily and Freund (2016) introduced the beautiful framework where the mechanism is required to provide accurate answers only as long as the analyst poses "concentrated queries". They obtained their results for this setting via a new notion they introduced, called *typical stability.* However, their analysis and definitions are quite complex. We show that essentially the same bounds can be obtained *in a significantly simpler way, using standard compression tools.* Specifically, our result (Theorem 1.3.5) recovers essentially the same bounds for all types of queries considered by Bassily and Freund (2016), including bounded-sensitivity queries, subgaussian queries, and subexponential queries. In addition to being significantly simpler, our result offers the following advantage: Using the results of Bassily and Freund (2016), we need to know *in advance* the parameter controlling the "concentration level" of the queries that will be presented in runtime, and this parameter is used by their algorithm. In contrast, our algorithm is oblivious to this parameter, and the guarantee is that our accuracy depends on the "concentration level" of the given queries. Furthermore, with our algorithm, different queries throughout the execution can have different "concentration levels", a feature which is not directly supported by Bassily and Freund (2016).

2. **Results regarding Gibbs-dependence (Section 7.1).** These results are in a different setting than that of Bassily and Freund, and the results are not directly comparable. In particular, Bassily and Freund do not study any specific dependence notion, such as Gibbs dependence. Our results show that assuming low Gibbs dependence allows for improved bounds. Specifically, Bassily and Freund (2016) can answer at most $\widetilde{\mathcal{O}}\left(n\right)$ adaptive queries efficiently, even if the dependencies within the sample are very weak. Using our notion of Gibbs-dependency, we can answer $\widetilde{\mathcal{O}}\left(n^2\right)$ adaptive queries efficiently, while accommodating small (but non-negligible) dependencies.

The work in this chapter is joint with Aryeh Kontorovich and Uri Stemmer (ICML 2022).

Chapter 2

Literature Review

2.1 Compression Schemes

Background. It appears that generalization bounds based on sample compression were independently discovered by Devroye and Wagner (1979) and Littlestone and Warmuth (1986) (when the former dealt with nearest-neighbor rules and the latter with generic learning algorithms), and further elaborated upon by Graepel et al. (2005); see Floyd and Warmuth (1995) for background and discussion. A more general kind of Occam learning was discussed in Blumer et al. (1989). Computational lower bounds on sample compression were obtained in Gottlieb et al. (2014), and some communication-based lower bounds were given in Kane et al. (2017).

Compression and Boosting. The idea of constructing compression schemes using the boosting technique is known in the literature. The first mention of this connection was made by Freund and Schapire (1997). In their seminal work they proved that boosting is possible by answering an open question, suggested by Kearns and Valiant (1994). Starting with Freund (1990) and later on with the construction of the famous `AdaBoost` algorithm by Freund and Schapire (1997), the boosting mechanism relied on an intermediate construction, providing a compression scheme of size $\mathcal{O}\left(d_{\mathcal{C}} \log n\right)$ for binary function classes $\mathcal{C}$ with VC-dimension $d_{\mathcal{C}}$. Continuing this line of work, Moran and Yehudayoff (2016) discuss the idea of leveraging this connection between boosting and compression, and recognize that their main result can in fact be seen as a refinement of this connection.

Real-Valued Functions. Beginning with Freund and Schapire (1997)'s `AdaBoost.R` algorithm, there have been numerous attempts to extend AdaBoost to the real-valued case (Bertoni, Campadelli, and Parodi (1997); Drucker (1997); Avnimelech and Intrator (1999); Karakoulas and Shawe-Taylor (2000); Duffy and Helmbold (2002); Kégl (2003); Nock and Nielsen (2007)) along with various theoretical and heuristic constructions of particular weak regressors (Mason et al., 1999; Friedman, 2001; Mannor and Meir, 2002); see also the survey Mendes-Moreira et al. (2012).

An explanation for the challenge of defining a good weak-learner was pointed and explained by Duffy and Helmbold (2002, Remark 2.1) we discuss this issue on 4.1.2. The (η, γ)-weak learner, which has appeared, among other works, in Anthony et al. (1996); Simon (1997); Avnimelech and Intrator (1999); Kégl (2003), gets around this difficulty, but provable general constructions of such learners have been lacking. Likewise, the heart of our sample compression engine, `MedBoost`, has been widely in use since Freund and Schapire (1997) in various guises. Our Theorem 1.1.4 supplies the remaining piece of the puzzle: *any* sample-consistent regressor applied to some random sample of bounded size yields an (η, γ)-weak hypothesis. The closest analogue we were able to find was Anthony et al. (1996, Theorem 3), which is non-trivial only for function classes with finite pseudo-dimension, and is inapplicable, e.g., to classes of 1-Lipschitz or bounded variation functions (see 4.2.3).

The literature on general sample compression schemes for real-valued functions is quite sparse. There are well-known narrowly tailored results on specifying functions or approximate versions of functions using a finite number of points, such as the classical fact that a polynomial of degree p can be perfectly recovered from $p + 1$ points. To our knowledge, the only *general* results on sample compression for real-valued functions (applicable to *all* learnable function classes) is Theorem 4.3 of David, Moran, and Yehudayoff (2016). They propose a general technique to convert any learning algorithm into a compression scheme. However, their notion of compression scheme is significantly weaker than ours: namely, they only guarantee a bound on the average error rather than than our *uniform* approximation requirement. In particular, in the special case of a family of *binary*-valued functions, their notion of sample compression does *not* recover the usual notion of sample compression schemes for classification, whereas our uniform ε-approximate compression notion *does* recover it as a special case. We therefore consider our notion to be a more fitting generalization of the definition of sample compression to the real-valued case.

Ashtiani et al. (2020) adopted the notion of a compression scheme to the distribution learning problem. They showed that if a class of distributions admits robust compressibility then it is agnostically learnable. They used those results in order to provide state-of-the-art sample-complexity bounds for learning a mixture of Gaussians.

2.2 Privacy

2.2.1 Background

The formal connection between differential privacy and learning theory was made by Kasiviswanathan et al. (2011), who proposed the Privately Probably Approximately Correct learning model (PPAC for short). One of the fundamental results In the field of machine learning is that the sample complexity of PAC learning is proportional to the Vapnik-Chervonenkis (VC) dimension of the concept class being learned. Kasiviswanathan et al. demonstrated that for any finite concept class, there exists a privacy-preserving learner with sample complexity that is logarithmic in the size of the class. However, Beimel et al. (2010) showed that while a non-private learner can properly learn the concept class of point functions (which consists of functions that evaluate to 1 for a single element and 0 anywhere else) with a sample complexity of $O(1)$, a pure differential private learner requires a sample complexity of $\Omega(\log(|\mathcal{X}|))$ for proper learning. Later on, Feldman and Xiao (2015) demonstrated that this separation holds even for improper learning. It was shown by Beimel et al. (2016a) that this gap can be made significantly smaller by relaxing the privacy requirement to approximate privacy (i.e. $\delta > 0$); nevertheless, there is still a crucial gap for some classes.

Recently, it was proven by Alon et al. (2019); Bun et al. (2020b); Golowich and Livni (2021) that private learning and online learning are equivalent. As online learning can be characterized by the Littlestone dimension (Littlestone, 1988; Ben-David et al., 2009) thus implying that private learning is possible if and only if for classes with finite *Littlestone dimension*. Specifically, it was shown that the sample complexity of privately learning a class $\mathcal{C}$ is at most $poly(Ldim(\mathcal{C}))$ and at least $\log^*(Ldim(\mathcal{C}))$, when $Ldim(\mathcal{C})$ represents the Littlestone dimension of the class. Within this large quantity gap, the exact dependency of sample complexity on the Littlestone dimension is unknown. Exploring the relationship between sample complexity and measures such as Littlestone dimension, VC dimension, and potentially other measures is an essential open question in the field.

2.2.2 Alternative models

In recent years several variants and modifications have been proposed for the initial definition of differential privacy.

Rényi differential privacy (Mironov (2017)). One main variant of differential privacy is the *Rényi differential privacy* (RDP for short). This definition utilizes the notion of *Rényi divergence*, a measure of the difference between two probability distributions.

Definition 2.2.1 (Rényi divergence). *Given two discrete distributions* $\mathcal{P} = \{p_1, \dots, p_n\}$,

$\mathcal{Q} = \{q_1, \ldots, q_n\}$ *The α-Rényi divergence of $\mathcal{P}$ and $\mathcal{Q}$ is defined as*

$$\mathcal{D}_\alpha(\mathcal{P}, \mathcal{Q}) := \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{n} \frac{p_i^\alpha}{q_i^{\alpha-1}} \right)$$

When for $\alpha \in \{0, 1, \infty\}$ is defined by taking a limit [1].

Recall that pure epsilon-differential privacy can be stated as a bound on the privacy loss function, i.e. $\log \frac{\Pr(\mathcal{M}(S) \in F)}{\mathcal{M}(S') \in F} \leq \varepsilon$ for any two neighboring data sets S, S'. In a similar manner, an algorithm will be said to preserve (α, ε)-Rényi differential privacy, requiring instead that the Rényi divergence between $\mathcal{M}(S)$ and $\mathcal{M}(S')$ be bounded by ε, $\mathcal{D}_\alpha(\mathcal{M}(S), \mathcal{M}(S')) \leq \varepsilon$. It can be shown that if α is taken to infinity we get $\mathcal{D}_\infty(\mathcal{P}, \mathcal{Q}) = \log \sup_{i \in [n]} \frac{p_i}{q_i}$ (van Erven and Harremoës, 2012); hence, (∞, ε)-Rényi differential privacy is equivalent to ε-differential privacy.

It was shown by Mironov (2017) that any (α, ε)-Rényi differential private algorithm is also (ε', δ)-differentially private for $\varepsilon' = \varepsilon + \frac{\log(1/\delta)}{1-\alpha}$, hence, Rényi differential privacy can be seen as an intermediate notion between pure and approximate differential privacy. Moreover, its composition properties are easy to calculate, making it practically appealing. In addition to this definition, additional variants to the basic definition of differential privacy have been defined in recent years. Among them, we can mention concentrated differential privacy (zCDP) (Bun and Steinke, 2016), Gaussian differential privacy (f-DP) (Dong et al., 2019) and fuzzy differential privacy (Hou et al., 2022).

Local differential privacy (Kasiviswanathan et al. (2011)). Recall that differential privacy is based on the assumption that a trustworthy curator with access to all private information is responsible for collecting the data. *Local differential privacy* (Local-DP or LDP for short) is a variant of differential privacy that removes this assumption, requiring that the process of data collection itself preserves privacy. Historically, The first well-known privacy-preserving statistical mechanism, the "randomized response", proposed by Warner (1965) and Greenberg et al. (1969), was in fact locally private. In recent years, since its first formulation by Evfimievski et al. (2003) and Kasiviswanathan et al. (2011), the local model has gained great popularity in the world of research. Furthermore, it has found many uses in industry and in practical applications (see Acharya et al. (2020); Bureau (2021a); Erlingsson et al. (2014b); Murakami and Kawamoto (2019)). The model allows companies to guarantee their users that the personal information collected while using the various services cannot be exposed even in the event of malicious use of their own servers. The main caveat of the local model is that it is a stricter notion of privacy and thus can result in significantly reduced utility.

[1] The definition can be extended to continuous random variables but for the sake of simplicity we present it on its discrete form.

The shuffle model of differential privacy. This problem is addressed by another model, known as the *shuffle model* Cheu et al. (2018); Bittau et al. (2017); Erlingsson et al. (2018). In this model, each user adds some auxiliary randomness to their information before sending it to an intermediate server. This randomness is combined with the actual data, and the intermediate server shuffles all of the inputs together before transmitting them to the main server. By including random data from each user, the shuffle model helps to obscure any individual user's contribution to the final result. It was shown by Cheu et al. (2018) that for some problems this model can achieve significantly better accuracy than the local model (see Cheu (2021) for a survey of results on such separations). The main disadvantage of the model, compared to the above local model, is that users have to trust the auxiliary server, both in terms of its reliability and in terms of the correctness of its implementation. Nevertheless, the shuffle model is of main interest in the privacy community (see Balle et al. (2020); Feldman et al. (2020b); Balle et al. (2019)).

2.2.3 Basic Differentially Private Tools

The inclusion of a privacy requirement necessitates the adaptation or reconstruction of statistical analyses and learning mechanisms, as they can no longer be utilized in their current form. As a result, in the past several years there was a tremendous effort of the research community to provide the fundamental building blocks of private statistics and private learning.

The Laplace mechanism Dwork et al. (2006b). In their seminal paper, Dwork et al. proposed the Laplace mechanism, which provides a way of answering queries as long as their sensitivity to changes in input is not high. One useful basic example which can be solved using the Laplace mechanism is answering counting queries, which simply asks for the number of points in the data which satisfies a certain property. It can be easily seen that this type of query has low sensitivity, any single change in the data will not change the count by more than 1, implying the need for a relatively small amount of noise added by the Laplace mechanism. Although very basic, counting queries are important primitives which can be composed into more complex algorithms, hence the ability to perform them in a privacy-preserving manner is highly important. Differential privacy can be composed by running several such primitives, resulting in a complex private mechanism which privacy parameters can be computed using the composition theorem (Dwork et al., 2006b, 2010a). Another, commonly used, set of functions that also enjoys low sensitivity is the histogram query. This defines a partitioning of the space into bins and then asks how many points in the data fall inside a specific bin. As in the case of counting queries, histogram queries enjoy low sensitivity to changes in the input data, and therefore each query can be answered with a small amount of noise added. In the case of large or even infinite spaces, the number of bins can be too large to answer all. In such a case, the notion of approximate privacy can be utilized using the Stability-based histogram algorithm Bun et al. (2019c), which essentially ignores almost the entire space and takes into account only the part of the support which has high probability mass.

The Exponential Mechanism McSherry and Talwar (2007). Later on, McSherry and Talwar proposed the Exponential Mechanism, a generic technique which able to tackle a wide range of tasks. The exponential mechanism, although it is often inefficient, as its run-time depends on the size of the space, can be used in an even broader set of tasks. A significant example is the one of private median estimation. The median, as opposed to the average, can be highly sensitive and hence can not be computed privately using the Laplace mechanism, yet it can be computed using the exponential mechanism.

The Sparse Vector technique (Dwork et al., 2009). Another primitive technique is the sparse vector technique by Dwork et al. (2009). Its main purpose is to answer only queries whose value is above some given threshold. The sparse vector technique allows the algorithm to "pay", in privacy, only for the queries which do pass the threshold and not for the entire stream of queries, which may be much larger. This technique lies at the core of several important works, including the private multiplicative weights mechanism by Hardt and Rothblum (2010).

Empirical Risk Minimization Another highly important task is the one of *empirical risk minimization* (ERM) which lies at the core of learning theory. The main setting involves an instance space and data set and a metric of the closeness of a prediction to the true label value called the loss function (see Section 3 for precise setting and definitions). The problem is to find a hypothesis or a vector, in the case of linear prediction, in which cumulative loss on the data set is minimal.

For bounded loss functions which are also Lipschitz, this task can be done using the exponential mechanism (Bassily et al., 2014). In order to solve the task in a broader setting and with better parameters, Song et al. (2013) introduced the *Differentially-Private Stochastic Gradient Descent* (DP-SGD), a private variant of the celebrated gradient descent algorithm used widely for almost every deep-learning framework. The DP-SGD mechanism has been thoroughly studied, and various modifications suggested through the years, in order to improve the utility-privacy trade-off in general or for specific settings (Abadi et al., 2016; Wang et al., 2019; Jayaraman and Evans, 2019; Kairouz et al., 2021; Liu and Lu, 2021).

2.2.4 Current research directions

In recent years, research and development in the field of privacy have experienced significant growth, both in theoretical and practical aspects. On the theoretical side, several significant results can be mentioned:

Characterization of Private Learning. First, as mentioned above, the line of research attempting to characterize what can be learned in a way that preserves privacy still includes some crucial gaps. Despite the recent breakthroughs that have shown that private learning is equivalent to online learning, i.e., is possible if and only if the Littlestone dimension of the class is finite, there is still a gap when it comes to the exact quantification of the connections (Alon et al., 2022; Ghazi et al., 2020).

Private Convex Optimization. As a continuation of the results concerning empirical risk minimization, which often relies on the DP-SGD algorithm, a fairly significant part of the research deals with the subject of private convex optimization. This area of research deals with the different ways to find optimal solutions (maximum or minimum) of convex functions. These problems, which are at the heart of many fields even beyond the ERM problem, are of great importance to the research world. Many natural scenarios can be formulated in this framework, such as the support vector machine algorithm (SVM) (Shalev-Shwartz and Ben-David, 2014), network-flow optimization (Wei et al., 2017), image denoising (Thanh et al., 2019) and much more. Therefore, it is important to find solutions that also meet the demand for privacy. To name a few key results from the last few years, we can point to Kifer et al. (2012); Iyengar et al. (2019); Bassily et al. (2021, 2019a); Feldman et al. (2020a); Kulkarni et al. (2021); Wu et al. (2016).

2020 United States census. On the practical side, the latest highlight was the transition of the *US Central Bureau* to the use of differential privacy in order to preserve the privacy of those surveyed in the 2020 census. The census, which takes place once a decade, is required by law to prevent harm to the privacy of those who participate (Bureau, 2021c). In previous surveys various privacy heuristics were used, but after a long discussion it was decided to change the methodology and use differential privacy. The project, perhaps the largest routine statistical project in the United States, required the Central Bureau of Statistics to integrate various tools from the literature within an analysis platform that would be accessible to the general public of researchers, most of whom had no prior knowledge of differential privacy (Bureau, 2021b). The great applied challenge constitutes a significant milestone on the way to the assimilation of the paradigm as an essential tool in additional concrete statistical analyses.

Deep Learning. As part of the applied research in the field of privacy, similar to the situation in the machine learning community, the focus of interest is in deep learning. As we mentioned above, at the center of the research is the DP-SGD algorithm and attempts to improve it so that it enables increasingly better performance, under the same privacy constraints. The fundamental problem of image classification is considered a central test case. Prior work on differentially-private deep learning by Abadi et al. (2016) demonstrated its use on the standard image classification benchmark data sets - MNIST and CIFAR-10. On the MNIST data set, which is considered an easy task, the authors achieved 0.9, 0.95, and 0.97 test set accuracy for $(0.5, 10^{-5})$, $(2, 10^{-5})$ and $(8, 10^{-5})$-differential privacy, respectively. On the more challenging CIFAR-10, the authors achieved accuracy of 0.67, 0.7, and 0.73, for $(2, 10^{-5})$, $(4, 10^{-5})$, and $(8, 10^{-5})$-differential privacy, respectively. This is a significant gap from the non-private state-of-the-art which is currently 0.9991 for MNIST and 0.995 for CIFAR-10 (An et al., 2020; Dosovitskiy et al., 2020). These results had repeatedly been improved up until the recent work by Balle et al. (2022a), which introduced a differentially private model based on a 40-layer Wide-ResNet neural network, achieving 0.814 accuracy under $(8, 10^{-5})$-differential privacy for CIFAR-10. Some more recent works propose different techniques in order to even further improve those results, using pre-training on non-private data, hyperparameter tuning, and so on (see Kurakin et al. (2022)).

Attacks. Parallel to this direction, many studies deal with presenting the problems of using algorithms that do not preserve privacy. The research in this direction revolves around the idea of reconstruction attacks and membership attacks. These attacks recover parts of the information from the models and the information that is traditionally released to the network openly, together with the non-private models. Certain attacks succeed in identifying, with high probability, whether information about a certain individual appeared in the database that was used (Shokri et al., 2017; Hu et al., 2022; Carlini et al., 2021; Choquette-Choo et al., 2020). Other attacks perform the reconstruction of images that appeared in the database on which models were trained to classify images (Balle et al., 2022b) and some attacks reconstruct words that appeared in the database on which language models were trained (Carlini et al., 2020). Recently, studies have also been done that leverage the idea of reconstruction attacks to try and quantify the relationship between privacy parameters and the likelihood or extent of information leakage from algorithms that preserve differential privacy (Hannun et al., 2021; Guo et al., 2022).

2.2.5 Open questions on the private learning domain

Research on differential privacy continues to advance with significant momentum in recent years. At the same time, many questions are still open and serious challenges are still facing the field. Fundamental questions are at the center of attention and require in-depth research. Problems exist, such as quantitative characterization of private learning, learning in models other than the classic PAC model, and understanding the meaning and connections between the various models, to name a few central issues.

In the applied direction, there is still a long way to go. A series of open-source libraries for implementing privacy-preserving algorithms are gaining momentum, but they are not yet complete. There is a requirement to understand the appropriate parameters for the various research and statistical needs. In addition, there is a fundamental need to find ways to improve the performance of privacy-preserving deep learning algorithms, so that they can be used in the various fields at the center of technological practice today, such as advanced classification models, large language models, and generative models.

2.3 Universal learning

Despite the dominance of the PAC model since its definition in the 1980s, over the years many criticisms have been heard (Buntine; Sarrett and Pazzani, 1989; Cohn and Tesauro, 1992a, 1990a). The main argument was that the model is too pessimistic, and that in reality the results are often fundamentally different from those obtained in the theory that derives from this definition.

For these reasons, Bousquet et al. proposed a relaxed model which they called *Universal Learning*. In this paper, the authors point out that in very natural cases the rate of convergence of different learning algorithms is several orders of magnitude faster than that which results from a pessimistic calculation under the definition of the PAC model. Hence, the main idea of the universal learning model is to allow the studied learning rate to depend not only on the class at hand but also on the distribution in the background. This idea also corresponds to the practical conduct in which the distribution does not change throughout the learning process but is fixed in the background, and only the sample size increases as needed.

This concept of learning rates that depend on the distribution (in contrast to the distribution-free nature of the PAC model) is not new. The classical notion of *bayes-consistency* defines a close model. The main difference is that Bayes consistency sets the class of functions to be learned to be the class of all measurable functions, as opposed to the universal learning model which allows for a more specific choice of classes.

The notion of Bayes consistency is one of the initial definitions of learning theory and can be found in the classical works of Fisher (1922); Stone (1977); Cover and Hart (1967) and Fix and Hodges (1989). After the introduction of the PAC model, the consistency definition was pushed aside and its study became secondary and less common. For this reason, the current literature dealing with this subject is quite sparse.

As for the study of private learning in a distribution-dependent context, prior work in this vein focused on obtaining better utility guarantees under the assumption that the underlying distribution adheres to certain "niceness" assumptions; for example margin assumptions. That is, these works do not aim to learn under *any* underlying distribution as we do, only under "nice" distributions. For example, private learning under margin assumptions was considered by Blum et al. (2005); Chaudhuri et al. (2014); Bun et al. (2020a); Nguyen et al. (2020), and private clustering under data stability assumptions was previously considered by Nissim et al. (2007); Wang et al. (2015); Huang and Liu (2018);Shechner et al. (2020); Cohen et al. (2021); Tsfadia et al. (2021). Another related work is by Haghtalab et al. (2020) who studied smooth analysis in the context of private learning, where the input points are perturbed slightly by nature. This is equivalent to assuming that the underlying distribution is not overly concentrated on any single point, which is similar in spirit to margin assumptions.

Parallel to our research and independently, this definition was investigated in a series of articles by Györfi and Kroll (2023, 2022) and Berrett and Butucea (2019). The results in these papers are very similar to ours, but with several differences. In terms of privacy, while we have focused on analyzing the curator model of differential privacy, Györfi and Kroll (2023, 2022) and Berrett and Butucea (2019) have primarily focused on local privacy. Also, they did a minimax analysis of the convergence rates for a specific family of density functions, while our focus was to demonstrate the possibility of private learning under the universal definition. In addition, they have proven results for regression problems, which is not the case in our research. Furthermore, while we used approximate privacy for density estimation, they focused on pure privacy only. On the other hand, for the classification case, we have shown results for general, metric (doubling) unbounded spaces, a property from which the construction and results of Györfi and Kroll (2023, 2022) and Berrett and Butucea (2019) do not benefit. Overall, our results provide a complementary perspective to their findings. Together, they provide a comprehensive perspective on the privacy analysis of machine learning algorithms, using this universal learning model to highlight the importance of considering different privacy metrics, problem types, and analysis techniques.

2.4 Adaptive data analysis

A classical technique used for multiple hypothesis testing is the Dunn-Bonferroni correction (Shaffer, 1995; Dunn, 1961), whose usage is limited due to the fact that it significantly reduces the number of discoveries. A more robust type suite of techniques is the false-discovery-rate (FDR) control method, such as the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) and the Bayesian approach by Storey (2003). Those methods focus mainly on statistical inference approaches such as hypothesis testing, confidence intervals etc., and are still a subject of active research (see also Li and Barber (2017); Javanmard and Montanari (2015)).

The different generic approach, which we also take here, uses the notion of algorithmic stability. Algorithmic stability is known to be intimately connected (and, in some settings, equivalent) to learnability (Bousquet and Elisseeff, 2002b; Shalev-Shwartz et al., 2010). Most of the existing stability notions, however, are not sufficient for our goal of adaptive learnability. For example, *uniform stability*, which has recently been the subject of several interesting results, is not closed under post-processing and does not yield the same type of adaptive generalization bounds as we study in this paper (Bousquet and Elisseeff, 2002a; Shalev-Shwartz et al., 2010; Hardt et al., 2016; Feldman and Vondrák, 2018, 2019). A notable exception is *local statistical stability*, which was shown to be both necessary and sufficient for adaptive generalization (Shenfeld and Ligett, 2019). However, so far, local statistical stability has not yielded new algorithmic insights.

A different line of research employs information-theoretic techniques, whereby overfitting is prevented by bounding the amount of mutual information between the input sample and the output hypothesis. However, these techniques generally only guarantee generalization in expectation, rather than high probability bounds (Russo and Zou, 2016; Xu and Raginsky, 2017; Rogers et al., 2016; Raginsky et al., 2016; Russo and Zou, 2019; Steinke and Zakynthinou, 2020).

The formulation of the adaptive data analysis we consider was introduced by Dwork et al. (2015b) (in the context of i.i.d. sampling), and has since then been the subject of many interesting papers (Bassily et al., 2016; Bun et al., 2018; Hardt and Ullman, 2014; Ullman et al., 2018; Shenfeld and Ligett, 2019; Jung et al., 2020; Shenfeld and Ligett, 2021). The connection between differential privacy and adaptive generalization also originated from Dwork et al. (2015b). Interestingly, this connection has recently been repurposed for different settings, such as adversarial streaming and dynamic algorithms (Hassidim et al., 2020; Attias et al., 2021; Kaplan et al., 2021; Beimel et al., 2021), and also in a preprint by the author, Uri Stemmer and Moshe Shechner, named *Streaming with advice* which is currently under review. We note that while this work by the author is relevant to the broader topic of privacy and compression, it was not included in this thesis due to its focus on a different aspect of the subject. Nonetheless, the findings have potential for further exploration and development and may be a valuable contribution to the field.

Chapter 3

Background and Preliminaries

This chapter details some preliminaries needed for a proper presentation of the results and discussion in the main part of the thesis. Some more specific preliminaries will be presented on the dedicated chapter in which they are relevant. An *instance space* is an abstract set $\mathcal{X}$ and a classifier is a binary function mapping points from the space to either zero or one $f : \mathcal{X} \to \{0, 1\}$.

3.1 Learning

3.1.1 The PAC Model and the VC Dimension

We use standard definitions from statistical learning theory. See, e.g., Shalev-Shwartz and Ben-David (2014).

The main goal of a classifier is to correctly predict the label of future points. Nevertheless, its performance in the current given sample serves as a starting point and a measure which might indicate its true performance.

Definition 3.1.1 (Sample error)**.** *The empirical error of a classifier h w.r.t. a labeled-sample $S \in (\mathcal{X} \times \{0, 1\})^n$ is defined as* $\mathrm{err}_S(h) = \frac{1}{n} \sum_{(x,y)\in S} \mathbb{1}[h(x) \neq y]$.

Definition 3.1.2 (True error)**.** *Given a data distribution μ and a hypothesis h we denote* $\mathrm{err}_\mu(h) = \mathbb{E}_{(x,y)\sim\mu}[1[h(x) \neq y]]$. *Given a data distribution μ, an algorithm $\mathcal{A}$, and sample size n, define*

$$\mathrm{err}_\mu(\mathcal{A}, n) = \mathop{\mathbb{E}}_{S\sim\mu^n} \mathop{\mathbb{E}}_{h_S\leftarrow\mathcal{A}(S)} [\mathrm{err}_\mu(h_S)] = \mathop{\mathbb{E}}_{S\sim\mu^n} \mathop{\mathbb{E}}_{h_S\leftarrow\mathcal{A}(S)} \mathop{\mathbb{E}}_{(x,y)\sim\mu} [1[h_S(x) \neq y]].$$

In words, $\mathrm{err}_\mu(\mathcal{A}, n)$ *is the expected loss of $\mathcal{A}$ given n labeled examples from μ.*

The main definition at the core of statistical learning literature is PAC learning. The definition aims to capture, in a quantifiable manner, what it essentially means *to learn* from the algorithmic perspective.

Definition 3.1.3 (PAC learnability Valiant (1984)). *Let $\alpha, \beta \in [0,1]$ and let $m \in \mathbb{N}$. An algorithm $\mathcal{A}$ is an (α, β, m)-PAC-learning algorithm for a class $\mathcal{C}$ if for every distribution μ over $\mathcal{X} \times \{0,1\}$ s.t. $\exists h^* \in \mathcal{C}$ with $\mathrm{err}_\mu(h^*) = 0$, it holds that $\Pr_{S \sim \mu^m}[\mathrm{err}_\mu(\mathcal{A}(S)) > \alpha] < \beta$. We refer to m as the* the sample complexity *of $\mathcal{A}$.*

In simple terms, a PAC learner is an algorithm that takes a labeled dataset as input and produces a classifier as output, which is guaranteed to predict the label of new instances with high probability. Since the seminal work by Vapnik and Chervonenkis, one main idea in the research of learnability of function classes is by trying to measure the complexity or the expressability of a function class. This is usually done using the combinatory idea of *shattering*.

Definition 3.1.4 (Shattering). *Let $\mathcal{C}$ be a class of functions over a domain $\mathcal{X}$. A set $S = (s_1, \ldots, s_k) \subseteq \mathcal{X}$ is said to be* shattered *by $\mathcal{C}$ if $|\{(f(s_1), \ldots, f(s_k)) : f \in \mathcal{C}\}| = 2^k$.*

Using the idea of shattering, Vapnik and Chervonenkis defined their famous notion of dimension for function classes.

Definition 3.1.5 (VC Dimension Vapnik and Chervonenkis (1971)). *The* VC dimension *of a class $\mathcal{C}$, denoted as $d_\mathcal{C}$, is the cardinality of the largest set shattered by $\mathcal{C}$. If $\mathcal{C}$ shatters sets of arbitrary large cardinality, then it is said that $d_\mathcal{C} = \infty$.*

When the roles of $\mathcal{X}$ and $\mathcal{C}$ are exchanged, i.e., an $x \in \mathcal{X}$ acts on $f \in \mathcal{C}$ via $x(f) = f(x)$, we refer to $\mathcal{X} = \mathcal{C}^*$ as the *dual* class of $\mathcal{C}$. Its VC-dimension is then $d^*_\mathcal{C} := d_{\mathcal{C}^*}$, and referred to as the *dual VC dimension.* Assouad (1983) showed that $d^*_\mathcal{C} \leq 2^{d_\mathcal{C}+1}$.

To illustrate this concept intuitively, consider the simple function class of one-dimensional intervals on the real line. In this case, the VC dimension can be easily grasped. Let $\mathcal{H}$ be the set of all intervals of the form $[a, b]$, where a and b are real numbers. An interval classifier labels by 1 any point which lays inside the interval and 0 for any other point. It is clear that if we take two distinct points x_1 and x_2 in the real line such that, w.l.o.g, $x_1 < x_2$, we can find intervals in $\mathcal{H}$ that can shatter them. specifically:

- The interval $[x_1 - 2, x_1 - 1]$ labels the pair by $(0, 0)$.
- The interval $[x_1 - 1, x_1]$ labels the pair by $(0, 1)$.
- The interval $[x_2 + 1, x_2 + 2]$ labels the pair by $(1, 1)$.
- The interval $[x_2, x_2 + 1]$ labels the pair by $(1, 0)$.

In other words the pair x_1 and x_2 can be labeled by any possible sequence of binary sequence of the same length. But, given a triplet $x_1 < x_2 < x_3$ there is not interval in $\mathcal{H}$ that can label them $(1, 0, 1)$. As the size of the smallest set that can be shattered by $\mathcal{H}$ is 2, the VC dimension of $\mathcal{H}$ is 2.

The main result that emerged from the VC-theory is that, in classes with a finite VC-dimension, it is possible to bound, with high probability, the gap between the empirical risk and the true risk of the classifier, also known as the generalization gap, meaning that such classes are PAC-learnable.

Theorem 3.1.6 (VC Dimension Generalization Bound Vapnik and Chervonenkis (1971); Blumer et al. (1989))**.** *Let $\mathcal{C}$ be a function-class and let μ be a probability measure over $\mathcal{X} \times \{0,1\}$. There exist a constant c s.t. for every $n \geq \frac{c}{\alpha}\big(d_{\mathcal{C}} \log(\frac{1}{\alpha}) + \log(\frac{1}{\beta})\big)$ every $\alpha, \beta > 0$, and every $f \in \mathcal{C}$ it holds that* $\Pr_{S\sim\mu^n}[\exists f \in \mathcal{C} : \mathrm{err}_\mu(f) \geq \alpha \wedge \mathrm{err}_S(f) \leq \alpha/10] \leq \beta$.

When dealing with real-valued functions, in contrast to the binary function of VC-theory, the notion found to be best suited as an analog is *fat-shattering.*

Definition 3.1.7 (Fat-Shattering (Alon et al., 1997))**.** *For $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ and $t > 0$, we say that $\mathcal{F}$ t-shatters a set $\{x_1, \ldots, x_k\} \subset \mathcal{X}$ if there is an $r \in \mathbb{R}^m$ such that for all $y \in \{-1,1\}^m$ there is an $f \in \mathcal{F}$ such that $\min_{i\in[k]} y_i(f(x_i) - r_i) \geq t$, when $[k] := \{1, \ldots, k\}$.*

That is, fat-shattering looks for expressability of the functions in the class as threshold functions with a given margin or scale. Again, this shattering notion is the basis for a dimension which characterizes learnability for real-valued function classes, just as the VC-dimension does for binary functions.

Definition 3.1.8 (Fat-Shattering dimension (Alon et al., 1997))**.** *The t-fat-shattering dimension $Fat_t(\mathcal{F})$ is the size of the largest t-shattered set (possibly ∞) . Again, the roles of $\mathcal{X}$ and $\mathcal{F}$ may be switched, in which case $\mathcal{X} = \mathcal{F}^*$ becomes the dual class of $\mathcal{F}$. Its t-fat-shattering dimension is then $Fat^*_t(\mathcal{F})$.* [1]

To demonstrate this concept intuitively, consider the class of affine functions on the real line. Given a pair of points x_1 and x_2 the pair is γ-shattered by the class for any given γ. A visual illustration of this is given in Figure 3.1 on which two example points x_1, x_2 are given and the 4 affine functions shatter the pair (each function labeled by the resulting pair's labeling) and the points r_1, r_2 witness the shattering as described in the above definition.

3.1.2 Compression Schemes

The formal notion of *compression scheme*, which was defined by Littlestone and Warmuth (1986) is the following: A *sample compression scheme* (κ, ρ) for a hypothesis class $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ is defined as follows. A *k-compression* function κ maps sequences $\mathcal{S} = ((x_1, y_1), \ldots, (x_m, y_m)) \in \bigcup_{\ell\geq 1}(\mathcal{X} \times \mathcal{Y})^\ell$ to elements in $\mathcal{K} = \bigcup_{\ell\leq k'}(\mathcal{X} \times \mathcal{Y})^\ell \times \bigcup_{\ell\leq k''} \{0,1\}^\ell$,

[1] It was proven by Kleer and Simon (2021) that it is impossible to obtain general bounds on the dual-fat-shattering dimension, similar to the one proven by Assouad (1983) for VC-dimension. Nevertheless, bounds do exist for some natural classes. To demonstrate this, in Section 4.2.3 we prove such bounds for the dual class of two fundamental classes: Lipschitz functions and bounded-variation functions.

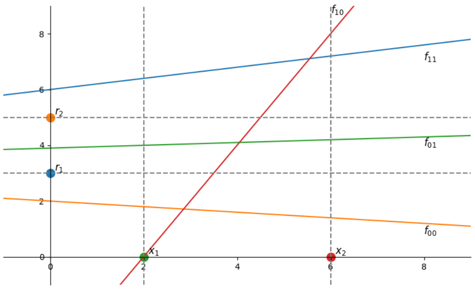

Figure 3.1: Fat-shattering illustration

where $\kappa(\mathcal{S}) \subseteq \mathcal{S}$ and $k' + k" \leq k$. A *reconstruction* is a function $\rho : \mathcal{K} \to \mathcal{Y}^{\mathcal{X}}$. We say that (κ, ρ) is a k-size sample compression scheme for $\mathcal{F}$ if κ is a k-compression and for all $h^* \in \mathcal{F}$ and all $S = ((x_1, h^*(x_1)), \ldots, (x_m, h^*(y_m)))$, we have $\hat{h} := \rho(\kappa(S))$ satisfies $\hat{h}(x_i) = h^*(x_i)$ for all $i \in [m]$.

For real-valued functions, there are several notions of compression-schemes. We say it is a *uniformly ε-approximate* compression scheme if

$$\max_{1 \leq i \leq m} |\hat{h}(x_i) - h^*(x_i)| \leq \varepsilon.$$

We note that a somewhat similar definition was proposed by David et al. (2016): Let $S = (x_1, y_i), \ldots, (x_m, y_m)$ be a tagged sample drawn i.i.d from some unknown distribution, an let $l : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ be some loss function. We say that (κ, ρ) is an *agnostic sample compression scheme* for $\mathcal{H}$ if, for every sample S, $f_S := \rho(\kappa(S))$, achieves $\mathcal{F}$-competitive empirical loss:

$$\frac{1}{m}\sum_{i=1}^{m} l(f_S(x_i), y_i) \leq \inf_{h \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} l(f_S(x_i), y_i),$$

and we say that it is ϵ-*Approximate Agnostic Sample Compression Scheme* for $\mathcal{H}$ if for every sample S

$$\frac{1}{m}\sum_{i=1}^{m} l(f_S(x_i), y_i) \leq \inf_{h \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} l(f_S(x_i), y_i) + \epsilon.$$

3.1.3 Bayes optimal Classifier and Plug-in Estimators

PAC learnability and VC-theory have been the main vanilla settings in the research community in the last decades. Nevertheless, there are other notions and regimes in the learning literature. The principal notion of learning, which predated the PAC-learning idea, is *consistency*. The main difference between consistency and PAC is that, while PAC aims for distribution-free and uniform rates for a given class of functions, consistency looks for distribution dependent rates without looking at a specific types of functions. This direction is most known in the context of analyzing classical learning algorithms such as the Nearest Neighbor rule, for which the arguments and tools from the PAC-model and VC-theory are irrelevant (Devroye and Györfi, 1985; Cover and Hart, 1967; Gottlieb et al., 2013).

Given a probability measure μ over $\mathcal{X} \times \{0,1\}$, we denote by η the *regression function*, also known as the *posteriori probability function*, defined as $\eta(x) = \Pr(y = 1 \mid x)$. The Bayes-optimal classifier is, then,

$$h^*(x) = \begin{cases} 1 & \eta(x) > 1/2 \\ 0 & \text{otherwise} \end{cases},$$

and we denote its error probability by $L^* = \mathrm{err}_\mu(h^*)$. It can be easily shown that h^* achieves the lowest error-rate among all the possible classifiers.

Definition 3.1.9. *An algorithm $\mathcal{A}$ is said to be* universally consistent *if for any distribution μ it holds that* $\lim_{n\to\infty} \{\mathrm{err}_\mu(\mathcal{A}, n)\} = L^*$.

As η is generally unknown, a possible approach for designing universally consistent algorithms is to create an approximation $\hat{\eta}$.

Definition 3.1.10. *Let $\hat{\eta} : \mathcal{X} \to [0,1]$ be any function. A* plug-in classification rule *w.r.t. $\hat{\eta}$ is defined as* $\hat{h}(x) = \begin{cases} 1 & \hat{\eta}(x) > 1/2 \\ 0 & \textit{otherwise} \end{cases}$.

The following theorem provides a bound on the error-rate of such a construction.

Theorem 3.1.11 (Devroye et al. (2013, Theorem 2.2))**.** *Let $\hat{\eta} : \mathcal{X} \to [0,1]$ be any function and let $\hat{h}$ be its corresponding plug-in classification rule. Then,* $\mathrm{err}_\mu(\hat{h}) - \mathrm{err}_\mu(h^*) \leq 2\,\mathbb{E}\left[|\eta(x) - \hat{\eta}(x)|\right]$, *where the expectation is over sampling x from the marginal distribution on unlabeled examples from μ.*

In Devroye et al. (2013), the above theorem is stated only for $\mathbb{R}^d$. The extension to arbitrary spaces is immediate; the details are given in Section 6.4 for completeness.

3.1.4 Density Estimation

In the problem of density estimation, given a sample containing n iid (unlabeled) elements from an (unknown) underlying distribution μ, our goal is to output a distribution $\hat{\mu}$ that is close (in L_1 distance) to the underlying distribution μ.

Definition 3.1.12. *An algorithm is said to be* universally consistent for density estima-

Algorithm 1 `Game`$(\mathcal{M}, k, \mathbb{A}, S)$

```
Inputs: Mechanism $\mathcal{M}$, interaction length $k$, adversary $\mathbb{A}$, dataset $S$.
The dataset $S$ is given to $\mathcal{M}$.
for $i \in [k]$ do
    $\mathbb{A}$ picks a query $q_i$.
    The query $q_i$ is given to $\mathcal{M}$.
    $\mathcal{M}$ outputs an answer $a_i$.
    The answer $a_i$ is given to $\mathbb{A}$.
```

tion in L_1 norm *if for any underlying distribution μ the following holds.*

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{S\sim\mu^n} \mathop{\mathbb{E}}_{\mu_n\leftarrow\mathcal{A}(S)} \|\mu-\mu_n\|_1 = \lim_{n\to\infty} \mathop{\mathbb{E}}_{S\sim\mu^n} \mathop{\mathbb{E}}_{\mu_n\leftarrow\mathcal{A}(S)} \int |\mu(x)-\mu_n(x)|dx = 0.$$

Our main metric for similarity between probability measures will be the *total variation distance.*

Definition 3.1.13 (Total Variation Distance)**.** *Given two measures ν and μ on the same space Ω, the* total variation distance *between them is defined as* $\|\nu-\mu\|_{\mathsf{TV}} := \sup_{A\subseteq\Omega}|\nu(A)-\mu(A)|$, *where the supremum is over the Borel sets of Ω. Equivalently,* $\|\nu-\mu\|_{\mathsf{TV}} = \frac{1}{2}\sum_{a\in\Omega}|\nu(a)-\mu(a)| = \frac{1}{2}\|\mu-\nu\|_{\ell_1}$.

3.1.5 Adaptive Data Analysis

The standard formulation of adaptive data analysis is defined as a game involving some (adversary) analyst and a query-answering mechanism. In the interests of this part of the thesis, queries are *statistical queries*, meaning they are functions of the form $q : \mathcal{X} \to [0,1]$. The goal of the mechanism is to make sure that the answers provided to the analyst are accurate w.r.t. the expected value of the corresponding queries over the underlying distribution. The idea is to formalize a utility notion that holds for *any* strategy of the data analyst. As a way of dealing with *worst-case analysts*, the analyst is assumed to be *adversarial* in that it tries to cause the mechanism to fail. If a mechanism can maintain utility against any such *adversarial* analyst, then it maintains utility against any analyst. This game is specified in Algorithm 1.

Definition 3.1.14 (Adaptive Empirical Accuracy)**.** *A mechanism $\mathcal{M}$ is* (α,β)-empirically accurate *for k rounds given a dataset of size n, if for every dataset S of size n and every adversary $\mathbb{A}$, it holds that* $\Pr_{\mathtt{Game}(\mathcal{M},k,\mathbb{A},S)}\left[\max_{i\in[k]}|q_i(S)-a_i| > \alpha\right] \le \beta$, *where* $q_i(S) := \frac{1}{|S|}\sum_{x\in S} q_i(x)$.

Definition 3.1.15 (Adaptive Statistical Accuracy)**.** *A mechanism $\mathcal{M}$ is* (α,β)-statistically accurate *for k rounds given n samples with Gibbs-dependence ψ, if for every distribution*

μ over n-tuples with Gibbs dependency ψ, and every adversary $\mathbb{A}$, it holds that

$$\Pr_{\substack{S\sim\mu \\ \mathtt{Game}(\mathcal{M},k,\mathbb{A},S)}}\left[\max_{i\in[k]}|q_i(\mu)-a_i|>\alpha\right]\leq\beta,$$

where $q_i(\mu):=\mathbb{E}_{T\sim\mu}[q_i(T)]=\mathbb{E}_{T\sim\mu}\left[\frac{1}{|T|}\sum_{x\in T}q_i(x)\right]$.

Remark 3.1.16. *The above definition is stated in general form, but in fact it is sufficient to show that a mechanism $\mathcal{M}$ exhibits the above guarantee for every* deterministic *adversary $\mathbb{A}$. The reason is that for a randomized adversary one can fix the adversary's random coins and use the* total probability law *in order to get the same result.*

3.2 Differential privacy

Differential privacy (Dwork et al., 2006b) is a mathematical definition of privacy that aims to enable statistical analyses of datasets, while providing strong guarantees that individual-level information does not leak. Informally, an algorithm that analyzes data satisfies differential privacy if it is robust in the sense that its outcome distribution does not depend "too much" on any single data point. Formally,

Definition 3.2.1 (Differential Privacy (Dwork et al., 2006b)). *A randomized algorithm $\mathcal{A}:\mathcal{X}^n\to\mathcal{Y}$ is (ε,δ)-differentially private if for every two datasets S,S' which differ on a single element and for any event F we have* $\Pr[\mathcal{A}(S)\in F]\leq e^{\varepsilon}\cdot\Pr[\mathcal{A}(S')\in F]+\delta$.

Two datasets $S,S'\in\mathcal{X}$ are said to be neighboring if they differ exactly on one element, formally, $d_H(S,S')=1$.

Definition 3.2.2 (Dwork et al. (2006c)). *Let f be a function mapping databases to real vectors. The* global sensitivity *of f is defined as* $GS(f)=\max_{d_H(S,S')=1}\left\|f(S)-f(S')\right\|_1$.

3.2.1 Properties and basic tools of Differential privacy

When dealing with the privacy of released information and statistics, we must account for future unknown attacks. This implies that every reasonable protection guarantee must not break under post-process of any kind. Hence, an important property of differential privacy is the following:

Theorem 3.2.3 (post-processing). *Let $\mathcal{A}:\mathcal{X}^n\to\mathcal{Y}$ be an (ε,δ)-differentially private algorithm, and let $f:\mathcal{Y}\to\mathcal{Y}'$ be any arbitrary mapping. Then $f\circ\mathcal{A}:\mathcal{X}^n\to\mathcal{Y}'$ is also (ε,δ)-differentially private.*

We will later make use of several differentially private algorithms combined, all applied upon the same data, in parallel or even in an adaptive-sequential manner. Even so, the combined process will remain privacy preserving (for different parameters) due to differential privacy been preserved under composition. Formally,

Theorem 3.2.4 (Advanced Composition, (Dwork et al., 2010a))**.** *Let* $0 < \delta', \varepsilon \leq 1$*, and let* $0 \leq \delta', \varepsilon \leq 1$*. An algorithm which permits k adaptive interactions with (various) mechanisms which preserve* (ε, δ)*-differential privacy, is then* $(\varepsilon', k\delta + \delta')$*-differentially private by itself, when* $\varepsilon' = \varepsilon\sqrt{sk \ln(1/\delta')} + 2k\varepsilon^2$*.*

We write $\mathrm{Lap}(\mu, b)$ to denote the *Laplace distribution* with mean μ and scale b. When the mean is zero, we will simply write $\mathrm{Lap}(b)$. The Laplace distribution lies at the core of many algorithms from the differential private literature, and most notably as used in the generic Laplace mechanism.

Definition 3.2.5 (The Laplace mechanism (Dwork et al., 2006c))**.** *Let* f *be a function mapping databases to vectors in* $\mathbb{R}^k$*, and let* ε *be a privacy parameter. Given an input database* S*, the Laplace mechanism outputs* $\mathbb{M}_\varepsilon(f, S) = f(S) + (a_1, \ldots, a_k)$*, when* a_i *are sampled i.i.d. from* $\mathrm{Lap}(GS(f)/\varepsilon)$*.*

Theorem 3.2.6 (Dwork et al. (2006c))**.** *The Laplace mechanism is* ε*-differentially private.*

One of the most basic and generic tools in the literature on differential privacy is the exponential mechanism of McSherry and Talwar (2007), defined as follows. Consider a "quality function" f that, given a dataset S, assigns every possible solution a (coming from some predefined solution-set A) a real valued number, identified as the "score" of the solution a w.r.t. the input dataset S. The goal is to privately identify a solution $a \in A$ with a high score $f(S, a)$. The mechanism itself simply picks a solution at random, where the probability for solution a is proportional to $e^{\varepsilon f(S,a)}$. As shown by McSherry and Talwar (2007) the exponential mechanism is $(\varepsilon, 0)$-differentially private.

3.2.2 Private Learning

Combining the ideas from privacy and learning research yields the natural idea of private learning. The vast majority of the research in recent years showing interested in this link revolves around the definition of Private-PAC learnability.

Definition 3.2.7 (Private-PAC learnability)**.** *An algorithm* $\mathcal{A}$ *is an* $(\alpha, \beta, \varepsilon, \delta, m)$*-PPAC learner for a class* $\mathcal{C}$ *if: (i)* $\mathcal{A}$ *is* (ε, δ)*-differentially private; and, (ii)* $\mathcal{A}$ *is an* (α, β, m)*-PAC learning algorithm for* $\mathcal{C}$*.*

In this thesis, we will also investigate the other notion of learnability in the light of privacy aspects. Namely, we define the natural definition combining privacy and consistency as follows:

Definition 3.2.8. *An algorithm* $\mathcal{A}$ *is said to be* (ε, δ)-Privately universally consistent*, or PUC for short, if it is* (ε, δ)*-differentially private and universally consistent.*

Remark 3.2.9. *Note that the utility requirement and the privacy requirement in the above definition are fundamentally different: Utility is only required to hold in the limiting regime when the sample size goes to infinity. In contrast, the privacy requirement is a worst-case kind of requirement that must hold for* any *two neighboring inputs, no matter how they were generated, even if they were not sampled from any distribution.*

3.3 Additional Notation

Given a number $\ell \in \mathbb{N}$ and a dataset S containing points from an ordered domain, we use $\min(S, \ell)$ (or $\max(S, \ell)$) to indicate the subset of ℓ minimal (or maximal) values within S. When S contains points from a d-dimentional domain, we write $\min_i(S, \ell)$ (or $\max_i(S, \ell)$) to denote the subset of ℓ minimal (or maximal) values within S w.r.t. the i^{th} axis.

Chapter 4

Real Valued Compression

In the study of machine learning theory, the standard definitions of learning, as PAC-learning for the binary case, require the learner to achieve arbitrary small accuracy. It is often difficult to be able to supply such a strong requirement, but nevertheless it may be much simpler, for a large set of problems, to construct learners which are somewhat better than a random labeling. Those learners are called *weak-learners* as opposed to the standard *strong-learners*. The idea of combining weak-learners in a way that produces a strong-learner is called Boosting.

As mentioned above (2.1), boosting has been shown to be a powerful technique for constructing compression schemes. In this chapter, we will first explore theoretical concepts regarding boosting and the notion of weak learning. We will then show how these ideas can be applied to the construction of compression schemes, and present our main results in this area.

4.1 Boosting Real-Valued Functions

The idea of leveraging or boosting weak-learners in order to achieve stronger learning guarantees started as a question proposed by Kearns, and reached a positive result in the seminal works by Schapire (1990) and Freund and Schapire (1997). The latter contained the well-known Adaboost algorithm, which is widely used in practice.

4.1.1 The MedBoost Algorithm

In the context of boosting for real-valued functions, the notion of an (η, γ)-weak hypothesis, defined above at Definition 1.1.3, plays a role analogous to the usual notion of a weak hypothesis in boosting for classification. Using this notion, the following boosting algorithm was proposed by Kégl (2003) as an extension to the classic Adaboost algorithm.

Intuitively, the MedBoost algorithm uses the weak learner in order iteratively produce

Algorithm 2 `MedBoost`$(\{(x_i, y_i)\}_{i\in[m]}, T, \gamma, \eta)$

```
Define $P_0$ as the uniform distribution over $\{1, \dots, n\}$
for $t = 0, \dots, T$ do
    Call weak learner to get $h_t$ and $(\eta/2, \gamma)$-weak hypothesis
     w.r.t. $(x_i, y_i) : i \sim P_t$ (repeat until it succeeds)
    for $i = 1, \dots, m$ do
        $\theta_i^{(t)} \leftarrow 1 - 2\mathbb{I}[|h_t(x_i) - y_i| > \eta/2]$
    $\alpha_t \leftarrow \frac{1}{2}\ln\left(\frac{(1-\gamma)\sum_{i=1}^m P_t(i)\mathbb{I}[\theta_i^{(t)}=1]}{(1+\gamma)\sum_{i=1}^m P_t(i)\mathbb{I}[\theta_i^{(t)}=-1]}\right)$
    if $\alpha_t = \infty$ then
        Return $T$ copies of $h_t$, and $(1, \dots, 1)$
    for $i = 1, \dots, m$ do
        $P_{t+1}(i) \leftarrow P_t(i)\frac{\exp\{-\alpha_t\theta_i^{(t)}\}}{\sum_{j=1}^m P_t(j)\exp\{-\alpha_t\theta_j^{(t)}\}}$
Return $(h_1, \dots, h_T)$ and $(\alpha_1, \dots, \alpha_T)$
```

weak hypotheses. It maintains a distribution on the initial sample which adaptively gives more weight to hard data points, enabling it to focus on those problematic points. Finally it assigns weight to each learner according to its performance. We can then use the series of learners and weights in order to construct a single strong hypothesis by taking the weighted median or the weighted quantile.

As it will be convenient for our later results, we expressed the algorithm output as a sequence of functions and weights; the boosting guarantee from Kégl (2003) applies to the weighted quantiles (and in particular, the weighted median) of these function values.

Here we define the weighted median as

$$\text{Median}(y_1, \dots, y_T; \alpha_1, \dots, \alpha_T) = \min\left\{ y_j : \frac{\sum_{t=1}^T \alpha_t \mathbb{I}[y_j < y_t]}{\sum_{t=1}^T \alpha_t} < \frac{1}{2} \right\}.$$

Also define the weighted *quantiles*, for $\gamma \in [0, 1/2]$, as

$$Q_\gamma^+(y_1, \dots, y_T; \alpha_1, \dots, \alpha_T) = \min\left\{ y_j : \frac{\sum_{t=1}^T \alpha_t \mathbb{I}[y_j < y_t]}{\sum_{t=1}^T \alpha_t} < \frac{1}{2} - \gamma \right\}$$
$$Q_\gamma^-(y_1, \dots, y_T; \alpha_1, \dots, \alpha_T) = \max\left\{ y_j : \frac{\sum_{t=1}^T \alpha_t \mathbb{I}[y_j > y_t]}{\sum_{t=1}^T \alpha_t} < \frac{1}{2} - \gamma \right\},$$

and abbreviate $Q_\gamma^+(x) = Q_\gamma^+(h_1(x), \dots, h_T(x); \alpha_1, \dots, \alpha_T)$ and $Q_\gamma^-(x) = Q_\gamma^-(h_1(x), \dots, h_T(x); \alpha_1, \dots, \alpha_T)$ for $h_1, \dots, h_T$ and $\alpha_1, \dots, \alpha_T$ the values returned by `MedBoost`.

After proposing the algorithm, Kégl (2003) proves the following:

Lemma 4.1.1. *(Kégl (2003)) For a training set* $Z = \{(x_1, y_1), \dots, (x_m, y_m)\}$ *of size* m*,* *the return values of* `MedBoost` *satisfy*

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\Big[\max\Big\{\Big|Q^{+}_{\gamma/2}(x_i)-y_i\Big|,\Big|Q^{-}_{\gamma/2}(x_i)-y_i\Big|\Big\}>\eta/2\Big]\leq\prod_{t=1}^{T}e^{\gamma\alpha_t}\sum_{i=1}^{m}P_t(i)e^{-\alpha_t\theta_i^{(t)}}.$$

We note that, in the special case of binary classification, `MedBoost` is closely related to the well-known AdaBoost algorithm (Freund and Schapire, 1997), and the above results correspond to a standard margin-based analysis of Schapire et al. (1998).

For our purposes, we will need the following corollary, which we prove below.

Corollary 4.1.2. *For* $T=\Theta\Big(\frac{1}{\gamma^2}\ln(m)\Big)$*, every* $i\in\{1,\ldots,m\}$ *has*

$$\max\Big\{\Big|Q^{+}_{\gamma/2}(x_i)-y_i\Big|,\Big|Q^{-}_{\gamma/2}(x_i)-y_i\Big|\Big\}\leq\eta/2.$$

In the proof, we use the following technical lemma

Lemma 4.1.3. *For* $x\geq\frac{1}{2}+\gamma$ *it holds that*

$$x^{1+\gamma}(1-x)^{1-\gamma}\leq\left(\frac{1}{2}+\gamma\right)^{1-\gamma}\left(\frac{1}{2}-\gamma\right)^{1+\gamma}.$$

Proof. Denote the left side as a function f and take log of f

$$\log(f(x))=(1+\gamma)\log(x)+(1-\gamma)\log(1-x).$$

Observe that the derivative with respect to x which is $(\log(f(x)))'=(1+\gamma)/x-(1-\gamma)/(1-x)$ is negative for $x\geq(1+\gamma)/2$. Since $x\geq\frac{1}{2}+\gamma>(1+\gamma)/2$ this condition holds. So the function $\log(f(a)):=\log\left(a^{1+\gamma}(1-a)^{1-\gamma}\right)$ is monotonically decreasing and by that also f itself is monotonically decreasing. Hence

$$x^{1+\gamma}(1-x)^{1-\gamma}\leq(\frac{1}{2}+\gamma)^{1+\gamma}(1-\frac{1}{2}+\gamma)^{1-\gamma}.$$

□

Proof of Corollary 4.1.2. By the definition of α_t we know that

$$e^{\alpha_t}=\left(\frac{(1-\gamma)\sum_{\theta_i(t)=1}P_t(i)}{(1+\gamma)\sum_{\theta_i(t)=-1}P_t(i)}\right)^{1/2}.$$

Split the sum within the RHS into $\{i \mid \theta_i(t) = 1\}$ and $\{i \mid \theta_i(t) = -1\}$ to get that

$$\begin{aligned}
&\prod_{t=1}^{T} e^{\gamma\alpha_t} \sum_{i=1}^{m} P_t(i) e^{-\alpha_t \theta_i(t)} a \\
&= \prod_{t=1}^{T} e^{\gamma\alpha_t} \left[\sum_{\theta_i(t)=1} P_t(i) e^{-\alpha_t} + \sum_{\theta_i(t)=-1} P_t(i) e^{\alpha_t} \right] \\
&= \prod_{t=1}^{T} e^{\gamma\alpha_t} \left[e^{-\alpha_t} \sum_{\theta_i(t)=1} P_t(i) + e^{\alpha_t} \sum_{\theta_i(t)=-1} P_t(i) \right] \\
&= \prod_{t=1}^{T} \left[e^{-\alpha_t(1-\gamma)} \sum_{\theta_i(t)=1} P_t(i) + e^{\alpha_t(1+\gamma)} \sum_{\theta_i(t)=-1} P_t(i) \right].
\end{aligned}$$

Plug-in e^{α_t}

$$\begin{aligned}
&= \prod_{t=1}^{T} \left[\left(\frac{(1+\gamma) \sum_{\theta_i(t)=-1} P_t(i)}{(1-\gamma) \sum_{\theta_i(t)=1} P_t(i)} \right)^{\frac{1-\gamma}{2}} \sum_{\theta_i(t)=1} P_t(i) \right. \\
&\quad \left. + \left(\frac{(1-\gamma) \sum_{\theta_i(t)=1} P_t(i)}{(1+\gamma) \sum_{\theta_i(t)=-1} P_t(i)} \right)^{\frac{1+\gamma}{2}} \sum_{\theta_i(t)=-1} P_t(i) \right] \\
&= \prod_{t=1}^{T} \left[\left(\sum_{\theta_i(t)=1} P_t(i) \right)^{\frac{1+\gamma}{2}} \left(\sum_{\theta_i(t)=-1} P_t(i) \right)^{\frac{1-\gamma}{2}} \left(\frac{1+\gamma}{1-\gamma} \right)^{\frac{1-\gamma}{2}} \right. \\
&\quad \left. + \left(\sum_{\theta_i(t)=1} P_t(i) \right)^{\frac{1+\gamma}{2}} \left(\sum_{\theta_i(t)=-1} P_t(i) \right)^{\frac{1-\gamma}{2}} \left(\frac{1-\gamma}{1+\gamma} \right)^{\frac{1+\gamma}{2}} \right].
\end{aligned}$$

By the (ε, γ)-weak-learning guarantee we know that

$$\sum_{\theta_i(t)=1} P_t(i) \geq \frac{1}{2} + \gamma$$

and

$$\sum_{\theta_i(t)=-1} P_t(i) < \frac{1}{2} - \gamma$$

and by Lemma 4.1.3

$$\begin{aligned}
&\leq \prod_{t=1}^{T} \left[\left(\frac{1+\gamma}{1-\gamma} \right)^{\frac{1-\gamma}{2}} + \left(\frac{1-\gamma}{1+\gamma} \right)^{\frac{1+\gamma}{2}} \right] \left(\frac{1}{2} + \gamma \right)^{\frac{1-\gamma}{2}} \left(\frac{1}{2} - \gamma \right)^{\frac{1+\gamma}{2}} \\
&= \prod_{t=1}^{T} \frac{1}{2} \left(\frac{1-\gamma}{1+\gamma} \right)^{\frac{\gamma}{2}} \left(\frac{1+2\gamma}{1-2\gamma} \right)^{\frac{\gamma}{2}} (1-4\gamma^2)^{1/2} \left(\left(\frac{1+\gamma}{1-\gamma} \right)^{1/2} + \left(\frac{1-\gamma}{1+\gamma} \right)^{1/2} \right),
\end{aligned}$$

noting that for every $\gamma \in (0, 1/3)$.

$$\frac{1}{2}\left(\frac{1-\gamma}{1+\gamma}\right)^{\frac{\gamma}{2}}\left(\frac{1+2\gamma}{1-2\gamma}\right)^{\frac{\gamma}{2}}(1-4\gamma^2)^{1/2}\left(\left(\frac{1+\gamma}{1-\gamma}\right)^{1/2}+\left(\frac{1-\gamma}{1+\gamma}\right)^{1/2}\right) < e^{-\gamma^2/4},$$

we get that

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\Big[\max\Big\{\Big|Q^{+}_{\gamma/2}(x_i)-y_i\Big|,\Big|Q^{-}_{\gamma/2}(x_i)-y_i\Big|\Big\} > \eta/2\Big]$$
$$\leq \prod_{t=1}^{T} e^{\gamma\alpha_t}\sum_{i=1}^{m}P_t(i)e^{-\alpha_t\theta_i^{(t)}} < e^{-T\gamma^2/4}.$$

Finally for $T = \frac{4}{\gamma^2}\ln(m)$ the last bound is equal to $\frac{1}{m}$ and hence the corollary holds. □

4.1.2 The Sample Complexity of Weak Learning

This section reveals our intent in choosing this notion of weak hypothesis, rather than using, for example, an ε-good strong learner under absolute loss. In addition to being a strong enough notion for boosting to work, we show here that it is also a weak enough notion for the sample complexity of weak learning to be of reasonable size: namely, a size quantified by the fat-shattering dimension. This result is also relevant to an open question posed by Simon (1997), which we discuss on Subsection 4.1.2.

The Notion of "Weak Learning"

As mentioned above, the notion of a *weak learner* for learning real-valued functions must be formulated carefully. The naïve thought that we could take any learner guaranteeing, for example, absolute loss at most $\frac{1}{2}-\gamma$, is known to be not strong enough to enable boosting to ε loss. However, if we make the requirement too strong, such as in Freund and Schapire (1997) for `AdaBoost.R`, then the sample complexity of weak learning will be so high that weak learners cannot be expected to exist for large classes of functions.

Starting with Kearns and Schapire, the notion of weak learning was tied to the notion of PAC learnability. Weak learning is, as one may expect, the weak version of PAC learning. This relation meant that weak-learning also was defined using a loss-function and a (weak) upper-bound on the loss of the resulting hypothesis, namely a fixed, yet bounded away from 1/2, bound on the expected loss.

Normally when extending the PAC paradigm to the real-valued/continuous case, we just replace the loss-function. Thus, we get the following

Definition 4.1.4 ("Standard"-Weak-Hypothesis). *For $\gamma \in [0, 1/2]$, we say that $f : \mathcal{X} \to$*

$\mathbb{R}$ *is an an* γ-weak hypothesis *(with respect to distribution* D *and target* $f^* \in \mathcal{F}$*) if*

$$\mathop{\mathbb{E}}_{X\sim D}\left[l(f_S(x), f^*(x))\right] \le \frac{1}{2} - \gamma.$$

Unfortunately, this extension for the problem of boosting essentially fails. Duffy and Helmbold (2002, Remark 2.1) point out that, using this notion of weak learning, one can not guarantee that using the method of modifying the distribution over the sample will force the learner to establish a good hypothesis. This is due to the fact that, unlike the binary-case, the error can be spread evenly over all the sample, meaning that the error remains the same regardless of the distribution on the sample. This might result in the learner outputting the same hypothesis on each iteration, and hence not improving the error of the final output regressor. Some lines of work, including Freund and Schapire's `AdaBoost.R`, used more complex boosting ideas in order to bypass this problem. Those algorithms are either problematic in their runtime, or, as in the `AdaBoost.R` case, based on weak learners whose sample complexity depends on the Pseudo-dimension of the class [1], which tends to be so high that weak learners cannot be expected to exist for large classes of functions.

For this reason, we use a different notion. Recall the definition

Definition 4.1.5 ((η, γ)-Weak-Hypothesis)**.** *For* $\eta \in [0,1]$ *and* $\gamma \in [0, 1/2]$*, we say that* $f : \mathcal{X} \to \mathbb{R}$ *is an an* (η, γ)-weak hypothesis *(with respect to distribution* D *and target* $f^* \in \mathcal{F}$*) if*

$$\Pr_{X\sim D}(|f(X) - f^*(X)| > \eta) \le \frac{1}{2} - \gamma.$$

The (η, γ)-weak-learner, which has appeared, among other works, in Anthony et al. (1996); Simon (1997); Avnimelech and Intrator (1999); Kégl (2003), gets around this difficulty by demanding a bound on the measure of the points in which the hypothesis has "big" local error. Furthermore, this notion was in fact proven useful in various, quite simple, boosting mechanisms, but, to our knowledge, provable general constructions of such learners have been lacking. Note that, as in other definitions of weak-learning, this definition also uses a "strong" definition of learning, which was proposed by Simon.

Definition 4.1.6 ((ε, γ)-good-model)**.** *For* $\varepsilon, \eta \in [0,1]$ *and* $\gamma \in [0, 1/2]$*, we say that* $f : \mathcal{X} \to \mathbb{R}$ *is an an* (ε, γ)-good model *(with respect to distribution* D *and target* $f^* \in \mathcal{F}$*) if*

$$\Pr_{X\sim D}(|f(X) - f^*(X)| > \eta) \le \varepsilon.$$

and a $\mathcal{A}$ is γ-learner if *for every* ε, δ and sample S of size $m = m(\varepsilon, \delta)$, with probability at least $1 - \delta$, $f = \mathcal{A}(S)$ is a (ε, γ)-good-model. So (η, γ)-weak-learner is simply a γ-learner

[1] A different combinatorial dimension for real-valued function classes, first defined by Pollard.

with the error parameter ε fixed, and bounded away from 1/2.

Although there exist several uses of this type of "weak-learning", to our knowledge there exist no provable constructions of such algorithms. We now present a provable and very natural, namely ERM based, (η, γ)-learner. From this result, we are also able to construct our (η, γ)-weak-learner, which was used by our compression-boosting mechanism.

Upper Bound on The Sample Complexity of (ε, γ)-Good-Learning

The following result is stated in the notion of the more general case of (ε, γ)-good-model. in order to apply it to our boosting mechanism, we later fix the error parameter ε as previously discussed, which then yields an Upper Bound on the sample complexity of (ε, γ)-weak-learner.

Define $\rho_\eta(f, g) = P_{2m}(x : |f(x) - g(x)| > \eta)$, where P_{2m} is the empirical measure induced by $X_1, \ldots, X_{2m}$ iid P-distributed random variables (the m data points and m ghost points). Define $N_\eta(\beta)$ as the β-covering numbers of $\mathcal{F}$ under the ρ_η pseudo-metric.

Theorem 4.1.7. *Fix any $\eta, \beta \in (0,1)$, $\alpha \in [0,1)$, and $m \in \mathbb{N}$. For $X_1, \ldots, X_m$ iid P-distributed, with probability at least $1 - \mathbb{E}\big[N_{\eta(1-\alpha)/2}(\beta/8)\big]\, 2e^{-m\beta/96}$, every $f \in \mathcal{F}$ with $\max_{1 \le i \le m} |f(X_i) - f^*(X_i)| \le \alpha\eta$ satisfies $P(x : |f(x) - f^*(x)| > \eta) \le \beta$.*

Proof. This proof roughly follows the usual symmetrization argument for uniform convergence (Vapnik and Červonenkis, 1971; Haussler, 1992), with a few important modifications to account for this (η, β)-based criterion. If $\mathbb{E}\big[N_{\eta(1-\alpha)/2}(\beta/8)\big]$ is infinite, then the result is trivial, so let us suppose it is finite for the remainder of the proof. Similarly, if $m < 8/\beta$, then $2e^{-m\beta/96} > 1$ and hence the claim trivially holds, so let us suppose $m \ge 8/\beta$ for the remainder of the proof. Without loss of generality, suppose $f^*(x) = 0$ everywhere and every $f \in \mathcal{F}$ is non-negative (otherwise subtract f^* from every $f \in \mathcal{F}$ and redefine $\mathcal{F}$ as the absolute values of the differences; note that this transformation does not increase the value of $N_{\eta(1-\alpha)/2}(\beta/8)$ since applying this transformation to the original $N_{\eta(1-\alpha)/2}(\beta/8)$ functions remains a cover).

Let $X_1, \ldots, X_{2m}$ be iid P-distributed. Denote by P_m the empirical measure induced by $X_1, \ldots, X_m$, and by P'_m the empirical measure induced by $X_{m+1}, \ldots, X_{2m}$. We have

$$\begin{aligned}
&\Pr(\exists f \in \mathcal{F} : P'_m(x : f(x) > \eta) > \beta/2 \text{ and } P_m(x : f(x) \le \alpha\eta) = 1) \\
&\ge \Pr\Big(\exists f \in \mathcal{F} : P(x : f(x) > \eta) > \beta \text{ and } P_m(x : f(x) \le \alpha\eta) = 1 \\
&\qquad\quad \text{and } P'_m(x : f(x) > \eta) > \beta/2\Big).
\end{aligned}$$

Denote by A_m the event that there exists $f \in \mathcal{F}$ satisfying $P(x : f(x) > \eta) > \beta$ and $P_m(x : f(x) \le \alpha\eta) = 1$, and on this event let $\tilde{f}$ denote such an $f \in \mathcal{F}$ (chosen solely based on $X_1, \ldots, X_m$); when A_m fails to hold, take $\tilde{f}$ to be some arbitrary-fixed element of $\mathcal{F}$. Then the expression on the right-hand side above is at least as large as

$$\Pr\left(A_m \text{ and } P'_m(x : \tilde{f}(x) > \eta) > \beta/2\right),$$

and noting that the event A_m is independent of $X_{m+1}, \ldots, X_{2m}$, this equals

$$\mathbb{E}\Big[\mathbb{I}_{A_m} \cdot \Pr\Big(P'_m(x : \tilde{f}(x) > \eta) > \beta/2 \Big| X_1, \ldots, X_m\Big)\Big]. \tag{4.1}$$

Then note that for any $f \in \mathcal{F}$ with $P(x : f(x) > \eta) > \beta$, a Chernoff bound implies

$$\begin{aligned}&\Pr\left(P'_m(x : f(x) > \eta) > \beta/2\right)\\ &= 1 - \Pr\left(P'_m(x : f(x) > \eta) \leq \beta/2\right) \geq 1 - \exp\{-m\beta/8\} \geq \frac{1}{2},\end{aligned}$$

where we have used the assumption that $m \geq \frac{8}{\beta}$ here. In particular, this implies that the expression in (4.1) is no smaller than $\frac{1}{2}\Pr(A_m)$. Altogether, we have established that

$$\begin{aligned}&\Pr(\exists f \in \mathcal{F} : P(x : f(x) > \eta) > \beta \text{ and } P_m(x : f(x) \leq \alpha\eta) = 1)\\ &\leq 2\Pr(\exists f \in \mathcal{F} : P'_m(x : f(x) > \eta) > \beta/2 \text{ and } P_m(x : f(x) \leq \alpha\eta) = 1).\end{aligned} \tag{4.2}$$

Now let $\sigma(1), \ldots, \sigma(m)$ be independent random variables (also independent of the data), with $\sigma(i) \sim$ *Uniform*$(\{i, m+i\})$, and denote $\sigma(m+i)$ as the sole element of $\{i, m+i\} \setminus \{\sigma(i)\}$ for each $i \leq m$. Also denote by $P_{m,\sigma}$ the empirical measure induced by $X_{\sigma(1)}, \ldots, X_{\sigma(m)}$, and by $P'_{m,\sigma}$ the empirical measure induced by $X_{\sigma(m+1)}, \ldots, X_{\sigma(2m)}$. By exchangeability of $(X_1, \ldots, X_{2m})$, the right-hand side of (4.2) is equal

$$\Pr\big(\exists f \in \mathcal{F} : P'_{m,\sigma}(x : f(x) > \eta) > \beta/2 \text{ and } P_{m,\sigma}(x : f(x) \leq \alpha\eta) = 1\big).$$

Now let $\hat{\mathcal{F}} \subseteq \mathcal{F}$ be a minimal subset of $\mathcal{F}$ such that $\max\limits_{f \in \mathcal{F}} \min\limits_{\hat{f} \in \hat{\mathcal{F}}} \rho_{\eta(1-\alpha)/2}(\hat{f}, f) \leq \beta/8$. The size of $\hat{\mathcal{F}}$ is at most $N_{\eta(1-\alpha)/2}(\beta/8)$, which is finite almost surely (since we have assumed above that its expectation is finite). Then note that (denoting by $X_{[2m]} = (X_1, \ldots, X_{2m})$) the above expression is at most

$$\begin{aligned}&\Pr\Big(\exists f \in \hat{\mathcal{F}} : P'_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) > (3/8)\beta \text{ and } P_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) \leq \beta/8\Big)\\ &\leq \mathbb{E}\Bigg[N_{\eta(1-\alpha)/2}(\beta/8) \max_{f \in \hat{\mathcal{F}}} \Pr\big(P'_{m,\sigma}(x{:}f(x){>}\eta(1+\alpha)/2) > (3/8)\beta\\ &\qquad\qquad \text{and } P_{m,\sigma}(x{:}f(x){>}\eta(1+\alpha)/2) \leq \beta/8 \big| X_{[2m]}\big)\Bigg].\end{aligned} \tag{4.3}$$

Then note that for any $f \in \mathcal{F}$, we have almost surely

$$\begin{aligned}
&\Pr\bigl(P'_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) > (3/8)\beta \text{ and } P_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) \le \beta/8 \big| X_{[2m]}\bigr) \\
&\le \Pr\bigl(P_{2m}(x : f(x) > \eta(1+\alpha)/2) > (3/16)\beta \text{ and } P_{m,\sigma}(x : f(x) > \eta(1+\alpha)/2) \le \beta/8 \big| X_{[2m]}\bigr) \\
&\le \exp\{-m\beta/96\},
\end{aligned}$$

where the last inequality is by a Chernoff bound, which (as noted by Hoeffding (1963)) remains valid even when sampling without replacement. Together with (4.2) and (4.3), we have that

$$\begin{aligned}
&\Pr(\exists f \in \mathcal{F} : P(x : f(x) > \eta) > \beta \text{ and } P_m(x : f(x) \le \alpha\eta) = 1) \\
&\le 2\,\mathbb{E}\bigl[N_{\eta(1-\alpha)/2}(\beta/8)\bigr]\, e^{-m\beta/96}.
\end{aligned}$$

□

Lemma 4.1.8. *There exist universal numerical constants $c, c' \in (0,\infty)$ such that $\forall \eta, \beta \in (0,1)$,*

$$N_\eta(\beta) \le \left(\frac{2}{\eta\beta}\right)^{cFat_{c'\eta\beta}(\mathcal{F})},$$

where $Fat.(\cdot)$ is the fat-shattering dimension.

Proof. Mendelson and Vershynin (2003, Theorem 1) establishes that the $\eta\beta$-covering number of $\mathcal{F}$ under the $L_2(P_{2m})$ pseudo-metric is at most

$$\left(\frac{2}{\eta\beta}\right)^{cFat_{c'\eta\beta}(\mathcal{F})} \tag{4.4}$$

for some universal numerical constants $c, c' \in (0,\infty)$. Then note that for any $f, g \in \mathcal{F}$, Markov's and Jensen's inequalities imply $\rho_\eta(f,g) \le \frac{1}{\eta}\|f-g\|_{L_1(P_{2m})} \le \frac{1}{\eta}\|f-g\|_{L_2(P_{2m})}$. Thus, any $\eta\beta$-cover of $\mathcal{F}$ under $L_2(P_{2m})$ is also a β-cover of $\mathcal{F}$ under ρ_η, and therefore (4.4) is also a bound on $N_\eta(\beta)$. □

Combining the above two results yields the following theorem.

Theorem 4.1.9. *For some universal numerical constants $c_1, c_2, c_3 \in (0,\infty)$, for any $\eta, \delta, \beta \in (0,1)$ and $\alpha \in [0,1)$, letting $X_1, \ldots, X_m$ be iid P-distributed, where*

$$m = \left\lceil \frac{c_1}{\beta}\left(Fat_{c_2\eta\beta(1-\alpha)}(\mathcal{F})\ln\left(\frac{c_3}{\eta\beta(1-\alpha)}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right\rceil,$$

with probability at least $1-\delta$, every $f \in \mathcal{F}$ with $\max_{i\in[m]}|f(X_i) - f^*(X_i)| \le \alpha\eta$ satisfies $P(x : |f(x) - f^*(x)| > \eta) \le \beta$.

Proof. The result follows immediately from combining Theorem 4.1.7 and Lemma 4.1.8.

□

In particular, the specific case of weak-learners, as stated in Theorem 1.1.4, follows immediately from this result by taking $\beta = 1/2 - \gamma$ and $\alpha = \gamma/2$.

Tightness of The Upper Bound

To discuss tightness of Theorem 4.1.9, we note that in addition to the definition of a (β, η)-*good model* Simon (1997) also proved the following lower bound

Theorem 4.1.10 (Simon (1997)). *Let A be an algorithm which learns function class F with an (β, η)-good model*

1. *If F is nontrivial* [2] *, $\beta < 1/2$ and $\eta < \Delta(F)/2$. then A needs $\Omega(\ln(1/\delta)/\beta)$ examples.*
2. *If $\beta \leq 1/8$, $0 < \delta \leq 1/100$. then A needs $\Omega((d_F^N(\eta) - 1)/\beta)$ examples.*

When $\Delta(F) = sup\{\|g - f\|_\infty \mid \exists x \in X : f(x) = g(x)\}$.

Combining the two, we get that a sample complexity lower bound for the same criterion of

$$\Omega\left(\frac{d_F^N(c\eta)}{\beta} + \frac{1}{\beta}\log\frac{1}{\delta}\right),$$

where $d_F^N(\cdot)$ is a quantity somewhat smaller than the fat-shattering dimension, essentially representing a fat Natarajan dimension.

Simon showed that this lower bound is tight and placed an open question

Open Problem: For every function class F there exists an algorithm A which learns F with an (β, η)-good model, using

$$\mathcal{O}\left(\frac{d_F^N(\eta)}{\beta} + \frac{1}{\beta}\ln(1/\delta)\right)$$

examples.

Thus, aside from the differences in the complexity measure (and a logarithmic factor), we establish an upper bound of a similar form to Simon's lower bound, hence making significant progress towards solving Simon's open question.

4.2 From Boosting to Compression

Generally, our strategy for converting the boosting algorithm `MedBoost` into a sample compression scheme of smaller size follows a strategy of Moran and Yehudayoff for binary classification, based on arguing that because the ensemble makes its predictions with a *margin* (corresponding to the results on *quantiles* in Corollary 4.1.2), it is possible

[2] Meaning: there exist $f, g \in F$ which are not pairwise disjoin, namely $\exists x \in X : f(x) = g(x)$.

to recover the same proximity guarantees for the predictions while using only a smaller *subset* of the functions from the original ensemble. Specifically, we use the following general *sparsification* strategy.

For $\alpha_1, \dots, \alpha_T \in [0,1]$ with $\sum_{t=1}^{T} \alpha_t = 1$, denote by $Cat(\alpha_1, \dots, \alpha_T)$ the *categorical distribution*: i.e., the discrete probability distribution on $\{1, \dots, T\}$ with probability mass α_t on t.

Algorithm 3 `Sparsify`$(\{(x_i, y_i)\}_{i\in[m]}, \gamma, T, n)$

```
Run MedBoost$(\{(x_i, y_i)\}_{i\in[m]}, T, \gamma, \eta)$
Let $h_1, \dots, h_T$ and $\alpha_1, \dots, \alpha_T$ be its return values
Denote $\alpha'_t = \alpha_t / \sum_{t'=1}^{T} \alpha_{t'}$ for each $t \in [T]$
repeat
    Sample $(J_1, \dots, J_n) \sim Cat(\alpha'_1, \dots, \alpha'_T)^n$
    Let $F = \{h_{J_1}, \dots, h_{J_n}\}$
until $\max_{1\le i\le m} |\{f \in F : |f(x_i) - y_i| > \eta\}| < n/2$
Return $F$
```

The Sparsify procedure samples a subset of the hypotheses produced by the MedBoost algorithm with probability which is proportional to the outputted weight corresponding to each hypothesis. The procedure keeps sampling until it finds a subset s.t. most of the hypotheses in the subset have low empirical error (with respect to the points in the sample).

For any values $a_1, \dots, a_n$, denote the (unweighted) median

$$Med(a_1, \dots, a_n) = Median(a_1, \dots, a_n; 1, \dots, 1).$$

Our intention in discussing the above algorithm is to argue that, for a sufficiently large choice of n, the above procedure returns a set $\{f_1, \dots, f_n\}$ such that

$$\forall i \in [m], |Med(f_1(x_i), \dots, f_n(x_i)) - y_i| \le \eta.$$

We analyze this strategy separately for binary classification and real-valued functions, since the argument in the binary case is much simpler (and demonstrates more directly the connection to the original argument of Moran and Yehudayoff), and also because we arrive at a tighter result for binary functions than for real-valued functions.

4.2.1 Binary Classification

We begin with the simple observation about binary classification (i.e., where the functions in $\mathcal{F}$ all map into $\{0,1\}$). The technique here is quite simple, and follows a similar line of reasoning to the original argument of Moran and Yehudayoff. The argument for real-valued functions below will diverge from this argument in several important ways, but the high level ideas remain the same.

The compression function is essentially the one introduced by Moran and Yehudayoff, except applied to the classifiers produced by the above `Sparsify` procedure, rather than a set of functions selected by a minimax distribution over all classifiers produced by $O(d_{\mathcal{F}})$ samples each. The weak hypotheses in `MedBoost` for binary classification can be obtained using samples of size $O(d_{\mathcal{F}})$. Thus, if the `Sparsify` procedure is successful in finding n such classifiers whose median predictions are within η of the target y_i values for all i, then we may encode these n classifiers as a compression set, consisting of the set of $k = O(nd_{\mathcal{F}})$ samples used to train these classifiers, together with $k \log k$ extra bits to encode the order of the samples.[3] To obtain Theorem 1.1.1, it then suffices to argue that $n = \Theta(d^*_{\mathcal{F}})$ is a sufficient value. The proof follows.

Proof of Theorem 1.1.1. Recall that $d^*_{\mathcal{F}}$ bounds the VC dimension of the class of sets $\{\{h_t : t \leq T, h_t(x_i) = 1\} : 1 \leq i \leq m\}$. Thus for the iid samples $h_{J_1}, \ldots, h_{J_n}$ obtained in `Sparsify`, for $n = 64(2309 + 16d^*_{\mathcal{F}}) > \frac{2304+16d^*_{\mathcal{F}}+\log(2)}{1/8}$, by the VC uniform convergence inequality of Vapnik and Červonenkis (1971), with probability at least 1/2 we get that

$$\max_{1\leq i\leq m} \left| \left(\frac{1}{n} \sum_{j=1}^{n} h_{J_j}(x_i) \right) - \left(\sum_{t=1}^{T} \alpha' h_t(x_i) \right) \right| < 1/8.$$

In particular, if we choose $\gamma = 1/8$, $\eta = 1$, and $T = \Theta(\log(m))$ appropriately, then Corollary 4.1.2 implies that every $y_i = \mathbb{I}\left[\sum_{t=1}^{T} \alpha' h_t(x_i) \geq 1/2\right]$ and $\left|\frac{1}{2} - \sum_{t=1}^{T} \alpha' h_t(x_i)\right| \geq 1/8$ so that the above event would imply every

$$y_i = \mathbb{I}\left[\frac{1}{n} \sum_{j=1}^{n} h_{J_j}(x_i) \geq 1/2 \right] = Med(h_{J_1}(x_i), \ldots, h_{J_n}(x_i)).$$

Note that the `Sparsify` algorithm need only try this sampling $\log_2(1/\delta)$ times to find such a set of n functions. Combined with the description above (from Moran and Yehudayoff, 2016) of how to encode this collection of h_{J_i} functions as a sample compression set plus side information, this completes the construction of the sample compression scheme. □

4.2.2 Real-Valued Functions

Next, we turn to the general case of real-valued functions (where the functions in $\mathcal{F}$ may generally map into $[0, 1]$). We have the following result, which says that the `Sparsify` procedure can reduce the ensemble of functions from one with $T = O(\log(m)/\gamma^2)$ functions in it, down to one with a number of functions *independent of m.*

[3] In fact, $k \log n$ bits would suffice if the weak learner is permutation-invariant in its data set.

Theorem 4.2.1. *Choosing*

$$n = \Theta\bigg(\frac{1}{\gamma^2} Fat^*_{c\eta}(\mathcal{F}) \log^2(Fat^*_{c\eta}(\mathcal{F})/\eta)\bigg)$$

suffices for the **`Sparsify`** *procedure to return* $\{f_1, \dots, f_n\}$ *with*

$$\max_{1 \le i \le m} |Med(f_1(x_i), \dots, f_n(x_i)) - y_i| \le \eta.$$

Proof. Recall from Corollary 4.1.2 that `MedBoost` returns functions $h_1, \dots, h_T \in \mathcal{F}$ and $\alpha_1, \dots, \alpha_T \ge 0$ such that $\forall i \in \{1, \dots, m\}$,

$$\max\bigg\{\Big|Q^+_{\gamma/2}(x_i) - y_i\Big|, \Big|Q^-_{\gamma/2}(x_i) - y_i\Big|\bigg\} \le \eta/2,$$

where $\{(x_i, y_i)\}_{i=1}^m$ is the training data set. We use this property to sparsify $h_1, \dots, h_T$ from $T = O(\log(m)/\gamma^2)$ down to k elements, where k will depend on η, γ, and the dual fat-shattering dimension of $\mathcal{F}$ (actually, just of $H = \{h_1, \dots, h_T\} \subseteq \mathcal{F}$) but **not** sample size m.

Letting $\alpha'_j = \alpha_j / \sum_{t=1}^T \alpha_t$ for each $j \le T$, we will sample k hypotheses $\Big\{\tilde{h}_1, \dots, \tilde{h}_k\Big\} =: \tilde{H} \subseteq H$ with each $\tilde{h}_i = h_{J_i}$, where $(J_1, \dots, J_k) \sim Cat(\alpha'_1, \dots, \alpha'_T)^k$ as in `Sparsify`. Define a function $\hat{h}(x) = Med(\tilde{h}_1(x), \dots, \tilde{h}_k(x))$. We claim that for any fixed $i \in [m]$, with high probability

$$|\hat{h}(x_i) - f^*(x_i)| \le \eta/2. \tag{4.5}$$

Indeed, partition the indices $[T]$ into the disjoint sets

$$\begin{aligned} L(x) &= \left\{j \in [T] : h_j(x) < Q^-_\gamma(h_1(x), \dots, h_T(x); \alpha_1, \dots, \alpha_T)\right\}, \\ M(x) &= \left\{j \in [T] : Q^-_\gamma(h_1(x), ..., h_T(x); \alpha_1, ..., \alpha_T) \le h_j(x) \le Q^+_\gamma(h_1(x), ..., h_T(x); \alpha_1, ..., \alpha_T)\right\}, \\ R(x) &= \left\{j \in [T] : h_j(x) > Q^+_\gamma(h_1(x), \dots, h_T(x); \alpha_1, \dots, \alpha_T)\right\}. \end{aligned}$$

Then the only way (4.5) can fail is if half or more indices $J_1, \dots, J_k$ sampled fall into $R(x_i)$ — or if half or more fall into $L(x_i)$. Since the sampling distribution puts mass less than $1/2 - \gamma$ on each of $R(x_i)$ and $L(x_i)$, Chernoff's bound puts an upper estimate of $\exp(-2k\gamma^2)$ on either event. Hence,

$$\mathbb{P}\Big(|\hat{h}(x_i) - f^*(x_i)| > \eta/2\Big) \le 2\exp(-2k\gamma^2). \tag{4.6}$$

Next, our goal is to ensure that with high probability, (4.5) holds simultaneously for all $i \in [m]$. Define the map $\boldsymbol{\xi} : [m] \to \mathbb{R}^k$ by $\boldsymbol{\xi}(i) = (\tilde{h}_1(x_i), \dots, \tilde{h}_k(x_i))$. Let $G \subseteq [m]$ be a minimal subset of $[m]$ such that

$$\max_{i \in [m]} \min_{j \in G} \|\boldsymbol{\xi}(i) - \boldsymbol{\xi}(j)\|_\infty \le \eta/2.$$

This is just a minimal ℓ_∞ covering of $[m]$. Then

$$\begin{aligned}
&\Pr\left(\exists i \in [m] : |Med(\boldsymbol{\xi}(i)) - f^*(x_i)| > \eta\right) \leq \\
&\sum_{j\in G} \Pr\left(\exists i : |Med(\boldsymbol{\xi}(i)) - f^*(x_i)| > \eta, \|\boldsymbol{\xi}(i) - \boldsymbol{\xi}(j)\|_\infty \leq \eta/2\right) \leq \\
&\sum_{j\in G} \Pr\left(|Med(\boldsymbol{\xi}(j)) - f^*(x_j)| > \eta/2\right) \leq 2N_\infty([m], \eta/2)\exp(-2k\gamma^2),
\end{aligned}$$

where $N_\infty([m], \eta/2)$ is the $\eta/2$-covering number (under ℓ_∞) of $[m]$, and we used the fact that

$$|Med(\boldsymbol{\xi}(i)) - Med(\boldsymbol{\xi}(j))| \leq \|\boldsymbol{\xi}(i) - \boldsymbol{\xi}(j)\|_\infty .$$

Finally, to bound $N_\infty([m], \eta/2)$, note that $\boldsymbol{\xi}$ embeds $[m]$ into the dual class $\mathcal{F}^*$. Thus, we may apply the bound in (Rudelson and Vershynin, 2006, Display (1.4)):

$$\log N_\infty([m], \eta/2) \leq C Fat^*_{c\eta}(\mathcal{F}) \log^2(k/\eta),$$

where C, c are universal constants and $Fat^*_{\cdot}(\mathcal{F})$ is the dual fat-shattering dimension of $\mathcal{F}$. It now only remains to choose a k that makes $\exp\left(C Fat^*_{c\eta}(\mathcal{F}) \log^2(k/\eta) - 2k\gamma^2\right)$ as small as desired. □

To establish Theorem 1.1.2, we use the weak learner from above, with the booster `MedBoost` from Kégl, and then apply the `Sparsify` procedure. Combining the corresponding theorems, together with the same technique for converting to a compression scheme discussed above for classification (i.e. encoding the functions with the set of training examples from which they were obtained, plus a string of bits to denote from which examples, and in what order, each weak hypothesis was obtained), this immediately yields the result claimed in Theorem 1.1.2, which represents our main new result for sample compression of general families of real-valued functions.

4.2.3 Examples

As an example of the generality and usefulness of the above schemes, we present two interesting and efficient compression schemes that can then be derived. The main technical result needed in order to apply our method to those cases was to find and prove the dual Fat-Shattering dimension of the function-classes at hand, a problem which is not trivial most of the time, requiring using tools from various domains. Leveraging novel and relatively new algorithmic results from learning theory yields the final desired compression-schemes.

Sample compression for BV functions

The function class $\mathrm{BV}(v)$ consists of all $f : [0,1] \to \mathbb{R}$ for which

$$V(f) := \sup_{n\in\mathbb{N}} \sup_{0=x_0<x_1<\ldots<x_n=1} \sum_{i=1}^{n-1} |f(x_{i+1}) - f(x_i)| \le v.$$

It is known (Anthony and Bartlett, 1999, Theorem 11.12) that $Fat_t(\mathrm{BV}(v)) = 1 + \lfloor v/(2t) \rfloor$. In Theorem 4.2.3 below, we show that the dual class has $Fat_t^*(\mathrm{BV}(v)) = \Theta(\log(v/t))$. Long (2004) presented an efficient, proper, consistent learner for the class $\mathcal{F} = \mathrm{BV}(1)$ with range restricted to $[0,1]$, with sample complexity $m_{\mathcal{F}}(\varepsilon,\delta) = O(\frac{1}{\varepsilon}\log\frac{1}{\delta})$. Combined with Theorem 1.1.2, this yields

Corollary 4.2.2. *Let $\mathcal{F} = \mathrm{BV}(1) \cap [0,1]^{[0,1]}$ be the class $f : [0,1] \to [0,1]$ with $V(f) \le 1$. Then the proper, consistent learner $\mathcal{L}$ of Long (2004), with target generalization error ε, admits a sample compression scheme of size $O(k\log k)$, where*

$$k = \mathcal{O}\left(\frac{1}{\varepsilon}\log^2\frac{1}{\varepsilon}\cdot\log\left(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}\right)\right).$$

The compression set is computable in expected runtime

$$\mathcal{O}\left(n\frac{1}{\varepsilon^{3.38}}\log^{3.38}\frac{1}{\varepsilon}\left(\log n + \log\frac{1}{\varepsilon}\log\left(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}\right)\right)\right).$$

The remainder of this section is devoted to proving

Theorem 4.2.3. *For $\mathcal{F} = \mathrm{BV}(v)$ and $t < v$, we have $Fat_t^*(\mathcal{F}) = \Theta(\log(v/t))$.*

First, we define some preliminary notions:

Definition 4.2.4. *For a binary $m \times n$ matrix M, define*

$$\begin{aligned} V(M,i) &:= \sum_{j=1}^{m} \mathbb{I}[M_{j,i} \ne M_{j+1,i}], \\ G(M) &:= \sum_{i=1}^{n} V(M,i), \\ V(M) &:= \max_{i\in[n]} V(M,i). \end{aligned}$$

Lemma 4.2.5. *Let M be a binary $2^n \times n$ matrix. If for each $b \in \{0,1\}^n$ there is a row j in M equal to b, then*

$$V(M) \ge \frac{2^n}{n}.$$

In particular, for at least one row i, we have $V(M,i) \ge 2^n/n$.

Proof. Let M be a $2^n \times n$ binary such that for each $b \in \{0,1\}^n$ there is a row j in M equal to b. Given M's dimensions, every $b \in \{0,1\}^n$ appears exactly in one row of M, and hence the minimal Hamming distance between two rows is 1. Summing over the $2^n - 1$ adjacent row pairs, we have

$$G(M) = \sum_{i=1}^{n} V(M,i) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{I}[M_{j,i} \neq M_{j+1,i}] \geq 2^n - 1,$$

which averages to

$$\frac{1}{n}\sum_{i=1}^{n} V(M,i) = \frac{G(M)}{n} \geq \frac{2^n-1}{n}.$$

By the pigeon-hole principle, there must be a row $j \in [n]$ for which $V(M,i) \geq \frac{2^n-1}{n}$, which implies $V(M) \geq \frac{2^n-1}{n}$. □

We split the proof of Theorem 4.2.3 into two estimates:

Lemma 4.2.6. *For $\mathcal{F} = \mathrm{BV}(v)$ and $t < v$, $Fat_t^*(\mathcal{F}) \leq 2\log_2(v/t)$.*

Lemma 4.2.7. *For $\mathcal{F} = \mathrm{BV}(v)$ and $4t < v$, $Fat_t^*(\mathcal{F}) \geq \lfloor \log_2(v/t) \rfloor$.*

Proof of Lemma 4.2.6. Let $\{f_1, \ldots, f_n\} \subset \mathcal{F}$ be a set of functions that are t-shattered by $\mathcal{F}^*$. In other words, there is an $r \in \mathbb{R}^n$ such that for each $b \in \{0,1\}^n$ there is an $x_b \in \mathcal{F}^*$ such that

$$\forall i \in [n], x_b(f_i) \begin{cases} \geq r_i + t, & b_i = 1 \\ \leq r_i - t, & b_i = 0 \end{cases}.$$

Let us order the x_bs by magnitude $x_1 < x_2 < \ldots < x_{2^n}$, denoting this sequence by $(x_i)_{i=1}^{2^n}$. Let $M \in \{0,1\}^{2^n \times n}$ be a matrix whose ith row is b_j, the latter ordered arbitrarily.

By Lemma 4.2.5, there is $i \in [n]$ s.t.

$$\sum_{j=1}^{2^n} \mathbb{I}[M(j,i) \neq M(j+1,i)] \geq \frac{2^n}{n}.$$

Note that if $M(j,i) \neq M(j+1,i)$ shattering implies that

$$x_j(f_i) \geq r_i + t \text{ and } x_{j+1}(f_i) \leq r_i - t$$

or

$$x_j(f_i) \leq r_i - t \text{ and } x_{j+1}(f_i) \geq r_i + t;$$

either way,

$$|f_i(x_j) - f_i(x_{j+1})| = |x_j(f_i) - x_{j+1}(f_i)| \geq 2t.$$

So for the function f_i, we have

$$\sum_{j=1}^{2^n}|f_i(x_j)-f_i(x_{j+1})| = \sum_{j=1}^{2^n}|x_j(f_i)-x_{j+1}(f_i)| \geq \sum_{j=1}^{2^n}\mathbb{I}[b_{j_i} \neq b_{j+1_i} \cdot 2t \geq \frac{2^n}{n}\cdot 2t.$$

As $\{x_j\}_{j=1}^{2^n}$ is a partition of $[0,1]$ we get

$$v \geq \sum_{j=1}^{2^n}|f_i(x_j)-f_i(x_{j+1})| \geq \frac{t2^{n+1}}{n} \geq t2^{n/2}$$

and hence

$$v/t \geq 2^{n/2}$$
$$\Rightarrow 2\log_2(v/t) \geq n.$$

□

Proof of Lemma 4.2.7. We construct a set of $n = \lfloor \log_2(v/t) \rfloor$ functions that are t-shattered by $\mathcal{F}^*$. First, we build a balanced Gray code (Flahive and Bose, 2007) with n bits, which we arrange into the rows of M. Divide the unit interval into 2^n segments and define, for each $j \in [2^n]$,

$$x_j := \frac{j}{2^n}.$$

Define the functions $f_1, \ldots, , f_{\lfloor \log_2(v/t) \rfloor}$ as follows:

$$f_i(x_j) = \begin{cases} t, & M(j,i)=1 \\ -t, & M(j,i)=0 \end{cases}.$$

We claim that each $f_i \in \mathcal{F}$. Since M is balanced Gray code,

$$V(M) = \frac{2^n}{n} \leq \frac{v}{t\log_2(v/t)} \leq \frac{v}{2t}.$$

Hence, for each f_i, we have

$$V(f_i) \leq 2tV(M,i) \leq 2t\frac{v}{2t} = v.$$

Next, we show that this set is shattered by $\mathcal{F}^*$. Fix the trivial offest $r_1 = ... = r_n = 0$ For every $b \in \{0,1\}^n$ there is a $j \in [2^n]$ s.t. $b = b_i$. By construction, for every $i \in [n]$, we have

$$x_j(f_i) = f_i(x_j) = \begin{cases} t \geq r_i + t, & M(j,i)=1 \\ -t \leq r_i - t, & M(j,i)=0 \end{cases}.$$

□

Sample compression for nearest-neighbor regression

Let $(\mathcal{X},\rho)$ be a metric space and define, for $L \geq 0$, the collection $\mathcal{F}_L$ of all $f : \mathcal{X} \to [0,1]$ satisfying

$$|f(x) - f(x')| \leq L\rho(x,x');$$

these are the L-Lipschitz functions. Gottlieb et al. (2017b) showed that

$$Fat_t\left(\mathcal{F}_L\right) = O\left(\lceil Ldiam\left(\mathcal{X}\right)/t\rceil^{ddim(\mathcal{X})}\right),$$

where $diam\left(\mathcal{X}\right)$ is the diameter and $ddim$ is the *doubling dimension* (see Definition 6.3.1. The proof is achieved via a packing argument, which also shows that the estimate is tight. Below we show that $Fat^*_t\left(\mathcal{F}\right) = \Theta(\log\left(M(\mathcal{X},2t/L)\right))$, where $M(\mathcal{X},\cdot)$ is the packing number of $(\mathcal{X},\rho)$. Applying this to the efficient nearest-neighbor regressor[4] of Gottlieb et al. (2017a), we obtain

Corollary 4.2.8. *Let $(\mathcal{X},\rho)$ be a metric space with hypothesis class $\mathcal{F}_L$, and let $\mathcal{L}$ be a consistent, proper learner for $\mathcal{F}_L$ with target generalization error ε. Then $\mathcal{L}$ admits a compression scheme of size $O(k\log k)$, where*

$$k = \mathcal{O}\left(D(\varepsilon)\log\frac{1}{\varepsilon}\cdot\log D(\varepsilon)\log\left(\frac{1}{\varepsilon}\log D(\varepsilon)\right)\right)$$

and

$$D(\varepsilon) = \left\lceil\frac{Ldiam\left(\mathcal{X}\right)}{\varepsilon}\right\rceil^{ddim(\mathcal{X})}.$$

We now prove our estimate on the dual fat-shattering dimension of $\mathcal{F}$:

Lemma 4.2.9. *For $\mathcal{F} = \mathcal{F}_L$, $Fat^*_t\left(\mathcal{F}\right) \leq \log_2\left(\mathcal{M}(\mathcal{X},2t/L)\right)$.*

Proof. Let $\{f_1,\ldots,f_n\} \subset \mathcal{F}_L$ a set that is t-shattered by $\mathcal{F}^*_L$. For $b \neq b' \in \{0,1\}^n$, let i be the first index for which $b_i \neq b'_i$, say, $b_i = 1 \neq 0 = b'$. By shattering, there are points $x_b, x_{b'} \in \mathcal{F}^*_L$ such that $x_b(f_i) \geq r_i + t$ and $x_{b'}(f_i) \leq r_i - t$, whence

$$f_i(x_b) - f_i(x_{b'}) \geq 2t$$

and

$$L\rho(x_b,x_{b'}) \geq f_i(x_b) - f_i(x_{b'}) \geq 2t.$$

It follows that for $b \neq b' \in \{0,1\}^n$, we have $\rho(x_b,x_{b'}) \geq 2t/L$. Denoting by $M(\mathcal{X},\varepsilon)$ the

[4] In fact, the technical machinery in Gottlieb et al. (2017a) was aimed at achieving *approximate* Lipschitz-extension, so as to gain a considerable runtime speedup. An *exact* Lipschitz extension is much simpler to achieve. It is more computationally costly but still polynomial-time in sample size.

ε-packing number of $\mathcal{X}$, we get

$$2^n = |\{x_b \mid b \in \{0,1\}^n\}| \leq \mathcal{M}(\mathcal{X}, 2t/L).$$

□

Lemma 4.2.10. *For $\mathcal{F} = \mathcal{F}_L$ and $t < L$, $Fat_t^*(\mathcal{F}) \geq \log_2(\mathcal{M}(\mathcal{X}, 2t/L))$.*

Proof. Let $S = \{x_1, ..., x_m\} \subseteq \mathcal{X}$ be a maximal $2t/L$-packing of $\mathcal{X}$. Suppose that $c : S \to \{0,1\}^{\lfloor \log_2 m \rfloor}$ is one-to-one. Define the set of function $F = \{f_1, \ldots, f_{\lfloor \log_2(m) \rfloor}\} \subseteq \mathcal{F}_L$ by

$$f_i(x_j) = \begin{cases} t, & c(x_j)_i = 1 \\ -t, & c(x_j)_i = 0 \end{cases}.$$

For every $f \in F$ and every two points $x, x' \in S$ it holds that

$$|f(x) - f(x')| \leq 2t = L \cdot 2t/L \leq L\rho(x, x').$$

This set of functions is t-shattered by S and is of size $\lfloor \log_2 m \rfloor = \lfloor \log_2(\mathcal{M}(\mathcal{X}, 2t/L)) \rfloor$. □

Chapter 5

Privately Leading Axis Aligned Rectangles

We now investigate the problem of privately learning the class of axis-aligned rectangles, defined as follows.

Definition 5.0.1 (Axis Aligned Rectangles)**.** *Let* $\mathcal{X} = \{0, \ldots, X\}^d$ *be a finite discrete d-dimensional domain. Every* $p = (p_1, \ldots, p_d) \in \mathcal{X}$*, induces a classifier* $h_p : \mathcal{X} \to \{0, 1\}$ *s.t for a given input* $x \in \mathcal{X}$ *we have*

$$h_p(x) = \begin{cases} 1, & \forall i \in [d] : x_i \leq p_i \\ 0, & \textit{otherwise} \end{cases}$$

Define the class of all axis-aligned and origin-placed rectangles *as* $REC_d^X = \{h_p : p \in \mathcal{X}\}$*.*

We focus on the *realizable* setting in which for a class $\mathcal{C}$ of potential classifiers, there exist some $h^* \in REC_d^X$, s.t $\mathrm{err}_\mu(h^*) = 0$.

Without the privacy requirement, learning axis-aligned rectangles is a simple task. As it is described in classical books such as Kearns and Vazirani (1997) and Shalev-Shwartz and Ben-David (2014) we can consider the simple algorithm which removes all the negative labeled points and picks the tightest rectangle containing the remaining points. It can also be seen from a compression scheme perspective, by applying the following simple algorithm:

1. For every axis $i \in [d]$
   (a) Remove all the negative-labeled points.
   (b) Project all the remaining points onto the i^{th} axis.
   (c) Pick a point a_i from the lowest valued projected point.
   (d) Pick a point b_i from the highest valued projected point.

2. Return the axis-aligned rectangle defined by the intervals $[a_i, b_i]$ at the different axes.

This is a valid compression scheme which yields a simple and efficient PAC-learner for the class.

As for privately learning, this idea is no longer valid due to its high sensitivity. Moreover, as mentioned in 1.2.1, it is not a matter of any specific algorithm or technique, but under privacy constraints this learning task is inherently harder, and the sample complexity must depend on the domain size at least by a $\log^*$ factor.

5.1 Baseline

The best prior result for the problem, which we use as our baseline, obtains sample complexity $\tilde{O}\big(d^{1.5} \cdot (\log^* |\mathcal{X}|)^{1.5}\big)$. This baseline algorithm is based on a reduction to (privately) solving the following problem, called the *interior point problem.*

Definition 5.1.1 (Bun et al. 2015)**.** *An algorithm $\mathcal{A}$ is said to solve the* Interior Point Problem *for domain $\mathcal{X}$ with failure probability β and sample complexity n, if for every $m \geq n$ and every database S containing m elements from $\mathcal{X}$ it holds that:* $\Pr[\min(S) \leq \mathcal{A}(S) \leq \max(S)] \geq 1 - \beta$.

That is, given a database S containing (unlabeled) elements from a (one dimensional) grid $\mathcal{X}$, the interior point problem asks for an element of $\mathcal{X}$ between the smallest and largest elements in S. The baseline we consider for privately learning axis-aligned rectangles is as follows: Suppose we have a differentially private algorithm $\mathcal{B}$ for the interior point problem over domain $\mathcal{X}$ with sample complexity n (let us ignore the failure probability for simplicity). We now use $\mathcal{B}$ to construct the following algorithm $\mathcal{A}$ that takes a database S containing labeled elements from $\mathcal{X}^d$. For simplicity, we assume that S contains "enough" positive elements, as otherwise we could simply return the all-zero hypothesis.

1. For every axis $i \in [d]$:
   (a) Project the positive points in S onto the ith axis.
   (b) Let A_i and B_i denote the smallest n and the largest n (projected) points, without their labels.
   (c) Let $a_i \leftarrow \mathcal{B}(A_i)$ and $b_i \leftarrow \mathcal{B}(B_i)$.
2. Return the axis-aligned rectangle defined by the intervals $[a_i, b_i]$ at the different axes.

Now, recall that each application of algorithm $\mathcal{B}$ returns an *interior point* of its input points. Hence, for every axis i, it holds that the interval $[a_i, b_i]$ contains (the projection) of all, but at most $2n$, of the positive examples in the ith axis. Therefore, the rectangle returned in Step 2 contains all, but at most $2nd$, of the positive points (and it does not contain any of the negative points, because this rectangle is *contained* inside the target rectangle). So algorithm $\mathcal{A}$ errs on at most $2nd$ of its input points.

Assuming that $|S| \gg 2nd$, we therefore get that algorithm $\mathcal{A}$ has small empirical error. As the VC dimension of the class of axis-aligned rectangles is $O(d)$, this means that algorithm $\mathcal{A}$ is a PAC learner for this class with sample complexity $O(nd)$. The issue here is that algorithm $\mathcal{A}$ executes algorithm $\mathcal{B}$ many times (specifically, $2d$ times). Hence, in order to argue that $\mathcal{A}$ is (ε, δ)-differentially private, standard composition theorems for differential privacy require each execution of algorithm $\mathcal{B}$ to be done with a privacy parameter of $\approx \varepsilon/\sqrt{2d}$. This, in turn, would mean that n (the sample complexity of algorithm $\mathcal{B}$) needs to be at least $\sqrt{2d}$, which means that algorithm $\mathcal{A}$ errs on $2nd \approx d^{1.5}$ input points, which translates to sample complexity of $|S| \gg d^{1.5}$.

The takeaway from this baseline learner is that in order to reduce the sample complexity to be linear in d, we want to bypass the costs incurred from composition. That is, we still want to follow the same strategy (apply algorithm $\mathcal{B}$ twice on every axis), but we want to do it without appealing to composition arguments in the privacy analysis. This was the starting point of our thought process.

5.2 The Algorithm

We now briefly survey two intuitive attempts that fail to achieve this, but are useful for the presentation of our final algorithm.

Failed Attempt #1. As before, let $\mathcal{B}$ denote an algorithm for the interior point problem over domain $\mathcal{X}$ with sample complexity n. Consider the following modification to algorithm $\mathcal{A}$ (marked in red). As before, algorithm $\mathcal{A}$ takes a database S containing labeled elements from $\mathcal{X}^d$, where we assume for simplicity that S contains "enough" positive elements.

1. For every axis $i \in [d]$:
   (a) Project the positive points in S onto the ith axis.
   (b) Let A_i and B_i denote the smallest n and the largest n (projected) points, without their labels.
   (c) Let $a_i \leftarrow \mathcal{B}(A_i)$ and $b_i \leftarrow \mathcal{B}(B_i)$.
   (d) Delete from S all points (with their labels) that correspond to A_i and B_i.
2. Return the axis-aligned rectangle defined by the intervals $[a_i, b_i]$ at the different axes.

The (incorrect) idea here is that by adding Step 1d we make sure that each datapoint from S is "used only once", and hence we do not need to pay in composition. In other words, the hope is that if every execution of algorithm $\mathcal{B}$ is done with a privacy parameter

ε, then the whole construction would satisfy differential privacy with parameter $O(\varepsilon)$.

The failure point of this idea is that by deleting *one* point from the data, we can create a "domino effect" that affects (one by one) many of the sets A_i, B_i throughout the execution. Specifically, consider two neighboring datasets S and $S' = S \cup \{(x', y')\}$ for some labeled point $(x', y') \in X^d \times \{0, 1\}$. Suppose that during the execution on S' it holds that $x' \in A_1$. So the additional point x' participates "only" in the first iteration of the algorithm, and gets deleted afterwards. However, since the size of the sets A_i, B_i is fixed, during the execution on S (without the point x') it holds that *a different point* z gets included in A_1 instead of x', and this point z is then deleted from S (but it is not deleted from S' during the execution on S'). Therefore, also during the second iteration we have that S and S' are not identical (they still differ on one point) and this domino effect can continue throughout the execution. That is, a single data point can affect many of the executions of $\mathcal{B}$, and we would still need to pay in composition to argue privacy.

Failed Attempt #2. In order to overcome the previous issue, one might try the following variant of algorithm $\mathcal{A}$.

1. For every axis $i \in [d]$:
   - (a) Project the positive points in S onto the ith axis.
   - (b) Let $\text{size}_{A_i} = 2n + \text{Noise}$ and let $\text{size}_{B_i} = 2n + \text{Noise}$.
   - (c) Let A_i and B_i denote the smallest size_{A_i} and the largest size_{B_i} (projected) points, respectively, without their labels.
   - (d) Let $a_i \leftarrow \mathcal{B}(A_i)$ and $b_i \leftarrow \mathcal{B}(B_i)$.
   - (e) Delete from S all points (with their labels) that correspond to A_i and B_i.
2. Return the axis-aligned rectangle defined by the intersection of the intervals $[a_i, b_i]$ at the different axes.

The idea now is that the noises we add to the sizes of the A_i's and the B_i's would "mask" the domino effect mentioned above. Specifically, the hope is as follows: Consider the execution of (the modified) algorithm $\mathcal{A}$ on S and on $S' = S \cup \{(x', y')\}$, and let i be the first axis such that $x' \in A_i \cup B_i$ during the execution on S'. Suppose w.l.o.g. that $x' \in B_i$. Now, the hope is that if during the execution on S we have that the noisy size_{B_i} is smaller by 1 than its value during the execution on S', then this eliminates the domino effect we mentioned, because we would not need to add another point instead of x'. Specifically, during time i, the point x' gets deleted from S', and every other point is either deleted from both S, S' or not deleted from any of them. So after time i the two executions continue identically. Thus, the hope is that by correctly "synchronizing" the noises between the two executions (such that only the size of the "correct" set gets modified by 1) we can make sure that only one application of $\mathcal{B}$ is affected (in the last example – only the execution of $\mathcal{B}(B_i)$ is affected), and so we would not need to apply composition arguments.

Although very convincing, this idea fails. The (very subtle) issue here is that it is not clear how to synchronize the noises between the two executions. To see the problem, let us try to formalize the above argument.

Fix two neighboring databases S and $S' = S \cup \{(x', y')\}$. Let us write A_i, B_i and A'_i, B'_i to denote these sets during the executions on S and on S', respectively. Aiming to synchronize the two executions, let us define a mapping $\pi : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ from noise vectors during the execution on S' to noise vectors during the execution on S (determining the values of $\text{size}_{A_1}, \text{size}_{B_1}, \dots, \text{size}_{A_d}, \text{size}_{B_d}$), such that throughout the execution we have that $A_i = A'_i$ and $B_i = B'_i$ for all i except for a single pair, say $B_j \neq B'_j$, of neighboring sets.

The straightforward way for defining such a mapping is as follows: Let j be the first time step in which the additional point x' gets included in a set A'_j or B'_j, and say that it is included in B'_j. Then the mapping would be to reduce (by 1) the value of size_{B_j} (the noisy size of B_j during the execution on S). This would indeed make sure that, conditioned on the noise vectors v' and $v = \pi(v')$, the two executions differ only in a single application of the interior point algorithm $\mathcal{B}$, and hence the outcome distribution of these two (conditioned) executions is very similar (in the sense of differential privacy). That is, for any noise vector v and any event F,

$$\Pr[\mathcal{A}(S') \in F | v] \leq e^{\varepsilon} \cdot \Pr[\mathcal{A}(S) \in F | \pi(v)] + \delta.$$

Furthermore, (assuming an appropriate noise distribution) we can make sure that the probability of obtaining the noise vectors v and $\pi(v)$ is similar, with densities differing by at most an e^{ε} factor (as is standard in the literature of differential privacy). Therefore, *had the mapping π we defined was a bijection*, for any event F we would have that

$$\begin{aligned}
\Pr[\mathcal{A}(S') \in F] &= \sum_{v} \Pr[v] \cdot \Pr[\mathcal{A}(S') \in F | v] \\
&\leq \sum_{v} e^{\varepsilon} \cdot \Pr[\pi(v)] \cdot (e^{\varepsilon} \cdot \Pr[\mathcal{A}(S) \in F | \pi(v)] + \delta) \\
&= \sum_{\pi(v)} e^{\varepsilon} \cdot \Pr[\pi(v)] \cdot (e^{\varepsilon} \cdot \Pr[\mathcal{A}(S) \in F | \pi(v)] + \delta) \\
&= e^{2\varepsilon} \cdot \Pr[\mathcal{A}(S) \in F] + e^{\varepsilon} \cdot \delta,
\end{aligned}$$

which would be great. Unfortunately, the mapping π we defined is *not* a bijection, and hence the second-to-last equality above is incorrect. To see that it is not a bijection, suppose that $d = 2$ and consider a database S containing the following positively labeled points: Many copies of the point $(0, 0)$, as well as 10 copies of the point $(1, 0)$ and 10 copies of the point $(0, 1)$. The neighboring database S' contains, in addition to all these points, also the point $\left(\frac{1}{2}, \frac{1}{2}\right)$. Now suppose that during the execution on S' we have that $|B'_1| = 5$ and $|B'_2| = 4$. That is, the additional point is included in B'_1. During the execution on S we therefore reduce (by 1) the size of B_1 and so $|B_1| = |B_2| = 4$. Now suppose that during the execution on S' we have that $|B'_1| = 4$ and $|B'_2| = 5$. Here, during the execution on S we reduce the size of B_2 and so, again, $|B_1| = |B_2| = 4$. This shows that the mapping π we defined is *not* a bijection. In general, in d dimensions, it is only a d-to-1 mapping, which would would break our analysis completely (it will not allow us to avoid the extra factor in d).

5.2.1 Our Solution - A Technical Overview

We now present a simplified version of our construction that overcomes the challenges mentioned above. We stress that the actual construction is somewhat different. Consider the following (simplified) algorithm.

1. For every axis $i \in [d]$:
   (a) Project the positive points in S onto the ith axis.
   (b) Let $\text{size}_{A_i} = 100n + \text{Noise}$ and let $\text{size}_{B_i} = 100n + \text{Noise}$, where the standard deviation of these noises is, say, $10n$.
   (c) Let A_i and B_i denote the smallest size_{A_i} and the largest size_{B_i} (projected) points, respectively, without their labels.
   (d) Let $A_i^{\text{inner}} \subseteq A_i$ be the n *largest* points in A_i. Similarly, let $B_i^{\text{inner}} \subseteq B_i$ be the n *smallest points in* B_i.
   (e) Let $a_i \leftarrow \mathcal{B}(A_i^{\text{inner}})$ and $b_i \leftarrow \mathcal{B}(B_i^{\text{inner}})$.
   (f) Delete from S all points (with their labels) whose projection onto the ith is *not* in the interval $[a_i, b_i]$.
2. Return the axis-aligned rectangle defined by the intersection of the intervals $[a_i, b_i]$ at the different axes.

There are *two* important modifications here. First, we still add noise to the size of the sets A_i, B_i, but we only use the n "inner" points from these sets. Second, we delete elements from S not based on them being inside A_i or B_i, but only based on the (privately computed) interval $[a_i, b_i]$. We now elaborate on these ideas, and present a (simplified) overview for the privacy analysis. Any informalities made herein are removed in the sections that follow.

Let S and $S' = S \cup \{(x', y')\}$ be neighboring databases, differing on the labeled point (x', y'). Consider the execution on S and on S'. The privacy analysis is based on the following two lemmas.

Lemma 5.2.1 (informal). *With probability at least $1 - \delta$, throughout the execution it holds that x' participates in at most $O(\log(1/\delta))$ sets A_i, B_i.*

This lemma holds because of our choice for the noise magnitude. In more detail, given that $x' \in A_i$, there is a constant probability that $x' \in A_i \setminus A_i^{\text{inner}}$. Since the interior point a_i is computed from A_i^{inner}, in such a case we will have that $x' < a_i$, and hence, x' is deleted from the data during this iteration. This means that every time x' is included in A_i, there is a constant probability that x' will be deleted from the data. Thus, one can show (using concentration bounds) that the number of times i such that $x' \in A_i$ is bounded (w.h.p.). A similar argument also holds for B_i.

Lemma 5.2.2 (informal). *In iterations i in which x' is* not *included in A_i or B_i, we have that a_i and b_i are distributed* exactly *the same during the execution on S and on S'.*

Indeed, in such an iteration, the point x' has no effect on the outcome distribution of $\mathcal{B}$ (who computes a_i, b_i). Overall, w.h.p., there are at most $O(\log \frac{1}{\delta})$ axes the point x' effects. We pay in composition only for those axes, while in all other axes we get privacy "for free". This allows us to save a factor of $\sqrt{d}$ in the sample complexity, and obtain an algorithm with sample complexity linear in d.

Note that the definition of privacy we work with is that of (ε, δ)-differential privacy. In contrast to the case of $(\varepsilon, 0)$-differential privacy, where it suffices to analyze the privacy loss w.r.t. every *single* possible outcome, with (ε, δ)-differential privacy we must account for arbitrary events. To tackle this, we had to perform a more explicit and meticulous analysis than that outlined above. Our analysis draws its structure from the proof of the advanced-composition theorem (Dwork et al., 2010b), but instead of composing everything we aim to preform *effective composition*, meaning that we incur a privacy loss only on a small fraction of the iterations. To achieve this, as we mentioned, we partition the iterations into several types – iteration on which we "pay" in privacy and iterations on which we do not. However, this partition must be done carefully, as the partition itself is random and needs to be different for different possible outcomes.

We believe that ideas from our work can be used more broadly, and hope that they find new applications in avoiding (or reducing) composition costs in other settings.

Remark 5.2.3. To simplify the presentation, in the technical sections of this paper we assume that the target rectangle is placed at the origin. Our results easily extend to arbitrary axis-aligned rectangles.

5.2.2 Formal Construction

Let $\mathcal{A}$ be an (ε, δ)-differentially private algorithm for solving the interior point problem over domain $\{0, \ldots, X\}$ with failure probability β and sample complexity $IP_{\mathcal{A}}(\varepsilon, \delta, \beta)$. We propose Algorithm 4, which we call `RandMargins`, and prove the following theorem.

Algorithm 4 `RandMargins`

```
**Input:** Data $S \subseteq \mathbb{R}^d$ of size $n$, and parameters $\beta < \frac{1}{4}$ and $\delta < 1/e^2, \varepsilon$
**Tool used:** An $(\varepsilon, \delta)$-private algorithm $\mathcal{A}$ for solving the interior point problem with failure probability $\beta$ and sample complexity $IP_{\mathcal{A}}(\varepsilon, \delta, \beta)$.
Denote $\Delta = IP_{\mathcal{A}}(\varepsilon, \delta, \beta)$
Denote $\mu = 4\Delta \log(1/\beta)$
Initialize $\bar{S} \leftarrow S$
**for** $i = 1$ **to** $d$ **do**
    $w_i \sim Lap(2\Delta)$
    $B_i = \max_i(\bar{S}, \lceil \mu + w_i \rceil)$
    $D_i = \min_i(B_i, \Delta)$
    $p_i \leftarrow \mathcal{A}(D_i, \varepsilon, \delta, \beta)$
    $R_i = \{y \in \bar{S} : y[i] \geq p_i\}$
    $\bar{S} \leftarrow \bar{S} \setminus R_i$
**Return** $(p_1, \ldots, p_d)$
```

Theorem 5.2.4. *Let $\varepsilon < 1, \delta < \frac{1}{e^2}, \alpha, \beta$. Algorithm 4 is $(\alpha, \beta, \tilde{\varepsilon}, \tilde{\delta})$-PPAC learner, for the REC_d class, given a labeled sample of size $\mathcal{O}\left(IP_{\mathcal{A}}(\varepsilon, \delta, \beta) \cdot \frac{d}{\alpha} \log\left(\frac{1}{\alpha}\right) \log\left(\frac{1}{\beta}\right)\right)$, for $\tilde{\delta} = (d+2)\delta$, and $\tilde{\varepsilon} = \mathcal{O}(\varepsilon \log(1/\delta))$.*

Remark 5.2.5. Kaplan et al. (2020a) introduced an algorithm $\mathcal{A}$ for the interior point problem with sample complexity $IP_{\mathcal{A}}(\varepsilon, \delta, \beta) = \widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon} \log^{1.5}\left(\frac{1}{\delta}\right) (\log^*(|\mathcal{X}|))^{1.5}\right)$. Hence, using their algorithm within Algorithm 4 provides the result of Theorem 1.2.2.

We analyze the privacy guarantees of Algorithm 4 in Section 5.3, and show the following lemma.

Lemma 5.2.6. *For every ε and every $\delta < \frac{1}{e^2}$. Then, given a labeled sample of size $\mathcal{O}\left(IP_{\mathcal{A}}(\varepsilon, \delta, \beta) \cdot \frac{d}{\alpha} \log\left(\frac{1}{\alpha}\right) \log\left(\frac{1}{\beta}\right)\right)$, Algorithm 4 is $(\tilde{\varepsilon}, \tilde{\delta})$-differentially private, for $\tilde{\delta} = (d+2)\delta$, and $\tilde{\varepsilon} = \mathcal{O}(\varepsilon \log(1/\delta))$.*

We analyze the utility guarantees of Algorithm 4 in Section 5.4, and show the following lemma.

Lemma 5.2.7. *For any choice of $\alpha, \beta, \varepsilon, \delta$, given a labeled sample of size*

$$\mathcal{O}\left(IP_{\mathcal{A}}(\varepsilon, \delta, \beta) \cdot \frac{d}{\alpha} \log\left(\frac{1}{\alpha}\right) \log\left(\frac{1}{\beta}\right)\right)$$

then, with probability at least $1 - \beta$ Algorithm 4 is α-accurate.

5.3 Privacy Analysis

Proof of Lemma 5.2.6. Let S and $S' = S \cup \{(x', y')\}$ be neighboring databases, differing on the labeled point (x', y'). Consider the execution on S and on S'.

We denote by $ind_i(x)$ the position of the point x in the remaining data $\bar{S}$, when the data is sorted by the i^{th} coordinate.

Denote by i^* the first iteration on which $x'[i] > p_i$, note that i^* is a random variable. For an input set S, denote by $\bar{S}_i$ the remaining set at the beginning of the i^{th} iteration and its size by $\bar{n}$.

Partition the iterations in the following way

- $\mathcal{I}_{in} = \{i \leq i^* \mid x' \in B'_i\}$
- $\mathcal{I}_{out} = \{i < i^* \mid x' \notin B'_i\}$
- $\mathcal{I}_{after} = \{i \mid i > i^*\}$

We first argue that $|\mathcal{I}_{in}|$ is small (with high probability). Intuitively, this follows from the fact that conditioned on $x' \in B'_i$, with constant probability, we get that $x' \in B'_i \setminus D_i$. Note that in such a case, projecting on the i^{th} axis, x' is bigger (or equal) than any point in D_i. Furthermore, as the interior point p_i is computed from D_i, w.h.p. we get that $x'[i] \geq p_i$, and hence x' is removed from the data. To summarize, conditioned on $x' \in B'_i$ there is a constant probability that x' is removed from the data, and hence the number of times such that $x' \in B'_i$ must be small (w.h.p.). We make this argument formal in the appendix, obtaining the following claim.

Claim 5.3.1.

$$\Pr[|\mathcal{I}_{in}| > 35 \log(1/\delta)] \leq \delta.$$

Next, we will denote by $\mathcal{B}$ the inner steps of the loop in the algorithm. Meaning, the input is $\bar{S}_i$, which $\mathcal{B}$ uses, along with the random noise and the mechanism $\mathcal{A}$, in order to output p_i. Note that $\mathcal{B}$ can be seen as a stand-alone (ε, δ)-differentially private algorithm (essentially amounts to a single execution of algorithm $\mathcal{A}$). For convenience, we will assume that the $\mathcal{B}$'s output includes the noise value w_i, and that the final output of `RandMargins` includes the noise vector $w = (w_1, \ldots, w_d)$. As will be proven below, algorithm `RandMargins` remains differentially private even when releasing this noise vector (in addition to the output $(p_1, \ldots, p_d)$).

Lemma 5.3.2 (Vadhan (2017))**.** *For every (ε, δ)-private algorithm M and every two neighboring datasets S, S', there exists an event $G = G(M, S, S')$ such that*

i) $\Pr[M(S) \in G] > 1 - \delta$

ii) $\Pr[M(S') \in G] > 1 - \delta$

iii) $\forall x \in G : \left| \ln \left(\frac{\Pr(M(S)=x)}{\Pr(M(S')=x)} \right) \right| \leq \varepsilon.$

Define the event $G = \{(p, w) \mid \forall j \in [d] : (p_j, w_j) \in G(\mathcal{B}, \bar{S}_j, \bar{S}'_j)\}$, where $G(\mathcal{B}, \bar{S}_j, \bar{S}'_j)$ is the event guaranteed to exist by applying Lemma 5.3.2 to $\mathcal{B}, \bar{S}_j, \bar{S}'_j$.

Note that by Lemma 5.3.2 and the union bound $\Pr[G] \geq 1 - d\delta$.

We wish to prove that for any possible output set P, it holds that

$$\Pr[\texttt{RandMargins}(S) \in P] \leq e^{\tilde{\varepsilon}} \cdot \Pr[\texttt{RandMargins}(S') \in P] + \tilde{\delta}.$$

Define the set

$$R = \left\{ (p, w) \left| \ln \left(\frac{\Pr[\mathcal{RM}(S) = (p, w)]}{\Pr[\mathcal{RM}(S') = (p, w)]} \right) > \tilde{\varepsilon} \right. \right\},$$

where $\mathcal{RM}$ is an abbreviation for `RandMargins`.

Now note that for every event P,

$$\begin{aligned}
&\Pr[\mathcal{RM}(S) \in P] \\
&\quad \leq \Pr[\mathcal{RM}(S) \in R] + \Pr[\mathcal{RM}(S) \in P \setminus R] \\
&\quad \leq \Pr[\mathcal{RM}(S) \in R] + e^{\tilde{\varepsilon}} \Pr[\mathcal{RM}(S') \in P \setminus R] \\
&\quad \leq \Pr[\mathcal{RM}(S) \in R| + e^{\tilde{\varepsilon}} \Pr[\mathcal{RM}(S') \in P]
\end{aligned}$$

So it is down to show that $\Pr[\mathcal{RM}(S) \in R] \leq \tilde{\delta}$. That is, we need to prove that

$$\Pr_{p,w \leftarrow \mathcal{RM}(S)} \left[\ln \left(\frac{\Pr(\mathcal{RM}(S) = p, w)}{\Pr(\mathcal{RM}(S') = p, w)} \right) > \tilde{\varepsilon} \right] \leq \tilde{\delta}.$$

We calculate,

$$
\begin{aligned}
&\Pr_{p,w\leftarrow\mathcal{RM}(S)}\left[\ln\left(\frac{\Pr(\mathcal{RM}(S)=p,w)}{\Pr(\mathcal{RM}(S')=p,w)}\right)>\tilde{\varepsilon}\right]\\
&=\Pr_{p,w\leftarrow\mathcal{RM}(S)}\Bigg[\left(\ln\left(\frac{\Pr(\mathcal{RM}(S)=p,w)}{\Pr(\mathcal{RM}(S')=p,w)}\right)\cdot\mathbb{1}_{p,w\in G}>\tilde{\varepsilon}\right)\ \text{OR}\\
&\qquad\qquad\left(\ln\left(\frac{\Pr(\mathcal{RM}(S)=p,w)}{\Pr(\mathcal{RM}(S')=p,w)}\right)\cdot\mathbb{1}_{p,w\notin G}>\tilde{\varepsilon}\right)\Bigg]\\
&\le\Pr_{p,w\leftarrow\mathcal{RM}(S)}\left[\ln\left(\frac{\Pr(\mathcal{RM}(S)=p,w)}{\Pr(\mathcal{RM}(S')=p,w)}\right)\cdot\mathbb{1}_{p,w\in G}>\tilde{\varepsilon}\right]\\
&\quad+\Pr_{p,w\leftarrow\mathcal{RM}(S)}\left[\ln\left(\frac{\Pr(\mathcal{RM}(S)=p,w)}{\Pr(\mathcal{RM}(S')=p,w)}\right)\cdot\mathbb{1}_{p,w\notin G}>\tilde{\varepsilon}\right]\\
&\le\Pr_{p,w\leftarrow\mathcal{RM}(S)}\left[\ln\left(\frac{\Pr(\mathcal{RM}(S)=p,w)}{\Pr(\mathcal{RM}(S')=p,w)}\right)\cdot\mathbb{1}_{p,w\in G}>\tilde{\varepsilon}\right]+(1-\Pr[G])\\
&\le\Pr_{p,w\leftarrow\mathcal{RM}(S)}\left[\ln\left(\frac{\Pr(\mathcal{RM}(S)=p,w)}{\Pr(\mathcal{RM}(S')=p,w)}\right)\cdot\mathbb{1}_{p,w\in G}>\tilde{\varepsilon}\right]+d\delta.
\end{aligned}
$$

It remains to prove that $\Pr_{p,w\leftarrow\mathcal{RM}(S)}\left[\ln\left(\frac{\Pr(\mathcal{RM}(S)=p,w)}{\Pr(\mathcal{RM}(S')=p,w)}\right)\cdot\mathbb{1}_{p,w\in G}>\tilde{\varepsilon}\right]\le 2\delta$.

We calculate,

$$
\begin{aligned}
&\Pr_{p,w\leftarrow\mathcal{RM}(S)}\left[\ln\left(\frac{\Pr(\mathcal{RM}(S)=p,w)}{\Pr(\mathcal{RM}(S')=p,w)}\right)\cdot\mathbb{1}_{p\in G}>\tilde{\varepsilon}\right]\\
&=\Pr_{p,w\leftarrow\mathcal{RM}(S)}\left[\ln\left(\prod_{i=1}^{d}\frac{\Pr(\mathcal{RM}(S)_i=p_i,w_i\mid p_{<i},w_{<i})}{\Pr(\mathcal{RM}(S')_i=p_i,w_i\mid p_{<i},w_{<i})}\right)\cdot\mathbb{1}_{p,w\in G}>\tilde{\varepsilon}\right]\\
&=\Pr_{p,w\leftarrow\mathcal{RM}(S)}\left[\sum_{i=1}^{d}\ln\left(\frac{\Pr(\mathcal{RM}(S)_i=p_i,w_i\mid p_{<i},w_{<i})}{\Pr(\mathcal{RM}(S')_i=p_i,w_i\mid= p_{<i},w_{<i})}\right)\cdot\mathbb{1}_{p,w\in G}>\tilde{\varepsilon}\right]\\
&\le\Pr_{p,w\leftarrow\mathcal{RM}(S)}\left[\sum_{i=1}^{d}\left(\ln\left(\frac{\Pr(\mathcal{RM}(S)_i=p_i,w_i\mid p_{<i},w_{<i})}{\Pr(\mathcal{RM}(S')_i=p_i,w_i\mid p_{<i},w_{<i})}\right)\cdot\mathbb{1}_{p_i,w_i\in G_i(\mathcal{B},\bar{S}_i,\bar{S}_i')}\right)>\tilde{\varepsilon}\right]\\
&=\Pr_{p,w\leftarrow\mathcal{RM}(S)}\Bigg[\sum_{i\in\mathcal{I}_{in}}\left(\ln\left(\frac{\Pr(\mathcal{RM}(S)_i=p_i,w_i\mid p_{<i},w_{<i})}{\Pr(\mathcal{RM}(S')_i=p_i,w_i\mid p_{<i},w_{<i})}\right)\cdot\mathbb{1}_{p_i,w_i\in G_i(\mathcal{B},\bar{S}_i,\bar{S}_i')}\right)\\
&\quad+\sum_{i\in\mathcal{I}_{out}}\left(\ln\left(\frac{\Pr(\mathcal{RM}(S)_i=p_i,w_i\mid p_{<i},w_{<i})}{\Pr(\mathcal{RM}(S')_i=p_i,w_i\mid p_{<i},w_{<i})}\right)\cdot\mathbb{1}_{p_i,w_i\in G_i(\mathcal{B},\bar{S}_i,\bar{S}_i')}\right)\\
&\quad+\sum_{i\in\mathcal{I}_{after}}\left(\ln\left(\frac{\Pr(\mathcal{RM}(S)_i=p_i,w_i\mid p_{<i},w_{<i})}{\Pr(\mathcal{RM}(S')_i=p_i,w_i\mid p_{<i},w_{<i})}\right)\cdot\mathbb{1}_{p_i,w_i\in G_i(\mathcal{B},\bar{S}_i,\bar{S}_i')}\right)>\tilde{\varepsilon}\Bigg].^{1}
\end{aligned}
\tag{5.1}
$$

[1] Note that the outer probability is over p and w. This allows the partition of the iterations into $\mathcal{I}_{in},\mathcal{I}_{out},\mathcal{I}_{after}$ to be well-defined, as this partition depends on p,w.

We will prove the following

(i) $\Pr\left[\sum_{i\in\mathcal{I}_{after}} \ln\left(\frac{\Pr(\mathcal{RM}(S)_i=p_i,w_i|p_{<i},w_{<i})}{\Pr(\mathcal{RM}(S')_i=p_i,w_i|p_{<i},w_{<i})}\right)\cdot \mathbb{1}_{p_i,w_i\in G_i(\mathcal{B},\bar{S}_i,\bar{S}'_i)} = 0\right] = 1$

(ii) $\Pr\left[\sum_{i\in\mathcal{I}_{out}} \ln\left(\frac{\Pr(\mathcal{RM}(S)_i=p_i,w_i|p_{<i},w_{<i})}{\Pr(\mathcal{RM}(S')_i=p_i,w_i|p_{<i},w_{<i})}\right)\cdot \mathbb{1}_{p_i,w_i\in G_i(\mathcal{B},\bar{S}_i,\bar{S}'_i)} = 0\right] = 1$

(iii) $\Pr\left[\sum_{i\in\mathcal{I}_{in}} \ln\left(\frac{\Pr(\mathcal{RM}(S)_i=p_i,w_i|p_{<i},w_{<i})}{\Pr(\mathcal{RM}(S')_i=p_i,w_i|p_{<i},w_{<i})}\right)\cdot \mathbb{1}_{p_i,w_i\in G_i(\mathcal{B},\bar{S}_i,\bar{S}'_i)} \leq \tilde{\varepsilon}\right] \geq 1-\delta$

Combining the above three claims implies a bound on (5.1) and finishes the proof.

Proof of (i). After i^*, by the algorithm definition, x' gets removed from S'. Hence, for every $i > i^*$, conditioning on $\mathcal{RM}(S)_{<i} = p_{<i}$, it holds that $B'_i = B_i$. This implies that, for every $i \in \mathcal{I}_{after}$,

$$\Pr(\mathcal{RM}(S)_i = p_i, w_i \mid p_{<i}, w_{<i}) = \Pr(\mathcal{RM}(S')_i = p_i, w_i \mid p_{<i}, w_{<i})$$

which yields

$$\begin{aligned}&\frac{\Pr(\mathcal{RM}(S)_i = p_i, w_i \mid p_{<i}, w_{<i})}{\Pr(\mathcal{RM}(S')_i = p_i, w_i \mid p_{<i}, w_{<i})} = 1\\ &\Rightarrow \Pr\left[\sum_{i\in\mathcal{I}_{after}} \ln\left(\frac{\Pr(\mathcal{RM}(S)_i = p_i, w_i \mid p_{<i}, w_{<i})}{\Pr(\mathcal{RM}(S')_i = p_i, w_i \mid p_{<i}, w_{<i})}\right)\cdot \mathbb{1}_{p_i,w_i\in G_i(\mathcal{B},\bar{S}_i,\bar{S}'_i)} = 0\right] = 1\end{aligned}$$

Proof of (ii). Recall that by the definition of $\mathcal{I}_{out}$ for every $i \in \mathcal{I}_{out}$ it holds that $x' \notin B'_i$, and hence, conditioning on the previous outputs, $B'_i = B_i$. We therefore get that the distribution of the i^{th} output is also the same. Formally,

$$\Pr(\mathcal{RM}(S)_i = p_i, w_i \mid p_{<i}, w_{<i}) = \Pr(\mathcal{RM}(S')_i = p_i, w_i \mid p_{<i}, w_{<i}).$$

This results in

$$\begin{aligned}&\ln\left(\frac{\Pr(\mathcal{RM}(S)_i = p_i, w_i \mid p_{<i}, w_{<i})}{\Pr(\mathcal{RM}(S')_i = p_i, w_i \mid p_{<i}, w_{<i})}\right)\cdot \mathbb{1}_{p_i,w_i\in G_i(\mathcal{B},\bar{S}_i,\bar{S}'_i)} = 0\\ &\Rightarrow \Pr\left[\sum_{i\in\mathcal{I}_{out}} \ln\left(\frac{\Pr(\mathcal{RM}(S)_i = p_i, w_i \mid p_{<i}, w_{<i})}{\Pr(\mathcal{RM}(S')_i = p_i, w_i \mid p_{<i}, w_{<i})}\right)\cdot \mathbb{1}_{p_i,w_i\in G_i(\mathcal{B},\bar{S}_i,\bar{S}'_i)} = 0\right] = 1.\end{aligned}$$

Proof of (iii). Note that, as we assume that the output of $\mathcal{RM}$ includes the random

Laplasian noise, then by fixing the past output-point $p_{<i}, w_{<i}$ we also fix $\bar{S}_i, \bar{S}'_i$. So,

$$\begin{aligned}&\ln\left(\frac{\Pr(\mathcal{RM}(S)_i = p_i, w_i \mid p_{<i}, w_{<i})}{\Pr(\mathcal{RM}(S')_i = p_i, w_i \mid p_{<i}, w_{<i})}\right) \cdot \mathbb{1}_{p_i,w_i \in G_i(\mathcal{B},\bar{S}_i,\bar{S}'_i)} \\ &= \ln\left(\frac{\Pr(\mathcal{B}(\bar{S}_i) = p_i, w_i)}{\Pr(\mathcal{B}(\bar{S}'_i) = p_i, w_i)}\right) \cdot \mathbb{1}_{p_i,w_i \in G_i(\mathcal{B},\bar{S}_i,\bar{S}'_i)}.\end{aligned}$$

Moreover, by the definition of the events G_i it holds that

$$\Pr\left[|\ln\left(\frac{\Pr(\mathcal{B}(\bar{S}_i) = p_i, w_i)}{\Pr(\mathcal{B}(\bar{S}'_i) = p_i, w_i)}\right) \cdot \mathbb{1}_{p_i,w_i \in G_i(\mathcal{B},\bar{S}_i,\bar{S}'_i)}| \leq 2\varepsilon\right] = 1.$$

which yields

$$\begin{aligned}&\Pr\left[\sum_{i\in\mathcal{I}_{in}} \ln\left(\frac{\Pr(\mathcal{B}(\bar{S}_i) = p_i, w_i)}{\Pr(\mathcal{B}(\bar{S}'_i) = p_i, w_i)}\right) \cdot \mathbb{1}_{p_i,w_i \in G_i(\mathcal{B},\bar{S}_i,\bar{S}'_i)} > \tilde{\varepsilon}\right] \\ &\leq \Pr\left[\sum_{i\in\mathcal{I}_{in}} 2\varepsilon > \tilde{\varepsilon}\right] \leq \Pr\left[|\mathcal{I}_{in}| > \frac{\tilde{\varepsilon}}{2\varepsilon}\right] \leq \delta,\end{aligned}$$

where the last inequality follows from Claim 5.3.1 and from our choice of

$$\tilde{\varepsilon} = \mathcal{O}\left(\varepsilon \log(1/\delta)\right).$$

□

5.4 Utility

Proof of Lemma 5.2.7. First, we must ensure that at every iteration, with high probability, we have enough points left in $\bar{S}$. At the same time, we must ensure that the auxiliary algorithm $\mathcal{A}$ will output an inner point of the given subset. Denote $a_j = w_j + \mu$. By the definition of the noise w and the mean μ, we get that for every iteration i: $\Pr[a_i > 6\Delta\log(1/\beta)] < \beta$. Hence, with probability $\geq 1 - d\beta$, it holds that for every i $a_i \leq 6\Delta\log(1/\beta)$. This means that the total number of removed points is at most $6d\Delta\log(1/\beta)$. Therefore, for a sample of size $6d\Delta\log(1/\beta)$ with high probability $\bar{S}$ will contain enough points.

Regarding the algorithm's accuracy, we notice that at every iteration j, $\mathcal{A}$ outputs a point which is at least the a_j-th largest point from *the points left in the set.* This means that, in the worst case, we delete a_j points from the data set at this iteration. Hence, again in worst case, we will output the $\sum_{j=1}^{i} a_j$-th largest point in the j^{th} axis.

By the above reasoning, with high probability we can say that for every i it holds that $a_j \leq 6\Delta \log(1/\beta)$. Meaning that every p_j is at least the $\sum_{j=1}^{d} a_j \leq 6d\Delta \log(1/\beta)$ largest point in the axis. This implies that, for sample of size $\mathcal{O}\left(\frac{d\Delta}{\alpha} \log(1/\alpha) \log(1/\beta)\right)$, denoting the by h_p the hypothesis induces by the output of Algorithm 4 $\Pr_{S\sim\mu^n}[\mathrm{err}_S(h_p) \geq \alpha/2] \leq \beta/2$. Since the VC-dimension of the class REC_d is $2d$, by Theorem 3.1.6 and the fact that the sample size is at least as the sample complexity bound $\mathcal{O}\left(\frac{1}{\alpha}\left(d \log\left(\frac{1}{\alpha}\right) + \log\left(\frac{1}{\beta}\right)\right)\right)$ it holds that: $\Pr_{S\sim\mu^n}[\mathrm{err}_\mu(h_p) \geq \mathrm{err}_S(h_p) + \alpha/2] \leq \beta/2$. Combining the two bounds concludes the proof. $\square$

Chapter 6

Universal Private Learning

We prove the existence of universal private learners. As mentioned above, the existence of such algorithm is in sharp contrast to the impossibility results for PAC-learning. Before introducing the algorithms and the results, we detail one tool which we will be using for doing private *histogram count*, which is one of the most fundamental statistical tasks. The task is, given a dataset, to count how many times each unique datum appears in the data. The most common private solution is the Laplace mechanism, which guarantees $(\varepsilon, 0)$-differential privacy. The main caveat of this approach is that the error, in some cases, might over-accumulate, since we add noise to every possible domain point. A different technique, specified in Algorithm 5, is to ignore zero-counts and also zero-out counts which do not exceed a certain (noisy) threshold. This allows us to avoid the above accumulation of error, at the cost of guaranteeing privacy with $\delta > 0$. Formally,

Algorithm 5 Stability based Histogram Bun et al. (2019c)

```
Input: Dataset $S \in \mathcal{X}^n$
for $x \in \mathcal{X}$ do
    if $count_S(x) = 0$ then
        $\hat{c}(x) \leftarrow 0$
    else
        $\hat{c}(x) \leftarrow count_S(x) + \mathrm{Lap}(2/\varepsilon)$
        if $\hat{c}(x) < \frac{2}{\varepsilon}\log\left(\frac{2}{\delta}\right) + 1$ then
            $\hat{c}(x) \leftarrow 0$
Return $\hat{c}$
```

Theorem 6.0.1 (Bun et al. (2019c)). *The* `Stability based Histogram` *algorithm is (ε, δ)-differentially private. Moreover, for every domain point $x \in \mathcal{X}$, the resulting count $\hat{c}(x)$ is such that if $count_S(x) = 0$ then $\hat{c}(x) = 0$, and otherwise $\mathbb{E}|\hat{c}(x) - count_S(x)| \leq O\left(\frac{1}{\varepsilon} \cdot \min\left\{\log\frac{1}{\delta}, count_S(x)\right\}\right)$.*

6.1 Classification

Algorithm 6 PCL

```
1: Input: Sample $S_n = \{(x_i, y_i)\}_{i=1}^n$
2: Set $r = \frac{1}{n^{1/(2d)}}$
3: Partition the space into equally sized cubes $\mathcal{C} = C_1, C_2, \dots$ with side length $r$
4: For any $x$ denote $C(x)$ the cube s.t. $x \in C(x)$
5: Define the hypothesis $h_{\mathcal{C}}$ s.t. $h_{\mathcal{C}}(x) = \mathbb{1}_{\sum_{x_i \in C(x)} y_i + Lap(1/\varepsilon) > \frac{|C(x)|}{2}}$
6: Return $h_{\mathcal{C}}$
```

We begin by studying UC learning over the bounded euclidean space $[0,1]^d$. Our classification algorithm is presented in Algorithm 6. In words: we partition the space into equally sized cubes with side length r. To classify a new point, take the bucket into which it falls and compute a noisy majority vote within this bucket.

Theorem 6.1.1. *Algorithm 6 is ε-differentially private.*

Proof. Histogram counts as used in Algorithm 6 have global sensitivity 1 (see (Dwork et al., 2014)). Hence, adding Laplace noise of scale $1/\varepsilon$ results in ε-differential privacy. Note that although Step 5 in the algorithm seems to access the data twice, which might require the scale of the noise to be larger, this is not the case. To see this, notice that a different way of calculating the same majority-vote is by looking at the following sum $\sum_{x \in C_j}(y_i - 1/2) + w_j$, where w_j is the noise added to the cube C_j, and outputting 1 if it is greater than 0 and output 0 otherwise. As such, this amounts to a single calculation with global sensitivity 1. Hence, by Theorem 3.2.6 the addition of Laplace noise of scale $1/\varepsilon$ ensures that the noisy counts are private. As the final output is merely a post-processing of these counts, it is also ε-private. □

Theorem 6.1.2. *Algorithm 6 is universally-consistent.*

Proof of Theorem 6.1.2. Given a test point $x \in [0,1]^d$, denote by $A(x) = \{X_i \in S \cap C(x)\}$ the set of points from S in the same bucket with x, and denote the size of that bucket as $N(x) = \Sigma_{i=1}^n \mathbb{1}_{X_i \in A(x)}$. Also define

- $\hat{\eta}_n(x) := \frac{1}{N(x)} \Sigma_{i: x_i \in A(x)} y_i$
- $\hat{\eta}_n^\varepsilon(x) := \hat{\eta}_n(x) + w_j$, where w_j is the noise added to $C(x)$.

Note, that algorithm `PCL` is a *plug-in classifier* w.r.t. $\hat{\eta}_n^\varepsilon$. Hence, by Theorem 3.1.11, in order to prove that it is consistent it suffices to show that

$$\lim_{n\to\infty} \mathbb{E}\left[|\hat{\eta}_n^\varepsilon(x) - \eta(x)|\right] = 0.$$

By the triangle inequality, $\mathbb{E}\left[|\hat{\eta}_n^{\varepsilon}(x) - \eta(x)|\right] \leq \mathbb{E}\left[|\hat{\eta}_n^{\varepsilon}(x) - \hat{\eta}_n(x)|\right] + \mathbb{E}\left[|\hat{\eta}_n(x) - \eta(x)|\right]$. In order to show that the first term goes to zero we use the following lemma.

Lemma 6.1.3 (Devroye et al. (2013))**.** *For any $k \in \mathbb{N}$ we have $\Pr[N(x) \leq k] \xrightarrow[n\to\infty]{} 0$, where the probability is over sampling $S_n \sim \mu^n$ and sampling $x \sim \mu$.*

Using the above lemma, we can bound the expected gap caused by the noise as follows.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S,x}\mathop{\mathbb{E}}_{\mathcal{A}}\left[|\hat{\eta}_n^{\varepsilon}(x) - \hat{\eta}_n(x)|\right] \leq \mathop{\mathbb{E}}_{S,x}\mathop{\mathbb{E}}_{\mathcal{A}}\left[|\hat{\eta}_n^{\varepsilon}(x) - \hat{\eta}_n(x)| \cdot \mathbb{1}_{N(x)>0}\right] + \Pr[N(x) = 0] \\
&= \mathop{\mathbb{E}}_{S,x}\ \mathop{\mathbb{E}}_{w_j\sim\mathrm{Lap}}\left[\frac{|w_j|}{N(x)} \cdot \mathbb{1}_{N(x)>0}\right] + \Pr[N(x) = 0] \\
&= \mathop{\mathbb{E}}_{S,x}\left[\frac{1}{\varepsilon N(x)} \cdot \mathbb{1}_{N(x)>0}\right] + \Pr[N(x) = 0] \\
&= \frac{1}{\varepsilon}\Bigg(\mathbb{E}\left[\frac{1}{N(x)} \mid 0 < N(x) < M\right] \cdot \Pr(0 < N(x) < M) \\
&\qquad + \mathbb{E}\left[\frac{1}{N(x)} \mid N(x) \geq M\right] \cdot \Pr(N(x) \geq M)\Bigg) \\
&\quad + \Pr[N(x) = 0] \leq \frac{1}{\varepsilon}\left(\Pr(N(x) < M) + \frac{1}{M}\right).
\end{aligned} \tag{6.1}$$

Since this is true for every choice of M and by using Lemma 6.1.3 again, this also can be made arbitrarily small using a sufficiently large sample size. Hence,

$$\mathbb{E}\left[|\hat{\eta}_n^{\varepsilon}(x) - \hat{\eta}_n(x)|\right] \xrightarrow{n\to\infty} 0. \tag{6.2}$$

Furthermore, we recall the following result by Devroye et al. (2013)

Theorem 6.1.4 (Devroye et al. (2013))**.** *For any r and n s.t. $\lim_{n\to\infty} r = 0$ and $\lim_{n\to\infty} nr^d = \infty$ we get that $\lim_{n\to\infty} \mathbb{E}\left[|\hat{\eta}_n(x) - \eta(x)|\right] = 0$.*

Hence, the choice of $r = \frac{1}{n^{1/(2d)}}$, together with (6.2) completes the proof. □

As we mentioned, in the supplementary material we extend this construction to metric spaces with finite doubling dimension.

6.2 Density Estimation

We now turn to the problem of density estimation over $\mathbb{R}^d$. In particular, this implies private UC learning over $\mathbb{R}^d$ (rather than over $[0, 1]^d$ as in the previous section). We present the following histogram-based approximation algorithm for density function.

Algorithm 7 PCDE

```
Input: Sample $S_n = \{(x_i)\}_{i=1}^{n}$.
Set $r = \frac{1}{n^{1/(2d)}}$
Partition the space into equally sized cubes $\mathcal{C} := C_1, C_2, \ldots$ with side length $r$
Apply Stability based Histogram with input $S_n$ to obtain estimates $\hat{c}_1, \hat{c}_2, \ldots$ for
  $c_1, c_2, \ldots$, where $c_j := |\{x \in S_n : x \in C_j\}|$ denotes the number of input points in the
  cube $C_j$.
For $x \in C_j$ denote $c(x) = c_j$ and $\hat{c}(x) = \hat{c}_j$.
Return the function $\hat{f}_S$ defined as $\hat{f}_S(x) := \frac{1}{nr^d}\hat{c}(x)$.
```

The algorithm make use of the *Stability based Histogram* algorithm in order to produce counting estimates while preserving privacy. It then defines the estimator to be the estimated count normalized with respect to the sample size and the partition-cubes' size.

Remark 6.2.1. *Due to the noises in the counts, the output $\hat{f}_S$ of algorithm* `PCDE` *might not be a density function: one needs to zero out negative terms it might contain and then to normalize it. This has a negligible effect on the distance from the underlying distribution, and we ignore it for simplicity.*[1]

Theorem 6.2.2. *Algorithm 7 is (ε, δ)-differentially private.*

Proof. As `Stability based Histogram` is (ε, δ)-differentially private, and since differential privacy is closed under post-processing, the output of `PCDE` is also (ε, δ)-differentially private. □

Theorem 6.2.3. *The output of Algorithm 7, denoted by $\hat{f}_S$, is universally consistent for density estimation in L_1 norm. Namely, for every distribution μ over $\mathbb{R}^d$ with density function f we have*

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{S\sim\mu^n} \mathop{\mathbb{E}}_{\hat{f}_S \leftarrow \mathcal{A}(S)} \int |f(x) - \hat{f}_S(x)| dx = 0.$$

Proof. For sample S and the corresponding partition $\mathcal{C}$, define the classic histogram-density estimation

$$f_S(x) := \frac{1}{nr^d} \sum_{i=1}^{n} \mathbb{1}_{x_i \in C(x)}. \tag{6.3}$$

We will be using the following theorem

Theorem 6.2.4 (Devroye et al. (2013), Devroye and Györfi (1985))**.** *Let f_S denote the*

[1] In more detail, let f denote the target distribution, let $\hat{f}$ denote the outcome of the algorithm, and suppose that the L_1 distance between f and $\hat{f}$ is w. Now let g denote $\hat{f}$ after zeroing out negative terms and after normalizing it (as in Remark 6.2.1). An easy calculation (follows from the triangle inequality) shows that the L_1 distance between f and g is at most $O(w)$. This means that if the L_1 distance between f and $\hat{f}$ goes to zero, then so does the distance between f and g.

standard histogram estimator (defined as in (6.3)*). Then,*

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{S\sim P^n} \int |f(x) - f_S(x)|dx = 0.$$

Now, by the triangle inequality,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S\sim\mu^n} \mathop{\mathbb{E}}_{\hat{f}_S\leftarrow\mathcal{A}(S)} \int |f(x) - \hat{f}_S(x)|dx \\
&\leq \mathop{\mathbb{E}}_{S\sim\mu^n} \mathop{\mathbb{E}}_{\hat{f}_S\leftarrow\mathcal{A}(S)} \int |f_S(x) - \hat{f}_S(x)|dx + \mathop{\mathbb{E}}_{S\sim\mu^n} \int |f(x) - f_S(x)|dx. \qquad (6.4)
\end{aligned}$$

Hence, by Theorem 6.2.4, it suffices to show that $\lim_{n\to\infty} \mathbb{E}_{S,\hat{f}_S} \int |f_S(x) - \hat{f}_S(x)|dx = 0$. To this end, let $\tau > 0$ be some parameter, let T_0 be such that there exist a cube $\mathcal{T}_0$ of side-length T_0 satisfying $\mu(\mathcal{T}_0) > 1 - \tau$. Now let $\mathcal{T}$ denote the cube $\mathcal{T}_0$ after extending it by 1 in each direction (so $\mathcal{T}$ is a cube of side length $T := T_0 + 2$).

Remark 6.2.5. *Recall that the cubes C_j defined by Algorithm 7 are of side length $r \leq 1$. Thus, any cube C_j that intersects $\mathcal{T}_0$ is contained in $\mathcal{T}$.*

The interior of $\mathcal{T}$ will be partitioned into $\frac{T^d}{r^d}$ cubes of volume r^d. If we restrict our calculation to $\mathcal{T}$, we get that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S\sim\mu^n} \mathop{\mathbb{E}}_{\hat{f}_S\leftarrow\mathcal{A}(S)} \int_{\mathcal{T}} |f_S(x) - \hat{f}_S(x)|dx = \int_{\mathcal{T}} \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{\hat{f}_S} |f_S(x) - \hat{f}_S(x)|dx \\
&\lesssim \int_{\mathcal{T}} \frac{1}{nr^d} \cdot \frac{1}{\varepsilon} \log\left(\frac{1}{\delta}\right) dx = \frac{1}{nr^d} \cdot \frac{1}{\varepsilon} \log\left(\frac{1}{\delta}\right) \int_{\mathcal{T}} dx \\
&= T^d \frac{1}{nr^d} \cdot \frac{1}{\varepsilon} \log\left(\frac{1}{\delta}\right) = \frac{T^d}{\varepsilon\sqrt{n}} \log\left(\frac{1}{\delta}\right), \qquad (6.5)
\end{aligned}$$

where the inequality is by Theorem 6.0.1 (after neglecting the constant hiding in the O-notation) and the last equality is by the choice of $r = \frac{1}{n^{1/(2d)}}$.

Outside $\mathcal{T}$, by its definition, we have $\mu(\bar{\mathcal{T}}) \leq \mu(\bar{\mathcal{T}}_0) < \tau$ and therefore

$$\mathbb{E}\left[|S \cap \bar{\mathcal{T}}|\right] \leq \mathbb{E}\left[|S \cap \bar{\mathcal{T}}_0|\right] < n\tau. \qquad (6.6)$$

We can calculate that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S\sim\mu^n}\ \mathop{\mathbb{E}}_{\hat{f}_S\leftarrow\mathcal{A}(S)}\int_{\bar{\mathcal{T}}}|f_S(x)-\hat{f}_S(x)|dx = \mathop{\mathbb{E}}_{S}\int_{\bar{\mathcal{T}}}\mathop{\mathbb{E}}_{\hat{f}_S}|f_S(x)-\hat{f}_S(x)|dx\\
&\quad\le \mathop{\mathbb{E}}_{S}\int_{\bar{\mathcal{T}}}\frac{1}{\varepsilon n r^d}\cdot c(x)\ dx = \frac{1}{\varepsilon n r^d}\cdot\mathop{\mathbb{E}}_{S}\int_{\bar{\mathcal{T}}}c(x)\ dx\\
&\quad\le \frac{1}{\varepsilon n r^d}\cdot\mathop{\mathbb{E}}_{S}\sum_{C_j:C_j\cap\bar{\mathcal{T}}\neq\emptyset}|S\cap C_j|\cdot r^d \le \frac{1}{\varepsilon n r^d}\cdot\mathop{\mathbb{E}}_{S}\sum_{C_j:C_j\subseteq\bar{\mathcal{T}}_0}|S\cap C_j|\cdot r^d\\
&\quad\le \frac{1}{\varepsilon n r^d}\cdot\mathop{\mathbb{E}}_{S}|S\cap\bar{\mathcal{T}}_0|\cdot r^d \le \frac{\tau}{\varepsilon},
\end{aligned}\tag{6.7}$$

where the first inequality follows from the properties of `Stability based Histogram`, and the last inequality follows from (6.6).

Finally, combining (6.5) and (6.7) yields

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S\sim\mu^n}\ \mathop{\mathbb{E}}_{\hat{f}_S\leftarrow\mathcal{A}(S)}\int|f_S(x)-\hat{f}_S(x)|dx\\
&= \mathop{\mathbb{E}}_{S\sim\mu^n}\ \mathop{\mathbb{E}}_{\hat{f}_S\leftarrow\mathcal{A}(S)}\int_{T}|f_S(x)-\hat{f}_S(x)|dx + \mathop{\mathbb{E}}_{S\sim\mu^n}\ \mathop{\mathbb{E}}_{\hat{f}_S\leftarrow\mathcal{A}(S)}\int_{\bar{T}}|f_S(x)-\hat{f}_S(x)|dx\\
&\lesssim \frac{T^d}{\varepsilon\sqrt{n}}\log\left(\frac{1}{\delta}\right)+\frac{\tau}{\varepsilon}.
\end{aligned}$$

As $\frac{T^d}{\sqrt{n}}\xrightarrow{n\to\infty}0$ and τ can be arbitrarily small we get that

$$\lim_{n\to\infty}\mathop{\mathbb{E}}_{S,\hat{f}_S}\int|f_S(x)-\hat{f}_S(x)|dx = 0.$$

This completes the proof. □

6.2.1 Consistent and Private Semi-Supervised Learning

We next show that the above result yields an application to the setting of semi-supervised private learning. Let $\mathcal{C}$ be a class of concepts. Recall that in the semi-supervised setting, we are given two samples $S\in(\mathcal{X}\times\{0,1\})^m$ and $U\in(\mathcal{X}\times\{\perp\})^n$. For simplicity, we will restrict our discussion in this subsection to the realizable setting. Let us first recall the definition of semi-supervised learning (SSL) in the distribution free PAC model.

Definition 6.2.6. *An algorithm $\mathcal{A}$ is said to be an* SSL learning algorithm *for a class $\mathcal{C}$ if for every α,β there exist $m=m(\alpha,\beta,\mathcal{C})$ and $n=n(\alpha,\beta,\mathcal{C})$ such that for every distribution μ it holds that $\Pr_{S\sim\mu^m,U\sim\bar{\mu}^n,h\sim\mathcal{A}(S,U)}[\mathrm{err}_\mu(h)>\alpha]<\beta$, where $\bar{\mu}$ is the marginal distribution of the unlabeled samples.*

Definition 6.2.7 (Private SSL)**.** *An algorithm is said to be a* PSSL-learning algorithm *for a class $\mathcal{C}$ if it is an SSL-learner for $\mathcal{C}$ and also it is (ε,δ)-differentially private.*

As in the standard learning model (where all examples are labeled), semi-supervised learning can be defined in the distribution-dependent setting, or consistent setting, as follows.

Definition 6.2.8. *An algorithm $\mathcal{A}$ is said to be a* consistent semi-supervised learner *(CSSL for short) for a class $\mathcal{C}$ if for every α, β there exist $m = m(\alpha, \beta, \mathcal{C})$ such that for every distribution μ there is some $n = n(\alpha, \beta, \mathcal{C}, \mu)$ for which*

$$\Pr_{S\sim\mu^m, U\sim\bar{\mu}^n, h\sim\mathcal{A}(S,U)}[\mathrm{err}_\mu(h) > \alpha] < \beta,$$

where $\bar{\mu}$ is the marginal distribution of the unlabeled samples.

Note that in the above definition, we required the labeled sample complexity to be *uniform* over all possible underlying distributions, while allowing the unlabeled sample complexity to depend on the underlying distribution. This is interesting because with differential privacy there are cases where semi-supervised learning cannot be done in the distribution-free setting. We show that it suffices for the unlabeled sample complexity to depend on the underlying distribution, while keeping the labeled sample complexity independent of it.

Definition 6.2.9. *An algorithm is an (ε, δ)-*private consistent semi-supervised learner *(private-CSSL for short) if it is a consistent semi-supervised learner and (ε, δ)-differentially private.*

For the following result, we will be using the notion of *semi-private learning.* The notion captures a scenario in which the data is sensitive, but the underlying distribution is not. This is modeled by defining a semi-supervised learning task in which the learner is required to preserve privacy only for the labeled part of the sample. Formally, a *semi-private* SSL algorithm is an SSL algorithm that satisfies differential privacy w.r.t. its labeled database (for every fixture of its unlabeled database).

Theorem 6.2.10 (Beimel et al. (2016b); Bassily et al. (2019b))**.** *for any concept class $\mathcal{C}$, there exists a semi-private SSL algorithm which have a labeled sample complexity of $m = \mathcal{O}\left(\frac{1}{\varepsilon\alpha}VC(\mathcal{C})\log\left(\frac{1}{\alpha\beta}\right)\right)$ and unlabeled sample complexity $n = \mathcal{O}\left(\frac{1}{\alpha}VC(\mathcal{C})\log\left(\frac{1}{\alpha\beta}\right)\right)$.*

As an application of our results for density estimation, we get the following corollary.

Theorem 6.2.11. *For every class $\mathcal{C}$ over $\mathbb{R}^d$ with $VC(\mathcal{C}) < \infty$ and for every ε, δ, there exists a proper (ε, δ)-private-CSSL for $\mathcal{C}$ whose (labeled) sample complexity is $m = \mathcal{O}\left(\frac{1}{\varepsilon\alpha}VC(\mathcal{C})\log\left(\frac{1}{\alpha\beta}\right)\right)$.*

Remark 6.2.12. *Notice that the labeled sample complexity is optimal, as a sample of size $\mathcal{O}\left(d_(\mathcal{C})\right)$ is necessary in order to learn a concept class $\mathcal{C}$ even without the privacy requirement.*

Proof of Theorem 6.2.11. Let $\mathcal{C}$ be some class with $d_(\mathcal{C}) < \infty$. Let $\mathcal{A}$ be a semi-private SSL algorithm for $\mathcal{C}$, as guaranteed by Theorem 6.2.10, and let m_{semi} and n_{semi} denote its labeled and unlabeled sample complexities, respectively.

Now fix an underlying distribution μ and let f denote its marginal distribution over unlabeled examples. By Theorem 6.2.3 there is some $n = n\left(\frac{\beta}{n_{\text{semi}}}, \beta, f\right)$ s.t. we can privately generate a function $\hat{f}$, which is $\frac{\beta}{n_{\text{semi}}}$ close (in *total variation distance*) to the density function f w.p. $1-\beta$. We proceed with the analysis assuming that this is the case.

Let $U \sim f^{n_{\text{semi}}}$ denote a sample containing n_{semi} samples from f and let $\hat{U} \sim \hat{f}^{n_{\text{semi}}}$ denote a sample containing n_{semi} samples from $\hat{f}$. As $f, \hat{f}$ are $\frac{\beta}{n_{\text{semi}}}$ close in total variation distance, we get that $f^{n_{\text{semi}}}$ and $\hat{f}^{n_{\text{semi}}}$ are β close in total variation distance. By Theorem 6.2.10 we know that

$$\Pr_{\substack{S\sim\mu^{m_{\text{semi}}},\\ U\sim f^{n_{\text{semi}}}\\ h\leftarrow\mathcal{A}(S,U)}}[\text{err}_\mu(h) > \alpha] < \beta,$$

and so,

$$\Pr_{\substack{S\sim\mu^{m_{\text{semi}}},\\ \hat{U}\sim \hat{f}^{n_{\text{semi}}}\\ h\leftarrow\mathcal{A}(S,\hat{U})}}[\text{err}_\mu(h) > \alpha] < 2\beta.$$

The unlabeled sample is accessed only via the private-density estimation algorithm, and the labeled sample is accessed only via the semi-private learning method. The algorithm is therefore differentially private by composition and post-processing. □

6.3 Metric Spaces with Finite Doubling Dimension

In this section, we extend our results to the more general setting of metric spaces with bounded doubling dimension. We first present some additional preliminaries.

Definition 6.3.1 (Doubling dimension). *For a metric space $(\mathcal{X}, \rho)$, let $\lambda > 0$ be the smallest integer such that every ball in $\mathcal{X}$ can be covered by λ balls of half the radius. The* doubling dimension *of $(\mathcal{X}, \rho)$ is $ddim(\mathcal{X}) = \log_2(\lambda)$.*

Definition 6.3.2. *For a metric space $(\mathcal{X}, \rho)$, a set of points $\mathcal{M}$ in $\mathcal{X}$ is said to be r-cover of $\mathcal{X}$ if for every $x \in \mathcal{X}$ there exist some $x' \in \mathcal{M}$ s.t. $\rho(x, x') \leq r$.*

Definition 6.3.3. *For a metric space $(\mathcal{X}, \rho)$, a set of points $\mathcal{N}$ in $\mathcal{X}$ is said to be r-packing of $\mathcal{X}$ if for every $x, x' \in \mathcal{N}$ $\rho(x, x') \geq r$.*

An r-packing $\mathcal{N}$ is said to be maximal *if for any $x \in \mathcal{X} \setminus \mathcal{N}$ it holds that $\mathcal{N} \cup \{x\}$ is not an r-packing of $\mathcal{X}$. Namely, it means that there is some $x' \in \mathcal{N}$ s.t. $\rho(x, x') < r$.*

We will be leveraging the following classical connection between packing and covering.

Theorem 6.3.4 (Vershynin (2018))**.**

1. *Let $\mathcal{N}$ be a* maximal r-packing *of $\mathcal{X}$, then $\mathcal{N}$ is also an* r-cover *of $\mathcal{X}$.*
2. *If there exists an* r-cover *of $\mathcal{X}$ of size m, then any* $2r$-packing *of $\mathcal{X}$ is of size at most m.*

Definition 6.3.5. *A metric space $(\mathcal{X}, \rho)$ is* separable *if it has a countable dense set. That is, there exists a* countable *set $Q \subseteq \mathcal{X}$ such that every nonempty open subset of $\mathcal{X}$ contains at least one element from Q.*

6.3.1 Bounded Doubling Metric Spaces

We begin by proving the following theorem.

Theorem 6.3.6. *Let $\varepsilon \leq 1$ be a constant. There is an $(\varepsilon, 0)$-differentially private universal consistent learner for every* bounded *and* separable *metric space with finite doubling dimension.*

Remark 6.3.7. *The separability requirement is in fact necessary. It has been shown by Hanneke et al. (2021) that metric spaces which are not essentially separable have no consistent learning rules, even non-private ones.*

Let $(\mathcal{X}, \rho)$ be a bounded and separable metric space with doubling dimension d. Note that as $\mathcal{X}$ has finite doubling dimension and is bounded, it has a finite covering for every r. Therefore, a *maximal* packing of $\mathcal{X}$ will also be of finite size.

Consider Algorithm 8, which is an extension of the PCL algorithm. The algorithm partition the space to Voronoi cells centered in the points corresponding to a maximal packing. The outputted classifier is then defined as the noisy majority vote for the cell. The privacy properties of this algorithm are straightforward; we now proceed with its utility analysis.

Algorithm 8 PCL2

```
Input: Sample $S_n = \{(x_i, y_i)\}_{i=1}^n$
Set $r = \frac{1}{n^{1/(4d)}}$
Let $\mathcal{N}$ be an $r$ maximal packing of $\mathcal{X}$.
Partition the space into Voronoi cells centered in the elements of $\mathcal{N}$: $\mathcal{C} = V_1, V_2, \dots$.
For any $x$ denote $V(x)$ the cell s.t. $x \in V(x)$
Define $h_{\mathcal{C}}(x) = \mathbb{1}_{\sum_{x_i \in V(x)} y_i + Lap(1/\varepsilon) > \frac{|V(x)|}{2}}$
Return $h_{\mathcal{C}}$
```

Given a test point $x \in \mathcal{X}$, denote by $A(x) = \{X_i \in S \cap V(x)\}$ the set of points from S in the same bucket with x, denote the size of that bucket as $N(x) = |A(x)|$, and lastly, $N(V) := \frac{1}{n}\Sigma_{i=1}^n \mathbb{1}_{X_i \in V}$ which is the relative size of the sample points in V from the entire sample.

Lemma 6.3.8. *For every $V \in \mathcal{C}$ it holds that diam $(V) \leq 2r = \frac{2}{n^{1/(4d)}}$*

Proof. Given a center point p_i from $\mathcal{N}$, denote by $\hat{V}_i$ the ball of radius r around it, and by V_i the Voronoi cell induced by it. Let $a, b \in V_i$ be two points on V_i. By the definition of Voronoi cells for any other center point p_j, it holds that $\rho(a, p_j) \geq \rho(a, p_i)$ and $\rho(b, p_j) \geq \rho(b, p_i)$. Therefore, if $\rho(a, p_i) > r$ or $\rho(b, p_i) > r$, we will get that $\forall p \in \mathcal{N} : \rho(a, p) > r$ or $\forall p \in \mathcal{N} : \rho(b, p) > r$, which is a contradiction to the covering property of $\mathcal{N}$. Hence, we get that $\rho(a, p_i) \leq r$ and $\rho(b, p_i) \leq r$ which, by the triangle inequality, result in $\rho(a, b) \leq r$. □

Lemma 6.3.9. *For any $k = k(n)$ such that $k(n) = o(n^{1/4})$ we have $\Pr[N(x) \leq k] \xrightarrow[n \to \infty]{} 0$, where the probability is over sampling $S_n \sim \mu^n$ and sampling $x \sim \mu$.*

Remark 6.3.10. *This theorem (and its proof) holds for both bounded and unbounded spaces. We decided to provide it in this general form as we also make use of it to analyze the unbounded case later on.*

Proof. Let $\theta = n^{1/(2d)}$, let $T \subseteq \mathcal{X}$ be a ball of radius θ, and denote $\bar{T} := \mathcal{X} \setminus T$. Also let T_{big} be a ball cantered at the same point as T, but with twice the radius.

By the doubling dimension of the domain, it is possible to cover T_{big} with $\left(\frac{4\theta}{r}\right)^d$ small balls each of radius $r/2$. By Theorem 6.3.4, this implies that *any* r-packing of T is of size at most $\left(\frac{4\theta}{r}\right)^d$. In particular, $\mathcal{N} \cap T_{\text{big}}$ is of size at most $\left(\frac{4\theta}{r}\right)^d$. Now observe that any Voronoi cell that intersects T is contained in T_{big}. As every such Voronoi cell corresponds to a unique point in $\mathcal{N} \cap T_{\text{big}}$, we get that there are at most $\left(\frac{4\theta}{r}\right)^d$ Voronoi cell that intersects T. As we set $r = \frac{1}{n^{1/(4d)}}$, this quantity equals $(4\theta \cdot n^{1/(4d)})^d$.

$$\begin{aligned}
\Pr[N(x) \leq k] &\leq \sum_{V \in \mathcal{C} : V \cap T \neq \emptyset} \Pr(N(x) \leq k, x \in V) + \Pr(\bar{T}) \\
&\leq \sum_{V \cap T \neq \emptyset, \Pr(V) \leq 2k/n} \Pr(V) + \sum_{\substack{V \cap T \neq \emptyset \\ \Pr(V) > 2k/n}} Pr(V) \Pr\left(N(V) \leq \frac{k}{n}\right) + \Pr(\bar{T}) \\
&\leq \frac{2k}{n}(4\theta \cdot n^{1/(4d)})^d + \Pr(\bar{T}) + \sum_{\substack{V \cap T \neq \emptyset \\ \Pr(V) > 2k/n}} Pr(V) \Pr\left(N(V) - \mathbb{E}\left[N(V)\right] \leq \frac{k}{n} - \Pr(V)\right) \\
&\leq \frac{2k}{n}(4\theta \cdot n^{1/(4d)})^d + \Pr(\bar{T}) + \sum_{\substack{V \cap S \neq \emptyset \\ \Pr(V) > 2k/n}} Pr(V) \Pr\left(N(V) - \mathbb{E}\left[N(V)\right] \leq -\frac{\Pr(V)}{2}\right) \qquad (6.8)
\end{aligned}$$

From this point, the proof proceeds in the same steps as in Devroye et al. (2013, Theorem

6.2). By Chebyshev's inequality,

$$\begin{aligned}(6.8) &\leq \frac{2k}{n}(4\theta \cdot n^{1/(4d)})^d + \Pr(\bar{T}) + \sum_{V\cap S\neq\emptyset, \Pr(V)\geq 2k/n} 4\Pr(V)\frac{Var(N(V))}{\Pr(V)^2} \\ &\leq \frac{2k}{n}(4\theta \cdot n^{1/(4d)})^d + \Pr(\bar{T}) + \sum_{V\cap S\neq\emptyset, \Pr(V)\geq 2k/n} 4\Pr(V)\frac{\Pr(V)(1-\Pr(V))}{n\Pr(V)^2}\end{aligned}$$

$$\leq \frac{2k}{n}(4\theta \cdot n^{1/(4d)})^d + \Pr(\bar{T}) + \sum_{V\cap S\neq\emptyset, \Pr(V)\geq 2k/n} 4\Pr(V)\frac{\Pr(V)}{n\Pr(V)^2} \tag{6.9}$$

When the second inequality is due to the variance of the binomial variable $N(V)$.

$$\begin{aligned}(6.9) &\leq \frac{2k+4}{n}(4\theta \cdot n^{1/(4d)})^d + \Pr(\bar{T}) \\ &= \frac{(4\theta)^d}{n^{3/4}}(2k+4) + \Pr(\bar{T}) = \frac{4^d}{n^{1/4}}(2k+4) + \Pr(\bar{T})\end{aligned}$$

Clearly, the first summand goes to zero when $n \to \infty$ (recall that $k = o(n^{1/4})$). As for the second summand, recall that θ goes to ∞ when $n \to \infty$, and so $\Pr(\bar{T})$ goes to zero when $n \to \infty$. □

We will make use of the following theorem.

Theorem 6.3.11. *Given a separable metric space, a partition based classification rule is universally-consistent if*

1. *diam* $(V(x)) \xrightarrow[n\to\infty]{} 0$
2. *For every constant* $k \in \mathbb{N}$ *it holds that* $\Pr[N(x) \leq k] \xrightarrow[n\to\infty]{} 0$

This theorem is an extension of Devroye et al. (2013, Theroem 6.1), where it is stated only for $\mathbb{R}^d$. The proof of this theorem appears in Section 6.4 for completeness.

Putting it all together, we now prove the following theorem.

Theorem 6.3.12. *Algorithm 8 is universally-consistent.*

Proof. Define

- $\hat{\eta}_n(x) := \frac{1}{N(x)}\Sigma_{i:x_i\in A(x)} y_i$
- $\hat{\eta}_n^{\varepsilon}(x) := \hat{\eta}_n(x) + w_j$, where w_j is the noise added to $V(x)$.

The proof is close in nature to the proof of Theorem 6.1.2. We note, that algorithm PCL2 is a *plug-in classifier* w.r.t. $\hat{\eta}_n^{\varepsilon}$. Hence, by Theorem 3.1.11, in order to prove that it is consistent it suffices to show that

$$\lim_{n\to\infty} \mathbb{E}\left[|\hat{\eta}_n^{\varepsilon}(x) - \eta(x)|\right] = 0.$$

By the triangle inequality, $\mathbb{E}\left[|\hat{\eta}_n^\varepsilon(x) - \eta(x)|\right] \leq \mathbb{E}\left[|\hat{\eta}_n^\varepsilon(x) - \hat{\eta}_n(x)|\right] + \mathbb{E}\left[|\hat{\eta}_n(x) - \eta(x)|\right]$. By the same arguments as in Theorem 6.1.2 we get that

$$\mathop{\mathbb{E}}_{S,x}\mathop{\mathbb{E}}_{\mathcal{A}}\left[|\hat{\eta}_n^\varepsilon(x) - \hat{\eta}_n(x)|\right] \leq \frac{1}{\varepsilon}\left(\Pr(N(x) < M) + \frac{1}{M}\right) \tag{6.10}$$

Since this is true for every choice of M and by using Lemma 6.3.9, this also can be made arbitrarily small using sufficiently large sample size. Hence,

$$\mathbb{E}\left[|\hat{\eta}_n^\varepsilon(x) - \hat{\eta}_n(x)|\right] \xrightarrow{n\to\infty} 0. \tag{6.11}$$

In order to show that $\lim_{n\to\infty}\mathbb{E}\left[|\hat{\eta}_n(x) - \eta(x)|\right] = 0$, by Theorem 6.3.11, it suffices to show that the following two conditions hold:

1. $\mathit{diam}\,(V(x)) \xrightarrow[n\to\infty]{} 0$
2. $\Pr[N(x) \leq k] \xrightarrow[n\to\infty]{} 0$

The first condition follows from Lemma 6.3.8 and the second condition follows from Lemma 6.3.9. □

6.3.2 Unbounded Doubling Metric Spaces

We now extend the previous result to the case of *unbounded* doubling metric spaces. This extension comes at the cost of relaxing the privacy requirement from pure-privacy to approximated-privacy. Formally, we show the following theorem.

Theorem 6.3.13. *Let $\varepsilon \leq 1$ be a constant and let $\delta : \mathbb{N} \to [0,1]$ be a function satisfying $\delta(n) = \omega(2^{-n^{1/4}})$. There is an $(\varepsilon, \delta(n))$-differentially private universal consistent learner for every separable (possibly unbounded) metric space with finite doubling dimension.*

Let $(\mathcal{X}, \rho)$ be a separable doubling metric space with doubling dimension d. Consider Algorithm 9.

Note that, as $\mathcal{X}$ is separable, it has a countable covering and countable maximal packing for every r, and hence step 3 is well-defined. [2] Moreover, by Theorem 6.0.1 the number of non-empty cells will be finite, hence the hypothesis defined at step 10 is well-defined.

Theorem 6.3.14. *Algorithm 9 is $(2\varepsilon, 2\delta)$-differentially private.*

Proof of Theorem 6.3.14. As `Stability based Histogram` is (ε, δ)-differentially private, and since differential privacy is closed under post-processing, by standard composition

[2] Clearly, every separable space has a countable covering. As the cardinality of a packing can be bounded by the cardinality of a cover, we get that the cardinality of every packing must also be countable. Formally, given a $2r$-packing $\mathcal{N}$ and an r-cover $\mathcal{M}$, as every r-ball centered around a point in $\mathcal{M}$ contains at most one point from $\mathcal{N}$, we get that there is an injection from $\mathcal{N}$ to $\mathcal{M}$. Hence the cardinality of $\mathcal{N}$ is bounded by that of $\mathcal{M}$.

Algorithm 9 PCL2b

```
Input: Sample $S_n = \{(x_i, y_i)\}_{i=1}^n$
Set $r = \frac{1}{n^{1/(4d)}}$
Let $\mathcal{N}$ be a countable $r$ maximal packing of $\mathcal{X}$.
Partition the space into Voronoi cells centered in the elements of $\mathcal{N}$: $\mathcal{C} = V_1, V_2, \dots$.
For any $x$ denote $V(x)$ the cell $V$ s.t. $x \in V$
Apply Stability based Histogram with input $S_n$ to obtain estimates $\hat{c}_1, \hat{c}_2, \dots$ such
    that $\hat{c}_j \approx |\{x \in S_n : x \in V_j\}|$.
For any $x$ denote $\hat{c}(x) = \hat{c}_j$ such that $x \in V_j$.
Apply Stability based Histogram with input $S_n^1 := \{x \in S_n : y = 1\}$ to obtain
    estimates $\hat{y}_1, \hat{y}_2, \dots$ such that $\hat{y}_j \approx |\{x \in S_n : y = 1, x \in V_j\}|$.
For any $x$ denote $\hat{y}(x) = \min\{\hat{y}_j, \hat{c}_j\}$ such that $x \in V_j$.
Define the hypothesis $h_{\mathcal{C}}$ s.t. $h_{\mathcal{C}}(x) = \mathbb{1}_{\hat{y}(x) > \frac{\hat{c}(x)}{2}}$
Return $h_{\mathcal{C}}$
```

theorems the output of `PCL2b` is $(2\varepsilon, 2\delta)$-differentially private. □

Theorem 6.3.15. *Algorithm 9 is universally-consistent.*

Proof. Define

- $\hat{\eta}_n(x) := \frac{1}{N(x)} \sum_{i: x_i \in A(x)} y_i$
- $\hat{\eta}_n^{\varepsilon,\delta}(x) := \begin{cases} \frac{\hat{y}(x)}{\hat{c}(x)} & \hat{c}(x) \neq 0 \\ 0 & \hat{c}(x) = 0 \end{cases}$.

Most of the arguments which were made for Algorithm 8 in the proof of Theorem 6.3.12 can be made also for Algorithm 9. The only part of the proof that requires attention is to show that $\lim_{n\to\infty} \mathbb{E}\left[|\hat{\eta}_n^{\varepsilon,\delta}(x) - \hat{\eta}_n(x)|\right] = 0$. We calculate,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S,x,\mathcal{A}}\left[|\hat{\eta}_n^{\varepsilon,\delta}(x) - \hat{\eta}_n(x)|\right] \\
&= \mathop{\mathbb{E}}_{S,x}\left[\mathop{\mathbb{E}}_{\mathcal{A}}\left[|\hat{\eta}_n^{\varepsilon,\delta}(x) - \hat{\eta}_n(x)|\right]\right] \\
&\leq \mathop{\mathbb{E}}_{S,x}\left[\mathop{\mathbb{E}}_{\mathcal{A}}\left[|\hat{\eta}_n^{\varepsilon,\delta}(x) - \hat{\eta}_n(x)|\right] \cdot \mathbb{1}_{N(x)>0}\right] + \Pr[N(x) = 0] \\
&= \mathop{\mathbb{E}}_{S,x}\left[\mathop{\mathbb{E}}_{\mathcal{A}}\left[\left|\frac{\hat{y}(x)}{\hat{c}(x)} - \frac{\sum_{i:x_i \in A(x)} y_i}{N(x)}\right|\right] \cdot \mathbb{1}_{N(x)>0}\right] + \Pr[N(x) = 0] \\
&\leq \mathop{\mathbb{E}}_{S,x}\left[\mathop{\mathbb{E}}_{\mathcal{A}}\left[\left|\frac{\sum_{i:x_i \in A(x)} y_i}{N(x)} - \frac{\hat{y}(x)}{N(x)}\right| + \left|\frac{\hat{y}(x)}{N(x)} - \frac{\hat{y}(x)}{\hat{c}(x)}\right|\right] \cdot \mathbb{1}_{N(x)>0}\right] \\
&\quad + \Pr[N(x) = 0] \\
&\leq \mathop{\mathbb{E}}_{S,x}\left[O\left(\frac{\frac{1}{\varepsilon}\log\frac{1}{\delta}}{N(x)}\right) \cdot \mathbb{1}_{N(x)>0}\right] + \Pr[N(x) = 0]
\end{aligned}$$

$$
\begin{aligned}
&\approx \frac{1}{\varepsilon}\log\frac{1}{\delta}\cdot \mathop{\mathbb{E}}_{S,x}\left[\frac{1}{N(x)}\cdot \mathbb{1}_{N(x)>0}\right] + \Pr[N(x)=0] \\
&= \frac{1}{\varepsilon}\log\frac{1}{\delta}\cdot \Bigg(\mathbb{E}\left[\frac{1}{N(x)}\cdot \mathbb{1}_{N(x)>0}\middle| N(x) < M\right] \\
&\quad \cdot \Pr[N(x) < M] + \mathbb{E}\left[\frac{1}{N(x)}\middle| N(x)\geq M\right]\cdot \Pr[N(x)\geq M]\Bigg) \\
&\quad + \Pr[N(x)=0] \\
&\leq \frac{1}{\varepsilon}\log\frac{1}{\delta}\cdot \left(\Pr[N(x)<M] + \frac{1}{M}\right) + \Pr[N(x)=0] \\
&\leq \frac{2}{\varepsilon}\log\frac{1}{\delta}\left(\Pr(N(x)<M)+\frac{1}{M}\right) \qquad (6.12)
\end{aligned}
$$

Since this is true for every choice of M and by using Lemma 6.3.9, this also can be made arbitrarily small using sufficiently large sample size [3]. Hence,

$$
\mathbb{E}\left[|\hat{\eta}_n^{\varepsilon,\delta}(x) - \hat{\eta}_n(x)|\right] \xrightarrow{n\to\infty} 0. \qquad (6.13)
$$

□

Remark 6.3.16. *Unlike our results for the (unbounded) euclidean case, where we showed a construction for a* density estimator*, for (unbounded) metric spaces with finite doubling dimension we only show a* learner*. The reason is that in our construction of a density estimator for the euclidean case we needed to compute* volumes *of the cells in the partition. In general metric spaces, however, we do not have a canonical analogue for the volume of a cell.*

6.4 Additional Details for Completeness

The proofs provided in this section are taken from Devroye et al. (2013). We include them here for completeness, as in Devroye et al. (2013) these theorems are stated only for $\mathbb{R}^d$.

Theorem 6.4.1. *For a probability space $(\mathcal{X},\mu)$, let $\hat{\eta}:\mathcal{X}\to[0,1]$ be any function, and let $\hat{h}$ be the* plug-in classification rule *w.r.t. $\hat{\eta}$. Then the following holds*

$$
\Pr_{(X,Y)}[\hat{h}(X)\neq Y] - L^* \leq 2\,\mathbb{E}\left[|\eta(X)-\hat{\eta}(X)|\right]
$$

[3] Note that, δ can decay exponentially fast as a function of M and hence also as a function of n, allowing the same $\delta(n)=\omega(2^{-\sqrt{n}})$ dependency as in Theorem 1.2.6

Proof of Theorem 6.4.1. Given $x \in \mathcal{X}$, if $h^*(x) = \hat{h}(x)$, then

$$\Pr_{(X,Y)}[\hat{h}(X) \neq Y \mid X = x] = \Pr_{(X,Y)}[h^*(X) \neq Y \mid X = x].$$

On the other hand if $h^*(x) \neq \hat{h}(x)$, then

$$|\eta(X) - \hat{\eta}(X)| \geq |\eta(X) - \frac{1}{2}|.$$

Therefore

$$\begin{aligned}&\Pr_{(X,Y)}[h^*(X) \neq Y \mid X = x] - \Pr_{(X,Y)}[\hat{h}(X) \neq Y \mid X = x] \\ &= (2\eta(x) - 1)(\mathbb{1}_{h^*(X)=1} - \mathbb{1}_{\hat{h}(x)=1}) = |2\eta(x) - 1| \cdot \mathbb{1}_{h^*(X)\neq\hat{h}(x)}\end{aligned}$$

By the law of total probability

$$\begin{aligned}&\Pr_{(X,Y)}[\hat{h}(X) \neq Y] - L^* \\ &= \int_{x\in\mathcal{X}} \Pr_{(X,Y)}[h^*(X) \neq Y \mid X = x] - \Pr_{(X,Y)}[\hat{h}(X) \neq Y \mid X = x]dx \\ &= \int_{x\in\mathcal{X}} |2\eta(x) - 1| \cdot \mathbb{1}_{h^*(X)\neq\hat{h}(x)}\mu(x)dx \\ &= \int_{x\in\mathcal{X}} 2|\eta(x) - \frac{1}{2}| \cdot \mathbb{1}_{h^*(X)\neq\hat{h}(x)}\mu(x)dx \\ &= \mathbb{E}\left[2|\eta(X) - \frac{1}{2}| \cdot \mathbb{1}_{h^*(X)\neq\hat{h}(X)}\right] \\ &\leq 2\,\mathbb{E}\left[|\eta(X) - \hat{\eta}(X)|\right]\end{aligned}$$

□

Proof of Theorem 6.3.11. As any partition rule is a special case of a plug-in estimator, we need to show that $\mathbb{E}\left[|\eta(x) - \hat{\eta}_n(x)|\right] \xrightarrow[n\to\infty]{} 0$. Define $\bar{\eta}(x) := \mathbb{E}[\eta(z) \mid z \in V(x)]$. By the triangle inequality

$$\mathbb{E}\left[|\eta(x) - \hat{\eta}_n(x)|\right] \leq \mathbb{E}\left[|\eta(x) - \bar{\eta}(x)|\right] + \mathbb{E}\left[|\bar{\eta}(x) - \hat{\eta}_n(x)|\right]$$

Examine the random variable $N(x)\hat{\eta}_n(x)$, which is the number of labeled-one points falling in the same "bucket" as x. By conditioning upon which points fall in this bucket, the remaining randomness in this r.v. is only which of them will be labeled one. This is then simply a binomial random variable with "success" probability $\bar{\eta}(x)$ and $N(x)$ trials.

Thus,

$$\begin{aligned}
&\mathbb{E}\left[|\bar{\eta}(x)-\hat{\eta}_n(x)| \mid \mathbb{1}_{x_1\in V(X)},\dots,\mathbb{1}_{x_n\in V(X)}\right]\\
&\le \mathbb{E}\left[|\frac{N(x)\hat{\eta}_n(x)}{N(x)}-\bar{\eta}(x)| \mid \mathbb{1}_{N(x)>0},\mathbb{1}_{x_1\in V(X)},\dots,\mathbb{1}_{x_n\in V(X)}\right]\\
&\le \mathbb{E}\left[\left(\frac{N(x)\hat{\eta}_n(x)}{N(x)}-\bar{\eta}(x)\right)^2 \mid \mathbb{1}_{N(x)>0},\mathbb{1}_{x_1\in V(X)},\dots,\mathbb{1}_{x_n\in V(X)}\right]^{1/2}\\
&\le \mathbb{E}\left[\frac{\bar{\eta}(x)(1-\bar{\eta}(x))}{N(x)}\mathbb{1}_{N(x)>0} \mid \mathbb{1}_{x_1\in V(X)},\dots,\mathbb{1}_{x_n\in V(X)}\right]^{1/2} \qquad (6.14)
\end{aligned}$$

When the second inequality is by the Jensen inequality and the third by the variance of a binomial distribution. Next, note that $\bar{\eta}(x)(1-\bar{\eta}(x))\le\frac{1}{4}$ and hence,

$$\begin{aligned}
(6.14) &\le \mathbb{E}\left[\frac{1}{4N(X)} \mid N(X)>0\right]^{1/2}\Pr[N(n)>0]+\Pr[N(X)=0]\\
&\le \mathbb{E}\left[\frac{1}{4N(X)} \mid N(X)>0\right]^{1/2}\Pr[N(n)>0]+\Pr[N(X)=0]\\
&\le \frac{1}{2}\Pr[N(x)\le k]+\frac{1}{2\sqrt{k}}+\Pr[N(X)=0] \qquad (6.15)
\end{aligned}$$

This is true for any k. Therefore (6.15) can be made arbitrarily small by choosing k large enough and then by condition (2) in the theorem's conditions.

Moving on to the first summand. For any $\tau>0$, there exists a uniform continuous real-valued function η_τ, such that $\mathbb{E}\left[|\eta(x)-\eta_\tau(x)|\right]<\tau$. Such a function exists, since for a separable metric space, the set of uniformly continuous, real valued functions is dense, in ℓ_1 norm, in the set of all continuous, real valued, functions. Define $\bar{\eta}_\tau(x):=\mathbb{E}[\eta_\tau(z)\mid z\in V(x)]$ and by the triangle inequality,

$$\begin{aligned}
&\mathbb{E}\left[|\eta(x)-\bar{\eta}(x)|\right]\\
&\le \mathbb{E}\left[|\eta(x)-\eta_\tau(x)|\right]+\mathbb{E}\left[|\eta_\tau(x)-\bar{\eta}_\tau(x)|\right]+\mathbb{E}\left[|\bar{\eta}_\tau(x)-\bar{\eta}(x)|\right]\\
&=: (*)+(**)+(***).
\end{aligned}$$

By the choice of $\eta_\tau(x)$, the $(*)\le\tau$. Also, by the definitions for $\bar{\eta}$ and $\bar{\eta}_\tau(x)$ the $(***)\le(*)\le\tau$. Finally, as $\eta_\tau(x)$ is uniformly continuous, there exist some θ s.t. the difference between points which are θ-close is bounded by τ. Hence, we get that $(**)\le\tau+\Pr(\mathit{diam}\,(V(x))>\theta)$, when by condition (1) of the theorem's conditions, can be made less than τ for large enough n. All in all, we showed that for any given τ we can ensure that $\mathbb{E}\left[|\eta(x)-\bar{\eta}(x)|\right]<\tau$, for large enough n. □

Chapter 7

Adaptive Data Analysis

In this chapter, we investigate various extensions enabling adaptive data analysis for correlated observations. We show that some of the directions and tools from the literature can be used for this setting, if we look at the right setting or the appropriate measurements.

7.1 Adaptive Generalization via Differential Privacy

We start by extending the connection between differential privacy and adaptive data analysis into settings where the data is not sampled in an i.i.d. fashion, but rather there are some small/bounded dependencies. We start by proving the following lemma, showing that differential privacy guarantees generalization in expectation. The proof of this lemma mimics the analysis of Bassily et al. (2016) for the i.i.d. setting. We extend the proof to the case where there are dependencies in the data, and show that we can "pay" for these dependencies in a way that scales with ψ.

Lemma 7.1.1 (Expectation bound). *Let $\mathcal{A}' : (\mathcal{X}^n)^T \to 2^{\mathcal{X}} \times [T]$ be an (ε, δ)-differentially private algorithm. Let μ be a distribution over $\mathcal{X}^n$ which has ψ-Gibbs-dependence. Let $\vec{S} = (S_1, \ldots, S_T)$ where for every i $S_i \sim \mu$. Denote by (h, t) the output of $\mathcal{A}'(\vec{S})$. Then $\left|\mathbb{E}_{\vec{S},\mathcal{A}'}\left[h(\mu) - h(S_t)\right]\right| \leq e^{\varepsilon} + T\delta + \psi - 1$.*

Proof. We consider a *multi sample* $\vec{S} = (S_1, \ldots, S_T)$, where $S_t = (x_{t,1}, \ldots, x_{t,n}) \sim \mu$. We calculate,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T}\Bigg[\mathop{\mathbb{E}}_{(h.t)\sim\mathcal{A}'(\vec{S})}[h(S_t)]\Bigg] \\
=&\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T}\Bigg[\mathop{\mathbb{E}}_{(h.t)\sim\mathcal{A}'(\vec{S})}\Bigg[\frac{1}{n}\sum_{i=1}^{n}h(x_{t,i})\Bigg]\Bigg] \\
=&\frac{1}{n}\sum_{i=1}^{n}\Bigg[\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T}\Bigg[\mathop{\mathbb{E}}_{(h.t)\sim\mathcal{A}'(\vec{S})}[h(x_{t,i})]\Bigg]\Bigg] \\
=&\frac{1}{n}\sum_{i=1}^{n}\Bigg[\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T}\Bigg[\mathop{\Pr}_{(h.t)\sim\mathcal{A}'(\vec{S})}[h(x_{t,i})=1]\Bigg]\Bigg] \\
=&\frac{1}{n}\sum_{i=1}^{n}\Bigg[\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T}\Bigg[\sum_{m=1}^{T}\mathop{\Pr}_{(h.t)\sim\mathcal{A}'(\vec{S})}[h(x_{m,i})=1\wedge t=m]\Bigg]\Bigg] \\
=&\frac{1}{n}\sum_{i=1}^{n}\Bigg[\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T}\Bigg[\mathop{\mathbb{E}}_{\vec{z}\sim\vec{\mu}_i|\vec{S}}\Bigg[\\
&\sum_{m=1}^{T}\mathop{\Pr}_{(h.t)\sim\mathcal{A}'(\vec{S})}[h(x_{m,i})=1\wedge t=m]\Bigg]\Bigg]\Bigg],
\end{aligned} \tag{7.1}$$

where $\vec{z}=(z_1,\ldots,z_T)$ is a vector s.t. $z_t\sim\mu_i(\cdot\mid S_t^{-i})$. Given a multi-sample $\vec{S}$ and an element z, we write $\vec{S}^{(m,i)\leftarrow z}$ to denote the multi-sample $\vec{S}$ after replacing the i^{th} element in the m^{th} sample S_m with z. Since $\mathcal{A}'$ is (ε,δ)-differentially private we get that the above is at most

$$\begin{aligned}
(7.1)\le&\frac{1}{n}\sum_{i=1}^{n}\Bigg[\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T}\Bigg[\mathop{\mathbb{E}}_{\vec{z}\sim\vec{\mu}_i|\vec{S}}\Bigg[\sum_{m=1}^{T}e^{\varepsilon}\mathop{\Pr}_{(h.t)\sim\mathcal{A}'(\vec{S}^{(m,i)\leftarrow z_m})} \\
&\Bigg[h(x_{m,i})=1\wedge t=m\Bigg]+\delta\Bigg]\Bigg]\Bigg] \\
=&T\delta+e^{\varepsilon}\cdot\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{T}\Bigg[\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T}\Bigg[\mathop{\mathbb{E}}_{\vec{z}\sim\vec{\mu}_i|\vec{S}}\Bigg[\\
&\mathop{\Pr}_{(h.t)\sim\mathcal{A}'(\vec{S}^{(m,i)\leftarrow z_m})}\Bigg[h(x_{m,i})=1\wedge t=m\Bigg]\Bigg]\Bigg]\Bigg] \\
=&T\delta+e^{\varepsilon}\cdot\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{T}\Bigg[\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T}\Bigg[\mathop{\mathbb{E}}_{\vec{z}\sim\vec{\mu}_i|\vec{S}}\Bigg[\\
&\mathop{\Pr}_{(h.t)\sim\mathcal{A}'(\vec{S})}[h(z_m)=1\wedge t=m]\Bigg]\Bigg]\Bigg]
\end{aligned}$$

$$
\begin{aligned}
&= T\delta + e^{\varepsilon} \cdot \frac{1}{n} \sum_{i=1}^{n} \sum_{m=1}^{T} \left[\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T} \left[\mathop{\mathbb{E}}_{\vec{z}\sim\vec{\mu}_i|\vec{S}} \left[\right.\right.\right. \\
&\quad \left.\left.\left. \Pr_{(h.t)\sim\mathcal{A}'(\vec{S})} [h(z_t) = 1 \wedge t = m] \right]\right]\right] \\
&= T\delta + e^{\varepsilon} \cdot \frac{1}{n} \sum_{i=1}^{n} \left[\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T} \left[\mathop{\mathbb{E}}_{\vec{z}\sim\vec{\mu}_i|\vec{S}} \left[\right.\right.\right. \\
&\quad \left.\left.\left. \sum_{m=1}^{T} \Pr_{(h.t)\sim\mathcal{A}'(\vec{S})} [h(z_t) = 1 \wedge t = m] \right]\right]\right] \\
&= T\delta + e^{\varepsilon} \cdot \frac{1}{n} \sum_{i=1}^{n} \left[\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T} \left[\right.\right. \\
&\quad \left.\left. \mathop{\mathbb{E}}_{\vec{z}\sim\vec{\mu}_i|\vec{S}} \left[\Pr_{(h.t)\sim\mathcal{A}'(\vec{S})} [h(z_t) = 1] \right] \right]\right] \\
&= T\delta + e^{\varepsilon} \frac{1}{n} \sum_{i=1}^{n} \left[\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T} \left[\mathop{\mathbb{E}}_{\vec{z}\sim\vec{\mu}_i|\vec{S}} \left[\mathop{\mathbb{E}}_{(h.t)\sim\mathcal{A}'(\vec{S})} [h(z_t)] \right]\right]\right] \\
&= T\delta + e^{\varepsilon} \frac{1}{n} \sum_{i=1}^{n} \left[\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T} \left[\mathop{\mathbb{E}}_{(h.t)\sim\mathcal{A}'(\vec{S})} \left[\mathop{\mathbb{E}}_{\vec{z}\sim\vec{\mu}_i|\vec{S}} [h(z_t)] \right]\right]\right] \\
&= T\delta + e^{\varepsilon} \cdot \frac{1}{n} \sum_{i=1}^{n} \left[\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T} \left[\right.\right. \\
&\quad \left.\left. \mathop{\mathbb{E}}_{(h.t)\sim\mathcal{A}'(\vec{S})} \left[\mathop{\mathbb{E}}_{z\sim\mu_i(\cdot|S_t^{-i})} [h(z)] \right]\right]\right]. \qquad (7.2)
\end{aligned}
$$

Since total variation is a special case of the Wasserstein metric $\mathcal{W}_1$, Kantorovich-Rubinstein duality implies that for two probability measures μ, ν on a space $\mathcal{X}$ and any function $h : \mathcal{X} \to [0,1]$, we have $|\mathbb{E}_{z\sim\mu}[h(x)] - \mathbb{E}_{z\sim\nu}[h(z)]| \le \|\mu - \nu\|_{\mathsf{TV}}$. Applying this to $\mu_i(\cdot \mid S_t^{-i})$ and μ_i we get that the above is at most

$$
\begin{aligned}
(7.2) &\le T\delta + e^{\varepsilon} \cdot \frac{1}{n} \sum_{i=1}^{n} \left[\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T} \left[\mathop{\mathbb{E}}_{(h.t)\sim\mathcal{A}'(\vec{S})} \left[\right.\right.\right. \\
&\quad \left.\left.\left. \mathop{\mathbb{E}}_{z\sim\mu_i} [h(z)] + \left\|\mu_i(\cdot \mid S_t^{-i}) - \mu_i\right\|_{\mathsf{TV}} \right]\right]\right] \\
&\le \psi + T\delta + e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{\vec{S}\sim\mu^T} \left[\mathop{\mathbb{E}}_{(h.t)\sim\mathcal{A}'(\vec{S})} \left[\right.\right. \\
&\quad \left.\left. \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{z\sim\mu_i} [h(z)] \right]\right] \\
&= \psi + T\delta + e^{\varepsilon} \cdot \mathop{\mathbb{E}}_{\vec{S},\mathcal{A}'(\vec{S})} [h(\mu)] \\
&\le \psi + T\delta + e^{\varepsilon} - 1 + \mathop{\mathbb{E}}_{\vec{S},\mathcal{A}'(\vec{S})} [h(\mu)],
\end{aligned}
$$

where the last inequality is due to the fact that $ye^{\varepsilon} \leq e^{\varepsilon} - 1 + y$ for $y \leq 1$ and $\varepsilon \geq 0$. In summary,

$$\begin{aligned} &\mathop{\mathbb{E}}_{\vec{S}\sim\mu^T}\left[\mathop{\mathbb{E}}_{(h.t)\sim\mathcal{A}'(\vec{S})}[h(S_t)]\right] \\ &\qquad \leq \psi + T\delta + e^{\varepsilon} - 1 + \mathop{\mathbb{E}}_{\vec{S},\mathcal{A}'(\vec{S})}[h(\mu)]\,. \end{aligned}$$

The other direction is symmetric. □

We use Lemma 7.1.1 to prove the following high-probability generalization bound for differentially private algorithms.

Theorem 7.1.2 (High probability bound)**.** *Let $\varepsilon \in (0, 1/3)$, $\delta \in (0, \varepsilon/4)$ and $n \geq \frac{\log(2k\varepsilon/\delta)}{\varepsilon^2}$. Let $\mathcal{A} : \mathcal{X}^n \to (2^{\mathcal{X}})^k$ be an (ε, δ)-differentially private algorithm. Let μ be a distribution over $\mathcal{X}^n$ and S be a sample of size n drawn from μ, and let $h_1, \ldots, h_k$ be the output of $\mathcal{A}(S)$. Then*

$$\Pr_{S,\mathcal{A}(S)}\left[\max_{i\in[k]}|h_i(\mu) - h_i(S)| \geq 10\varepsilon + 2\psi\right] \leq \frac{\delta}{\varepsilon}.$$

The proof of Theorem 7.1.2 is almost identical to the analysis of Bassily et al. (2016). It appears in the appendix for completeness. Intuitively, the proof is as follows. We assume, towards contradiction, that there may be a differentially private algorithm that does not enjoy strong generalization guarantees. We then use this mechanism to describe a different differentially private algorithm with a "boosted inability" to generalize. That is, the proof goes by saying that if there is a differentially private algorithm whose generalization properties are not "very good" then there must exist a differentially private algorithm whose generalization properties are "bad", to the extent that contradicts Lemma 7.1.1.

Our connection between Gibbs-dependence and differential privacy (Theorem 1.3.2) now follows as a corollary of Theorem 7.1.2.

Proof of Theorem 1.3.2. $\mathcal{M}$ is (ε, δ)-differentially private. Since $\mathbb{A}$ can only access the data via $\mathcal{M}$, we can treat the pair $\mathbb{A}, \mathcal{M}$ as a single algorithm $\mathcal{A}$, which gets a sample $S \sim \mu$ as input and returns k predicates, as output. By closure to post-processing, $\mathcal{A}$ is also (ε, δ)-differentially private. Applying Theorem 7.1.2 on $\mathcal{A}$ we get that

$$\Pr\left[\max_{i\in[k]}|h_i(\mu) - h_i(S)| \geq 10\varepsilon + 2\psi\right] \leq \frac{\delta}{\varepsilon}.$$

Since $\mathcal{M}$ is (α,β)-empirically-accurate it holds that

$$\Pr\left[\max_{i\in[k]}|q_i(S)-a_i|>\alpha\right]\le\beta.$$

Combining these two bounds with the triangle inequality, we get

$$\Pr\left[\max_{i\in[k]}|q_i(\mu)-a_i|>\alpha+10\varepsilon+2\psi\right]<\beta+\frac{\delta}{\varepsilon}.$$

□

7.1.1 A Tight Negative Result for Differential Privacy and Gibbs-Dependence

In this section, we construct a distribution which is ψ-Gibbs-Dependant, and describe a differentially-private algorithm whose generalization gap w.r.t. this distribution is at least ψ. Hence, in general, the ψ factor attained on Theorem 1.3.2 is tight up to a constant. Let $\mathcal{X}=[0,1]$ and define a measure μ over $\mathcal{X}^n$ by the following random process:

1. Sample a point $x^*\sim U([0,1])$.
2. For every $i\in[n]$:
   (a) Sample $\sigma\sim\text{Ber}(\psi)$.
      i. If $\sigma=1$ then $x_i=x^*$.
      ii. Otherwise $x_i\sim U([0,1])$
3. Return $S=(x_1,\dots,x_n)$

Lemma 7.1.3. *The measure defined by the above process has ψ-Gibbs-dependency.*

Proof. Initially, every marginal distribution is just uniform, i.e. $\mu_i\sim U([0,1])$ and hence, for every $A\subseteq[0,1]$ it holds that $\mu_i(A)=|A|$. After conditioning, for every possible x^{-i} and x^*, we get that

$$\begin{aligned}&\mu_i(A\mid x^{-i},x^*)\\=&\mu_i(A\setminus\{x^*\}\mid x^{-i},x^*)+\mu_i(A\cap\{x^*\}\mid x^{-i},x^*)\\&\in\Big(|A|(1-\psi),\ |A|(1-\psi)+\psi\Big).\end{aligned}$$

Since the above holds for every choice of x^*, we also have that

$$\mu_i(A\mid x^{-i})\in\Big(|A|(1-\psi),\ |A|(1-\psi)+\psi\Big).$$

Therefore, for every $A \subseteq [0,1]$ it holds that

$$\begin{aligned}&|\mu_i(A) - \mu_i(A \mid x^{-i})| \\ &\leq \max\{|A| - |A|(1-\psi)\ ,\ |A|(1-\psi) + \psi - |A|\} \leq \psi.\end{aligned}$$

So $\|\mu_i(\cdot) - \mu_i(\cdot \mid x^{-i})\|_{\mathsf{TV}} \leq \psi$. Plugging this bound to the Gibbs-dependency definition yields

$$\psi(\mu) = \sup_{x \in \mathcal{X}^n} \mathop{\mathbb{E}}_{i \sim [n]} \left\|\mu_i(\cdot) - \mu_i(\cdot \mid x^{-i})\right\|_{\mathsf{TV}} \leq \psi.$$

□

We next describe an algorithm that, despite being differentially private, performs "badly" when executed on samples from the above measure μ. Specifically, this algorithm is capable of identifying a predicate with generalization error $\Omega(\psi)$. This shows that our connection between differential privacy and generalization (in the correlated setting) is tight, in the sense that the generalization error of differentially private algorithms *can* grow with ψ. This matches our positive result (see Theorem 1.3.2).

Our algorithm is specified in Algorithm 10. As a subroutine, we use the following result of Bun et al. (2019b) for privately computing histograms.

Theorem 7.1.4 (Private histograms, Bun et al. (2019b)). *There exists an (ε, δ)-differentially private algorithm that takes an input dataset $S \in \mathcal{X}^n$ and returns an a list $L \subseteq \mathcal{X}$ such that with probability at least $1 - \beta$, every $x \in \mathcal{X}$ that appears at least $\mathcal{O}\left(\frac{1}{\varepsilon}\log\frac{1}{\beta\delta}\right)$ times in S is included in L, and furthermore, every $x \in L$ appears at least* twice *in S.*

Algorithm 10 Deviating Private Algorithm

```
Input: A sample $S$, privacy parameters $\varepsilon, \delta$.
Tool used: An $(\varepsilon, \delta)$-DP algorithm $\mathcal{H}$ for histograms.
$L \leftarrow \mathcal{H}(S, \varepsilon, \delta)$
if $L$ is empty then
    Return $h \equiv 0$
else
    Let $x$ be an arbitrary element in $L$
    Define $h : \mathcal{X} \to [0,1]$ as $h = \mathbb{1}_x$
    Return $h$
```

Lemma 7.1.5. *For every $\beta > 0$, every $n \geq \mathcal{O}\left(\frac{1}{\psi\varepsilon}\log\frac{1}{\beta\delta}\right)$, and for every $\psi < 1$ Algorithm 10 is (ε, δ)-differentially private and it outputs a predicate $h : \mathcal{X} \to [0,1]$ s.t.*

$$\Pr\left[|h(S) - h(\mu)| \geq \frac{\psi}{2}\right] > 1 - \beta - \exp\left(-\frac{n}{8}\right).$$

Proof. First observe that Algorithm 10 is (ε, δ)-differentially private, as it merely postprocesses the outcome of the private histogram algorithm.

Next observe that, by the definition of the underlying measure μ, and by our choice of n, w.h.p., there are many copies of x^* in the dataset S. Formally, by the Chernoff bound,

$$\begin{aligned}&\Pr\left[\frac{1}{n}|\{x' \in S \mid x' = x^*\}| < \frac{1}{2}\psi\right]\\&= \Pr\left[\frac{1}{n}\sum_{i=1}^{n}\sigma_i < \frac{1}{2}\psi\right] \leq \exp\left(-\frac{n}{8}\right).\end{aligned}$$

In addition, the probability of any element $x \neq x^*$ appearing more than once in S is simply zero. Thus, with probability at least $1 - \exp\left(-\frac{n}{8}\right)$ we have that x^* appears in S at least $n\psi/2 = \Omega(\frac{1}{\varepsilon}\log(\frac{1}{\beta}\delta))$ times, and every other element appears in S at most once. By the properties of the private histogram algorithm (see Theorem 7.1.4), in such a case, with probability at least $1 - \beta$ we have that $L = \{x^*\}$, and Algorithm 10 returns the hypothesis $h = \mathbb{1}_{x^*}$. As x^* appears many times in S, this predicate has "large" empirical value. On the other hand, for such predicate it holds that

$$\begin{aligned}h(\mu) &= \mathop{\mathbb{E}}_{\bar{x^*},\bar{x_1},\ldots,\bar{x_n}}\left[\frac{1}{n}\left(\sum_{i=1}^{n}h(\bar{x_i})\right)\right]\\&= \frac{1}{n}\left(\sum_{i=1}^{n}\Pr_{\bar{x^*},\bar{x_1},\ldots,\bar{x_n}}[\bar{x_i} = x^*]\right) = 0\end{aligned}$$

as the probability that for a fresh new sampling we will get $\bar{x^*} = x^*$ is zero, implying that the probability that any point in the sample to be x^* is also zero.

Overall, with probability at least $1 - \beta - \exp\left(-\frac{n}{8}\right)$, the algorithm returns a predicate h such that $h(S) \geq \psi/2$ but $h(\mu) = 0$.

□

7.2 Adaptive Learning Via Transcript Compression

In this section we show how the notion of *transcript compressibility* can be used to derive generalization bounds even if the data is not i.i.d. distributed. We start by recalling the notion of *transcript compression* by Dwork et al. (2015a). We denote by $AG_{n,k}(\mathcal{A}, \mathcal{M}, S)$ the *transcript* of the interaction between the mechanism $\mathcal{M}$ and the analysis $\mathcal{A}$ during the adaptive accuracy game defined in Algorithm 1 with sample of size n and k queries.

Definition 7.2.1 (Transcript Compression (Dwork et al., 2015a)). *We say that a mechanism $\mathcal{M}$ enables* transcript compression *to $b(n,k)$-bits, if for every deterministic analyst $\mathcal{A}$ there exist a set of possible transcripts $\mathcal{H}_{\mathcal{A}}$, of size $|\mathcal{H}_{\mathcal{A}}| \leq 2^{b(n,k)}$, s.t. for every sample S it holds that* $\Pr[AG_{n,k}(\mathcal{A}, \mathcal{M}, S) \in \mathcal{H}] = 1$.

Following Bassily and Freund (2016), in this section we aim to design mechanisms that answer adaptively chosen queries while providing statistical accuracy, under the assumption that the given queries are *concentrated* around their expected value. Unlike Bassily and Freund (2016), we aim to achieve this goal using the notion of *transcript compression*, rather than *typical-stability.* As we show, this allows for a significantly simpler analysis (and definitions). Formally,

Definition 7.2.2. *Given a measure μ over $\mathcal{X}$, a query $q : \mathcal{X}^n \to \mathbb{R}$, and a parameter $\delta \in [0,1]$, we write $\gamma(q,\mu,\delta)$ to denote the minimal number $\gamma \in [0,1]$ such that*

$$\Pr_{S\sim\mu}\left[\left|q(S) - \mathop{\mathbb{E}}_{T\sim\mu}[q(T)]\right| > \gamma\right] < \delta.$$

That is, $\gamma(q,\mu,\delta)$ denotes the minimal number such that, without adaptivity, $q(S)$ deviates from its expectation by more than $\gamma(q,\mu,\delta)$ with probability at most δ when sampling $S \sim \mu$.

Remark 7.2.3. *The results in this section are not restricted to statistical queries. The results in this section hold for arbitrary queries (mapping n-tuples to the reals).*

Consider again Algorithm 1 and Definition 3.1.15 (the definition of statistical accuracy). We now use Definition 7.2.2 in order to introduce a relaxation for statistical accuracy, in which the mechanism is allowed to incur $\gamma(q,\mu,\delta)$ as an additional error.

Definition 7.2.4. *A mechanism $\mathcal{M}$ is (α,β,δ)-*statistically-query-accurate *for k rounds given n samples, if for every distribution μ over n-tuples, and every adversary $\mathbb{A}$, it holds that*

$$\Pr_{\substack{S\sim\mu \\ \mathtt{Game}(\mathcal{M},k,\mathbb{A},S)}}\left[\max_{i\in[k]}|q_i(\mu) - a_i| > \alpha + \gamma(q_i,\mu,\delta)\right] \leq \beta.$$

Remark 7.2.5. *For a statistical query q and a product measure μ, by Hoeffding's inequality, we get that $\gamma(q,\mu,\delta) = \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}$. Hence, for the i.i.d. regime, for large enough samples, the definition of (α,β,δ)-statistical-query-accuracy is in fact equivalent (up to factor 2) to the original definition of (α,β)-statistical-accuracy (Definition 3.1.15).*

We observe that the analysis of Dwork et al. (2015a) for transcript compression easily extends to non-i.i.d. measures when given concentrated queries. Somewhat surprisingly, this simple technique essentially matches the bounds obtained using typical stability (Bassily and Freund, 2016). In the next lemma we show that (w.h.p.) an analyst interacting with a transcript-compressing mechanism cannot identify a query that overfits to the data.

Lemma 7.2.6. *Let $\mathcal{M}$ be a mechanism which enables transcript compression to $b(n,k)$-*

bits. For every measure μ and every analyst $\mathcal{A}$,

$$\Pr_{S,AG_{n,k}}[\exists i : |q_i(S) - q_i(\mu)| \geq \gamma(q,\mu,\delta)] \leq \delta \cdot k \cdot 2^{b(n,k)}$$

Proof. Fix an analyst $\mathcal{A}$. By Definition 7.2.1, there exist a set of transcripts $H_{\mathcal{A}}$ of size at most $2^{b(n,k)}$. As every transcript consists of at most k queries, there can be at most $k2^{b(n,k)}$ possible queries over all possible interactions between $\mathcal{A}$ and $\mathcal{M}$. Denote this set of possible queries as $Q_{\mathcal{A}}$. By a union bound we get that

$$\Pr_{S\sim\mu}\left[\bigvee_{q\in Q_{\mathcal{A}}} |q_i(S) - q_i(\mu)| \geq \gamma(q,\mu,\delta)\right] \leq \delta \cdot k \cdot 2^{b(n,k)},$$

and hence

$$\Pr_{S,AG_{n,k}}[\exists i : |q_i(S) - q_i(\mu)| \geq \gamma(q,\mu,\delta)] \leq k \cdot \delta \cdot 2^{b(n,k)}.$$

□

Using the above lemma, we prove our main theorem for this section.

Theorem 7.2.7. *Let $\mathcal{M}$ be a mechanism which enables transcript compression to $b(n,k)$ bits and also exhibits (α,β)-empirical-accuracy for k rounds given n samples. Then $\mathcal{M}$ is also $(\alpha,\beta+\delta k2^{b(n,k)},\delta)$-statistically-query-accurate, for every choice of δ.*

Proof. As $\mathcal{M}$ is (α,β)-empirically-accurate and also enables transcript compression to $b(n,k)$ bits, by Lemma 7.2.6 and the union bound

$$\begin{aligned}\Pr_{S,AG_{n,k}}\big[\,&(\exists i : |q_i(S) - q_i(\mu)| > \gamma(q,\mu,\delta))\\ &\vee (\exists i : |q_i(S) - a_i| > \alpha)\,\big]\\ \leq\ &\beta{+}\delta \cdot k \cdot 2^{b(n,k)}.\end{aligned}$$

Hence by the triangle inequality

$$\begin{aligned}&\Pr_{S,AG_{n,k}}[\exists i : |a_i - q_i(\mu)| \geq \alpha + \gamma(q,\mu,\delta)]\\ &\quad\leq \beta + \delta \cdot k \cdot 2^{b(n,k)}.\end{aligned}$$

□

Applying Theorem 7.2.7 together with the transcript-compressing mechanisms of Dwork et al. (2015a), we get the following two results.

Theorem 7.2.8. *For every α,δ, there exists an (α,β,δ)-statistically-query-accurate mechanism for k rounds given n samples, where $\beta = k \cdot \delta \cdot 2^{k\cdot\log\frac{1}{\alpha}}$. The mechanism is compu-*

tationally efficient.

Theorem 7.2.9. *For every δ, there exists an (α, β, δ)-statistically-query-accurate mechanism for k rounds given n samples, where*

$$\alpha = \mathcal{O}\left(\left(\frac{\ln k}{n}\right)^{1/4}\right) \text{ and } \beta = k \cdot \delta \cdot 2^{\tilde{\mathcal{O}}\left(\sqrt{n}\cdot\log|\mathcal{X}|\cdot(\log k)^{3/2}\right)}.$$

The mechanism is computationally inefficient.

7.3 Application to Markov Chains

In this section, we demonstrate an application of our tools and results regarding Gibbs-dependency and differential privacy to the problem of learning *Markov chains* adaptively. For our notion of dependence, it will be more convenient to analyze the *Undirected Markov Chains.* By the Hammersley-Clifford theorem (Hammersley and Clifford, 1971; Clifford, 1990), every Markov measure on a chain graph with nonzero transition probabilities can be factorized according to pairwise potential functions (formalized below), which we refer to as the *undirected Markov chain* formalization (Kontorovich, 2012).

The formal definition of an undirected Markov chain measure is as follows.

Definition 7.3.1. *A measure μ over Ω^n is an* undirected Markov chain *if there are positive functions $\{g_i\}_{i\in[n-1]}$, called* potential functions*, such that for any $x \in \Omega^n$*

$$\mu(x) = \frac{\prod_{i=1}^{n-1} g_i(x_i, x_{i+1})}{\sum_{x'\in\Omega^n}\prod_{i=1}^{n-1} g_i(x'_i, x'_{i+1})}.$$

This is a special case of the more general *undirected graphical model* (see Lauritzen (1996)). For the sake of convenience, we will use the following notations.

Definition 7.3.2. *Let μ be an undirected Markov chain with potential functions $\{g_i\}_{i\in[n-1]}$. We denote the maximal and minimal potentials as follows.*

- $R_i(\mu) = \max_{a,b\in\Omega} g_i(a, b)$,
- $r_i(\mu) = \min_{a,b\in\Omega} g_i(a, b)$,
- $R(\mu) = \max_i\{R_i(\mu)\}$,
- $r(\mu) = \min_i\{r_i(\mu)\}$,
- $\bar{R}(\mu) := \frac{R(\mu)^2 - r(\mu)^2}{R(\mu)^2 + r(\mu)^2}$.

When μ is clear from the context, we simply write $R_i, r_i, R, r, \bar{R}$ instead of $R_i(\mu)$, $r_i(\mu)$, $R(\mu)$, $r(\mu)$, $\bar{R}(\mu)$.

In order to apply our techniques to the case where the underlying distribution is an undirected Markov chain, we need to bound the Gibbs-dependency of undirected Markov chains. We first show the following lemma. (The proof of this lemma is deferred to a later part of this section.)

Lemma 7.3.3. *For every undirected Markov chain μ we have*

$$\psi(\mu) \leq \bar{R} := \frac{R^2 - r^2}{R^2 + r^2}.$$

That is, the above lemma bounds the Gibbs-dependency of undirected Markov chains as a function of the potential functions. Combining this bound with Corollary 1.3.4, we obtain the following result.

Corollary 7.3.4. *There exists a computationally efficient mechanism for answering k adaptively chosen queries with the following properties. When given $n \geq m = \tilde{O}\left(\frac{\sqrt{k}}{\alpha^2} \log \frac{1}{\beta}\right)$ samples (an n-tuple) from an (unknown) undirected Markov chain μ, the mechanism guarantees $\big(\alpha + 2\bar{R}(\mu), \beta\big)$-statistical-accuracy (w.r.t. the underlying distribution μ).*

In particular, Corollary 7.3.4 shows that if the underlying chain μ satisfies $\bar{R}(\mu) \leq \alpha$, then the dependencies in μ can be "accommodated for free", in the sense that we can efficiently answer the same amount of adaptive queries as if the underlying distribution is a product distribution. We are not aware of an alternative method for answering this amount of adaptive queries under these conditions. As we next explain, we can broaden the applicability of our techniques even further, by reducing dependencies in the data as follows. The idea is to access only a part of the chain, obtained by "skipping" a fixed number of elements between two random samples. Formally,

Definition 7.3.5 (Skipping Samples). *Given a measure μ over n-tuples, and an integer t, we define the measure $\mu_{\times t}$ over $\frac{n}{t}$-tuples as follows.*[1] *To sample from $\mu_{\times t}$, let $(x_0, x_1, x_2, x_3, \ldots, x_{n-1}) \sim \mu$, and return $(x_0, x_t, x_{2t}, x_{3t}, \ldots, x_{n-t})$.*

Intuitively, as Markov chains are "memoryless processes", skipping points in our sample (as in Definition 7.3.5), should significantly reduce dependencies within the remaining points. We formalize this intuition and prove the following theorem. (The proof of this theorem is deferred to a later part of this section.)

Theorem 7.3.6. *For every undirected Markov chain μ and for every t we have*

$$\psi(\mu_{\times t}) \leq \psi(\mu)^t.$$

That is, Theorem 7.3.6 states that by reducing our sample size *linearly* with t, we could reduce dependencies within our sample *exponentially* in t. Combining this bound with Corollary 1.3.4, we obtain the following result.

[1] We assume here for simplicity that t divides n.

Corollary 7.3.7. *There exists a computationally efficient mechanism that is $(3\alpha, \beta)$-statistically-accurate for k adaptively chosen queries, given a sample (an n-tuple) drawn from an underlying distribution $\mu_{\times t}$, where μ is an undirected Markov-chain, and where*

$$n \geq \widetilde{\mathcal{O}}\left(\frac{\log(1/\beta)\sqrt{k}}{\alpha^2}\right) \qquad \textit{and} \qquad t \geq \frac{\log(1/\alpha)}{\log(1/\bar{R})}.$$

Remark 7.3.8. *As a baseline, one can choose the "skipping parameter" t to be sufficiently big s.t. the Gibbs-dependency would drop below β/n. As we mentioned in Section 1.3, in that case the dependencies in the data would be small enough to the extent we could simply apply existing tools for answering queries* w.r.t. product distributions*, in order to answer adaptive queries w.r.t. $\mu_{\times t}$. However, this would require the skipping parameter t to be as big as $\frac{\log(n/\beta)}{\log(1/\bar{R})}$, i.e., to increase by (roughly) a* $\log(n)$ *factor, which in turn, would result in a larger sample complexity.*

We next prove Lemma 7.3.3 and Theorem 7.3.6.

Proof of Lemma 7.3.3. For any $i \in [2, n-1]$, [2] $a \in \Omega$ and $u, v \in \Omega^n$

$$\mu_i(a \mid v^{-i}) = \frac{g_{i-1}(v_{i-1}, a) g_i(a, v_{i+1})}{\sum_{a'} g_{i-1}(v_{i-1}, a') g_i(a', v_{i+1})}$$

We will be using the following lemma of Kontorovich (2012):

Lemma 7.3.9. *For $n \in \mathbb{N}$ and $0 \leq r \leq R$, consider the vectors $\alpha \in [0, \infty)^n$ and $f, g \in [r, R]^n$. Then,*

$$\frac{1}{2}\sum_{i=1}^{n} \left| \frac{\alpha_i f_i}{\sum_{j=1}^{n} \alpha_j f_j} - \frac{\alpha_i g_i}{\sum_{j=1}^{n} \alpha_j g_j} \right| \leq \frac{R-r}{R+r}.$$

We apply the lemma using

- $f_a = g_{i-1}(v_{i-1}, a) g_i(a, v_{i+1})$
- $h_a = g_{i-1}(u_{i-1}, a) g_i(a, u_{i+1})$
- $\alpha_a = 1$

and get that

$$\frac{1}{2}\sum_{a} |\mu_i(a \mid u^{-i}) - \mu_i(a \mid v^{-i})| \leq \frac{R_{i-1}R_i - r_{i-1}r_i}{R_{i-1}R_i + r_{i-1}r_i}.$$

[2] The case of $i \in \{1, n\}$ has an almost identical argument; only the $g_{i-1}(v_{i-1}, a)$ (respectively, $g_n(a, v_{i+1})$) factor is omitted. This does not affect the rest of the argument for the upper bound.

It follows that

$$
\begin{aligned}
&\|\mu_i(\cdot) - \mu_i(\cdot \mid v^{-i})\|_{\mathsf{TV}} \\
=&\frac{1}{2}\sum_a |\mu_i(a) - \mu_i(a \mid v^{-i})| \\
=&\frac{1}{2}\sum_a |\sum_{u^{-i}} \mu_i(a \mid u^{-i})\mu^{-i}(u^{-i}) - \mu_i(a \mid v^{-i})| \\
=&\frac{1}{2}\sum_a \Bigg| \sum_{u^{-i}} \mu_i(a \mid u^{-i})\mu^{-i}(u^{-i}) \\
&\quad - \sum_{u^{-i}} \mu^{-i}(u^{-i})\mu_i(a \mid v^{-i}) \Bigg| \\
=&\frac{1}{2}\sum_a |\sum_{u^{-i}} \mu^{-i}(u^{-i})\left[\mu_i(a \mid u^{-i}) - \mu_i(a \mid v^{-i})\right]| \\
\leq&\frac{1}{2}\sum_a \sum_{u^{-i}} \mu^{-i}(u^{-i})|\mu_i(a \mid u^{-i}) - \mu_i(a \mid v^{-i})| \\
=&\sum_{u^{-i}} \mu^{-i}(u^{-i})\frac{1}{2}\sum_a |\mu_i(a \mid u^{-i}) - \mu_i(a \mid v^{-i})| \\
\leq&\sum_{u^{-i}} \mu^{-i}(u^{-i})\frac{R_{i-1}R_i - r_{i-1}r_i}{R_{i-1}R_i + r_{i-1}r_i} = \frac{R_{i-1}R_i - r_{i-1}r_i}{R_{i-1}R_i + r_{i-1}r_i}.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
\psi(\mu) &= \sup_v \mathop{\mathbb{E}}_i \left\|\mu_i(\cdot) - \mu_i(\cdot \mid v^{-i})\right\|_{\mathsf{TV}} \\
&\leq \frac{R^2 - r^2}{R^2 + r^2}.
\end{aligned}
$$

□

In order to prove Theorem 7.3.6, we first establish the following notations:

- $x_{i\pm t} = x_{i-1}, x_{i+t}$
- $c_i^t := \sup_{x_{i\pm t}, y_{i\pm t}} \|\mu_{i\pm(t-1)}(\cdot \mid x_{i\pm t})$ $-\mu_{i\pm(t-1)}(\cdot \mid y_{i\pm t})\|_{\mathsf{TV}}$
- $\gamma_i^t := \sup_{x_{i\pm t}, y_{i\pm t}} \|\mu_i(\cdot \mid x_{i\pm t}) - \mu_i(\cdot \mid y_{i\pm t})\|_{\mathsf{TV}}$

Note that

$$
c_i^1 = \gamma_i^{i+1} = \sup_{x_{i\pm 1}, y_{i\pm 1}} \|\mu_i(\cdot \mid x_{i\pm 1}) - \mu_i(\cdot \mid y_{i\pm 1})\|_{\mathsf{TV}}.
$$

We will be using the following two lemmas (we prove these two lemmas after the proof of Theorem 7.3.6).

Lemma 7.3.10. $\gamma_i^t \leq \prod_{j=1}^t c_i^j$

Lemma 7.3.11. *For every t and every i, there exist some j s.t. $c_i^t \leq c_j^1$.*

We now prove Theorem 7.3.6 using Lemmas 7.3.10 and 7.3.11.

Proof of Theorem 7.3.6. Combining Lemma 7.3.10 and Lemma 7.3.11 yields that for every undirected Markov measure μ and for every t,

$$\begin{aligned}\max_i \gamma_i^t &\leq \max_i \prod_{j=1}^t c_i^j \leq \max_i \prod_{j=1}^t c_{l(j)}^1 \\ &\leq \max_i \max_l (c_l^1)^t = (\max_i c_l^1)^t = (\psi(\mu))^t,\end{aligned} \tag{7.3}$$

where the first inequality is due to Lemma 7.3.10, the second is by Lemma 7.3.11 [3]. The Last equality holds by the definitions of ψ and c_l^1. Since

$$\mu_i(\cdot) = \sum_{x_{i\pm t} \in \Omega^2} \mu_i(\cdot \mid x_{i\pm t}) \mu_{i\pm t}(x_{i\pm t}),$$

we have, by the undirected Markov property,

$$\begin{aligned}\psi(\mu_{\times t}) &= \max_i \sup_{y_{i\pm t}} \|\mu_i(\cdot) - \mu_i(\cdot \mid y_{i\pm t})\|_{\mathsf{TV}} \\ &= \max_i \sup_{y_{i\pm t}} \| \sum_{x_{i\pm t}\in\Omega^2} (\mu_i(\cdot \mid x_{i\pm t}) - \mu_i(\cdot \mid y_{i\pm t}))\mu(x_{i\pm t}) \|_{\mathsf{TV}} \\ &\leq \max_i \sup_{y_{i\pm t}} \sum_{x_{i\pm t}\in\Omega^2} \|\mu_i(\cdot \mid x_{i\pm t}) - \mu_i(\cdot \mid y_{i\pm t})\|_{\mathsf{TV}}\, \mu(x_{i\pm t}) \\ &\leq \max_i \gamma_i^t \leq (\psi(\mu))^t,\end{aligned}$$

where the last inequality is due to (7.3). □

Proof of Lemma 7.3.10. Let $x_{i\pm t}, y_{i\pm t}$ be some pairs of realization for the $i-t, i+t$ variable in the chain. By the law of total probability,

$$\begin{aligned}&\|\mu_i(\cdot \mid x_{i\pm t}) - \mu_i(\cdot \mid y_{i\pm t})\|_{\mathsf{TV}} \\ &= \| \sum_{x_{i\pm(t-1)}} \mu_i(\cdot \mid x_{i\pm(t-1)}) \mu_{i\pm(t-1)}(x_{i\pm(t-1)} \mid x_{i\pm t}) \\ &- \sum_{y_{i\pm(t-1)}} \mu_i(\cdot \mid y_{i\pm(t-1)}) \mu_{i\pm(t-1)}(y_{i\pm(t-1)} \mid y_{i\pm t}) \|_{\mathsf{TV}}.\end{aligned} \tag{7.4}$$

Define a coupling measure $\Pi_{i\pm(t-1)}(\cdot, \cdot \mid x_{i\pm t}, y_{i\pm t})$ whose marginals are $\mu_{i\pm(t-1)}(\cdot \mid x_{i\pm t})$

[3] The function $l : [n] \to [n]$ returns for every coordinate i the appropriate coordinate $l(i)$ which is guaranteed by Lemma 7.3.11 to bound it.

and $\mu_{i\pm(t-1)}(\cdot \mid y_{i\pm t})$. Then

$$\begin{aligned}
&(7.4) = \\
&\quad \| \sum_{x_{i\pm(t-1)}} \sum_{y_{i\pm(t-1)}} (\mu_i(\cdot \mid x_{i\pm(t-1)}) - \mu_i(\cdot \mid y_{i\pm(t-1)})) \\
&\quad \Pi_{i\pm(t-1)}(x_{i\pm(t-1)}, y_{i\pm(t-1)} \mid x_{i\pm t}, y_{i\pm t})\|_{\mathsf{TV}} \\
&\quad \leq \sum_{x_{i\pm(t-1)}} \sum_{y_{i\pm(t-1)}} \left\|\mu_i(\cdot \mid x_{i\pm(t-1)}) - \mu_i(\cdot \mid y_{i\pm(t-1)})\right\|_{\mathsf{TV}} \\
&\quad \Pi_{i\pm(t-1)}(x_{i\pm(t-1)}, y_{i\pm(t-1)} \mid x_{i\pm t}, y_{i\pm t}) \\
&\quad \leq \gamma_i^{t-1} \sum_{x_{i\pm(t-1)}} \sum_{y_{i\pm(t-1)}} 1_{x_{i\pm(t-1)} \neq y_{i\pm(t-1)}} \\
&\quad \Pi_{i\pm(t-1)}(x_{i\pm(t-1)}, y_{i\pm(t-1)} \mid x_{i\pm t}, y_{i\pm t}).
\end{aligned}$$

By the dual form of the total variation distance,[4] we can choose $\Pi_{i\pm(t-1)}$ to be such that

$$\begin{aligned}
&\left\|\mu_{i\pm(t-1)}(\cdot \mid x_{i\pm t}) - \mu_{i\pm(t-1)}(\cdot \mid y_{i\pm t})\right\|_{\mathsf{TV}} \\
&= \sum_{x_{i\pm(t-1)}} \sum_{y_{i\pm(t-1)}} 1_{x_{i\pm(t-1)} \neq y_{i\pm(t-1)}} \\
&\quad \Pi_{i\pm(t-1)}(x_{i\pm(t-1)}, y_{i\pm(t-1)} \mid x_{i\pm t}, y_{i\pm t})
\end{aligned}$$

and therefore

$$\begin{aligned}
&\|\mu_i(\cdot \mid x_{i\pm t}) - \mu_i(\cdot \mid y_{i\pm t})\|_{\mathsf{TV}} \\
&\leq \gamma_i^{t-1} \left\|\mu_{i\pm(t-1)}(\cdot \mid x_{i\pm t}) - \mu_{i\pm(t-1)}(\cdot \mid y_{i\pm t})\right\|_{\mathsf{TV}} \\
&\leq \gamma_i^{t-1} c_i^t.
\end{aligned}$$

Hence we get that

$$\gamma_i^t = \sup_{x_{i\pm t}, y_{i\pm t}} \|\mu_i(\cdot \mid x_{i\pm t}) - \mu_i(\cdot \mid y_{i\pm t})\|_{\mathsf{TV}} \leq \gamma_i^{t-1} c_i^t$$

and by induction we get the lemma's result. □

Proof of Lemma 7.3.11. First we will show that for any j, k the following holds

$$\begin{aligned}
&\sup \|\mu_j(\cdot \mid x_{j-1}, x_{j+k}) - \mu_j(\cdot \mid y_{j-1}, y_{j+k})\|_{\mathsf{TV}} \\
&\quad \leq \sup \|\mu_j(\cdot \mid x_{j-1}, x_{j+k-1}) - \mu_j(\cdot \mid y_{j-1}, y_{j+k-1})\|_{\mathsf{TV}}. \qquad (7.5)
\end{aligned}$$

[4] By the Kantorovich-Rubinstein duality of the specific case of total-Variation distance $\|P - Q\|_{\mathsf{TV}} = \min_{\Pi \in \Delta(P,Q)} \int_\Omega \int_\Omega 1_{x \neq y} d\Pi(x, y)$ when $\Delta(P, Q)$ is the set of all the possible coupling of P and Q.

Indeed,

$$\begin{aligned}
&\sup \|\mu_j(\cdot \mid x_{j-1}, x_{j+k}) - \mu_j(\cdot \mid y_{j-1}, y_{j+k})\|_{\mathsf{TV}} \\
&= \sup \| \sum_{x_{j+k-1}} \mu_j(\cdot \mid x_{j-1}, x_{j+k-1}) \mu_{j+k-1}(x_{j+k-1} \mid x_{j+k}) \\
&- \sum_{y_{j+k-1}} \mu_j(\cdot \mid y_{j-1}, y_{j+k-1}) \mu_{j+k-1}(y_{j+k-1} \mid y_{j+k}) \|_{\mathsf{TV}}.
\end{aligned}$$

Let $\Pi_{j+k-1}(\cdot,\cdot \mid x_{j+k}, y_{j+k})$ be a coupling distribution whose marginal distributions are $\mu_{j+k-1}(y_{j+k-1} \mid y_{j+k})$ and $\mu_{j+k-1}(x_{j+k-1} \mid x_{j+k})$, we get that the above is equal to

$$\begin{aligned}
&\sup \| \sum_{x_{j+k-1}} \sum_{y_{j+k-1}} (\mu_j(\cdot \mid x_{j-1}, x_{j+k-1}) \\
&\qquad - \mu_j(\cdot \mid y_{j-1}, y_{j+k-1})) \\
&\qquad \Pi_{j+k-1}(x_{j+k-1}, y_{j+k-1} \mid x_{j+k}, y_{j+k}) \|_{\mathsf{TV}} \\
&\leq \sup \textstyle\sum_{x_{j+k-1}} \displaystyle\sum_{y_{j+k-1}} \|\mu_j(\cdot \mid x_{j-1}, x_{j+k-1}) \\
&\qquad - \mu_j(\cdot \mid y_{j-1}, y_{j+k-1}) \|_{\mathsf{TV}} \\
&\qquad \Pi_{j+k-1}(x_{j+k-1}, y_{j+k-1} \mid x_{j+k}, y_{j+k}) \\
&\leq \sup \|\mu_j(\cdot \mid x_{j-1}, x_{j+k-1}) - \mu_j(\cdot \mid y_{j-1}, y_{j+k-1})\|_{\mathsf{TV}}.
\end{aligned}$$

Now we turn to the quantity of interest:

$$\begin{aligned}
&\sup_{x_{i\pm t}, y_{i\pm t}} \left\|\mu_{i\pm(t-1)}(\cdot \mid x_{i\pm t}) - \mu_{i\pm(t-1)}(\cdot \mid y_{i\pm t})\right\|_{\mathsf{TV}} \\
&= \sup_{x_{i\pm t}, y_{i\pm t}} \| \sum_{x_{i+t-1}} \mu_{i\pm(t-1)}(\cdot \mid x_{i\pm t}, x_{i+t-1}) \\
&\qquad \mu_{i+t-1}(x_{i+t-1} \mid x_{i\pm t}) \\
&\qquad - \sum_{y_{i+t-1}} \mu_{i\pm(t-1)}(\cdot \mid y_{i\pm t}, y_{i+t-1}) \\
&\qquad \mu_{i+t-1}(y_{i+t-1} \mid y_{i\pm t}) \|_{\mathsf{TV}} \\
&= \sup_{x_{i\pm t}, y_{i\pm t}} \| \sum_{x_{i+t-1}} \mu_{i-t+1}(\cdot \mid x_{i-t}, x_{i+t-1}) \\
&\qquad \mu_{i+t-1}(x_{i+t-1} \mid x_{i\pm t}) \\
&\qquad - \sum_{y_{i+t-1}} \mu_{i-t+1}(\cdot \mid y_{i-t}, y_{i+t-1}) \\
&\qquad \mu_{i+t-1}(y_{i+t-1} \mid y_{i\pm t}) \|_{\mathsf{TV}}. \qquad (7.6)
\end{aligned}$$

Let $\Pi_{i+t-1}(\cdot,\cdot \mid x_{i\pm t}), y_{i\pm t})$ be a coupling distribution whose marginals are $\mu_{i+t-1}(x_{i+t-1} \mid$

$x_{i\pm t})$ and $\mu_{i+t-1}(y_{i+t-1} \mid y_{i\pm t})$. Then the above is then equal to

$$\begin{aligned}
(7.6) = \sup\nolimits_{x_{i\pm t}, y_{i\pm t}} \| \sum_{x_{i+t-1}} \sum_{y_{i+t-1}} & \mu_{i-t+1}(\cdot \mid x_{i-t}, x_{i+t-1}) \\
& - \mu_{i-t+1}(\cdot \mid y_{i-t}, y_{i+t-1}) \\
& \Pi_{i+t-1}(x_{i+t-1}, y_{i+t-1} \mid x_{i\pm t}, y_{i\pm t}) \|_{\mathsf{TV}} \\
\leq \sup\nolimits_{x_{i\pm t}, y_{i\pm t}} \sum_{x_{i+t-1}} \sum_{y_{i+t-1}} & \| \mu_{i-t+1}(\cdot \mid x_{i-t}, x_{i+t-1}) \\
& - \mu_{i-t+1}(\cdot \mid y_{i-t}, y_{i+t-1}) \|_{\mathsf{TV}} \\
& \Pi_{i+t-1}(x_{i+t-1}, y_{i+t-1} \mid x_{i\pm t}, y_{i\pm t}).
\end{aligned}$$

Plugging $j = i - t + 1$ and $k = t - 2$ into (7.5) yields

$$\begin{aligned}
& \sup_{x_{i\pm t}, y_{i\pm t}} \left\| \mu_{i\pm(t-1)}(\cdot \mid x_{i\pm t}) - \mu_{i\pm(t-1)}(\cdot \mid y_{i\pm t}) \right\|_{\mathsf{TV}} \\
& \quad \leq \sup_{x_{i-t,i-t+2}, y_{i-t,i-t+2}} \| \mu_{i-t+1}(\cdot \mid x_{i-t,i-t+2}) \\
& \qquad - \mu_{i-t+1}(\cdot \mid y_{i-t,i-t+2}) \|_{\mathsf{TV}}
\end{aligned}$$

which completes the proof. □

7.4 Additional proofs

7.4.1 Product measure

We show the following claim

Claim 7.4.1. *For a measure $\mu \sim \mathcal{X}^n$, if for every $i \in [n]$ and for every possible $x \in \mathcal{X}^n$ it holds that $\mu_i = \mu_i(\cdot \mid x^{-i}$, then μ is a product measure.*

Proof. For convenience, we denote for every $i \leq j$ the following notation $a_{i:j} = a_i, \ldots, a_j$. Now, for every $a \in \mathcal{X}^n$, $\mu(a) = \prod_{i \in [n]} \mu(a_i \mid a_{1:i-1}$. For start, we show that $\mu(a_2 \mid a_1) = \mu(a_2)$. Indeed,

$$\begin{aligned}
\mu(a_2 \mid a_1) &= \sum_{a_3, \ldots, a_n} \mu(a_2 \mid a_1, a_{3:n}) \cdot \mu(a_{3:n} \mid a_1) \\
&= \sum_{a_3, \ldots, a_n} \mu(a_2) \cdot \mu(a_{3:n} \mid a_1) = \mu(a_2)
\end{aligned}$$

In the same way it can be shown that $\mu(a_3) = \mu(a_3 \mid a_{1:2})$ and so on.

□

Algorithm 11 Auxiliary Algorithm $\mathcal{A}'$

```
**Input:** $\vec{S} = (S_1, \ldots, S_T)$, where $T = \frac{\varepsilon}{\delta}$.
$F \leftarrow \emptyset$
**for** $t \in [T]$ **do**
    $(h_1^t, \ldots, h_k^t) \leftarrow \mathcal{A}(S_t)$
    $H_t \leftarrow \{(h_1^t, t), \ldots, (h_k^t, t)\}$
    $\bar{H}_t \leftarrow \{1 - h \mid h \in H_t\}$
    $F \leftarrow F \cup H_t \cup \bar{H}_t$
Sample $(h^*, t^*)$ from $F$ using the exponential mechanism. Specifically, sample $(h^*, t^*) \in F$ with probability proportional to $\exp\left(\frac{\varepsilon n}{2}\left(h^*(S_{t^*}) - h^*(\mu)\right)\right)$.
Return $(h^*, t^*)$
```

7.4.2 Proofs from Section 7.1

Proof of Theorem 7.1.2. Fix a measure μ on $\mathcal{X}^n$ with Gibbs-dependence ψ_n, and fix an (ε, δ)-differentially private algorithm that takes a sample $S \in \mathcal{X}^n$ and returns k predicates $h_1, \ldots, h_k : \mathcal{X} \to \{0, 1\}$. Assume towards contradiction that

$$\Pr_{S, \mathcal{A}(S)}\left[\max_{i \in [k]} |h_i(\mu) - h_i(S)| \geq 10\varepsilon + 2\psi\right] \geq \frac{\delta}{\varepsilon}. \tag{7.7}$$

Consider the procedure described in Algorithm 11. As differential private algorithms are immune to post-processing and by the composition theorem, $\mathcal{A}'$ is by itself $(2\varepsilon, \delta)$-differentially private. Given a multi-set $\vec{S}$ sampled from μ^T, by (7.7) we get that

$$\forall t : \Pr_{S_t, \mathcal{A}(S_t)}\left[\max_{i \in [k]} |h_i^t(\mu) - h_i^t(S_t)| \geq 10\varepsilon + 2\psi\right] \geq \frac{\delta}{\varepsilon},$$

and hence, by setting $T = \frac{\varepsilon}{\delta}$, we have that

$$\begin{aligned}&\Pr_{\vec{S}, \mathcal{A}'(\vec{S})}\left[\max_{t \in [T], i \in [k]} |h_i^t(\mu) - h_i^t(S_t)| \geq 10\varepsilon + 2\psi\right] \\ &\qquad \geq 1 - \left(1 - \frac{\delta}{\varepsilon}\right)^T \geq \frac{1}{2}.\end{aligned}$$

By Markov's inequality,

$$\mathop{\mathbb{E}}_{\vec{S}, \mathcal{A}'(\vec{S})}\left[\max_{t \in [T], i \in [k]} |h_i^t(\mu) - h_i^t(S_t)|\right] \geq 5\varepsilon + \psi.$$

Now the set constructed in the algorithm's run, F, contains also the negation of each predicate, and hence

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\vec{S},\mathcal{A}'(\vec{S})}\left[\max_{(h,t)\in F}\left\{h(S_t)-h(\mu)\right\}\right]\\
&=\mathop{\mathbb{E}}_{\vec{S},\mathcal{A}'(\vec{S})}\left[\max_{t\in[T],i\in[k]}|h_i^t(\mu)-h_i^t(S_t)|\right]\geq 5\varepsilon+\psi.
\end{aligned}$$

By the properties of the exponential mechanism (see McSherry and Talwar (2007) or Bassily et al. (2016)), denoting the output of the algorithm by (h^*,t^*) we get that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(h^*,t^*)}\left[h^*(S_{t^*})-h^*(\mu)\right]\\
&\geq\max_{(h,t)\in F}\{h^*(S_{t^*})-h^*(\mu)\}-\frac{2}{\varepsilon n}\log(2Tk).
\end{aligned}$$

Taking expectation on both sides yields

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\vec{S},\mathcal{A}'(\vec{S})}\left[h^*(S_{t^*})-h^*(\mu)\right]\\
&\geq\mathop{\mathbb{E}}_{\vec{S},\mathcal{A}'(\vec{S})}\left[\max_{(h,t)\in F}\{h^*(S_{t^*})-h^*(\mu)\}\right]-\frac{2}{\varepsilon n}\log(2Tk)\\
&\geq 5\varepsilon+\psi-\frac{2}{\varepsilon n}\log(2k\varepsilon/\delta).
\end{aligned}$$

For $n\geq\frac{\log(2k\varepsilon/\delta)}{\varepsilon^2}$, this is at least $2\varepsilon+\psi$ which contradicts Lemma 7.1.1. □

Chapter 8

Conclusion

In this thesis, we have delved into the intricate interplay between theoretical machine learning, differential privacy, and compression. Throughout our investigation, we have explored the conceptual and quantitative connections between these three domains, offering novel insights and solutions to fundamental challenges in the realm of data-driven problems. Our exploration has led us to uncover new perspectives on privacy-preserving algorithms, compression schemes, and the symbiotic relationship between learning and compression.

8.1 Summary of Contributions

Our research has spanned across multiple aspects, with each component contributing to our broader understanding of the intricate connections between differential privacy, compression, and machine learning.

We initiated our inquiry by examining the intertwinement of machine learning and compression, seeking to understand the qualitative and quantitative connections between these domains. Our work illuminated the equivalence between learning and compression in certain settings, providing another bridge between the realms of compression schemes and algorithmic learning. We extended this relationship to encompass regression problems, presenting a pioneering compressed regression result that uses minimal sample size while maintaining efficient running-time. This foundational contribution lays the groundwork for future developments in efficient, information-preserving data analysis.

Diving into the realm of privacy, we addressed the fundamental question of how much data is necessary to learn while ensuring privacy remains uncompromised. We tackled the problem of privately learning axis-aligned rectangles, a case which serves as a basic building block for a range of complex algorithms. The concept of sample complexity emerged as a pivotal point of investigation. Through innovative algorithmic design, we achieved an almost optimal trade-off, minimizing sample complexity. This achievement not only offers practical benefits in privacy-aware learning but also contributes to the broader discourse on the balance between learning efficiency and privacy guarantees.

Our exploration further led us to challenge the traditional definition of learning, advocating for a more flexible paradigm termed "Universal Learning." By questioning the pessimistic worst-case assumptions, we provided insights into aligning theoretical models with real-world data properties. This departure from the conventional approach bears the potential to bridge the gap between theoretical and practical aspects of machine learning, particularly under privacy constraints.

In the realm of adaptive data analysis, we navigated the complex landscape of inquiries emerging from evolving data accumulation processes. Bridging ideas from privacy and compression, we tackled the challenge of extending adaptive tools to encompass correlated examples. Our results showcased the feasibility of adapting privacy-based and compression-based algorithms to these intricate scenarios, thereby expanding the toolbox for reliable data analysis in adaptive settings.

8.2 Implications and Future Research

The contributions of this thesis pave the way for several exciting avenues of future research.

Firstly, the relationship between learning and compression, while extensively explored, still harbors uncharted territories. Investigating the precise limits of this connection is a fundamental problem which aligns with some long standing problems in the field. Namely, finding the optimal compression size for regression problems extends the known open question regarding the possibility of linear relation between compression size and the sample complexity of learning. One possible path might include extending and analyzing some specific well-studies classes such as *Maximal Classes* and *Duddly Classes*.

The exploration of sample complexity within a privacy-preserving context remains a fertile ground for further investigation. Investigating the trade-offs between privacy requirements and learning efficiency across a wider array of learning tasks could lead to more comprehensive and nuanced results, enabling practitioners to strike a fine balance between utility and privacy. On the particular scope of this thesis, a recent work emerged to definitively address the challenge of privately learning axis-aligned rectangles. This work introduced an optimal algorithm for the problem, culminating in a comprehensive solution that builds upon the foundation laid by our exploration.

The proposition of Universal Learning as a more flexible paradigm has significant implications for bridging the gap between theory and practice. Extending this approach to various machine learning paradigms and addressing its applicability in real-world scenarios could contribute to the development of more adaptive and resilient learning algorithms. Moreover, quantifying the precise sample complexity for fundamental problems and algorithms under this framework could provide a more nuanced understanding of the trade-offs between privacy and learning efficiency.

Lastly, adaptive data analysis, as one of the novel challenges posed in this thesis, holds untapped potential for future exploration. Investigating more complex adaptive scenarios and devising mechanisms to integrate privacy and compression in such settings can foster the development of robust and practical data analysis tools. Concurrently with our theoretical research, a paramount objective is the creation of effective algorithmic implementations in the realm of adaptive data analysis. Such implementations are pivotal in rendering these algorithms accessible to researchers and statisticians. Their availability could serve to minimize errors resulting from the improper utilization of conventional tools, and may help research areas where collecting information is a difficult and expensive job.

8.3 Closing Remarks

In conclusion, this thesis has embarked on a journey through the intricate landscapes of theoretical machine learning, differential privacy, and compression. We have explored the interplay between these domains, unraveling novel connections, and proposing innovative solutions to various challenges. As data-driven technologies continue to evolve, the insights gained from this research join a broad and cumulative effort to continue the development of efficient, privacy-preserving algorithms and drive the advancement of machine learning theory and practice. I am grateful for the privilege of being part of this research community that propels our understanding and knowledge towards a more secure, efficient, and ethical data-driven future.

Bibliography

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, oct 2016. doi: 10.1145/2976749.2978318.

Jayadev Acharya, Kallista Bonawitz, Peter Kairouz, Daniel Ramage, and Ziteng Sun. Context aware local differential privacy. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 52–62. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/acharya20a.html`.

Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997. URL `citeseer.ist.psu.edu/alon97scalesensitive.html`.

Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite littlestone dimension. In *STOC*, pages 852–860. ACM, 2019.

Noga Alon, Amos Beimel, Shay Moran, and Uri Stemmer. Closure properties for private classification and online prediction. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 119–152. PMLR, 2020.

Noga Alon, Mark Bun, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private and online learnability are equivalent. *ACM Journal of the ACM (JACM)*, 2022.

Sanghyeon An, Minjun Lee, Sanglee Park, Heerin Yang, and Jungmin So. An ensemble of simple convolutional neural network models for mnist digit recognition, 2020. URL `https://arxiv.org/abs/2008.10400`.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999. ISBN 0-521-57353-X. doi: 10.1017/CBO9780511624216. URL `http://dx.doi.org/10.1017/CBO9780511624216`.

Martin Anthony, Peter L. Bartlett, Yuval Ishai, and John Shawe-Taylor. Valid generalisation from approximate interpolation. *Combinatorics, Probability & Computing*, 5: 191–214, 1996. doi: 10.1017/S096354830000198X. URL https://doi.org/10.1017/S096354830000198X.

Aristotle. *ON THE HEAVENS: Text and Translation*, pages 47–169. Liverpool University Press, 1995. ISBN 9780856686634. URL http://www.jstor.org/stable/j.ctv1228h9n.7.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

Hassan Ashtiani, Shai Ben-David, Nicholas JA Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Journal of the ACM (JACM)*, 67(6):1–42, 2020.

Patrice Assouad. Densité et dimension. *Ann. Inst. Fourier (Grenoble)*, 33(3):233–282, 1983. ISSN 0373-0956. URL http://www.numdam.org/item?id=AIF_1983__33_3_233_0.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. In Aurélien Garivier and Satyen Kale, editors, *Algorithmic Learning Theory, ALT 2019, 22-24 March 2019, Chicago, Illinois, USA*, volume 98 of *Proceedings of Machine Learning Research*, pages 162–183. PMLR, 2019. URL http://proceedings.mlr.press/v98/attias19a.html.

Idan Attias, Edith Cohen, Moshe Shechner, and Uri Stemmer. A framework for adversarial streaming via differential privacy and difference estimators. *CoRR*, abs/2107.14527, 2021.

Ran Avnimelech and Nathan Intrator. Boosting regression estimators. *Neural computation*, 11(2):499–520, 1999.

Borja Balle, James Bell, Adrià Gascón, and Kobbi Nissim. The privacy blanket of the shuffle model. In *Annual International Cryptology Conference*, 2019.

Borja Balle, James Bell, Adrià Gascón, and Kobbi Nissim. Private summation in the multi-message shuffle model. *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020.

Borja Balle, Leonard Berrada, Soham De, Jamie Hayes, Samuel L Smith, and Robert

Stanforth. JAX-Privacy: Algorithms for privacy-preserving machine learning in jax, 2022a. URL http://github.com/deepmind/jax_privacy.

Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. *CoRR*, abs/2201.04845, 2022b. URL https://arxiv.org/abs/2201.04845.

Raef Bassily and Yoav Freund. Typicality-based stability and privacy. *CoRR*, abs/1604.03336, 2016. URL http://arxiv.org/abs/1604.03336.

Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private empirical risk minimization, revisited. *CoRR*, abs/1405.7085, 2014. URL http://arxiv.org/abs/1405.7085.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059, 2016.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Thakurta. Private stochastic convex optimization with optimal rates. In *Neural Information Processing Systems*, 2019a.

Raef Bassily, Shay Moran, and Noga Alon. Limits of private learning with access to public data. In *NeurIPS*, pages 10342–10352, 2019b.

Raef Bassily, Crist'obal Guzm'an, and Anupama Nandi. Non-euclidean differentially private stochastic convex optimization. *ArXiv*, abs/2103.01278, 2021.

Amos Beimel, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. In *Theory of Cryptography Conference*, pages 437–454. Springer, 2010.

Amos Beimel, Hai Brenner, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. *Machine Learning*, 94(3):401–437, 2014.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. *Theory of Computing*, 12(1):1–61, 2016a. doi: 10.4086/toc.2016.v012a001. URL http://www.theoryofcomputing.org/articles/v012a001.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. *Theory Comput.*, 12(1):1–61, 2016b. doi: 10.4086/toc.2016.v012a001. URL https://doi.org/10.4086/toc.2016.v012a001.

Amos Beimel, Shay Moran, Kobbi Nissim, and Uri Stemmer. Private center points and learning of halfspaces. In *COLT*, volume 99 of *Proceedings of Machine Learning Research*, pages 269–282. PMLR, 2019a.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of pure private learners. *J. Mach. Learn. Res.*, 20:146:1–146:33, 2019b. URL `http://jmlr.org/papers/v20/18-269.html`.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Learning privately with labeled and unlabeled examples. *Algorithmica*, pages 1–39, 2020.

Amos Beimel, Haim Kaplan, Yishay Mansour, Kobbi Nissim, Thatchaphol Saranurak, and Uri Stemmer. Dynamic algorithms against an adaptive adversary: Generic constructions and lower bounds. *CoRR*, abs/2111.03980, 2021.

Shai Ben-David and Ami Litman. Combinatorial variability of vapnik-chervonenkis classes with applications to sample compression schemes. *Discrete Applied Mathematics*, 86(1):3–25, 1998.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Thomas Berrett and Cristina Butucea. Classification under local differential privacy. *arXiv preprint arXiv:1912.04629*, 2019.

Alberto Bertoni, Paola Campadelli, and M Parodi. A boosting algorithm for regression. In *International Conference on Artificial Neural Networks*, pages 343–348. Springer, 1997.

Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnés, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. *CoRR*, abs/1710.00901, 2017. URL `http://arxiv.org/abs/1710.00901`.

Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In *PODS*, pages 128–138. ACM, 2005.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4): 929–965, 1989.

Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002a. URL `http://jmlr.org/papers/v2/bousquet02a.html`.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002b.

Olivier Bousquet, Steve Hanneke, Shay Moran, Ramon van Handel, and Amir Yehudayoff. A theory of universal learning. 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 532–541. ACM, 2021. doi: 10.1145/3406325.3451087. URL `https://doi.org/10.1145/3406325.3451087`.

Olivier Bousquet, Haim Kaplan, Aryeh Kontorovich, Yishay Mansour, Shay Moran, Menachem Sadigurschi, and Uri Stemmer. Differentially-private bayes consistency, 2022. URL `https://arxiv.org/abs/2212.04216`.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. *CoRR*, abs/1605.02065, 2016. URL `http://arxiv.org/abs/1605.02065`.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil P. Vadhan. Differentially private release and learning of threshold functions. In *FOCS*, pages 634–649, 2015.

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. *SIAM Journal on Computing*, 47(5):1888–1938, 2018.

Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. *Journal of Machine Learning Research*, 20(94):1–34, 2019a. URL `http://jmlr.org/papers/v20/18-549.html`.

Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. *J. Mach. Learn. Res.*, 20:94:1–94:34, 2019b. URL `http://jmlr.org/papers/v20/18-549.html`.

Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. *Journal of Machine Learning Research*, 20(94):1–34, 2019c. URL `http://jmlr.org/papers/v20/18-549.html`.

Mark Bun, Marco Leandro Carmosino, and Jessica Sorrell. Efficient, noise-tolerant, and private learning via boosting. *CoRR*, abs/2002.01100, 2020a. URL `https://arxiv.org/abs/2002.01100`.

Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In *FOCS*, pages 389–402. IEEE, 2020b.

Wray Lindsay Buntine. *A theory of learning classification rules.* PhD thesis, Citeseer.

Census Bureau. Census bureau sets key parameters to protect privacy in 2020 census results, Jan 2021a. URL `https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html`.

US Census Bureau. Federal law, Oct 2021b. URL `https://www.census.gov/about/policies/privacy/data_stewardship/federal_law.html`.

US Census Bureau. Federal law, Oct 2021c. URL `https://www.census.gov/about/policies/privacy/data_stewardship/federal_law.html`.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. *CoRR*, abs/2012.07805, 2020. URL `https://arxiv.org/abs/2012.07805`.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. *CoRR*, abs/2112.03570, 2021. URL `https://arxiv.org/abs/2112.03570`.

Kamalika Chaudhuri, Daniel J. Hsu, and Shuang Song. The large margin mechanism for differentially private maximization. In *NIPS*, pages 1287–1295, 2014.

Artem Chernikov and Pierre Simon. Externally definable sets and dependent pairs. *Israel J. Math.*, 194(1):409–425, 2013. ISSN 0021-2172. URL `https://doi.org/10.1007/s11856-012-0061-9`.

Albert Cheu. Differential privacy in the shuffle model: A survey of separations. *CoRR*, abs/2107.11839, 2021. URL `https://arxiv.org/abs/2107.11839`.

Albert Cheu, Adam D. Smith, Jonathan R. Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via mixnets. *CoRR*, abs/1808.01394, 2018. URL `http://arxiv.org/abs/1808.01394`.

Christopher A. Choquette-Choo, Florian Tramèr, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. *CoRR*, abs/2007.14321, 2020. URL `https://arxiv.org/abs/2007.14321`.

Peter Clifford. Markov random fields in statistics. In Geoffrey Grimmett and Dominic Welsh, editors, *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley*, pages 19–32. Oxford University Press, Oxford, 1990.

Edith Cohen, Haim Kaplan, Yishay Mansour, Uri Stemmer, and Eliad Tsfadia.

Differentially-private clustering of easy instances. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 2049–2059. PMLR, 2021.

David Cohn and Gerald Tesauro. Can neural networks do better than the vapnik-chervonenkis bounds? In R.P. Lippmann, J. Moody, and D. Touretzky, editors, *Advances in Neural Information Processing Systems*, volume 3. Morgan-Kaufmann, 1990a. URL `https://proceedings.neurips.cc/paper/1990/file/816b112c6105b3ebd537828a39af4818-Paper.pdf`.

David Cohn and Gerald Tesauro. How Tight Are the Vapnik-Chervonenkis Bounds? *Neural Computation*, 4(2):249–269, 03 1992a. ISSN 0899-7667. doi: 10.1162/neco.1992.4.2.249. URL `https://doi.org/10.1162/neco.1992.4.2.249`.

David A. Cohn and Gerald Tesauro. Can neural networks do better than the vapnik-chervonenkis bounds? In *NIPS*, pages 911–917. Morgan Kaufmann, 1990b.

David A. Cohn and Gerald Tesauro. How tight are the vapnik-chervonenkis bounds? *Neural Comput.*, 4(2):249–269, 1992b.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20 (3):273–297, 1995.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. doi: 10.1109/TIT.1967.1053964.

Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In *Conference on Learning Theory*, pages 772–814, 2016.

Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, and Siddhartha Jayanti. Learning from weakly dependent data under dobrushin's condition. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 914–928. PMLR, 2019. URL `http://proceedings.mlr.press/v99/dagan19a.html`.

Constantinos Daskalakis, Nishanth Dikkala, and Ioannis Panageas. Regression from dependent observations. In Moses Charikar and Edith Cohen, editors, *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 881–889. ACM, 2019. doi: 10.1145/3313276.3316362. URL `https://doi.org/10.1145/3313276.3316362`.

Ofir David, Shay Moran, and Amir Yehudayoff. Supervised learning through the lens of compression. In *Advances in Neural Information Processing Systems*, pages 2784–2792, 2016.

L. Devroye and T. Wagner. Distribution-free performance bounds with the resubstitution error estimate (corresp.). *IEEE Transactions on Information Theory*, 25(2):208–210, 1979. doi: 10.1109/TIT.1979.1056018.

Luc Devroye and László Györfi. *Nonparametric Density Estimation: The L1 View*. Wiley Interscience Series in Discrete Mathematics. Wiley, 1985. ISBN 9780471816461. URL `https://books.google.co.il/books?id=ZVALbrjGpCoC`.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, page 202–210, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581136706. doi: 10.1145/773153.773173. URL `https://doi.org/10.1145/773153.773173`.

Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy. *CoRR*, abs/1905.02383, 2019. URL `http://arxiv.org/abs/1905.02383`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL `https://arxiv.org/abs/2010.11929`.

Harris Drucker. Improving regressors using boosting techniques. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 107–115, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. ISBN 1-55860-486-3. URL `http://dl.acm.org/citation.cfm?id=645526.657132`.

Nigel Duffy and David Helmbold. Boosting methods for regression. *Machine Learning*, 47:153–200, 2002. ISSN 0885-6125.

Brian Duignan. Occam's razor, 1998.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Serge Vaudenay, editor, *Advances in Cryptology - EUROCRYPT 2006*, pages 486–503, Berlin, Heidelberg, 2006a. Springer Berlin Heidelberg. ISBN 978-3-540-34547-3.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006b.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference*, pages 265–284, 2006c.

Cynthia Dwork, Moni Naor, Omer Reingold, Guy N Rothblum, and Salil Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 381–390, 2009.

Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60, 2010a. doi: 10.1109/FOCS.2010.12.

Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and Differential Privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60, Las Vegas, NV, USA, October 2010b. IEEE. ISBN 978-1-4244-8525-3. doi: 10.1109/FOCS.2010.12. URL `http://ieeexplore.ieee.org/document/5670947/`.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015b.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and

Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126, 2015c.

Úlfar Erlingsson, Aleksandra Korolova, and Vasyl Pihur. RAPPOR: randomized aggregatable privacy-preserving ordinal response. *CoRR*, abs/1407.6981, 2014a. URL `http://arxiv.org/abs/1407.6981`.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014b.

Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. *CoRR*, abs/1811.12469, 2018. URL `http://arxiv.org/abs/1811.12469`.

Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, page 211–222, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581136706. doi: 10.1145/773153.773174. URL `https://doi.org/10.1145/773153.773174`.

Vitaly Feldman and Jan Vondrák. Generalization bounds for uniformly stable algorithms. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 9770–9780, 2018. URL `http://papers.nips.cc/paper/8182-generalization-bounds-for-uniformly-stable-algorithms`.

Vitaly Feldman and Jan Vondrák. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 1270–1279. PMLR, 2019. URL `http://proceedings.mlr.press/v99/feldman19a.html`.

Vitaly Feldman and David Xiao. Sample complexity bounds on differentially private learning via communication complexity. *SIAM J. Comput.*, 44(6):1740–1764, 2015.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization:

optimal rates in linear time. *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, 2020a.

Vitaly Feldman, Audra McMillan, and Kunal Talwar. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 954–964, 2020b.

Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.

Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3):238–247, 1989.

Mary Flahive and Bella Bose. Balancing cyclic r-ary gray codes. *the electronic journal of combinatorics*, 14(1):R31, 2007.

Sally Floyd. Space-bounded learning and the vapnik-chervonenkis dimension. In *Proceedings of the second annual workshop on Computational learning theory*, pages 349–364. Morgan Kaufmann Publishers Inc., 1989.

Sally Floyd and Manfred K. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

Yoav Freund. Boosting a weak learning algorithm by majority. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, COLT '90, page 202–216, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc. ISBN 1558601465.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997. ISSN 0022-0000. doi: http://dx.doi.org/10.1006/jcss.1997.1504.

Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. *Ann. Statist.*, 29(5):1189–1232, 2001. ISSN 0090-5364. URL `https://doi.org/10.1214/aos/1013203451`.

Badih Ghazi, Noah Golowich, Ravi Kumar, and Pasin Manurangsi. Sample-efficient proper PAC learning with approximate differential privacy. *CoRR*, abs/2012.03893, 2020. URL `https://arxiv.org/abs/2012.03893`.

Ned Glick. Sample-based multinomial classification. *Biometrics*, 29(2):241–256, 1973. ISSN 0006341X, 15410420. URL `http://www.jstor.org/stable/2529389`.

Noah Golowich and Roi Livni. Littlestone classes are privately online learnable. *CoRR*, abs/2106.13513, 2021. URL `https://arxiv.org/abs/2106.13513`.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014. URL `https://arxiv.org/abs/1412.6572`.

Louis Gordon and Richard A. Olshen. Asymptotically efficient solutions to the classification problem. *The Annals of Statistics*, 6(3):515–533, 1978. ISSN 00905364. URL `http://www.jstor.org/stable/2958556`.

Louis Gordon and Richard A. Olshen. Consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 10(4):611–627, 1980.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. *CoRR*, abs/1306.2547, 2013. URL `http://arxiv.org/abs/1306.2547`.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors. In *Advances in Neural Information Processing Systems*, pages 370–378, 2014.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Nearly optimal classification for semimetrics. In *Artificial Intelligence and Statistics (AISTATS)*, 2016.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate lipschitz extension. *IEEE Trans. Information Theory*, 63(8):4838–4849, 2017a. doi: 10.1109/TIT.2017.2713820. URL `https://doi.org/10.1109/TIT.2017.2713820`.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate lipschitz extension. *IEEE Transactions on Information Theory*, 63(8):4838–4849, Aug 2017b. ISSN 0018-9448. doi: 10.1109/TIT.2017.2713820.

Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. PAC-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

Bernard G Greenberg, Abdel-Latif A Abul-Ela, Walt R Simmons, and Daniel G Horvitz. The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, 64(326):520–539, 1969.

Chuan Guo, Brian Karrer, Kamalika Chaudhuri, and Laurens van der Maaten. Bounding training data reconstruction in private (deep) learning. *arXiv preprint arXiv:2201.12383*, 2022.

László Györfi and Martin Kroll. On rate optimal private regression under local differential privacy. *arXiv preprint arXiv:2206.00114*, 2022.

László Györfi and Martin Kroll. Multivariate density estimation from privatised data: universal consistency and minimax rates. *Journal of Nonparametric Statistics*, pages 1–23, 2023.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis of online and differentially private learning. In *NeurIPS*, 2020.

J. M. Hammersley and P. E. Clifford. Markov random fields on finite graphs and lattices. Unpublished manuscript, 1971.

Samuel Haney, William Sexton, Ashwin Machanavajjhala, Michael Hay, and Gerome Miklau. Differentially private algorithms for 2020 census detailed DHC race \& ethnicity. *CoRR*, abs/2107.10659, 2021. URL `https://arxiv.org/abs/2107.10659`.

Steve Hanneke, Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Universal bayes consistency in metric spaces. *CoRR*, abs/1906.09855, 2019a. URL `http://arxiv.org/abs/1906.09855`.

Steve Hanneke, Aryeh Kontorovich, and Menachem Sadigurschi. Sample compression for real-valued learners. In Aurélien Garivier and Satyen Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 466–488. PMLR, 22–24 Mar 2019b. URL `https://proceedings.mlr.press/v98/hanneke19a.html`.

Steve Hanneke, Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Universal Bayes consistency in metric spaces. *The Annals of Statistics*, 49(4):2129 – 2150, 2021. doi: 10.1214/20-AOS2029. URL `https://doi.org/10.1214/20-AOS2029`.

Steve Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Awni Hannun, Chuan Guo, and Laurens van der Maaten. Measuring data leakage in machine-learning models with fisher information. In *Uncertainty in Artificial Intelligence*, pages 760–770. PMLR, 2021.

Moritz Hardt and Guy N Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 61–70. IEEE, 2010.

Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 454–463. IEEE, 2014.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `http://proceedings.mlr.press/v48/hardt16.html`.

Avinatan Hassidim, Haim Kaplan, Yishay Mansour, Yossi Matias, and Uri Stemmer. Adversarially robust streaming algorithms via differential privacy. In *NeurIPS*, 2020.

David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992. doi: 10.1016/0890-5401(92)90010-D. URL `http://dx.doi.org/10.1016/0890-5401(92)90010-D`.

David Helmbold, Robert Sloan, and Manfred K Warmuth. Learning integer lattices. *SIAM Journal on Computing*, 21(2):240–266, 1992.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, 58:13–30, 1963.

Naoise Holohan, Stefano Braghin, Pól Mac Aonghusa, and Killian Levacher. Diffprivlib: the IBM differential privacy library. *ArXiv e-prints*, 1907.02444 [cs.CR], July 2019.

Yongchao Hou, Xiaofang Xia, Hui Li, Jiangtao Cui, and Abbas Mardani. Fuzzy differential privacy theory and its applications in subgraph counting. *IEEE Transactions on Fuzzy Systems*, pages 1–1, 2022. doi: 10.1109/TFUZZ.2022.3157385.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

Zhiyi Huang and Jinyan Liu. Optimal differentially private algorithms for k-means clustering. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, SIGMOD/PODS '18, page 395–408, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450347068. doi: 10.1145/3196959.3196977. URL `https://doi.org/10.1145/3196959.3196977`.

Roger Iyengar, Joseph P. Near, Dawn Xiaodong Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. *2019 IEEE Symposium on Security and Privacy (SP)*, pages 299–316, 2019.

Adel Javanmard and Andrea Montanari. On online control of false discovery rate. *arXiv preprint arXiv:1502.06197*, 2015.

Bargav Jayaraman and David E. Evans. Evaluating differentially private machine learning in practice. In *USENIX Security Symposium*, 2019.

Biing Hwang Juang and Laurence R Rabiner. Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. In *ITCS*, volume 151 of *LIPIcs*, pages 31:1–31:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning*, 2021.

Daniel M. Kane, Roi Livni, Shay Moran, and Amir Yehudayoff. On communication complexity of classification problems. *CoRR*, abs/1711.05893, 2017. URL `http://arxiv.org/abs/1711.05893`.

Haim Kaplan, Yishay Mansour, Yossi Matias, and Uri Stemmer. Differentially private learning of geometric concepts. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 3233–3241. PMLR, 2019.

Haim Kaplan, Katrina Ligett, Yishay Mansour, Moni Naor, and Uri Stemmer. Privately learning thresholds: Closing the exponential gap. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 2263–2285. PMLR, 2020a.

Haim Kaplan, Yishay Mansour, Uri Stemmer, and Eliad Tsfadia. Private learning of halfspaces: Simplifying the construction and reducing the sample complexity. In *NeurIPS*, 2020b.

Haim Kaplan, Yishay Mansour, Kobbi Nissim, and Uri Stemmer. Separating adaptive streaming from oblivious streaming using the bounded storage model. In *CRYPTO (3)*, volume 12827 of *Lecture Notes in Computer Science*, pages 94–121. Springer, 2021.

Grigoris Karakoulas and John Shawe-Taylor. Towards a strategy for boosting regressors. In Alexander J. Smola, Peter L. Bartlett, and Schölkopf, editors, *Advances in Large Margin Classifiers*, Advances in Neural Information Processing Systems, pages 43–54. MIT Press, Cambridge, MA, USA, 2000. ISBN 0-262-19448-1.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, 2011.

M. Kearns. Thoughts on hypothesis boosting. Unpublished, December 1988.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/kearns18a.html`.

Michael J. Kearns and Leslie G. Valiant. Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM (JACM)*, 41(1):67–95, 1994.

Micheal Kearns and Umesh Vazirani. *An Introduction to Computational Learning Theory.* The MIT Press, 1997.

Balázs Kégl. Robust regression by boosting the median. In *Learning Theory and Kernel Machines*, pages 258–272. Springer, 2003.

Daniel Kifer, Adam D. Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Annual Conference Computational Learning Theory*, 2012.

Pieter Kleer and Hans Simon. Primal and dual combinatorial dimensions. *CoRR*, abs/2108.10037, 2021. URL `https://arxiv.org/abs/2108.10037`.

Aryeh Kontorovich. Obtaining measure concentration from Markov contraction. *Markov Processes and Related Fields*, 4:613–638, 2012.

Aryeh Kontorovich and Maxim Raginsky. Concentration of measure without independence: a unified approach via the martingale method. In *Convexity and Concentration*, pages 183–210. Springer, 2017.

Aryeh Kontorovich and Roi Weiss. Uniform Chernoff and Dvoretzky-Kiefer-Wolfowitz-type inequalities for Markov chains and related processes. *Journal of Applied Probability*, 51(4):1100 – 1113, 2014. doi: jap/1421763330. URL `https://doi.org/`.

Aryeh Kontorovich, Menachem Sadigurschi, and Uri Stemmer. Adaptive data analysis with correlated observations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11483–11498. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/kontorovich22a.html`.

Leonid (Aryeh) Kontorovich and Kavita Ramanan. Concentration Inequalities for Dependent Random Variables via the Martingale Method. *Ann. Probab.*, 36(6):2126–2158, 2008.

Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth empirical risk minimization and stochastic convex optimization in subquadratic steps. *ArXiv*, abs/2103.15352, 2021.

Alexey Kurakin, Steve Chien, Shuang Song, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy. *CoRR*, abs/2201.12328, 2022. URL `https://arxiv.org/abs/2201.12328`.

Dima Kuzmin and Manfred K. Warmuth. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 8:2047–2081, 2007. URL `http://dl.acm.org/citation.cfm?id=1314566`.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Ang Li and Rina Foygel Barber. Accumulation tests for fdr control in ordered hypothesis testing. *Journal of the American Statistical Association*, 112(518):837–849, 2017.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. Technical report, 1986.

Daogao Liu and Zhou Lu. Lower bounds for differentially private erm: Unconstrained and non-euclidean. 2021.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

Roi Livni and Pierre Simon. Honest compressions and their application to compression schemes. In *Conference on Learning Theory*, pages 77–92, 2013.

Philip M. Long. Efficient algorithms for learning functions with bounded variation. *Inf. Comput.*, 188(1):99–115, 2004. doi: 10.1016/S0890-5401(03)00164-0. URL `https://doi.org/10.1016/S0890-5401(03)00164-0`.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian

Vladu. Towards deep learning models resistant to adversarial attacks, 2017. URL `https://arxiv.org/abs/1706.06083`.

Shie Mannor and Ron Meir. On the existence of linear weak learners and applications to boosting. *Machine Learning*, 48(1-3):219–251, 2002. doi: 10.1023/A:1013959922467. URL `https://doi.org/10.1023/A:1013959922467`.

K. Marton. A measure concentration inequality for contracting markov chains. *Geometric & Functional Analysis GAFA*, 6(3):556–571, May 1996. ISSN 1420-8970. doi: 10.1007/BF02249263. URL `https://doi.org/10.1007/BF02249263`.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 512–518, Cambridge, MA, USA, 1999. MIT Press. URL `http://dl.acm.org/citation.cfm?id=3009657.3009730`.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Invent. Math.*, 152(1):37–55, 2003. ISSN 0020-9910. doi: 10.1007/s00222-002-0266-3. URL `http://dx.doi.org/10.1007/s00222-002-0266-3`.

João Mendes-Moreira, Carlos Soares, Alípio Mário Jorge, and Jorge Freire De Sousa. Ensemble approaches for regression: A survey. *ACM Comput. Surv.*, 45(1):10:1–10:40, December 2012. ISSN 0360-0300. doi: 10.1145/2379776.2379786. URL `http://doi.acm.org/10.1145/2379776.2379786`.

Ilya Mironov. Renyi differential privacy. *CoRR*, abs/1702.07476, 2017. URL `http://arxiv.org/abs/1702.07476`.

Omar Montasser, Steve Hanneke, and Nathan Srebro. VC classes are adversarially robustly learnable, but only improperly. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 2512–2530. PMLR, 2019a. URL `http://proceedings.mlr.press/v99/montasser19a.html`.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2512–2530. PMLR, 25–28 Jun 2019b. URL `https://proceedings.mlr.press/v99/montasser19a.html`.

Shay Moran and Amir Yehudayoff. Sample compression schemes for vc classes. *Journal of the ACM (JACM)*, 63(3):1–10, 2016.

Shay Moran, Amir Shpilka, Avi Wigderson, and Amir Yehudayoff. Teaching and compressing for low vc-dimension. In *A Journey Through Discrete Mathematics*, pages 633–656. Springer, 2017.

Takao Murakami and Yusuke Kawamoto. {Utility-Optimized} local differential privacy mechanisms for distribution estimation. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1877–1894, 2019.

Huy Le Nguyen, Jonathan R. Ullman, and Lydia Zakynthinou. Efficient private algorithms for learning large-margin halfspaces. In *ALT*, volume 117 of *Proceedings of Machine Learning Research*, pages 704–724. PMLR, 2020.

Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, pages 75–84. ACM, 2007.

Richard Nock and Frank Nielsen. A real generalization of discrete adaboost. *Artificial Intelligence*, 171(1):25 – 41, 2007. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2006.10.014. URL `http://www.sciencedirect.com/science/article/pii/S0004370206001111`.

David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *2016 IEEE Information Theory Workshop, ITW 2016, Cambridge, United Kingdom, September 11-14, 2016*, pages 26–30. IEEE, 2016. doi: 10.1109/ITW.2016.7606789. URL `https://doi.org/10.1109/ITW.2016.7606789`.

Ryan M. Rogers, Aaron Roth, Adam D. Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In Irit Dinur, editor, *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 487–494. IEEE Computer Society, 2016. doi: 10.1109/FOCS.2016.59. URL `https://doi.org/10.1109/FOCS.2016.59`.

Benjamin I. P. Rubinstein and J. Hyam Rubinstein. A geometric approach to sample compression. *Journal of Machine Learning Research*, 13:1221–1261, 2012. URL `http://dl.acm.org/citation.cfm?id=2343686`.

Benjamin I. P. Rubinstein, Peter L. Bartlett, and J. Hyam Rubinstein. Shifting: One-inclusion mistake bounds and sample compression. *J. Comput. Syst. Sci.*, 75(1):37–59, 2009. doi: 10.1016/j.jcss.2008.07.005. URL `https://doi.org/10.1016/j.jcss.2008.07.005`.

M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *Ann. of Math. (2)*, 164(2):603–648, 2006. ISSN 0003-486X. URL `https://doi.org/10.4007/annals.2006.164.603`.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1232–1240, Cadiz, Spain, 09–11 May 2016. PMLR. URL `http://proceedings.mlr.press/v51/russo16.html`.

Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

Bruce Sacerdote. Peer Effects with Random Assignment: Results for Dartmouth Roommates*. *The Quarterly Journal of Economics*, 116(2):681–704, 05 2001. ISSN 0033-5533. doi: 10.1162/00335530151144131. URL `https://doi.org/10.1162/00335530151144131`.

Menachem Sadigurschi and Uri Stemmer. On the sample complexity of privately learning axis-aligned rectangles. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28286–28297. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/ee0e95249268b86ff2053bef214bfeda-Paper.pdf`.

Wendy E Sarrett and Michael J Pazzani. Average case analysis of empirical and explanation-based learning algorithms. 1989.

Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Statist.*, 26(5):1651–1686, 1998. ISSN 0090-5364. doi: 10.1214/aos/1024691352. URL `http://dx.doi.org/10.1214/aos/1024691352`.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Juliet Popper Shaffer. Multiple hypothesis testing. *Annual review of psychology*, 46(1): 561–584, 1995.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, USA, 2014. ISBN 978-1-107-05713-5.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670, 2010. URL `http://portal.acm.org/citation.cfm?id=1953019`.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11: 2635–2670, 2010.

Moshe Shechner, Or Sheffet, and Uri Stemmer. Private k-means clustering with stability assumptions. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 2518–2528. PMLR, 2020.

Moshe Shenfeld and Katrina Ligett. A necessary and sufficient stability notion for adaptive generalization. In *NeurIPS*, pages 11481–11490, 2019.

Moshe Shenfeld and Katrina Ligett. Generalization in the face of adaptivity: A bayesian perspective. *CoRR*, abs/2106.10761, 2021.

Reza Shokri, Marco Stronati, and Vitaly Shmatikov. Membership inference attacks against machine learning models. *CoRR*, abs/1610.05820, 2016. URL `http://arxiv.org/abs/1610.05820`.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

Hans Ulrich Simon. Bounds on the number of examples needed for learning functions. *SIAM J. Comput.*, 26(3):751–763, 1997. doi: 10.1137/S0097539793259185. URL `https://doi.org/10.1137/S0097539793259185`.

Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248, 2013. doi: 10.1109/GlobalSIP.2013.6736861.

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. URL http://proceedings.mlr.press/v125/steinke20a.html.

Charles J Stone. Consistent nonparametric regression. *The annals of statistics*, pages 595–620, 1977.

John D Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *The annals of statistics*, 31(6):2013–2035, 2003.

Dang Thanh, Prasath Surya, et al. A review on ct and x-ray images denoising methods. *Informatica*, 43(2), 2019.

Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, and Uri Stemmer. Friendlycore: Practical differentially private aggregation. *CoRR*, abs/2110.10132, 2021.

Jonathan Ullman, Adam Smith, Kobbi Nissim, Uri Stemmer, and Thomas Steinke. The limits of post-selection generalization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6400–6409. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7876-the-limits-of-post-selection-generalization.pdf.

Salil Vadhan. The Complexity of Differential Privacy. In Yehuda Lindell, editor, *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer International Publishing, Cham, 2017. ISBN 978-3-319-57047-1 978-3-319-57048-8. doi: 10.1007/978-3-319-57048-8_7. URL http://link.springer.com/10.1007/978-3-319-57048-8_7. Series Title: Information Security and Cryptography.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Tim van Erven and Peter Harremoës. Rényi divergence and kullback-leibler divergence. *CoRR*, abs/1206.2459, 2012. URL http://arxiv.org/abs/1206.2459.

V. N. Vapnik and A. Ja. Červonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.*, 16:264–279, 1971. ISSN 0040-361x.

V. N. Vapnik and A. Ya. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, January 1971. ISSN 0040-585X. doi: 10.1137/1116025. URL https://epubs.siam.org/doi/10.1137/1116025. Publisher: Society for Industrial and Applied Mathematics.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. ISBN 9781108415194. URL `https://books.google.co.il/books?id=J-VjswEACAAJ`.

Bao Wang, Quanquan Gu, March Tian Boedihardjo, Farzin Barekat, and S. Osher. Dp-lssgd: A stochastic optimization method to lift the utility in privacy-preserving erm. In *Mathematical and Scientific Machine Learning*, 2019.

Yining Wang, Yu-Xiang Wang, and Aarti Singh. Differentially private subspace clustering. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL `https://proceedings.neurips.cc/paper/2015/file/051e4e127b92f5d98d3c79b195f2b291-Paper.pdf`.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Wei Wei, Jianhui Wang, Na Li, and Shengwei Mei. Optimal power flow of radial networks and its variations: A sequential convex optimization approach. *IEEE Transactions on Smart Grid*, 8(6):2974–2987, 2017.

Geoffrey Wolfer and Aryeh Kontorovich. Minimax learning of ergodic markov chains. In Aurélien Garivier and Satyen Kale, editors, *Algorithmic Learning Theory, ALT 2019, 22-24 March 2019, Chicago, Illinois, USA*, volume 98 of *Proceedings of Machine Learning Research*, pages 903–929. PMLR, 2019. URL `http://proceedings.mlr.press/v98/wolfer19a.html`.

Xi Wu, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey F. Naughton. Differentially private stochastic gradient descent for in-rdbms analytics. *ArXiv*, abs/1606.04722, 2016.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2524–2533. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/6846-information-theoretic-analysis-of-generalization-capability-of-learning-algori pdf`.